Performance Evaluation Methodology for Historical Document Image Binarization

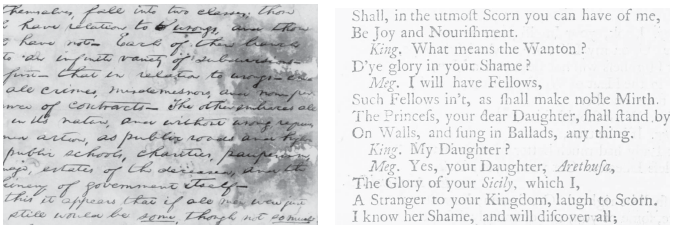
Konstantinos Ntirogiannis, Basilis Gatos, and Ioannis Pratikakis, *Senior Member, IEEE*

***Abstract*— Document image binarization is of great importance in the document image analysis and recognition pipeline since it affects further stages of the recognition process. The evaluation of a binarization method aids in studying its algorithmic behavior, as well as verifying its effectiveness, by providing qualitative and quantitative indication of its performance. This paper addresses a pixel-based binarization evaluation methodology for historical handwritten/machine-printed document images. In the proposed evaluation scheme, the recall and precision evaluation measures are properly modified using a weighting scheme that diminishes any potential evaluation bias. Additional performance metrics of the proposed evaluation scheme consist of the percentage rates of broken and missed text, false alarms, background noise, character enlargement, and merging. Several experiments conducted in comparison with other pixel-based evaluation measures demonstrate the validity of the proposed evaluation scheme.**

***Index Terms*— Document image binarization, ground truth, performance evaluation.**

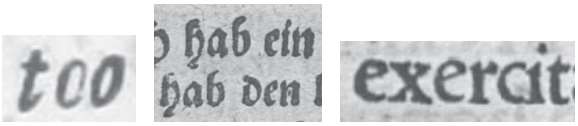
I. Introduction

HISTORICAL documents suffer from various degradations due to ageing, extended use, several attempts of acquisition and environmental conditions [1]–[4]. The main artefacts encountered in historical documents are shadows, non-uniform illumination, smear, strain, bleed-through and faint characters (Fig. 1). Those artefacts are problematic for document image analysis methods which assume smooth background and uniform quality of writing [1]. In handwritten documents (Fig. 1a), the writer may use different amount of ink and pressure and generate characters of different intensity or thickness, as well as faint characters. The same writer may write in different ways even within the same document. Similar problems, such as faint characters (Fig. 1b) and non-uniform



(a) (b)

Fig. 1. Historical documents with various degradations. (a) Handwritten document image with smears, faint characters, and characters of uneven intensity and thickness. (b) Machine-printed document image with bad illumination, faint characters, and some bleed-through.



(a) (b) (c)

Fig. 2. (a)–(c) Example historical machine-printed images with deviations from standard font characteristics (size, structure, etc.) that result in nonuniform appearance of the same font.

appearance of characters of the same font (Fig. 2), are also encountered in historical machine-printed documents [3], [4].

Document image binarization is a critical stage of the document image analysis and recognition pipeline that affects the final recognition results [5]. Therefore, it is imperative to have an evaluation methodology which will account for the performance of the binarization not only in qualitative but also in quantitative terms. Several efforts have been presented for the evaluation of document image binarization techniques that can be classified in three main categories.

In the first category, evaluation is performed by the visual inspection of one or many human evaluators [6]–[12]. For example, in [6], the amount of symbols that are broken or blurred, the loss of objects and the noise in background and foreground are used as visual evaluation criteria. The symbols that are broken, lost, etc. can be roughly counted since those criteria cannot be quantitatively measured with satisfactory precision by humans.

In the second category, evaluation is addressed taking into account the OCR performance. The binarization outcome is subject to OCR and the corresponding result is evaluated with respect to character and word accuracy [2], [23]–[25]. Evaluation using OCR (supported by contemporary state-of-the-art OCR engines, e.g. ABBYY FineReader [22]) concerns mainly modern machine-printed documents since handwritten or historical document OCR does not always yield satisfactory

Manuscript received March 12, 2012; revised June 26, 2012; accepted August 27, 2012. Date of publication September 18, 2012; date of current version January 10, 2013. This work was supported by the EU 7th Framework Programme through Project IMPACT under Grant 215064. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Mark (Hong-Yuan) Liao.

K. Ntirogiannis is with the Department of Informatics and Telecommunications, National and Kapodistrian University of Athens, Athens GR-15784, Greece, and also with the Institute of Informatics and Telecommunications, National Center for Scientific Research "Demokritos," Athens GR-15310, Greece (e-mail: kntir@iit.demokritos.gr).

B. Gatos is with the Institute of Informatics and Telecommunications, National Center for Scientific Research "Demokritos," Athens GR-15310, Greece (e-mail: bgat@iit.demokritos.gr).

I. Pratikakis is with the Department of Electrical and Computer Engineering, Democritus University of Thrace, Xanthi GR-67100, Greece (e-mail: ipratika@ee.duth.gr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2012.2219550

results [9], [23]–[25]. Even though OCR-based evaluation is important, the OCR performance does not only depend on binarization but also on the effectiveness of several tasks for the recognition [26]–[28]. Consequently, the OCR does not provide a direct evaluation of binarization.

In the third category, direct evaluation of binarization is performed by taking into account the pixel-to-pixel correspondence between the ground truth and the binarized image. Evaluation is based either on synthetic images [23], [24], [29]–[31] or on real images [25]–[28], [31]–[41]. Synthetic images can be produced on a large scale but they do not reflect the degradations encountered in real degraded documents [26], [27]. Furthermore, as shown in [33], degradations that appear in historical machine-printed documents cannot be handled by degradation models used mainly in the image acquisition [42]–[44] or in the OCR development area [45], [46]. On the contrary, ground truth images from real degraded images [37]–[39] which correspond to real "challenging" cases, are not available in large quantities.

Concerning pixel-based evaluation, several measures have been used for the evaluation of document image binarization techniques, such as F-Measure (Recall, Precision), PSNR [32], Negative Rate Metric (NRM) [37], [38], Misclassification Penalty Metric (MPM) [37]–[39], chi-square metric [47], geometric-mean accuracy [24] and the normalized cross-correlation metric [27], even though they have not been specifically designed for that purpose. Nevertheless, there are evaluation measures developed for document images, such as the DRD (Distance Reciprocal Distortion) [48] which is used to measure the distortion on binary document images generated from the image acquisition process. Furthermore, in [35], [36], the dual representation of the ground truth (dualGT) is considered, by which the skeletonized ground truth is used for the computation of Recall (Recall$_{skel}$). However, some researchers have stated the need for an improved pixel-based evaluation measure for document image binarization. For instance, in [27], wherein the ground truth generation from several users was studied, it was stated that there is a need for a weighted measure in relation to the ground truth borders in order to compensate the subjectivity of the ground truth. Furthermore, in [21] it was stated that there is a gap between the pixel-based and the OCR evaluation results and that pixel-based evaluation metrics should be improved.

In this paper, a pixel-based evaluation methodology is presented which relies upon new measures that extend the typical Recall and Precision measures. The two newly proposed measures, namely pseudo-Recall R_{ps} and pseudo-Precision P_{ps}, are based upon a weighted penalization of the pixels around the ground truth character borders and take into account the local stroke width and the distance from the contour of the ground truth text. Several experiments made in comparison with other pixel-based evaluation measures demonstrate the validity of the proposed measures.

In the remainder of the paper, the motivation for the proposed evaluation measures is presented in Section II, the proposed evaluation measures are detailed in Section III, while in Section IV the experimental results are presented. Finally, in Section V, conclusions are drawn.

Fig. 3. (a) Original image. (b) Indicated area of (a) in which the white mark belongs to "foreground" and the black mark belongs to "background." Ambiguity exists along a path which connects those two marks. (c) Text boundary as determined by the Canny edge detector.

II. Motivation

The core motivation for the proposed evaluation measures, was the ambiguity in text boundary localization that mainly occurs due to the document digitization process. As it is shown in Fig. 3, there is not a rapid change of the grey values between foreground and background, but ambiguity exists within a small region between them. Several authors have already stated that the location of the correct boundaries between two different regions is a subjective matter [27], [49], [50]. In [49], where humans marked the boundaries to segment image regions, it was mentioned that the ground truth data contained boundary localization errors and *"would not tolerate any localization error and would consequently overpenalize algorithms that generate usable, though slightly mislocalized boundaries"*. Furthermore, in [27], the author discussed the subjective nature of binarization addressed at the ground truth construction stage and it was concluded that the ground truth from different users was not identical near the contour of the characters. In the context of the aforementioned ambiguity, different binarization techniques could produce outputs that differ mainly along their contour. In this case, for a historical document with faint characters (Fig. 4), representative measures that make use of the number of true/false positives/negatives such as F-Measure (FM) and PSNR, could rank in a better position a binarized image with more broken characters and false alarms as in Fig. 4b (FM = 94.37, PSNR = 22.89) than a better binarized image as in Fig. 4c (FM = 93.69, PSNR = 22.56). For Fig. 4c that contains less broken characters, higher Recall is expected than Fig. 4b. However, the binarized image of Fig. 4c achieves lower Recall = 89.78 compared to the Recall = 93.77 of Fig. 4b, as a result of the more missing foreground pixels (false negatives) which are mainly situated along the borders of the characters, making their absence less obvious.

In similar case to the aforementioned (Fig. 4), the use of the skeletonized ground truth for the computation of Recall [35], [36], provides better evaluation results. For Fig. 4b, false negatives corresponding to broken characters are taken into account (FM$_{dualGT}$ = 95.29, Recall$_{skel}$ = 95.62), while false negatives situated near the contour as in Fig. 4c, are not considered at all (FM$_{dualGT}$ = 98.79, Recall$_{skel}$ = 99.64). However, the dual representation of the ground truth [35], [36] could mislead the evaluation results when the binarized image is deformed while the skeletonized ground truth can be completely detected, as shown in Fig. 5. In those cases, both Recall$_{skel}$ and Precision are 100 (FM$_{dualGT}$ = 100), leading to erroneous evaluation.

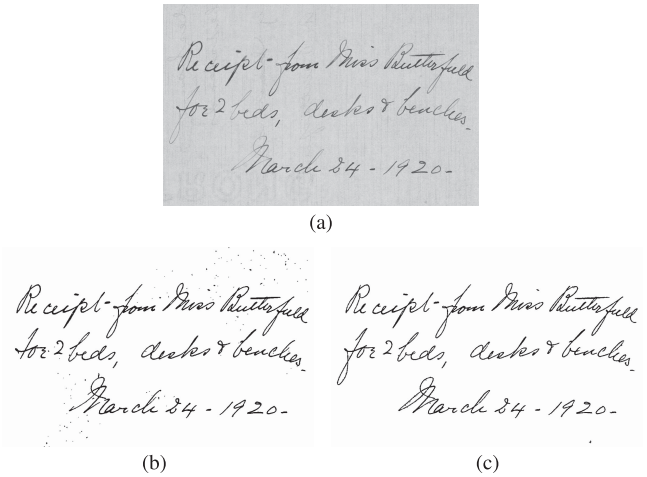


Fig. 4. Deviation between quantitative and qualitative evaluation using F-measure (FM) and PSNR measures (a) original image, (b) binarized image with broken characters and false alarms, FM = 94.37 (recall = 93.77), PSNR = 22.89, and (c) binarized image of better quality but with lower performance, FM = 93.69 (recall = 89.78), PSNR = 22.56.

