Serdica J. Computing **6** (2012), 111–120

Serdica
Journal of Computing

Bulgarian Academy of Sciences
Institute of Mathematics and Informatics

ADAPTIVE DOCUMENT IMAGE BINARIZATION WITH APPLICATION IN PROCESSING ASTRONOMICAL LOGBOOKS*

Lasko Laskov

Abstract. Recently, the digitalization of the astronomical scientific heritage has been considered an important task that can facilitate much researches in astronomy. The creation of digital libraries and databases of astronomical photographic plates brings up the problem of digitalization astronomical logbooks , since the data contained in them is crucial for the usage of the plates . An optical character recognition (OCR) system for the handwritten numerical data is needed in order to speed up the process of database creation and extension.

In this paper document image binarization is considered since it is a critical stage for the subsequent steps in an OCR software system. A specific method is proposed which outmatches the state-of-the-art techniques in the case of the images of interest.

ACM Computing Classification System (1998): I.7, I.7.5.

Key words: astronomical logbooks, document image processing, adaptive image binarization.

*This work has been partially supported by Grant No. DO02-275/2008, Bulgarian NSF, Ministry of Education and Science.

1. Introduction. During the last few decades the digitalization of cultural and scientific heritage has become an important and challenging task with application in many scientific fields. The development of digital databases and libraries of various content has improved the accessibility, distribution and usage of data sources, previously accessible only by few specialists. Such digital libraries are also being developed in the field of astronomy, such as the database of scanned astronomical photographic plates [9], which are a priceless source of information for much research.

The digitalization of astronomical photographic plates involves two basic steps:

- digitalization of the plate itself;
- digitalization of the accompanying meta-data, usually contained in a handwritten logbook.

It is important to note that the photographic plate itself cannot be used in a digital database efficiently without the meta-data contained in the astronomical logbooks. Since the digitalization of a plate to produce a digital image involves mainly scanning, the step which slows down the process is the digitalization of astronomical logbooks, which currently is performed manually. This implies that if a software system for digitalization of automatical logbook content is developed, the whole process of astronomical plate database development and enrichment is going to be vastly accelerated.

The meta-data contained in the logbooks includes the number of the plate, date and time of image acquisition, celestial coordinates of the centre of the plate, etc., where all these elements are represented by handwritten digits (see Fig. 1). Hence, the software system that can speed up the process of logbook digitalization is OCR (optical character recognition) of the handwritten digits contained in the logbooks.

Document image binarization is an important preprocessing step to document image analysis and recognition. Also, it can be considered as a critical stage in OCR software systems since the result of the subsequent steps is highly dependent on its effectiveness. This is the reason why document image binarization has been a subject of extensive research during the last decades [8]. Nevertheless it can be considered as a solved problem in the case of document images with good quality (uniform illumination, absence of noise due to both data source degradation and image acquisition, etc.), it is still an open question in the case of images of degraded, old, or even historical documents.

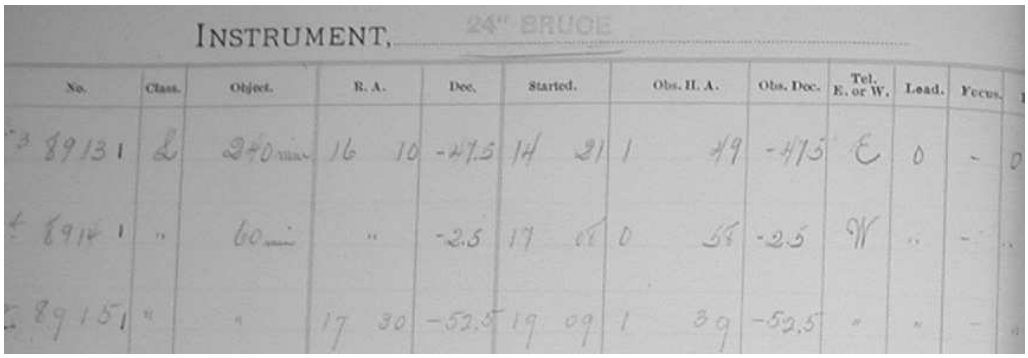


Fig. 1. A fragment of an astronomical logbook containing photographic plate meta-data. The image is a part of the Harvard University collection, published on their web-page `http://tdc-www.cfa.harvard.edu/plates/a/logs/`

