ISSN 1102-3321 «AL-FARG'ONIY AVLODLARI» ELECTRONIC SCIENTIFIC JOURNAL № 1 (1) | 2023

FEATURES OF HANDWRITING IMAGE PROCESSING AND ANALYSIS

https://doi.org/10.5281/zenodo.7711389

Sobir Radjabov,
Doctor of technical sciences (DSc) of "Tashkent Institute of Irrigation and Agricultural Mechanization Engineers" National Research University

Musokhan Dadakhanov,
Namangan State University, head of the «Informatics» department, candidate of technical sciences (PhD)

Muhammadmulla Asraev,
Senior lecturer of the Fergana branch of the Tashkent University of Information Technologies named after Muhammad al-Khorazmi

Abstract. During the digitization stage, the manuscript text may be corrupted or in some cases interfered with due to technical or human factors. In addition, in some cases the source itself (for example, an ancient manuscript) is in poor condition.

Keywords: Geometric distortion, uneven lighting, contrast variation, ink bleed distortion, faded ink, staining, paper ghosting, blurring, fine or faint text, poor quality documents.

Introduction. Non-text elements include ink stains, paper stains, pictures, etc. that were not removed during the binarization process. Their extraction can be done, for example, by extracting the connection components in the image, calculating geometric features and classifying the connection component as part of the text or as a defect using machine learning or heuristic methods.

At the stage of identifying segmented text features, a set of features is formed that allows solving one or another problem of the analysis of the manuscript text, and it depends on the practical problem to be solved. Thus, when solving the problem of recognizing a handwritten text, features of handwritten signs that are common to one character are determined, and when determining the author of a handwritten text, features specific to the author of the text are determined.

In the character space formed at the stage of recognition, recognition is carried out based on one or another recognition method.

Literature review and methodology. At the moment, the task of automating the process of analyzing handwritten text images has not been fully solved. One of the most important stages is the preliminary processing of manuscript images, which is underdeveloped.

This article examines such interferences, which are common in practice, and the available algorithms for their elimination or reduction.

Geometric distortion. In the digitization of manuscript text images, sources such as changes in the lens of the digitizing device and distortion of the geometric similarity of the image with respect to its original appear.

The main types of initial image distortion include distortion (pillow and barrel distortion), warping, and perspective. One of the ways to eliminate such distortions is to reduce the nonlinear predistortion of the raster during image spreading, which allows to compensate for the expected distortions. In addition, the distorted raster can be a posteriori corrected based on a polynomial approximation of each horizontal and vertical line. The approximations are then used to calculate the inverse correction functions for each grid cell. The spatial curvature method allows to eliminate perspective distortions. By tilting the image of an elongated object viewed from the side, it can be made to appear as an image viewed at a right angle. Another important application of this method is to correct paired images of the same scene taken at different viewing angles.

Uneven lighting. Due to incident light in an optical medium, the scattering of image particles in the light path decreases exponentially, and as a result, the image

quality obtained by light microscopy deteriorates and defects such as uneven illumination appear [1]. As a result of light absorption and scattering, light spectra change and it causes background objects to be unevenly illuminated [2]. Uneven illumination of background objects causes various problems in document image analysis. If the lighting in the handwritten text images is uneven, the recognition of the texts in the image will often lead to ineffective results. In general, high-accuracy text recognition involves converting the grayscale image to a binary image and extracting the text. At the same time, due to uneven illumination, the binary image has artifacts, which causes incorrect extraction of text from the regions of these artifacts.

Contrast variability. Contrast is defined as a change in brightness. In most cases, contrast refers to the difference between high and low intensity pixels in an image. In addition, contrast can be obtained as the difference between the pixel values of the top or bottom of the object and the background in the image [3]. Factors such as interference environment, sunlight, illumination and occlusion often lead to contrast variability [4]. Contrast variation in handwritten text images causes various problems in extracting the foreground text from the background of document images and using traditional thresholding methods and algorithms in their analysis. Such problems are overcome by applying image quality improvement techniques before image binarization.

