Handwritten Document Image Binarization: An Adaptive K-Means Based Approach

Prithwish Jana, Soulib Ghosh, Suman Kumar Bera and Ram Sarkar

Department of Computer Science and Engineering

Jadavpur University, Kolkata, India

jprithwish@gmail.com, ghoshsoulib@gmail.com, berasuman007@gmail.com, raamsarkar@gmail.com

*Abstract*—Degraded historical document images face many challenges in the process of optical character recognizing or word spotting, even after applying the traditional binarization techniques. In this paper, we propose a K-means based clustering technique for adaptive binarization of degraded document images. For validation of test results, we have used the recent dataset of Handwritten counterpart of Document Image Binarization Contest (H-DIBCO’16) comprising of highly degraded handwritten document images and computed detailed results of each image. In order to corroborate verification and validation, the experimental results are compared with three top winning ones in the contest and other prominent techniques in the literature. Experimental results reveal outstanding performance in the four evaluation measures compared with the top winners of the competition, claiming its effectiveness and validity conformance.

Keywords— document image binarization, background estimation, global and adaptive thresholding, K-means, H-DIBCO.

# Introduction

Document image binarization (DIB) attempts to categorize pixels in a degraded input image into two subdivisions based on gray-level pixel intensity: *foreground* and *background*. Foreground subdivision contains texts in documents represented by black pixels having low intensity whereas document background segment is represented by white pixels of high intensity. DIB has been used as a very important preprocessing stage in recognition and analysis of document images and helps in the downstream processing that deal with information retrieval and optical character recognition (OCR). Image binarization facilitates a number of ensuing tasks such as detection of line, slant, slope, skew and word estimation. Other application areas include restoration of historical document and verification of signature. DIB is still considered as a major research area because of its complex challenges due to appearance of noise, complex background, uneven illumination and faint foreground often caused by smear, strain, bleed through, blur, ageing factor, seepage of ink etc.

Handwritten Document Image Binarization Contest (H-DIBCO) [1] is an established and popular forum in the field of DIB. Since 2009, DIBCO and H-DIBCO provide publicly available benchmarking dataset not only to the participants in the contest, but also to the research community for the purpose of upcoming work in the said field. This paper is organized in five sections. In following part, we present a brief literature survey related to several binarization methods. In section three, we present the proposed methodology through five major steps. This is followed by section four comprising experimental results and verification. Finally, the conclusions and future work are narrated.

