Evaluating Glyph Binarizations Based on Their Properties

Shira Faigenbaum[§], Arie Shaus[§], Barak Sober[§], Eli Turkel
The Department of Applied Mathematics
Tel Aviv University
Tel Aviv 69978, Israel
+972-3-640-6024
alecsan1@post.tau.ac.il, ashaus@post.tau.ac.il,
baraksov@post.tau.ac.il, turkel@post.tau.ac.il
§ These authors contributed equally to this work.

Eli Piasetzky
The Sackler School of
Physics and Astronomy
Tel Aviv University
Tel Aviv 69978, Israel
+972-3-640-9428
eip@tauphy.tau.ac.il

ABSTRACT
Document binary images, created by different algorithms, are commonly evaluated based on a pre-existing ground truth. Previous research found several pitfalls in this methodology and suggested various approaches addressing the issue. This article proposes an alternative binarization quality evaluation solution for binarized glyphs, circumventing the ground truth. Our method relies on intrinsic properties of binarized glyphs. The features used for quality assessment are stroke width consistency, presence of small connected components (stains), edge noise, and the average edge curvature. Linear and tree-based combinations of these features are also considered. The new methodology is tested and shown to be nearly as sound as human experts' judgments.

Categories and Subject Descriptors
I.7.5 [**Document Capture**]: Document analysis

General Terms
Algorithms, Measurement, Performance, Experimentation.

Keywords
Binarization, glyph, evaluation, quality measure, ground truth.

1. INTRODUCTION
The plethora of available binarization algorithms results in different outputs for the same document image. The ensuing need for comparing binarizations, gives rise to the existing ground truth-based (GT) evaluation methodology [1-3]. The evaluation is based on a manual GT creation, and on various GT-versus-binarization measures (e.g., F-measure, PSNR, Distance Reciprocal Distortion, Misclassification Penalty, etc.). Several recent papers [4-6] performed a detailed analysis of this approach, stressing its inherent weaknesses such as subjectivity and the inherent inconsistency within the GT creation process. Among the alternative solutions suggested, are skeleton-based GT variants (maintaining some degree of human intervention) [7-8], automatic GT creation (via another binarization procedure) [9], creation of synthetic document images out of existing GT (applicable if noise model exists) [10-11] and goal-directed approach, e.g. assessing OCR results (applicable if an OCR engine is available) [12]. Trier and Taxt [13] proposed a method somewhat reminiscent of the one specified herein, yet it was performed manually upon visual inspection of binarizations.

DocEng'13, September 10–13, 2013, Florence, Italy.

This article provides an approach which eliminates the need for GT. The document binarizations are judged automatically, based on the intrinsic properties of their glyphs. 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, dependent on the task under consideration (human epigraphic analysis, alphabet reconstruction, OCR, etc.).

The purpose of this study is to provide the best available binary image on a glyph scale. The challenging problem of glyph regions extraction, along with its related topics of concern such as broken strokes and touching characters, is outside the scope of this article (the papers [14-16] deal with some of these issues).

2. SUGGESTED GLYPH MEASURES
2.1 Measures Definitions
We start by defining independent binarization quality measures, correlating to common human perception. Four measures, pertaining to different aspects of binarized images, are proposed and formalized. We will work on small binarized images, each containing a single glyph. This can be an outcome of any segmentation algorithm, such as [14-16]. The foreground (valued at 0) and the background (valued at 255) will be denoted respectively as F and B, with $p=(x,y)$ a pixel coordinate.

2.1.1 Stroke width consistency
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. 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 [17], 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 as $seg(p,\alpha)$.
- Define $SW(p) = \min_{\alpha} \left\| seg(p,\alpha) \right\|_2$.

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

- Calculate directional derivatives $G_x(p)$ and $G_y(p)$.
- Define the gradient magnitude with respect to L_∞ norm:

$$G(p) = \max(|G_x(p)|, |G_y(p)|)$$

Step 3 – Apply the measure: $M_{SWC} = \underset{p \in F}{mean}(G(p))$

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

2.1.2 Stains proportion

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 $\|...\|$. 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: $SCC = \{cc_i \in CC \mid \|cc_i\| \le T\}$. Throughout our experiments, the value of T was set to 0.5% of the glyph image size.

