Indonesian Journal of Electrical Engineering and Computer Science
Vol. 15, No. 3, September 2019, pp. 1329~1336
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v15.i3.pp1329-1336

Pixel-wise classification using support vector machine for binarization of degraded historical document image

Fauziah Kasmin, Zuraini Othman, Sharifah Sakinah Syed Ahmad
Department of Intelligent Computing and Analytics, Faculty of Information and Communications Technology, Universiti Teknikal Malaysia Melaka, Malaysia

Article Info

Article history:

Received Dec 11, 2018
Revised Mar 20, 2019
Accepted Apr 25, 2019

Keywords:

Binarization
Grey level
Historical document
Local neighbourhood
RGB

ABSTRACT

Binarization of historical documents nowadays is very important as digital archiving has become the best and preferred solution for the retrieval and storage of valuable archives. However, the process becomes more challenging due to the degradation of historical documents. Hence, this paper described a method on binarization of historical documents using the learning concept. Support vector machine (SVM) learning was used as a classifier in this work. After training some images with the help of ground truth images, a model was developed. Testing images then used the model to segregate each pixel as text or non-text. The grey level and RGB values were chosen as descriptors for a particular pixel and comparisons were made between these two descriptors. The intensities of the local neighbourhood for every pixel were used in the experiment. To compare these descriptors, standard dataset HDIBCO2014, DIBCO2012 and DIBCO2016 were used in the training and testing phase. The results from the experiment clearly showed that grey level values gave better performance compared to RGB values.

Corresponding Author:

Fauziah Kasmin,
Department of Intelligent Computing and Analytics,
Faculty of Information and Communications Technology,
Universiti Teknikal Malaysia Melaka,
76100, Durian Tunggal, Melaka, Malaysia.
Email: fauziah@utem.edu.my

1. INTRODUCTION

Historical documents are often valuable and need to be protected and well preserved. Damage to any original historical documentation, or rare texts, can cause degradation and result in deterioration. Libraries, archivists, and those who have stewardship over precious texts, and in preserving and protecting them, often find that circulating these original treasures almost becomes impossible. In recent years, digital archiving has been the optimal and preferred solution for retrieval and storage of these valuable archives. Historical document image analysis requires several steps to be performed, consisting of layout analysis, followed by text line, word segmentation and finally optical character recognition (OCR) [1]. Binary image representation is said to be the most preferred image format, and the process of obtaining a binary image is called binarization [1].

However, many problems are associated with historical documents as they tend to consist of handwritten and machine printed documents. Also, [2] handwritten documents are more difficult to process compared to machine printed documents. One of the main reasons for this is that these documents lack a specific structure and typically the style of writing is exclusive to a particular individual. Furthermore, the characters may be attached or connected depending on the style of the calligraphic writing. Another reason

associated with handwritten documents is that many of the documents are written using pen quills. In this instance, several degradations can occur, for example, large ink stains that bleed through the paper and as well as producing faint characters. As stated in [3], historical documents are more complex to binarize compare to recent documents due to some factors like color, paper aging, stains, transludicity, texture and many more.

The literature in this field of research has revealed that the binarization algorithm consists of two groups [4]. The first group is using the thresholding method. Threshold value is obtained from an algorithm that separates text from non-text. Some of the earlier work carried out by researchers was by Otsu [5], Kittler et al. [6], Li & Lee [7], Niblack [8], Sauvola [9] and many more. The thresholding method includes global and local thresholding, although one of the weaknesses in the global thresholding method is that the approach fails when background of a particular image is various or the patterns of certain images have heterogeneous background. In other words, the global thresholding method cannot be adjusted with varying illumination images and do not work well in low-quality images [10]. While for local thresholding, the main weakness is challenging to estimate the parameters used for the algorithms. Likewise, they work poorly on images with a high degree of variability [11].

The second group category is the binarization method that used classification approach which allows clustering and selecting features in order to produce binarization [10]. However, the drawback of this method is that it needs a training set to improve the binarization performance and the results depend mainly on the quality of the training set [10]. Although, many low-level features can be used for training each pixel. Some of the features use grey level intensities [12], gradients [13], red, green and blue (RGB) [14] and many more.

Therefore, the main objective of this work is to compare two descriptors used in the classification of each pixel in document binarization; RGB colour intensities and grey level intensities. Intensities in a 3 by 3 window are used as features to determine each pixel as text or non-text. To classify the pixels, Support Vector Machine (SVM) learning is used.

