EDCS 2019 IOP Publishing
Journal of Physics: Conference Series **1488** (2020) 012014 doi:10.1088/1742-6596/1488/1/012014

Methodology for obtaining textual information from images and its analysis

A A Allaberganov and M Yu Kataev
Tomsk State University of Control Systems and Radioelectronics (TUSUR) 40 Lenin ave., 634050, Tomsk, Russia

E-mail: nsk-kapital@mail.ru

Abstract. The paper proposes a solution of the issue of extracting textual information from a document image. There are many interesting solutions of this problem thanks to practical applications, such as text recognition, translation, forensics, etc. The quality of the text information of documents converted into images depends on the type of scene (simple or complex), type of digital camera (low-cost devices, mobile devices), the magnitude and direction of illumination, type of substrate (paper, plastic). These points lead to the fact that the automatic extraction of text from the image is extremely difficult.

1. Introduction

The informational side of extracting textual information consists of three stages: image segmentation, localization and text extraction, as well as character recognition [1-3]. The text in the images contains significant and useful information for understanding the content of images, which plays an important role in the analysis of the documents themselves. The image of the document contains various information, such as texts, drawings, and graphics, which are solid or dashed lines. The difficulty of extracting these lines (which make up the text) consists in changing the image quality by scanning, the long history of the document itself or its image, etc. In the case of a colour image of the document, extracting the text becomes difficult and it is not always possible to distinguish between individual components of the text (lines that make up letters) due to mixing colours of text and background.

Text detection methods can be classified into three categories. The first consists of interconnected methods that assume that image areas containing text must have uniform colours, satisfy certain sizes, a given shape, etc. These methods are effective only in the case of high contrast between the colour of the text and the background, and in other cases, the efficiency is low. The second consists of methods that compute texture metrics, which assume that areas of the image containing text have a texture different from the background. Although these methods are relatively less sensitive to background colour, they may not distinguish between text and text-like backgrounds. The third consists of methods based on the calculation of lines on the image and the search for connectivity between them. Areas of text are detected under the assumption that the marginal edges of the background are smaller than those of the text areas. However, such approaches are not very effective for detecting texts with large font sizes.

Currently, along with the expanding use of the Internet and e-mail, it is the transmission of an image of a document obtained with a digital camera or formed using a scanning device, simply a PDF or Word document. At first glance, the presence of an electronic document is convenient, but upon closer examination, without a digital original, the image is difficult to copy not the image itself, but the information presented on it. The inconvenience of storing electronic versions of the document as images is also obvious due to the large file size and inaccessibility of the content for automatic processing. Document processing may include the selection of documents by keywords, determining the subject of the document by finding words specific to any subject area, automatic indexing and translation, as well as the classification of documents according to the affiliation of the sending organization. To solve all these problems, an accessible text of the document is needed.

The relevance of solving the problem of highlighting textual information in a digital image format is confirmed by the latest numerous publications around the world. Even the generally recognized leaders among optical character recognition (OCR) packages, which are designed specifically for solving tasks of this kind, cannot cope with the recognition of the usual image of a test document, despite the fact that the text can be easily read visually. Existing recognition systems do not always allow efficient recognition of images of printed texts of low quality, typical of documents received, for example by fax or received in poor lighting, glare, etc.

Manufacturers of software for information systems provide solutions for extracting textual information from an image, but the algorithms they offer require detailed tuning for the needs of a particular customer. In scientific research, the tendency to minimize human labor by the widespread introduction of machine learning methods prevails. Researchers do not set as their goal to create a comprehensive and universal technology for processing images of scanned documents, but solve specific narrow problems. Pairing many heterogeneous algorithms within the framework of a single technology leads either to a decrease in the overall speed due to high computational complexity, or to a decrease in the quality of image processing due to the weak consistency of individual algorithms.

In this regard, the urgent problem is the development and coordination of algorithms for the analysis and processing of images of text documents. This paper proposes an algorithm for extracting textual information based on physical signs of reflection. This allows you to more accurately select areas of the image that contain the ink with which the document was printed. The developed methodology for processing text documents, allowing achieving high classification accuracy and processing speed.

