IMAGE ANALYSIS OF OLD GLAGOLITIC BOOKS BASED ON PHOTOGRAPHIC DIGITALIZATION

SLIKOVNA ANALIZA STARIH GLAGOLJSKIH KNJIGA BAZIRANA NA FOTOGRAFSKOJ DIGITALIZACIJI

Damir Modrić*, Danijel Radošević**
*Faculty of Graphic Arts, Zagreb, Croatia
**Faculty of Organisation and Information, Varaždin,Croatia

Sažetak: Područje glagoljičkog pisma i prvih knjiga tiskanih glagoljicom je tehnički i tehnološki slabo istraženo. Ovaj rad ispituje mogućnost tehničke podrške (ili eventualnu negaciju), povijesnog znanja i pretpostavki o počecima tiskarstva u Hrvatskoj pomoću fotografske digitalizacije i analize digitalnih zapisa programima „Maglica“ i ImageJ. Program "Maglica" izrađen je za potrebe ovog istraživanja i služi za statističku obradu dobivenih slika fotografskom digitalizacijom. Digitalni zapisi pojedinih stranica knjiga i njihovih pseudo3D prikazi ukazuju na to da se prikazanom metodom može odrediti karakteristični otisak - "fingerprint" tiskarske preše te da se na temelju ove metode određena knjiga može „pridružiti“ određenoj tiskarskoj preši ili tiskari.

Ključne riječi: *fotografska digitalizacija, pseudo3D, glagoljičke knjige, tiskarska preša*

Abstact: The area of Glagolitic script and the first printed books in Glagolitic has been technically and technologically poorly explored. The possibility of technical support (or possibly a denial), by photographic digitization and analysis of the digital records with so called “Vapour” and ImageJ program, of historical knowledge and assumptions about the origins of the printing in Croatia is presented in the paper. A computer program "Vapour" has been developed for the purposes of this study and used for statistical processing of digital records. Digital records of individual pages of the books and their pseudo-3D outlook indicate that the presented method can identify the typical impression – ‘fingerprint’ of the printing presses and, based on this method, a particular book can be “assigned” to a particular printing press or printing house.

Key words: *photographic digitization, pseudo3D, glagolitic books, printing press*

Introduction

In the late 1980's, the prevalence of fast computers, large computer memory, and inexpensive scanners fostered an increasing interest in document image analysis. With many paper documents being sent and received via fax machines and being stored digitally in large document databases, the interest grew to do more with these images than simply view and print them. Just as humans extract information from these images, research was performed and commercial systems built to read text on a page, to find fields on a form, and to locate lines and symbols on a diagram. Today, the results of research work in document processing and optical character recognition (OCR) can be seen and felt every day. OCR is used by the post offices to automatically route mail. Engineering diagrams are extracted from paper for computer storage and modification. Handheld computers recognize symbols and handwriting for use in niche markets such as inventory control. In the future, applications such as these will be improved, and other document applications will be added. For instance, the millions of old paper volumes now in libraries will be replaced by computer files of page images that can be searched for content and accessed by many people at the same time — and will never be mis-shelved. Business people will carry their file cabinets in their portable computers, and paper copies of new product literature, receipts, or other random notes
will be instantly filed and accessed in the computer. Signatures will be analyzed by the computer for verification and security access.

This paper describes some of the technical methods and systems used for document processing of text and graphics images. The methods have grown out of the fields of digital signal processing, digital image processing, and pattern recognition. The objective is to give an understanding of what approaches are used for application to documents. Since the field of document processing is relatively new, it is also dynamic, so current methods have room for improvement, and innovations are still being made. In addition, there are rarely definitive techniques for all cases of a certain problem. The intended audience is executives, managers, and other decision makers whose business requires some acquaintance or understanding of document processing. (We call this group "executives" in accordance with the *Executive Briefing* series.)

A large number of challenges is associated with the digitization of old books. They should be handled with immense care because they are delicate and irreplaceable. Until now, old books and documents are mostly photographed because existing methods scanning causes severe damage to them. Processing of digitized images of old books and historical documents requires special algorithms must take into account the noise, and possible effects of splitting and cutting of the paper. But the image analysis can potentially provide some information about the document especially in the case of old documents for which is not known technical-technological framework in which it originated. In this paper, we focused on the possibility of obtaining some data related to the printing surface using image analysis.

Collections of ancient and historical documents available in libraries around the world are of great cultural and scientific importance. Transformation of documents into digital format is necessary to maintain the quality of the original and provide scientists full access to this information. It is normal for such documents to suffer from the problem of degradation[1]. Let us mention only a few of such as the presence of spots, deformation, large variations in the background, smearing ink, etc. These are factors that interfere with (in many cases can also disable) the readability of documents. Therefore, appropriate methods should be developed to remove noise from images of historical documents and improve their quality before they are exposed to public in libraries. In this context, noise is considered to be all that is not tied to the textual or pictorial information in document images. In contrast to the image processing, most image analysis procedures attempts to draw only "important" information from an image such as to identify and count the features in the image, reducing the amount of data, with perhaps one million bytes to several tens.

The beginnings of printing in Croatia

The establishment of the first Croatian printing house attributed Blaž Baromić, who was born about 1450. in Vrbnik at island Krk in Croatia. Blaž Baromić for the needs of the Republic of Dubrovnik, using Glagolitic alphabet Glagolitic breviary wrote and it is assumed that, because of the duration and complexity of this work, became interested in the possibility of publishing a books. It is believed that because of this Blaž Baromić went to where the Senj Glagolitic alphabet was in official use and in which there could be a cultural and financial interest in the foundation of Glagolitic printing house[2].