The remainder of this the paper is organised into the following sections. Section 2 presents and discusses the literature review. Section 3 describes the proposed method which is followed by Section 4 which discusses the experimental results. Lastly, Section 5 presents the overall findings of this work and conclusions.

2. RELATED WORKS

A lot of studies have been done in the area of historical document image binarization using a supervised approach. One study by [1] used hierarchical deep supervised network (DSN) architecture in the classification of the text's pixels at two groups of feature levels, namely; high level and low-level features. Low-level features were used to obtain foreground maps, where at the boundary area the visual quality was found to be much better. Whereas, the high-level features enabled differentiation between the foreground and background which was shown to manage severe degradations quite well. Degraded document images have been binarized by [4] using a structured classifier; a conditional random field (CRF). Markov random fields generated conditional probability distribution from the binarized image, and the CRF then modelled the probability. Also, the training phase was used to estimate the model parameters. The final results of the binarization output were lastly chosen by looking at the most probable binary image. The binary image chosen is the one that give maximum accuracy in trained model.

Next, the researcher in [15] proposed a learning-based binarization method, claiming that the proposed method could increase the accuracy of the binarization method for documents of the same type to stabilise the quality. At the stage of learning, knowledge of binarization evaluation and optimisation were first obtained. Then, the results of the binarization were used as input towards the binarization process again. This process was performed so that the binarization parameters could be adjusted and the process increased the binarization accuracy.

Learning framework for obtaining optimised parameter values was proposed by [16] to cater for any binarization method. The framework consists of two sections; learn and apply. In this framework, a ground truth dataset is used for learning process with three consecutive steps to extract the features, obtaining the optimal parameter from the training images and using support vector regression to perform the classification. The researcher also used support vector regression (SVR) as structural risk can be minimised due to SVR's intrinsic ability. In another study, Su et al. [17] proposed document image binarization using a self-training approach. In this approach, pixels were grouped into three classes, i.e. foreground, background and uncertain pixels. Then a classifier was used to classify the foreground and background pixels after learning from the document image pixels. Uncertain pixels were then classified by the rules generated from the learned pixels. On the other hand, Xiong et al. [18] used SVM to segment text from degraded document images applying

different global thresholds for a different background where the images were divided into w × w regions. Local contrast enhancement was used in order to pre-process every image block. This was then followed by using SVM to obtain a threshold value for every block. The process was conducted for the entire image using a locally adaptive thresholding method. In another study by [19], they have proposed to binarize musical documents using convolutional neural network. They have used patches which contain region that surround target pixels as descriptors. They have done several modifications by testing number of layers, filters, size of kernels, activations type and number of dropout units for convolutional neural network classifier. The main advantage of using learning framework is its ability to be generalizable.

3. PROPOSED METHOD

Many researchers have conducted binarization using a supervised approach as it has been proven to be quite successful in segmenting several images, for example; documents, retinal blood vessels and more. In this study, the method used a classifier to classify the features used for every pixel in the image. As various features can be used to classify a pixel, in this work, the RGB values and grey level values were compared as a representative for a particular pixel. Next, the neighbourhood using 3 x 3 window was chosen to represent the pixel for each feature. This is illustrated in Figure 1 for the RGB and grey level values.

R_1 B_1 G_1	R_2 B_2 G_2	R_3 B_3 G_3
R_4 B_4 G_4	R_5 B_5 **I** G_5	R_6 B_6 G_6
R_7 B_7 G_7	R_8 B_8 G_8	R_9 B_9 G_9

P_1	P_2	P_3
P_4	P_5 **I**	P_6
P_7	P_8	P_9

Figure 1. Neighbourhood of 3 x 3 window for every target pixel

Where I = target pixel; R = intensity values for red; G = intensity values for green; B = intensity values for blue; and P= intensity values for grey level.

Each pixel was classified as text for label 1 or non-text for label 0. Figure 2 shows the RGB values used as descriptors for a target pixel. Figure 3 shows the value of intensities in the grey level used as descriptors for a target pixel. The value for the label was the value of the target pixel in the respective ground truth image.

