An Objective Evaluation Methodology for Handwritten Image Document Binarization Techniques

K. Ntirogiannis, B. Gatos and I. Pratikakis
Computational Intelligence Laboratory,
Institute of Informatics and Telecommunications,
National Center for Scientific Research "Demokritos",
GR-153 10 Agia Paraskevi, Athens, Greece
http://www.iit.demokritos.gr/cil,
{kntir,bgat,ipratika}@iit.demokritos.gr

Abstract

This paper presents an objective evaluation methodology for handwritten document image binarization techniques that aims to reduce the human involvement in the ground truth construction and consecutive testing. A detailed description of the methodology along with a benchmarking of the state-of-the-art binarization algorithms based on the proposed methodology is presented.

Keywords: binarization, handwritten, objective evaluation.

1. Introduction

Document image binarization is an important step in the document image analysis and recognition pipeline. The performance of a binarization technique directly affects the required analysis or recognition outcome. Therefore, it is imperative to have an objective evaluation which will account for the performance of the binarization.

Several efforts have been presented that strive towards evaluating the performance of binarization techniques. These efforts can be divided in three categories. In the first category evaluation is performed by a human evaluator [1], [2], [3] while in the second category, the binary result is subject to OCR and the corresponding result is evaluated with respect to accuracy [4], [5], [6]. The third category uses a combination of human-oriented evaluation and OCR results accuracy [7].

Evaluation performed by a human expert is not only subjective but also time consuming. Furthermore, it lacks robustness since it has been observed that in fuzzy situations, the same observer may make different selections for the same dataset in different sessions. The use of OCR as a means for evaluation is widely used in the evaluation of modern printed documents only, supported by contemporary OCR engines (e.g. ABBYY FineReader).

In this paper, we propose an objective evaluation methodology for handwritten image document binarization techniques that avoids the dependence of an OCR engine which cannot be applied to handwritten documents [1] and reduce the human interference based upon a semi-automated ground truth construction.

Figure 1 shows all stages of the proposed methodology. Each stage is analyzed in the following sections. Specifically, the construction of the ground truth image is described in Section 2. The evaluation metrics of recall and precision are detailed in Sections 3 and 4 respectively. The evaluation results of representative state-of-the-art binarization techniques are discussed in Section 5 and Section 6 contains the conclusion of our methodology.

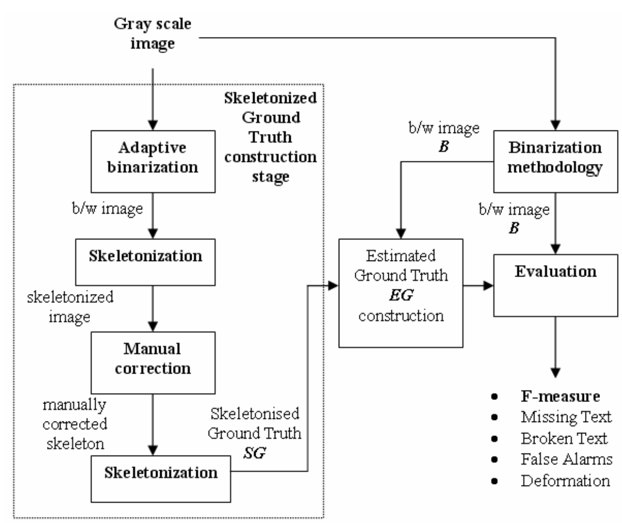


Figure 1. The proposed methodology stages

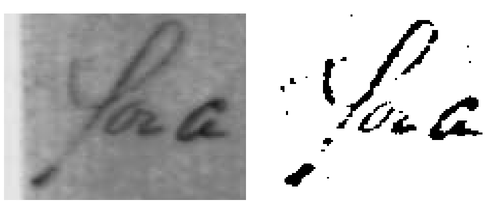
2. Construction of Ground Truth Image

In the proposed methodology, the construction of ground truth plays an important role, since it helps to automate the evaluation procedure. It consists of two distinct steps, namely *skeletonized Ground Truth* (SG) stage and *Estimated Ground Truth* (EG) stage. These stages will be described in detail, in the following sections.

