Beyond the Ground Truth: Alternative Quality Measures of Document Binarizations

Arie Shaus, Barak Sober, Eli Turkel
The Department of Applied Mathematics
Tel Aviv University
Tel Aviv, Israel.
ashaus@post.tau.ac.il, baraksov@post.tau.ac.il,
turkel@post.tau.ac.il

Eli Piasetzky
The Sackler School of Physics and Astronomy
Tel Aviv University
Tel Aviv, Israel.
eip@tauphy.tau.ac.il

Abstract— **This article discusses the quality assessment of binary images. The customary, ground truth based methodology, used in the literature is shown to be problematic due to its subjective nature. Several previously suggested alternatives are surveyed and are also found to be inadequate in certain scenarios. A new approach, quantifying the adherence of a binarization to its document image is proposed and tested using six different measures of accuracy. The measures are evaluated experimentally based on datasets from DIBCO and H-DIBCO competitions, with respect to different kinds of binarization degradations.**

Keywords— Ground truth, binarization, evaluation, quality measure.

I. Introduction

The established methodology of document binarization assessment relies upon ground truth (GT) images (see competitions [1-6]). 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. 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. [7-10]).

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 [1-6] 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.

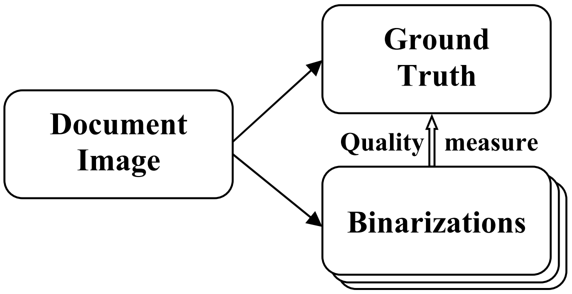

Figure 1. Standard binarization quality evaluation process. The document image is gray-scale, while the binarization and the ground truth are black and white images. The quality metric measures the adherence of the binarization to the ground truth.

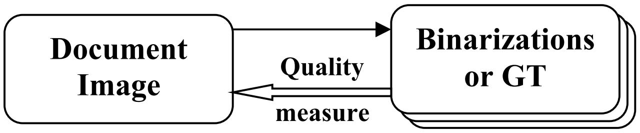

Figure 2. Proposed binarization quality evaluation process. The quality of binarization or ground truth is assessed by measuring their adherence to the document image.

The main contribution of the article is the **suggestion of several new measures, enabling the assessment of the accuracy of black and white images (binarizations or GT) directly vs. the document image** (see Fig. 2).

The rest of the paper is organized as follows: Section II deals with the pitfalls of the existing methodology. Section III gives a brief survey of previous attempts to solve or avoid these drawbacks. Section IV specifies our solution, while Sections V deals with the experimental setting and results. Finally, Section VI summarizes the outcomes and proposes future research directions.

II. 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 [11], the variability of five binarization algorithms was compared to that of different manual GTs. Significant irregularities in the GTs of the same document was found. Surprisingly, the results revealed that the variance between the binarizations was smaller than the variance between the different GTs.

2167-6445/16 $31.00 © 2016 IEEE
DOI 10.1109/ICFHR.2016.90

Article [12] deals with GTs of First Temple period Hebrew inscriptions, created by several experts. Their GTs were shown to be of markedly different quality. Paper [13] performed a binarization classifier training, based on three variants of GT. The performance of the classifiers varied significantly with respect to the underlying GT.

We therefore conclude that [11-13] demonstrate 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 [14], where automatic systems were found to be more reliable than the human "ground truther".

III. Existing Solutions

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

Article [10] 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 [15-16], and **corrected manually**. The document image edges are extracted by the Canny method [17].

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 existing methodology. The similarity of the outcomes in [10] (as well as Occam's razor principle) suggests that the existing methodology should be favored. A later article [18] made attempts to improve upon [10], 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 [19] is an elaboration on the same theme. The main changes are dropping the manual correction phase, and dilating with respect to binarizations [20-22]. This avoids 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, which is not proved by the authors.