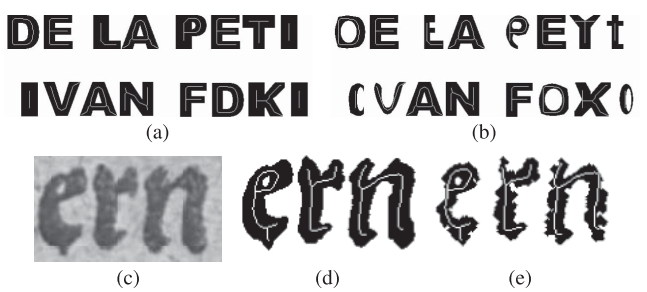


Fig. 5. Problematic cases concerning the skeletonized ground truth (a) original characters along with the skeletonized ground truth, (b) modified characters in which the skeletonized ground truth is fully detected, (c) original image, (d) ground truth image along with the skeletonized ground truth, and (e) binarization output with severely damaged characters, wherein the skeletonized ground truth is fully detected.

The distance-based evaluation measures, such as DRD and MPM (for which lower values indicate better performance), have the advantage of applying a weighting penalization starting from the ground truth borders. Indeed, MPM equals 0.85 and 0.07, while DRD equals 1.72 and 1.53 for the binarized images of Fig. 4b and Fig. 4c respectively, indicating better performance for Fig. 4c. However, those measures can overpenalize a binarized image with noise far from the text preserving the textual information as in Fig. 6b (MPM = 15.56, DRD = 32.71), compared to a binarized image where noise is situated among the characters destroying the textual information, as in Fig. 6c (MPM = 2.28, DRD = 29.96).

III. Proposed Evaluation Measures

The proposed evaluation measures take into account major issues that are responsible for erroneous pixel-based evaluation, such as the ambiguity in the text boundary localization and the location of the introduced noise. Specifically, those measures take into account the local stroke width and the distance from the contour of the ground truth text (Fig. 7).

Fig. 6. Deviation between quantitative and qualitative evaluation using DRD and MPM measures (a) original image, (b) binarization with noise outside the text, DRD = 32.71, MPM = 15.56, and (c) binarization with noise among the text but with better performance, DRD = 29.96, MPM = 2.28.

A. Pseudo-Recall

To measure the loss of information, evaluation measures based on the amount of true/false positives/negatives, such as Recall and PSNR or distance-based measures such as MPM [39] and DRD [48] can be used. However, the use of those measures could lead in erroneous evaluation, especially in the cases listed below.

1) A binarized character with missing foreground pixels (false negatives) from the contour that do not affect the character topology (Fig. 8a), compared to a binarized character for which the lack of the same amount of foreground pixels alters the character topology (Fig. 8b), achieves: a) equal performance (Table I) when the typical measures of Recall or PSNR are used, because of the same amount of false negative pixels, b) better performance (Table I) when the distance-based measures MPM and DRD are used, because those measures apply lower penalization near the ground truth contour (Fig. 9).
2) A breaking at a thin section could seriously affect the character instance, as shown in Fig. 8c, while the lack of the same amount of foreground pixels from a thicker section has a smaller impact on the character instance (Fig. 8d). Although the character breaking is considered more serious, equal performance is achieved through the typical measures Recall and PSNR, while less penalization is applied by the distance-based measures DRD and MPM (Table I), since when a breaking occurs there are more false negative pixels situated near the ground truth contour where lower weights are assigned by those measures (Fig. 9).

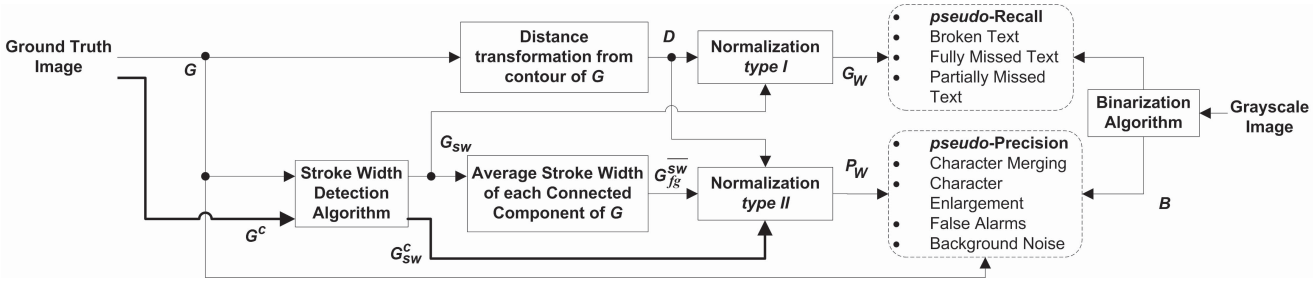


Fig. 7. Flowchart of the proposed evaluation methodology.

Fig. 8. Missing pixels illustrated by shaded black areas. (a) Missing pixels close to the character borders. (b) Same number of missing pixels results in character breaking. (c) Missing pixels results in breaking of a thin but important character part. (d) Equal number of missing pixels from a character part that have lower impact on the character instance.

TABLE I
COMPARISON OF DIFFERENT MEASURES CONCERNING THE LOSS OF INFORMATION AS DEPICTED IN FIG. 8

	Recall	PSNR	MPM	DRD	R_{ps}
Fig. 8a	93.00	15.30	**2.74**	**2.75**	**98.20**
Fig. 8b	93.00	15.30	15.85	3.46	94.11
Fig. 8c	98.73	21.83	**1.10**	**0.30**	98.97
Fig. 8d	98.73	21.83	2.74	0.37	**99.71**

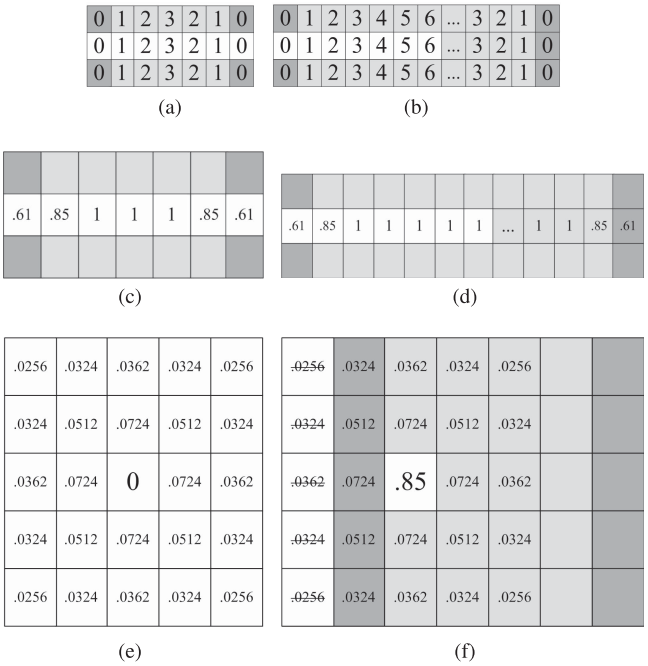


Fig. 9. (a) and (b) Weights to which the MPM is based start from the ground truth contour. False negatives (in white) that (a) result in character breaking or (b) do not result in character breaking. (c) and (d) Weights to which the DRD is based are lower than one only close to the ground truth contour. White pixels denote the false negatives that (c) result in character breaking or (d) do not result in character breaking. (e) DRD is based on the 5 × 5 normalized weight matrix shown. (f) example application of the DRD matrix on a false negative pixel near the contour.

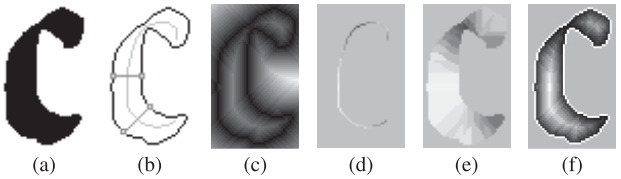


Fig. 10. (a) Ground truth character and (b) contour points in black along with the skeleton $S(x, y)$ in grey. Examples of line segments connecting anti-diametric contour points are also given. (c) Distance map $D(x, y)$ from the contour points. (d) Local thickness assigned to $S(x, y)$. (e) Stroke width image $G_{sw}(x, y)$. (f) Weighted ground truth image $G_W(x, y)$.

To overcome the aforementioned inconsistencies of the typical measures (Recall,PSNR) and the distance-based measures (MPM, DRD), the pseudo-Recall R_{ps} is proposed, for which the foreground ground truth image domain is weighted by attributing to each pixel its distance from the contour. In this way, weights of the contour equal to zero while the maximum values are taken by these pixels that belong to the skeleton $S(x, y)$ of the ground truth text. Furthermore, those weights are normalized along each line segment that connects two anti-diametric contour points and it is normal to the skeleton $S(x, y)$ (Fig. 10b), so that each character breaking is equally penalized regardless of the local stroke width. In order to achieve the aforementioned characteristics, pseudo-Recall is defined as follows:

$$R_{ps} = \frac{\sum_{x=1,y=1}^{x=I_x,y=I_y} B(x, y) \cdot G_W(x, y)}{\sum_{x=1,y=1}^{x=I_x,y=I_y} G_W(x, y)} \quad (1)$$

where $B(x, y)$ denotes the binarized image under evaluation (with 0 corresponding to 'background' and 1 to 'foreground') and $G_W(x, y)$ denotes the weighted ground truth image (Fig. 11f, 12f) which is defined in Eq. 2:

$$G_W(x, y) = \begin{cases} \frac{D(x,y)}{N_R(x,y)}, & \text{if } G_{sw}(x, y) > 2 \\ \frac{1}{G_{sw}(x,y)}, & \text{otherwise} \end{cases} \quad (2)$$

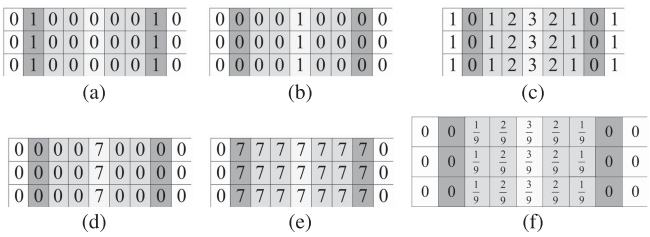


Fig. 11. (a) Contour points of ground truth, (b) skeletonized ground truth $S(x, y)$, (c) distance map $D(x, y)$, (d) local thickness assigned to $S(x, y)$, (e) stroke width image $G_{sw}(x, y)$, and (f) weighted $G_W(x, y)$ image.

where: $D(x, y)$ denotes the distance map based on the 'Chebyshev' distance metric of the foreground ground truth image using the contour points as starting points (Fig. 10c, 11c).

$G_{sw}(x, y)$ denotes the stroke width image of the ground truth characters which is calculated based on [51]. For a given connected component of the ground truth, each point of the skeletonized ground truth $S(x, y)$ (Fig. 10b, 11b) is assigned to the local stroke width (Fig. 10d, 11d). Finally, the remaining ground truth points inherit the value of the nearest weighted skeleton point (Fig. 10e, 11e).

$N_R(x, y)$ (Eq. 3) denotes the pixel-wise normalization factor of the distance map $D(x, y)$, so that the summation of the normalized $D(x, y)$ values along a line segment which is normal to the skeleton $S(x, y)$ and connects two anti-diametric contour points of the ground truth, equals to 1 regardless of the local stroke width (Normalization *type I*, Fig. 7).

$$N_R(x, y) = \begin{cases} \lfloor \frac{G_{sw}(x,y)}{2} \rfloor^2, & \text{if } G_{sw}(x, y) mod(2) = 1 \\ (\frac{G_{sw}(x,y)}{2})(\frac{G_{sw}(x,y)}{2} - 1), & \text{otherwise.} \end{cases} \quad (3)$$

For the sake of clarity, a distinction should be made between the proposed pseudo-Recall measure R_{ps} and the pseudo-Recall measure used in H-DIBCO 2010 contest [37]. In [37], the term pseudo-Recall was used to discriminate the modified Recall (that uses only the skeleton of the ground truth) from the standard Recall.