2.1. Skeletonized Ground Truth Image

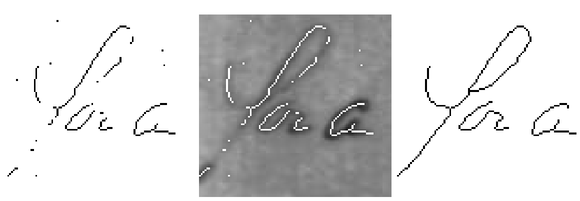
The idea of building a skeletonized ground truth image originates from a user's natural interaction with a character that is required to be presented by its silhouette. For this task, a user would directly draw a one pixel wide curve approximately in the medial axis of the character. Our proposed skeletonized ground truth image construction stage strives toward automating the aforementioned procedure. To accomplish this, we follow the consecutive stages described in the following.

The grey scale image (see Figure 2a) is binarized using an adaptive binarization technique (see Figure 2b). Then, a skeletonization method [8] is used and the resulting skeleton image has one pixel wide text (see Figure 3a). Due to artifacts in the character skeletonization does not always represent the complete character. In this case the user is required to delineate the remaining character or remove spurious parts. To further aid the user during the correction stage, we show both the layers of the skeleton and the grey level image to guide him/her in the correction process (see Figure 3b). Finally, a second skeletonization pass guaranties that ground truth text one pixel wide (see Figure 3c).



(a) (b)

Figure 2. (a) Original image and (b) binary image for skeletonization



(a) (b) (c)

Figure 3. (a) Binary image after skeletonization (skeleton image); (b) simultaneous viewing of skeleton and grey level image layer; (c) skeletonized ground truth image which text is one pixel wide

The skeletonized ground truth image (see Figure 3c) and the respective connected component labeled image are defined by the following equations:

$$SG(x,y)=\begin{cases}0, & \text{background}\\ 1, & \text{text}\end{cases} \tag{1}$$

$$SGC(x,y)=\begin{cases}0, & if\ \ SG(x,y)=0\\ i, & otherwise\end{cases} \tag{2}$$

where $i \in K$, K = {1, 2…, M} and M denotes the number of the connected components found in ground truth image.

After the end of the skeletonized ground truth construction stage, we are able to automatically measure the performance of any binarization algorithm in terms of recall (see Section 3).

2.2. Estimated Ground Truth Image

To complete the evaluation process we should calculate the performance of binarization algorithms in terms of precision (see Section 4). Precision requires considering ground truth characters as much close as the original ones. In this paper we present a methodology to automatically estimate the ground truth for the computation of precision taking into account that a skeletonized ground truth image has been achieved.

Given the skeletonized ground truth image, we apply a dilation constrained by the edge image ***CE*** (see Figure 4a) and the binary image ***B*** under evaluation (see Figure 4b) where

$$CE(x,y)=\begin{cases}1, & if\ \ (x,y)\in \partial I(x,y)\\ 0, & otherwise\end{cases} \tag{3}$$

$$B(x,y)=\begin{cases}0, & \text{background}\\ 1, & \text{text}\end{cases} \tag{4}$$

where $\partial I(x,y)$ denotes the resulting image after Canny Edge detection [9] on the original grey scale image *I(x,y).*

(a) (b) (c)

Figure 4. (a) Edges of original image; (b) binary image under evaluation and (c) estimated ground truth image painted in grey

For the sake of clarity, we provide the algorithm in pseudo code for the estimated ground truth image stage. For a better understanding, the following definitions should be considered.

Let ***A*** be a binary image then $\boldsymbol{A}^{r}$ is the respective dilated image after r iterations. Let ***BC*** denotes the connected component labeled image of ***B***, defined as:

$$BC(x,y) = \begin{cases} 0, & if\ B(x,y) = 0 \\ j, & otherwise \end{cases} \tag{5}$$

where $j \in L$, L = {1, 2…, N} and N denotes the number of the connected components found in the binarized image.