2. Proposed methodology

In the tasks of analysing documents, the problem arises of obtaining images of this document in various spectral ranges in order to reveal the specifics of the reflection properties of paper and ink. The optical properties of paper and ink make it possible to determine with high accuracy their belonging to certain classes. To solve the problem, it is necessary to systematize the stages of the study associated with the processing and analysis of the results, as well as the preparation of an expert opinion.

Application of the systematic approach consists in the fact that a paper document (paper medium) converted into digital form in various spectral ranges (ultraviolet, visible (blue, green, red), near infrared, infrared) allows a multi-aspect research approach to be applied using a technical system. The need for such a volume of data and a set of methods that lead to obtaining results that allow you to get the amount of information necessary for analysing a document. Unlike existing approaches, we propose to obtain this data set and apply processing methods consistent with measurements in a single complex. Existing approaches to the analysis of documents, as a rule, are examined sequentially, from one method to another, depending on the existing set of technical means and the availability of a specific methodological base. This creates a lot of variability of examinations of the same document. Digital doubles being created allow improving the quality of recognition due to the processing of components in a single plan, taking into account the physical characteristics of the reflection of paper and ink [4-6]. A known fact is the peculiarity of the reflection of paper and ink depending on the portion of the spectrum range. Taking this fact into account allows us to assume that the created set of data on the reflection of a document (paper and ink) in different parts of the spectrum has broader analysis capabilities than just an RGB image or a received image in a different range of the spectrum.

Analysis of paper reflection properties can be explained by the image formation formula:

$$I(\lambda)=Io(\lambda)\cdot cos(\varphi)\cdot \rho(\lambda)\cdot W, \tag{1}$$

Here $I(\lambda)$ is the brightness of the measured image pixels ($R= I(\lambda_3)\sim 600\ nm,\ G= I(\lambda_2)\sim 500\ nm,\ B= I(\lambda_1)\sim 400\ nm$), $Io(\lambda)$ is the brightness of the radiation source, $cos(\varphi)$ is the angle of illumination, $\rho(\lambda)$ is the spectral reflection coefficient, and W is the paper roughness. Note that expression (1) does not take into account the features of the measuring technique, since it is shown only for the specifics of the proposed approach.

From additional studies, we can obtain information on the properties of paper $\rho(\lambda)$ and W. Thus, having an unknown type of paper, but knowing the lighting characteristics, we can obtain $\rho(\lambda)\cdot$and compare this product with known data. Selecting sections of a document with textual information, by performing similar manipulations, you can get similar information.

Digital copies of text documents created in various areas of the spectrum, as well as for a given angle of illumination and brightness, can be obtained using various optical filters and radiation sources [7]. Naturally, knowing the features of the reflective properties of paper and ink, from a comparative analysis of the data of digital twins, it is possible to obtain information about the type of paper and ink with high accuracy. This method is intended for full and effective research, analysis, and for recognition of text in documents, images, identifying authenticity, identification, authentication, methods of creation, history of making changes to documents and images.

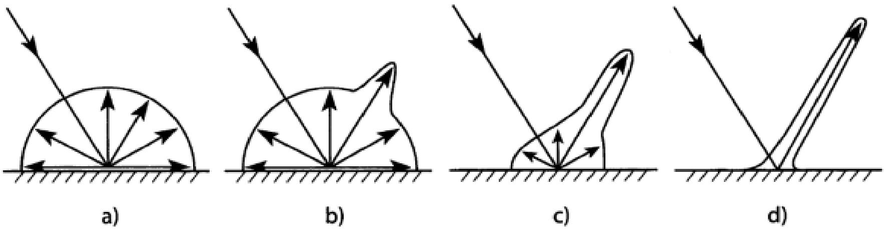

Figure 1. Various types of reflection from the paper (a – lambert reflection, b – quasi-Lambert reflection, c – quasi-specular reflection and d – specular reflection).

Note that any image is formed from the physical laws of reflection of radiation, for example, the visible spectral range, from the surface of the paper. In this case, depending on the quality of the paper, various types of reflection appear. The figure 1 shows how the reflection of visible light depends on the roughness W of the paper (a - very rough paper and d - smooth glossy paper).

