

Improving Binarization Methods for Historical Handwritten Documents

David Asatryan
Institute for Informatics and
Automation Problems of NAS RA
Russian-Armenian University,
Yerevan, Armenia
e-mail: dasat@iip.sci.am

Mariam Haroutunian
Institute for Informatics and
Automation Problems of NAS RA
Yerevan, Armenia
e-mail: armar@sci.am

Grigor Sazhumyan
Institute for Informatics and
Automation Problems of NAS RA
Yerevan, Armenia
e-mail: grigorsazhumyan@gmail.com

Alexander Kupriyanov
Samara National Research University,
IPSI RAS
Samara, Russia
e-mail: akupr@ssau.ru

Rustam Paringer
Samara National Research University,
IPSI RAS
Samara, Russia
e-mail: rusparinger@gmail.com

Dmitriy Kirsh
Samara National Research University,
IPSI RAS
Samara, Russia
e-mail: limitk@mail.ru

Abstract— Binarization of historical documents is a rather complex task, which is intensively dealt with by researchers all over the world. A large number of approaches, procedures, and binarization algorithms have been proposed, but methods that work equally well in all cases have not yet been proposed. Various criteria for evaluating the quality of binarization result are proposed in the literature. In the case of the binarization of ancient handwritten texts, the degree of readability of the text by a visual method or using technical means is taken as a criterion for the quality of the binarization algorithm. One of the approaches to improve the quality of the binarization result proposed in the literature is the preliminary processing of the original image, using filtering methods, morphological analysis, spectral analysis, etc., and the selection of segments with certain sizes. The proposed procedure allows selecting objects in the image with certain sizes, in particular, artifacts existing in the binarized image. In this work, the possibility of improving the quality of a binary image obtained by Niblack's method by applying the above-mentioned procedure is investigated experimentally. For comparison, the results of Otsu's and Sauvola's methods are used.

Keywords—Handwritten documents, binarization, segmentation, artefacts, Niblack's method, Sauvola's method.

I. INTRODUCTION

All over the world, there is a written heritage of an enormous volume of people and nationalities. Ancient historical handwritten and printed documents are of the greatest value, they contain useful, sometimes unique information about the history of peoples.

Unfortunately, many documents, especially ancient ones, are difficult to read due to the poor quality of the source material, degraded by external distortions, such as fading of paper and other media, smearing and staining of ink, uneven color tone, torn or wrinkled paper, etc.

The most common and effective way to improve the quality of written documents is to photograph or scan the

document, obtain a suitable digital image of it, and process the resulting image on a computer. In this case, as a preliminary method of information processing, a binarization algorithm is used that converts the image into a two-color one, in which the background consists of white pixels and the foreground consists of black ones.

An important characteristic of all applied binarization algorithms is the quality of the result obtained, which ensures a sufficiently complete extraction of the necessary information from the image. At the same time, the concept of binary image quality in each situation is determined in accordance with the statement of the problem, purpose, and method of extracting the necessary information [1, 2]. Therefore, various formal criteria for assessing the quality of binarization are proposed in the literature, with the help of which a meaningful and comparative analysis of the methods used is carried out. However, in the case of binarization of distorted images of handwritten texts, many researchers take the degree of readability of the text by a visual method or with the help of technical means as the main quality criterion.

There are two types of binarization algorithms in the literature: global (or thresholding) and local (or adaptive). Global binarization involves applying a fixed threshold to the entire image. The well-known method of global binarization is the Otsu method [3]. The efficiency of the global binarization method was quite thoroughly studied in [4], in which 40 binarization algorithms were analyzed and a comparative analysis was carried out. In [5], global binarization methods based on the optimization of various quality criteria for the results obtained were studied. Two types of quality criteria are considered based on the properties of the histogram of the tested image and evaluating the similarity of the latter with the binarized image.

It is known that the global binarization method works well with images the pixel intensity distribution of which has clearly defined modes, i.e., there is a clear contrast between the brightness of the background and the foreground. Since

many images not only have an uneven distribution of pixel intensity but also contain differently distorted areas, the use of global binarization methods in such cases is not recommended.