Algorithm description

1. $A^0(x,y) = SG(x,y)$,
   $(x,y) \in B(x,y)$: $BC(x,y) = i$ AND $\sum_{BC(x,y)=i} B(x,y) \cdot SG(x,y) > 0$
2. Stop = false
3. *while* (*NOT*(*Stop*))
4. $A^{r+1}(x,y) = (A^r(x,y) \oplus B(x,y)) \cap B(x,y)$
5. *if* $\dfrac{\sum_{BC(x,y)=i} A^{r+1}(x,y) \cdot CE(x,y)}{\sum_{BC(x,y)=i} CE(x,y)} > \dfrac{1}{2}$ OR $A^{r+1}(x,y) = A^r(x,y)$
6. Stop = true
7. End *if*
8. End *while*

where $i \in L$ and $\oplus$ denotes dilation.

The aforementioned procedure is applied for each component of the binary image ***B*** and the estimated ground truth image ***EG*** is fully constructed, as defined in Eq. 6. An example estimated ground truth image is shown in Figure 4c.

$$EG(x,y) = \begin{cases} 0, & \text{background} \\ 1, & \text{text} \end{cases} \tag{6}$$

3. Recall

Recall is defined as the percentage of the skeletonized ground truth image ***SG*** that is detected in the resulting binary image ***B***. Recall is given by the following formula:

$$\text{Recall} = \frac{\sum_{x=1,y=1}^{x=Ix,y=Iy} SG(x,y) \cdot B(x,y)}{\sum_{x=1,y=1}^{x=Ix,y=Iy} SG(x,y)} \ 100\ \% \tag{7}$$

All ***SG*** parts that are not detected can be classified as broken or missing text. Broken text is related with ***SG*** components that are partially detected while missing text denotes all ***SG*** components that have not been detected at all (see Figure 5d). In the following sections, we further analyze the recall result by calculating broken and missing texts information.

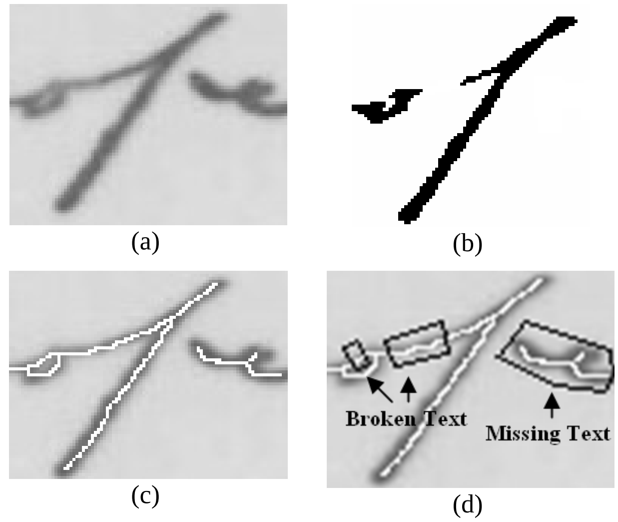


Figure 5. (a) Original gray scale image; (b) resulting binary image ***B***; (c) skeletonized ground truth ***SG*** image overlay and (d) broken and missing text parts

3.1. Broken Text

Let function *f* denotes whether part of a skeletonized ground truth component is partially detected in the binary image ***B***. *f is* calculated as follows:

$$f(i) = \begin{cases} 1, & if \sum_{\substack{x=1,y=1 \\ SGC(x,y)=i}}^{x=Ix,y=Iy} SG(x,y) \cdot B(x,y) > 0 \\ 0, & otherwise \end{cases} \tag{8}$$

where $i \in K$.

Broken text can be defined by the percentage of the skeletonized ground truth image ***SG*** parts which are not detected in the resulting binary image ***B*** while belonging to components that are partial detected (see Figure 5d). Broken text is given by the following equation:

$$Broken = \frac{\sum_{x=1,y=1}^{x=Ix,y=Iy} f(SGC(x,y)) \cdot (1 - B(x,y))}{\sum_{x=1,y=1}^{x=Ix,y=Iy} SG(x,y)} 100\% \tag{9}$$

3.2. Missing Text

Missing text is defined by the percentage of the skeletonized ground truth image ***SG*** parts which are not detected in the resulting binary image ***B*** while belonging to components that are not detected at all (see Figure 5d). Missing text is given by the following equation:

$$Missing = \frac{\sum_{x=1,y=1}^{x=Ix,y=Iy} (1 - f(SGC(x,y))) \cdot SG(x,y)}{\sum_{x=1,y=1}^{x=Ix,y=Iy} SG(x,y)} 100\% \quad (10)$$

4. Precision

Considering the binary image ***B***, precision estimates the foreground that is actually text. In our method, the actual text is considered as the estimated ground truth image ***EG*** as described in Section 2.