Given a grey-scale image with pixel values usually in the integer range $[0, 255]$ and assuming that the image contains two types of pixels (background and object), binarization can be viewed as the process of assigning of each pixel one of the two labels, 0 for background and 1 for object pixel. Basically there are two broad types of binarization algorithms: global [6], [4] and locally adaptive [1], [2], [3]. Global methods are usually based on the global characteristics of the entire input image and they apply a global threshold which divides the pixels in to object and background pixels according to their value. These algorithms are relatively fast and can produce satisfactory results, provided that the input image is of good quality. On the other hand, if this condition is not satisfied, usually the global approaches fail to separate the object from the background pixels (see Fig. 4(b) and Fig. 5(c)). In this case locally adaptive methods tend to produce better results and to overcome the difficulties which follow from the low quality of the input images.

The locally adaptive methods for image binarization are based on the local characteristics of the neighborhood of each pixel or image parts. The size of the filter structuring element can be predefined according to some criterion as in the case of [7], where the filter size depends on the width of the objects of interest. Another approach is to determine the size of the sub-images being processed dynamically during the execution of the algorithm, for example as in the case of the variable thresholding using locally Otsu's method, as described in [3], Chapter 10, Section: Image partitioning. The locally adaptive methods have proved to be more efficient than the global ones if the input images are of bad quality, as is in the case of the examined astronomical logbooks.

In this paper a specific locally adaptive method for document image binarization is proposed, which outmatches the state-of-the-art techniques in the case of processing astronomical logbooks. The method is based on horizontal and vertical one-dimensional filters which are combined to extract the two-dimensional characteristics of the local neighborhood of each pixel. This paper is organized as follows: in Section 2 two locally adaptive methods which are the basis of the proposed algorithm are described; Section 3 contains a detailed description of the proposed algorithm; Section 4 is dedicated to the experimental results; and finally, in Section 6 the conclusions and some directions for future work are given.

2. Related work. In this section two methods are discussed which are used as the basis of the algorithm described in Section 3. The first one is proposed by Gonzalez in [3], Chapter 10, Section: Using moving averages. The method is based on one-dimensional mean filter which is moving along the scan lines of the image in a zigzag pattern (see Fig. 2: (a) initial state of the filtering process, where the values outside the image boundaries are considered zeroes; (b) the filter moves rightwards along the image scan line; (c) when the right boundary of the image is reached, the filter turns on the next scan line starting from its last pixel; (d) the filter continues to move leftwards; (e) when the filter reaches the left boundary, it turns on the next scan line starting from its first pixel.

Since the image is scanned only once this method is fast and it is particularly efficient for document image processing.

On each step of the filtering process, the method calculates the local mean μ of the intensity values within the range of the filter and calculates a threshold t for the pixel which lies under the first element of the filter. Let us denote the input image with I and the intensity level at step k of the image scan with $I(k)$. The local mean at step $(k+1)$ is given by:

$$\begin{aligned} \mu(k+1) &= \frac{1}{n} \sum_{i=k+2-n}^{k+1} I(i) \\ &= \mu(k) + \frac{1}{n} \left(I(k+1) - I(k-n) \right) \end{aligned} \tag{1}$$

where n is the filter window size which depends on the average stroke width. The values outside the image boundaries are considered zeroes. Then for each pixel a local threshold is calculated using:

$$t(k) = s\mu(k), \tag{2}$$

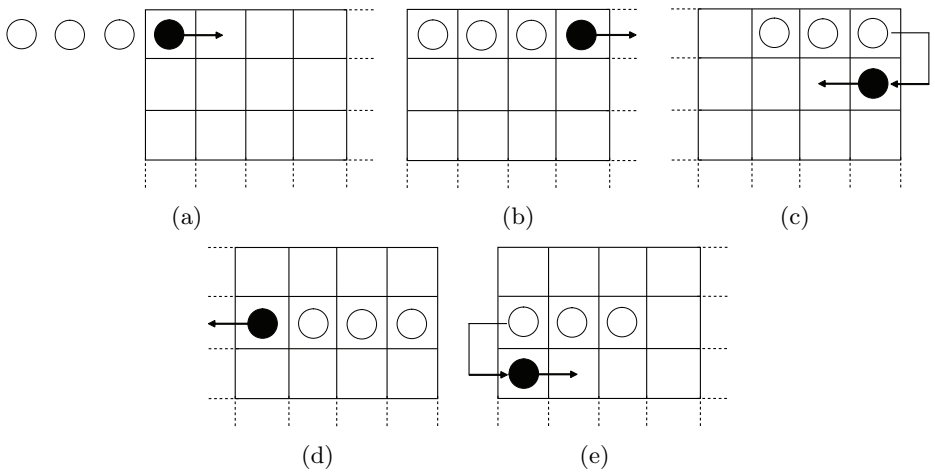


Fig. 2. The pixels of the image as squares and the structuring element of the one-dimensional filter as circles, where the blackened circle is the element of the filter, which lies on the pixel currently processed

where s is a scaling coefficient, $s \in (0, 1]$.

