A Comprehensive Review on Document Image (DIBCO) Database

W A Mustafa¹, Wan Khairunizam², I Zunaidi³, Z M Razlan² and Shahriman A B²

¹Faculty of Engineering Technology, Kampus Sg. Chuchuh, Universiti Malaysia Perlis, 02100 Padang Besar, Perlis, Malaysia.
²School of Mechatronic Engineering, Universiti Malaysia Perlis, Arau, Perlis, MALAYSIA
³Faculty of Technology, University of Sunderland, St Peter's Campus, Sunderland, SR6 0DD, United Kingdom

Abstract. One of the most significant current discussions in image processing is a document analysis. Now, many types of document database were established in order to address the issue of binarization effectiveness. In this paper, a comprehensive review of the document database was presented. Review based on an image from Document Image Binarization Contest (DIBCO) from 2013 to 2017 which consists of handwritten and printed image. The best algorithm for each year is discussed and analyzed. Based on the results, the technique using background estimation and stroke edges is better performance for the overall database. Besides, the method using the combination of Laplacian operator and canny edge detection also shows the successful result, especially in the printed image. Implications of the review give the direction for future binarization approach developments.

1. Introduction

In the recent decades, the binarization technique has been one of the major interesting research subjects due to detect the text images. However, there has been little discussion and challenges about the degradation of input document images caused by large background ink stains, contrast variation and illumination problem [1–3]. Binarization is one of the important process in order to separate between the object (text) and background by removing all unwanted noised [4,5]. A large and growing body of literature has investigated a new method in order to segment the background and foreground document images [6,7]. However, selecting the most optimum method for binarization is a difficult task due to the presence of a variety of degradation in document images.

There are many algorithms that can be used for binarization process but the earliest and most enduring image binarization methods is Otsu’s method [12,13]. This method works by finding a threshold value that differentiates pixels into two components which is lighter and darker pixels. The lighter pixel will be classified as white meanwhile for the darker pixel it is black. The method has its own weakness, it cannot binarize image with almost the same pixels color. Besides, Ntirogiannis et al. [14] have done a combined approach for the binarization of handwritten document images. The research was inspired by the low detection rate of faint characters in binarization of handwritten document images. Firstly, image normalization based on background compensation and estimation is applied. For the second step, global thresholding is applied to the image. On the binarized images, very small components are discarded and representative characteristics of a document image such as the stroke width and the contrast are computed. Furthermore, local adaptive binarization is performed on the normalized image taking into account the aforementioned characteristics. Finally, the two binarization outputs are combined at connecting component level. For DIBCO 2011 dataset, the average F-measure and PSNR are 94.05% and 21.65 respectively.

In this paper, a comprehensive review of the document image database from 2012 to 2017 was presented. The database consists of 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 2.5 described the database image. Finally, Section 3 explained the conclusion of this work.

2. Overview Database

Recently, researchers have shown an increased interest publication in document binarization techniques [15–17]. Every year, a new document database was provided by a specialist in order to increase the challenging for researchers. Based on the review and publication number, many research used a document database from http://www.iit.demokritos.gr/~bgat and http://utopia.duth.gr/~ipratika/ website [18,19]. The contest of binarization using a popular document database was organized called as International Document Image Binarization Contest (DIBCO). In this paper, a comprehensive review of the document database from DIBCO 2013 to DIBCO 2017 was analyzed.

2.1 DIBCO 2013

Database 2013 presents are 8 handwritten images and 8 printed images (shown in Figure 1) and can be download at http://utopia.duth.gr/~ipratika/DIBCO2013/benchmark. Overall, the database images show the different size and same resolution as show in Table 1. However, two (2) image from the handwritten group consists of different resolution which is 96 ppi (HW06 and HW08). Based on image size, database 2013 is largest compared to the previous database (2009-2012) [20]. Example, ‘HW01’ image size is 4161 x 1094 pixel.

