

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/3710586>


Adaptive Document Binarization.

Conference Paper in Pattern Recognition · September 1997
DOI: 10.1109/ICDAR.1997.619831 · Source: IEEE Xplore

CITATIONS
833

READS
2,187


4 authors, including:



Jaakko Jari Sauvola
University of Oulu

85 PUBLICATIONS 3,530 CITATIONS


SEE PROFILE



Tapio Seppänen
University of Oulu

376 PUBLICATIONS 10,197 CITATIONS

SEE PROFILE



Matti Pietikäinen
University of Oulu

322 PUBLICATIONS 59,784 CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:

- Project

Brain Signal Derived Directed Connections (vs. Graph Theory)/Neuromarketing [View project](#)
- Project

Digital Health Revolution (DHR) 2 [View project](#)

Adaptive Document Binarization

Jaakko Sauvola, Tapio Seppänen, Sami Haapakoski and Matti Pietikäinen
Machine Vision and Media Processing Group
Infotech Oulu, University of Oulu
FIN-90570 Oulu, Finland
e-mail: jjs@ee.oulu.fi

Abstract

A new method is presented for adaptive document image binarization, where the page is considered as a collection of subcomponents such as text, background and picture. The problems caused by noise, illumination and many source type related degradations are addressed. The algorithm uses document characteristics to determine (surface) attributes, often used in document segmentation. Using characteristics analysis, two new algorithms are applied to determine a local threshold for each pixel. An algorithm based on soft decision control is used for thresholding background and picture regions. An approach utilizing local mean and variance of gray values is applied to textual regions. Tests were performed with images including different types of document components and degradations. The results show that the method adapts and performs well in each case.

Keywords: Adaptive binarization, soft decision, histogram, document segmentation, document analysis, document understanding.

1: Introduction

Most document analysis algorithms are built on taking advantage of the underlying binarized image data [1]. The use of a bi-level information decreases the computational load and enables the utilization of simplified analysis methods compared to 256 levels of grey-scale or color image information. Document image understanding methods require logical and semantic content preservation during thresholding. For example, a letter connectivity must be maintained for optical character recognition and textual compression [2]. This requirement narrows down the use of a global threshold in many cases.

Binarization has been a subject of an intense research interest during the last ten years. Most of the developed algorithms rely on statistical methods, not considering the special nature of document images. However, recent devel-

opments on document types, for example documents with mixed text and graphics, call for more specialized binarization techniques. Some document directed binarization algorithms have been developed. In [3] O’Gorman proposes a global approach calculated from a measure of local connectivity information. The thresholds are found at the intensity levels aiming to preserve the connectivity of regions. Liu et. al. [4] propose a method for document image binarization focused on noisy and complex background problems. They use grey-scale and run-length histogram analysis in a method called ‘object attribute thresholding’. It identifies a set of global thresholds using global techniques which is used for final threshold selection utilizing local features. Yang et.al.’s [5] thresholding algorithm uses a statistical measurement, called ‘largest static state difference’. The method aims to track changes in the statistical signal pattern, dividing the level changes to static or transient according to a grey-level variation. The threshold value is calculated according to static and transient properties separately at each pixel. Stroke connectivity preservation issues in textual images are examined by Chang et. al. in [6]. They propose an algorithm that uses two different components: the background noise elimination using grey-level histogram equalization and enhancement of grey-levels of characters in the neighborhood using an edge image composition technique. The ‘binary partitioning’ is made according to a smoothed and equalized histogram information calculated in five different steps. Pavlidis [7] presents a technique based on the observation that after blurring a bi-level image, the intensity of original pixels is related with the sign of the curvature of the pixels of the blurred image. This property is used to construct the threshold selection of partial histograms in locations where the curvature is significant.

Our approach starts with two observations. First, a document image usually contains many regions with differing structure and semantic content, for example picture, text, background and linedrawing. Therefore, specialized methods are needed to analyze the various types of regions. Second, the state and degree of a degradation can

vary significantly within a document image due to various sources of error, such as scanning, copying and poor source material (paper and print quality etc.).

We propose a new method that first performs a rapid classification of the local contents of a page to background, pictures and text. Two different approaches are then applied to define a threshold for each pixel: a soft decision method (SDM) for background and pictures, and a specialized text binarization method (TBM) for textual and linedrawing areas. The SDM includes noise filtering and signal tracking capabilities, while the TBM is used to separate text components from background in bad conditions, caused by uneven (il)lumination or noise. Finally, the outcome of these algorithms are combined.

