Document Image Database (2009 - 2012): A Systematic Review

Wan Azani Mustafa¹*, Mohamed Mydin M. Abdul Kader¹
¹Faculty of Engineering Technology, Universiti Malaysia Perlis, UniCITI Alam Campus, Sungai Chuchuh, 02100 Padang Besar, Perlis, Malaysia

¹*wanazani@unimap.edu.my, ²mohamedm@unimap.edu.my

Abstract. Document image binarization contributes significantly to the success of the document image analysis and recognition challenging tasks. Image quality can play an important role in addressing the issue of binarization effectiveness. In this paper, a comprehensive review of document database was presented. Review based on image from Document Image Binarization Contest (DIBCO) 2009 to 2012 consists handwritten and printed image. The best algorithm for each year is discussed and analysed. Implications of the review give the direction for future binarization approach developments.

1. Introduction

There are many challenges addressed in handwritten document image binarization, such as faint characters, bleed-through and large background ink stains [1]–[3]. Document image binarization is the process that segments the document image into the text and background by removing any existing degradations [1]. Many document image binarization methods have been proposed in the literature. However, selecting the most optimum method for binarization is a difficult task due to the presence of a variety of degradation in document images.

In 2009, Reza and Mohamed published a paper in which they described the new model of a low quality document image using virtual diffusion processes [4]. This technique focuses on the shadow- through and bleed-through problem. Several studies investigating document binarization based on Otsu modification [5]–[7]. Starting 2010, Nina et al. mention the significant combination between the recursive extension of Otsu thresholding and selective bilateral filtering of scanned handwritten images [5]. This approach considers background estimation before applied the post-processing stage [5], [7]. The above findings contradict the study by Zhang and Wu. They examined the modification algorithm based on the Adaptive Otsu method [6]. The proposed based on three main steps; (1) applied the Wilner filter in order to eliminate noise, (2) improved adaptive Otsu’s method, and (3) dilation and erosion operators were performed to preserve stroke connectivity and fill possible breaks, gaps, and holes. The advantage of this approach is that faster processing time compared to Recursive Otsu Thresholding Method [5] and AdOtsu method [7]. Similarly, Reza and Mohamed also proposed a new novel method based on the Otsu modification known as AdOtsu method [7]. The main idea of this technique was considered parameterless behavior such as average stroke width and the average line height. The positive result was achieved compared to Sauvola method, Otsu Method, and Lu and Tan method.
In this paper, a comprehensive review of document image database from 2009 to 2016 was presented. The database consists two types of image which is handwritten and printed image. The specification in terms of size and resolution also was discussed. Furthermore, the best binarization method for each year from the first rank to the third rank was explained. The rest of this paper is organized as follows: Section 2 describes an overview of the document image. Sub-section 2.1 to 2.4 described the database image. Finally, Section 3 explained the conclusion of this work.

2. Literature Of Database

Document Image Binarization Contest (DIBCO) is established and popular database image. Many researchers used this database in their research and published in the local and international journal [8][9][10]. This research concentrated on DIBCO 2009, DIBCO 2010, DIBCO 2011, and DIBCO 2012 based on a number of images, size, and resolution. The three best methods for each year also was described.

2.1 DIBCO 2009

In International Document Image Binarization Contest (DIBCO 2009), thirty five (35) research groups have participated in the competition with forty three (43) different algorithms. The database of DIBCO 2009 contains two types which are; five (5) handwritten images and five (5) printed image. This database image can be reached at http://www.iit.demokritos.gr/~bgat/DIBCO2009/benchmark. Figure 1 shows the example of an image in the DIBCO 2009 database. The database image present different of resolution (Pixels Per Inch) and size of the pixel as shown in Table 1. The handwritten image labeled ‘H0’ and printed image ‘P0’.

| Handwritten Images | Printed Images |
|--------------------|---------------|
| ![Handwritten Image](image1.png) | ![Printed Image](image2.png) |

**Table 1. Specification of image database**

| Images | Size (pixels) | Resolution (ppi) |
|--------|---------------|------------------|
| H01    | 2025 x 426    | 300              |
| H02    | 946 x 1366    | 200              |
| H03    | 582 x 492     | 72               |
| H04    | 1091 x 581    | 72               |
| H05    | 1341 x 713    | 72               |
| P01    | 1268 x 263    | 96               |
| P02    | 1223 x 310    | 96               |
| P03    | 1153 x 493    | 96               |
| P04    | 1849 x 357    | 96               |
| P05    | 1218 x 259    | 300              |

Based on the contest at 10th International Conference on Document Analysis and Recognition [11], the first ranks winner comes from Institute for Infocomm Research, Singapore. They proposed a new binarization techniques based on four steps; (1) document background extraction, (2) stroke edge detection, (3) local thresholding, and (4) post-processing [12]. Fabrizio and Marcotegui from Université Pierre, France achieved second place with finding a new algorithm using toggle mapping operator [9]. They focused on the morphological operator such as erosion and dilation in order to divide the input
image into three classes; foreground, background and homogeneous. The third rank is method proposed by Henault et al. based on local probabilistic models and the calculus of variation [13], [14].