It is also assumed that, in order to master the craft of printing, he went to Venice where the famous typographer and printer Andreas Torresania Glagolitic participated in drafting a set of typographic and in printing of the first printed Glagolitic breviar[3] 1493. It is also assumed that this is just a set of typographic issued in Senj, where he founded a printing house in which he 1494. printed in Senj Glagolitic Missal. This printing house is, with certainty, attributed to press six more books: Spovid općena (1496.), Mirakuli slavne dive Marije (1507.), Naručnik plebanušev (1507.), Meštrija od dobra umrtija s Ritualom (1507.), Tranzit svetog Jerolima (1508.) and Korizmenjak (1508.)[4] . Today, it is assumed that all these books were printed with typographical set brought from Venice.

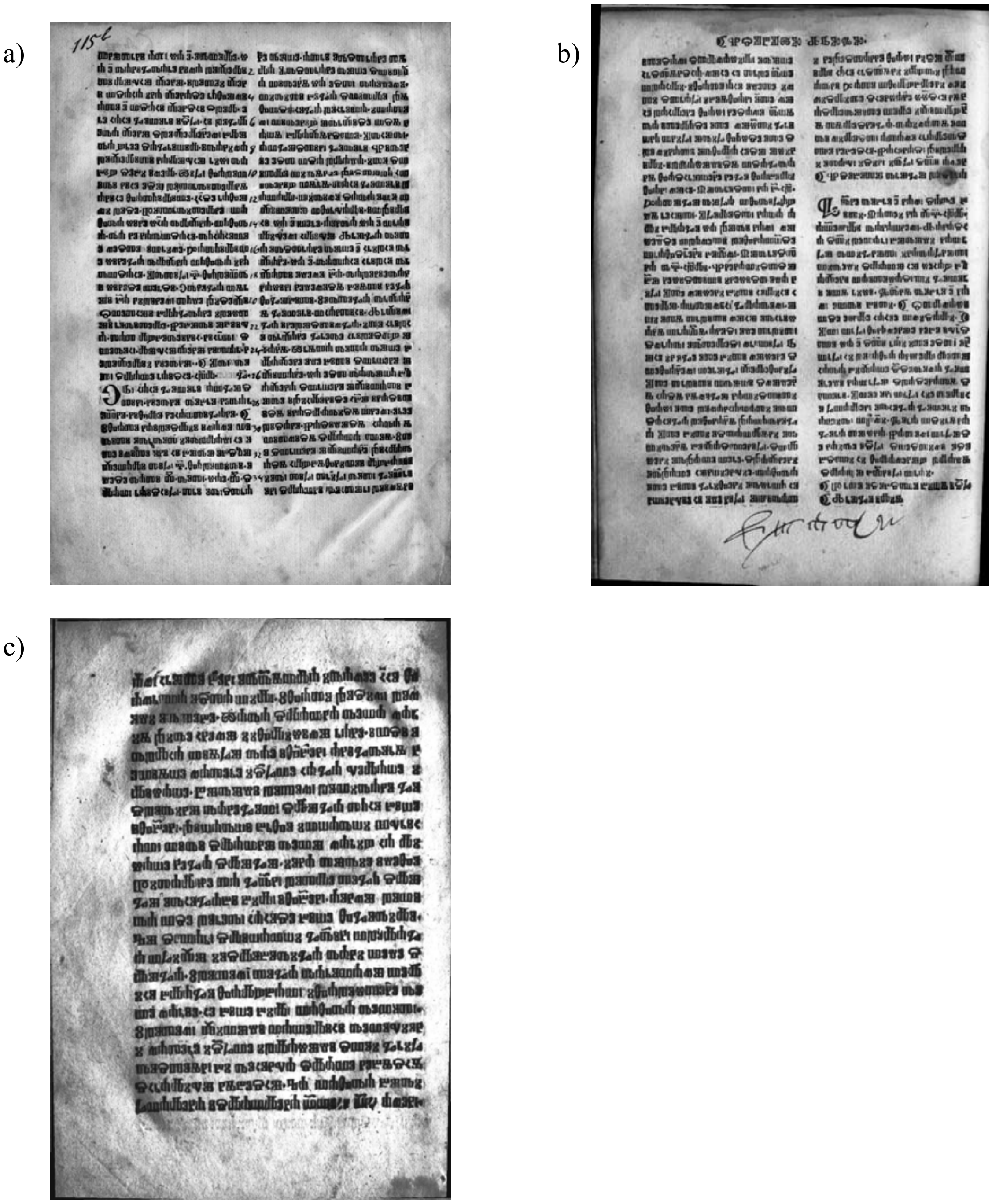


Figure 1. Examples of Glagolitic books - a) Naručnik plebanušev; b) Korizmenjak; c) Mirakuli slavne dive Marije;

Digitization of books, recording and analysis of the digital records

Throughout this work is to investigate the possibility of technical and technological support to (or eventually denial of) these assumptions. As these old books are very valuable and sensitive the direct tests are virtually impossible and the first step in studying is their digitization.

To determine the characteristic fingerprint of printing press, for the analysis of the obtained digital recording, we create a program that we called "Vapor" and made additional image analysis using ImageJ software and pseudo3D display.

Our idea was to use the photo digitizing, considering that due to the sensitivity of the available books too invasive scanning methods. When taking these motifs needed illumination uniform of the entire recorded object, and to the optical characteristics of the photographic apparatus as less impact on the results of the analysis, it is necessary to work with balanced use of the lens aperture that will provide the same sharpness across the entire surface of the object (a book page). Image sharpness is directly related with the ability to resolution line (horizontal and vertical), which assumes the use of optical-quality lenses and the use of photographic apparatus with high-resolution sensor. In order to obtain as complete photographing information, the optimal use of the RAW format records the great depth of color, and thus, potentially, a large dynamic range[5]. For the above reasons is need to record with the relatively low sensitivity (up to 100/21 ISO). For the analysis with ImageJ program, RAW images must be processed and transferred to a JPEG or TIFF file.