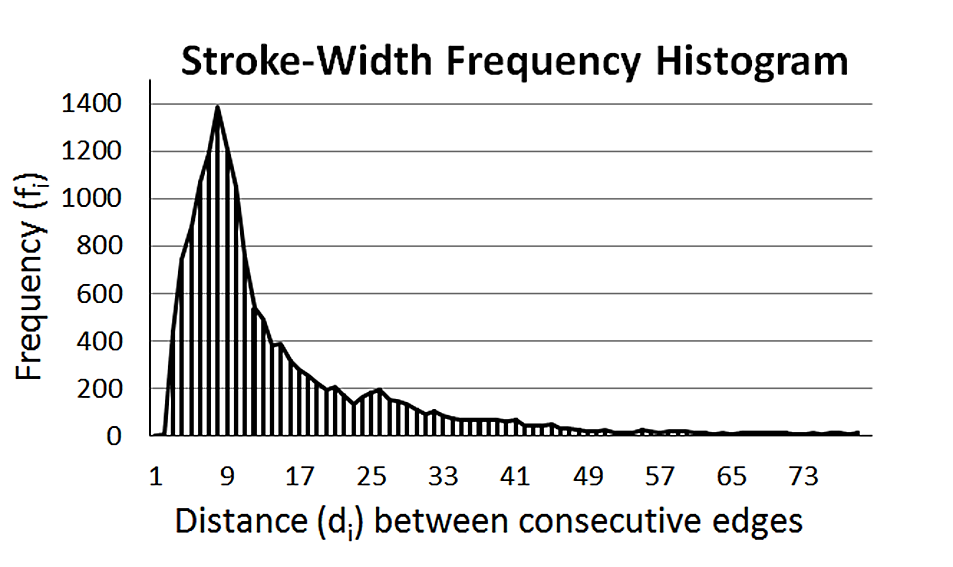
# Related Work

The desire to upsurge performance-wise efficiency of degraded document has appealed DIB to be explored colossally, resulting in publication of a large number of binarization algorithms. Generally speaking, the most commonly adopted techniques deal with threshold estimation for each pixel using *global* or *local* method. In *global thresholding* method, a single threshold value is determined to be applied to the whole image in order to put aside the pixels under consideration into *foreground* and *background*. The classical clustering based technique by Otsu [2] causes minimal intra-class variance using standard deviation and mean of bimodal histogram. It comprehensively searches a global threshold value taking weighted sum of variances. Other well-known methods include moment-preserving deterministic way by Tsai [3], entropy measure from gray level distribution by Johannsen and Bille [4] and entropy of histogram by Kapur et al. [5]. The main drawbacks of the global thresholding method are that it cannot handle complex background, faint foreground, bleed through etc. Moreover, global thresholding is not adaptive but the computational cost is comparatively very less.

In case of *local or adaptive thresholding*, same threshold is never used throughout the entirety of the image. Rather, the properties of a pixel and its neighbors in a sub-image help to determine the threshold. Niblack [6] uses rectangular window sliding over the gray-level image to determine pixel-wise threshold that adapts using standard deviation and local mean. Gatos et al. [7] use a low-pass Wiener filter, estimate foreground parts and interpolate background intensities of neighboring pixels to compute background surface. The computed background surface is combined with the original followed by a post-processing stage after image up-sampling for improved quality with retained stroke connectivity. *Adaptive* or *local thresholding* also helps to handle complex background, faint foreground, to an extent but it is computationally expensive and mostly manual parameter tuning driven.

Background separation method using efficient background surface computation is presented by Gatos et al. [8]. Moghaddam and Cheriet [9] adopt shrinkage of wavelet or time-stepping technique and use reverse diffusion process for both sides of document remain degraded. Drira et al. [10] suggest to apply local PDE based anisotropic diffusion filter to diminish the underlying rounding corner problem and also help reinforce character discontinuity. Ramirez-Ortegon et al. [11] introduce the idea of transition pixel categorized by extreme transition metric by calculating differences of pixel-intensity in a small sub-window neighborhood for use in adaptive gray-level threshold computation. Hu et al. [12] adopt a region-based segmentation to extract hieroglyph strokes from images of degraded ancient Maya codices while preserving delicate local details. The inherent drawback is the performance bottleneck due to innate sequential nature and finer segmentation. Moran [13] adopts a learning technique to binarize using hashing-based search of estimated adjacent neighbor at the cost of voluminous training data and storage. Shen et al. [14] formulate multiclass classification as an optimization problem using binary hash codes solving a sub-problem efficiently using either a linear program or a binary quadratic program (BQP). DIB has been a popular research topic for over three decades and even today produces interesting results over improvements focusing on several prejudice and limitations, adding impetus to our motivation.

# The Proposed Methodology

The proposed method consists of five distinct major steps that eventually culminate in effective binarization of the degraded document images. Initially, the average stroke-width of the candidate foreground regions is computed. Stroke-width at a certain pixel is defined as the thickness of curves and lines that make up a character, measured between two opposite contour pixels, along any direction, generally, horizontal, vertical or any diagonal. In our work, the horizontal stroke-width is considered. This value is used in selection of the size of structuring element to estimate the background. Finally, the background separated image is partitioned into uniformly symmetric blocks and based on a certain condition each block is binarized separately with the aid of K-Means clustering.

1. Histogram of frequency distribution of stroke-width

## Calculation of Stroke-Width in Text Regions

The average stroke-width () of characters in a document image helps to assess whether the image contains a majority of thick text-strokes or thin text-strokes. This value of is used in later parts of our work to automate the selection of the size of structuring element in morphological operations. At first, edge detection of the document image is done, using Sobel edge detection algorithm as presented by Vincent and Folorunso [15]. Sobel’s method is preferred because the use of an average filter applies a smoothening effect to random high intensity point-noises and edges appear thick, bright and enhanced.