2.2 DIBCO 2010
The document database 2010 consist ten (10) handwritten images with different size and specification. All the sample images are non-uniform and degraded caused by large background ink stains, shadow and other. The database image, benchmark and evaluation tool can be download at http://www.iit.demokritos.gr/~bgat/H-DIBCO2010/benchmark. Example document images as illustrated in Figure 2. Based on observation Figure 2, image show very low-quality with many noise and non-uniform background. Non-uniform illumination and noise can come from various sources such as aging filaments, sunlight, camera sensors or a noisy environment [15]. Many researchers have agreed that very complicated and difficult to segment the document image in low-grade condition [16]–[19]. The specification in terms of size and resolution are presented in Table 2. All the images show the different size with the large size came from ‘H02’ image (1570 x 841).

Table 2. Specification of image database

| Images | Size (pixels) | Resolution (ppi) |
|--------|--------------|------------------|
| H01    | 1489 x 380   | 300              |
| H02    | 1570 x 841   | 72               |
| H03    | 786 x 423    | 96               |
| H04    | 935 x 537    | 96               |
| H05    | 1726 x 391   | 72               |
| H06    | 945 x 366    | 300              |
| H07    | 1742 x 467   | 72               |
| H08    | 2280 x 326   | 72               |
| H09    | 1158 x 637   | 300              |
| H10    | 1768 x 624   | 72               |

The binarization contest using document database 2010 was done and published at 12th International Conference on Frontiers in Handwriting Recognition. The objective is to find the advanced and effective algorithm in term of measurement criteria such as F-Measure, Negative Rate Metric (NRM) and Misclassification penalty metric (MPM). According to Pratikakis et al. [20], Sixteen (16) distinct research groups have participated in the competition with seventeen (17) different algorithms. The winner is research by Yosef et al. [21] from Ben-Gurion University, Computer Science department, Israel. The proposed method based on adaptive technique and a few pre-processing steps such as removed unwanted objects and accurate local region-based active contour. Interestingly, this rank was shared with the method proposed by B. Su et al. [12]. Second rank achieved by T. Leleore and F. Bouchara from South University of Toulon-Var, France. This method quite complicated to estimate the edge and text location before performing the clustering algorithm [22]. Howe from Smith College, MA, USA at the third rank [23].
They suggest the combination between Laplacian operator and Canny edge detection to detect the text region. This approach is simple however very effective.

2.3 DIBCO 2011
Database 2011 consists eight (8) handwritten images and eight (8) printed images with different pixel size, however same resolution (72ppi). The database can be accessed at http://utopia.duth.gr/~ipratika/DIBCO2011/benchmark. A few sample images are shown in Figure 3. Based on observation, the major problem came from the non-uniform background. Table 3 shows the specification of the database image, where the handwritten image labeled ‘HW’ and printed image ‘PR’.

| Handwritten Images | Printed Images |
|--------------------|---------------|
| [Image]            | [Image]       |

**Table 3. Specification of image database**

| Images | Size (pixels) | Resolution (ppi) |
|--------|---------------|------------------|
| HW1    | 645 x 743     | 72               |
| HW2    | 1218 x 781    | 72               |
| HW3    | 1870 x 511    | 72               |
| HW4    | 469 x 597     | 72               |
| HW5    | 1623 x 261    | 72               |
| HW6    | 787 x 687     | 72               |
| HW7    | 982 x 657     | 72               |
| HW8    | 998 x 410     | 72               |
| PR1    | 1381 x 368    | 72               |
| PR2    | 1180 x 371    | 72               |
| PR3    | 1203 x 363    | 72               |
| PR4    | 1838 x 798    | 72               |
| PR5    | 690 x 682     | 72               |
| PR6    | 1315 x 1069   | 72               |
| PR7    | 600 x 564     | 72               |
| PR8    | 859 x 323     | 72               |

