2012 10th IAPR International Workshop on Document Analysis Systems

Effect of "Ground Truth" on Image Binarization

Elisa H. Barney Smith
Boise State University
Boise, Idaho, USA
EBarneySmith@BoiseState.edu

Chang An
Lehigh University
Bethlehem, Pennsylvania, USA
cha305@Lehigh.edu

***Abstract*—Image binarization has a large effect on the rest of the document image analysis processes in character recognition. Algorithm development is still a major focus of research. Evaluation of image binarization has been done by comparison of the result of OCR systems on images binarized by different methods. That has been criticized in that the binarization alone is not evaluated, but rather how it interacts with the downstream processes. Recently pixel accurate "ground truth" images have been introduced for use in binarization algorithm evaluation. This has been shown to be open to interpretation. The choice of binarization ground truth affects the binarization algorithm design, either directly if design is by automated algorithm trying to match the provided ground truth, or indirectly if human designers adjust their designs to perform better on the provided data. Three variations in pixel accurate ground truth were used to train a binarization classifier. The performance can vary significantly depending on choice of ground truth, which can influence binarization design choices.**

***Keywords*-Image Binarization, Ground Truthing, Degraded document images, Performance Evaluation**

I. Introduction

Optical Character Recognition (OCR) involves more than classification algorithms. All stages in the process contribute to the automatic recognition. This includes the acquisition, binarization, page segmentation or zoning, character segmentation, feature selection, training data, classification algorithm, and possibly post-processing. Many of these stages seem like they should be straight forward, and yet they have been researched for a long time, and continue to be topics of research. For ICDAR 2011 Lamiroy et al. [8] evaluated OCR system components in an end-to-end environment and asked groups to contribute new algorithms to see what component would have the greatest effect on the final recognition results. Two groups participated, contributing five total algorithms for evaluation. The conclusion was that while the choice of OCR algorithms had the greatest effect on results of the algorithms contributed for consideration in the contest, a binarization algorithm had the most significant positive impact of contributed algorithms on improving the end-to-end performance.

Early seminal papers evaluating document image binarization looked at the performance of the OCR when the images were binarized by different algorithms to determine which binarization algorithm performed the best [15]. This has been criticized as being a metric of how well the binarization output fits with the remainder of the OCR processing, and not a direct measure of the binarization algorithm itself. Later work generated synthetic images and degraded them with a degradation algorithm [14]. These images after binarization with several algorithms were then compared to the original non-degraded images to evaluate the binarization algorithms. This however doesn't compare just the binarization results, because the effects of stroke width changes and corner erosion on the character caused by the blurring degradation were not accounted for in the 'ground truth'. The next step was the pixel accurate ground truth proposed by Ntirogiannis et al. [11]. This was created by running a binarization algorithm developed by Kamel et al. [7] on the document, skeletonizing the ground truth [9], and then doing some manual correction on the skeleton as the authors deemed necessary.

The continued interest in image binarization was highlighted at ICDAR in 2009 when 35 research groups contributed 43 algorithms for evaluation in the first Document Image Binarization Contest (DIBCO 2009). Since then several more papers have appeared in the literature on the topic, and two more document image binarization contests have been held (H-DIBCO 2010 [12] & DIBCO 2011 [13]). The DIBCO contests have in many ways changed the way people are approaching image binarization. This paper starts to explore how the presence of pixel accurate ground truth could affect algorithm development.

II. Ground Truth Development

For DIBCO 2009 a set of 2 handwritten and 2 machine printed documents were provided, each with an accompanying ground truth image. From these the competitors were to fine-tune their algorithms and submit executable code that was evaluated on a set of 5 handwritten and 5 machine printed documents which also had accompanying ground truth. The ground truth images were created based on a semi-automated procedure [11]. The data set used in DIBCO 2009 has spawned much work in image binarization. It allows a different set of questions to be asked, and perhaps answered, in the field of document image binarization. As it produces a very precise and specific dataset, down almost to the microscopic level, in a way never before available to this

978-0-7695-4661-2/12 $26.00 © 2012 IEEE
DOI 10.1109/DAS.2012.32

research community, it is having an effect on binarization algorithm development. The algorithms being developed are attempting to match that dataset and achieve high evaluation metric scores. The algorithms likely to be adopted by others will likely be those that perform well on that dataset. But as the choice of ground truth was shown in [4] to be open to interpretation, and not a single choice, as the term ground truth implies and its status as an exemplar advocates, this paper explores what effect that might have on binarization algorithm design.