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

Figure 1: Sample image from DIBCO Database 2013

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

| Images | Size (pixels) | Resolution (ppi) |
|--------|---------------|------------------|
| HW01   | 4161 x 1094   | 72               |
| HW02   | 1136 x 559    | 72               |
| HW03   | 2290 x 504    | 72               |
| HW04   | 3297 x 1097   | 72               |
| HW05   | 1493 x 1613   | 72               |
| HW06   | 3088 x 1152   | 96               |
| HW07   | 1903 x 807    | 72               |
| HW08   | 2292 x 1082   | 96               |
| PR01   | 1557 x 965    | 72               |
| PR02   | 1485 x 847    | 72               |
| PR03   | 1192 x 956    | 72               |
| PR04   | 2287 x 870    | 72               |
| PR05   | 2251 x 429    | 72               |
| PR06   | 2275 x 737    | 72               |
| PR07   | 871 x 369     | 72               |
| PR08   | 1661 x 479    | 72               |

Table 1: Specification of the image database
In 2013, Pratikakis et al. [21] organized Document Image Binarization Contest at 12th International Conference on Document Analysis and Recognition. The objective is to select the best algorithm based on 4 measurements calculation such as F-measure and PSNR. Eighteen (18) research groups have participated in the competition with twenty-three (23) distinct algorithms. Again, B.Su et al. National University of Singapore and the Institute for Infocomm Research, Singapore achieved the first rank Therefore, this method is an improvement from the method was proposed in 2009. They used the local maximum and minimum and local image gradient to obtained the local contrast [22,23]. Second rank belonging by Howe from Smith College, MA, USA. The main idea is using the combination of Laplacian operator and Canny edge detection. The technique is described in previous section [24,25]. Moghaddam et al. University of Quebec, Canada and Smith College, MA, USA proposed combination techniques between Laplacian energy binarization method introduced in [25] and the Ensemble-of-Expert (EoE) framework [26].

2.2 DIBCO 2014

Refer on the http://users.iit.demokritos.gr/~bgat/HDIBCO2014/benchmark, the document database 2014 consists ten (10) handwritten image without any printed images. All images show the different pixel size but same resolution (96ppi). Figure 2 illustrated a few sample images from database 2014. Based on observation, the major problem of this database caused by ink fading and background artifact. The details specification was described in Table 2. The largest image is ‘H03’ (2675 x 1255) and the smallest image is ‘H06’ (775 x 460). All the image labeled from ‘H01’ to ‘H10’.

| Handwritten Images | Table 2. Specification of the image database |
|--------------------|--------------------------------------------|
| ![Sample Images](image1.png) | ![Specs Table](image2.png) |

Table 2. Specification of the image database

| Images | Size (pixels) | Resolution (ppi) |
|--------|---------------|------------------|
| H01    | 1761 x 707    | 96               |
| H02    | 1881 x 455    | 96               |
| H03    | 2675 x 1255   | 96               |
| H04    | 1105 x 339    | 96               |
| H05    | 1317 x 288    | 96               |
| H06    | 775 x 460     | 96               |
| H07    | 1449 x 463    | 96               |
| H08    | 1925 x 688    | 96               |
| H09    | 1474 x 482    | 96               |
| H10    | 1498 x 407    | 96               |

Figure 2. Image from DIBCO Database 2014

According to the Competition on Handwritten Document Image Binarization (H-DIBCO 2014), seven (7) distinct research groups have participated in the competition with eight (8) different algorithms since one participating research group has submitted two algorithms. Based on four evaluations such as F-measure, pseudo F-measure, Peak Signal Noise Ratio (PSNR) and Distance Reciprocal Distortion Metric (DRD), the winner is the method by Mesquita et al. [16] from Centro de Informática, Universidade Federal de Pernambuco, Brazil. This method based on human visual system perceives
and improved the Howe's binarization method [25], where Howe's method may fail in some situations in which its parametric choice is optimal for only a part of the image, as in cases of degradation generated by smudges or uneven illumination. Howe from Smith College, Department of Computer Science, Northampton (MA), USA achieved the second rank [25]. The main part is the canny threshold to separate the foreground and the background. The details technique can be referred to section 2.4. The third rank is Nafchi et al. [27] from Synchromedia Lab, Icole de technologies supérieure, University of Quebec, Montreal, Quebec, Canada. This method involved a few steps which are; (1) Denoised image with phase preserved, (2) Maximum moment of phase congruency covariance and (3) Locally weighted mean phase angle.

2.3 DIBCO 2015

Based on two established link for document database is http://vc.ee.duth.gr/ and http://users.iit.demokritos.gr/~bgat/ referred that no any document database provided in 2015. However, a few papers were published 2015 used an Indian document dataset [28,29]. In International Conference on Document Analysis and Recognition (ICDAR 2015) only used the datasets are generated from a set of historical manuscripts written between the 17th and 20th centuries, collected by Archives of Quebec (Canada) [30].