The second method discussed in this section is the one proposed by Sauvola and Pietikäinen in [7]. In this approach not only the local mean μ is considered, but also the standard deviation σ and its dynamic range R:

$$t(x, y) = \mu(x, y) \left[1 + s \left(\frac{\sigma(x, y)}{R} - 1 \right) \right]. \tag{3}$$

In (3) the local mean $\mu(x, y)$ and local standard deviation $\sigma(x, y)$ are accumulated in a two-dimensional neighborhood of the pixel with coordinates (x, y). The size of the neighborhood depends on the average stroke width. The dynamic range of the standard deviation is fixed to the constant $R = 128$ and $s \in (0, 1]$ is a scaling coefficient. This method has proved to be efficient in locally adaptive binarization of textual image areas which have close grey-scale values of the background and object pixels.

3. The proposed algorithm. The proposed approach combines the effectiveness of the implementation of (1) and robustness of (3) in the case of close grey-scale values of background and object pixels. The algorithm is composed by two binarization scans of the image: horizontal and vertical scan. It incorporates

a filtering step which removes the line-wise noise caused by the fact that the structuring element is one-dimensional. Finally the results of the horizontal and vertical scans are combined using a logical disjunction operation to produce the final binary image.

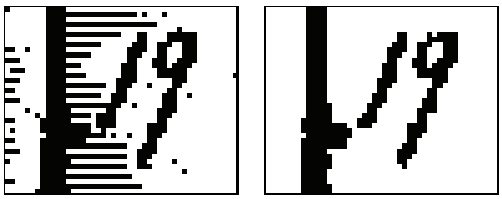
The *horizontal scan* is performed as shown in Fig. 2. On each step of the image scan the local mean is calculated using (1). Also, the dispersion is calculated by:

(4)

$$\begin{aligned}\sigma^2(k+1) &= \frac{1}{n}\sum_{i=k+-n}^{k+1}(I(i)-\mu(i))^2 \\ &= \sigma^2(k)+\frac{1}{n}\left[(I(k+1)-\mu(k+1))^2-(I(k-n)-\mu(k-n))^2\right]\end{aligned}$$

and the standard deviation σ is calculated from it. For each pixel, the local threshold is calculated using (3), but in this case the local statistics μ and σ are accumulated in one dimension of the image, depending on the filter window size n. The result of this stage is a binary image B_h.

The *vertical scan* is actually the horizontal scan of the rotated input image by 90°. After the horizontal scan of the rotated image is performed the output is rotated back to its initial position and the result is a binary image B_v.



(a) (b)

Fig. 3. The same fragment of a logbook image (a) without line-wise nose removal; (b) with line-wise noise filter applied to both B_h and B_v

After performing the horizontal and vertical scans, a simple procedure is applied to both binary images which removes the line-wise noise. This procedure simply checks each black pixel in an image, and if both its upper and lower neighbors are white, the pixel is transformed to white. On Fig. 3 an example is given with the same fragment of a logbook image, where Fig. 3(a) is produced without filtering and Fig. 3(b) is the result of the algorithm with line-noise filtering incorporated.

The last step of the proposed method is the combination of the two binary

images using the logical disjunction operation for each pixel:

$$B(x,y) = B_h(x,y) \vee B_v(x,y), \quad x = 0,1,\ldots,M-1, y = 0,1,\ldots,N-1, \tag{5}$$

where M is the number of rows and N is the number of columns of the input image I.

