STRUCTURAL COMPRESSION OF DOCUMENT IMAGES WITH PDF/A

Sergey Usilin
Innovations and High Technology Department
Moscow Institute of Physics and Technology
9, Institutskii lane, Dolgoprudny, Russia, 141700
usilin.sergey@gmail.com

Dmitry Nikolaev
Institute for Information Transmission Problems,
Russian Academy of Sciences
19, Bolshoy Karetny lane, Moscow, Russia, 127994
dimonstr@iitp.ru

Vassili Postnikov
Institute for System Analysis,
Russian Academy of Sciences
9, 60-letiya Oktyabrya ave., Moscow, Russia, 117312
vassili.postnikov@gmail.com

KEYWORDS

Document image processing, data compression, morphological operations, color saturation

ABSTRACT

This paper describes a new compression algorithm of document images based on separating the text layer from the graphics one on the initial image and compression of each layer by the most suitable common algorithm. Then compressed layers are placed into PDF/A, a standardizated file format for long-term archiving of electronic documents. Using the individual separation algorithm for each type of document makes it possible to save the image to the best advantage. Moreover, the text layer can be processed by an OCR system and the recognized text can also be placed into the same PDF/A file for making it easy to perform cut and paste and text search operations.

INTRODUCTION

Scanning is the most popular method of document conversion to electronic forms nowadays. However, images resulted from high quality scanning have large size and are not effective in electronic archives. Common techniques of image compression are not applicable because documents may contain monochrome text and full-color graphics at the same time. Lossless image compression algorithms which are effective for monochrome texts are ineffective for full-color graphics, while lossy image compression algorithms show good results for color images, but can corrupt text information.

A combined approach can be used to compress document images. The idea of this approach consists in extracting structural blocks in the image, combining these blocks into layers (i.e. image “separation” into textual, graphic and other layers), and compressing each layer using the most appropriate technique.

One of the structural document compression methods is implemented in the DjVu format (Lizardech 2005, Haffner et al. 1998). But in spite of high ratios of the document image compression, the DjVu-format has essential disadvantage: this format is not standardizated therefore its usage for creating electronic archives is complicated. Besides, using the same separation technique for all image types is not always worthwhile and may lead to significant corruption of the document.

This paper proposes a new method of structural compression document images into PDF/A (ISO 2005, Rog 2007), a standardizated file format. The method contains a set of image separation algorithms typical for images of different document types.

FORMAT PDF/A

PDF/A is a file format for long-term archiving of electronic documents. It is based on the PDF Reference Version 1.4 from Adobe System Inc. and is defined by ISO 19005-1:2005. It ensures the documents can be reproduced the exact same way in years to come. All of the information necessary for displaying the document in the same manner every time (all content, fonts, colors and etc.) is embedded in the file. These features of PDF/A make it possible to use it as a major file format in the electronic archives.

STRUCTURAL COMPRESSION ALGORITHM

In this section we describe briefly a scheme of the proposed structural compression algorithm (Figure 1). The algorithm is meant for compression of document images (images of financial documents, books and magazines pages, manuscripts, etc.). For each type of document there exists a unique separation algorithm. Therefore, first of all, it is necessary to determine the type of the source image and choose the appropriate separation algorithm. Using the chosen algorithm we separate the source image into text and graphics layers. According to the method architecture the text layer has only parts of the source image such that correspond to real text information. Hence it is easy to recognize this

Proceedings 24th European Conference on Modelling and
Simulation ©ECMS Andrzej Bargiela, Sayed Azam Ali
David Crowley, Eugène J.H. Kerckhoffs (Editors)
ISBN: 978-0-9564944-0-5 / ISBN: 978-0-9564944-1-2 (CD)