Much more widespread are local or adaptive methods, which involve dividing the image into blocks using sliding scanning and independent binarization of the central pixel of each block separately. In this case, the threshold value of a pixel is determined taking into account the characteristics of both this pixel and its neighbors in the block.

Currently, many adaptive binarization algorithms are proposed. The best known are the Niblack [6] and Sauvola [7] algorithms, which, along with numerous applications, have been improved by various authors in solving a number of problems with low-quality images. The purpose of these improvements is to improve the quality of the binarization result. Some works were also carried out on a comparative analysis of their various modifications using certain image properties (we point out, for example, [8]).

Another way to improve the binarization procedures is to apply a preliminary transformation of the original image. For example, in [9], it is proposed to preliminarily apply the histogram equalization process, in [10] - the filtering technique, in [11] - the gamma transform, and many other improvement techniques are applied.

However, two types of distortions are always observed in a binary image, namely the loss of certain properties and details of the tested image and the appearance of artifacts that not related to its content.

In this paper, we propose an approach based on selecting all segments of a binary image, analyzing the distribution of their sizes, and selecting segments with certain sizes. This approach will allow classifying objects in a binary image by size and, if necessary, selecting segments of a certain size, in particular, artifacts to exclude them.

As the basic binarization procedures, the Niblack and Sauvola algorithms are chosen, which are easy to implement and often give quite acceptable results, as well as the Otsu algorithm for comparison.

II. RESEARCH METHOD

The proposed technique for improving binarization methods consists of applying a special segmentation procedure to the result of global or adaptive binarization, analysis of the size distribution, and an appropriate choice of segments.

Binary image segmentation. When segmenting a binary image, only black pixels are considered. A subset of black pixels of the tested image is called a segment if all neighbors of each pixel of the subset are included in the same subset. Thus, the segments of the binary image do not intersect, so the implementation of the segmentation algorithm is simplified.

Denote S_1, S_2, \dots, S_K the segmentation results, and N_1, N_2, \dots, N_K the number of pixels of segments. Of interest is the histogram of the size of the segments, which characterizes the structure and properties of the image. So, on the image of the histogram, the existing clusters of segments with certain sizes are highlighted, in particular, artifacts that are small in size, but which, as a rule, are many.

Analyzing the histogram, the interval $[L_1, L_2]$ of segment size values required for further analysis and processing is selected. For this, the following operations are performed:

- Binarization by the chosen method,
- Full segmentation of a binary image with fixation of pixel coordinates and sizes of all found segments,
- Full histogram analysis by segment size distribution,
- Fixing the required interval $[L_1, L_2]$ of segment size values and selecting segments with these size values,
- Merges all selected segments into a single image with a white background.

The described procedure for segmentation and manipulation with segments of certain sizes is implemented as a special software system with the ability to analyze and visualize intermediate and final results.

III. EXPERIMENTAL RESULTS

Below are the results of experiments with fragments of images of ancient handwritten documents. The tested images are borrowed from the materials of the manuscripts stored in the Matenadaran - the Institute of Ancient Manuscripts named after Mesrop Mashtots in Yerevan. For experiments, both fragments of images of manuscripts are selected, on which the background brightness is visually distributed evenly, and images with an uneven distribution. All tested images are converted to Grayscale 8-bit format.

A. Application to images with global binarization. The purpose of this experiment is to evaluate the parameters L_1 and L_2 for proper separation of the image into artifacts and objects of interest.

Table 1 shows two examples of images by applying the above operations. The first column of the table shows fragments of images of ancient manuscripts, and the second column shows the results of binarization by the Otsu method with an indication of the binarization threshold Tr . Below is a graph of the initial fragment of the histogram of the sizes of binary image segments (third column), with the help of which the parameters L_1 and L_2 are visually estimated and segments with the corresponding sizes are selected (fourth and fifth columns). In this case, the quality of the result is assessed by visual analysis of the readability of the text. We see that in these examples the quality of binarization is indeed higher than before the application of the procedure. If necessary, the quality can be formally evaluated using the well-known quality criterion F-score [12].

