Novel Approach to Background-Text-Non-Text Separation in Ancient Degraded Document Images

David, Asatryan
Institute for Informatics and Automation Problems
Yerevan, Armenia
e-mail:
dasat@ipia.sci.am

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

Lusine, Aznauryan
Russian-Armenian (Slavonic) University
Yerevan, Armenia
e-mail:
lusine.aznauryan8@gmail.com

ABSTRACT

Nowadays lots of handwritten and printed ancient documents need to be digitized for automated processing and analysis. In this paper, an approach to background-text-non-text separation procedure based on differences of presented in a document image objects sizes which can be obtained by binarization and segmentation algorithms, is proposed. After binarization by proper method it is segmented and the distribution of segments sizes is obtained. It is assumed that the three types of objects presented in an image have significantly different sizes; therefore the problem of separation comes to discrimination of the set of segments into three groups. The thresholds for separation of these groups can be found by minimizing the intrasample variation which used in discriminant analysis. Some examples of images from Matenadaran collection are considered and the separated parts of the image are illustrated and interpreted.

Keywords

Ancient document image, degraded image, binarization, segmentation, distribution of segments sizes, discriminant analysis.

1. INTRODUCTION

There exist lots of handwritten and printed historical documents in libraries and museums in the world including ancient manuscripts, author writings, old newspapers, archives, etc. Nowadays old and historical documents are archived and preserved in large quantities worldwide in digital form. We can mention that the Matenadaran (Armenian Museum of Ancient Manuscripts, [1]) is in possession of a collection of nearly 17,000 manuscripts and 30,000 other documents which cover a wide array of subjects such as historiography, geography, philosophy, grammar, art history, medicine and science. Due to wide area and diversity of subjects the investigation of Armenian ancient documents has special interest for development of image processing technique.

A lot of research papers have been dedicated to optical character recognition (OCR) systems to convert paper information into digital. One of the important tasks for successful recognition is an adequate separation of text from other components, which are presented in the document image. Moreover, the objects separated from text also can present certain interest for analysis and understanding of ancient world history.

It is well known that an extraction of the text from a degraded document image is a very challenging task due to the high variation between the document background and the foreground text of document images and the existing there painting objects.

The most important procedures in pre-processing of poor quality scanned documents are binarization and segmentation which, in general, is not an easy task and often requires an individual approach to analysis of document of interest. Any binarization procedure becomes more difficult when the image is heterogeneous, there occur a varying illumination of objects and many other factors, so the literature devoted to text binarization problems is very large. Analyzing of the proposed approaches in the literature is not of special interest of this paper and we can refer to surveys [2, 4].

The problem of text extraction from images was also considered in a huge number of papers (see, for examples, survey [5]). The problem of text/background separation in images is closer to paper [6]. But the extraction or separation of objects is performed using the specific properties of the considered images with text, for instance, in these papers - text location in an image, text line properties, edges, etc. Therefore, the set of documents with different features has many classes, so the methods for problem's solution are also very different.

In this paper, we limit ourselves to a class of ancient degraded documents, in which there can be objects of three kinds: background (polluted texture), text and images (or paintings). The sizes of text symbols and images should be at least differ visually from the objects of the remained types. Understanding documents of such type is one of the most difficult tasks because it is necessary to apply simultaneously many image processing approaches, often contradicting each other.

We use a very simple approach based on three distinguishing types of specified objects on the document image by size of segments obtained by appropriate algorithms of binarization and segmentation. The visual analysis of images of many ancient handwritten documents from the Matenadaran collection show that, as a rule, they consist of these three types of objects. In Figure 1 some typical examples of such ancient handwritten document images are illustrated which are preliminarily converted into Gray Scale (8 bit) format for simplicity. The color images are processed in the similar manner after splitting them in a proper color space. Usually the images of such type have a two-mode histogram, so they can be successfully binarized by many existing algorithms (for example, by Otsu algorithm [7]).

The background in these images usually looks like a polluted texture or a set of polluted connected areas of small sizes. Text part of such images usually contains many handwritten symbols of small sizes located in a certain part of an image.

The number of text symbols is usually large and occupies the major portion of an image. The third type of object represents itself as a picture of size which noticeably exceeds the sizes of symbols.

Figure 1. Some documents from Matenadaran collection

In this paper we propose a separation procedure for three mentioned types of objects based on analysis of intrasample and intersample scatterings of segments sizes which are obtained after binarization and total segmentation.

2. PROPOSED PROCEDURE

A procedure and software tool were developed for binarization, segmentation and separation of a set of segments sizes into three subsets by choosing thresholds t_1 and t_2.

The procedure includes the following steps:
Step 1. Binarization of a document image. A great number of techniques have been proposed in the literature for the binarization of gray-scale or colored documents images, but, no one among them is generic and efficient for all types of documents. In this paper we assume that the binarization of images of considered type is performed by any appropriate way and we will consider them in a separate paper.

