

Information Density Based Image Binarization for Text Document Containing Graphics

Soma Datta, Nabendu Chaki, Sankhayan Choudhury

▶ **To cite this version:**

Soma Datta, Nabendu Chaki, Sankhayan Choudhury. Information Density Based Image Binarization for Text Document Containing Graphics. 15th IFIP International Conference on Computer Information Systems and Industrial Management (CISIM), Sep 2016, Vilnius, Lithuania. pp.105-115, 10.1007/978-3-319-45378-1_10 . hal-01637461

HAL Id: hal-01637461

https://inria.hal.science/hal-01637461

Submitted on 17 Nov 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

Information Density Based Image Binarization for Text Document Containing Graphics

Soma Datta, Nabendu Chaki, and Sankhayan Choudhury

Department of Computer Science and Engineering,University of Calcutta, JD2 Block,Sector III, Saltlake, Kolkata, India,
soma21dec@yahoo.co.in, nabendu@ieee.org, sankhayan@gmail.com

Abstract. In this work, a new clustering based binarization technique has been proposed. Clustering is done depending on the information density of the input image. Here input image is considered as a set of text, images as foreground and some random noises, marks of ink, spots of oil, etc. in the background. It is often quite difficult to separate the foreground from the background based on existing binarization technique. The existing methods offer good result if the input image contains only text. Experimental results indicate that this method is particularly good for degraded text document containing graphic images as well. USC-SIPI database is used for testing phase. It is compared with iterative partitioning, Otsu's method for seven different metrics.

Keywords: Iterative partitioning, NTSC color format, Wiener filter, Binarization, Entropy.

1 Introduction

It is very important to maintain the documents and the legacy of the document. To fulfill these purpose document image processing takes a vital role. Document image binarization is usually performed in optical character recognition (OCR) [1, 2] and image searching. This involves handwriting recognition, extracting logos and pictures from a graphical image. The main purpose of document image processing [3] is reduction of paper usage, easy access to the documents with lowest storage cost. At this point the most challenging task is to segment the region of interest (ROI) for further analysis. The simplest method for image segmentation [4] is thresholding based binarization which is also an essential technique in enhancement and biomedical image analysis. The output of this process is a binary image[5]. Though researchers work upon document image binarization for several years, the thresholding of compound document images still remains a challenging task due to its sensitivity to noise, illumination, variable intensity and sometimes insufficient contrast. It has been observed that some of the existing methods [6–10] offer very good result for text document. However, the performance degrades when a degraded text document contains some graphical images in it. We refer this type of document as compound document in the rest of this paper. In this research work, we aim to devise a new segmentation

methodology that would be good for the compound documents. We separate the entire image into three regions as the background, only text region and the graphical image. Our proposed method keeps a good balance both for text and graphics in the degraded compound documents. The proposed binarization method is based on cluster density information. It consists of six phases. These are noise removal with image normalization, entropy calculation, fuzzy c_mean clustering, segmenting of each region based on the clustering output, applying local threshold based binarization and finally integrating the segmented region. Each of these phases is described detail in the design methodology section.

2 Survey on Existing Techniques

Document image binarization has drawn lot of attention in the machine vision research community. Some of the highly cited methods are discussed in this section in a nutshell. Parker et al. proposed a method based on Shen-Castan edge detector to identify object pixels[7]. This method creates a surface using moving least squares method used to threshold images. Chen et al. proposed enhanced speed entropic threshold selection algorithm [8]. This method works upon the selection of global threshold value using maximin optimization procedure. O'Gorman proposed a global approach based on the measurement of information on local connectivity. The threshold values are incorporated at intensity level. Thus this method has advan-tages of local as well as global adaptive approaches. Liu et al. proposed a method based on grey scale and run length histogram. This method carefully handles noisy and complex background problems. Chang et al. worked upon stroke connectivity preservation issues for graphical images [11]. Their proposed algorithm is able to eliminate the background noise and enhancement of grey levels of texts. This method is used to extract the strokes from low level density as well as darker background. S. Shaikh, et al. [12] proposed iterative partitioning method. In this method, entire input image is divided into four equal sub images if the number of peaks is greater than two in the input image. This process will continue until the sub image contains less than two histogram peaks. This binarization method is offering good result for very old, faded, stained documented images but fails for medical image segmentation. In Otsu's [13] method, the thresholding is based on the class variance criterion and the histogram of the input image. This method segments the image into two classes, so that the total variance of different classes is maximized. Otsu's binarization technique produces good result for graphical images; however it can't properly binaries the old spotted document.