In the Section 2, we describe the overall structure of our algorithm. Experimental results and comparisons are presented in Section 3, with the different degradation conditions such as varying illumination, noise and stains. Section 4 concludes the paper.

2: Binarization algorithm

The document image contains different surface (texture) types that can be divided into uniform, differentiating and transiently changing. The texture contained in pictures and background can usually be classified to uniform or differentiating categories, while the text, line drawings etc. have more transient properties by nature.

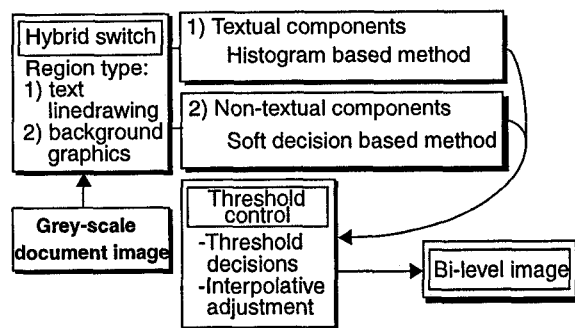


Fig. 1. Binarization algorithm.

Our approach is to analyze the local document image surface in order to decide on the binarization method needed (Fig. 1). During this decision, a 'hybrid switching' module selects one of two specialized binarization algorithms to be applied to the region. The goal of the binarization algorithms is to produce an optimal threshold value for each pixel. A fast option is to compute first a threshold for every n th pixel and then use interpolation for the rest of the pixels (Fig. 2).

The binarization method can also be set to bypass the hybrid switch phase. Then the user can choose which algorithm is selected for thresholding. All other modules function in the same way as in hybrid conditions.

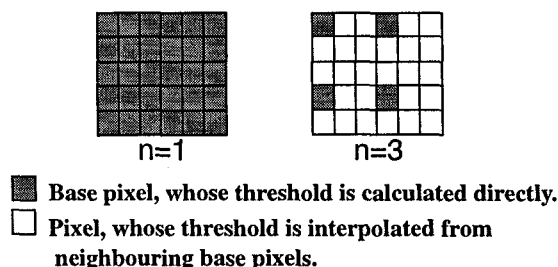


Fig. 2. Interpolation options.

The following subsection describes the region type and switching algorithms. The two different binarization algorithms are then discussed in detail. The final binarization is performed using the proposed threshold values. This process is depicted in the last subsection.

2.1: Region analysis and switching

Threshold computation is preceded by the selection of the proper binarization method based on an analysis of local image properties. First, the document image is tiled to equal sized rectangular windows of 10-20 pixels wide, corresponding to the resolution that linearly varies between >75dpi to <300dpi. Two simple features are then computed for each window; these results are used to select the method.

The first feature is simply the average grey value of a window. The second feature, 'transient difference', measures local changes in contrast (Eq. 4). The difference values are accumulated in each subwindow and then scaled between 0 and 1. Using the limits of 10, 15 and 30% of scaled values, the transient difference property is defined as 'uniform', 'near-uniform', 'differing' or 'transient'. This coarse division is made according to average homogeneity on the surface. According to these labels, a vote is given to corresponding binarization method that is to be used in a window. The labels 'uniform' and 'near-uniform' correspond to background and 'scene' pictures, and give votes to the SDM. The labels 'differing' and 'transient' give their votes to the TBM method.

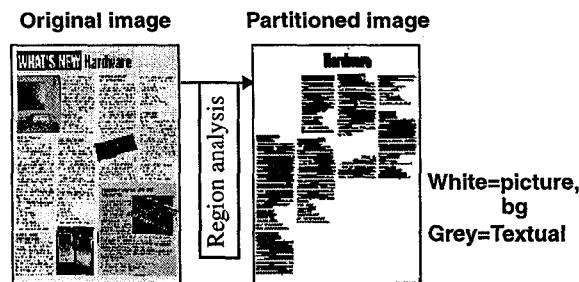


Fig. 3. Example of region partitioning for algorithm (SDM/TBM) selection.

Selection of a binarization algorithm is then performed as following example rules (1-2) show:

(1) If the average is high and a global histogram peak is in the same quarter of the histogram and transient difference is transient, then use SDM.

(2) If the average is medium and a global histogram peak is not in the same quarter of the histogram and transient difference is uniform, then use TBM.