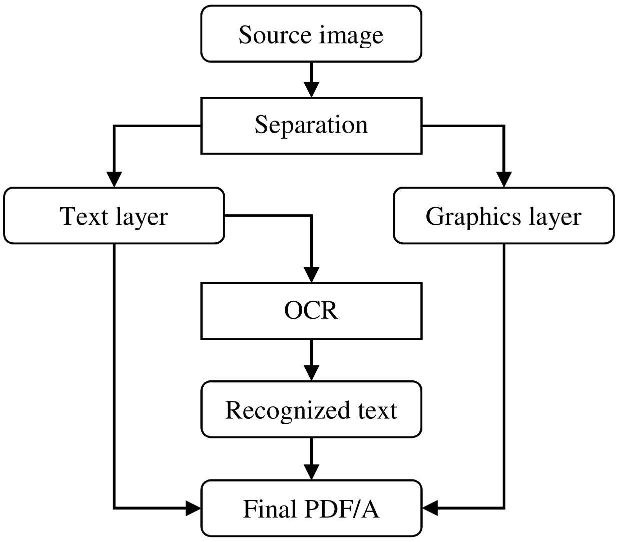


Figure 1. The scheme of the structure compression algorithm

layer by any OCR (optical character recognition) system. The obtained recognized text can be useful for seeking documents in electronic archives. Then both text and graphics layers as well as the recognized text are placed into a PDF/A file in a special way.

Thereby the main parts of the proposed algorithm are the image separation technique and the way of placing obtained layers and recognized text into PDF/A. These parts are described in the following sections.

IMAGE SEPARATION

Besides existence of text information, each document type has its own features. For example, financial documents usually have a lot of stamps and signatures; illustrated magazines can contain a complex multicolor background; books often include formulas, schemes, and diagrams. Therefore the proposed algorithm provides a unique separation method for each type of document. We would like to consider the separation algorithms for two practically important document types: books and financial documents.

An image separation of a common book page

A book page usually contains black text on white background and, possibly, schemes, diagrams, etc. (Figure 2). The areas with text and graphics usually are not intercrossed in the books. The second key feature of the book makeup is usage of the fonts of similar linear sizes.

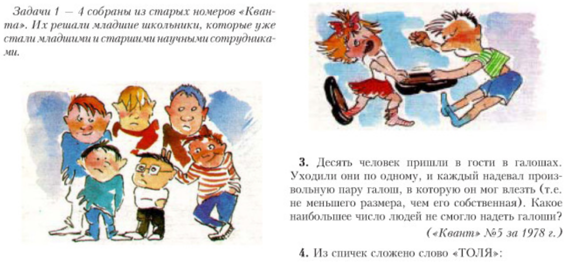
Taking into account these features, we will build the algorithm of book page image separation.

Step 1. Binarize the initial image using one of global binarization methods, for example, Otsu method (Otsu 1979, Nikolaev 2003, Trier and Taxt 1995). As the image contains black text on white background, such binarization should not significantly corrupt text information.

Step 2. Apply mathematical morphology (Gonzalez and Woods 2002, Hasan and Karam 2000) to join each word in the text to connected components (Figure 3). By w and h denote typical width and height of the characters correspondingly. Let us note that the letter spacing is approximately equal to the stem width and the word spacing is comparable with the character width. Therefore words can be joined to connected components by applying mathematical morphology with a rectangular structuring element $w \times 1$.

This operation is not required large computation power because only one simple morphological operation (opening) is needed. Furthermore, using van Herk's algorithm (Van Herk 1992) allows computing any simple morphology operations with a rectangular structuring element for time independent of the size of primitive.

Step 3. Build a heights histogram of the found connected component (Figure 3). As characters in the



Задачи 1 — 4 собраны из старых номеров «Кванта». Их решали младшие школьники, которые уже стали младшими и старшими научными сотрудниками.

3. Десять человек пришли в гости в галошах. Уходили они по одному, и каждый надевал произвольную пару галош, в которую он мог влезть (т.е. не меньшего размера, чем его собственная). Какое наибольшее число людей не смогло надеть галоши?