3 Design of New Information Density Based Binarization Technique

3.1 Image Acquisition and Enhancement

The design methodology follows a pipelined approach starting with image acquisition and ending with binarization. Histogram based Otsu's method may

provide satisfactory result when documented images are clear. However in reality the old and multiple times photocopied documents are not so good. Hence these documents are often not binarized properly using Otsu's method.

Otsu's method is a histogram based generalized binarization. Information density based segmentation is not done here. The algorithm assumes that the input image contains foreground and background pixels and it then calculates the optimal threshold that separates the two regions. USC-SIPI database [14] is used for testing phase. The quality and size of the original image is not changed. Instead of this database we have also tested our proposed method with some sample documented scanned images that consists of text and graphical images.

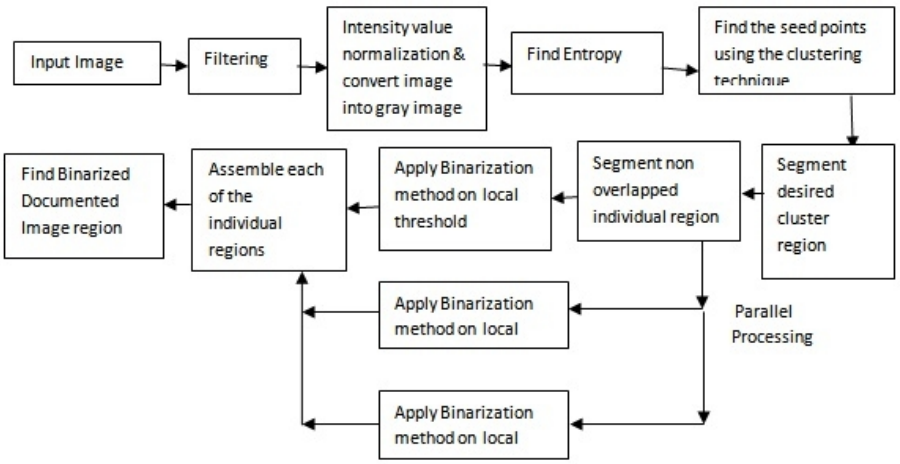


Fig. 1: Block diagram of proposed method

Image enhancement improves the visual quality of the input image for further processing techniques. The qualities of some images are not so good. Most of the image contains speckle noise, salt and pepper noises and some random noises. Here Wiener tilter is used to remove that kind of mixture noises. This filter is a linear filter to remove additive noise and blurring of the images. It offers good result to reduce the mean square error rate. These filters are often applied in the frequency domain. The corresponding output image is $W(f_1, f_2)$ as follows:

$$W(f_1, f_2) = \frac{H^*(f_1, f_2) S_x(f_1, f_2)}{|H(f_1, f_2)|^2 * S_x(f_1, f_2) + S_y(f_1, f_2)} \tag{1}$$

Here, $S_x(f_1, f_2)$ and $S_y(f_1, f_2)$ represent the power spectrum of the original image and the noisy image respectively and $H(f_1, f_2)$ means the blurred filter. The Wiener filter performs deconvolution during minimization of least square error as follows,

$$e_2 = E\{(f - \widehat{f})^2\} \tag{2}$$

Here E is the mean value and f is the un-degraded image.

3.2 Convert input color image into gray scale image

Color of a pixel is represented as the combination of chrominance and luminance. Chrominance is the color components of the input image and luminance is the intensity. This intensity is calculated as the weighted means of red, green and blue (RGB) component. Now image is in three dimensional that requires a massive computational time. Hence RGB images are converted into gray scale image using NTSC color format[2].

3.3 Intensity value normalization

Intensity normalization is very important towards handling the variable light intensity. Basically it is a method that maps the intensity values as per prerequisite. Normalization transforms an n dimensional grayscale image represented as, $Img = \{X \subseteq R^n\} \rightarrow \{Min, Max\}$. Image Img consist the intensity values in between Min and Max. This image is converted into a new image,$Img = \{X \subseteq R^n\} \rightarrow \{new_Min, \ldots., new_Max\}$. New image Img consist the intensity values in between new_Min and new_Max. Normalization is done by using the histogram. The following steps are made for normalization.