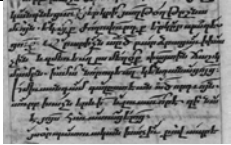
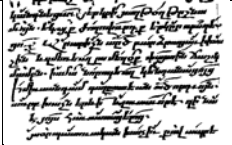
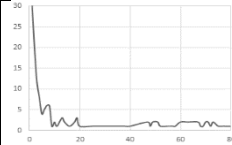

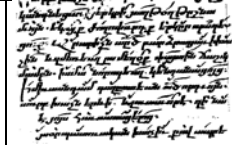
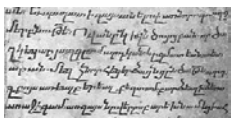
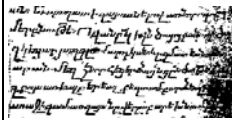
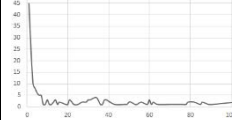

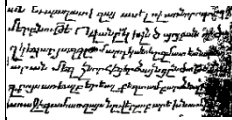
B. Application to images with adaptive binarization. The proposed method for improving the quality of binarization is also applicable in the case of adaptive binarization methods. Here we will look at the well-known and widely used Niblack and Sauvola algorithms.

Niblack's algorithm. The local binarization threshold $T(x, y)$ for a pixel with coordinates (x, y) is determined using the average value $m(x, y)$ and the standard deviation $\sigma(x, y)$ of all pixels of a sliding window with size $w \times w$ and centered at the point (x, y) , according to the formula

$$T(x, y) = m(x, y) + k * \sigma(x, y) \quad (1)$$

The quality of the binarized image depends on the values of the parameters w and k and, according to many authors, with $w = 15$ it is preferable $k = -0.2$.

Tab. 1. Examples of application of the proposed procedure in case of global binarization

				
Original	Otsu, Tr=104	Histogram fragment	$L_1=1, L_2=40$	$L_1=41, L_2=6000$
				
Original	Otsu, Tr=153	Histogram fragment.	$L_1=1, L_2=9$	$L_1=10, L_2=6000$

A characteristic feature of this method is the abundance of artifacts, sometimes even comparable in size to the objects in the image. This is due to the increased sensitivity of the method to noise. Some modifications of this method are proposed in the literature to improve the quality of binarization of images with uneven illumination, with areas that have a low contrast texture or with high values of variation. The well-known method in this sense is Sauwola's

method, which is an improvement on the Niblack algorithm. In this method, the formula for determining the local binarization threshold has the form

$$T(x, y) = m(x, y) \left[1 - k * \left(1 - \frac{\sigma(x, y)}{R} \right) \right] \quad (2)$$

where R is the dynamic range of values $\sigma(x, y)$ (usually $R = 128, k \in [0.2, 0.5]$).

Table 2. Examples of application of the proposed procedure for binarization by the Niblack method

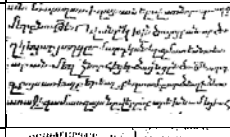
Original	Otsu	Niblack (w=15, k=0.2)	Niblack (w=15, k=0.2)	Sauvola (w=15, k=0.5)
				
				
				
				
				

Table 2 shows the results of the binarization of fragments of images of ancient handwritten texts by the Otsu method (second column), the Niblack method (third column), the improved Niblack method (fourth column) and Sauvola's method, for comparison. A visual comparison of the above images shows that the improved Niblack method leads to significantly better results than the Otsu method and the main

Niblack method, and also slightly outperforms Sauvola's method. The binarization parameters are indicated in the table, except for the image in the last row, which was processed with the values of the parameters $k=-0.5$ for Niblack's method and $k=0.8$ for Sauvola's method.

It should be noted that the proposed improvement procedure is also applicable to other methods of adaptive binarization.