Foreground information of the ground truth image that has not been detected by $B(x, y)$ is classified as *Fully Missed Text (E_{fmt})*, *Partially Missed Text (E_{pmt})* and *Broken Text (E_{bt})*. Those error measures are complementary to the pseudo-Recall ($R_{ps} + E_{fmt} + E_{pmt} + E_{bt} = 1$) and are detailed in the following.

1) Fully Missed Text: corresponds to connected components of the ground truth image that have been completely missed by the binarized image $B(x, y)$ (Fig. 12d).

$$E_{fmt} = \frac{\sum_{x=1,y=1}^{x=I_x,y=I_y} (1 - f_g(G_{cc}(x, y))) \cdot G_W(x, y)}{\sum_{x=1,y=1}^{x=I_x,y=I_y} G_W(x, y)} \quad (4)$$

where $G_{cc}(x, y)$ denotes the connected component labelled image of the ground truth $G(x, y)$ (hereafter subscript cc denotes the corresponding connected component labelled image) while $f_g(i)$ denotes whether the i-th connected component of the ground truth (i.e. $G_{cc}(x, y) = i$) is fully detected

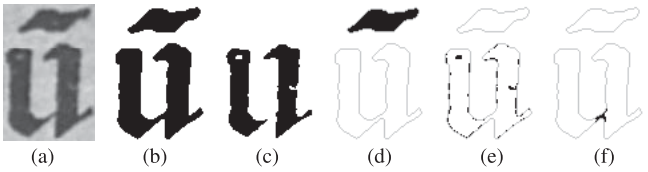


Fig. 12. (a) Original image. (b) Ground truth image, (c) Binarization output. Location of the pixels classified as (d) fully missed text, (e) partially missed text, and (f) broken text.

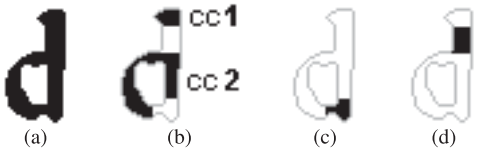


Fig. 13. (a) Ground truth character, (b) binarization output and the corresponding connected components cc1 and cc2 of the true positives image $I^{tp}(x, y)$, (c) missing component (I_{cc}^{fn}) that neighbors only with cc2 is classified as partially missed text, and (d) missing component (I_{cc}^{fn}) that neighbors with cc1 and cc2 is classified as broken text.

by the binarized image $B(x, y)$ and it is defined as follows:

$$f_g(i) = \begin{cases} 0, & \text{if } \sum_{\substack{x=1,y=1 \\ G_{cc}(x,y)=i}}^{x=I_x,y=I_y} G(x, y) \cdot B(x, y) = 0 \\ 1, & \text{otherwise.} \end{cases} \quad (5)$$

2) Partially Missed Text: concerns the false negative pixels that do not result in the local breaking of a ground truth component into two or more components (Eq. 7) (Fig. 12e, 13c). To formally define the *Partially Missed Text*, it is necessary to define: i) the false negatives image, i.e. $I^{fn}(x, y) = G(x, y) \cdot (1 - B(x, y))$ (Fig. 13c-13d), ii) the true positives image, i.e. $I^{tp}(x, y) = G(x, y) \cdot B(x, y)$ (Fig. 13b) and iii) the function $k(i)$ (Eq. 6) that indicates whether the i-th connected component of $I^{fn}(x, y)$ neighbours with more than one connected components of the $I^{tp}(x, y)$ image.

$$k(i) = \begin{cases} 1, & \text{if } h(i) > 1 \\ 0, & \text{otherwise} \end{cases} \quad (6)$$

where $h(i)$ denotes the number of different $I^{tp}(x, y)$ components which are neighbours to the i-th connected component of $I^{fn}(x, y)$.

$$E_{pmt} = \frac{\sum_{x=1,y=1}^{x=I_x,y=I_y} f_g(G_{cc}(x, y)) \cdot (1 - k(I_{cc}^{fn}(x, y))) \cdot G_W(x, y)}{\sum_{x=1,y=1}^{x=I_x,y=I_y} G_W(x, y)}. \quad (7)$$

3) Broken Text: It is defined (Eq. 8) by the false negative pixels that result in the local breaking of a ground truth

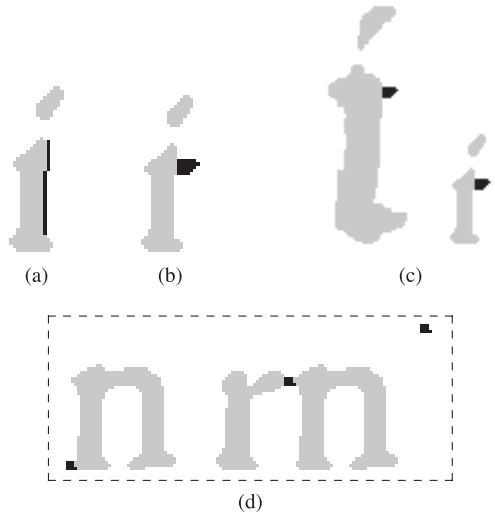


Fig. 14. Noise (black pixels) around (a) the contour of the character and (b) the same amount of noise altering the character. (c) Same amount of noise close to components of different sizes where the smaller one is more affected (ground truth is in grey). (d) Small amount of false positives situated i) at a single "n" character, ii) among an "n" and an "r" merging then into an "m", and iii) far from any character (ground truth is in grey).

component into two or more components (Fig. 12f, 13d).

$$E_{fmt} = \frac{\sum_{x=1,y=1}^{x=I_x,y=I_y} f_g(G_{cc}(x,y)) \cdot k(I_{cc}^{fn}(x,y)) \cdot G_W(x,y)}{\sum_{x=1,y=1}^{x=I_x,y=I_y} G_W(x,y)}. \tag{8}$$

B. Pseudo-Precision

To measure the amount of the introduced noise (excessive information), either typical evaluation measures that are based on the true/false positives/negatives, such as Precision and PSNR, or distance-based measures such as MPM and DRD, can be used. However, the use of those measures could lead to erroneous evaluation, especially in the cases listed below.

1) As shown in Fig. 14a-14b, a binarized character with false positive pixels detected along its contour without seriously affecting the character topology, compared to a binarized character where equal amount of false positives alters its topology and consequently textual information is missed, achieves: a) equal performance (Table II) when the typical measures of Precision or PSNR are used, because of the same amount of false positives pixels, b) better performance when the distance-based measures MPM and DRD are used, because those measures apply lower penalization near the ground truth contour, as already demonstrated in previous Sections II and III-A.
2) As shown in Fig. 14c, false positives detected on a thin character and equal amount of false positives detected on a thicker/larger character could result in equal performance (Table II) using either the typical measures (Precision, PSNR) or the distance-based measures (MPM, DRD), although the thin character gets a higher noise contamination. This behaviour is expected since none of the aforementioned measures considers the thickness of the characters.
3) As shown in Fig. 14d, a small amount of false positive pixels could merge adjacent characters, while when the same amount of noise affects a single or no character at all, the noise contamination is much lower. However, using the typical measures (Precision, PSNR) equal performance is achieved (Table II), while using the distance-based measures (MPM, DRD) the merging case gets lower penalty (under-penalization) since more false positives are situated near the ground truth contour where lower penalization is applied. Furthermore, using the distance-based measures (MPM, DRD) false positives detected far from the ground truth components preserving the textual information (see also Fig. 6) are over-penalized, especially by the MPM, as has already been described in Section II.

TABLE II
COMPARISON OF DIFFERENT MEASURES CONCERNING ERRONEOUSLY INTRODUCED NOISE AS DEPICTED IN FIG. 14

	Precision	PSNR	MPM	DRD	P_{ps}
Fig. 14a	91.62	19.33	**0.90**	**1.32**	**94.93**
Fig. 14b	91.62	19.33	3.05	1.82	93.48
Fig. 14c left	98.57	27.53	0.19	0.53	**99.09**
Fig. 14c right	98.57	27.53	0.19	0.53	98.91
Fig. 14d left	99.34	29.60	0.15	0.15	99.53
Fig. 14d middle	99.34	29.60	**0.10**	**0.13**	99.45
Fig. 14d right	99.34	29.60	1.06	0.18	**99.68**

To overcome the aforementioned inconsistencies of the typical measures (Precision, PSNR) and the distance-based measures (MPM, DRD), the pseudo-Precision P_{ps} is proposed, for which the background ground truth image domain is weighted based on the distance from the contour of the ground truth text. To handle the over-penalization of false positives pixels that are far from a ground truth component, the weights are constrained within a region that extends according to the stroke width of the corresponding component. Weights within this region are normalized to the interval (1,2] taking into account the stroke width of the corresponding ground truth component and the distance between any neighbouring component, while weights outside this region are set to 1. It should be noted that the distance between adjacent characters is required to handle the under-penalization that occurs in the character merging cases, since in those cases the distance-based weights are limited to low values. Specifically, the normalized weights along a line segment that connects two anti-diametric contour points (passing through background points) and it is normal to the skeleton of the ground truth background, take maximum values on pixels of the aforementioned skeleton. In order to achieve all these characteristics, the pseudo-Precision is defined as follows:

$$P_{ps}(x,y) = \frac{\sum_{x=1,y=1}^{x=I_x,y=I_y} G(x,y) \cdot B_W(x,y)}{\sum_{x=1,y=1}^{x=I_x,y=I_y} B_W(x,y)} \tag{9}$$

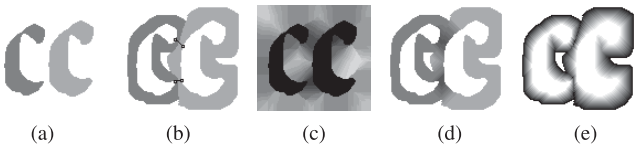


Fig. 15. (a) Average stroke width $G_{fg}^{\overline{sw}}(x, y)$ image, (b) $G_{bg}^{\overline{sw}}(x, y)$ image along with example background line segments that connect anti-diametric contour points, (c) $G_{sw}^{c}(x, y)$ image that aids in the normalization of background pixels, (d) normalization image $N_P(x, y)$, and (e) weighted image $P_W(x, y)$.

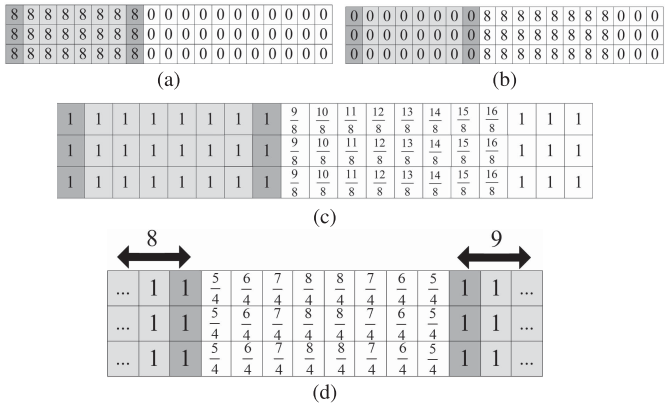


Fig. 16. (a) $G_{fg}^{\overline{sw}}(x, y)$ image with dark grey pixels denoting ground truth contour, (b) $G_{bg}^{\overline{sw}}(x, y)$ image of (a), (c) $P_W(x, y)$ image of (a), and (d) example $P_W(x, y)$ image with components/strokes close to each other.

where $B_W(x, y) = B(x, y) \cdot P_W(x, y)$ is the binarized image for evaluation after the proposed weighted map $P_W(x, y)$ is applied, for which the definition is given by Eq. 10.

$$P_W(x, y) = \begin{cases} 1, & \text{if } G_{bg}^{\overline{sw}}(x, y) = 0 \\ 1 + \frac{D(x,y)}{N_P(x,y)}, & \text{otherwise} \end{cases} \tag{10}$$

where: $G_{bg}^{\overline{sw}}(x, y)$ denotes the region that extends to the background by the average stoke width of the corresponding connected component $G_{fg}^{\overline{sw}}(x, y)$ (Fig. 15a, 16a). Within this region, pixels of $G_{bg}^{\overline{sw}}(x, y)$ are assigned to the average stroke width $G_{fg}^{\overline{sw}}(x, y)$ value of the nearest ground truth component found, while zero is assigned elsewhere (Fig. 15b, 16b).