III. DICE Classifier

An algorithm designed for document image content extraction (DICE) [1] was adapted for this analysis. The DICE classifier is a family of algorithms, able to find regions containing machine-printed text, handwriting, photographs, etc. in images of documents [3]. The algorithms handle a diverse set of document, image, and content types. Types of document images accepted include color, grey-level, and bilevel (black-and-white); also, many sizes or resolutions (digitization spatial sampling rates); and in a wide range of file formats (TIFF, JPEG, PNG, etc.). All images are converted into the HSL (Hue, Saturation, and Luminance) color space.

The DICE classifier operates on a trainable iterated classification technology, using a sequence of classifiers, each trained separately on the training-data results of the previous classifier, guided always by the same pixel accurate ground truth. Both training and test datasets consist of pixels labeled with their ground-truth class. The fast approximate 5 Nearest Neighbors using hashed k-d trees [5] is used for classification. Individual pixels, not regions, are classified in order to avoid the arbitrariness and restrictiveness of region shapes in page segmentation. Each pixel sample is represented by scalar features extracted by image processing of a small region centered on that pixel. The features are discussed in detail in [2]. As binarization is a form of page segmentation, separating text from background, this classifier is used to examine the effect of ground truth on the binarization output.

It is hypothesized that the characteristics of the ground truth provided in this dataset is affecting the development of binarization algorithms. It was found that when the DICE classifier was used for page segmentation, the algorithm was sensitive to the ground truthing (GT) policy, whether the GT was "loose", "tight" or pixel-accurate. It was concluded that pixel accurate ground truth provided the best segmentation results. As most classifiers are sensitive to the training data, this is not surprising. The objective is to begin to quantify how the development of a prominent ground truth dataset for image binarization might affect future binarization algorithm development. As the effect on indirect development by humans is not so easy to quantify, especially in a short period of time as has passed since the DIBCO 2009 dataset was introduced. Using a classification based binarization algorithm, while not a perfect substitute for years of human development, can mimic some of the iterated design processes humans will go through with a dataset.

IV. Experiment

For this study only two classes, machine printed text and blank space, were used for the DICE classifier. The five machine printed images that were provided as the test images for DIBCO 2009 and eight from DIBCO 2011 are used for experimentation. Partially due to ink seepage, especially for liquid inks, and significantly due to imaging system optics taking their response from an area on the paper around the sensor location, a zone of mid-range tones will be present in the gray level (or color) images, Figure 1. The places where binarization has the greatest uncertainty is along the boundaries where ink transitions to paper. This can lead to variable opinions about the ground truth [4]. Therefore in addition to considering the original machine printed DIBCO09 and DIBCO11 ground truths, two alternate ground truths were created for experimentation in this paper to simulate the effect of a difference of opinion on the ground truth. This is used to see what effect the choice of ground truth could have on image binarization. The first was by dilating the ground truth with a 3x3 structuring element, and the second by eroding the ground truth with the same structuring element. This produces one ground truth that is biased to be broad, one in the middle, and one that is very conservative.

Training of the DICE segmenter was done in a leave-one-out method on the images in the dataset under consideration. The trained segmenter was then run on the remaining gray level or color image from that set. The resulting binarized image was then compared to three possible ground truths for that image. The metric results were averaged over the 13 images. Eight metrics were used for evaluation.

A. Evaluation Metrics

There are many metrics used to evaluate the similarity (or difference) between a pair of images. Many are designed for natural scene pictures, but will return information useful for describing the difference between binary images. In the DIBCO 2009 competition four evaluation metrics were introduced: F-Measure, Negative Rate Metric, Peak SNR and Misclassification Penalty Measure. For the H-DIBCO competition in 2010 Recall was replaced with a pseudo-Recall term to produce a fifth metric, pseudo-F-Measure. DIBCO 2011 saw the introduction of the Distance Reciprocal Distortion Metric.

- *F-Measure (FM)*: This metric is the same as used in information retrieval and was used as the primary metric for [6]. F-measure is derived from the harmonic mean of Precision and Recall