The edge-detected image is henceforth termed as . Initially, all the connected components of are detected. Out of those, the small connected components are eliminated [16] as these components are isolated noise pixels in the original image. So it is wise not to count them in the computation of the stroke-width. For the large connected components, the height and width of the smallest rectangle bounding them is computed. Each row of the bounding-box is scanned from left to right. In each row, the horizontal distance (d) between 2 special kinds of pixels, is measured. The first is a pixel which is itself an edge-pixel while the pixel to its immediate right is a non-edge pixel, and the second is a pixel which is itself an edge-pixel while the pixel to its immediate left is a non-edge pixel. This distance d is considered as a single stroke-width. Ultimately a histogram (H) of the various stroke-widths () versus their frequency () of occurrence is formed, for . Average of the stroke-widths corresponding to three highest frequencies of H is considered as the average stroke-width (). A histogram showing stroke-width frequency distribution corresponding to Image-3 of H-DIBCO’16 [1] dataset is provided in Fig. 1.

## Background Estimation

Binarization algorithms, like that of Otsu’s [2], provide best outcome when the intensity histogram of an image has at most two sharp peaks. But in general, it fails to efficiently handle the documents with degradations like bleed-through, uneven background, background ink-stains, non-uniform illumination, and also those degraded by image artifacts. Background separation and image normalization help in providing a bimodal histogram to this type of image, thus making the further procedure of binarization pretty much easier. The proposed method employs a fast and proficient yet simple process for estimating an approximate background surface () from the original grayscale image, Initially, a structuring element [16] of size is formed with each of its element as 1, as shown in Eq. (1), where .



The structuring element defines the neighborhood of the pixel of interest, located at its well-defined center-pixel (because it is odd-dimensional). Grayscale erosion [17] of an image locates the minimum intensity of the points in neighborhood of the center-pixel, where the neighborhood is defined by the structuring element. Thus it is similar to a local-minimum operator. Consequently, this process removes all the small components and shrinks the dimension of the larger components. The grayscale erosion of by is denoted by Eq. (2), where D represents the domain of , and D = [-,+].



Grayscale dilation [17] of an image locates the maximum intensity of the points in neighborhood of the center-pixel, and returns this maximum value within the moving window of . Thus, it is similar to a local-maximum operator. Consequently, this process probes and expands the shapes contained in its input image. The grayscale dilation of by is denoted by Eq. (3), where D represents the domain of , and D = [-,+].



The proposed method estimates from by successive morphological closing and opening operation [18] on . This allows the eradication of undesirable regions without significantly affecting the remaining structures of the image. Thus it is beneficial for recovering structures which are not completely destroyed by erosion or dilation, which is effectively the approximate illumination of , i.e. .

|  |  |  |
| --- | --- | --- |
|  |  |  |
| (a) | (b) | (c) |
|  |  |  |
| (d) | (e) | (f) |

1. Original versus corresponding background images: (a)-(c) as original image and (d)-(f) background image of (a)-(c) respectively

The morphological closing of a grayscale image by a structuring element is dilation followed by erosion, using the same structuring element for both operations. On a counter note, the morphological opening of a grayscale image by a structuring element is erosion followed by dilation with the same structuring element. So, in accordance with the proposed method, initially is passed through a morphological closing operation by . This removes small objects from an image while preserving the shape and size of bigger objects in the image. This image is then subjected to a morphological opening operation by the same structuring element . This operation fills almost all the gaps in the closed image. The image obtained by performing these two processes successively, is Fig. 2 shows three variations of original images and their corresponding background images respectively.

## Normalization of Image

The variation in appearance induced by the changes in illumination levels across the image has been a formidable problem for binarization. Hence it is essentially a subtle necessity to do away with this variation. So, after the estimation of the approximate background of a document image possessing uneven background and non-uniform illumination, the next step undertaken is to approximately subtract the background image, from the original grayscale image, . This brings about a nearly uniform intensity for all the background pixels and thus effectuates a bi-modal histogram to the image. Lighting condition of the picture subsequently gets balanced, and the image becomes easier to binarize with considerable success. The original grayscale image is subtracted [16] from the background image estimated from background estimation stage and thus the variations of illumination in are practically eradicated and an image, is obtained. is acquired as in Eq. (4).



This results in an image with the 0-side representing background and 255-side representing foreground. So, the final background separated image (with the 0-side representing foreground and 255-side representing background), is obtained by complementing as in Eq. (5) where R,C represents the height and width of , respectively.



|  |  |  |
| --- | --- | --- |
|  |  |  |
| (a) | (b) | (c) |

1. Original images of Fig. 2 in their normalized forms

Fig.3 shows the results of normalization done on the original images given in Fig. 2.