Unfortunately access to our Glagolitic books is extremely complicated because it involves a lot of various permits from the centers and libraries, conservators, etc.. and although there is understanding and support of responsible persons in this field, until now books could not be accessed for digitalization.

After document input by digital scanning, pixel processing is performed. This level of processing includes operations that are applied to all image pixels. These include noise removal, image enhancement, and segmentation of image components into text and graphics (lines and symbols). The objective of document image analysis is to recognize the text and graphics components in images, and to extract the intended information.

The idea is to determine the fingerprint of printing in which the book was printed. The basic idea is that on every page of the book we have in fact two pieces of information. The first refers to the written text, pictures, etc. and this information is easily accessible and understandable, but differs from page to page. The other information common to all pages and gives the press, printer, and his skill, used ink and many other factors that affect the printing process itself. Since we did not have enough images to desired Glagolitic books a method that we develop we applied the following books:

- Carthusiensis Adrianus De remediis utriusque tempests printed in Cologne; Ulrich Zel, about 1470[6]
- Albertus Magnus: Sermo de tempore et de Sanctis; Cologne: Arnold Ther Hoernen, 24 Dec. 1474[7]
- Antoninus Florentinus: Confessionale: Defecerunt scrutantes scrutinio, Venice: Peregrinus de Pasqualibus, Bononiensis Bertochus and Dionysius, 25 Oct. 1484[8]

The books were taken because they made the same historical period as Baromić, so we assumed that the same is applied printing technology.

"Vapor" extraction

The idea in the base of the proposed method is to compare the set of optically extracted characteristics of observed book in order to compare them with the corresponding characteristics of other books. Results of such comparison could help in finding the origin of observed book, like printing house and approximate time period of printing.

Characteristics to be extracted

An optically scanned book consists of image files representing particular pages. Each page contains background (predominantly white, but, important for the proposed method, not entirely white) and symbols (letters, digits, punctuation; predominantly, but not entirely black). According to threshold value, it's possible to separate background from the symbols and to normalize page into a small figure, colloquially called as a "vapor". The experiments have been conducted with the "vapors" of 100·100 and 300·300 pixels. All the "vapors" from the same book result in an average "vapor", some kind of the book's "fingerprint". Actually, there currently two "fingerprints" of the book, one given from page backgrounds (colloquially called as "white vapor"), and another given from the symbols (colloquially called as "black vapor"), as shown in Figure 1.

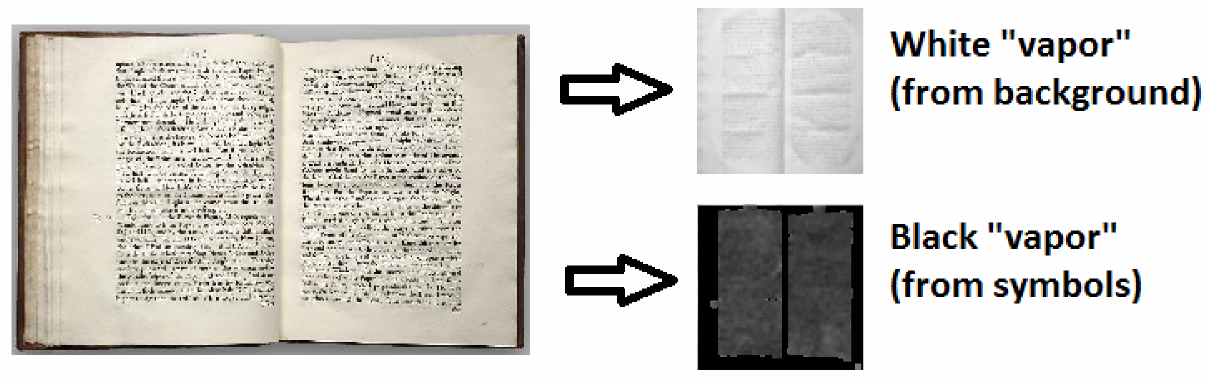
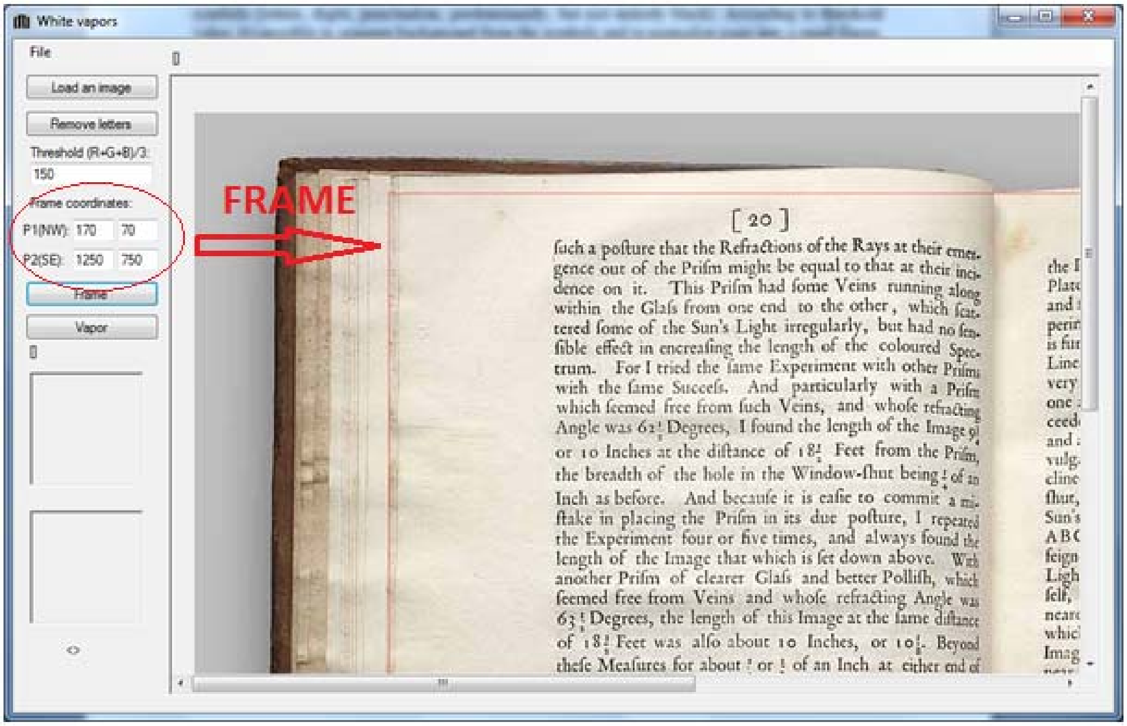