An example result of image partitioning is shown in Fig. 3. The white regions are guided to the SDM algorithm, while the grey regions are binarized with the TBM algorithm.

2.2: Binarization of non-textual components

As in soft control applications, our algorithm first analyses the window surface by calculating descriptive characteristics. Then, the soft control algorithm is applied to every n th pixel (Fig. 2). The result is a local threshold based on local region characteristics.

To ensure local adaptivity of threshold selection, we use two different types of locally calculated features: 'weighted bound' and 'threshold difference'. The membership function issues, soft decision rules and defuzzification algorithm are presented in the following paragraphs.

Weighted bound calculation - Histogram based analysis schemes and features are often used in binarization methods. In document analysis the histogram is very useful for detecting and differentiating domains in physical and logical analysis. We use a new approach developed for local detection and weighting of bounds in grey-scale image texture. A new feature called weighted bound (W_b) is introduced and utilized in the soft control algorithm. The W_b is used for characterization of local pixel value profiles by tracking low, medium and high pixels in a small area. In a given surface area of $n \times n$ pixels, three different measures are calculated. The values are collected in a two dimensional table used to weight and simplify the three envelope curves in soft control membership functions. The measures are minimum, medium and maximum averages given in equations (1)-(3).

Minimum average, A_{min}

$$A_{min} = \sum_{k=0}^{100/n} \frac{\min_{100/n}(P(i, j))}{100/k} \quad (1)$$

Medium average, A_{med}

$$A_{med} = \sum_{k=0}^{100/n} \frac{\text{med}_{100/n}(P(i, j))}{100/k} \quad (2)$$

Maximum average, A_{max}

$$A_{max} = \sum_{k=0}^{100/n} \frac{\max_{100/n}(P(i, j))}{100/k} \quad (3)$$

These values are stored in an $n \times n \times 3$ table, called a weighted average table (WAT). Using Eqs. (1)-(3), three different histograms are formed where the values are added to their respective bin values (value = bin index). These histograms are then separately partitioned to ten horizontal and three vertical sections, where the number of peaks from histograms are calculated to each section according to sectioning limits.

The horizontal borders are set between bins 0 and 255 with a formula $\text{int}((256/10) * m)$, where $m=1, 2, \dots, 9$. The number of borders was set to ten. Also a smaller number could be selected, but the penalty is that the original histogram is aliased more. Ten borders equals 25 bins of grey-scale. The two vertical borders are set between 0 and maximum, representing the number of votes calculated for each horizontal bin so that the limits are set to 80% of maximum number of votes and to 40% of the maximum number of votes, respectively. These limits are set according to the tests performed with a large set of images. The higher limit is relatively insensitive to $\pm 10\%$ change. Lowering the lower limit brings more votes to medium peak calculation, thus enhancing the envelope curve in bins where a medium peak appears.

After the peaks are calculated in a 3×10 table, the weighting is performed (Fig. 4). The result is a W_b envelope curve that is used in the soft decision process. The three W_b curves, calculated from A_{min} , A_{med} and A_{max} are used as membership functions.

Transient difference calculation - The transient difference is aimed at extracting the average amount of variations occurring between the neighbouring pixels (contrast difference) in an $n \times n$ area, i.e. to follow local surface changes. The differences between adjacent pixels are accumulated. The transient difference (TD) of the horizontal and vertical adjacent pixel values is calculated and accumulated. The gained value is then scaled between 0-1 (4). L represents the number of grey-levels in the image.

$$TD = \frac{\left(\sum_{i=1}^n \sum_{j=1}^n |2P(i, j) - [P(i-1, j) + P(i, j-1)]| \right)}{(Ln)^2} \quad (4)$$

The TD value is used in soft decision making to expose uniform, differential and transient area types when calculating the control value for threshold selection.