Algorithm 1: Intensity Value Normalization

Assumptions: Input image should be in gray scale.
Input: Gray scale image.
Output: Scaled intensity image.
Step

1: Read input image and find its size.
2: Find the cumulative distribution function as $Q(x) = \sum_{i=1}^{x} h(i)$ where x is the gray scale value and h is the image histogram.
3: Calculate new gray scale values using

$$h(i) = round(\frac{Q(1) - Q_{min}}{X * Y - Q_{min}} * (L - 1)) \tag{3}$$

where Qmin is the shortest value of the cumulative distribution function. X, Y represents the number of columns and rows of the input image and L is the number of gray levels that are used.
4: New image is created by replacing original gray values by the newly calculated gray values and compare this with the original.
5: Stop

3.4 Find the most informative regions

Input images may contain the actual information along with some distortion due to oil ink etc. Hence, it is very much important to segment the actual informative region. In order to find the most informative region, texture based information have been used. Among many texture properties, entropy is used to do the needful. Entropy [15, 16] refers to the disorder, uncertainty or randomness of the given dataset. The following algorithm is used to find the most informative regions. The covariance or probability of randomness is higher in the text area but it is less in non text area.

Algorithm 2: ROI segmentation

Assumptions: Input image should be in gray scale.
Input: Gray scale image.
Output: Only most wanted region.
Step

1: Find the sub image ($row1 - n/2$ to $row2 + n/2$ and $col1 - n/2$ to $col1 + n/2$) where row1 and col1 is two variable and n denotes the row, column of the sub-window.
2: Repeat step1 for all row1 is in between $row1 - n/2$ to $row1 + n/2$ and for all col1 is in between $col1 - n/2$ to $col1 + n/2$.
3: Repeat step 4 to 6 for all sub images and store the entropy value in row1 and col1.
4: Calculate histogram of n*n sub image and store the result in a vector hist.
5: Normalized histogram is stored at hist_normalized $hist_normalized = hist/(n * n)$
6: Find all the points at hist_normalized where value is 0 and replace that value by1.
7: Calculate entropy element and store them into cee vector $cee(i) = hist_normalized(i) * \log_2 normalized(i)$, i is used to denote index.
8: Add all entries from cee(i) and multiply the result by (-1)
9: Stop

3.5 Segment individual non-overlapped regions.

The next step is to segment the individual non-overlapped regions. Clustering method [17, 18] has been applied to find out different cluster seed points that are shown in Fig4.a. Here three cluster points have been found. Now our target is to segment each individual clustering region as shown in Fig2.b, Fig2.c and Fig2.d. The next step is to segment each of the regions.

$Single_Region_Set = \{P : Pisasubsetof Region_point_Set\}$

Elements of set P form a vector containing row and column number of a pixel. P contains the coordinates of all pixels of a single region. There is no common element between any two elements of $Single_Region_Set$.

$P_1 \cup P_2 \cup P_3 \cup P_n = Region_point_set$ Where $P_1, P_2, , P_n$ are all elements of $Single_Region_Set$, n is the total no of elements in $Single_Region_Set$. $P_i \cap P_j = \phi$where P_i, P_j are any two elements of $Single_Region_Set$ and value of i,

Algorithm 3: Clustering Segmentation

Assumptions: Let $E = \{e1, e2, e3..., en\}$is the set of entropy values in entropy matrix. $C = \{C1, C2, C3\}$ is the set of initial cluster centers.

Input: Gray scale image.

Output: Based on entropy calculation the image is segmented into three groups.

Step

1: Initial cluster centers are $\{C1, C2, C2\}$ i.e. c = 3 where, C1 is the average entropy value of most informative region in the document. C2 is the average entropy value of less informative region. C3 is the average entropy value of in between region.

2: Calculate the fuzzy membership 'F_{ij} ' using

$$F_{ij} = \frac{1}{\sum_{k=1}^{e=3}(\frac{D_{ij}}{D_{ik}})^{(2/m-l)}} \tag{4}$$

Where, F_{ij} are the elements of F_M matrix (fuzzy membership matrix), D_{ij} is the Euclidean distance between i^{th} object and j^{th} cluster centre. D_{ik} is the Euclidean distance be-tween i^{th} object and j^{th} cluster center, m is a constant.