## Quadtree-based partitioning

Post background separation and image normalization, the image is devoid of local illumination variations, like shadows and highlights, as well as global illumination variations. As a consequence, the image is henceforth enhanced, before an adaptive thresholding algorithm is applied. This proposed method puts forward a quadtree-based algorithm [19] for image partitioning. Thus, it combines the noble traits of global (high speed) and local (high accuracy) binarization methods. The task is to delineate regions in an image into a number of blocks of equal sizes which are thereafter binarized individually. The proposed method takes the advantage of the fact that the partitioning of the image minimizes the within-class variance, thus simplifying further binarization steps.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

1. Partitioning the normalized image into blocks

Initially the background separated image, is partitioned into four rectangular sub-images by splitting both the height and width into two equal halves. Each of the subdivisions is again subdivided into four equal partitions in a similar manner, by slicing across their half-height and half-width points. The process of quadtree-partitioning is stopped at the level of two because after this, the sub-image will become too small, and it may contain only background or only foreground, thus making the process of binarization erroneous. This proposed method of partition on one hand is simple and very fast, and on the other hand it reduces the significant variations in the gray levels in each block. Thus the process of binarization, in each of these blocks individually, is streamlined to a large extent. Fig. 4 shows the effect of partitioning on the normalized image into blocks.

## Binarization by clustering in the individual blocks

Finally, sixteen blocks are obtained from the partitioning of the background-separated image. In this stage, we employ a multivariate K-means clustering algorithm [20] for performing segmentation in each block based upon fulfillment of a condition, else, all pixel intensities are transformed to 255 (white). For each block, at first, the range (*rng*) i.e. the difference between the maximum (*max*) and minimum (min) intensity, mean (*avg*) and standard deviation (*std*) of all the pixel intensities are measured and the following condition is defined as Eq. (6). The condition, = 1 depicts that has a very high mean and low standard deviation, thus it is reasonably concluded to be an all-background block.



Accordingly, for best binarization result, the proposed method refrains itself from applying clustering algorithm to it, and hence it is left unbinarized and all pixel intensities in are set to 255 (white). We found optimally that and give best results when they are in range and respectively. Otherwise the block is binarized through K-means clustering algorithm. The proposed method uses the following variables for the multivariate K-Means clustering in a 3×3 window:

* Intensity of each central pixel
* Mean of intensities of the 8-connected neighboring pixels of each central pixel
* Standard deviation of intensities of the 8-connected neighboring pixels of each central pixel

Values outside non-overlapping bounds are computed by mirror-reflecting the intensities across the boundary.

Clustering algorithms are based upon the index of similarity or dissimilarity between each pair of data points. K-means is one of the unsupervised learning algorithms for clustering and make the decision whether the corresponding pixel intensity is text pixel or non-text pixel. Its major aim is to delineate regions in a sub-image into a number of clusters, which is carried out by identifying regions that are homogeneous in some set of local image attributes. The segregation is such that there exists high similarity of the samples inside the clusters (minimization of clusters within group) and high dissimilarity between samples belonging to distinct clusters (maximizing the difference between groups). Initially, the number of clusters (k) is defined which is 2 for the present scenario, since the proposed method aims to segregate the sub-image into two subdivisions, viz. foreground (represented as black) and background (represented as white). Then center of k clusters are randomly selected and the distance between the each pixel to each cluster centers is measured. For the comparison of the distance, the Euclidean distance between the centers is used. The pixel is relocated to cluster which has the shortest distance among the clusters. After re-estimation of the new centroid, each pixel is compared to the k centroids and reassigned to the nearest updated center. This process reiterates until the center converges to a solitary point. Finally, pixels of each (where equals zero) are segmented into k-clusters (k considered as 2 here) distinctly categorized into *background* and *foreground*.

# Experimental Results and Observation

For the test results, we have used the most recent dataset of H-DIBCO (H-DIBCO’16 [1]), the handwritten images of the DIBCO series. This publicly available database contains 12 diverse competing methods from nine different research groups respectively. All the images and their respective ground truths are considered from the competition’s site [21]. H-DIBCO is a recognized and prevalent forum in the DIB field. These datasets provide challenging assignment with handwritten set of characters alongside embedded noises and distortions to be barely readable in it’s as-is form.