$$FM = \frac{2 * Recall * Precision}{Recall + Precision}. \quad (1)$$

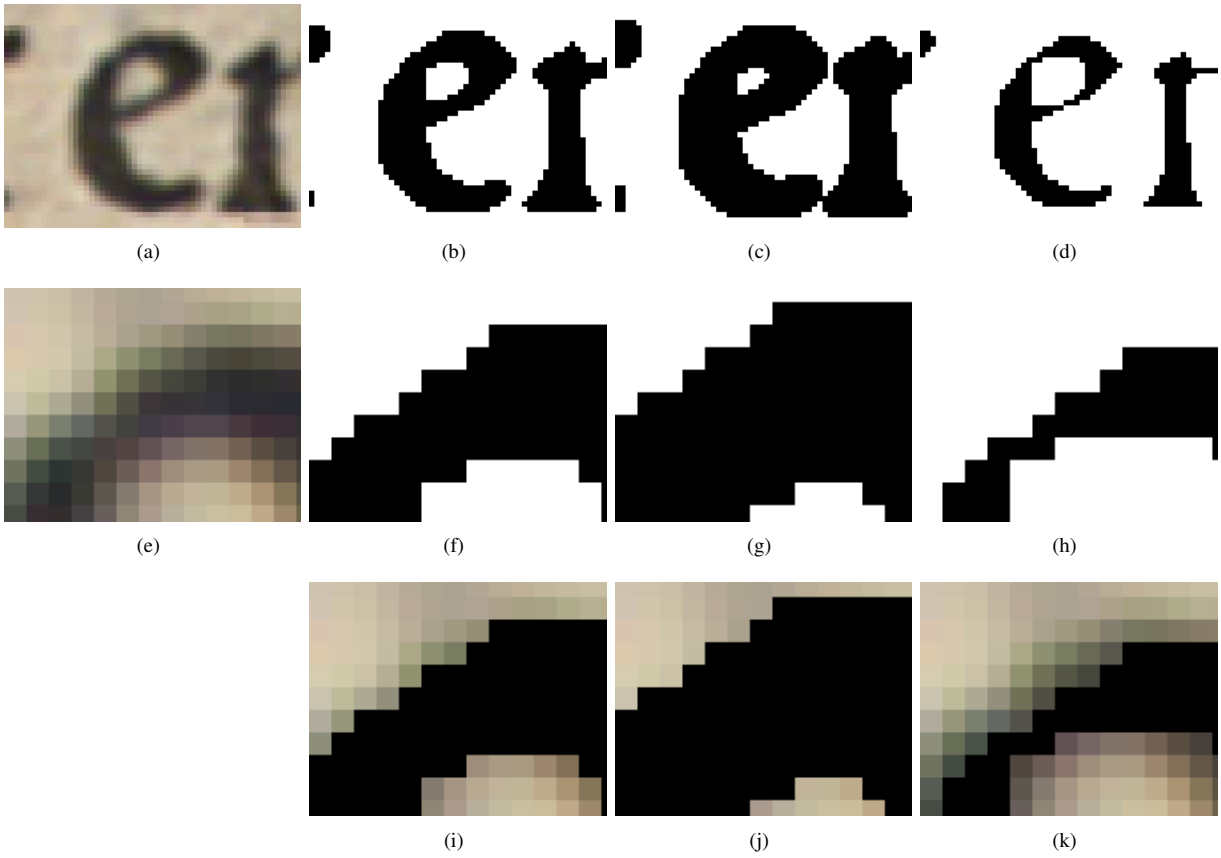

Figure 1. Example of (a) original image and possible ground truths. (b) original ground truth, (c) dilated, (d) eroded. (e)-(h) Images of a zoomed portion of the images in (a)-(d). (i)-(k) The image in (e) with the corresponding mask superimposed.

Recall is the proportion of correctly binarized foreground pixels within the true foreground pixels. When Recall is 100%, there are no false negatives and thus no ink elements were incorrectly classified as paper. Precision is the proportion of true foreground pixels within the binarized foreground pixels. When Precision is 100% there are no false positives and no paper elements were incorrectly classified as ink. A higher F-measure indicates a better match.

- *Peak SNR (PSNR)* looks at how many pixels in the test image differ from the ground truth image values, and by how much. This metric is based more directly on the image difference and is calculated by

$$PSNR = 10 * log_{10}(\frac{C^2}{MSE}), \tag{2}$$

where the Mean Square Error (MSE) is calculated from

$$MSE = \sum_{x=1}^{N}\sum_{y=1}^{M}\frac{(I_1(x,y) - I_2(x,y))^2)}{M * N} \tag{3}$$

and C is the difference between the foreground and background colors. A higher PSNR indicates a better match.

These two metrics look at misclassification of pixels in the image independent of their status as foreground, background or border pixels. Because border pixels are the ones that are hardest to definitively label as foreground or background, and have the greatest variance in labeling by human evaluators, three other metrics have been introduced that consider the location of an error pixel relative to the character boundary.

- *pseudo-F-Measure (p-FM)*. The H-DIBCO2011 contest used pseudo-Recall which was proposed in DAS 2008 [11]. The ground truth is skeletonized and the pixels in the skeleton are used as foreground pixels in the calculation of pseudo-Recall. Pseudo-Recall is used together with Precision to calculate the pseudo-F-Measure using Equation 1.
- *Distance-Reciprocal Distortion metric (DRD)* was proposed in 2004 [10]. The error is weighted by how far the flipped pixels are from other character pixels. The weighting is based on a 5x5 square where the values are the reciprocal of the Euclidean distance from the center. This metric was shown to be correlated with

human perception of degradation level. A low DRD score denotes that the algorithm is good at binarization.