A different approach [23-24] is to avoid the GT creation step altogether. 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, or if the noise model cannot be adequately deduced, the method is also inapplicable.

Yet another, "goal-directed" approach [25], tries to avoid ground-truthing altogether. 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.

IV. Suggested Solution

The main contribution of this article is the proposal of **several new metrics** assessing either the binarization or the GT. A first step in that direction was undertaken in [12], 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.

A similar methodology can be used in order to evaluate the quality of either the binarization (bypassing the GT), or the GT itself (therefore, adding a verification step to the existing scheme).

A. Preliminary Definitions

We assume:

1. A **black and white image** $BW(x,y)$ ($BW:[1,M]\times[1,N]\to\{0,255\}$) 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, e.g. [12]).
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 $\boldsymbol{D}$** .

We use the following notations: μ_F and μ_B are the foreground and background averages respectively, where: $\mu_S=\frac{\sum_{(x,y)\in S} D(x,y)}{\#S}$ for $S=F,B$ are the mean values of the foreground and the background; σ_F and σ_B are their respective standard deviations, defined in a similar fashion.

$n_F = \dfrac{\#F}{\#F + \#B}$ and $n_B = \dfrac{\#B}{\#F + \#B}$ are respectively the proportions of the foreground and the background pixels.

$f_i = \dfrac{\#\{(x,y) \in F \mid D(x,y) = i\}}{\#F}$ and

$b_i = \dfrac{\#\{(x,y) \in B \mid D(x,y) = i\}}{\#B}$, $i = 0...255$, are the empirical distributions (histograms) of F and B.

B. Proposed Measures

We consider the following measures:

Adapted Otsu: Article [26] 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:

$$m_{Otsu} = \mathrm{n_F} \cdot \sigma_{\mathrm{F}}{}^2 + \mathrm{n_B} \cdot \sigma_{\mathrm{B}}{}^2 . \qquad (1)$$

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

Adapted Kapur: Paper [27] 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 get:

$$m_{Kapur} = \sum_{i=0}^{255} f_i \log(f_i) + \sum_{j=0}^{255} b_j \log(b_j), \qquad (2)$$

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 authors of [28] 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:

$$\begin{aligned} m_{KI} &= 1 + 2 \cdot [\mathrm{n_B} \log(\sigma_{\mathrm{B}}) + \mathrm{n_F} \log(\sigma_{\mathrm{F}})] - \\ &-2 \cdot [\mathrm{n_B} \log(\mathrm{n_B}) + \mathrm{n_F} \log(\mathrm{n_F})] \end{aligned} . \qquad (3)$$

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

CMI: Paper [12] deals with the quality assessment of GTs of historical inscriptions. As such, this is not an adapted method, but a measure developed directly in order to handle similar tasks (also see [29-30] for additional usages):

$$m_{CMI} = \mu_B - \mu_F . \qquad (4)$$

Larger values of m_{CMI} should point to a better BW.

Potential Contrast (PC): The concept of "Potential Contrast" was presented in [31], 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:

$$m_{PC} = 255 \cdot \sum_{i: f_i \le b_i} (b_i - f_i) . \qquad (5)$$

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

Remark 1: Some of the above mentioned measures are adaptations of global binarization techniques. Indeed, 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.

Remark 2: As seen above, different approaches prefer either small or large measure values. For the sake of consistency, in the experimental 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) comparison can be also utilized for our purpose, in particular $\mathrm{L_1}$, $\mathrm{L_2}$ and PSNR measures.

L$_1$: Defined by:

$$m_{L_1} = \sum_{(x,y)} |D(x,y) - BW(x,y)| . \qquad (6)$$

L$_2$: Defined by:

$$m_{L_2} = \sqrt{\sum_{(x,y)} (D(x,y) - BW(x,y))^2} . \qquad (7)$$

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

PSNR (used in [1-6] vs. GT): Defined by:

$$m_{PSNR} = 10 \cdot \log_{10}\left(255^2 \,/\, \left(\frac{m_{L_2}{}^2}{MN}\right)\right) . \qquad (8)$$

Definition: 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, e.g.:

$$\begin{aligned} & m_1(BW, D) > m_1(BW^*, D) \Leftrightarrow \\ & m_2(BW, D) > m_2(BW^*, D) \end{aligned} \qquad (9)$$

Proposition 1: The PSNR measure is equivalent to the negated L$_2$, i.e. $m_{PSNR} \sim -m_{L_2}$.