Label										
0	1:R_1	2:G_1	3:B_1	4:R_2	5:G_2	6:B_2	7:R_3	8:G_3	9:B_3	10:R_4	11:G_4
	12:B_4	13:R_5	14:G_5	15:B_5	16:R_6	17:G_6	18:B_6	19:R_7	20:G_7	21:B_7	22:R_8
	23:G_8	24:B_8	25:R_9	26:G_9	27:B_9						

Figure 2. RGB representation

Label								
0	1:P_1	2:P_2	3:P_3	4:P_4	5:P_5	6:P_6	7:P_7	8:P_8	9:P_9

Figure 3. Grey level representation

Initially, the classifier was trained by SVM and used ground truth images to provide the correct label for each pixel. Since this is a classification by pixel, the redundant rows needed to be identified as most of the cases that happened. These cases were then removed in order to reduce the complexity and time. Then, 8000 sets of data were selected at random for training. From these 8000 data, 4000 data were labelled as 1, and 4000 were labelled as 0. Then to classify the pixels, LIBSVM was used in the experiment [20]. Radial basis function kernel is used in LIBSVM and hence, its' parameter values C and γ need to be determined. At first, data were trained across 5-folds cross-validation. Once the values C and γ were obtained, these values were then used to create a model for the trained data. To test the data, images used for the training session were excluded. Figure 4 illustrates the steps of the approach employed in this work.

4. EXPERIMENT AND RESULT

The standard databases HDIBCO2014 [21], DIBCO2012 [22] and DIBCO2016 [23] were used to compare these two features which consisted of historical document images with their respective ground truth. The databases are freely available to the public and have been used by many researchers. HDIBCO2014 consisted of 10 handwritten images where three images were used for training while another six images were used for testing in the experiment. DIBCO2012 consisted of 14 images and DIBCO2016 consisted of 10 images. Images in DIBCO2012 and DIBCO2016 used as test images in the experiment and all these images used model constructed from training images in HDIBCO2014. The ground truth of these images was available and used for a training session to develop the correct model by the classifier.

To evaluate the binary images obtained, several metrics were used. These metrics have also been used by many researchers to quantify the binarization procedure [24-25]. The metrics included the F-measure, Peak Signal Noise to Ratio (PSNR) and Negative Rate Metric (NRM). The pixels of the final image were then classified as text or non-text. Then the following expression was applied:

$$F-measure = \frac{2\times recall\times precision}{recall+precision} \text{ where } recall = \frac{TP}{TP+FN}, precision = \frac{TP}{TP+FP} \quad (1)$$

Where TP is true positive, FP is false positive, and FN is false negative. The F-measure shows the accuracy in percentage of the obtained binary image.

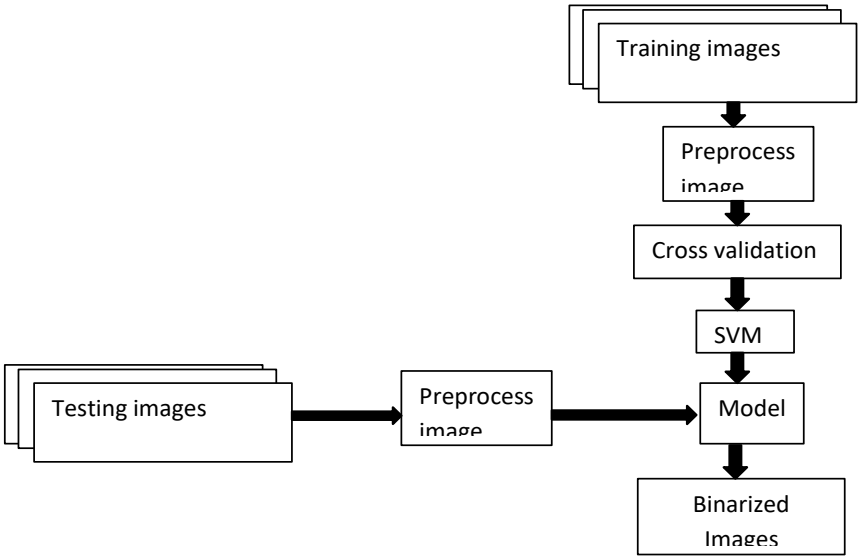


Figure 4. Steps of the proposed method

$$PSNR = 10\log\left(\frac{C^2}{MSE}\right) \text{where } MSE = \frac{\sum_x^M \sum_y^N (I(x,y)-I_2(x,y))^2}{MN} \quad (2)$$

PSNR determines how good a given image is similar to another image, and thus, a higher value of PSNR indicates a higher similarity between the two images. I and I_2 are the two images; ground truth image

and a final binary image were obtained. M and N are the height and width and C is the difference between the foreground and background.