3: New fuzzy centers New_F_j is calculated using,

$$New_F_j = (\frac{(\sum_{i=1}^{n}(F_{ij})^m X_i)}{(\sum_{i=1}^{n}(F_{ij})^m)})\forall j = 1, 2, ...c \tag{5}$$

4: Repeat Step 2 and 3 until $\|F_M(k+1)F_M(k)\| < \beta$ Where, 'k', 'β' are the iteration step and termination criterion between [0, 1] respec-tively. $F_M(k) = (F_{ij})n * c$ is a fuzzy membership matrix after k_{th} iteration and $F_M_(k+1)$ is used to denote after $(k+1)^{th}$ iteration. j is the objective function.

5: Find all the points which belong to the most informative region cluster. This cluster has C1 center and contains the pixel position of the image of most informative region.

6: Stop

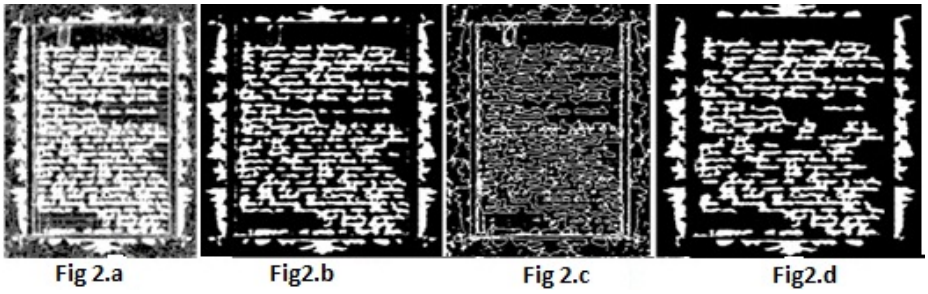


Fig. 2: Fig2.a: After applying clustering technique
Fig2.b: First most informative region whose cluster centre is 74.1604
Fig2.c: Second most informative region whose cluster centre is 167.7177
Fig2.d: Third most informative region whose cluster centre is 225.7846

j may be 1, 2, 3..n but $i \neq j$. Each component of *Single_Region_Set* contains the coordinates of a single contour, and they are found by applying procedure segment each region.

Algorithm 4: Segment region

Input: The binary matrix that does not contain any false contours.
Output: The coordinates of each contour.
Step

1: Find any coordinate from input binary matrix (*region_matrix*), containing value 1 and this point is called seed point.
2: Find all other coordinates of the binary matrix that belongs to the same contour with the seed point. It is implemented using the stack, 8-connectivity and 4 connectivity properties.
3: The points obtained from step2 are the points that represent a single contour. These points are sent to the next phase for further processing.
4: Repeat step2 to step3 using a newly found seed value that belongs to a separate region. This iterative process will continue until all disjoint regions are extracted.
5: Stop

After applying segment each region algorithm, the wanted regions are segmented. Now apply global binarization technique on the segmented regions. Threshold value for binarization is calculated as follows [19].

4 Performance Analysis

For performance analysis some degraded document images are used that are collected from multiple sites. Our method is tested on the USC-SIPI database arbitrarily for experimental verification purpose. The results are shown in Fig 3.

Algorithm 5: Binarization

Input: Gray scale image.
Output: binary image.
Step

1: Initial threshold value $T = (min_intensity + max_intensity)/2$
2: Partition the image into two regions, R1 consists of value (pixels)< T and R2 consists of value(pixels) > T.
3: For each region, calculate the average intensity value I1 and I2.
4: Preserve the old threshold value and update the threshold value with the new threshold value that is obtained from step3.
5: Continue this iterative process from step2 to step4 until the absolute difference between preserved value and new threshold value is less than tolerance parameter e.
6: Stop

In this work, we have compared seven different metrics in between Otsu's method, iterative partitioning and our proposed method. The seven metrics are recall, precision, F_measure, PERR or pixel error rate, MSE or mean squared error, SNR or signal to noise ratio and peak signal to noise (PSNR) [12, 20]. Theoretically recall is the probability of a relevant document is retrieved during search and precision is the probability of whether a retrieved document is relevant or not. However in both of these methods, some text region has been lost. In our method, none of the text region has lost and only the black shirt comes as white due to no change of entropy of that region. Iterative partitioning offers poor result in both the cases as it partitions the image into four sub block based on the histogram. The Table1 shows the average performance analysis between Otsu's method, iterative partitioning and our proposed method.