Further, any paper is illuminated by a light source and the reflected radiation enters the receiving lens of a digital camera and is fixed to the photo matrix in RGB format. In figure 2 shows the spectral patterns of reflection of different types of paper, and shown that, depending on the type of paper, there is a fairly significant difference in the magnitudes of reflection (between white and black). You can also notice that in the visible region of the spectrum, there is a significant increase in the reflection of green and red, with respect to blue. Figure 2 shows the reflection coefficients of different types of paper (white and black).

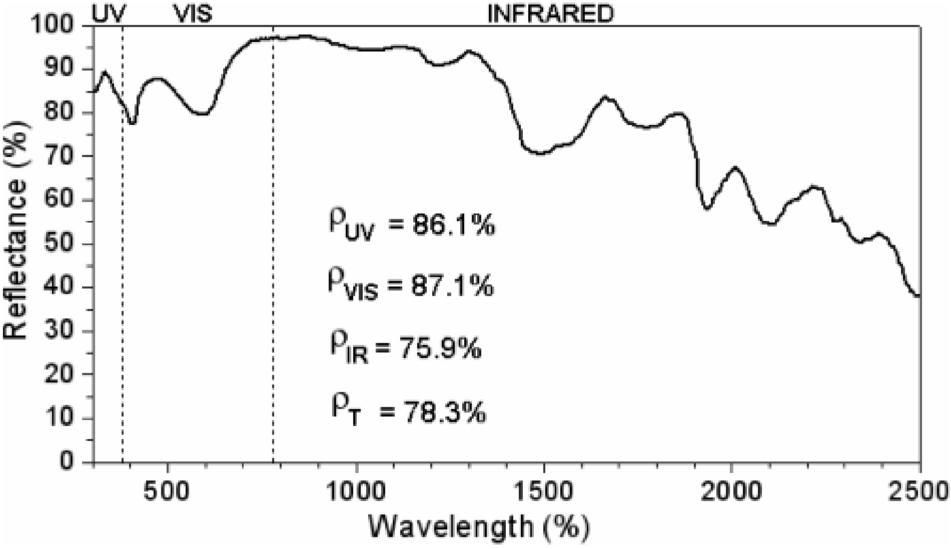

Figure 2. Reflection coefficients of white paper [1] (ρ_{UV} - ultraviolet reflection, ρ_{VIS} - visible reflection, ρ_{IR} - infrared reflection, ρ_T – thermal reflection) [8].

An analysis of Figure 2 shows that in the visible region of the spectrum (350-650 nm) a reflection coefficient is measured by about 10%, with an average reflection coefficient of 0.85%.

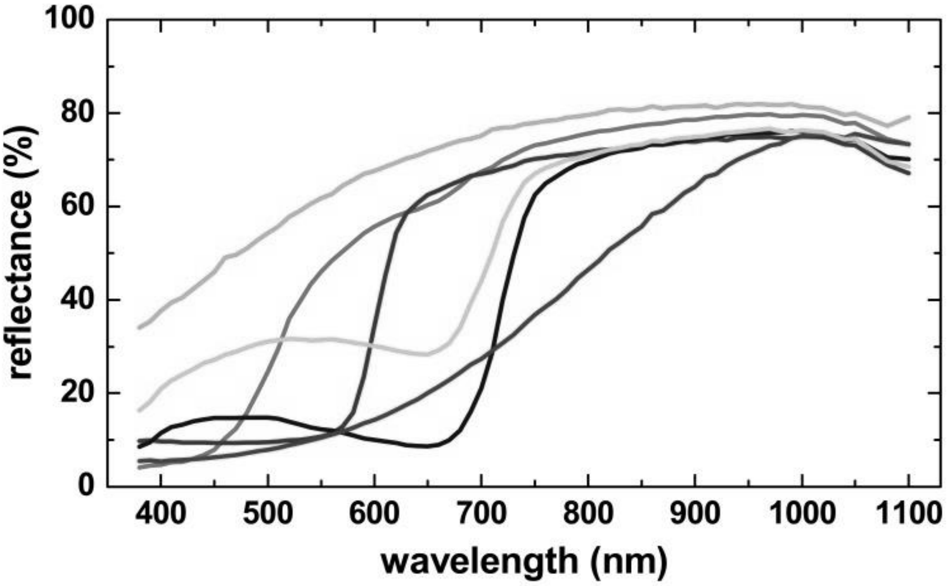

Figure 3. Reflection coefficients of different types of inks [9].