$N_P(x, y)$ (Eq. 11) denotes the normalization of the distance map $D(x, y)$ (Normalization *type II*, Fig. 7). It takes into account the $G_{bg}^{\overline{sw}}(x, y)$ and the $G_{sw}^{c}(x, y)$ image (Fig. 15c) which is used to handle the under-penalization when merging of adjacent components occurs (Fig. 14d). Normalization factor $N_P(x, y)$ is set according to the minimum between the $G_{bg}^{\overline{sw}}(x, y)$ value and the $G_{sw}^{c}(x, y)$ value (Fig. 15d) in order to create the final weight map $P_W(x, y)$ (Fig. 15e, 16c-16d).

$$N_P(x, y) = \begin{cases} min(G_{sw}^{c}(x, y), \quad G_{bg}^{\overline{sw}}(x, y)), \\ \qquad\qquad \text{if } G_{bg}^{\overline{sw}}(x, y) \neq 0 \\ 1, \quad \text{otherwise}. \end{cases} \tag{11}$$

Finally, $G_{sw}^{c}(x, y)$ is calculated similarly with the $G_{sw}(x, y)$ image but for the background points instead of the foreground points. In brief, the skeleton of the background is computed and at each skeleton point (i, j) the value of the distance map

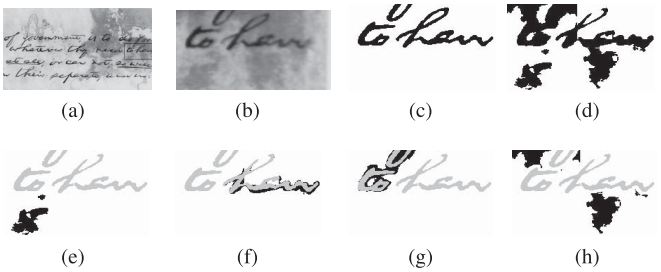


Fig. 17. (a) Original image, (b) region of (a) to demonstrate the different error types, (c) *G(x,y)* image, (d) binarization output, (e)–(h) location of the pixels classified as false alarms, character enlargement, character merging, and background noise, respectively.

$D(i, j)$ is assigned to it. Then, each point of the background inherit the value of the nearest skeleton point found.

Furthermore, foreground pixels of the binarized image $B(x, y)$ that do not correspond to foreground ground truth pixels are classified as *False Alarms* (E_{fa}), *Background Noise* (E_{bn}), *Character Enlargement* (E_{ce}) and *Character Merging* (E_{cm}). Those error measures are complementary to the pseudo-Precision ($P_{ps} + E_{fa} + E_{bn} + E_{ce} + E_{cm} = 1$) and are detailed in the following.

1) False Alarms: they refer to connected components of the binarized image under evaluation for which there is not any correct correspondence with the ground truth (Fig. 17e)

$$E_{fa}(x, y) = \frac{\sum_{x=1,y=1}^{x=I_x,y=I_y} (1 - f_b(B_{cc}(x, y))) \cdot B_W(x, y)}{\sum_{x=1,y=1}^{x=I_x,y=I_y} B_W(x, y)} \tag{12}$$

where $f_b(i)$ denotes whether ground truth foreground pixels can be detected by the i-th connected component of the binarized image $B(x, y)$:

$$f_b(i) = \begin{cases} 0, & \text{if } \sum_{\substack{x=1,y=1 \\ B_{cc}(x,y)=i}}^{x=I_x,y=I_y} G(x, y) \cdot B(x, y) = 0 \\ 1, & \text{otherwise}. \end{cases} \tag{13}$$

2) Character Enlargement: False positives pixels within the region around the ground truth components (where $P_W(x, y) \neq 1$) that are responsible for enlarging ground truth components without merging them (Fig. 17f) are quantified by the *Character Enlargement* measure (Eq. 14).

$$E_{ce}(x, y) = \frac{\sum_{\substack{x=1,y=1 \\ P_W(x,y)\neq 1}}^{x=I_x,y=I_y} d(B_{cc}(x, y)) \cdot B_W(x, y) \cdot (1 - G(x, y))}{\sum_{x=1,y=1}^{x=I_x,y=I_y} B_W(x, y)} \tag{14}$$

where $d(i)$ is a function that denotes whether the i-th connected component of $B(x, y)$ corresponds to one or more than

one ground truth connected components

$$d(i) = \begin{cases} 1, & if\ |q(i)| = 1 \\ 0, & if\ |q(i)| > 1 \end{cases} \quad (15)$$

where $q(i) = \{|G_{cc}(x, y)|,\ if\ B_{cc}(x, y) = i, 1 \le x \le I_x, 1 \le y \le I_y\}$ and $||$ denotes cardinality.

3) Character Merging: False positive pixels within the region around the ground truth components (where $P_W(x, y) \neq 1$) that are responsible for merging adjacent ground truth components (Fig. 17g) are quantified by the *Character Merging* measure (Eq. 16)

$$E_{cm}(x, y) = \frac{\sum_{\substack{x=1,y=1 \\ P_W(x,y)\neq 1}}^{x=I_x,y=I_y} (1 - d(B_{cc}(x, y))) \cdot B_W(x, y) \cdot (1 - G(x, y))}{\sum_{x=1,y=1}^{x=I_x,y=I_y} B_W(x, y)}. \quad (16)$$

4) Background Noise: It indicates the false positive pixels that are in the background region where the $P_W(x, y)$ values are equal to one (Fig. 17h).

$$E_{bn}(x, y) = \frac{\sum_{\substack{x=1,y=1 \\ P_W(x,y)=1}}^{x=I_x,y=I_y} f_b(B_{cc}(x, y)) \cdot B_W(x, y)}{\sum_{x=1,y=1}^{x=I_x,y=I_y} B_W(x, y)}. \quad (17)$$

It is worth mentioning that the proposed error measures described so far, are adapted to the ground truth components and are independent of a single font size value. On the contrary, similar error measures, such as the Touching Character Factor and the Broken Character Factor [52], [53], that were used to measure the relationship between the degradation effects, the filters and the OCR performance, were calculated as a function of a single font size that was detected from the binarized image under evaluation. For instance, "black connected components that are long and low" are considered for the Touching Character Factor while components that are three quarters of the font size are considered for the Broken Character Factor [52]. However, this type of evaluation could become problematic in documents of multiple font sizes or in handwritten documents with continuous writing.

IV. Experimental Results

In this section, representative examples are used to compare the proposed evaluation measures with representative measures. For the experiments, images from the DIBCO'09 [38] and DIBCO'11 [39] competitions were used, which contain non-uniform illumination, shadows, faint characters and other significant artefacts [54], [55]. Additionally, the consistency of those measures to the OCR evaluation results is measured for some machine-printed documents where the OCR application was feasible, while OCR ABBYY FineReader 8.1 [22] was applied with font types chosen for historical documents, e.g. "OldEnglish", "OldItalian". Furthermore, the ground truth construction procedure that was used for the DIBCO'09, H-DIBCO-10 and DIBCO'11 contests [37]–[39] is described in details, since it has not been previously detailed in any of the corresponding DIBCO papers. Finally, the proposed evaluation measures and all their complementary error measures are used to demonstrate how those measures can designate the shortcomings of the binarization method being evaluated.

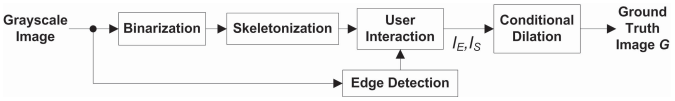


Fig. 18. Ground truth construction procedure used for the DIBCO series.

A. Ground Truth Construction Stage

For the ground truth creation from real images, there are several methodologies reported in the literature. According to our previous work [35], the original image is initially binarized and skeletonized. The user is required not only to erase all skeleton segments that correspond to false alarms of the binarization, but also to carefully draw all skeleton segments in order to form the complete skeleton for each character. The conditional dilation of the skeleton that follows, not only uses the Canny edges [56], but also the binarized image under evaluation, and hence the produced ground truth image could bias the final evaluation results. Furthermore, the variety of the local stroke width is not always preserved, since for each binarized component, the skeleton continues to dilate overpassing the edges, until half of the edges that are 'within' each binarized component are covered.

Recently, in [40] a similar procedure is proposed, as in [35]. However, in [40], the skeleton is not corrected by a user and hence skeleton segments that correspond to false alarms of the initial binarization remain. Additionally, the dilation of the skeleton is constrained by a state-of-the-art binarization result such as Sauvola [57], Gatos [2] or Lu [41]. Therefore, false alarms could remain at the ground truth image, and if faint components are entirely missed by both the initial binarization method and the final state-of-the-art method, those components are not considered part of the ground truth as they should. However, those faults in the ground truth images could be partially justified taking into account that the methodology of [40] was used for massive ground truth creation for books.

Furthermore, in [27], users marked the ground truth on real images using a 3×3 brush [58]. According to [27], the user has to paint over all the character instance and decisions concerning the boundaries of the character have to be made for all the contour of each character, increasing in this way the subjectivity degree. In [33], the user was assisted by software [59] to create the ground truth for machine-printed images, by merging and splitting clusters in the character clustering stage, as well as by adding and removing character models to degraded character instances in the character matching stage. However, the aforementioned procedure can be applied only on machine-printed documents without many different font types or many variations within the same font type (Fig. 2),

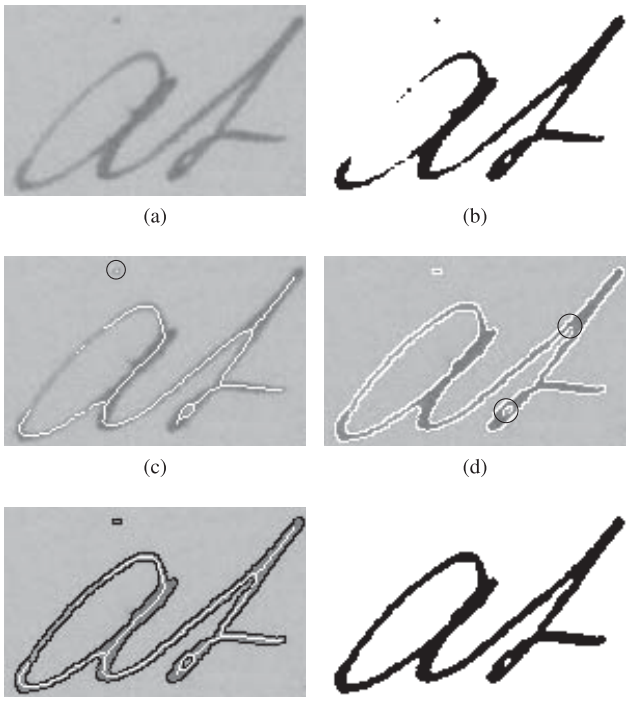


Fig. 19. (a) Original image, (b) binarization result, (c) skeletonization of (b) corresponding to the initial dilation markers (in white) and a false dilation marker that shall be erased indicated in circle, (d) edges of the image (in white) and edge disconnections indicated in circles, (e) final skeleton segments $I_S(x, y)$ (in white) and the edges $I_E(x, y)$ (in black) after user involvement, and (f) ground truth image $G(x, y)$.

since character segmentation, clustering and matching would be extremely difficult tasks. Finally, for historical documents with severe degradations, the use of global thresholding for the ground truth construction as in [32], [34], will fail.