2.4 DIBCO 2016

Latest, the database 2016 can be accessed at http://vc.ee.duth.gr/h-dibco2016/benchmark. Database 2016 provided 10 handwritten images with different size and resolution. The image ‘7’ show the lowest resolution which is 96ppi and the rest image is same resolution (300ppi). The problem of the images is the low quality background and fade ink as show in Figure 3. The specification details of the database are present in Table 3. Many researchers agreed that difficulties and challenges task in order to detect the text region in a non-uniform background condition [31,32].

| Handwritten Images |
|--------------------|
| ![Image 1](image1.jpg) |
| ![Image 2](image2.jpg) |
| ![Image 3](image3.jpg) |

Figure 3. Image from DIBCO Database 2016

| Images | Size (pixels) | Resolution (ppi) |
|--------|---------------|------------------|
| 1      | 1510 x 1067   | 300              |
| 2      | 2259 x 1023   | 300              |
| 3      | 2417 x 1064   | 300              |
| 4      | 2363 x 615    | 300              |
| 5      | 2628 x 867    | 300              |
| 6      | 1364 x 788    | 300              |
| 7      | 963 x 656     | 96               |
| 8      | 1782 x 334    | 300              |
| 9      | 1339 x 302    | 300              |
| 10     | 378 x 315     | 300              |

Table 3. Specification of the image database
In Handwritten Document Image Binarization Contest (H-DIBCO 2016), Pratikakis et al. [33] listed the ranking from nine (9) research groups with twelve (12) distinct algorithms. The winner is Nati Kligler and Ayellet Tal from Israel Institute of Technology, Israel. This method involved three stages; (1) pre-processing that used Creating the Visibility Score Map Hidden Point Removal (HPR) operator [34]. (2) Binarization – for the database 2016, they suggest to applied the binarization method presented by [25] and (3) Post-processing- use standard denoising. In this competition, Hassaine et al. from the University of Bordeaux, France and Qatar University submit are three methods where the ‘Method 3’ get the second place and ‘Method 2’ get the third place [35]. Method 2 based on Otsu binarization method [36], while Method 3 is the combination of Method 2 and the restoration of optical soundtracks [37].

2.5 DIBCO 2017

International Competition on Document Image Binarization (ICDAR 2017), eighteen (18) research groups have participated in the competition with twenty-six (26) different algorithms. The database of DIBCO 2017 contains two types which are; ten (10) handwritten images and ten (10) printed image. This database image can be reached at http://vc.ee.duth.gr/dibco2017/benchmark/. Figure 4 shows the example of an image in the DIBCO 2017 database. The database image present different of resolution (Pixels Per Inch) and the size of the pixel as shown in Table 4. The handwritten image labeled ‘H’ and the printed image ‘P’.

| Handwritten Images | Printed Images |
|--------------------|---------------|
| ![Handwritten Image](image1.png) | ![Printed Image](image2.png) |
| ![Handwritten Image](image3.png) | ![Printed Image](image4.png) |

Figure 4. Image from DIBCO Database 2016

| Images | Size (pixels) | Resolution (ppi) |
|--------|---------------|-----------------|
| H1     | 2397 x 1350   | 300             |
| H2     | 2095 x 1781   | 96              |
| H3     | 2379 x 1308   | 96              |
| H4     | 2439 x 1229   | 300             |
| H5     | 2291 x 874    | 300             |
| H6     | 351 x 292     | 300             |
| H7     | 593 x 376     | 300             |
| H8     | 1303 x 594    | 300             |
| H9     | 1744 x 809    | 300             |
| H10    | 1721 x 924    | 300             |
| P1     | 2092 x 951    | 72              |
| P2     | 1083 x 877    | 300             |
| P3     | 1139 x 433    | 72              |
| P4     | 1225 x 739    | 300             |
| P5     | 1179 x 516    | 72              |
| P6     | 1553 x 2073   | 72              |
| P7     | 1233 x 2206   | 72              |
| P8     | 1050 x 608    | 300             |
| P9     | 1320 x 768    | 96              |
| P10    | 1075 x 775    | 300             |

Table 4. Specification of the image database
Based on the contest at 14th IAPR International Conference on Document Analysis and Recognition [38], the first ranks winner comes from the Institute for Infocomm Research, Singapore. They proposed binarization techniques based on four steps; (1) document background extraction, (2) stroke edge detection, (3) local thresholding, and (4) post-processing [22]. 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 a method proposed by Henault et al. based on local probabilistic models and the calculus of variation [39,40].