- *Misclassification penalty metric (MPM).* Misclassified pixels are penalized by their l_∞ distance from the ground truth object's border. The distances are normalized by the sum of the pixel-to-contour distances of the GT object across the background of the image. A low MPM score denotes that the algorithm is good at the global binarization, such as with stains and lighting issues, and doesn't penalize border errors.

B. Results

The results of the experiment are shown in Tables I and II. The F-measure and the Peak SNR measures both show that the binarization method designed (trained) with the ground truth specified performed best on that type of ground truth. In pattern recognition, this is to be expected, but its appearance here highlights the possible effects of having a dominant ground truth.

The recall results shows that for each test policy, the highest recall is achieved by using the dilated ground truth for training and the highest precision is with an eroded ground truth training. This is as expected based on the definitions of precision and recall.

Table I
BINARIZATION EVALUATION BY FOUR METRICS. RESULTS BASED ON ERODED (E) ORIGINAL (O) AND DILATED (D) DIBCO 2009 & 2011 MACHINE PRINT GROUND TRUTH (GT).

Recall		Training Source		
		E	O	D
GT	E	77.84	95.89	**97.19**
	O	58.66	86.97	**92.43**
	D	43.48	66.45	**77.93**

Precision		Training Source		
		E	O	D
GT	E	**69.81**	53.46	42.85
	O	**90.10**	82.90	68.36
	D	**92.85**	88.66	80.23

F-Measure		Training Source		
		E	O	D
GT	E	**72.01**	66.64	56.92
	O	69.00	**82.85**	75.93
	D	57.11	73.62	**76.03**

Peak SNR		Training Source		
		E	O	D
GT	E	**14.09**	11.93	9.82
	O	12.04	**14.12**	11.95
	D	9.47	10.83	**10.89**

Table II shows the results for metrics that are designed to consider the character borders. For none of the metrics does the classifier perfectly prefer the ground truth on which it was trained. As the edge pixels contribute less penalty, and edge pixels are what changes between ground truth sets in this paper, this is reasonable. However the metrics are not consistent in their choices.

Table II
BINARIZATION EVALUATION BY AN ADDITIONAL FOUR METRICS. RESULTS BASED ON ERODED (E) ORIGINAL (O) AND DILATED (D) DIBCO GROUND TRUTH (GT).

Pseudo-Recall		Training Source		
		E	O	D
GT	E	80.42	96.22	**97.35**
	O	78.56	95.58	**97.03**
	D	74.19	91.61	**94.88**

Pseudo-F-Measure		Training Source		
		E	O	D
GT	E	**73.18**	66.81	56.99
	O	82.77	**87.10**	77.83
	D	81.28	**88.58**	84.74

Distance Recriprocal Distortion Metric		Training Source		
		E	O	D
GT	E	**11.32**	19.71	35.61
	O	12.85	**12.28**	23.07
	D	21.07	**17.23**	22.49

Misclass. Penalty Metric		Training Source		
		E	O	D
GT	E	**2.96**	8.17	15.66
	O	**3.11**	6.92	13.47
	D	**5.18**	7.48	12.92

Pseudo-Recall, similar to regular Recall favors training with a dilated GT. It produces a much greater response than regular Recall when the original or dilated GT is used in evaluation. As these two cases are both strong, the p-FM favors the training by original GT even when testing on dilated GT.

DRD and MPM imply that the eroded GT is better for overall performance based on the lowest of the nine scores. The three edge tolerant metrics all favor a thin GT and will penalize algorithms that produce wider strokes.

The results in Tables I and II do indicate that the original GT provided is better than the globally eroded or dilated results. It does not evaluate when a portion of the image stroke is broader or narrower.

V. CONCLUSION

As binarization is an important step for most OCR systems, it is crucial that it be implemented effectively. If a paradigm shift in the development of those algorithms is being introduced into the Document Image Analysis community, its possible effects need to be known.

The DICE classifier is designed to segment documents and while binarization is a type of segmentation, these documents are not the type for which it was designed, and the features used were not modified to specifically fit this dataset or to be particularly effective on stains or show through. The F-Measure performance on DIBCO'09 data alone would rank it 8^{th} in that contest. Its overall performance is not as good as the top algorithms in the DIBCO contests, still it is a useful tool to see how different binarization ground truths affect binarization choices on the pixel level. The performance of the classifier on DIBCO 2009 images and on DIBCO 2011 images was different. This needs to be evaluated more carefully to see the cause.