Step 2. Segmentation of binarized document image. We take into consideration comprehensive segmentation, when each "black" pixel of an image I belongs to one of the segments $S_1, S_2, ..., S_N$, where N is the general number of segments. This means that

a) $I = \bigcup_{i=1}^{N} S_i$,

b) $S_i \bigcap S_j = 0,\ i, j = 1,2,...,N$.

Let $n_i, i = 1,2,...,N$ be the number of pixels of segment S_i. All segments should localized to be processed or deleted in case of necessity.

Step 3. Discrimination of segments between three groups by sizes. At this step we create and analyze a histogram of segments sizes and split the set of all segments into small, medium and big sizes using two thresholds t_1 and t_2, $t_1 + 1 < t_2$. The interval $0 < z \le t_1$ corresponds to small sizes; interval $t_1 < z \le t_2$ corresponds to medium sizes, and interval $z > t_2$ to big sizes for the examined segments.

If we suppose that some objects in the binarized image have sizes significantly exceeding the others, then the second threshold t_2 can be found very easily. Sometimes the sequence z_i contains many consequent elements with $z_i = 0$. So, often some visual consideration of sequence z_i can be enough for choosing an appropriate value for the threshold t_2.

In the general case, we propose to use the three-class discrimination analysis, based on intrasample and intersample scatterings of classes.

We use the concepts, approach and the corresponding denotations of discriminant analysis [8] to find appropriate values of thresholds t_1 and t_2.

Let $z_i, i = 1,2,...,N$ be number of segments which consists of i pixels. The set of total N segments are divided into three subsets with elements belonging to intervals $[1, t_1]$, $(t_1, t_2]$ and $(t_2, N]$. Let $z_{\xi_\gamma}^{(\gamma)}$ be an element of, and $n_\gamma, \gamma = 1,2,3$ be the number of elements of γ-th interval, $N = n_1 + n_2 + n_3$.

Then let's denote $\overline{z} = \frac{1}{N}\sum_{i=1}^{N} z_i$, $\overline{z^{\gamma}} = \frac{1}{n_\gamma}\sum_{\xi_\gamma=1}^{n_\gamma} z_{\xi_i}^{\gamma}$,

$$S_W(t_1,t_2) = \sum_{\gamma=1}^{3}\sum_{\xi_\gamma=1}^{n_\gamma}\left(z_{\xi_\gamma}^{(\gamma)} - \overline{z^{\gamma}}\right)^2,$$

$$S_B(t_1,t_2) = \sum_{\gamma=1}^{3} n_\gamma\left(\overline{z^{(\gamma)}} - \overline{z}\right)^2,$$

$$S = \sum_{i=1}^{N}\left(z_i - \overline{z}\right)^2.$$

It can be shown that $S = S_W(t_1,t_2) + S_B(t_1,t_2)$. Here S is the scattering of total sample, $S_W(t_1,t_2)$ and $S_B(t_1,t_2)$ are intrasample and intersample scatterings correspondingly for three specified groups.

The values of thresholds t_1 and t_2 are calculated by minimization of ratio

$$Q(t_1,t_2) = S_W(t_1,t_2)/S.$$

Let's show the proposed separation procedure by the application to the document image [9] of Toros Taronatsi shown in the first column of Figure 1 (converted into Gray Scale image and shown in the second column). The histogram of this image is shown in Figure 2. Binarization of this image is reasonable to perform by Otsu method. The Otsu threshold is equal to 139, and the corresponding binary image is shown in the first column of Figure 2.

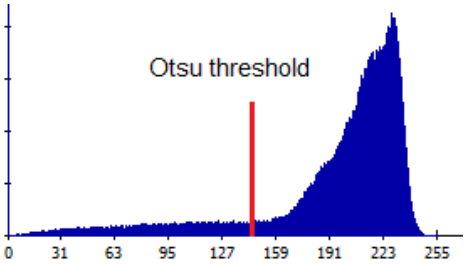


Figure 2. Image histogram and Otsu threshold.

The segmentation procedure is applied after binarization, and all segments are examined by the number of pixels belonging to each "black" segment. The number of segments N can be large because of presence of noised background of an image which can include artifacts. In this example we have 795 segments which include from 1 to 169298 pixels. A fragment of distribution of segments sizes is given in Tab. 1.

Application of separation algorithm described above gives the values of two thresholds $t_1 = 33$ and $t_2 = 201$.

The first interval includes 490 segments of size from 1 to 30 (see second column of Table 2). These segments present separate pixels of the background and various artifacts, which mostly are degraded parts of objects presented in the image, some meaningless connected sets of image pixels, etc.