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 sets out to review in detail the available information on the document image database from the established website. The review focuses on DIBCO 2013 to DIBCO 2017 based on quantity, pixel size and resolution. The best binarization method for each database image also was discussed. The results of this review indicate that the binarization technique using background estimation and stroke edges is best and efficient for the overall database. The method using the combination of Laplacian operator and canny edge detection also successful result especially for the printed image. Further research might explore the comparison effect between handwritten and printed image on binarization field.

4. References

[1] Singh B M, Sharma R, Ghosh D and Mittal A 2014 Adaptive binarization of severely degraded and non-uniformly illuminated documents Int. J. Doc. Anal. Recognit. 393–412
[2] Mustafa W A and Yazid H 2016 Illumination and Contrast Correction Strategy using Bilateral Filtering and Binarization Comparison J. Telecommun. Electron. Comput. Eng. 8 67–73
[3] Mustafa W A and Yazid H 2017 Image Enhancement Technique on Contrast Variation : A Comprehensive Review J. Telecommun. Electron. Comput. Eng. 9 199–204
[4] Rivest-Hénault D, Farrahi Moghaddam R and Cheriet M 2012 A local linear level set method for the binarization of degraded historical document images Int. J. Doc. Anal. Recognit. 15 101–24
[5] Mustafa W A and Kader M M M A 2018 Binarization of Document Images: A Comprehensive Review J. Phys. Conf. Ser. 1019 1–9
[6] Mustafa W A, Khairunizam W, Ibrahim Z, Ab S and M.Razlan Z 2018 Improved Feng Binarization Based on Max-Mean Technique on Document Image IEEE International Conference on Computational Approach in Smart Systems Design and Applications (ICASSDA) (IEEE) pp 1–6
[7] Mustafa W A, Aziz H, Khairunizam W, Ibrahim Z, Ab S and M.Razlan Z 2018 Review of Different Binarization Approaches on Degraded Document Images IEEE International Conference on Computational Approach in Smart Systems Design and Applications (ICASSDA) (IEEE) pp 1–8
[8] Moghaddam R F and Cheriet M 2009 Low quality document image modeling and enhancement Int. J. Doc. Anal. Recognit. 11 183–201
[9] Nina O, Morse B and Barrett W 2010 A Recursive Otsu Thresholding Method for Scanned Document Binarization IEEE Workshop on Applications of Computer Vision (WACV) pp 307–14
[10] Zhang Y and Wu L 2011 Fast Document Image Binarization Based on an Improved Adaptive Otsu’s Method and Destination Word Accumulation J. Comput. Inf. Syst. 6 1886–92
[11] Reza Farrahi Moghaddam and Cheriet M 2012 AdOtsu: An adaptive and parameterless generalization of Otsu’s method for document image binarization Pattern Recognit. 45 2419–
[12] Ntirogiannis K, Gatos B and Pratikakis I 2013 Performance evaluation methodology for historical document image binarization *IEEE Trans. Image Process.* **22** 595–609

[13] Lokhande S S and Dawande N A 2015 A survey on document image binarization techniques *Proceedings - 1st International Conference on Computing, Communication, Control and Automation, ICCUBEA 2015* pp 742–6

[14] Ntirogiannis K, Gatos B and Pratikakis I 2014 A combined approach for the binarization of handwritten document images *Pattern Recognit. Lett.* **35** 3–15

[15] Wagdy M, Faye I and Rohaya D 2013 Fast and efficient document image clean up and binarization based on retinex theory *Proceedings - 2013 IEEE 9th International Colloquium on Signal Processing and its Applications, CSPA 2013* pp 58–62

[16] Mesquita R G, Mello C A B and Almeida L H E V 2014 A new thresholding algorithm for document images based on the perception of objects by distance *Integr. Comput. Aided. Eng.* **21** 133–46

[17] Soua M, Kachouri R and Akil M 2015 Improved Hybrid Binarization based on Kmeans for Heterogeneous document processing *9th International Symposium on Image and Signal Processing and Analysis, ISPA 2015* pp 210–5