Figure 2: Scanned page from the book and appropriate "vapors"

Extraction of white "vapors"

In the first step, a frame for processing have to be specified (Figure 3) in order to avoid the impact of image parts outside of the observed page, like sheath of the book. The frame stays the same for all pages to be processed. It's important that frame content includes both symbols and page background.



Threshold (R+G+B)/3: 150

Frame coordinates:

P1(NW): 170 70

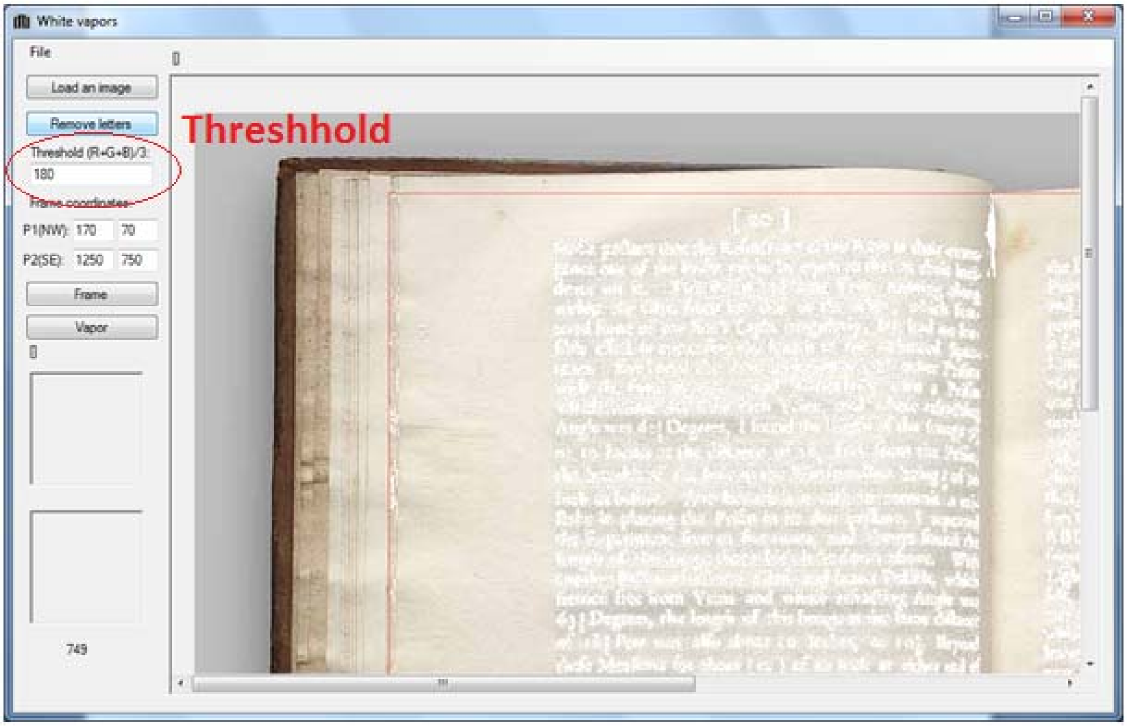
P2(SE): 1250 750

[20]

ſuch a poſture that the Refractions of the Rays at their emergence out of the Priſm might be equal to that at their incidence on it. This Priſm had ſome Veins running along within the Glaſs from one end to the other, which ſcattered ſome of the Sun's Light irregularly, but had no ſenſible effect in encreaſing the length of the coloured Spectrum. For I tried the ſame Experiment with other Priſms with the ſame Succeſs. And particularly with a Priſm which ſeemed free from ſuch Veins, and whoſe refracting Angle was 62½ Degrees, I found the length of the Image 9¾ or 10 Inches at the diſtance of 18½ Feet from the Priſm, the breadth of the hole in the Window-ſhut being ¼ of an Inch as before. And becauſe it is eaſie to commit a miſtake in placing the Priſm in its due poſture, I repeated the Experiment four or five times, and always found the length of the Image that which is ſet down above. With another Priſm of clearer Glaſs and better Polliſh, which ſeemed free from Veins and whoſe refracting Angle was 63½ Degrees, the length of this Image at the ſame diſtance of 18½ Feet was alſo about 10 Inches, or 10⅛. Beyond theſe Meaſures for about ¼ or ⅓ of an Inch at either end of

Figure 3: A frame for processing

For the purpose of extracting white "vapors", the impact of symbols has to be minimized. Symbols are to be eliminated according to the threshold value. The threshold value represents the average intensity of three basic colors (red, green and blue; each in the range 0-255). All pixels below the threshold are being replaced by white color in the process of removing, as shown in **Figure 4**.