(«Квант» №5 за 1978 г.)

4. Из спичек сложено слово «ТОЛЯ»:

Figure 2. An example of a common book page image

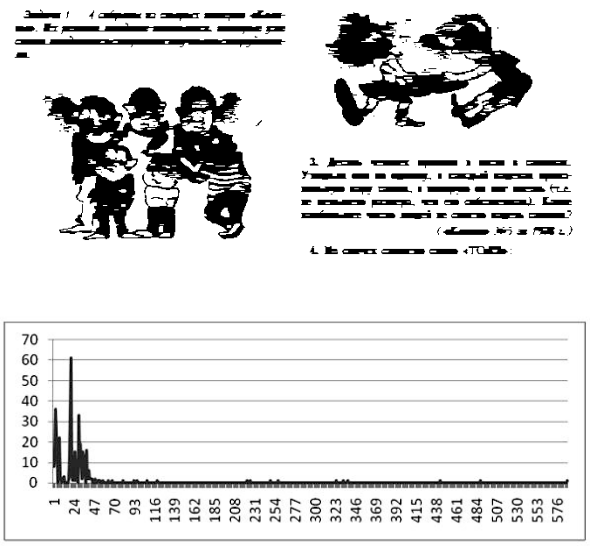


Figure 3.Connected components and a histogram of their heights

text on the image have similar size the connected components corresponding to the words shall be of similar height and create one or more maximums at the histogram (Fletcher and Kasturi 1988). Due to this it is possible to determine typical heights of the letters h_{font} and therefore find the areas of the source image corresponding to text blocks and areas corresponding to graphics and split the source image into the text and the graphics layers.

As we use fast algorithms of mathematical morphology with a rectangular structural element, the text blocks are strongly required to be aligned relatively axes of image coordinates. Therefore we will use the Fast Hough Transform (Nikolaev et al. 2008) to unskew the input image before applying morphological filtrations.

An image separation of a financial document

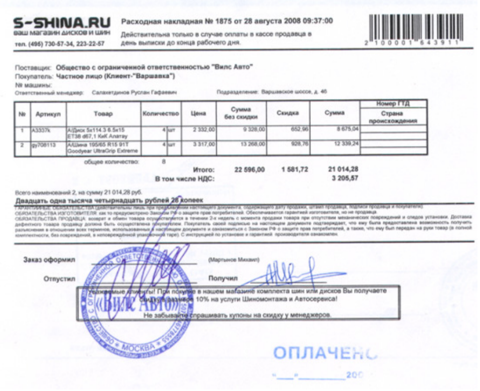
The above features of the book page image are not typical of the color images of financial documents (invoices, receipts, contracts, etc.) because the graphic elements (seals, signatures, manuscript notes) are often laid over the textual blocks (Figure 4). Therefore, it is unreasonable to use the above algorithm. We will build a separation algorithm based on color characteristics of the image. Color saturation of black text and white background is close to zero point, while color saturation of the blue seals and signatures is high. Taking these features into account we will build the following algorithm of separation for financial document image.

Step 1. Calculate color saturation (Gonzalez and Woods 2002) of each pixel on the source image using the formula:

$$s = \max(r, g, b) - \min(r, g, b)$$

Here r, g and b are red, green and blue components of the pixel concerned. It is obvious that s can take value between 0 and 255.

Step 2. Build a saturation histogram $y = \log N_x$, where N_x – number of the pixels whose saturation equals x (Figure 5).



S-SHINA.RU
ваш магазин дисков и шин
тел. (495) 730-57-34, 223-22-57

Расходная накладная № 1875 от 28 августа 2008 09:37:00
Действительна только в случае оплаты в кассе продавца в день выписки до конца рабочего дня.