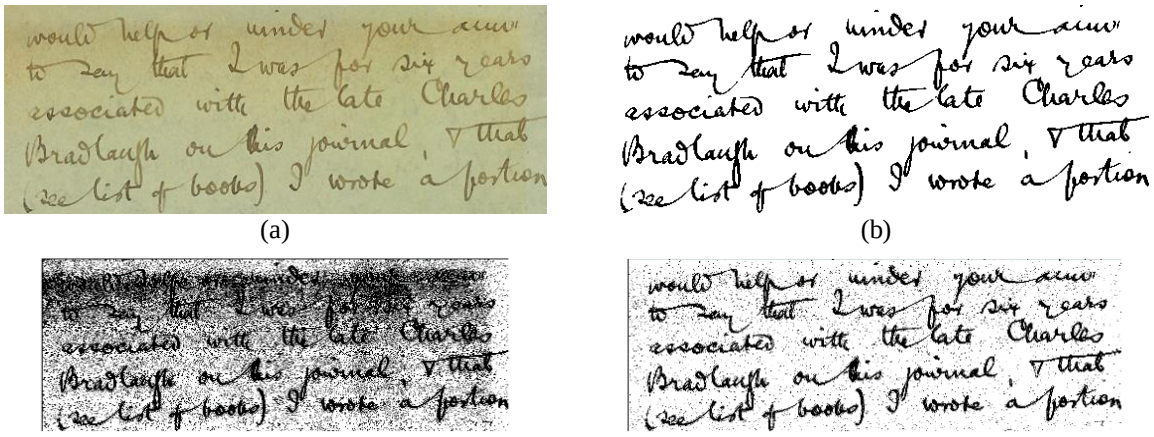
$$NRM = \frac{NR_{FN}+NR_{FP}}{2} \text{ where } NR_{FN} = \frac{FN}{FN+TP} \text{ and } NR_{FP} = \frac{FP}{FP+TN} \quad (3)$$

In this expression (3), NRM shows the mismatches between the final binary image and the ground truth image and thus, a lower value of NRM indicates a higher similarity between the two images.

The results obtained from the calculations are shown in Table 1. For HDIBCO2014, the F-measure value and NRM for the grey level were found to be much better compared to RGB. The F-measure for grey level values was 55.52 compared to 52.65 for RGB. This can also be seen from the NRM values, however, the NRM values for the grey level were much lower compared to RGB. Similar results also were obtained for DIBCO2012 and DIBCO2016. F-measure for grey level values was 44.42 compared to 42.13 for RGB for images in DIBCO2012. Whereas, for DIBCO2016, F-measure for grey level values is only slightly higher compared to RGB values. This is supported by NRM values by having smaller values for grey level compared to RGB for both datasets, DIBCO2012 and DIBCO2016. By observing the results of every image, some images were better if RGB was used, and some images were not. However, PSNR for the grey level was lower than RGB in HDIBCO2014 and DIBCO2016. This is due to trade off that occurs when some resultant images are better when using RGB compared to the grey level. Figures 5, 6 and 7, show the binarization results for the image H09.png in DIBCO2012, H09.png in HDIBCO2014 and 10.tif in DIBCO2016 respectively. Figures 5 and 7, shows that better resultant images were obtained when grey level values were used. Whereas, in Figure 6, better resultant images were obtained when RGB values were used.