Proof: Indeed, due to monotonicity of $\mathrm{C} \cdot \mathrm{x}$ ($0 \ne \mathrm{C} \in \mathbb{R}$), $\log_{10}$, $1/\mathrm{x}$ and x^2 (for $\mathrm{x} \ge 0$):

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

□

Proposition 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:

$$\begin{aligned} m_{L_1} &= \sum_{(x,y)} \left| D(x,y) - BW(x,y) \right| = \\ &= \sum_{(x,y)\in F} D(x,y) + \sum_{(x,y)\in B} \left(255 - D(x,y)\right) \end{aligned} \tag{11}$$

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

$$\begin{aligned} &\sim \sum_{F} D(x,y) + \sum_{B} \left(255 - D(x,y)\right) - \sum_{(x,y)} D(x,y) \\ &= \sum_{B} \left(255 - 2D(x,y)\right) \end{aligned} \tag{12}$$

On the other side:

$$\begin{aligned} m_{L_2} &= \sqrt{\sum_{(x,y)} \left(D(x,y) - BW(x,y)\right)^2} \\ &\sim \sum_{(x,y)} \left(D(x,y) - BW(x,y)\right)^2 \end{aligned} \tag{13}$$

And moreover:

$$\begin{aligned} &= \sum_{(x,y)\in F} D(x,y)^2 + \sum_{(x,y)\in B} \left(255 - D(x,y)\right)^2 = \\ &= \sum_{(x,y)\in F\cup B} D(x,y)^2 + 255 \sum_{B} \left(255 - 2D(x,y)\right) \end{aligned} \tag{14}$$

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

$$\sim \sum_{B} \left(255 - 2D(x,y)\right). \tag{15}$$

□

From Propositions 1 and 2 it follows that despite the seeming dissimilarity of the last three measures, they would in fact yield the same binarizations' ranking. Therefore, in what follows, we would only use the m_{PSNR} measure.

V. 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.

A. *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 [29], [32] 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.

Dataset: Heterogeneous and openly available data from several past binarization competitions were used, in particular DIBCO 2009 [1] (5 handwritten and 5 printed documents), H-DIBCO 2010 [2] (10 handwritten documents), DIBCO 2011 [3] (8 handwritten and 8 printed documents), H-DIBCO 2012 [4] (14 handwritten documents), DIBCO 2013 [5] (8 handwritten and 8 printed documents), and H-DIBCO 2014 [6] (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 gray-scale 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 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.

B. *Experimental Results*

A summary of the results for different types of deterioration are presented in Tables I, II and III.

TABLE I. RESULTS FOR SALT AND PEPPER DETERIORATION

Dataset[a]	#Files	% of Breaks of Monotonicity					
		Otsu	*Kapur*	*KI*	*CMI*	*PC*	*PSNR*
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.

Table I 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.

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 II. RESULTS FOR DILATION OF THE FOREGROUND

Dataset	#Files	% of Breaks of Monotonicity					
		Otsu	*Kapur*	*KI*	*CMI*	*PC*	*PSNR*
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%**

Table II 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 good 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 III. RESULTS FOR EROSION OF THE FOREGROUND

Dataset	#Files	% of Breaks of Monotonicity					
		Otsu	*Kapur*	*KI*	*CMI*	*PC*	*PSNR*
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%**

Table III 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.

VI. SUMMARY AND FUTURE DIRECTIONS

We presented several measures, which quantify the adherence of a binary image to its grayscale document image. The binary document can either be a GT, or a product of a binarization algorithm. 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.

The measures used in this article are of a global nature. Other such measures can be adapted from surveys such as [33]. Additionally, various measures operating on a local (i.e. "moving window") level, can also be considered.

Another research direction is the elimination of the reliance not only on the GT, but also on the document image itself. This may be possible utilizing the intrinsic properties of the binarization. Such a proposal is hinted by [34] (where it is performed manually) and [35].