For the ground truth construction of the images used in the DIBCO series [37]–[39], a semi-automatic procedure is followed (Fig. 18). Initially, the original image is binarized using the method of [2] which provides good results for degraded document images (Fig. 19b). The binarization result is further skeletonized using the method of [60] (Fig. 19c) and the skeleton segments are only used as dilation markers, the correct location of which is verified by the user. It should be noted that the use of the skeleton reduces the amount of the data that are to be processed by the user and from the human perspective, it provides a simplified version of the character instance [61]. In the next step, edge detection is performed on the original image using Canny's method [56] which is widely used for character boundary detection [19], [62], [63]. Finally, the user interacts with i) the skeleton and ii) the edges of the image as described in the following.

The skeleton segments are used as dilation markers, for which interactivity is required in the case of an erroneous or a missing marker. The user shall verify that at least one dilation marker exists within the borders of each ground truth component, as they are delimited by the edges (Fig. 19d). All skeleton segments produced after the aforementioned procedure form the draft skeleton image $I_S(x, y)$ (Fig. 19e). Similarly, the user shall close any edge disconnections (Fig. 19d, 19e) and

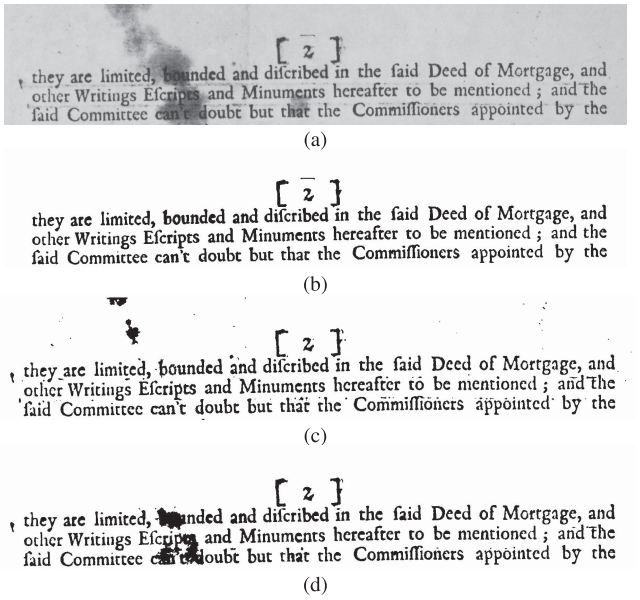


Fig. 20. (a) Original image, (b) ground truth image, (c) and (d) binarization outputs where noise is outside and inside the text area, respectively.

TABLE III
EXPERIMENTAL RESULTS USING PIXEL-BASED MEASURES

	FM	PSNR	MPM	DRD	F_{ps}
Fig. 4b	**94.37**	**22.89**	0.85	1.72	95.02
Fig. 4c	93.69	22.56	**0.07**	**1.53**	**98.24**
Fig. 6b	59.04	15.06	15.56	32.71	**60.18**
Fig. 6c	**61.37**	**15.10**	**2.28**	**29.96**	55.54
Fig. 20c	90.30	16.89	7.33	3.85	**93.38**
Fig. 20d	**91.48**	**17.37**	**0.66**	**3.39**	92.37
Fig. 21c	**98.05**	**21.00**	**0.30**	**1.13**	97.26
Fig. 21d	97.91	20.71	0.80	1.35	**98.50**

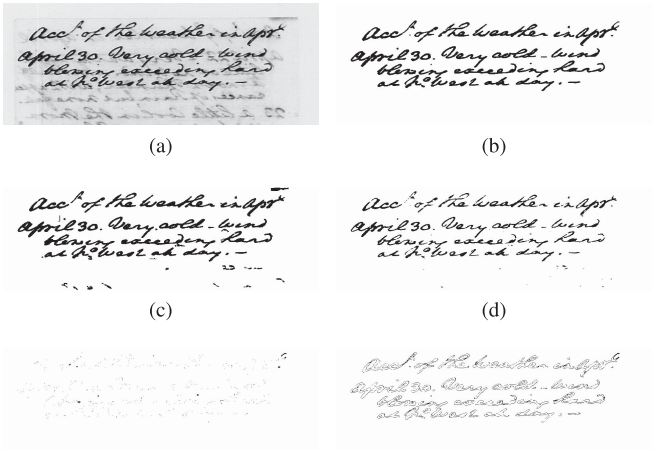
the edge image $I_E(x, y)$ is formed. Hence, decisions made by the user concerning the character boundaries are limited to the edge disconnections, decreasing in this way the subjectivity degree. In the case of great uncertainty about the correct location of the edges, it is preferable for the user to choose pixels with intensity closer to the background intensity, since Canny edges usually follow the same principle and even more, the zero penalty assigned on the ground truth contour would compensate this choice. Finally, the draft skeleton image $I_S(x, y)$ is dilated iteratively using a 3×3 cross-type mask and the dilation is constrained by the edge image $I_E(x, y)$. In this way, the ground truth image $G(x, y)$ is constructed (Fig. 19f). Example ground truth images can be found in the following sections through Fig. 20-25.

B. Comparison With Pixel-Based Evaluation Measures and OCR Consistency

Through Section II, it has been discussed that existing pixel-based evaluation measures can mislead the evaluation results. In the following, Table III presents the detailed evaluation results using all measures for the cases shown in Fig. 4 and Fig. 6 of Section II, as well as for other representative cases

(a) (b) (c) (d)

Fig. 21. (a) Original image, (b) ground truth, (c) and (d) binary image with eight and two characters missing (in grey), respectively.



(a) (b) (c) (d) (e) (f)

Fig. 22. (a) Original image, (b) ground truth, (c) and (d) example binarization outputs, (e) and (f) all missing pixels from the binarization outputs (c) and (d), respectively.

described below. Moreover, for the experiments, the proposed measures of pseudo-Recall R_{ps} and pseudo-Precision P_{ps} are combined into the pseudo-FMeasure F_{ps} following the same formula as F-Measure (Eq. 18).

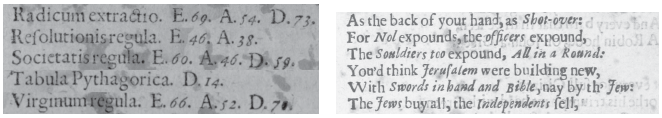
$$F_{ps} = \frac{2 \cdot R_{ps} \cdot P_{ps}}{R_{ps} + P_{ps}}. \tag{18}$$

In Fig. 20, a binarization output containing a stain that destroys the textual information (Fig. 20d) and a binarization output containing noise of similar size but far from the text preserving the useful information (Fig. 20c) are demonstrated. The distance-based MPM measure [38] gives high penalty to the false alarms of Fig. 20c because of their high distance from the ground truth and DRD measure [48] gives low penalty to foreground pixels that are erroneously inserted near/between the characters. According to the proposed methodology, the aforementioned shortcomings of the distance based-measures are eliminated by the definition of the areas around the ground truth components and the proposed normalization, during the pseudo-Precision calculation.

In Fig. 21, the loss of information from characters of different sizes is demonstrated. The distance-based measures favour the loss of small characters since the missing textual information is closer to the ground truth borders. It is clear that according either to MPM or DRD, the loss of eight characters (Fig. 21c) is less penalized than the loss of only two characters (Fig. 21d). Non-distance pixel-based evaluation measures (e.g. F-Measure, PSNR) are prone to erroneous evaluation results

TABLE IV
EXPERIMENTAL RESULTS OF FIG. 22(c) AND (d)

	FM	F_{ps}	Recall	R_{ps}	Precision	P_{ps}
Fig. 22c	92.58	92.48	97.37	99.00	88.23	86.76
Fig. 22d	82.47	96.85	70.55	94.61	99.23	99.19



(a) (b)

Fig. 23. (a) and (b) Representative historical images used for the OCR experiments.

(Table III) since all the textual information has the same importance for the evaluation. On the contrary, the proposed measure F_{ps} successfully managed to rank in a better position the binary image that has missed two characters rather than eight smaller characters. Those improved results rely upon the normalization of the distance weights according to the local stroke width during the pseudo-Recall calculation.

In the following, an example case of two binarization outputs (Fig. 22c-22d) that differ mainly around the character borders is presented. In this case, Recall is misleading the evaluation results (Table IV) because of the lack of tolerance around the ground truth borders. In both cases, the binarization example output shown in Fig. 22c achieves better Recall but due to the higher number of false alarms and enlarged characters (e.g. 'e' and 'b' are filled) it is ranked second according to the proposed measures. Additionally, the binarization example output shown in Fig. 22d is penalized by Recall because it differs from the ground truth image only by the contour pixels (Fig. 22f). This can also be noticed from Table IV where the binarization example output shown in Fig. 22d has very low Recall 70.5% while the pseudo-Recall (R_{ps}) is over 94.6%. The evaluation results using the proposed measures are more accurate since characters are penalized when they are broken (or much textual information is missing) and not when missing pixels correspond to the boundaries of the ground truth.

In the following, the OCR is included in the experiments. Additional images used for those experiments are shown are shown in Fig. 23. Furthermore, eight state-of-the-art global and adaptive binarization techniques were used, as listed below.

1) Adaptive Logical method (AL) [8], 2) Bernsen method (BER) [64], 3) ABBYY FineReader 8.1 (FR) [22], 4) Gatos method (GPP) [2], 5) Kim method (KIM) [7], 6) Niblack method (NIB) [65], 7) Otsu method (OTS) [66], 8) Sauvola method (SAU) [57]. In Fig. 24, an example is shown in which the binarization results are placed according to the ranking based upon the OCR Accuracy [67] (Eq. 19), while the corresponding evaluation results are presented in Table V. It should be noted that the OCR Accuracy equals 100 if perfect match is achieved between the text produced by the OCR and the ground truth text, while it could be even negative, e.g. -100, if all the necessary corrections, i.e. inserted/deleted/substituted characters, are twice the ground truth characters. Additionally, to compare the ranking lists of

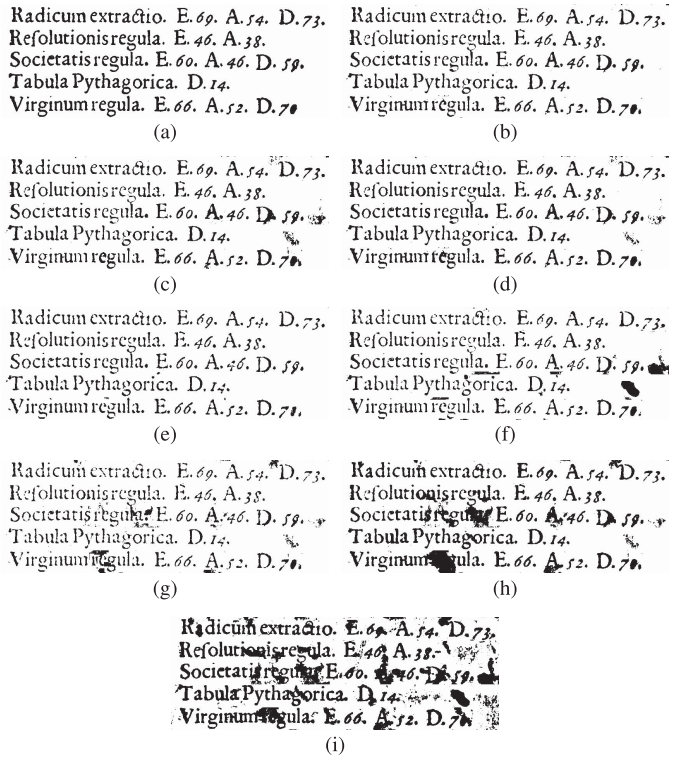


Fig. 24. (a) Ground truth of Fig. 23(a). (b)-(i) Binarization results according to the OCR ranking (i.e. GPP, KIM, SAU, AL, FR, BER, OTS, NIB, respectively).