Precision is as defined the percentage of the estimated ground truth image that is detected in the binary image (see Figure 6c).

$$\text{Precision} = \frac{\sum_{x=1,y=1}^{x=Ix,y=Iy} EG(x,y) \cdot B(x,y)}{\sum_{x=1,y=1}^{x=Ix,y=Iy} B(x,y)} 100\% \quad (11)$$

The foreground of the binary image that is not detected during precision estimation is considered as either false alarms or deformations which are described in the following section.

4.1. False Alarms

False alarms refer to foreground pixels of the binary image ***B*** that do not belong to estimated ground truth image (see Figure 6c). They are defined by the percentage of all pixels of the components of the binary image ***B*** that do not have any corresponding pixel with the skeletonized ground truth image ***SG***.

$$FAlarms = \frac{\sum_{x=1,y=1}^{x=Ix,y=Iy} h(BC(x,y)) \cdot B(x,y)}{\sum_{x=1,y=1}^{x=Ix,y=Iy} B(x,y)} 100\% \quad (12)$$

where *h(i)* is a function denoting whether a binary component is not detected in the skeletonized ground truth image.

$$h(i) = \begin{cases} 1, & if \sum\limits_{\substack{x=1,y=1 \\ BC(x,y)=i}}^{x=Ix,y=Iy} SG(x,y) \cdot B(x,y) = 0 \\ 0, & otherwise \end{cases} \quad (13)$$

where $i \in L$

4.2. Deformations

Components often merge with adjacent background information that was recognized as text during binarization. Deformations do not only enlarge (deform) components but they are also responsible for merging adjacent components (see Figure 6c).

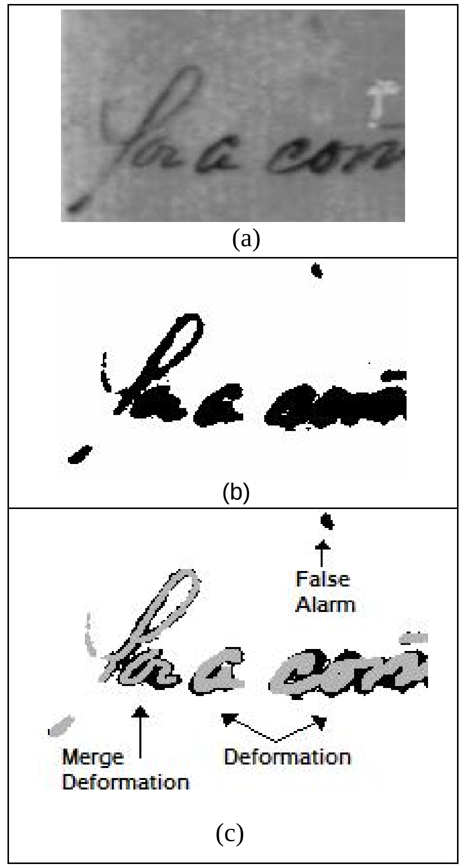


Figure 6. (a) Original gray scale image; (b) resulting binary image *B*; (c) estimated ground truth image is demonstrated in grey while false alarms and deformations in black.

In our method, deformations of the binary image are defined by the percentage of all text pixels of the binary image ***B*** that are not detected in the estimated ground truth image ***EG*** and do not belong to false alarms components as described in the previous section.