**Figure 3.** Image from DIBCO Database 2011
In 2010, Pratikakis et al. published a paper in which they described the advanced and winner in document binarization using document database 2011 [24]. Sixteen (17) research groups have participated in the competition with eighteen (18) different algorithms. Based on four types of measurement calculation, the best algorithm is proposed by T. Lelore and F. Bouchara from South University of Toulon-Var, France [22]. In 2010 competition, this approaches achieved the second place, however after applied slightly improvement on method became more successful. The improvement is applied a median filtering of the input image and then upscale it using linear interpolation. Furthermore, binarization method by B.Su et al. [12] from National University of Singapore drop to the second rank. The details method was explained in section 2.1. The third rank again achieved by Howe from Smith College, MA, USA [23]. The key element of this approaches is the minimized a global energy function for efficient exact computation. The details method was explained in section 2.2.

2.4 DIBCO 2012
In contrast, the document database 2012 provided 14 handwritten images with the same resolution (72ppi). However, every image shows the different pixel size and the database can be download at http://utopia.duth.gr/~ipratika/HDIBCO2012/benchmark. Table 4 described the specification of database image in terms of size and resolution. The largest size is ‘H01’ (2245 x 1317) and the smallest is ‘H04’ (961 x 854). Figure 4 illustrated a few sample images from database 2012. Based on Figure 6, the major problem images is caused by the fade ink and degrade of background.

| Images | Size (pixels) | Resolution (ppi) |
|--------|---------------|------------------|
| H01    | 2245 x 1317   | 72               |
| H02    | 1285 x 843    | 72               |
| H03    | 1709 x 1371   | 72               |
| H04    | 961 x 854     | 72               |
| H05    | 1661 x 633    | 72               |
| H06    | 1499 x 939    | 72               |
| H07    | 1221 x 297    | 72               |
| H08    | 1645 x 453    | 72               |
| H09    | 1389 x 527    | 72               |
| H10    | 1735 x 1021   | 72               |
| H11    | 1623 x 945    | 72               |
| H12    | 1841 x 433    | 72               |
| H13    | 1703 x 847    | 72               |
| H14    | 2800 x 768    | 72               |

Table 4. Specification of image database

Figure 4. Image from DIBCO Database 2012
Pratikakis et al. [25] indicated that nineteen (19) distinct research groups have participated in the competition with twenty four (24) different algorithms because certain participating research groups have submitted more than one algorithm. After the evaluation process, the first rank belonging to Howe from Smith College, Department of Computer Science, Northampton (MA), USA. Last year, this method placed at third ranking. The key is to optimize a global energy function based on the Laplacian image and dynamically setting the regularization coefficient [23]. Thibault Lelore and Frédéric Bouchara from South University of Toulon- Var, La Garde, France achieved the second rank [22]. The input image divided into three group image (Text, Background and Unknown). The details explanation method can be referred in section 2.2 and 2.3. The third rank is method proposed by B. Su et al. [12] from Institute for Infocomm Research, & School of Computing, National University of Singapore. This method described in Section 2.1.

3. Conclusion
Binarization plays a vital role in the document recognition and analysis. The low quality image such as degraded and non-uniform background will effect on the post-processing stage. This study set out to review in detail the available information on document image database. The review on DIBCO 2009 to DIBCO 2012 in terms pixel size and resolution. The best binarization method for each database image also was discussed. The results of this review indicate that binarization technique using background estimation and stroke edges is best and efficient for this database. Further research might explore the comparison effect between handwritten and printed image on binarization field.