1. Comparative Results in Applicable Measures

| Measures based on H-DIBCO’16 dataset | | | | |
| --- | --- | --- | --- | --- |
| Method | F-Measure |  | PSNR | DRD |
| 3rd | 88.47 ± 4.45 | 91.71 ± 4.38 | 18.29 ± 3.35 | 3.93 ± 1.37 |
| 2nd | 88.72 ± 4.68 | **91.84 ± 4.24** | 18.45 ± 3.41 | **3.86 ± 1.57** |
| 1st | 87.61 ± 6.99 | 91.28 ± 8.36 | 18.11 ± 4.27 | 5.21 ± 5.28 |
| Otsu | 86.61 ± 7.26 | 88.67 ± 7.99 | 17.80 ± 4.51 | 5.56 ± 4.44 |
| Sauvola | 82.52 ± 9.65 | 86.85 ± 8.56 | 16.42 ± 2.87 | 7.49 ± 3.97 |
| Proposed | **89.08 ± 5.31** | 90.26 ± 6.15 (4th) | **18.48 ± 4.22** | 4.474 ± 3.32 (3rd) |

1. Results of H-DIBCO’16 Images in Terms of Different Measuring Parmeters

| H-DIBCO’16 Images | | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Measures | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| F-Measure | 93.10 | 85.83 | 95.96 | 88.63 | 96.83 | 88.19 | 88.27 | 79.32 | 90.28 | 84.44 |
|  | 93.64 | 87.96 | 95.84 | 95.12 | 96.58 | 91.06 | 94.74 | 79.31 | 86.90 | 81.47 |
| PSNR | 20.05 | 22.40 | 23.94 | 19.31 | 23.52 | 18.23 | 16.55 | 11.58 | 16.28 | 12.91 |
| DRD | 4.69 | 4.57 | 1.58 | 4.16 | 1.13 | 5.70 | 3.02 | 12.92 | 2.23 | 4.69 |

Two foremost objectives for any performance assessment are certainly its usefulness for practice and efficiency when compared with other methods. In Table I, the comprehensive results of the three best contending techniques in the H-DIBCO contest of 2016 are provided. Values, in bold, denote the first rank among all the competitors, as well as, Otsu’s, Sauvola’s in each corresponding column. For the evaluation purpose, an ensemble of measures has been used. In the contest of H-DIBCO’16, the evaluation measures of *F-Measure* (*FM*), *pseudo-FMeasure* (), *Peak Signal to Noise Ratio* (*PSNR*), *Distance Reciprocal Distortion* (*DRD*) Metric are used [1]. It can be arbitrated that even without deploying any post-processing step after binarization by the suggested schematized mandate of stages, the method beats the best competing techniques in the H-DIBCO 2016 in two fields, and stands in close competition with the top three in the other two measures. The outcomes of the proposed technique are collated along with them, and compared with the former. It is evident from Table I that the proposed method outperforms all the participants of H-DIBCO’16 in the measuring parameters of *FM* and *PSNR*, while it secures the third and fourth position respectively in the fields of *DRD* and . In Table II, we present the detailed results using the DIBCO-provided evaluator applied on each image of H-DIBCO’16. For some images, we are not getting better results because specific challenge needs to be dealt separately. Several original images from H-DIBCO’16 dataset and their binarization results are accorded in Fig. 5.

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
| (a) | (b) |

1. (a) Few selected dataset images from H-DIBCO and (b) Binarization results obtained using proposed method.

# Conclusion and Future Work

We have used a multivariate K-means clustering algorithm after partitioning the background-separated normalized image. This effectively binarizes degraded document images, and thereby making them ready for better image retrieval and downstream processing. The contribution lies in the fact that superlative results are obtained without the use of any complex blocking or post-processing. Moreover it employs a simple clustering algorithm like K-Means, which is not even required in all the blocks. Our proposed method shows outstanding performance in experiments involving better than the best with two measures and close to first three results of the competition using the dataset provided by H-DIBCO 2016. However, a scope for further improvement still remains to bring out more translucent foreground texts from the noisy background, thereby providing room for our future work with improved results. In this regard, we may apply a hybrid approach of global and adaptive thresholding techniques as a tradeoff that could deal with specific challenges.