2 100001 643911

Поставщик: Общество с ограниченной ответственностью "Вилс Авто"
Покупатель: Частное лицо (Клиент-"Варшавка")
№ машины:
Ответственный менеджер: Салахетдинов Руслан Гафаевич
Подразделение: Варшавское шоссе, д. 46

№	Артикул	Товар	Количество		Цена	Сумма без скидки	Скидка	Сумма	Номер ГТД Страна происхождения
1	A3337k	А/Диск 5x114.3 6.5x15 ET38 d67,1 КиК Алатау	4	шт	2 332,00	9 328,00	652,96	8 675,04	
2	gy708113	А/Шина 195/65 R15 91T Goodyear UltraGrip Extreme	4	шт	3 317,00	13 268,00	928,76	12 339,24	

общее количество: 8

Итого: 22 596,00 1 581,72 21 014,28
В том числе НДС: 3 205,57

Всего наименований 2, на сумму 21 014,28 руб.
Двадцать одна тысяча четырнадцать рублей 28 копеек

ГАРАНТИЙНЫЕ ОБЯЗАТЕЛЬСТВА (действительны лишь при предъявлении настоящего документа, содержащего дату продажи, штамп продавца, подписи продавца и покупателя)
ОБЯЗАТЕЛЬСТВА ИЗГОТОВИТЕЛЯ: как то предусмотрено Законом РФ о защите прав потребителей. Обеспечивается гарантией изготовителя, но не продавца.
ОБЯЗАТЕЛЬСТВА ПРОДАВЦА: возврат и обмен товара осуществляется в течении 2-х недель с момента продажи товара при отсутствии механических повреждений и следов установки. Доставка дефектного товара продавцу должна быть осуществлена покупателем. Покупатель своей подписью на настоящем документе подтверждает, что ему была предоставлена возможность получить разъяснения в отношении всех терминов, использованных в настоящем документе и ознакомиться с Законом РФ о защите прав потребителей, а также, что ему был передан на руки товар (в полной комплектности, без повреждений, в неповреждённой упаковочной таре). С инструкцией по установке и гарантией производителя ознакомлен.

Заказ оформил ____________ (Мартынов Михаил)
Отпустил ____________ Получил ____________

Уважаемые клиенты! При покупке в нашем магазине комплекта шин или дисков Вы получаете скидку в размере 10% на услуги Шиномонтажа и Автосервиса!
Не забывайте спрашивать купоны на скидку у менеджеров.

«Вилс Авто»
МОСКВА

ОПЛАЧЕНО
«___»________200

Figure 4. An example of financial document

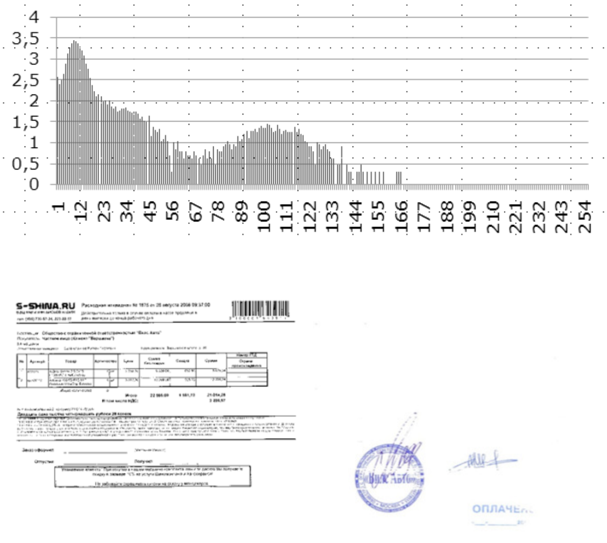


Figure 5. A saturation histogram and text and graphics layers after separation

Step 3. It is easy to notice two classes in this histogram: the first class is formed by pixels with low value of color saturation and the second one – by pixels with high value of color saturation. That is the first class corresponds to background and text areas in the source image and the second class corresponds to graphics areas. Find the threshold t^* between these classes by Otsu method (Otsu 1979).