ACKNOWLEDGMENT

The research reported here received initial funding from the Israel Science Foundation – F.I.R.S.T. (Bikura) Individual Grant no. 644/08, as well as the Israel Science

Foundation Grant no. 1457/13. The research was also funded the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013)/ERC grant agreement no. 229418, and by an Early Israel grant (New Horizons project), Tel Aviv University. 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. We would also like to acknowledge the valuable help of Shirly Ben-Dor Evian, Sivan Einhorn, Shira Faigenbaum-Golovin, Israel Finkelstein and David Levin.

References

[1] B. Gatos, K. Ntirogiannis and I. Pratikakis, “ICDAR 2009 document image binarization contest (DIBCO 2009),” in Proceedings of the 10th International Conference on Document Analysis and Recognition (ICDAR 2009), pp. 1375-1382, 2009.

[2] I. Pratikakis, B. Gatos and K. Ntirogiannis, “H-DIBCO 2010 – Handwritten document image binarization competition,” in Proceedings of the 12th International Conference on Frontiers in Handwriting Recognition (ICFHR 2010), pp. 727-732, 2010.

[3] I. Pratikakis, B. Gatos and K. Ntirogiannis, “ICDAR 2011 document image binarization contest (DIBCO 2011),” in Proceedings of the 11th International Conference on Document Analysis and Recognition (ICDAR 2011), pp. 1506-1510, 2011.

[4] I. Pratikakis, B. Gatos and K. Ntirogiannis, “H-DIBCO 2012 – Handwritten document image binarization competition,” in Proceedings of the 13th International Conference on Frontiers in Handwriting Recognition (ICFHR 2012), pp. 817-822, 2012.

[5] I. Pratikakis, B. Gatos and K. Ntirogiannis, “ICDAR 2013 document image binarization contest (DIBCO 2013),” in Proceedings of the 12th International Conference on Document Analysis and Recognition (ICDAR 2013), pp. 1471-1476, 2013.

[6] K. Ntirogiannis, B. Gatos and I. Pratikakis, “H-DIBCO 2014 – Handwritten document image binarization competition,” in Proceedings of the 14th International Conference on Frontiers in Handwriting Recognition (ICFHR 2014), pp. 809-813, 2014.

[7] E. Saund, J. Lind, and P. Sarkar. “PixLabeler: User interface for pixel-level labeling of elements in document images,” in Proceedings of the 10th International Conference on Document Analysis and Recognition (ICDAR 2009), pp. 646–650, 2009.

[8] A. Fischer, E. Indermühle, H. Bunke, G. Viehhauser and M. Stolz, “Ground truth creation for handwriting recognition in historical documents,” in Proceedings of the 9th IAPR Workshop on Document Analysis Systems (DAS 2010), pp. 3-10, 2010.

[9] C. Clausner, S. Pletschacher and A. Antonacopoulos, “Aletheia - An advanced document layout and text ground-truthing system for production environments,” in Proceedings of the 11th International Conference on Document Analysis and Recognition (ICDAR 2011), pp. 48-52, 2011.

[10] K. Ntirogiannis, B. Gatos and I. Pratikakis, “An objective evaluation methodology for document image binarization techniques,” in Proceedings of the 8th IAPR Workshop on Document Analysis Systems (DAS 2008), pp. 217-224, 2008.

[11] E. H. Barney Smith, “An analysis of binarization ground truthing,” in Proceedings of the 9th IAPR Workshop on Document Analysis Systems (DAS 2010), pp. 27-33, 2010.

[12] A. Shaus, E. Turkel and E. Piasetzky, “Quality evaluation of facsimiles of Hebrew First Temple period inscriptions”, in Proceedings of the 10th IAPR Workshop on Document Analysis Systems (DAS 2012), pp. 170-174, 2012.

[13] E. H. Barney Smith and C. An, “Effect of “Ground Truth” on image binarization,” in Proceedings of the 10th IAPR Workshop on Document Analysis Systems (DAS 2012), pp. 250-254, 2012.

[14] R. M. Brown, T. H. Fay and C. U. Walker, “Handprinted symbol recognition system”, Pattern Recognition 21, No. 2, pp. 91-118, 1988.