Threshold (R+G+B)/3: 180

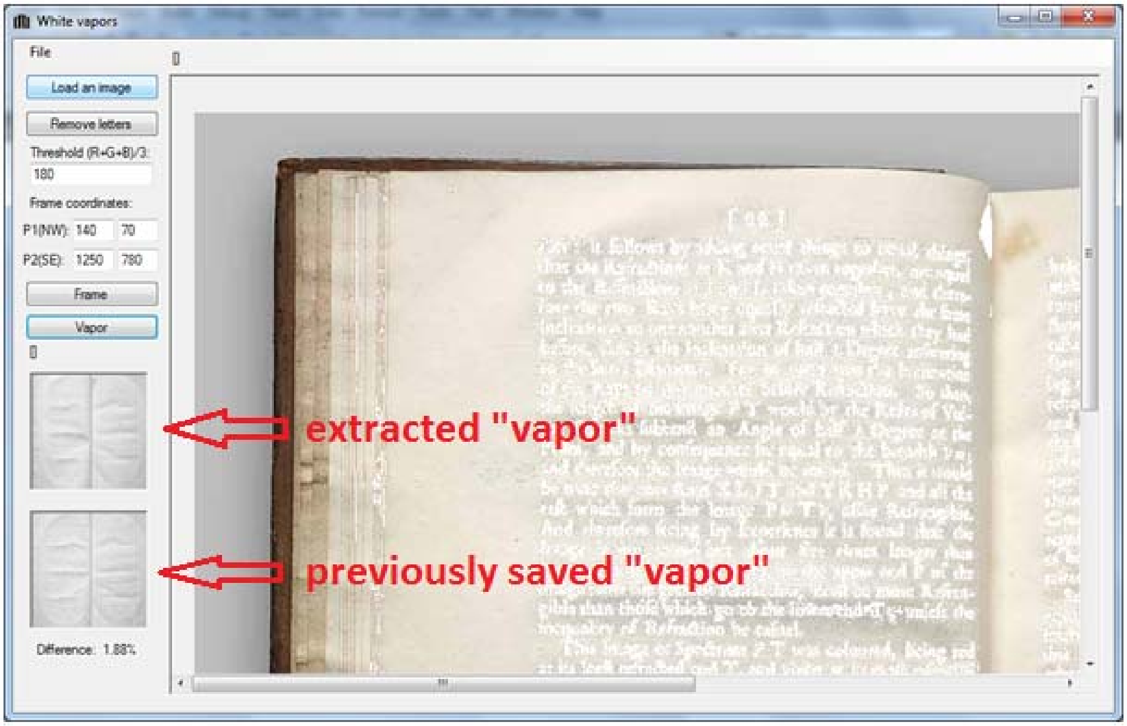
Frame coordinates

P1(NW): 170 70

P2(SE): 1250 750

Figure 4: Elimination of symbols, according to threshold value

The idea in the base of the extraction system is to resize the frame containing scanned page into normalized frame (e.g. 100·100 pixels). The resizing process should additionally reduce the impact of former symbols.



Threshold (R+G+B)/3: 180

Frame coordinates:

P1(NW): 140 70

P2(SE): 1250 780

Difference: 1.88%

Figure 5: Extracted white "vapor"

The intensity (0..255; gray value) of each pixel in the "vapor" is formed as an average value of appropriate field of pixels inside the frame (Figure 6) using these several rules:

- the size of each field to calculate an average value has the height (H) and width (W), where:

$$H = \frac{H_1}{H_2} \quad \text{and} \quad W = \frac{W_1}{W_2}$$

- both values are rounded on a higher integer value
- the average value of field gives the RGB intensity level (0-255) of a particular pixel of the vapour. It's important that this value is calculated only from non-white pixels, in order to reduce the impact of (former) symbols that has been removed.