The deformation of the binary image is defined by Eq. 14 while deformation leading to merging is defined by Eq. 15.

$$Deform = \frac{\sum_{x=1,y=1}^{x=Ix,y=Iy} d(BC(x,y)) \cdot B(x,y) \cdot (1 - EG(x,y))}{\sum_{x=1,y=1}^{x=Ix,y=Iy} B(x,y)} 100\% \quad (14)$$

$$MergeDeform = \frac{\sum_{x=1,y=1}^{x=Ix,y=Iy} m(BC(x,y)) \cdot B(x,y) \cdot (1-EG(x,y))}{\sum_{x=1,y=1}^{x=Ix,y=Iy} B(x,y)} 100\% \quad (15)$$

where *d(i)* and *m(i)* are functions denoting whether a binary component corresponds to one or more ground truth components respectively. Functions *d(i)* and *m(i)* are defined as follows:

$$d(i) = \begin{cases} 1, & if \quad |\ SGC(x,y)\ | = 1 \\ & \forall (x,y):\ BC(x,y) = i \\ 0, & otherwise \end{cases} \quad (16)$$

$$m(i) = \begin{cases} 1, & if \quad |\ SGC(x,y)\ | > 1, \\ & \forall (x,y):\ BC(x,y) = i \\ 0, & otherwise \end{cases} \quad (17)$$

where | *SGC(x, y)* | denotes the Cardinality of *SGC(x, y)*.

5. Results

The proposed objective evaluation methodology for handwritten image document binarization techniques was applied on a range of gray scale handwritten documents with low quality, shadows, non-uniform illumination, strains, presence of the handwriting from the other side of the page and other significant artifacts. Among all documents, we selected the ten (10) most representative and marked the skeletonized ground truth *SG* following the procedure described in section 2. Six (6) of the most promising global and adaptive binarization techniques were chosen for evaluation:

1. Otsu's method (OTS) [10]
2. Bernsen's method (BER) [11]
3. Niblack's method (NIB) [12]
4. Sauvola's method (SAU) [13]
5. Adaptive Logical method (AL) [14]
6. Adaptive Degraded Document method (ADD) [15]

An example of the application of all methodologies to a gray scale document image is given in Figure 7 where the skeletonized ground truth image ***SG*** is also demonstrated.

Our evaluation is based on F-Measure which is defined as follows:

$$\text{F-Measure} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (18)$$

Table 1, presents the evaluation results where the average values for all test images have been taken into account. According to these results, Adaptive Degraded Document method (ADD) [15] had the best overall performance and performed slightly better than the Sauvola's method (SAU) [13].

To provide an overall picture of the evaluations measures used, we give at Table 1, a detailed analysis of each of the factors which contribute to recall (broken and missing text) and precision (false alarms and deformations).

Table 1. The average values of all evaluation metrics concerning all test images for every binarization technique.

	ADD	AL	BER	NIB	OTS	SAU
F-Measure	**85.23**	81.99	77.70	50.87	77,02	84.61
Recall	85.36	79.82	87.83	**98.47**	89.4	86.51
Precision	87.98	**88.83**	75.86	35.42	73.59	85.89
Broken	14.04	19.53	11.90	**01.42**	10.30	12.89
Missing	00.60	00.65	00.27	**00.01**	00.30	00.60
MergeDeform	01.05	**00.29**	13.42	12.93	15.36	00.61
Deform	10.02	08.94	**06.89**	09.14	08.96	10.22
FAlarms	**00.95**	01.94	03.83	42.51	02.09	03.28