[18] Mustafa W A and Kader M M M A 2018 Binarization of Document Image Using Optimum Threshold Modification *J. Phys. Conf. Ser.* **1019** 1–8

[19] Mustafa W A, Yazid H and Jaafar M 2018 An Improved Sauvola Approach on Document Images Binarization *J. Telecommun. Electron. Comput. Eng.* **10** 43–50

[20] Mustafa W A and Kader M M M A 2018 Document Image Database (2009 - 2012): A Systematic Review *J. Phys. Conf. Ser.* **1019** 1–7

[21] Pratikakis I, Gatos B and Ntirogiannis K 2013 ICDAR 2013 document image binarization contest (DIBCO 2013) *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR* pp 1471–6

[22] Lu S, Su B and Tan C L 2010 Document image binarization using background estimation and stroke edges *Int. J. Doc. Anal. Recognit.* **13** 303–14

[23] Su B, Lu S and Tan C L 2013 Robust document image binarization technique for degraded document images *IEEE Trans. Image Process.* **22** 1408–17

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

[25] Howe N R 2013 Document binarization with automatic parameter tuning *Int. J. Doc. Anal. Recognit.* **16** 247–58

[26] Reza Farrahi Moghaddam, Fereydoun Farrahi Moghaddam M C 2013 Unsupervised ensemble of experts (EoE) framework for automatic binarization of document images *Comput. Vis. Pattern Recognit.* 703–7

[27] Ziaei Nafchi H, Farrahi Moghaddam R and Cheriet M 2013 Historical document binarization based on phase information of images *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* vol 7729 LNCS pp 1–12

[28] Kavitha S, Shivakumara P, Hemantha Kumar G and Tan C L 2015 A robust script identification system for historical Indian document images *Malaysian J. Comput. Sci.* **28** 283–300

[29] Ahmad E, Azimifar Z, Shams M, Famouri M and Shafiee M J 2015 Document image binarization using a discriminative structural classifier *Pattern Recognit. Lett.* **63** 36–42

[30] Hedjam R, Nafchi H Z, Moghaddam R F, Kalacska M and Cheriet M 2015 Multispectral Text Extraction Contest *13th Int. Confrence Doc. Anal. Recognit. - ICDAR’15* 1181–5

[31] Gatos B, Pratikakis I and Perantonis S J 2004 An Adaptive Binarization Technique for Low Quality Historical Documents *Doc. Anal. Syst.* **VI** 102–13

[32] Pai Y-T, Chang Y-F and Ruan S-J 2010 Adaptive thresholding algorithm: Efficient computation technique based on intelligent block detection for degraded document images
Pattern Recognit. 43 3177–87

[33] Pratikakis I, Zagoris K, Barlas G and Gatos B 2017 ICFHR 2016 handwritten document image binarization contest (H-DIBCO 2016) Proc. Int. Conf. Front. Handwrit. Recognition, ICFHR 619–23

[34] Katz S, Tal A and Basri R 2007 Direct visibility of point sets ACM Trans. Graph. 26 24

[35] Maadeed S Al, Ayoubi W, Hassaïne A and Aljaam J M 2012 QUWI: An Arabic and English handwriting dataset for offline writer identification Proceedings - International Workshop on Frontiers in Handwriting Recognition, IWFHR pp 746–51

[36] Hassaïne A, Al-Maadeed S and Bouridane A 2012 A set of geometrical features for writer identification Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) vol 7667 LNCS pp 584–91

[37] Hassaïne A, Decencière E and Besserer B 2009 Efficient restoration of variable area soundtracks Image Anal. Stereol. 28 113–9

[38] Pratikakis I, Zagoris K, Barlas G and Gatos B 2018 ICDAR2017 Competition on Document Image Binarization (DIBCO 2017) Proc. Int. Conf. Doc. Anal. Recognition, ICDAR 1 1395–403

[39] Farrahi Moghaddam R and Cheriet M 2009 RSLDI: Restoration of single-sided low-quality document images Pattern Recognit. 42 3355–64

[40] Rivest-Hénault D and Cheriet M 2007 Image Segmentation Using Level Set and Local Linear Approximations Image Analysis and Recognition: 4th International Conference,ICIAR 2007, Montreal, Canada, August 22-24, 2007. Proceedings ed M Kamel and A Campilho (Berlin, Heidelberg: Springer Berlin Heidelberg) pp 234–45