Membership function generation - Two different membership functions are used according to the extracted feature values for a given pixel: weighted bound (W_b) and

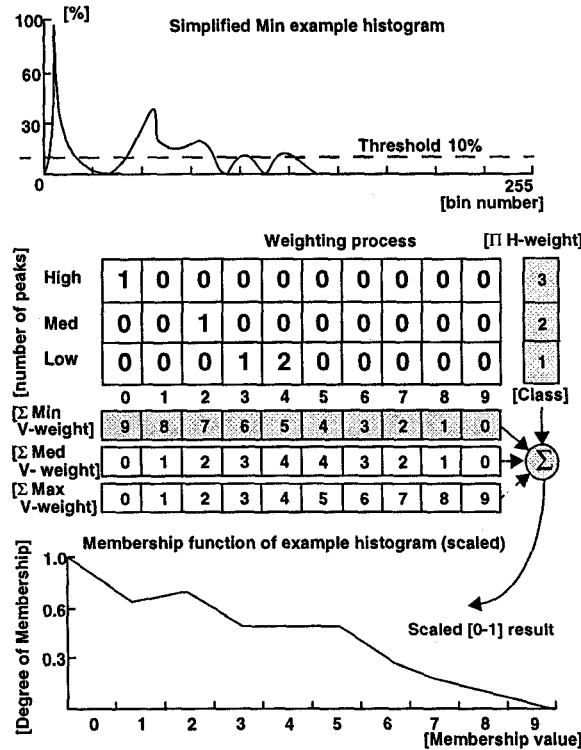


Fig. 4. An example of W_b membership function calculation using A_{min} histogram.

transient difference (TD_m). The first one is calculated dynamically from the image. The transient difference uses predefined membership functions. Fig. 5 depicts these functions using the ideal functions as W_b and the actual membership functions for TD_m .

Soft decision rules and defuzzification - In the soft decision process, we use nine different rules derived from the feature analysis and membership management. For W_b these are (LOW, MIDDLE, HIGH), denoting the local histogram properties. For TD_m we use (UNIFORM, DIFFERING, TRANSIENT), describing the local region property. The rule set is shown in Fig. 6. As in soft control problems, the rules are expressed with clauses, for example:

If W_b is $\langle P(i,j) \rangle$ and TD_m is $\langle TD(i,j) \rangle$
then $T_c(i,j) = \langle 0, 255 \rangle$

The current rule set is designed for pictorial and background type image regions. Using this set the noise and most illumination defects can be adaptively corrected in the processed areas.

For defuzzification we use Mamdani's method [8]. The result of the defuzzification is a unique threshold value for each pixel n .

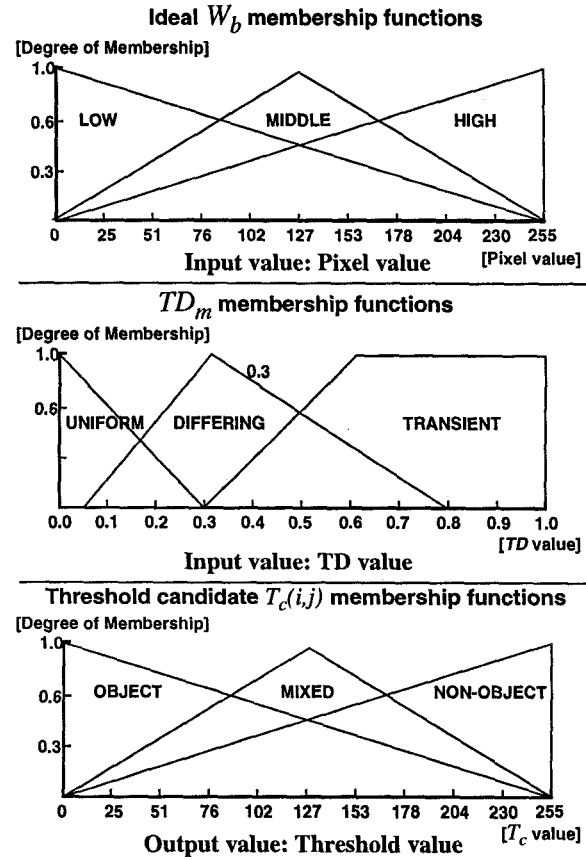


Fig. 5. Input and output membership functions: W_b (ideal), TD_m and T_c .

| $TD_m \backslash W_b$ | LOW | MIDDLE | HIGH |
|-----------------------|--------|--------|------------|
| UNIFORM | Object | MIXED | Non-object |
| DIFFERING | MIXED | MIXED | MIXED |
| TRANSIENT | Object | MIXED | Non-object |

Fig. 6. Example of soft decision rules for threshold candidate $T_c(i,j)$.

2.3: Binarization of textual components

For text binarization we use a modified version of Niblack's algorithm [9]. The idea of Niblack's method is to vary the threshold over the image, based on the local mean, m , and local standard deviation, s , computed in a small neighborhood of each pixel. A threshold for each pixel is computed from $T = m + k * s$, where k is a user defined parameter and gets negative values. This method does not work well for cases in which the background con-

tains light texture as gray values of these unwanted details easily exceed threshold values. This results in costly post-processing as demonstrated in [10].