4. Acknowledgment
This work was supported by Short Term Grant Scheme (Grant No 9001-00576)

5. References
[1] K. Ntirogiannis, B. Gatos, and I. Pratikakis, “A combined approach for the binarization of handwritten document images,” Pattern Recognit. Lett., vol. 35, pp. 3–15, Jan. 2014.
[2] B. M. Singh, R. Sharma, D. Ghosh, and A. Mittal, “Adaptive binarization of severely degraded and non-uniformly illuminated documents,” Int. J. Doc. Anal. Recognit., pp. 393–412, 2014.
[3] D. Rivest-Hénault, R. Farrahi Moghaddam, and M. Cheriet, “A local linear level set method for the binarization of degraded historical document images,” Int. J. Doc. Anal. Recognit., vol. 15, no. 2, pp. 101–124, 2012.
[4] R. F. Moghaddam and M. Cheriet, “Low quality document image modeling and enhancement,” Int. J. Doc. Anal. Recognit., vol. 11, no. 4, pp. 183–201, Feb. 2009.
[5] O. Nina, B. Morse, and W. Barrett, “A Recursive Otsu Thresholding Method for Scanned Document Binarization,” in IEEE Workshop on Applications of Computer Vision (WACV), 2010, pp. 307–314.
[6] Y. Zhang and L. Wu, “Fast Document Image Binarization Based on an Improved Adaptive Otsu’s Method and Destination Word Accumulation,” J. Comput. Inf. Syst., vol. 6, no. 7, pp. 1886–1892, 2011.
[7] Reza Farrahi Moghaddam Mohamed Cheriet, “AdOtsu: An adaptive and parameterless generalization of Otsu’s method for document image binarization,” Pattern Recognit., vol. 45, pp. 2419–2431, Jun. 2012.
[8] A. Shaus, B. Sober, E. Turkel, and E. Piasetzky, “Beyond the ground truth: Alternative quality measures of document binarizations,” in Proceedings of International Conference on Frontiers in Handwriting Recognition, ICFHR, 2017, pp. 495–500.
[9] I. Ben Messaoud, H. Amiri, H. El Abed, and V. Märgner, “New binarization approach based on text block extraction,” in *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, 2011, pp. 1205–1209.

[10] T. H. N. Le, T. D. Bui, and C. Y. Suen, “Ternary entropy-based binarization of degraded document images using morphological operators,” in *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, 2011, pp. 114–118.

[11] B. Gatos, K. Ntirogiannis, and I. Pratikakis, “ICDAR 2009 Document Image Binarization Contest (DIBCO 2009),” *2009 10th Int. Conf. Doc. Anal. Recognit.*, no. Dibco, pp. 1375–1382, 2009.

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

[13] R. Farrahi Moghaddam and M. Cheriet, “RSLDI: Restoration of single-sided low-quality document images,” *Pattern Recognit.*, vol. 42, no. 12, pp. 3355–3364, 2009.

[14] D. Rivest-Hénault and M. Cheriet, “Image Segmentation Using Level Set and Local Linear Approximations,” in *Image Analysis and Recognition: 4th International Conference, ICIAR 2007, Montreal, Canada, August 22-24, 2007. Proceedings*, M. Kamel and A. Campilho, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 234–245.

[15] W. A. Mustafa and H. Yazid, “Illumination and Contrast Correction Strategy using Bilateral Filtering and Binarization Comparison,” *J. Telecommun. Electron. Comput. Eng.*, vol. 8, no. 1, pp. 67–73, 2016.

[16] Y. Yang and H. Yan, “An adaptive logical method for binarization of degraded document images,” *Pattern Recognition*, vol. 33. pp. 787–807, 2000.

[17] J. G. Kuk and N. I. Cho, “Feature Based Binarization of Document Images Degraded by Uneven Light Condition,” *2009 10th Int. Conf. Doc. Anal. Recognit.*, pp. 748–752, 2009.

[18] W. A. Mustafa, “A proposed optimum threshold level for document image binarization,” *J. Adv. Res. Comput. Appl.*, vol. 1, no. 1, pp. 8–14, 2017.

[19] B. Su, S. Lu, and C. L. Tan, “Robust document image binarization technique for degraded document images,” *IEEE Trans. Image Process.*, vol. 22, pp. 1408–1417, 2013.

[20] I. Pratikakis, B. Gatos, and K. Ntirogiannis, “H-DIBCO 2010 - Handwritten Document Image Binarization Competition,” *12th Int. Conf. Front. Handwrit. Recognit.*, pp. 727–732, Nov. 2010.

[21] I. Bar-Yosef, N. Hagbi, K. Kedem, and I. Dinstein, “Fast and Accurate Skew Estimation Based on Distance Transform,” *2008 Eighth IAPR Int. Work. Doc. Anal. Syst.*, pp. 402–407, 2008.

[22] T. Lelore and F. Bouchara, “Document image binarisation using Markov field model,” in *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, 2009, pp. 551–555.

[23] N. R. Howe, “A Laplacian energy for document binarization,” in *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, 2011, pp. 6–10.

[24] I. Pratikakis, B. Gatos, and K. Ntirogiannis, “ICDAR 2011 Document Image Binarization Contest (DIBCO 2011),” *Int. Conf. Doc. Anal. Recognit.*, pp. 1506–1510, Sep. 2011.

[25] K. Ntirogiannis, B. Gatos, and I. Pratikakis, “ICFHR2012 Competition on Handwritten Document Image Binarization (H-DIBCO 2012),” *Int. Conf. Front. Handwrit. Recognit.*, pp. 809–813, 2012.