##### References

1. I. Pratikakis, K. Zagoris, G. Barlas and B. Gatos, "ICFHR2016 Handwritten Document Image Binarization Contest (H-DIBCO 2016)," *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, Shenzhen, 2016, 619-623.
2. N. Otsu, “A threshold selection method from graylevel histograms”, *Automatica*, vol. 20, no. 1, 1975, 62–66.
3. W. Tsai, “Moment-preserving thresholding: a new approach”, *Computer Vision, Graphics and Image Processing*, 29 (3), 1985, 377–393.
4. G. Johannsen and J. Bille, “A threshold selection method using information measures”, *6th ht. Conf. on Pattern Recognition*, Munich, Germany, 1982, 140–143.
5. J. N. Kapur. P.K.Sahoo, A. K. C. Wong, “A new method for gray-level picture thresholding using the entropy of the histogram”, *Computer Vision, Graphics, and Image Processing*, V. 29, Issue 3, 1985, 273-285.
6. W. Niblack, “An Introduction to Digital Image Processing”, *Strandberg Publ. Co*., Birkeroed, Denmark, 1985.
7. B. Gatos, I. Pratikakis, S. J. Perantonis, “Adaptive degraded document image binarization”, *Pattern Recogn.* 39 (3), 2006, 317–327.
8. B. Gatos,., Pratikakis I., Perantonis S.J., “An Adaptive Binarization Technique for Low Quality Historical Documents”, Marinai S., Dengel A.R. (eds) *Document Analysis Systems VI. DAS 2004*. Lecture Notes in Computer Science, vol 3163. Springer, Berlin, Heidelberg, 2004.
9. R. F. Moghaddam, M. Cheriet, “A variational approach to degraded document enhancement’, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32 (8) (2010) 1347–1361.
10. F. Drira, F. Lebourgeois, H. Emptoz, “A new PDE-based approach for singularity-preserving regularization: application to degraded characters restoration”, *Int. J. Doc. Anal. Recognit.* 15, 3, 2012, 183-212.
11. M. Ramirez-Ortegon, E. Tapia, L. Ramirez-Ramirez, R. Rojas, E. Cuevas, “Transition pixel: a concept for binarization based on edge detection and gray-intensity histograms”, *Pattern Recogn.* 43, 2010 1233–1243.
12. R. Hu, J. Odobez and D. Gatica-Perez, “Extracting Maya Glyphs from Degraded Ancient Documents via Image Segmentation”, *J. Comput. Cult. Herit.* 10, 2, Article 10, 2017, 23 pages.
13. S. Moran, “Learning to Project and Binarise for Hashing Based Approximate Nearest Neighbour Search”, *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval (SIGIR '16),* ACM, New York, NY, USA, 2016, 897-900.
14. F. Shen, Y. Mu, Y. Yang, W. Liu, L. Liu, J. Song, and H. T. Shen, “Classification by Retrieval: Binarizing Data and Classifiers”, *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '17)*. ACM, New York, NY, USA, 2017, 595-604.
15. O. R. Vincent, O. Folorunso, “A Descriptive Algorithm for Sobel Image Edge Detection”, *Proceedings of Informing Science & IT Education Conference (InSITE) 2009*, 97-107.
16. S. Mandal, S. Das, A. Agarwal and B. Chanda, "Binarization of degraded handwritten documents based on morphological contrast intensification," *2015 Third International Conference on Image Information Processing (ICIIP)*, Waknaghat, 2015, 73-78.
17. I. Bloch,. "Duality vs. adjunction for fuzzy mathematical morphology and general form of fuzzy erosions and dilations", *Fuzzy Sets and Systems*, 160, no. 13 (2009): 1858-1867.
18. M. Pesaresi and J. A. Benediktsson. "A new approach for the morphological segmentation of high-resolution satellite imagery." *IEEE transactions on Geoscience and Remote Sensing* 39, 2 (2001): 309-320.
19. Z. F. Muhsin, A. Rehman, A. Altameem, Tanzila Saba, and M. Uddin, "Improved quadtree image segmentation approach to region information." *The Imaging Science Journal* 62, no. 1 (2014): 56-62.
20. S. Na, L. Xumin, and G. Yong, "Research on k-means clustering algorithm: An improved k-means clustering algorithm", *Third IEEE International Symposium on Intelligent Information Technology and Security Informatics (IITSI)*, pp. 63-67, 2010.
21. H-DIBCO 2016: ICFHR 2016 Handwritten Document Image Binarization Contest, URL: http://vc.ee.duth.gr/h-dibco2016/, last accessed: July, 2017.