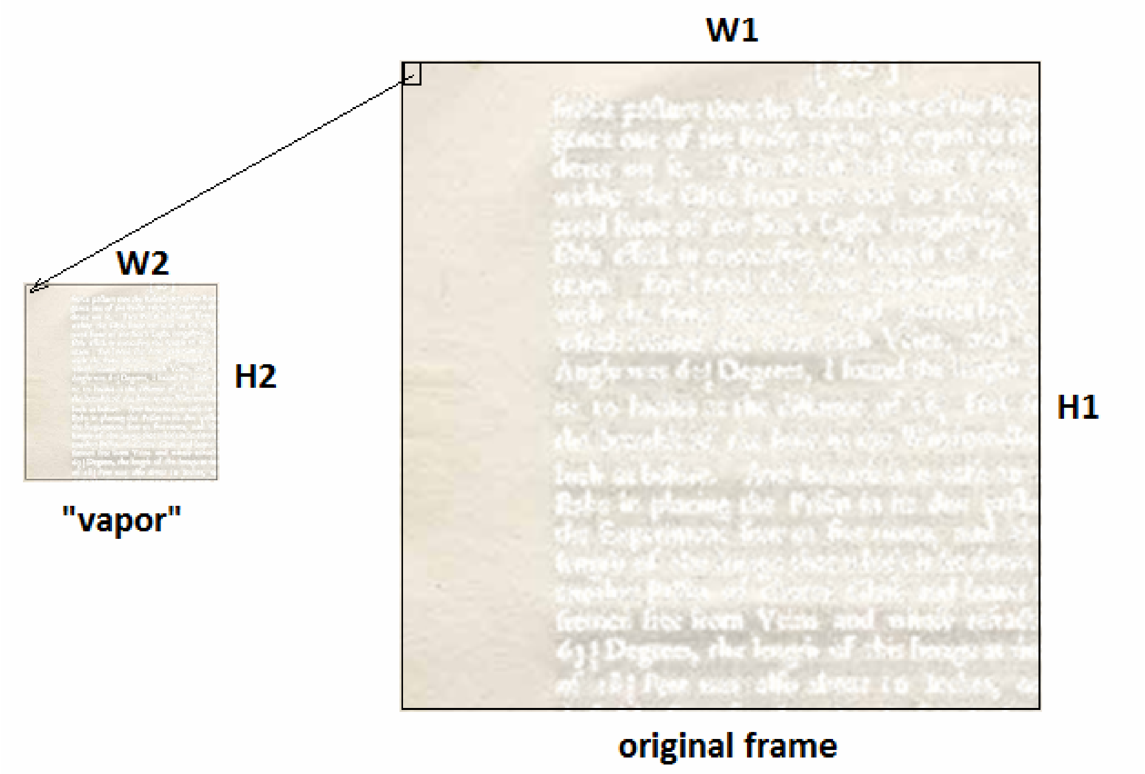
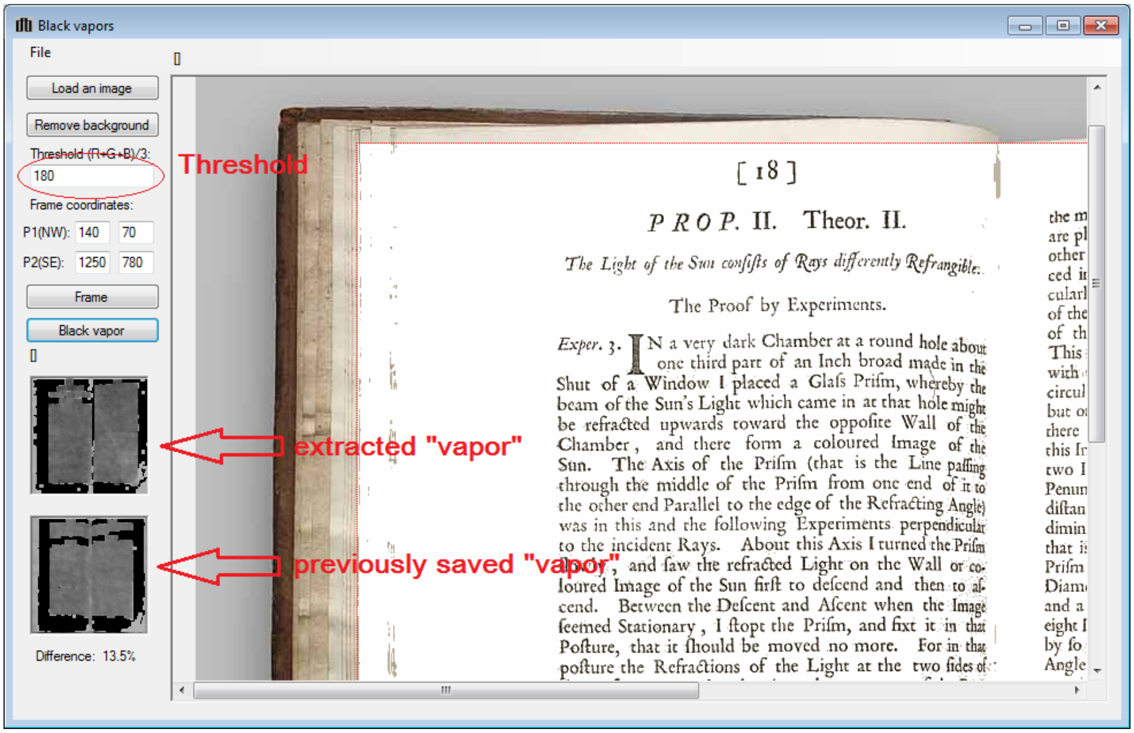


Figure 6: Vapor extraction process

Extraction of black "vapors"

The process of black "vapors" extraction is complementary to the process of white "vapors" extraction. Instead of page background, the goal is to extract the intensity of black ink over the page. So the process, after defining frame for processing (as described previously), has to eliminate the impact of page background, according to threshold value, as shown in Figure 7.

The intensity (0..255; gray value) of each pixel in the "vapor" is formed in an exactly the same manner as in case of white "vapors" extraction. The result of this process is that the page background is represented by black color, and symbols as different gray levels on the black "vapor".



Threshold (R+G+B)/3: 180

Frame coordinates:

P1(NW):	140	70
P2(SE):	1250	780

Difference: 13.5%

[18]

P R O P. II. Theor. II.

The Light of the Sun conſiſts of Rays differently Refrangible.

The Proof by Experiments.

Exper. 3. IN a very dark Chamber at a round hole about one third part of an Inch broad made in the Shut of a Window I placed a Glaſs Priſm, whereby the beam of the Sun's Light which came in at that hole might be refracted upwards toward the oppoſite Wall of the Chamber, and there form a coloured Image of the Sun. The Axis of the Priſm (that is the Line paſſing through the middle of the Priſm from one end of it to the other end Parallel to the edge of the Refracting Angle) was in this and the following Experiments perpendicular to the incident Rays. About this Axis I turned the Priſm [illegible], and ſaw the refracted Light on the Wall or coloured Image of the Sun firſt to deſcend and then to aſcend. Between the Deſcent and Aſcent when the Image ſeemed Stationary, I ſtopt the Priſm, and fixt it in that Poſture, that it ſhould be moved no more. For in that poſture the Refractions of the Light at the two ſides of

Figure 7: Elimination of background, according to threshold value

The average "vapor"

The average "vapor" contains average pixel intensity level for a group of "vapors" (white or black; representing e.g. some book). It represents a "fingerprint" of the observed group of pages. Examples of average "vapors" are given in Fig. 7.