Disruption due to ink leakage. Ink smears or ink bleeds occur when text is written on both sides of the paper, or when ink used on one side begins to show on the other side of the paper. The presence of ink leakage in the binarization of handwritten text images leads to increased interference. The ink spreads from one side of the page to the other, causing the quality of the text there to deteriorate. Many ideas have been proposed to prevent such disruptions, and researchers have faced two major challenges. The first problem is a corrupted document digitized in high resolution. This is except for cases involving a digitization project or a library. The second problem is found in all methods of recovery. This is due to the fact that the original and complete sample of historical document records is not available in the quantitative analysis of the results [5]. This problem can be solved by preparing a specific image with quality distortion based on the original truth [6] or by improving the image quality by knowing the original truth in the distorted images [7]. Although reliability is not available, efficiency can always be analyzed. It is necessary to determine the quantitative effect of the recovery result for the next stage. For example, characters in a handwritten text image can be evaluated by analyzing the output of optical recognition systems.

Faded ink or faded characters. There is considerable historical, social and political interest in analyzing large numbers of official documents and placing them in digital libraries and archives, where most of the official documents are typed. This creates various problems related to their recognition [8] . Each individual character in a document may appear thinner or thicker compared to other printed documents or characters around it. It directly depends on the printing part of the particular character button on the machine and its pressing force.

Second, most machine-printed documents survive only as copies printed on very thin paper (Japanese paper) with a unique texture. In addition, most machine-printed copies of documents (originals and carbon copies require a harder key press) are blurred [9]. Problems such as reuse and wear, tears, stains, paper clip rust, holes, document fragmentation, and discoloration adversely affect the quality of machine-printed historical documents.

View behind painting or paper. After the digitization of manuscripts, many problems arise due to distortions in images and low resolution digitization. Such problems have a negative impact on the visual appearance of the image [10]. Historical manuscripts can have various forms of corruption. All the disorders in them depend on the passage of time and have a different nature [11]. However, one of the biggest problems with documents is the quality degradation caused by one-sided writing being reflected on the back of the paper. In the past, many documents were written on both sides of the paper [12] . This problem occurs when the ink on one side shows through on the other side, making the document difficult to read. Such disruptive documents need to be restored to make them easier to read. Solving the problem of the appearance of ink behind the paper allows you to significantly reduce the time of image compression and download them faster over the Internet. If such distortions in the image are eliminated, a clear background can be achieved [10] .

Fading. There are two types of blur in handwritten text images, motion blur and focus blur. In general, motion blur artifacts are caused by the relative speed

between the camera and the subject, or sudden rapid camera movement. Out-of-focus blur occurs when the light beam fails to converge on the image. Research topics in solving the blurring problem are implemented using blurring evaluation tools in images to evaluate the accuracy of optical character recognition, thereby allowing the user to obtain new images and provide the opportunity to achieve the required recognition accuracy.

Fine or weak text. Manuscript documents written in the past consisted of very thin or faint text, written mostly in ink and sometimes through paint. Fading of the ink or paint used in the writing of historical manuscript documents results in a deterioration of the image quality of the manuscript. In other cases, the use of low-quality ink and the nature of the paper used will cause thin or weak text. This complicates the application of binarization methods and text recognition. Nowadays, researchers are more interested in the analysis of historical manuscript text images, which poses various challenges. Distortions in historical documents, such as delicate or weak texts, are prompting researchers to develop image enhancement and binarization algorithms that provide good enough results to solve these problems [13]. Based on the binarized images, successive steps such as joint displacement detection, page or line segmentation were created later.

Degraded documents. Typically, original documents written on paper come in a variety of media (ink, graphite, watercolor) and formats (maps, spreadsheets, and notebooks). Such documents may contain informational, evidential, associative and significant information of intrinsic value [3]. A document consisting of historical, legal or scientific information is considered to be of great evidentiary value if the original state of the media format and image has not undergone a drastic change or quality has not deteriorated [5]. Nevertheless, the careless use of documents is not the only factor that leads to the loss of their parts, deterioration of quality and various damages [3]. Factors such as poor storage and poor use and environmental conditions also affect the quality of documents. In addition, serious damage and quality deterioration can also be caused by environmental factors.