[15] M. Kamel, and A. Zhao, “Extraction of binary character/graphics images from grayscale document images”, CVGIP: Computer Vision Graphics and Image Processing 55, No. 3, pp. 203-217, 1993.

[16] H. J. Lee, and B. Chen, “Recognition of handwritten Chinese characters via short line segments”, Pattern Recognition 25, No. 5, pp. 543-552, 1992.

[17] J. Canny, “A Computational Approach to Edge Detection”, IEEE Transactions on Pattern Analysis and Machine Intelligence 8, No. 6, pp. 679-698, 1986.

[18] K. Ntirogiannis, B. Gatos and I. Pratikakis, "Performance evaluation methodology for historical document image binarization", IEEE Transactions on Image Processing 22, No. 2, pp. 595-609, 2012.

[19] I. Ben Messaoud, H. El Abed, H. Amiri and V. Märgner, “A design of a preprocessing framework for large database of historical documents,” in Proceedings of the 2011 Workshop on Historical Document Imaging and Processing, pp. 177-183, 2011.

[20] J. Sauvola and M. Pietikainen, “Adaptive document image binarization,” Pattern Recognition 33, No. 2, pp. 225–236, 2000.

[21] B. Gatos, I. Pratikakis, and S. Perantonis, “Adaptive degraded document image binarization,” Pattern Recognition 39, pp. 317–327, 2006.

[22] S. Lu and B. Su and C. L. Tan, “Document image binarization using background estimation and stroke edge,” Inter. Journal on Document Analysis and Recognition 13, No. 4, pp. 303–314, 2010.

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

[24] R. Paredes and E. Kavallieratou, “ICFHR 2010 contest: Quantitative evaluation of binarization algorithms,” in Proceedings of the 12th International Conference on Frontiers in Handwriting Recognition (ICFHR 2010), pp. 733-736, 2010.

[25] Ø. D. Trier and A. K. Jain, “Goal-directed evaluation of binarization methods”, IEEE Transactions On Pattern Analysis And Machine Intelligence 17, No. 12, pp. 1191-1201, 1995.

[26] N. Otsu, “A threshold selection method from gray-level histograms,” IEEE Transactions on Systems, Man and Cybernetics 9, No. 1, pp. 62-66, 1979.

[27] J. N. Kapur, P. K. Sahoo and A. K. C. Wong, “A new method for gray-level picture thresholding using the entropy of the histogram,” Computer Vision, Graphics, and Image Processing 29, No. 3, pp. 273-285, 1985.

[28] J. Kittler and J. Illingworth, “Minimum error thresholding,” Pattern Recognition 19, No. 1, pp. 41-47, 1986.

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

[30] A. Shaus, I. Finkelstein and E. Piasetzky, “Bypassing the eye of the beholder: Automated ostraca facsimile evaluation,” Maarav 17.1, pp. 7-20, 2010.

[31] S. Faigenbaum, B. Sober, A. Shaus, M. Moinester, E. Piasetzky, G. Bearman, M. Cordonsky and I. Finkelstein, “Multispectral Images of Ostraca: Acquisition and Analysis,” Journal of Archaeological Science, Vol. 39, Issue 12, pp. 3581–3590, 2012.

[32] A. Shaus, B. Sober, E. Turkel and E. Piasetzky, “Improving binarization via sparse methods”, in Proceedings of the 16th International Graphonomics Society Conference (IGS 2013), pp. 163-166, 2013.

[33] M. Athimethphat, “A review on global binarization algorithms for degraded document images”, Assumption University Journal of Technology 14, No. 3, pp. 188-195, 2011.

[34] Ø. D. Trier and T. Taxt, “Evaluation of binarization methods for document images”, IEEE Transactions on Pattern Analysis and Machine Intelligence 17, No. 3, 1995.

[35] S. Faigenbaum, A. Shaus, B. Sober, E. Turkel and E. Piasetzky, “Evaluating glyph binarizations based on their properties”, in Proceedings of the 13th ACM Symposium on Document Engineering (DocEng2013), pp. 127-130, 2013.