Step 4. If the obtained threshold value t^* is quite small (i.e. if t^* less than some in advance defined minimal value t_{min}) than the source image contains only black text and has no graphics. So the source image coincides with the text layer and the graphics layer is empty. Otherwise generate the following separation algorithm. For each pixel (x, y) calculate its saturation s and if $s < t^*$ then (x, y) is a part of text layer. Otherwise $s \geq t^*$ then (x, y) is a part of graphics layer.

An example of separation of a financial document image is shown in Figure 5.

PLACING LAYERS IN PDF/A

After the image is separated the text layer is binarized by Otsu binarization method. Two popular lossless algorithms, CCITT Group 4 and JBIG2 (ISO, 1993), were considered. In this case (compression of monochrome image with only text information) we chose the last one because JBIG2 typically generates files one third to one fifth the size of CCITT Group 4.

The graphics layer is down-scaled to 100 dpi and compressed by the JPEG (ISO, 1994) with middle compression quality. Such compression makes it possible to reach small size without significant distortion of graphics.

As it has been mentioned in the brief description of the algorithm, the text layer can be easily recognized by any OCR system. We used OpenOCR engine which is an omnifont multilingual open source system (OpenOCR.org). We split the recognized text into the words and place separately each word into PDF/A file using invisible style. In this case correct layout which could be left out during OCR processing will be established automatically by PDF parser program.

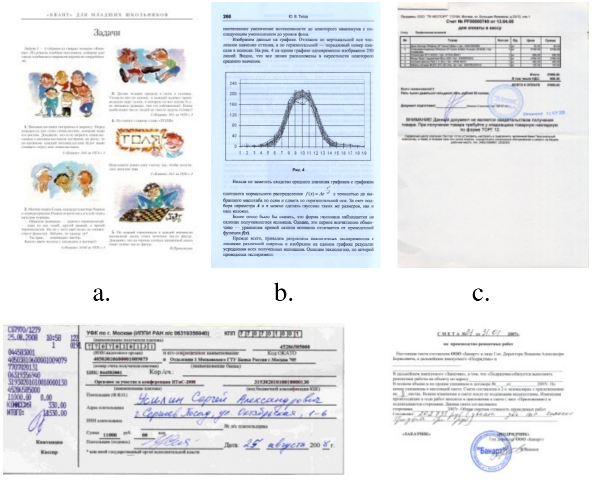
CONCLUSION

In the paper a new structural compression algorithm of document images is considered. Thanks to the using PDF/A as output format the compressed image can be used in electronic archives. Using different schemes of image separations makes it possible to save the document image face and achive the highest possible compression ratio. Including recognized text to the PDF/A file makes it easy to find and copy information in the documents.

Typically proposed algorithm compress document images at 300 dpi to 50-150 KB (approximately 3 to 10 times better than JPEG for a similar level of subjective quality). The following table presents the compression results of several images shown in Figure 6.

Table 1. Compression results

Image	JPEG	DJVU	PDF/A
Figure 6a	718 KB	50 KB	94 KB
Figure 6b	373 KB	29 KB	51 KB
Figure 6c	642 KB	23 KB	52 KB
Figure 6d	120 KB	20 KB	40 KB
Figure 6e	292 KB	17 KB	29 KB



a. b. c.

d. e.

Figure 6. Examples of different document images

Significantly difference in sizes between DJVU and PDF/A files is explained by the fact that besides the useful information (images and text) PDF/A file contains also some service information (fonts, color profiles, metadata and others).

In case of automated compression of a set of different type documents, automatic selection of the appropriate separation algorithm is necessary. This task is solved with the help of methods of preliminary determination of document type.

The considered technique of structural compression is implemented as a program and is embedded in Magnitogorsk Iron and Steel Works Open Joint Stock Company (MMK) workflow system.

REFERENCES

Fletcher, L.A. and R. Kasturi. 1988. “A Robust Algorithm for Text String Separation from Mixed Text/Graphics Images”. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.10, No.6, 910-918.