Figure 3 is provided only for a qualitative comparison of the reflectivity of various types of inks and clarification of the main spectral regions of absorption and reflection of visible light.

3. Results

From the figures 2, 3 it is seen that the main difference in the reflection of paper and ink is observed in the visible region of 350-650 nm. Thus, using this dependence, it is possible to detect the presence of ink on paper on separate channels R (550-650 nm), G (450-550 nm), and B (350-450 nm) for each pixel. In principle, the red channel makes it possible to identify individual types of ink. Thus, the spectral characteristics of each channel of a digital camera make it possible to unambiguously identify areas of the document where only paper is present and where there is text (ink). From figures 2 and 3, you can build an algorithm that allows you to clearly separate two objects of a text document - paper and text, including signature and printing. To do this, you need to translate the RGB image in grayscale, for example, using an expression and then using formula (1), calculate the reflection coefficient for each pixel of the document image. So, for paper, the reflection coefficient will be above 75% (see Fig. 2), and for ink below 50% (see Fig. 3). Next, a binary mask is constructed either for each pixel or for blocks of pixels (for example, 3x3): paper - 0, ink - 1. This allows you to identify all parts of the document where text, signature or print is present.

4. Conclusion

This paper shows the ability to divide a document into two areas, one of which contains only paper and the other ink. This is possible due to the use of paper reflection properties (high, 80-90% reflection in all three channels R, G, B) and ink (low reflection (10-40%) for channels B, G and higher (30-60 %) for the R channel). Comparison of the reflection value when using formula (1) allows you to accurately split the document into two parts. Using the spectral properties of the reflection of ink, it also appears to identify the type of some ink. Thus, the proposed technique can be used at the stage of preliminary processing of the document, to highlight areas of the text. The study of the geometric properties of letters is next to the task at hand. The article discusses the use of the spectral properties of paper and ink to classify the image of a text document, possibly containing, in addition to text, a signature or print. The proposed approach has the ability to detect portions of a document with text in good lighting conditions, close to a perpendicular drop on a paper document, upon receipt of the image. Possible changes in the illumination angle of the document during image acquisition were not considered in this article and are the subject of further research.

References

[1] Donaldson K and Myers G K 2005 Bayesian Super Resolution of Text in Video with Text-Specific Bimodal Prior, *International Journal on Document Analysis and Recognition* **7** 159–167

[2] Hase H, Yoneda M, Tokai Sh, Kato J and Ching Y S 2004 Color Segmentation for Text Extraction” International *Journal on Document Analysis and Recognition* 271–284

[3] Kim K I, Jung K, Park S H and Kim H J 2001 Support Vector Machine-based Text Detection *Digital* Video *Pattern Recognition Letters* **34(2)** 527–529

[4] Farnood R 2009 Review: Optical properties of paper: theory and practice *Advances in Pulp and Paper Research*, Oxford, pp 273–352.

[5] Zhao Y and Berns R S 2007 Image-based spectral reflectance reconstruction using the matrix r method *Color. Res. & Appl.* **32** 343–351

[6] Hebert M and Hersch R D 2004 Classical Print Reflection Models: A Radiometric Approach *J. Imaging Sci. Techonol.* **48** 363-374

[7] Neugebauer H E J 2005 The theoretical basis of multicolor letterpress printing *Color Res. Appl.* **30** 322–331

[8] Dornelles K and Roriz M 2006 A Method to Identify the Solar Absorptance of Opaque Surfaces with a Low-cost Spectrometer” PLEA2006 - The 23rd Conference on Passive and Low Energy Architecture, Geneva, Switzerland, 6-8 September, pp 17-23

[9] Klein M E, Aalderink B J, Padoan P, de Bruin G and Steemers A G 2008 Quantitative Hyperspectral Reflectance Imaging *Sensors* **8** 5576-18