TABLE V
RANKING USING DIFFERENT EVALUATION MEASURES FOR THE IMAGES OF FIG. 24

Rank	OCR Acc.	F_{ps}	FM	PSNR	MPM	DRD
1	**GPP** 78.44	**GPP** 96.14	**GPP** 88.17	**GPP** 14.91	**GPP** 0.861	**GPP** 2.922
2	**KIM** 78.44	**SAU** 94.48	**SAU** 87.52	**SAU** 14.58	**AL** 1.035	**SAU** 3.356
3	**SAU** 76.65	**AL** 93.48	**KIM** 85.68	**KIM** 14.01	**SAU** 6.054	**KIM** 3.971
4	**AL** 73.65	**KIM** 93.24	**AL** 83.63	**AL** 13.70	**KIM** 9.151	**AL** 4.031
5	**FR** 71.86	**FR** 88.23	**OTS** 81.94	**OTS** 12.31	**BER** 10.54	**BER** 6.519
6	**BER** 67.07	**BER** 87.57	**FR** 78.17	**FR** 12.23	**OTS** 15.21	**FR** 6.571
7	**OTS** 58.68	**OTS** 81.59	**BER** 76.40	**BER** 12.07	**FR** 25.44	**OTS** 6.963
8	**NIB** 40.12	**NIB** 67.39	**NIB** 71.27	**NIB** 9.30	**NIB** 78.61	**NIB** 15.72

the different evaluation measures, the Kendall's tau (τ) ranking correlation coefficient is used [68], that range from 1 in the case of perfect match, to -1 in the case of completely reversed list.

$$OCR\_Accuracy = 100 \cdot \frac{Correct\ Characters}{Ground\ Truth\ Characters} \qquad (19)$$

where *Correct Characters = Ground Truth Characters-(Insertions+Substitutions+Deletions)*.

TABLE VI
OCR OUTPUT AND OCR ACCURACY

	OCR Output	OCR Acc.
Fig. 24b	Radicum extraoho. E.69. A.f4. D.73. Refolutionisrec'ula. E 46. A.js. Socictatis regula. E. 60. A. 46.1). j9. Tabula Pythagorica. D. 14. Virginumreguia. E 66. D.70,	78.44
Fig. 24c	Radicum extradho. E.69. A./4/D.7J. Refolutionis recula. E.46. A.js. Societatis regula. E.60. k.,46. I> S9 Tabula .Pythagorica. D. 14. ˆ Virginum regula. E. 66. D. 7*	78.44
Fig. 20c	÷ C 2 J ĵhey are limited, **bounded** and diicribed in the (aid Deed of Mortgage, and other Writings **Efcripts** and Minuments hereafter to be mentioned; andlhc iaid Committee **can't doubt** but that the Commifnoncrs appointed by the	92.86
Fig. 20d	. they are limited, **ĵpndcd** and diicribed in the iaid Deed of Mortgage, and other Writings **EicntMA** and Minuments hereafter to be mentioned ; andlhe iaid Committee **c#rt!&ouSt** but that the Commidioners appointed by the	87.50
Fig. 6b	POWER gu RESEARCH DSPABTKSNT SAN FEANCISCO 1937	81.82
Fig. 6c	*	0.0

From the experiments (Fig. 24, Table V) it is shown high correlation between the proposed evaluation methodology and the OCR evaluation. The last four techniques, i.e. FR, BER, OTS and NIB, produce more distinctive results (Fig. 24f- 24i) and most of the introduced noise is among the text and hence that noise cannot be bypassed by the OCR. Therefore, the majority of the binarization errors are considered by both the OCR and the pixel-based measures and comparison can be performed in a better basis. For those techniques, the rankings of the OCR Accuracy and the F_{ps}, are identical. A critical mistake that all the measures except for the F_{ps} and DRD have made is that the OTS binarization (Fig. 24h) is ranked in a better place than the FR binarization (Fig. 24f). Moreover, according to the MPM and DRD measure, BER (Fig. 24g) is better than FR (Fig. 24f) and this is explained by the fact that the noise is closer to the characters of BER than the characters of FR. However, the noise among the characters of BER deteriorates the OCR and the legibility.

On the contrary, the first four techniques produce binarization results for which the text has been better binarized and any introduced noise does not interfere with the text. Therefore, the introduced noise could be bypassed by the OCR. For example, the GPP and KIM results shown in Fig. 24b-24c, achieve the same OCR Accuracy (Table V), although the GPP has better erased the background noise than KIM. Table VI demonstrates the OCR output along with the OCR Accuracy for those cases, as well as for the cases of Fig. 6 and Fig. 20.

In Table VI, the bold text of the cases shown in Fig. 20c-20d, indicate the location where the text is covered by stains.

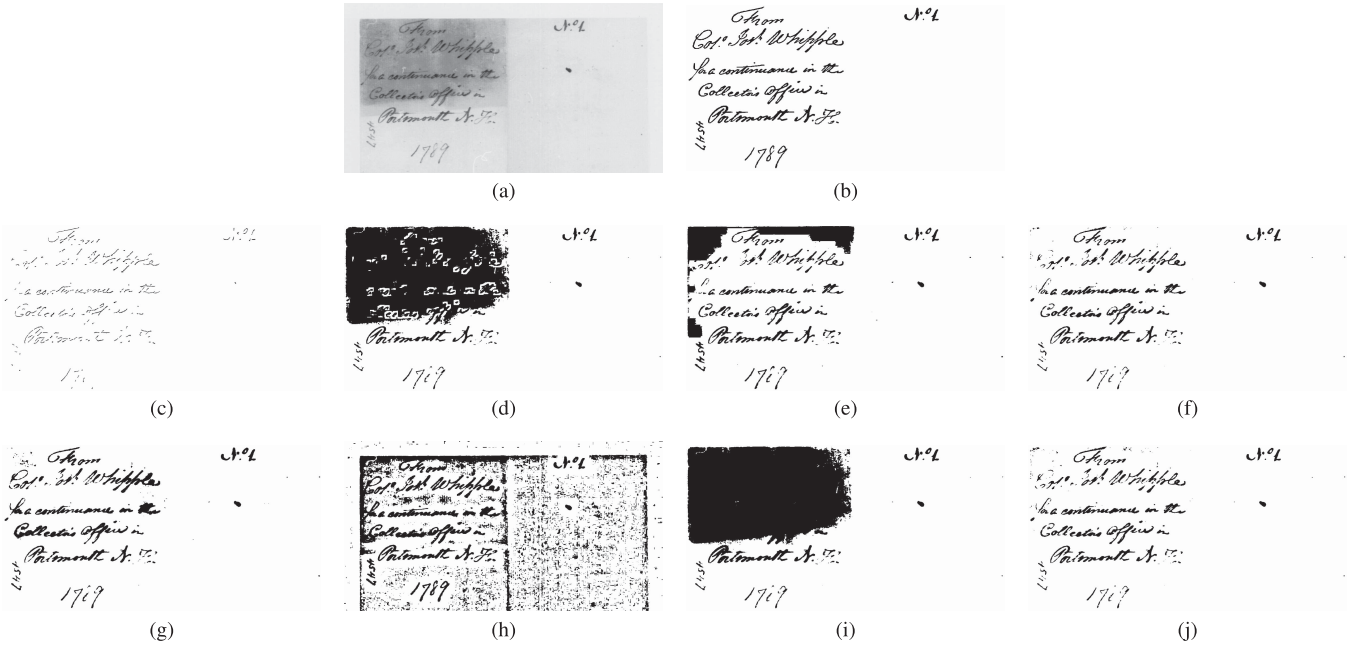


Fig. 25. (a) Original image, (b) ground truth, (c) AL [8] binarization, (d) BER [64] binarization, (e) FR [22] binarization, (f) GPP [2] binarization, (g) KIM [7] binarization, (h) NIB [65] binarization, (i) OTS [66] binarization, and (j) SAU [57] binarization.

TABLE VII
RANKING USING AVERAGE VALUES FOR ALL MEASURES FOR 8 HISTORICAL DOCUMENTS ALONG WITH THE KENDALL'S TAU (τ) COEFFICIENT

Rank	OCR Acc.	F_{ps}	FM	PSNR	MPM	DRD
1	**GPP**	**GPP**	**GPP**	**GPP**	**AL**	**GPP**
	73.52	93.99	88.31	15.36	4.129	3.862
2	**KIM**	**SAU**	**SAU**	**SAU**	**GPP**	**SAU**
	72.72	93.01	87.83	15.07	05.27	4.219
3	**SAU**	**AL**	**KIM**	**KIM**	**SAU**	**KIM**
	70.53	92.88	86.12	14.45	7.662	4.559
4	**AL**	**KIM**	**OTS**	**AL**	**KIM**	**AL**
	67.97	92.73	84.56	14.10	08.78	4.628
5	**FR**	**FR**	**AL**	**OTS**	**BER**	**OTS**
	66.59	88.20	83.64	13.82	11.89	6.604
6	**BER**	**BER**	**FR**	**FR**	**OTS**	**FR**
	66.47	86.81	81.88	13.45	13.05	6.914
7	**OTS**	**OTS**	**BER**	**BER**	**FR**	**BER**
	60.04	85.69	79.64	13.11	13.37	7.354
8	**NIB**	**NIB**	**NIB**	**NIB**	**NIB**	**NIB**
	41.51	70.66	72.53	09.97	88.33	17.60
τ	1	0.857	0.714	0.786	0.571	0.786

The binarized image of Fig. 20c had better removed the stains among the text and it is evaluated better than the binarized image of Fig. 20d only in terms of OCR Accuracy and of the proposed F_{ps} measure (Table III). Similar conclusions are drawn for the binarized images shown in Fig. 6b-6c. Notice that the OCR has failed to identify text areas at the binarized image of Fig. 6c, since the textual content is covered by noise.

Overall, Table VII presents the average values of OCR-Accuracy and of other pixel-based evaluation measures for 8 historical machine-printed images. These are 24-bit color images (Fig. 23) with average size of 1270 $\times$ 280 approximately containing 152 characters on average. From both Table V and Table VII it is depicted that the ranking of the F_{ps} is closer to the ranking of the OCR Accuracy than the rankings of the other measures. The F_{ps} ranking list achieves the highest correlation in terms of Kendall's tau (τ) coefficient with the OCR ranking list. The OCR Accuracy and the F_{ps} agree to the same ranking for all techniques except for KIM. However, in Table VII, the F_{ps} values of GPP and AL are quite close (93.99 and 92.88), while there is a bigger gap between the corresponding values of OCR Accuracy (73.52 and 67.97). This can be explained by the fact that the potential range of the OCR Accuracy is wider, while the range of the proposed F_{ps} is limited between 0 and 100. Indeed, according to the statistics shown in Table VIII, OCR Accuracy has a much wider range than F_{ps}; its minimum and maximum values are 0 and 91.50, respectively, while the corresponding values of F_{ps} are 59.89 and 98.43.