The measure definition is: $M_{SP} = \sum_{cc_j \in SCC} \|cc_j\| \Big/ \sum_{cc_i \in CC} \|cc_i\|$

2.1.3 Average edge curvature

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):

$$\kappa = \left|\frac{dT}{ds}\right| = \left|\frac{d\theta}{ds}\right| \cong \left|\frac{\Delta\theta}{\Delta s}\right| \quad (1)$$

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:

$$E = F \setminus erosion(F) \quad (2)$$

Step 2 – Calculate 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 $v_2(p)$:

$$\psi(p) = \arccos\langle v_1(p), v_2(p)\rangle \quad (3)$$

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

$$\Delta\theta(p) = \pi - \psi(p) \quad (4)$$

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

Step 3 – Approximate the local curvature:

$$\kappa(p) \cong \Delta\theta(p)/\Delta s(p) \quad (5)$$

Step 4 – Apply the measure:

$$M_{AEC} = \underset{p \in E}{mean}(\kappa(p)) = \sum_{p \in E} \Delta s(p)\kappa(p) \Big/ \sum_{p \in E} \Delta s(p) \quad (6)$$

Note that the following also holds:

$$M_{AEC} = \pi - \underset{p \in E}{mean}\left(\arccos\langle v_1(p), v_2(p)\rangle\right) \quad (7)$$

It should be 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.

2.1.4 Edge noise proportion

Another suggested property is the presence of typical edge noise, which often correlates with the overall quality of the binarization. The paper [18] suggests a procedure involving 12 different convolution kernels, approximating the amount of such noise. Below, we suggest a simplified method, involving 4-connectivity morphological operations.

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

$$\bar{E} = dilation(F) \setminus erosion(F) \quad (8)$$

Step 2 – Calculate a noise estimate (cl=closure, op=opening):

$$N = \left(cl(F) \setminus F\right) \cup \left(F \setminus op(F)\right) = cl(F) \setminus op(F) \quad (9)$$

The closure attaches isolated B pixels to F, while the opening performs a dual operation. N provides a set of all isolated pixels.

Step 3 – Apply the measure: $M_{ENP} = \|N\| / \|\bar{E}\|$

2.1.5 Monochromatic binarizations

In general, undesirable scenarios of an almost completely black or white binarization (e.g. due to illumination conditions) should also be addressed for all four measures. Accordingly, cases where an insufficient number of either F or B pixels exist, were detected and handled in the following fashion. Assuming 4-connectivity, if a double-dilation of F left no B pixels, or if a double-erosion of F left no F pixels, all the measures were set to $Inf = 32768$.

2.2 Measure Combinations

The measures presented above can be applied on their own right, each assessing a different glyph characteristic. In fact, in certain settings, we have seen some of them (in particular M_{ENP}) 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, glyph 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 section.

3. EXPERIMENTAL SECTION

3.1 Motivation and Data Set

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.

Our database consisted of segmented glyphs, along with their binarizations. We used glyphs originating from two different First Temple Period Hebrew inscriptions: 50 images (glyphs) were taken from Arad #1 [19], while 47 images (glyphs) were obtained from Lachish #3 [20]. The segmentation into individual characters was performed via algorithm [16]. The state of preservation of these ink-over clay samples was poor, presenting a challenge for our methodology.

The 9 binarizations in use were: Otsu [21], Bernsen [22] with window sizes (in pixels) of $w=50$ and $w=200$, Niblack [23] with $w=50$ and $w=200$, Sauvola [24] with $w=50$ and $w=200$, as well as our own binarization [16] with or without unspeckle stage.

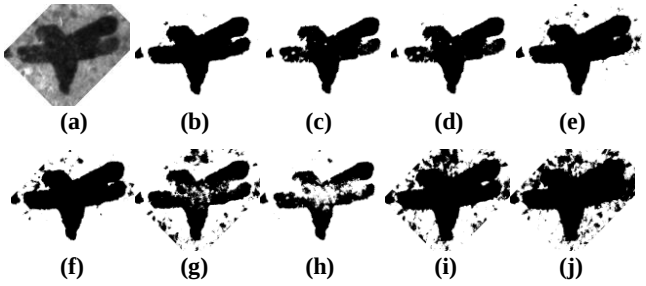

Figure 1. Expert's ranking of one glyph, in decreasing quality order. (a) Original image, (b) Sauvola $w=200$, (c) Shaus et al. [16] inc. unspeckle stage, (d) Shaus et al., (e) Otsu, (f) Niblack $w=200$, (g) Niblack $w=50$, (h) Sauvola $w=50$, (i) Bernsen $w=50$, (j) Bernsen $w=200$.