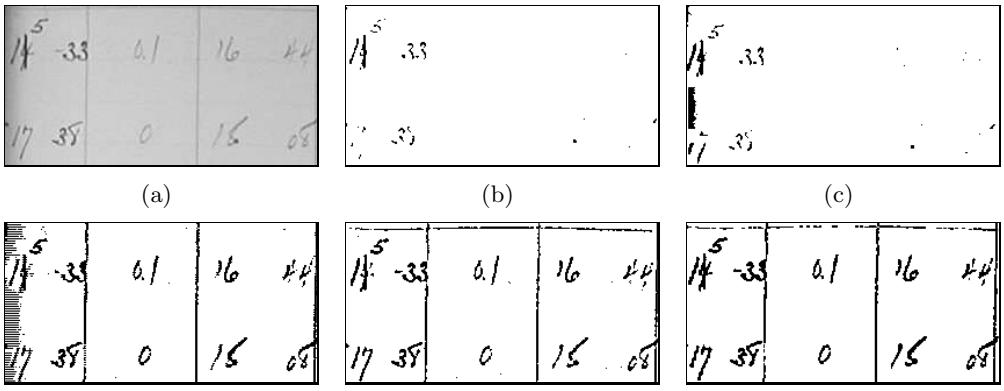


Fig. 4. A fragment of a logbook (a) processed with: (b) Otsu's global method; (c) Otsu's local method based on image partitioning; (d) moving averages, $n = 20$, $s = 0.95$; (e) Sauvola's method, 9×9 structuring element, $R = 128$, $s = 0.05$; (f) the proposed algorithm, $n = 20$, $R = 128$, $s = 0.05$

4. Experimental results. Two types of experiments were conducted in order to evaluate the effectiveness of the proposed method: (i) on a broad set of astronomical logbooks and (ii) on synthetic data which allows to calculate the recall, precision and F1 performance evaluation measures defined in [5]. Also, a comparison with the following state-of-the-art binarization techniques was made: Otsu's global method [6], Otsu's locally adaptive method based on image partitioning, moving averages [3] and Sauvola's method [7].

On Fig. 4, a fragment of a logbook is given, processed by the four state-of-the art methods and by the proposed algorithm. The evaluation of the results in this case can be strictly subjective since no ground true data is available. It is obvious that the global Otsu's method in Fig. 4(b) and its locally adaptive version in Fig. 4(c) does not produce good results with the images of interest. Much better results are produced with the moving averages' Fig. 4(d), even though

line-wise noise is present. Sauvola's method given in Fig. 4(e) produces better results than all the standard algorithms but it produces slightly more noise pixels than the proposed method in Fig. 4(f), which also gives better character stroke connectivity. It is important to note that the computational complexity of the proposed algorithm is much less than the implementation of the Sauvola's method since it is based on a combination of one-dimensional filters on the input image.

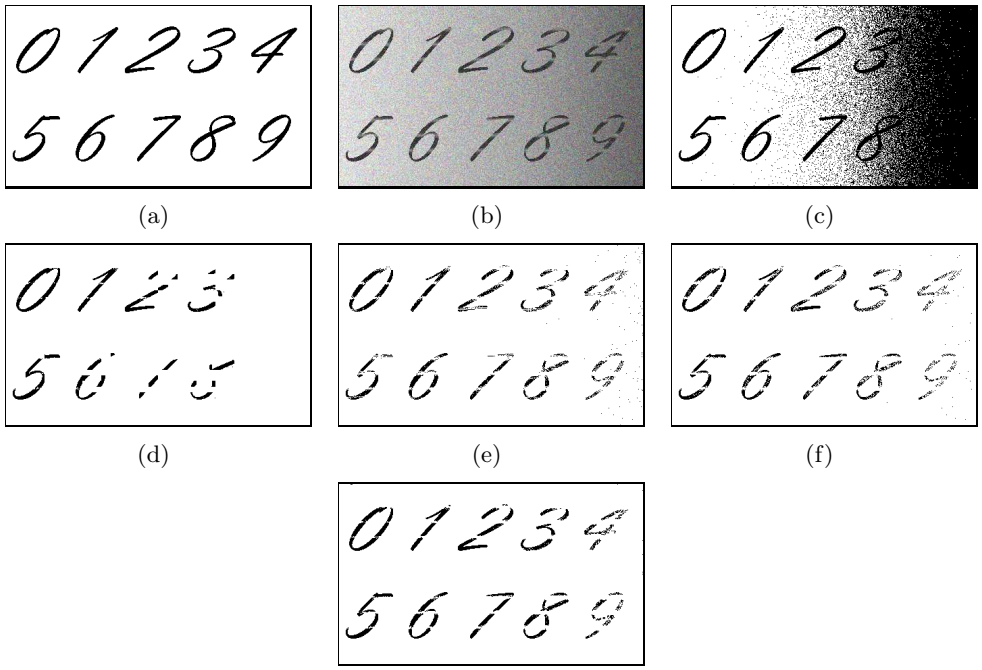


Fig. 5. A synthetic image (a) degraded with gradient and Gaussian noise image (b) processed with: (c) Otsu's global method; (d) Otsu's local method based on image partitioning; (e) moving averages, $n = 20$, $s = 0.6$; (f) Sauvola's method, 9×9 structuring element, $R = 128$, $s = 0.5$; (g) the proposed algorithm, $n = 20$, $R = 128$, $s = 0.5$