Table 1: Performance Analysis

Performance mea-surement Matrices	Otsu Binarization technique	Iterative Parti-tioning	Our proposed method
Recall	82.96	99.44	99.50
Precision	97.87	95.88	96.45
$F_measure$	89.82	97.62	98.89
PERR	14.17	3.69	3.91
MSE	9.32	2.39	2.10
SNR	31.12	36.08	35.55
PSNR	57.20	62.13	63.67

Performance analysis table shows that the probability of retrieval of relevant document is 99.5% and the precision is 96.45% which are quite good. On the

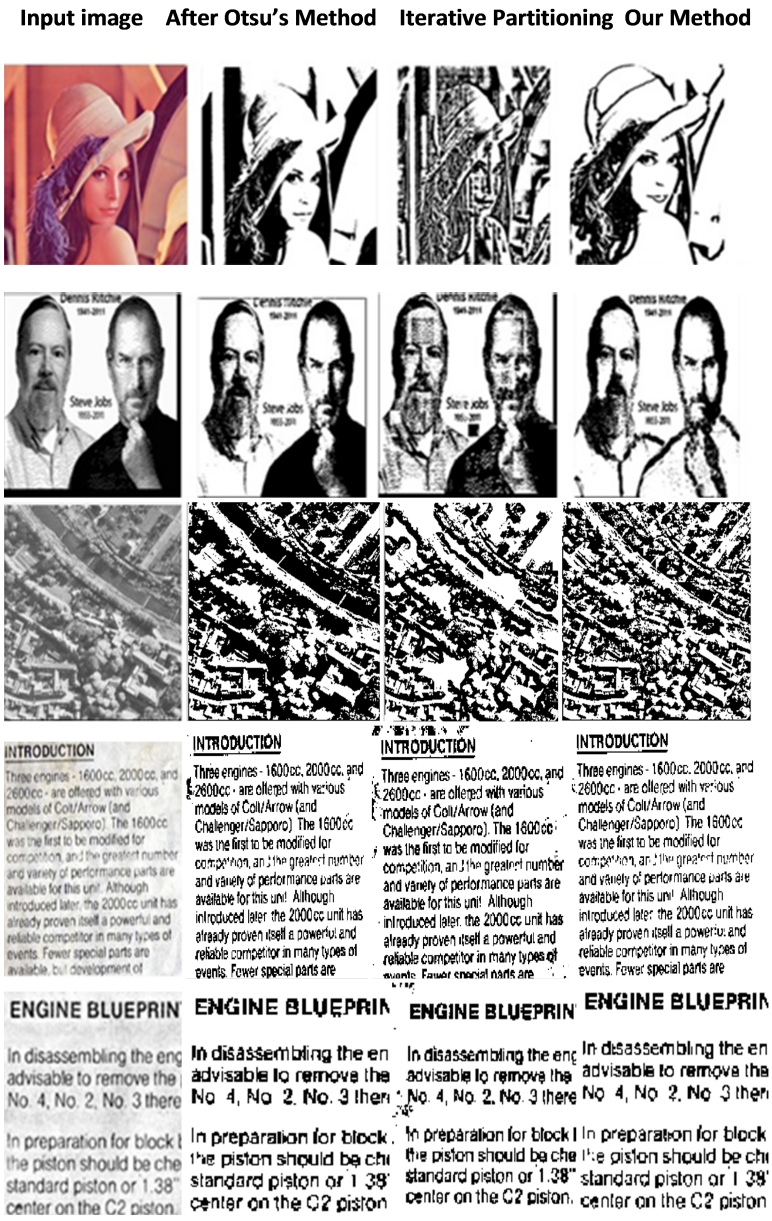


Fig. 3: Comparative Study

other hand misclassification rate is also less than Otsu's method and iterative partitioning. Our method performs much better than the other two methods in the presence of Gaussian and random noises. It makes obvious that the informa-

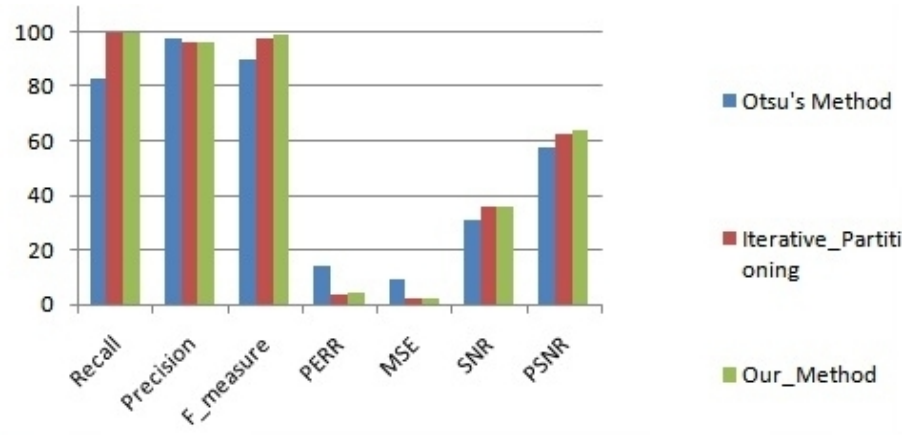


Fig. 4: Average performance graph on 7 matrices for the said methods

tion density based binarization technique is more noise resistive than other two binarization techniques.