TABLE VIII
STATISTICS PER MEASURE USING ALL BINARIZATION METHODS ON 8 HISTORICAL DOCUMENTS

	OCR Acc.	F_{ps}	FM	PSNR	MPM	DRD
Average	64.92	88.00	83.07	13.67	19.06	7.08
St. dev.	19.64	9.72	8.22	2.18	28.79	5.74
Min	0.00	59.89	61.14	8.79	0.52	2.11
Max	91.52	98.43	94.09	18.58	120.80	31.76

C. Evaluation Analysis Using the Proposed Measures

To present the performance evaluation analysis when the proposed evaluation measures are used, experiments were

TABLE IX
EVALUATION RESULTS OF FIG. 25

	AL	BER	FR	GPP	KIM	NIB	OTS	SAU
F_{ps}	52.46	27.06	56.87	87.37	81.82	33.93	25.46	87.21
R_{ps}	35.59	94.72	86.89	85.91	92.20	99.26	96.54	87.02
E_{bt}	46.97	3.47	9.03	10.26	5.06	0.31	2.56	9.41
E_{pmt}	15.50	1.80	3.81	3.83	2.74	0.43	0.90	3.57
E_{fmt}	1.95	0.00	0.26	0.00	0.00	0.00	0.00	0.00
P_{ps}	99.74	15.79	42.27	88.88	73.54	20.46	14.67	87.40
E_{cm}	0.00	22.45	3.83	0.00	7.53	6.53	28.57	0.00
E_{ce}	0.02	0.12	3.97	9.33	16.50	8.73	0.92	9.87
E_{fa}	0.24	0.72	6.11	1.79	2.31	44.57	0.62	2.73
E_{bn}	0.00	60.92	43.82	0.00	0.11	19.71	55.23	0.00

TABLE X
EVALUATION RESULTS OF FIG. 25 USING PREVIOUS METHOD [35]

	AL	BER	FR	GPP	KIM	NIB	OTS	SAU
FM	62.56	29.11	57.69	87.99	84.39	35.11	28.06	87.93
Rec.	45.57	94.52	86.58	85.50	91.85	99.22	96.32	86.65
Br.	52.32	5.48	13.29	14.50	8.15	0.78	3.68	13.35
Mis.	2.11	0.00	0.13	0.00	0.00	0.00	0.00	0.00
Prec.	99.74	17.21	43.25	90.64	78.05	21.33	16.42	89.24
Mer.	0.00	81.90	46.53	0.00	6.02	22.98	82.07	0.00
Enl.	0.01	0.11	3.97	7.54	13.50	9.26	0.82	7.97
F.Al.	0.24	0.78	6.25	1.82	2.43	46.43	0.69	2.79

conducted using all images from DIBCO'09 contest [38] and the eight state-of-the-art global and adaptive binarization techniques as presented in the previous section IV-B. Representative results of those binarization techniques are given in Fig. 25. The corresponding evaluation results are given in Table IX and Table X using the proposed and the previous method [35], respectively. Table XI presents the average performance of the binarization algorithms.

The proposed measures are useful for analysing the evaluation results and studying the benefits and drawbacks of each technique. For instance, in AL (Fig. 25c), the background is correctly eliminated but at the cost of increased thinning, broken and lost characters (Tables IX-XI). Moreover, as demonstrated in Fig. 25c, the binarization of AL produces very thin characters and that cannot be measured by the previous method [35] in which the skeletonized ground truth was used for the computation of Recall. As a consequence, the AL receives a better rating than the FR using the previous method [35] (Table X) even though the FR has better preserved the textual information (Fig. 25e). Additionally, from Fig. 25e and Tables IX-X more detailed error analysis is performed than [35] in which the merging rate is 46.53% (Table X) and hence almost half of the characters are expected to be merged. However, only three characters are merged and the additional noise is located in the background. According to the proposed evaluation results (Table IX), this case is better described with merging rate 3.83% and background noise 43.82%.

The behavior of the binarization techniques that our methodology implies can be verified by other works. For example,

TABLE XI
AVERAGE EVALUATION RESULTS CONCERNING ALL TEST IMAGES

	AL	BER	FR	GPP	KIM	NIB	OTS	SAU
F_{ps}	85.09	83.89	85.17	91.85	90.63	56.80	82.21	90.09
R_{ps}	79.39	96.10	95.66	95.09	96.11	99.83	98.46	95.53
E_{bt}	14.90	2.14	2.29	3.04	1.77	0.05	0.76	2.64
E_{pmt}	5.50	1.74	2.00	1.87	2.13	0.12	0.77	1.84
E_{fmt}	0.21	0.02	0.04	0.00	0.00	0.00	0.02	0.00
P_{ps}	91.66	74.44	76.76	88.83	85.73	39.70	70.57	85.24
E_{cm}	0.04	3.59	0.85	0.20	1.77	5.24	7.74	0.43
E_{ce}	1.97	2.48	3.95	6.37	6.89	8.95	6.28	6.51
E_{fa}	6.26	6.63	8.99	4.46	5.32	39.79	3.38	7.65
E_{bn}	0.07	12.87	9.46	0.14	0.29	6.32	12.04	0.18

in [69] it is stated that, "The Bernsen method is better (than Otsu), but there are artefacts and broken strokes". Indeed, BER is better than OTS (Table XI) even if it has more broken text E_{bt} and more false alarms E_{fa}. The aforementioned attributes of BER are verified by other works [6], [70].

V. CONCLUSION

In this work, a pixel-based evaluation methodology for document image binarization techniques was presented, with a particular focus on historical documents containing complex degradations and complex font types. Two new measures were defined, namely pseudo-Recall and pseudo-Precision, that make use of the distance from the contour of the ground truth to minimize the penalization around the character borders, as well as the local stroke width of the ground truth components to provide improved document-oriented evaluation results. The proposed measures can be used to record the binarization performance in a better and more efficient way than other pixel-based measures and have also better correlation with the OCR. In addition, useful error measures (broken and missed text, character enlargement and merging, background noise and false alarms) were defined that make more evident the weakness of each binarization technique being evaluated.

REFERENCES

[1] A. Antonacopoulos and D. Karatzas, "Semantics-based content extraction in typewritten historical documents," in *Proc. Int. Conf. Document Anal. Recognit.*, 2005, pp. 48–53.

[2] B. Gatos, I. Pratikakis, and S. J. Perantonis, "Adaptive degraded document image binarization," *Pattern Recognit.*, vol. 39, no. 3, pp. 317–327, Mar. 2006.

[3] A. Antonacopoulos, D. Karatzas, H. Krawczyk, and B. Wiszniewski, "The lifecycle of a digital historical document: Structure and content," in *Proc. ACM Symp. Document Eng.*, Milwaukee, WI, 2004, pp. 147–154.

[4] A. Antonacopoulos and D. Karatzas, "Document image analysis for world war II personal records," in *Proc. Int. Workshop Document Image Anal. Libraries*, 2004, pp. 336–341.

[5] B. Lamiroy, D. Lopresti, and T. Sun, "Document analysis algorithm contributions in end-to-end applications: Report on the ICDAR 2011 contest," in *Proc. Int. Conf. Document Anal. Recognit.*, Beijing, China, Sep. 2011, pp. 1521–1525.

[6] D. Trier and T. Taxt, "Evaluation of binarization methods for document images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 17, no. 3, pp. 312–315, Mar. 1995.

[7] I. K. Kim, D. W. Jung, and R. H. Park, "Document image binarization based on topographic analysis using a water flow model," *Pattern Recognit.*, vol. 35, no. 1, pp. 265–277, 2002.

[8] Y. Yang and H. Yan, "An adaptive logical method for binarization of degraded document images," *Pattern Recognit.*, vol. 33, no. 5, pp. 787–807, 2000.

[9] E. Kavallieratou and S. Stathis, "Adaptive binarization of historical document images," in *Proc. Int. Conf. Pattern Recognit.*, vol. 3. Hong Kong, Aug. 2006, pp. 742–745.

[10] J. G. Kuk, N. I. Cho, and K. M. Lee, "Map-MRF approach for binarization of degraded document image," in *Proc. IEEE Int. Conf. Image Process.*, San Diego, CA, Oct. 2008, pp. 2612–2615.

[11] J. G. Kuk and N. I. Cho, "Feature based binarization of document images degraded by uneven light condition," in *Proc. Int. Conf. Document Anal. Recognit.*, Jul. 2009, pp. 748–752.

[12] Q. Huang, W. Cao, and W. Cai, "Thresholding technique with adaptive window selection for uneven lighting image," *Pattern Recognit. Lett.*, vol. 26, no. 6, pp. 801–808, 2005.

[13] J. He, Q. D. M. Do, A. C. Downton, and J. H. Kim, "A comparison of binarization methods for historical archive documents," in *Proc. Int. Conf. Document Anal. Recognit.*, vol. 1. Seoul, South Korea, Aug. 2005, pp. 538–542.

[14] H. Cao and V. Govindaraju, "Preprocessing of low-quality handwritten documents using Markov random fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 7, pp. 1184–1194, Jul. 2009.

[15] M. Ramirez, E. Tapia, M. Block, and R. L. Rojas, "Quantile linear algorithm for robust binarization of digitalized letters," in *Proc. Int. Conf. Document Anal. Recognit.*, vol. 2. Curitiba, Brazil, Sep. 2007, pp. 1158–1162.

[16] A. Dawoud and M. S. Kamel, "Iterative multimodel subimage binarization for handwritten character segmentation," *IEEE Trans. Image Process.*, vol. 13, no. 9, pp. 1223–1230, Sep. 2004.

[17] H. Cecotti and A. Belaid, "Dynamic filters selection for textual document image binarization," in *Proc. Int. Conf. Pattern Recognit.*, Tampa, FL, Dec. 2008, pp. 1–4.

[18] L. Likforman-Sulem, J. Darbon, and E. H. B. Smith, "Pre-processing of degraded printed documents by non-local means and total variation," in *Proc. Int. Conf. Document Anal. Recognit.*, Barcelona, Spain, Jul. 2009, pp. 758–762.

[19] Z. Zhou, L. Li, and C. L. Tan, "Edge based binarization for video text images," in *Proc. Int. Conf. Pattern Recognit.*, Istanbul, Turkey, Aug. 2010, pp. 133–136.

[20] Y. T. Pai, Y. F. Chang, and S. J. Ruan, "Adaptive thresholding algorithm: Efficient computation technique based on intelligent block detection for degraded document images," *Pattern Recognit.*, vol. 43, no. 9, pp. 3177–3187, 2010.

[21] R. F. Moghaddam and M. Cheriet, "A multi-scale framework for adaptive binarization of degraded document images," *Pattern Recognit.*, vol. 43, no. 6, pp. 2186–2198, 2010.

[22] *ABBYY FineReader OCR*. (2011) [Online]. Available: http://finereader.abbyy.com/

[23] P. Stathis, E. Kavallieratou, and N. Papamarkos, "An evaluation survey of binarization algorithms on historical document images," in *Proc. Int. Conf. Pattern Recognit.*, Dec. 2008, pp. 1–4.

[24] R. Paredes, E. Kavallieratou, and R. D. Lins, "ICFHR 2010 contest: Quantitative evaluation of binarization algorithms," in *Proc. Int. Conf. Frontiers Handwrit. Recognit.*, Kolkata, India, Nov. 2010, pp. 733–736.

[25] R. Hedjam, R. F. Moghaddam, and M. Cheriet, "A spatially adaptive statistical method for the binarization of historical manuscripts and degraded document images," *Pattern Recognit.*, vol. 44, no. 9, pp. 2184–2196, 2011.