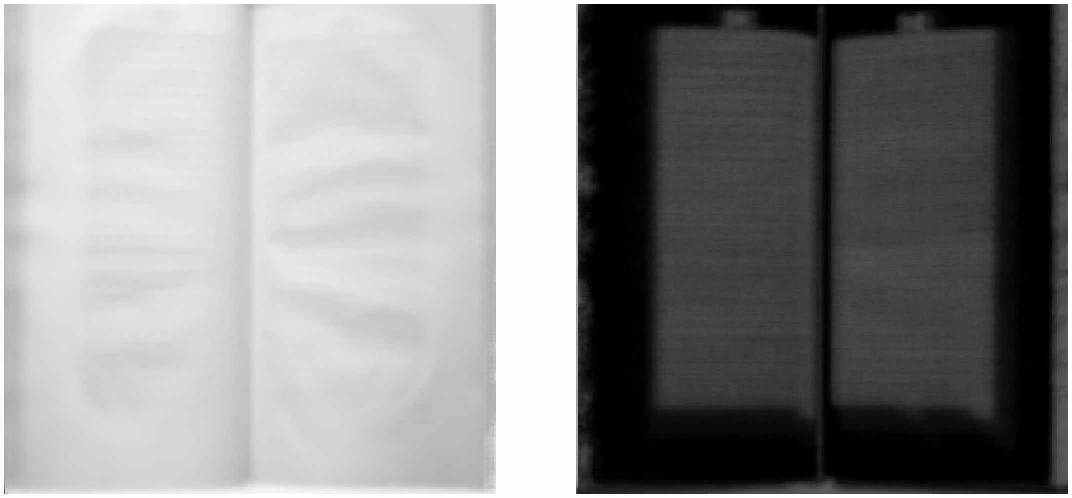


Figure 8: Average "vapors"

Results

Before analysis it was necessary to divide the digitized pages on odd and even ones. This was done because the sheets are first printed on one then the other side. These pages definitely differ among themselves because in the second case, the paper is no longer intact. We assumed that, during the first printing, press left some kind of trace upon it. Examples of averaging of white and black "vapor" are shown in Figure 9. Pseudo 3D display indicates that there are some similarities, but also that it takes much more effort on getting the quality of conclusions. It is certainly necessary to do more computing experiments to optimize the parameters of the "Vapor".

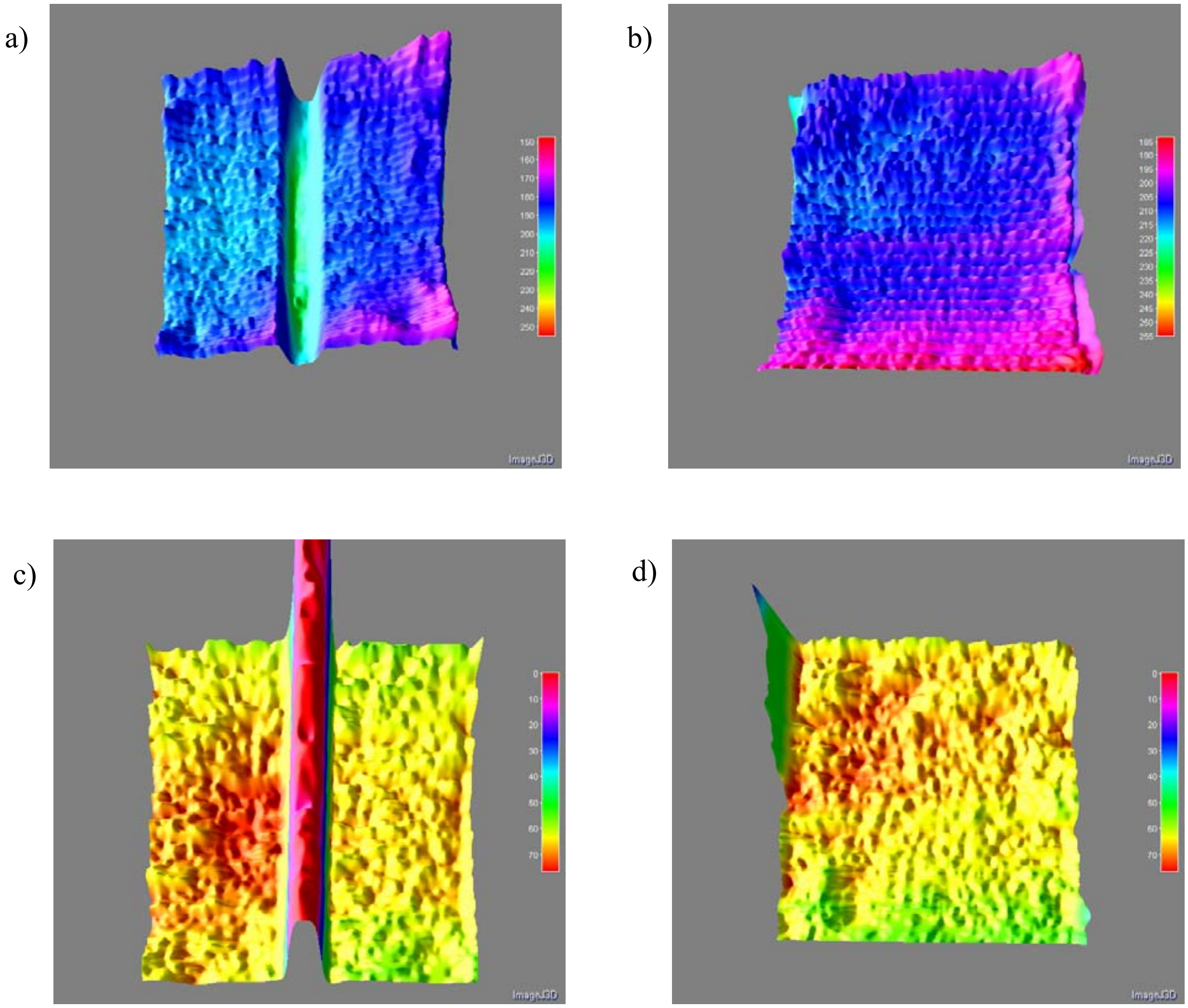


Figure 9: Examples of pseudo 3D images o averaged even white “vapor” of book a) K1 and b) K2 and even black “vapor” of book c) K1 and d) K2. (K1 - Carthusiensis Adrianus De remediis utriusque tempests printed in Cologne; Ulrich Zel, about 1470; K2 - Albertus Magnus: Sermo de tempore et de Sanctis; Cologne: Arnold Ther Hoernen, 24 Dec. 1474)

These two books are taken because they come from the same printing house, so we assumed it would contain some common features that will give us a fingerprint of that printing house.