5 Conclusion

In this research work, we have proposed a new method using clustering technique of a gray scale image. Here random noises, pepper and salt noises, Gaussian noises are removed. The proposed method has been tested on the benchmarked image data-base. This method easily separates the compound document and produce better result than Otsu's method, iterative partitioning. The proposed method also offers good result for the above said evaluation metrics. However, experimental observation finds limitation of the proposed method towards binarization of X-ray type of medical images. The work may be extended to address this aspect.

References

1. Thillou, C., Gosselin, B.: Segmentation-based binarization for color degraded images. In: Computer Vision and Graphics. Springer (2006) 808–813
2. Gonzalez, R.C., Woods, R.E.: Digital image processing. Pearson Education India (2009)
3. Namboodiri, A.M., et al.: Document structure and layout analysis, digital document processing, mar. 13, 2007
4. Dinan, R.F., Dubil, J.F., Malin, J.R., Rodite, R.R., Rohe, C.F., Rohrer, G.D.: Document image processing system (December 19 1989) US Patent 4,888,812.
5. Jaimes, A., Mintzer, F.C., Rao, A.R., Thompson, G.: Segmentation and automatic descreening of scanned documents. In: Electronic Imaging'99, International Society for Optics and Photonics (1998) 517–528

6. Ghosh, P., Bhattacharjee, D., Nasipuri, M.: Blood smear analyzer for white blood cell counting: A hybrid microscopic image analyzing technique. Applied Soft Computing **46** (2016) 629–638
7. Parker, J.R., Jennings, C., Salkauskas, A.G.: Thresholding using an illumination model. In: Document Analysis and Recognition, 1993., Proceedings of the Second International Conference on, IEEE (1993) 270–273
8. Chen, W.T., Wen, C.H., Yang, C.W.: A fast two-dimensional entropic thresholding algorithm. Pattern Recognition **27**(7) (1994) 885–893
9. Yanowitz, S.D., Bruckstein, A.M.: A new method for image segmentation. In: Pattern Recognition, 1988., 9th International Conference on, IEEE (1988) 270–275
10. Ghosh, P., Bhattacharjee, D., Nasipuri, M., Basu, D.K.: Medical aid for automatic detection of malaria. In: Computer Information Systems–Analysis and Technologies. Springer (2011) 170–178
11. Yang, J.D., Chen, Y.S., Hsu, W.H.: Adaptive thresholding algorithm and its hardware implementation. Pattern Recognition Letters **15**(2) (1994) 141–150
12. Shaikh, S.H., Maiti, A.K., Chaki, N.: A new image binarization method using iterative partitioning. Machine vision and applications **24**(2) (2013) 337–350
13. Otsu, N.: A threshold selection method from gray-level histograms. Automatica **11**(285-296) (1975) 23–27
14. California, S.: Usc-sipi image database, university of southern california http://sipi.usc.edu/database/.
15. Jain, A.K.: Fundamentals of digital image processing. Prentice-Hall, Inc. (1989)
16. Abutaleb, A.S.: Automatic thresholding of gray-level pictures using two-dimensional entropy. Computer vision, graphics, and image processing **47**(1) (1989) 22–32
17. Datta, S., Chaki, N.: Person identification technique using rgb based dental images. In: Computer Information Systems and Industrial Management. Springer (2015) 169–180
18. Han, Y., Shi, P.: An improved ant colony algorithm for fuzzy clustering in image segmentation. Neurocomputing **70**(4) (2007) 665–671
19. Chaki, N., Shaikh, S.H., Saeed, K.: Exploring image binarization techniques. Springer (2014)
20. Su, B., Lu, S., Tan, C.L.: Binarization of historical document images using the local maximum and minimum. In: Proceedings of the 9th IAPR International Workshop on Document Analysis Systems, ACM (2010) 159–166