In our modification, a threshold is computed with $T=m*[1+k*(s/R-1)]$, where R is the dynamic range of standard deviation, and parameter k gets positive values. Use of a local mean to multiply both terms. This has the effect of amplifying the contribution of standard deviation in an adaptive manner. Consider, for example, dark text on light dirty-looking background (e.g., stains in a bad copy). The m -coefficient decreases threshold value in background areas. This efficiently removes the effect of stains in a thresholded image. In our experiments, we used $R=128$ with 8-bit gray level images and $k=0.5$ to obtain good results. The algorithm is not too sensitive to the value of parameter k . An example of the threshold selection using the 'Stain' image (Fig. 9) is depicted in Fig. 7.

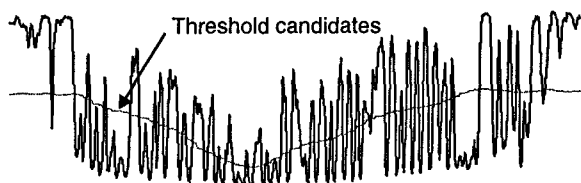


Fig. 7. Example of threshold candidate selection of 'Stain' image.

2.4: Interpolative threshold selection

After the surface type guided thresholding, the final thresholds are calculated for background, textual, graphics and line drawing regions. A fast option is to compute first a threshold for every n th pixel and then using interpolation for the rest of the pixels.

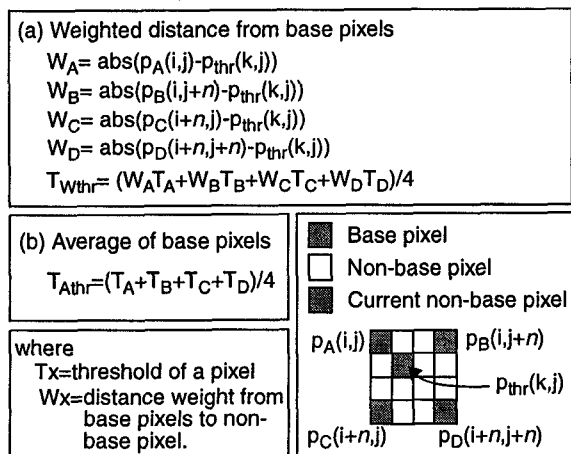


Fig. 8. Two interpolation choices for threshold selection of non-base pixels.

The control algorithm has two modes depending on the value of n (Section 2). If $n=1$, the threshold values

gained from SDM and TBM algorithms are combined directly. If $n>1$, threshold values for non-base pixels are calculated using the surrounding threshold values.

We have two options to calculate the non-base pixel thresholds: bilinear interpolation and simple averaging. In the bilinear interpolation method, the threshold value for a non-base pixel is gained by computing the surrounding base pixels distance to the current one, and using these values as weights, Fig. 8 (a). This approach gives a more precise, weighted threshold value for each pixel. In the simple averaging method, the average of the surrounding four n pixel threshold candidate values is calculated and used as a final threshold for each non-base pixel between the selected base pixels, Fig. 8 (b). This approach is used to lower the computational load and is suitable for most images, especially for those with random noise and n larger than five pixels.

3: Results

We tested our algorithm with several different cases of document degradation caused by various sources.



Fig. 9. Examples of binarization results for textual and background images.

The main emphasis was on textual and background binarization, i.e. the regions most affecting document analysis algorithms and optical character recognition. The evaluation was performed by running the same images with the Niblack's [9] basic algorithm which was recommended by Trier and Jain [10]. They compared eleven

locally adaptive binarization methods. Niblack's method ranked the best with the postprocessing step used in Yanowitz and Bruckstein algorithm [11].

First, we present visual results of the degradation situations taking place caused by errors in scanning, copying or poor source quality, Fig. 9. The sample of original images are on the left, the comparison results with Niblack's algorithm in the middle and the results with our algorithm on the right. The results on pictures with corresponding comparisons are presented in Fig. 10.