Gonzalez, R.C. and R.E. Woods. 2002. *Digital Image Processing, second edition (2nd Edition)*. Prentice Hall.

Haffner, P.; L. Bottou; P.G. Howard; P. Simard; Y. Bengio; and Y. Le Cun. 1998. “Browsing through High Quality Document Images with DjVu”. In *Proceedings of the Advances in Digital Libraries Conference*. IEEE, Washington, DC, USA, 309-319.

Hasan, Y.M.Y. and L.J. Karam. 2000. “Morphological Text Extraction from Images”. *IEEE Transactions on Image Processing*, Vol.9, No.11, 1978-1983.

International Organization for Standardization. 1994. *Information technology – Coded representation of picture and audio information – Progressive bi-level image compression (ISO/IEC 11544).*

International Organization for Standardization. 2005. *Document management – Electronic document file format for long-term preservation – Part 1: Use of PDF 1.4 (PDF/A) (ISO 19005-1).*

Lizardech. 2005. *Lizardtech DjVu Reference*.

Nikolaev, D.P. 2003. “Segmentation-based binarization method for color document images”. In *Proceedings of 6th Open Russian-German Workshop on Pattern Recognition and Image Understanding*, 190-193.

OpenOCR.org. *http://en.openocr.org/*.

Organization for Standardization. 1994. *Information technology – Digital compression and coding of continuous-tone still images: Requirements and guidelines (ISO/IEC 10918-1).*

Otsu, N. 1979. “A threshold selection method from gray-level histograms”. *IEEE Transactions on Systems, Man, and Cybernetics*, Vol.9, No.1 (Jan), 62-66.

Rog, J. 2007. *PDF Guidelines: Recommendations for the creation of PDF files for long-term preservations and access.*

Trier, P.D. and T. Taxt. 1995. “Evaluation of binarization methods for document images”. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.17, No.3, 312-315.

Van Herk, M. 1992. “A Fast Algorithm for Local Minimum and Maximum Filters on Rectangular and Octagonal Kernels”, *Pattern Recognition Letters*. Elsevier Science Inc, NY, USA, 517-521.

Nikolaev D.P., S.M. Karpenko, I.P. Nikolaev and P.P Nikolayev. 2008. “Hough Transform: Underestimated

Tool in the Computer Vision Field". In *Proceedings of 22nd European Conference on Modelling and Simulation (ECMS 2008)*. Dudweiler, Germany, 238-243.

AUTHOR BIOGRAPHIES

SERGEY A. USILIN was born in Sergiev Posad, Russia. He studied mathematics and physics, obtained his Master degree in 2009 from Moscow Institute of Physics and Technology (MIPT). Since 2009 he has been a Ph.D. student at Moscow Institute of Physics and Technology. His research activities are in the areas of image processing and object detection. His e-mail address is usilin.sergey@gmail.com.

DMITRY P. NIKOLAEV was born in Moscow, Russia. He studied physics, obtained his Master degree in 2000 and Ph.D. degree in 2004 from Moscow State University. Since 2007 he has been a head of the sector at the Institute for Information Transmission Problems, RAS. His research activities are in the areas of computer vision with primary application to colour image understanding. His e-mail address is dimonstr@iitp.ru and his Web page can be found at http://chaddt.net.ru/Lace/cv_ip/index.html

VASSILI V. POSTNIKOV was born in Sverdlovsk, USSR. He studied mathematics and physics, obtained his Master degree in 1990 from Moscow Institute of Physics and Technology (MIPT). His Ph.D. thesis was named "Structural documents identification and recognition". He obtained his Ph.D. degree in 2002 in the Institute for System Analysis (ISA RAS). Since 2006 he was deputy head of the Cognitive technologies department at MIPT and leading scientist at ISA RAS. His research activities are in the areas of document image analysis, processing and understanding. His e-mail address is vassili.postnikov@gmail.com.