Thus, the presented experimental results show the productivity of the proposed approach to the issue of improving the quality of a binary image by both global and adaptive methods.

IV. CONCLUSIONS

A large number of binarization algorithms for historical handwritten documents have been proposed in the literature, but methods that work equally well in all cases have not been found. Various criteria for evaluating the quality of the binarization result were also proposed. In the case under consideration with the binarization of images of handwritten texts, the degree of readability of the text by a visual method or using technical means is used as a criterion for the quality of the binarization algorithm. One of the approaches to improve the quality of the binarization result proposed in the literature is the preliminary processing of the original image using filtering, morphological analysis, spectral analysis, etc., methods. In this paper, we propose a method for improving the quality of a binary image based on its segmentation, analysis of the distribution of segment sizes and selection of segments with certain sizes. The proposed procedure allows to select the artifacts existing in the image, resulting from the background binarization, as well as objects of interest with certain sizes.

To test the effectiveness of the proposed procedure, we used binary images of handwritten documents obtained by Niblack's algorithm. The results of Otsu's and Sauvola's algorithms were also used for comparison. The results indicate an improvement in the quality of the binary image and the emergence of additional opportunities for classifying objects in the image.

ACKNOWLEDGMENT

This work was supported by the Russian Foundation for Basic Research and RA Science Committee in the frames of the joint research project RFBR 20-51-05008 Arm_a and SCS 20RF-144, accordingly.

REFERENCES

- [1] J. Ambily, S.B. Jaini, J. Poorna, K.V. Beena. "Objective Quality Measures in Binarization", *International Journal of Computer Science and Information Technologies*, vol. 3, no. 4, pp. 4784-4788, 2012.
- [2] K. Nürogiannis, B. Gatos and I. Pratikakis, "An Objective Evaluation Methodology for Document Image Binarization Techniques," *The Eighth IAPR International Workshop on Document Analysis Systems*, pp. 217-224, 2008.
- [3] N. Otsu. "A threshold selection method from gray-level histograms", *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no 1, pp. 62-66, 1979.
- [4] M. Sezgin, and B. Sankur. "Survey over image thresholding techniques and quantitative performance evaluation", *Journal of Electronic Imaging*, vol. 13, no. 1, pp. 146 - 165, 2004.
- [5] D. Asatryan, M. Haroutunian, G. Sazhumyan, A. Kupriyanov, R. Paringer and D. Kirsh, "Comparative Quality Analysis of Image Global Binarization Procedures", *IX International Conference on Information Technology and Nanotechnology (ITNT), Samara, Russian Federation*, pp. 1-5, 2023.
- [6] W. Niblack, *An Introduction to Digital Image Processing*, Englewood Cliffs, N.J. Prentice Hall, pp. 115-116, 1956.
- [7] J. Sauvola and M. Pietikäinen, "Adaptive document image binarization", *Pattern Recognition*, vol. 33, no. 2, pp. 225-236, 2000.
- [8] J. He, Q. D. M. Do, A. C. Downton and J. H. Kim, "A comparison of binarization methods for historical archive documents," *Eighth International Conference on Document Analysis and Recognition (ICDAR'05)*, vol. 1, pp. 538-542, 2005.
- [9] P. D. Ingle and P. Kaur, "Adaptive thresholding to robust image binarization for degraded document images," *1st International Conference on Intelligent Systems and Information Management (ICISIM)*, pp. 189-193, 2017.
- [10] B. Gatos, I. Pratikakis, and S.J. Perantonis, "Adaptive degraded document image binarization", *Pattern Recognition*, vol. 39, pp. 317 - 327, 2006.
- [11] Cunzhao Shiy, Yanna Wangy, Baihua Xiaoy and Chunheng Wang, "OTSU Guided Adaptive Binarization of CAPTCHA Image using Gamma Correction", *23rd International Conference on Pattern Recognition (ICPR)*, pp. 3951-3956, 2016.
- [12] <https://en.wikipedia.org/wiki/F-score#mw-head>