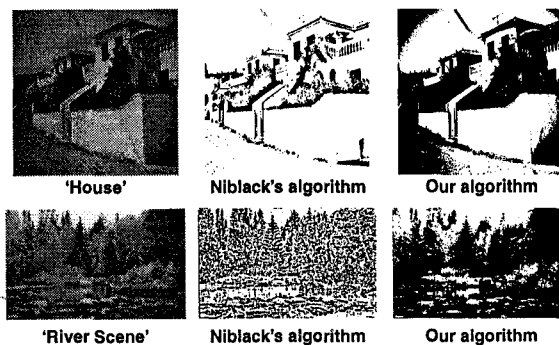


Fig. 10. Examples of binarization results for scene images.

The results were visually evaluated by a group of test viewers. They gave their grade independently, using the scale (Good, Average, Poor or Bad). The overall results for a number of test images are shown in Fig. 11.

| Image name | Type | Niblack | Our method |
|-------------|---------|---------|------------|
| Stain | Textual | Average | Good |
| Design2 | Textual | Poor | Average |
| Text1 | Textual | Good | Good |
| Frame | Textual | Average | Good |
| House | Picture | Average | Average |
| River scene | Picture | Poor | Average |

Scale:

Good - Best result taking the original quality into consideration.
Average - Result is average and is regarded sufficient for OCR.
Poor - Result is poor and is regarded not to be sufficient for OCR.
Bad - Result is bad, no clear visual recognition of objects is possible.

Fig. 11. Overall comparison results performed on a set of test images.

4: Conclusion

We have proposed a new method for adaptive binarization of document images. The main features of our approach include locally adaptive threshold selection, information content preserving analysis, and seamless applicability to various types of documents. The grey-level document image is first analyzed to determine the surface properties. Then, according to analysis information the recognized surface properties are treated with two different binarization methods. For background and 'scene' type areas an algorithm utilizing the soft decision method in a new con-

text is performed. For textual and badly illuminated regions a histogram method is applied.

Our algorithm was tested with several severe cases of degradations, natural and synthetic. The test results show that the algorithm adapts well to even severe degradations, enhancing for example the OCR rate in badly degraded images from non-recognizable to correct or near-correct results. Furthermore the algorithm's custom parametrization is minimized with the use of soft decision methods and special analysis procedures.

References

- [1] Sauvola J. and Pietikäinen M. (1995) Page Segmentation and Classification Using Fast Feature Extraction and Connectivity Analysis. In: International Conference on Document Analysis and Recognition, ICDAR '95, Montreal, Canada, pp.1127-1131.
- [2] Sauvola J. and Pietikäinen M. (1996) A Document Management Interface Utilizing Page Decomposition and Content-Based Compression. In: International Conference for Pattern Recognition, ICPR '96, Vienna, Austria, pp. 752-757.
- [3] O'Gorman L. (1994) Binarization and Multithresholding of Document Images Using Connectivity. In: CVGIP: Graphical Models and Image Processing, Vol. 56, No. 6, November, pp. 496-506.
- [4] Liu Y., Fenrich R. and Srihari S.N. (1993) An Object Attribute Thresholding Algorithm for Document Image Binarization. In: International Conference on Document Analysis and Recognition, ICDAR '93, Japan, pp. 278-281.
- [5] Yang J., Chen Y. and Hsu W. (1994) Adaptive thresholding algorithm and its hardware implementation. In: Pattern Recognition Letters, No. 15, pp. 141-150.
- [6] Chang M., Kang S., Rho W., Kim H. and Kim D. (1995) Improved Binarization Algorithm for Document Image by Histogram and Edge Detection. In: International Conference for Document Analysis and Recognition 1995, Montreal, Canada, pp. 636-643.
- [7] Pavlidis T. (1993) Threshold Selection Using Second Derivatives of the Gray Scale Image. In: International Conference on Document Analysis and Recognition, ICDAR '93, Japan, pp. 274-277.
- [8] Welstead S.T. (1995) Neural Network and Fuzzy Logic Applications in C/C++. John Wiley & Sons, Inc., 494 pages.
- [9] Niblack W. (1986) An introduction to image processing. Englewood Cliffs, N.J.: Prentice Hall, pp. 115-116.
- [10] Trier O.D. and Jain A.K. (1995) Goal-directed evaluation of binarization methods, IEEE Transactions on pattern analysis and machine intelligence, Vol. 17, No. 12, December 1995.
- [11] Yanowitz S.D. and Bruckstein A.M. (1989) A new method for image segmentation. In: Computer Vision, Graphics and Image Processing, Vol. 46, no.1, pp. 82-95.