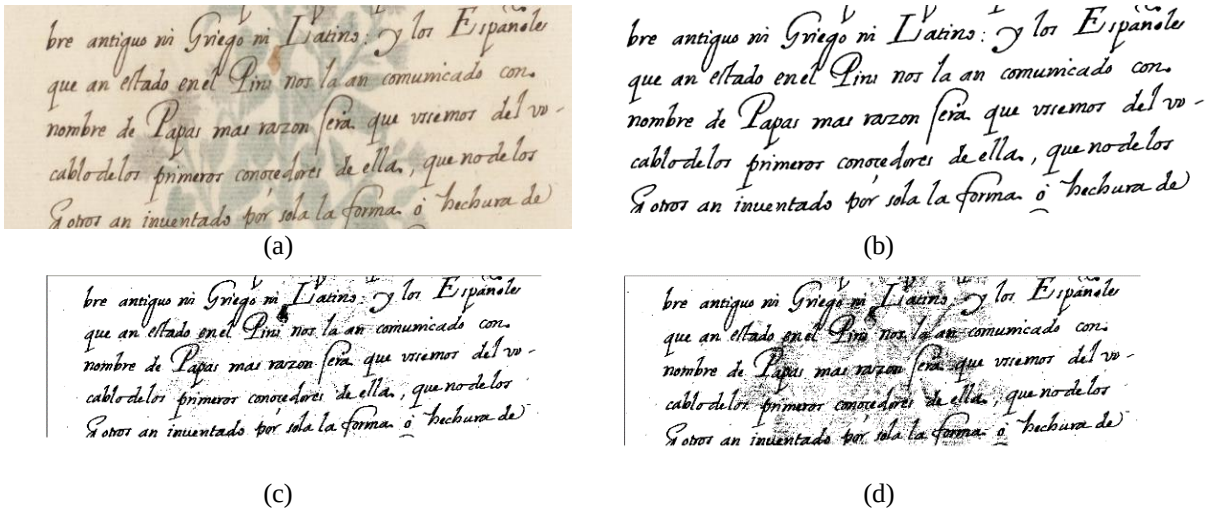
Table 1. Results of Average F-Measure, PSNR and NRM Values with Standard Deviation for HDIBCO2014, DIBCO2012 and DIBCO2016

Database	RGB			Grey Level		
	F-measure	PSNR	NRM	F-measure	PSNR	NRM
HDIBCO2014	52.65±13.18	10.32±2.58	0.26±0.09	55.52±14.96	10.21±2.53	0.22±0.09
DIBCO2012	42.13±17.29	8.72±4.33	0.26±0.10	44.42±11.36	9.41±2.20	0.21±0.07
DIBCO2016	56.04±12.47	11.27±4.30	0.22±0.07	56.12±9.79	10.91±3.54	0.19±0.07



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

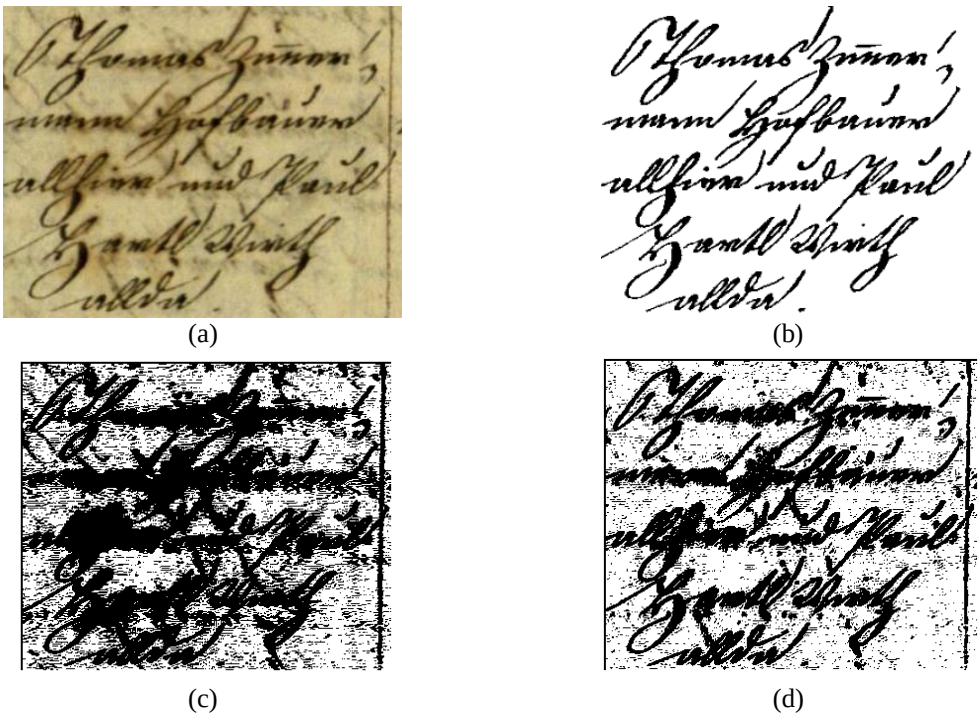
Figure 5. Binarization results for H09.png in DIBCO2012 (a) original image (b) ground truth image (c) binarization result for RGB with F- measure = 35.43361 (d) binarization result for grey level with F-measure = 57.92860