[26] E. H. B. Smith, L. Likforman-Sulem, and J. Darbon, "Effect of pre-processing on binarization," *Proc. SPIE*, vol. 7534, p. 75340H, Jan. 2010.

[27] E. H. B. Smith, "An analysis of binarization ground truthing," in *Proc. Int. Workshop Document Anal. Syst.*, Boston, MA, Jun. 2010, pp. 27–33.

[28] B. Bataineh, S. N. H. S. Abdullah, and K. Omar, "An adaptive local binarization method for document images based on a novel thresholding method and dynamic windows," *Pattern Recognit. Lett.*, vol. 32, no. 14, pp. 1805–1813, 2011.

[29] M. Sezgin and B. Sankur, "Survey over image thresholding techniques and quantitative performance evaluation," *J. Electron. Imag.*, vol. 13, no. 1, pp. 146–168, 2004.

[30] L. Y. Fan, C. L. Tan, and L. X. Fan, "Edge-preserving prefiltering for document image binarization," in *Proc. Int. Conf. Image Process.*, vol. 1. Thessaloniki, Greece, Oct. 2001, pp. 1070–1073.

[31] F. Kleber, M. Diem, and R. Sablatnig, "Scale space binarization using edge information weighted by a foreground estimation," in *Proc. Int. Conf. Document Anal. Recognit.*, Beijing, China, Sep. 2011, pp. 1180–1184.

[32] C. Bastos, C. Mello, J. Andrade, D. Falcao, M. Lima, W. Santos, and A. Oliveira, "Thresholding images of historical documents with back-to-front interference based on color quantization by genetic algorithms," in *Proc. IEEE Int. Conf. Tools Artif. Intell.*, vol. 1. Patras, Greece, Oct. 2007, pp. 488–491.

[33] T. Obafemi-Ajayi, G. Agam, and O. Frieder, "Ensemble LUT classification for degraded document enhancement," *Proc. SPIE*, vol. 6815, pp. 681509-1–681509-9, Jan. 2008.

[34] R. D. Lins and J. M. M. da Silva, "Assessing algorithms to remove back-to-front interference in documents," in *Proc. Int. Telecommun. Symp.*, Fortaleza, Brazil, 2006, pp. 868–873.

[35] K. Ntirogiannis, B. Gatos, and I. Pratikakis, "An objective evaluation methodology for document image binarization techniques," in *Proc. Int. Workshop Document Anal. Syst.*, Nara, Japan, Sep. 2008, pp. 217–224.

[36] A. Clavelli, D. Karatzas, and J. Llados, "A framework for the assessment of text extraction algorithms on complex colour images," in *Proc. Int. Workshop Document Anal. Syst.*, 2010, pp. 19–26.

[37] I. Pratikakis, B. Gatos, and K. Ntirogiannis, "H-DIBCO 2010 - handwritten document image binarization competition," in *Proc. Int. Conf. Frontiers Handwrit. Recognit.*, Kolkata, India, Nov. 2010, pp. 727–732.

[38] B. Gatos, K. Ntirogiannis, and I. Pratikakis, "DIBCO 2009: Document image binarization contest," *Int. J. Document Anal. Recognit.*, vol. 14, no. 1, pp. 35–44, 2011.

[39] I. Pratikakis, B. Gatos, and K. Ntirogiannis, "ICDAR 2011 document image binarization contest - DIBCO 2011," in *Proc. Int. Conf. Document Anal. Recognit.*, Beijing, China, Sep. 2011, pp. 1506–1510.

[40] I. B. Messaoud, H. E. Abed, V. Maergner, and H. Amiri, "A design of a preprocessing framework for large database of historical documents," in *Proc. Workshop Historical Document Imag. Process.*, Beijing, China, Sep. 2011, pp. 177–183.

[41] S. Lu, B. Su, and C. L. Tan, "Document image binarization using background estimation and stroke edges," *Int. J. Document Anal. Recognit.*, vol. 13, no. 4, pp. 303–314, 2010.

[42] H. S. Baird, "Document image defect models," in *Proc. IAPR Workshop Syntactic Struct. Pattern Recognit.*, Murray Hill, NJ, Jun. 1990, pp. 38–46.

[43] T. Kanungo, R. M. Haralick, and I. Phillips, "Global and local document degradation models," in *Proc. Int. Conf. Document Anal. Recognit.*, Tsukuba, Japan, Oct. 2003, pp. 730–734.

[44] E. H. B. Smith, "Characterization of image degradation caused by scanning," *Pattern Recognit. Lett.*, vol. 19, no. 13, pp. 1191–1197, 1998.

[45] P. Sarkar, H. S. Baird, and X. Zhang, "Training on severely degraded text-line images," in *Proc. Int. Conf. Document Anal. Recognit.*, vol. 1. Edinburgh, Scotland, Aug. 2003, pp. 38–43.

[46] T. K. Ho and H. S. Baird, "Large-scale simulation studies in image pattern recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 10, pp. 1067–1079, Oct. 1997.

[47] E. Badekas and N. Papamarkos, "Automatic evaluation of document binarization results," in *Proc. Iberoamer. Congr. Pattern Recognit.*, Havana, Cuba, Nov. 2005, pp. 1005–1014.

[48] H. Lu, A. C. Kot, and Y. Q. Shi, "Distance-reciprocal distortion measure for binary document images," *IEEE Signal Process. Lett.*, vol. 11, no. 2, pp. 228–231, Feb. 2004.

[49] D. R. Martin, C. C. Fowlkes, and J. Malik, "Learning to detect natural image boundaries using local brightness, color, and texture cues," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 5, pp. 530–549, May 2004.

[50] R. Kasturi, D. Goldgof, P. Soundararajan, V. Manohar, J. Garofolo, R. Bowers, M. Boonstra, V. Korzhova, and J. Zhang, "Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics, and protocol," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 319–336, Feb. 2009.

[51] K. Ntirogiannis, B. Gatos, and I. Pratikakis, "A modified adaptive logical level binarization technique for historical document images," in *Proc. Int. Conf. Document Anal. Recognit.*, Barcelona, Spain, 2009, pp. 1171–1175.

[52] D. K. Reed and E. H. B. Smith, "Correlating degradation models and image quality metrics," *Proc. SPIE*, vol. 6815, pp. 681508-1–681508-8, Jan. 2008.

[53] A. Souza, M. Cheriet, S. Naoi, and C. Y. Suen, "Automatic filter selection using image quality assessment," in *Proc. Int. Conf. Document Anal. Recognit.*, 2003, pp. 508–512.

[54] *Library of Congress*. (2011) [Online]. Available: http://memory.loc.gov/ammem
[55] *IMPACT Project*. (2011) [Online]. Available: http://www.impact-project.eu
[56] J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 8, no. 6, pp. 679–698, Nov. 1986.
[57] J. Sauvola and M. Pietikainen, "Adaptive document image binarization," *Pattern Recognit.*, vol. 33, no. 2, pp. 225–236, 2000.
[58] E. Saund, J. Lind, and P. Sarkar, "Pixlabeler: User interface for pixel-level labeling of elements in document images," in *Proc. Int. Conf. Document Anal. Recognit.*, 2009, pp. 646–650.
[59] G. Bal, G. Agam, and O. Frieder, "Interactive degraded document enhancement and ground truth generation," *Proc. SPIE*, vol. 6815, pp. 68150Z-1–68150Z-9, Mar. 2008.
[60] H. J. Lee and B. Chen, "Recognition of handwritten Chinese characters via short line segments," *Pattern Recognit.*, vol. 25, no. 5, pp. 543–552, May 1992.
[61] A. Dawoud and M. Kamel, "New approach for the skeletonization of handwritten characters in grey-level images," in *Proc. Int. Conf. Document Anal. Recognit.*, vol. 2. Edinburgh, Scotland, Aug. 2003, pp. 1233–1237.
[62] J. van Beusekom, F. Shafait, and T. M. Breuel, "Automated OCR ground truth generation," in *Proc. Int. Workshop Document Anal. Syst.*, Nara, Japan, Sep. 2008, pp. 111–117.
[63] J. Zhang, D. Goldgof, and R. Kasturi, "A new edge-based text verification approach for video," in *Proc. Int. Conf. Pattern Recognit.*, Tampa, FL, Dec. 2008, pp. 1–4.
[64] J. Bernsen, "Dynamic thresholding of grey-level images," in *Proc. Int. Conf. Pattern Recognit.*, 1986, pp. 1251–1255.
[65] W. Niblack, *An Introduction to Digital Image Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1986, pp. 115–116.
[66] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst., Man, Cybern.*, vol. 9, no. 1, pp. 62–66, Jan. 1979.
[67] S. V. Rice, "Measuring the accuracy of page-reading systems," Ph.D. dissertation, Dept. Comput. Sci., Univ. Nevada, Las Vegas, 1996.
[68] M. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, nos. 1–2, pp. 81–89, 1938.
[69] J. Bai, Y. Q. Yang, and R. L. Tian, "Complicated image's binarization based on method of maximum variance," in *Proc. Int. Conf. Mach. Learn. Cybern.*, Dalian, China, Aug. 2006, pp. 3782–3785.
[70] M. Pan, F. Zhang, and H. Ling, "An image binarization method based on HVS," in *Proc. IEEE Int. Conf. Multimedia Expo*, Beijing, China, Dec. 2007, pp. 1283–1286.

Konstantinos Ntirogiannis received the Degree from the Department of Informatics and Telecommunications, National and Kapodistrian University, Athens, Greece, in 2006, where he is currently pursuing the Ph.D. degree, conducting his doctoral research in collaboration with the National Center for Scientific Research "Demokritos," Athens.

He is currently involved in research on historical document image processing with a particular focus on document image binarization and its performance evaluation. He has participated in the EU funded project IMPACT (IMProving ACcess to Text) (http://www.impact-project.eu/).

Basilis Gatos received the Ph.D. degree with thesis "Optical Character Recognition Techniques."

He was a Director in digital preservation of old newspapers with the Research Division, Lambrakis Press Archives, from 1998 to 2001. He was a Managing Director in document management and recognition with the R&D Division, BSI S.A., Athens, Greece, from 2001 to 2003. He is currently a Researcher with the Institute of Informatics and Telecommunications, National Center for Scientific Research "Demokritos," Athens, Greece. He has participated in several research programs funded by the European community. His current research interests include image processing and document image analysis, OCR, and pattern recognition.

Dr. Gatos is a member of the Technical Chamber of Greece and the Editorial Board of the *International Journal on Document Analysis and Recognition*. He is a Program Committee Member of several international conferences and workshops.

Ioannis Pratikakis (SM'12) received the Ph.D. degree in 3-D image analysis from the Electronics and Informatics Engineering Department, Vrije Universiteit Brussels, Brussels, Belgium, in 1999.

He is currently an Assistant Professor with the Department of Electrical and Computer Engineering, Democritus University of Thrace, Xanthi, Greece. From 1999 to 2000, he was with IRISA/ViSTA Group, Rennes, France, as an INRIA Post-Doctoral Fellow. From 2003 to 2010, he was an Adjunct Researcher with the Institute of Informatics and Telecommunications, National Centre for Scientific Research "Demokritos," Athens, Greece. His current research interests include image processing, pattern recognition, vision and graphics, document image analysis and recognition, medical image analysis, multimedia content analysis, and search and retrieval with a particular focus on visual content.

Dr. Pratikakis is a member of the European Association for Computer Graphics (Eurographics). He has served as a Co-Chair of the Eurographics Workshop on 3-D object retrieval from 2008 to 2009 as well as a Guest Editor for the Special issue on 3-D object retrieval at the *International Journal of Computer Vision*. He has been a member of the Board of the Hellenic Artificial Intelligence Society from 2010 to 2012.