Results. To achieve this goal, it is necessary to perform the following tasks:

- analysis of the modern state of problems of processing and recognition of handwritten text images;
- researching the problem of quantitative assessment of the quality of the original image of the handwritten text;
- separation of periodically repeating straight lines in the image;
- handwritten text image binarization;
- segmentation of handwritten text lines and words in the image;
- experimental research and evaluation of the effectiveness of algorithms created for preliminary processing of handwritten text images;
- creation and practical application of a set of preprocessing programs for handwritten text images.

Conclusion.The main practical problems to be solved based on the analysis of handwritten text images were studied. In this case, it was proved that the initial processing stage of given images is important in the creation of systems for the analysis of handwritten text images and has a significant impact on the final results of the system.

Specific features of processing and analysis of handwritten text images are identified and the main problems are described. Algorithms that allow solving the stated problems were analyzed. As a result of the analysis, their achievements and shortcomings were determined. It was found that the existing algorithms completely or partially solve some of the identified problems. This situation shows the need to develop and research algorithms for pre-processing of handwritten text images.

References:

1. Ploem J.S., Tanke H.J. Introduction to Fluorescence Microscopy. Wiley Liss, Inc.: New York, NY, USA, 2001.
2. Van der Kempen, G.M.P.; van Vliet, L.J.; Verveer, P.J.; Van der Voort, H.T.M. A quantitative comparison of image restoration methods for confocal microscopy. J. Microsc. 1997, 185, 354–365.
3. Mustafa W.A., Yazid H. Image Enhancement Technique on Contrast Variation: A Comprehensive Review. J. Telecommun. Electron. Comput. Eng. 2017, 9, 199–204.
4. Mustafa W.A., Yazid H. Illumination and Contrast Correction Strategy using Bilateral Filtering and Binarization Comparison. J. Telecommun. Electron. Comput. Eng. 2016, 8, 67–73.
5. Hadjadj Z., Meziane A., Cheriet M., Cherfa Y. An active contour-based method for image

binarization: Application to degraded historical document images. In Proceedings of the 14th International Conference on Frontiers in Handwriting Recognition (ICFHR'14), Crete Island, Greece, 1–4 September 2014; pp. 655–660.

6. Huangy Y., Brown M.S., Xuy D. A Framework for Reducing Ink-Bleed in Old Documents. In Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 23–28 June 2008.

7. Leedham, G.; Varma, S.; Patankar, A.; Govindaraju, V. Separating text and background in degraded document images – A comparison of global thresholding techniques for multi-stage thresholding. In Proceedings of the 8th International Workshop on Frontiers in Handwriting Recognition, Niagara-on-the-Lake, ON, Canada, 6–8 August 2002; pp. 244–249.

8. Smigiel, E.; Belaid, A.; Hamza, H. Self-organizing Maps and Ancient Documents. In Proceedings of the 6th International Workshop on Document Analysis Systems VI, Florence, Italy, 8–10 September 2004; pp. 125–134.

9. Sehad, A.; Chibani, Y.; Cheriet, M.; Yaddaden, Y. Ancient degraded document image binarization based on texture features. In Proceedings of the 2013 8th International Symposium on Image and Signal Processing and Analysis (ISPA), Trieste, Italy, 4–6 September 2013.

10. Quraishi, M.I.; De, M.; Dhal, K.G.; Mondal, S.; Das, G. A novel hybrid approach to restore historical degraded documents. In Proceedings of the 2013 International Conference on Intelligent Systems and Signal Processing (ISSP), Gujarat, India, 1–2 March 2013; Volume 1, pp. 185–189.

11. Shirani, K.; Endo, Y.; Kitadai, A.; Inoue, S.; Kurushima, N. Character Shape Restoration of Binarized Historical Documents by Smoothing via Geodesic Morphology. In Proceedings of the 2013 12th International Conference on Document Analysis and Recognition, Washington, DC, USA, 25–28 August 2013; Volume 12, pp. 1285–1289.

12. Xu, L.; Yan, Q.; Xia, Y.; Jia, J. Structure extraction from texture via relative total variation. ACM Trans. Graphics 2012, 31, 139:1–139:10.

13. Nagendhar, G.; Rajani, D. China Venkateswarlu SonagiriV.Sridhar. Text Localization in Video Data Using Discrete Wavelet Transform. Int. J. Innov. Res. Sci. Eng. Technol. 2012, 1, 118–127.