(a) (b)

(c) (d)

Figure 6. Binarization results for H09.png in HDIBCO2014 (a) original image (b) ground truth image (c) binarization result for RGB with F- measure = 48.81100 (d) binarization result for grey level with F-measure = 45.37836



(a) (b)

(c) (d)

Figure 7. Binarization results for 10.bmp in DIBCO2016 (a) original image (b) ground truth image (c) binarization result for RGB with F- measure = 36.04240 (d) binarization result for grey level with F-measure = 41.23519

5. DISCUSSION AND CONCLUSION

The experimental results demonstrated that RGB values and grey level values could be used as a descriptor for a particular pixel. For the overall conclusion, the F-measure and NRM values were found to be better when using grey level values compared to RGB. By observing the original image, the colour of the image did not show varieties in colour and was more towards a grey level image. Hence, it was not suitable to use RGB values as descriptors for every pixel. Furthermore, by using grey level values, the algorithm simplifies, and computational requirements are reduced. Whereas, the RGB colour introduces unnecessary information and the amount of training data increases in order to achieve better performance. Also, the RGB

values do not have colour difference sensitivity and cannot measure small colour difference [26]. One of the advantage of the proposed method is that it is easy, can use same model for different database if used for same type of images and competent of obtaining good results. However, this method consumes much time, and the ground truth images will need to be used for training the models. For future study, it is proposed to use other types of colour features, for example, HSV and CIELAB and to observe their impact on the binarization of historical documents.

ACKNOWLEDGEMENTS

Our deepest gratitude and thanks to Universiti Teknikal Malaysia Melaka (UTeM) and the Ministry of Higher Education Malaysia for funding this research grant under Short Term Research Grant (Grant no: S01629-PJP/2018/FTMK(2B)).