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. 1.

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

3.2 Ranking Prediction

The experiment attempted at creating a model matching the three experts' ranking. The model types under consideration were linear and tree-based regressions [25]. These models used the 4 *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 necessitate the prerequisites specified in the next sub-section.

3.2.1 Prerequisites

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 873, containing rankings of binarization blocks in a stacked manner, 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, $R_{mean} = mean(R_1, R_2, R_3)$, was calculated (coordinate-wise), possibly containing non-integer values. Then, a re-ranking of R_{mean} enforced scores of 1…9 within each binarization block, resulting in $R_{experts}$. Such process is denoted below as "re-ranking procedure". In addition, 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}.

Model score:

A model m is scored in the following fashion. A prediction produced by the model is re-ranked, resulting in R_m, which is then compared with the experts ranking via standard linear (cor) or Kendall (τ) [26] correlations:

$$c_m = \min_{i=1..3}\left(cor(R_i, R_m)\right),\ \tau_m = \min_{i=1..3}\left(\tau(R_i, R_m)\right)$$

3.2.2 Model selection stage

Model specifications:

Both linear and tree-based regression models were considered. 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*). 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 [25], as well as a "forced" tree with 9 leaves). This resulted in a total of *30 tree models* under consideration.

Selection procedure:

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.

Success criteria:

Since even human experts differ in their judgments, we do not expect the best model to perform flawlessly, but in a "human-like" fashion. Our golden standards are the *minimal* correlations between pairs of human experts, denoted as c_{expert} and τ_{expert}. Hence, our optimal model is expected to adhere to:

$$c_m \le 0.8 \cdot \min_{1\le i<j\le 3}\left(cor(R_i, R_j)\right) = 0.8 \cdot c_{expert} \quad (10)$$

$$\tau_m \le 0.8 \cdot \min_{1\le i<j\le 3}\left(\tau(R_i, R_j)\right) = 0.8 \cdot \tau_{expert} \quad (11)$$

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 $\tau_m = 0.543$. Since $c_{expert} = 0.768$ and $\tau_{expert} = 0.634$, the criteria was met.

3.2.3 Model verification stage

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 glyph 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 pvalue=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.

4. SUMMARY AND FUTURE RESEARCH DIRECTIONS

Following inherent obstacles in GT-based quality evaluation of binary images, we proposed a solution based on several intrinsic properties of binary glyphs. 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, glyph reconstruction or OCR. For our uses, a tree-based model produced adequate and robust results. Some shortcomings and potential enhancements can be proposed:

- The results of different binarization algorithms, as well as comparison with other methodologies, can be elaborated upon.
- The approach is not limited to the glyph level. If an extraction of words, sentences, or text areas are given, the measures remain applicable. However, this might involve issues such as illumination equalization and text size normalization.
- The size of our training/testing set is limited due to the reasons stated above. A further enlargement of our database is planned in the near future. In particular, testing in different settings (e.g. printed characters) may provide interesting insights related to our methodology. Moreover, if a labeled database is available, an individual combination of measures can be trained for every character, taking into account their different features.
- A potential hazard is an undesired "tailoring" of the binarization algorithms according to the evaluation methodologies employed (e.g. post-processing via median filter). Indeed, any quality measure can result in a binarization algorithm trained (in fact, over-fitted) to target the measure.

5. ACKNOWLEDGMENTS

The research was partially funded by the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013)/ERC grant agreement no. 229418. This study was also supported by a generous donation of Mr. Jacques Chahine, made through the French Friends of Tel Aviv University. Arie Shaus is grateful to the Azrieli Foundation for the award of an Azrieli Fellowship.