The results of these experiments do show that if F-Measure and Peak SNR are primary metrics, it is likely that differences in opinion about character edge boundaries will appear in binarization algorithms directly following the GT. The other metrics will cause a bias in the algorithm choice, but not in a direct fashion.

Humans design their algorithms in ways that share some characteristics with the DICE classifier. They check their intermediate algorithms performance against a dataset and adjust it hoping that it will also work better on as yet unseen datasets or images. And while the exact dataset may not specifically play a role in the binarization algorithm parameters or processes, it will have an influence on the design.

Mixing pixel level results with overall system performance is likely a better way to evaluate the binarization algorithm. Perhaps having available end-to-end systems like the DAE system will allow the binarization algorithms to be tested in a 'goal directed' fashion in conjunction with the pixel accurate ground truth. The multiple follow-on stages that are available might mitigate the worry that only the specific interaction is being evaluated and not the power of the binarization algorithm itself.

Future studies should try other classifier based binarization algorithms, or modify the classification features to better support binarization goals. Long term the fit of binarization results to DIBCO datasets and the influence on algorithm development should be monitored. With the uncertainty of pixel accurate ground truth, the binarization algorithm effectiveness should be correlated with OCR accuracy.

REFERENCES

[1] C. An and H. S. Baird. High recall document content extraction. In *Proc. Document Recognition and Retrieval*, page 787405, Burlingame, CA, USA, 2011.

[2] H. S. Baird, M. A. Moll, and C. An. Document image content inventories. In *Proc. Document Recognition and Retrieval XIV*, pages 285–296, San Jose, CA, USA, 2007.

[3] H. S. Baird, M. A. Moll, J. Nonnemaker, M. R. Casey, and D. L. Delorenzo. Versatile document image content extraction. In *Proc. Document Recognition and Retrieval XIII*, pages 215–221, San Jose, CA, USA, 2006.

[4] E. H. Barney Smith. An analysis of binarization ground truthing. In *Proc. Workshop on Document Analysis Systems*, pages 27–33, Boston, MA, USA, 2010.

[5] M. R. Casey and H. S. Baird. Towards versatile document analysis systems. In *Proc. 7th IAPR Document Analysis Workshop (DAS'06)*, pages 280–290, Nelson, New Zealand, 2006.

[6] B. Gatos, K. Ntirogiannis, and I. Pratikakis. ICDAR 2009 document image binarization contest (DIBCO 2009). In *Proc. International Conference on Document Analysis and Recognition*, pages 1375–1382, Barcelona, Spain, July 2009.

[7] M. Kamel and A. Zhao. Extraction of binary character/graphics images from grayscale document images. *CVGIP: Computer Vision Graphics and Image Processing*, 55(3):203–217, 1993.

[8] B. Lamiroy, D. Lopresti, and T. Sun. Document analysis algorithm contributions in end-to-end applications: Report on the ICDAR 2011 contest. In *Proc. International Conference on Document Analysis and Recognition*, pages 1521–1525, Beijing, China, 2011.

[9] H. J. Lee and B. Chen. Recognition of handwritten chinese characters via short line segments. *Pattern Recognition*, 25(5):543–552, 1992.

[10] H. Lu, A. C. Kot, and Y. Q. Sun. Distance-reciprocal distortion measure for binary document images. *IEEE Signal Processing Letters*, 11(2):228–231, 2004.

[11] K. Ntirogiannis, B. Gatos, and I. Pratikakis. An objective evaluation methodology for document image binarization techniques. In *Proceedings of the 8th International Workshop on Document Analysis Systems (DAS'08)*, pages 217–224, Nara Japan, September 2008.

[12] I. Pratikakis, B. Gatos, and K. Ntirogiannis. H-DIBCO 2010 handwritten document image binarization competition. In *2010 12th International Conference on Frontiers in Handwriting Recognition*, pages 727–732, Kolkata, India, 2010.

[13] I. Pratikakis, B. Gatos, and K. Ntirogiannis. ICDAR 2011 document image binarization contest (DIBCO 2011). In *Proc. International Conference on Document Analysis and Recognition*, pages 1506–1510, Beijing, China, 2011.

[14] M. Sezgin and B. Sankur. Survey over image thresholding techniques and quantitative performance evaluation. *Journal of Electronic Imaging*, 13(1):146–168, 2004.

[15] Øivind Due Trier and T. Taxt. Evaluation of binarization methods for document images. *Transactions on Pattern Analysis and Machine Intelligence*, 17(3), March 1995.