Somewhat improved representation is given with contours, which indicate that there are some common features of these books are, as we assume, a consequence of the printing house, which printed the master, his skills, presses, and many other factors. These books are more than five hundred years to and time, their storage and use over time etc. affected that the requested information elude.

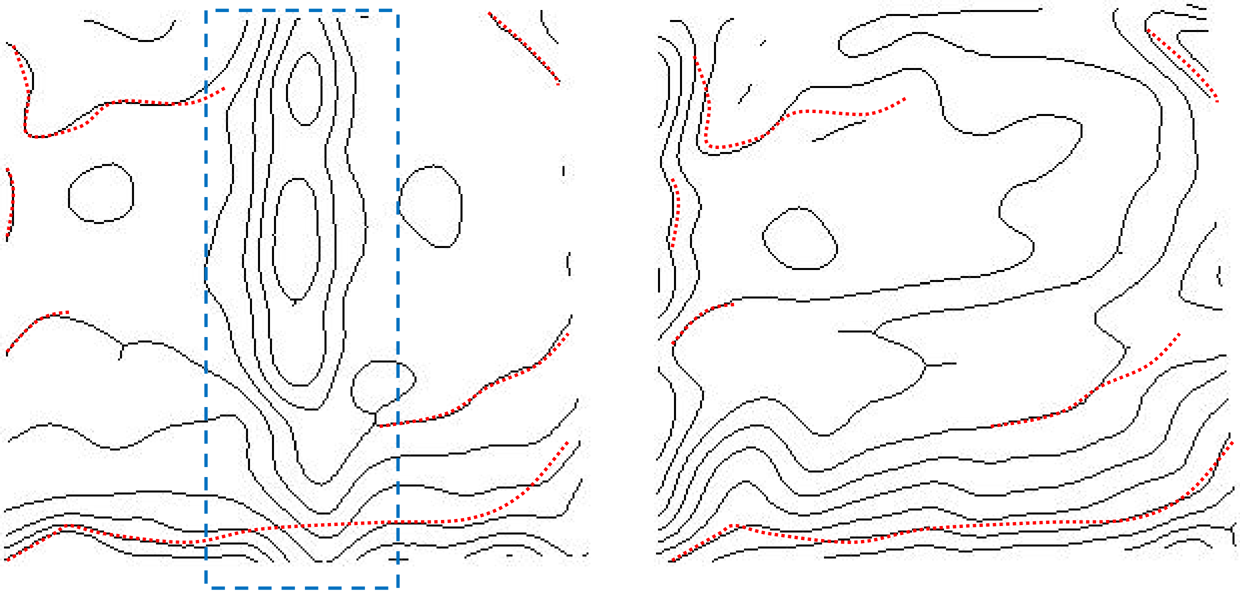


Figure 10: On figures a) K1 and b) K2 representation of white "vapors" are given with contours. Red doted lines represent observed similarities in both books.

To obtain average "Vapors" we took 100 odd and even pages. This gave us the well averaging, and deviations from the mean of each page "vapor" are between 1-3%. This indicates that the site does not differ too much if averaged as shown above.

Conclusion

The idea of this work was using non-destructive method to get information that will enable us to more easily locate in time and place a book for which this data is unknown. The results are inconclusive, although still showing that we are on the right track and that should be perform more experiments that would give the desired result. But even then, we will be unable to say with certainty whether the result is exactly fingerprint of a particular printing house.
On the other hand, we assume that such research will help to increase the technical and technological knowledge of the procedures used during the printing in the 15th and 16th century. At that time the book was a product of high technology that required support of many other

professions, such as goldsmiths, manufacturers of paper and ink, tannery, etc. These craftsmen were able to gather only the major centers where there was a need for the written word.

Literature

[1] Baird, H.S. (2004.) "Difficult and Urgent Open Problems in Document Image Analysis for Libraries" DIAL'04, pages (25-32)

[2] http://www.croatianhistory.net/etf/senj4.html Accessed: 3. 6. 2011.

[3] Nazor, A., (2005). Hrvatske glagoljske knjige tiskane u Veneciji u XV. i XVI. stoljeću, *Zbornik radova 9. međunarodnoa savjetovanje tiskarstva, dizajna i grafičkih komunikacija Blaž Baromić*, Bolanča, Z., Mikota, M. (ur.), pp.11-16, Lovran, Grafički fakultet Sveučilišta u Zagrebu, Ogranak Matice hrvatske Senj, Inštitut za celulozo in papir Ljubljana, Zagreb

[4] Nazor, A., (2008)., Glagolitic Books Printed in Senj in 1508, *Zbornik radova 12.. međunarodnog savjetovanae tiskarstva, dizajna i grafičkih komunikacija Blaž Baromić*, Bolanča, Z. (ur.), pp.11-16, Zadar, Grafički fakultet Sveučilišta u Zagrebu, Ogranak Matice hrvatske Senj, University of Ljubljana Faculty of Natural Science and Engineering, Inštitut za celulozo in papir Ljubljana, Zagreb

[5] Mikota, M. (2000.). *Kreacija fotografijom,* VDT Publishing, Zagreb

[6] http://www.europeana.eu/portal/record/09428/566FA8612024A07BD56EDC4A9229031EF658F38B.html ;

[7] http://www.europeana.eu/portal/record/09428/4C670B43B6F33E557D0B10D4554F046E94F5A2E3.html;

[8] http://www.europeana.eu/portal/record/09428/52F1F45CECE168C6350D4A0E7584D92134F2D2C4.html;