REFERENCES

[1] Q. N. Vo, S. H. Kim, H. J. Yang, and G. Lee, "Binarization of Degraded Document Images based on Hierarchical Deep Supervised Network," *Pattern Recognit.*, vol. 74, pp. 568-586, 2018.
[2] 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.
[3] M. M. Almeida, R. D. Lins, R. B. Bernardino, D. Jesus, and B. Lima, "A New Binarization Algorithm for Historical Documents," *J. Imaging*, no. Special Issue, pp. 1-12, 2018.
[4] A. Ehsan, A. Zohreh, S. Maryam, F. Mahmoud, and S. Mohammad Javad, "Document Image Binarization using A Discriminative Structural Classifier," *Pattern Recognit. Lett.*, vol. 63, pp. 36-42, 2015.
[5] N. Otsu, "A Threshold Selection Method from Gray-Level Histograms," *IEEE Trans. Syst. Man, Cybern.*, vol. SMC-9, no. 1, pp. 62-66, 1979.
[6] J. Kittler and J. Illingworth, "Minimum Error Thresholding," *Pattern Recognit.*, vol. 19, no. 1, pp. 41-47, 1986.
[7] C. H. Li and C. K. Leet, "Minimum Cross Entropy Thresholding," vol. 26, no. 4, 1993.
[8] W. Niblack, "An Introduction to Digital Image Processing". *Strandberg Publishing Company*, 1985.
[9] J. Sauvola and M. Pietikainen, "Adaptive Document Image Binarization," *Pattern Recognit.*, vol. 33, pp. 225-236, 2000.
[10] A. Ehsan, A. Zohreh, S. Maryam, F. Mahmoud, and S. M. Javad, "Document Image Binarization using A Discriminative Structural Classifier," *Pattern Recognit. Lett.*, vol. 63, pp. 36-42, 2015.
[11] R. Farrahi Moghaddam and M. Cheriet, "AdOtsu: An Adaptive and Parameterless Generalization of Otsu's Method for Document Image Binarization," *Pattern Recognit.*, vol. 45, no. 6, pp. 2419-2431, Jun. 2012.
[12] F. Kasmin, A. Abdullah, and A. S. Prabuwono, "Ensemble of Steerable Local Neighbourhood Grey-Level Information for Binarization," *Pattern Recognit. Lett.*, vol. 98, pp. 8-15, 2017.
[13] C. Becker and R. Rigamonti, "KernelBoost : Supervised Learning of Image Features For Classification," *Epfl Tr*, 2013.
[14] T. Ã. Hosaka, T. Kobayashi, and N. Otsu, "Image Segmentation using MAP-MRF Estimation and Support Vector Machine," *Interdiscip. Inf. Sci.*, vol. 13, no. 1, pp. 33-42, 2007.
[15] Y. Zhu, "Augment Document Image Binarization by Learning," *2008 19th Int. Conf. Pattern Recognit.*, pp. 1-4, 2008.
[16] C. Mohamed, M. R. Farrahi, and H. Rachid, "A Learning Framework for the Optimization and Automation of Document Binarization Methods," *Comput. Vis. Image Underst.*, vol. 117, pp. 269-280, Mar. 2013.
[17] B. Su, S. Lu, and C. Lim Tan, "A Self-Training Learning Document Binarization Framework," *Proc. - Int. Conf. Pattern Recognit.*, pp. 3187-3190, 2010.
[18] W. Xiong, J. Xu, Z. Xiong, J. Wang, and M. Liu, "Degraded Historical Document Image Binarization using Local Features and Support Vector Machine (SVM)," *Optik (Stuttg).*, vol. 164, pp. 218-223, 2018.
[19] J. Calvo-Zaragoza, G. Vigliensoni, and I. Fujinaga, "Pixel-Wise Binarization of Musical Documents with Convolutional Neural Networks," *Proc. 15th IAPR Int. Conf. Mach. Vis. Appl. MVA 2017*, pp. 362-365, 2017.
[20] C. Hsu, C. Chang, and C. Lin, "A Practical Guide to Support Vector Classification," 2010.
[21] K. Ntirogiannis, B. Gatos, and I. Pratikakis, "ICFHR2014 Competition on Handwritten Document Image Binarization," 2014, pp. 809-813.
[22] I. Pratikakis, B. Gatos, and K. Ntirogiannis, "2012 International Conference on Frontiers in Handwriting Recognition ICFHR 2012 Competition on Handwritten Document Image Binarization," 2012.
[23] I. Pratikakis, K. Zagoris, G. Barlas, and B. Gatos, "ICFHR 2016 Handwritten Document Image Binarization Contest (H-DIBCO 2016)," 2016, pp. 619-623.
[24] 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, Oct. 2011.
[25] A. Kefali, T. Sari, and H. Bahi, "Foreground-Background Separation by Feed-forward Neural Networks in Old Manuscripts," *Informatica*, vol. 38, pp. 329-338, 2014.
[26] X. Y. Wang, T. Wang, and J. Bu, "Color Image Segmentation using Pixel Wise Support Vector Machine Classification," *Pattern Recognit.*, vol. 44, no. 4, pp. 777-787, 2011.

BIOGRAPHIES OF AUTHORS

FAUZIAH KASMIN received B. Sc (Mathematics) and M. Sc (Applied Statistics) at Universiti Putra Malaysia. She received the PhD degree from Universiti Kebangsaan Malaysia in 2018. She is currently a senior lecturer in Department of Intelligent Computing and Analytics, Faculty of Information and Communications Technology, Universiti Teknikal Malaysia Melaka. Her current research focuses on the algorithmic aspects of image processing.

ZURAINI OTHMAN is currently a senior lecturer in the Department of Intelligent Computing and Analytics, Faculty of Information & Communication Technology, Universiti Teknikal Malaysia Melaka (UTeM). She received her Bachelors and Masters degrees of Applied Mathematics in School of Mathematics from University Science Malaysia. Her current research focuses on the mathematical modelling and algorithmic aspects of image processing.

SHARIFAH SAKINAH SYED AHMAD is currently an associate professor in the Department of Intelligent Computing and Analytics, Faculty of Information & Communication Technology, Universiti Teknikal Malaysia Melaka (UTeM). She received her Bachelors and Masters degrees of Applied Mathematics in School of Mathematics from University Science Malaysia. Following this, she received her Ph. D from the University of Alberta, Canada in 2012 in Intelligent System. Her research in graduate school focused on the granular computing and fuzzy modeling. Her current research interests including evolutionary optimization, fuzzy system, and granular computing. Her current research work is on the granular fuzzy rule-based system, Evolutionary Method, Modeling and Data Science.