Table 1. A histogram fragment of segments sizes

Interval	$\sum z_i$	Interval	$\sum z_i$
1 - 10	217	101 - 110	4
11 - 20	40	111 - 120	4
21 - 30	26	121 - 130	7
31 - 40	77	131 - 140	2
41 - 50	86	141 - 150	2
51 - 60	49	151 - 160	2
61 - 70	20	161 - 170	1
71 - 80	20	171 - 180	0
81 - 90	16	181 - 190	0
91-100	6	191 - 200	0

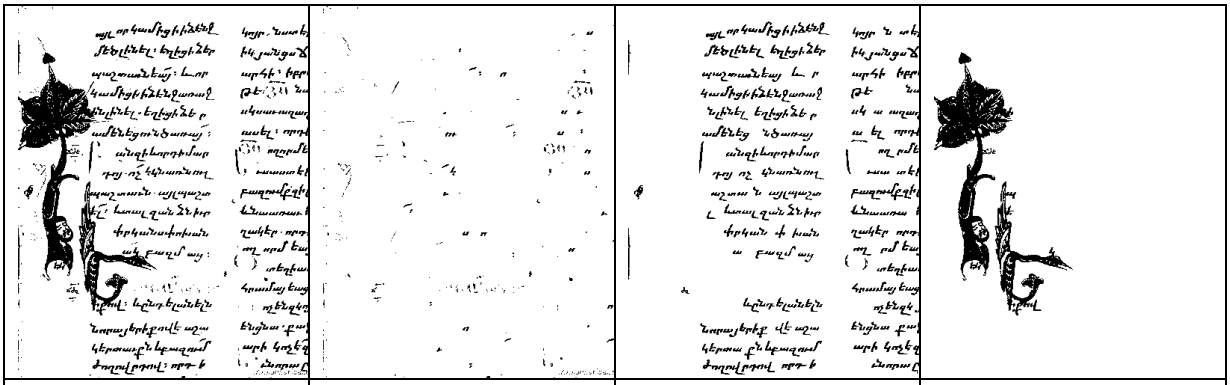
Second interval includes 302 segments of size from 31 to 209 (see the third column of Table 2). These segments present the text symbols (or, maybe some specific nonsignificant details of the image).

Finally, the third interval includes 3 segments of size from 210 to 169298. We can notice that the biggest segments fall into the third group. The first two segments are shown in the fourth column of Table 2, but the third (the biggest one) usually presents the background of an image.

It can be noted that by using another binarization methods or its combinations, *better results can be achieved* for many images of ancient documents which have inhomogeneous background or painting objects.

A software system was created to perform the proposed procedure. It has the possibility to choose any interval of segment sizes and separate them from the image.

Table 2. Separated image parts of different sizes



1 - 300000	1 - 33	34 - 201	>201

CONCLUSIONS

Nowadays lots of handwritten and printed ancient documents need to be digitized for automated processing and analysis. There are a lot of papers devoted to the problem of text extraction or separation from background in document images. In this paper, an approach to background-text-non-text separation procedure based on differences of the presented in a document image objects sizes which can be obtained by binarization and segmentation algorithms, is considered. The procedure is applicable to documents, in which three types of objects can be presented, namely - the background, text symbols and paintings or pictures. The distribution of segments sizes is analyzed and a procedure for separation of these three types of objects is proposed. The separation method is adopted from discrimination analysis methodology. Some examples of ancient degraded documents from the Matenadaran collection are considered, separated parts of an image are illustrated and interpreted.

REFERENCES

[1] http:// http://www.matenadaran.am/

[2] Nabendu Chaki , Soharab Hossain Shaikh, Khalid Saeed. "A Comprehensive Survey on Image Binarization Techniques". in Exploring Image Binarization Techniques. Volume 560 of the series Studies in Computational Intelligence, pp. 5-15, 2014.

[3] Rashmi Saini. Document Image Binarization Techniques, Developments and Related Issues: A Review. *International Journal of Computer Applications* Volume 116 , No. 7, pp. 41-44, April 2015.

[4] Sangeetha R., Rajkumar T.C, Amali Asha A. Survey on Document Image Binarization for Enhancing Degraded Document Image. *International Journal of Innovative Research in Computer and Communication Engineering*, Vol. 5, Issue 3, pp. 3826-3829, March 2017.

[5] C.P. Sumathi, T. Santhanam and G.Gayathri Devi. A Survey on Various Approaches of Text Extraction in Images. *International Journal of Computer Science & Engineering*, Vol.3, No.4, pp. 27-42, August 2012.

[6] Abderahmane Kefali, Toufik Sari, Halima Bah. Text/ Background separation in the degraded document images by combining several thresholding techniques. *WSEAS Transactions on Signal Processing*, Volume 10, pp. 436-443, 2014

[7] N. Otsu, ―A threshold selection method from gray-level histograms‖, *IEEE Transaction on Systems Man and Cybernetics (SMC)*, Vol. 9, pp. 62-66, 1979.

[8] Wilks, S. S.: Mathematical Statistics. J. Wiley and Sons, New York–London, 1962.

[9] E. Korkhmazian, I. Drampian, G. Hakopian, "Armenian Manuscripts of the 13th and 14th centuries", Matenadaran Collection, Leningrad, 1984