6. REFERENCES

[1] Gatos, B., Ntirogiannis, K. and Pratikakis, I. 2009. ICDAR 2009 document image binarization contest (DIBCO 2009). In *Proc. of ICDAR '09*, 1375-1382.

[2] Pratikakis, I., Gatos, B. and Ntirogiannis, K. 2010. H-DIBCO 2010 – Handwritten document image binarization competition. In *Proc. of ICFHR '10*, 727-732.

[3] Pratikakis, I., Gatos, B. and Ntirogiannis, K. 2011. ICDAR 2011 document image binarization contest (DIBCO 2011). In *Proc. of ICDAR '11*, 1506-1510.

[4] Barney Smith, E. H. 2010. An analysis of binarization ground truthing. In *Proc. of DAS '10*, 27-33.

[5] Barney Smith, E. H. and An, C. 2012. Effect of "Ground Truth" on image binarization. In *Proc. of DAS '12*, 250-254.

[6] Shaus, A., Turkel, E. and Piasetzky, E. 2012. Quality evaluation of facsimiles of Hebrew First Temple period inscriptions. In *Proc. of DAS '12*, 170-174.

[7] Ntirogiannis, K., Gatos, B., and Pratikakis, 2008. An objective evaluation methodology for document image binarization techniques. In *Proc. of DAS '08*, 217-224.

[8] Ntirogiannis, K., Gatos, B., and Pratikakis, 2013. Performance evaluation methodology for historical document image Binarization. *IEEE Transactions On Image Processing*, Vol. 22(2).

[9] Ben Messaoud, I., El Abed, H., Amiri, H. and Märgner, 2011. A design of a preprocessing framework for large database of historical documents. In *Proc. of HIP '11*, 177-183.

[10] Stathis, P., Kavallieratou, E. and Papamarkos, N. 2009. An evaluation technique for binarization algorithms. *J. of Universal Computer Science 14*, No. 18, pp. 3011-3030.

[11] Paredes, R. and Kavallieratou, E. 2010. ICFHR 2010 contest: Quantitative evaluation of binarization algorithms. In *Proc. of ICFHR 2010*, 733-736.

[12] Trier, Ø. D. and Jain, A. K. 1995. Goal-directed evaluation of binarization methods, *IEEE PAMI 17*, No. 12, 1191-12

[13] Trier, Ø. D. and Taxt, T. 1995. Evaluation of binarization methods for document images. *IEEE PAMI 17*, *No. 3.* 31-36.

[14] Breuel, T. M., 2001. Segmentation of handprinted letter strings using a dynamic programming algorithm. In *Proc. of DAS 2001*, 821-826.

[15] Casey, R.G. and Lecolinet, E, 1996. A survey of methods and strategies in character segmentation, *IEEE PAMI 18*, No. 7, 690-706

[16] Shaus, A., Turkel, E. and Piasetzky E. 2012. Binarization of First Temple Period inscriptions - performance of existing algorithms and a new registration based scheme. In *Proc. of ICFHR '12*, 641-646.

[17] Epshtein, B., Ofek, E. and Wexler, Y. 2010. Detecting Text in Natural Scenes with Stroke Width Transform. In *Proc. of CVPR '10*.

[18] McGillivary, C., Hale, C. and Barney Smith, E. H. 2009. Edge Noise in Document Images. In *Proc. of AND '09*, 17-24.

[19] Aharoni, Y. 1981. *Arad Inscriptions*. Israel Exploration Society.

[20] Torczyner, H. et al. 1938. *Lachish I: The Lachish Letters.* London.

[21] Otsu, N. 1979. A threshold selection method from gray-level histograms. *IEEE Trans. Systems Man Cybernet. Vol. 9 (1)*. 62–66.

[22] Bernsen, J. 1986. Dynamic thresholding of grey-level images. In *Proc. of ICPR '86*. 1251–1255.

[23] Niblack, W. 1986. *An Introduction to Digital Image Processing*. Prentice-Hall, 115–116.

[24] Sauvola, J. and Pietikainen, M. 2000. Adaptive document image binarization. *Pattern Recognition, Vol. 33.* 225–236.

[25] Tree model, R version 2.12.2. http://www.r-project.org

[26] Kendall, M. 1938. A New Measure of Rank Correlation. Biometrika 30 (1–2), 81–93.