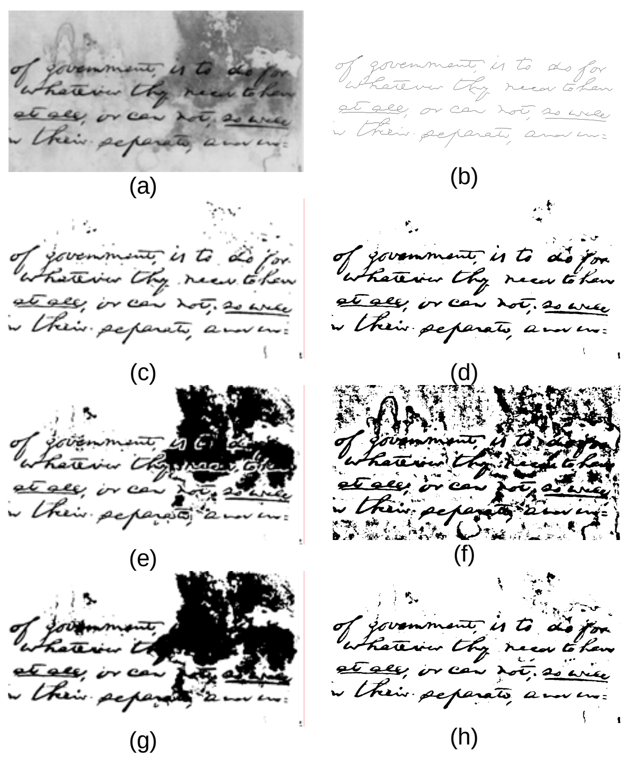


Figure 7. (a) Original Image, (b) skeletonized ground truth image, (c) AL result image, (d) ADD result image, (e) BER result image, (f) NIB result image, (g) OTS result image and (h) SAU result image

6. Conclusions

This work presents an objective evaluation methodology for handwritten document image binarization techniques. It is based on a semi-automatic procedure for the construction of the ground truth as well as a fully automated evaluation scheme.

Acknowledgements

The research leading to these results has received funding from the European Community's Seventh Framework Programme under grant agreement n° 215064 (project IMPACT).

References

[1] E. Kavallieratou and S. Stathis, “Adaptive Binarization of Historical Document Images”, *18th International Conference on Pattern Recognition (ICPR’06)*, Hong Kong, China, 2006, vol.3, pp. 742-745.

[2] D. Trier and T. Taxt, “Evaluation of Binarization Methods for Document Images”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, March 1995, vol.17, no.3, pp. 312-315.

[3] Q. Wang and C. L. Tan, “Matching of Double Sided Document Images to Remove Interference”, *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR ’01)*, Kauai island of Hawaii, USA, 2001, vol.1, pp. 1084-1089.

[4] M. Sezgin and B. Sankur, “Survey over Image Thresholding Techniques and Quantitative Performance Evaluation”, *Journal of Electronic Imaging*, January 2004, vol. 13, no. 1, pp. 146-168.

[5] J. He, Q. D. M. Do, A.C. Downton, and J.H Kim, “A Comparison of Binarization Methods for Historical Archive Documents”, *Proceedings of the Eighth International Conference on Document Analysis and Recognition (ICDAR ’05)*, Seoul, South Korea, 2005, vol. 1, pp. 538-542.

[6] Y. Zhu, C. Wang, and R. Dai, “Document image Binarization Based on Stroke Enhancement”, *Proceedings of the 18th International Conference on Pattern Recognition (ICPR ’06)*, Hong Kong, China, 2006, vol. 1, pp. 955-958.

[7] F. Chang, K. Liang, T. Tan, and W. Hwang, “Binarization of Document Images Using Hadamard Multiresolution Analysis”, *Proceedings of the Fifth International Conference on Document Analysis and Recognition (ICDAR ’99)*, Bangalore, India, 1999, pp. 157-160.

[8] H. J. Lee, and B. Chen, “Recognition of Handwritten Chinese Characters via Short Line Segments”, *Pattern Recognition*, 1992, vol. 25, no. 5, pp. 543-552.

[9] J. Canny, “A computational Approach to Edge Detection”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, November 1986, vol.8, no.6, pp. 679-698.

[10] N. Otsu, “A Thresholding Selection Method from Gray-level Histogram”, *IEEE Transactions on Systems, Man and Cybernetics*, March 1979, vol. 9, pp. 62-66.

[11] J. Bernsen, “Dynamic Thresholding of Grey-level Images”, *Proceedings of the Eighth International Conference on Pattern Recognition*, Paris, France, 1986, pp. 1251-1255.

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

[13] J. Sauvola, and M.Pietikainen, “Adaptive Document Image Binarization”, *Pattern Recognition*, 2000, vol. 33, no. 2, pp. 225-236.

[14] Y. Yang, and H. Yan, “An Adaptive Logical Method for Binarization of Degraded Document Images”, *Pattern Recognition*, 2000, vol. 33, no. 5, pp. 787-807.

[15] B. Gatos, I. Pratikakis, and S. J. Perantonis, “Adaptive Degraded Document Image Binarization”, *Pattern Recognition*, March 2006, vol. 39, no. 3, pp. 317-327.