In Fig. 5 the experiment with synthetic data is given. The ground true image Fig. 5(a) is artificially degraded, Fig. 5(b), with gradient noise to simulate non uniform illumination and also Gaussian noise is added in order to simulate the low quality of the astronomical logbooks images.

Fig. 5(c) shows the result of Otsu's global method, which produces precision 0.181, recall 0.989 and F1 measure 0.307. The performance evaluation

measures show that many of the background pixels are recognized as object pixels by this method.

In Fig. 5(d) the result of the locally adaptive application of Otsu's method is given, which produces precision 0.998, recall 0.537 and F1 measure 0.699. The result is much better then the global Otsu's method, but much of the object pixels are recognized as background pixels.

Fig. 5(e) contains the result of the moving averages algorithm. It produces precision 0.985, recall 0.610 and F1 measure 0.754. In this experiment this is the standard method which produces results, even better then Sauvola's algorithm in Fig. 5(f) which has precision 0.818, recall 0.587 and F1 measure 0.683.

In Fig. 5(g) the result produced by the proposed method is given. In this case the performance evaluation measures are: precision 0.998, recall 0.735 and F1 measure 0.847. They prove that the proposed algorithm gives the best results, compared with the sate-of-the-art techniques above.

6. Conclusion In this paper a new method for adaptive document image binarization is proposed. The method is based on the incorporation of the effectiveness of the floating mean algorithm and robustness of Sauvola's method. The method is compared with four state-of-the art techniques and outmatches them in both the experiments with astronomical logbooks images and the experiment with synthetic data. The experiment with synthetic data allows formal performance evaluation using recall, precision and F1 measures, and the results prove the superiority of the proposed method. Since the approach is based on a combination of one-dimensional filters, it is also computationally effective, which is an important feature in document image processing.

Currently the described approach is part of an image processing and recognition system and is constantly in use in the experiments with astronomical logbooks and other non-standard documents. As future work, the method will be involved as a preprocessing step in an OCR software especially designed for handwritten digit recognition in astronomical logbooks images which is expected to increase the performance of the process of digitalization of astronomical plates .

REFERENCES

[1] Gatos B., I. Pratikakis, S. Perantonis. Adaptive degraded document image binarization. *Pattern Recognition*, **39** (2006), No 3, 317–327.

[2] Gatos B., I. Pratikakis, S. Perantonis. Improved document image binarization by using a combination of multiple binarization techniques and adapted edge information. In:Proceedings of the 19th International Conference on Pattern Recognition ICPR, Tampa, Florida, USA, 2008, IEEE, 1–4.

[3] Gonzalez R., R. Woods. Digital Image Processing (3rd Edition). Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 2007.

[4] Kittler J., J. Illingworth. Minimum error thresholding. *Pattern Recognition*, **19** (1986), No 1, 41–47.

[5] Olson D. L., D. Delen. Advanced Data Mining Techniques. Springer, 2008.

[6] Otsu N. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man and Cybernetics*, **9** (1979), No 1, 62–66.

[7] Sauvola J., M. Pietikainen. Adaptive document image binarization. *Pattern Recognition*, **33** (2000), No 2, 225–236.

[8] Sezgin M., B. Sankur. Survey over image thresholding techniques and quantitative performance evaluation. *Journal of Electronic Imaging*, **13** (2004), No 1, 146–165.

[9] Tsvetkov M. Making astronomical photographic data available: the european perspective. In: Preserving Astronomy's Photographic Legacy: Current State and the Future of North American Astronomical Plates, Vol. **410**, ASP Conference Series, 2009, 15–29.

Lasko Laskov
Department of Informatics
New Bulgarian University
21, Montevideo Str.
1618 Sofia, Bulgaria
e-mail: `llaskov@nbu.bg`

Received October 31, 2011
Final Accepted March 15, 2012