An OCR Assessment of the Quality of Document Images Acquired with Portable Digital Cameras

Rafael Dueire Lins, Brenno Miro, Gabriel Pereira e Silva
Departamento de Eletrônica e Sistemas – Universidade Federal de Pernambuco - Brazil
rdl@ufpe.br, gfps@cin.ufpe.br

Abstract

This article analyses the quality of documents acquired with portable digital cameras for Optical Character Recognition. The results obtained are compared with same documents after border removal, perspective and skew correction and their scanned equivalent with different resolutions and saved into distinct file formats.

1. Motivation

Students and professionals of many different areas now use portable digital cameras for digitalizing documents, taking advantage of their low weight, portability, low cost, small dimensions, etc. This new research area [1][2] is evolving fast in many different directions and claims for new algorithms, tools and processing environments that are able to provide users in general with simple ways of visualizing, printing, transcribing, compressing, storing and transmitting through networks such images. Reference [3] points out some particular problems that arise in this document digitalization process: the first of all is background removal. Very often the document photograph goes beyond the document size and incorporates parts of the area that served as mechanical support for taking the photo of the document. The second problem is due to the skew often found in the image in relation to the photograph axes, as documents have no fixed mechanical support very often there is some degree of inclination in the document image. The third problem is non-frontal perspective, due to the same reasons that give rise to skew. A fourth problem is caused by the distortion of the lens of the camera. This means that the perspective distortion is not a straight line but a convex line, depending on the quality of the lens and the relative position of the camera and the document. The fifth difficulty in processing document images acquired with portable cameras is due to non-uniform illumination. This paper focuses on assessing the output of a commercially OCR (Optical Character Recognition) software for such documents. The results obtained are compared with the results obtained of processing the same batch of documents with PhotoDoc [4] a freely available software environment for processing document images acquired with portable cameras. The results of unprocessed and PhotoDoc Processed camera images are compared with the transcription obtained for the scanned version of the same documents. This work besides updating the results presented in [5] to more modern camera models of current use today, it applies a much better assessment methodology.

2. The assessment methodology

Assessing image quality in general is a complex subjective task. A quantitative assessment that avoids such subjectivity is of great importance. Similarly to the experiments reported in [5], in this paper the assessment methodology was limited to analyze the performance of commercial OCR tools. ABBYY FineReader Professional Edition 9 [6] was used, because it is possibly the best general purpose tool available today, able to process formatted texts.

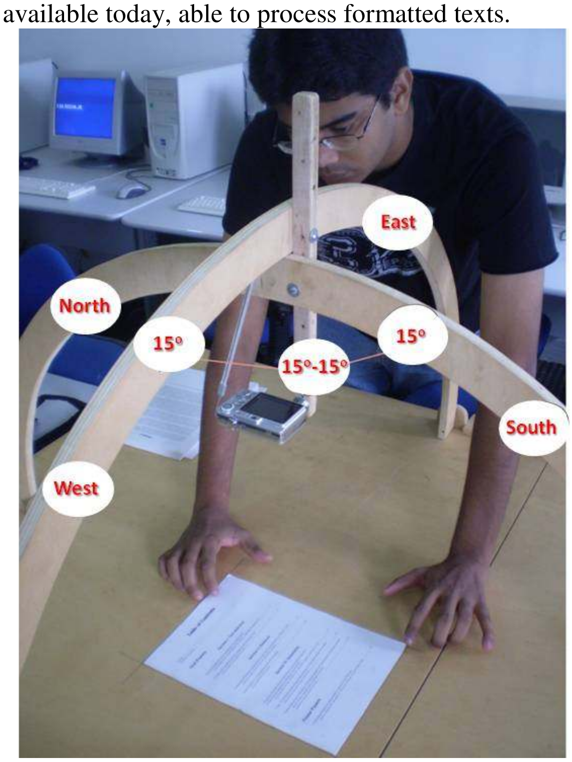


Figure 01. The “Planetarium” test bed

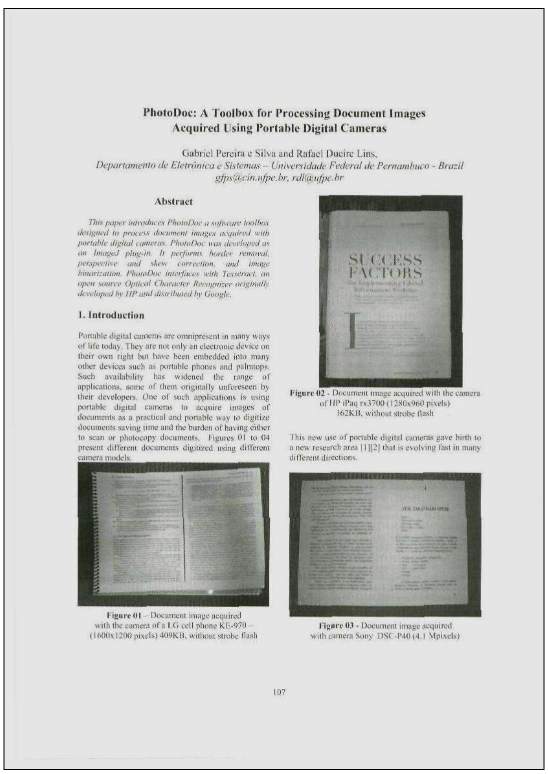
The “Planetarium” test bed shown in Figure 01 allows a controlled way to measure the angles and height of the camera to verify in extreme cases the effects of the perspective into the document transcription. These results are later used to assess the gains obtained with the documents after each processing step. On its turn, analyzing the results of OCRs is far from being a trivial task. The methodology presented in reference [7] which takes into account the nature of the errors in transcription was adopted here.

The errors were classified according to:

1. Character replacement.
2. Missing characters.
3. Character insertion.
4. Punctuation errors.

3. Test images features

The 168 pages of the proceedings of CBDAR 2007 were used as test document images for this work. Several pages include photographs, graphs, tables and other illustrations. They are printed in black in opaque white paper, where negligible back-to-front [8] interference was observed.



PhotoDoc: A Toolbox for Processing Document Images Acquired Using Portable Digital Cameras

Gabriel Pereira e Silva and Rafael Dueire Lins.
Departamento de Eletrônica e Sistemas – Universidade Federal de Pernambuco - Brazil
gfps@cin.ufpe.br, rdl@ufpe.br

Abstract

This paper introduces PhotoDoc a software toolbox designed to process document images acquired with portable digital cameras. PhotoDoc was developed as an ImageJ plug-in. It performs border removal, perspective and skew correction, and image binarization. PhotoDoc interfaces with Tesseract, an open source Optical Character Recognizer originally developed by HP and distributed by Google.

1. Introduction

Portable digital cameras are omnipresent in many ways of life today. They are not only an electronic device on their own right but have been embedded into many other devices such as portable phones and palmtops. Such availability has widened the range of applications, some of them originally unforeseen by their developers. One of such applications is using portable digital cameras to acquire images of documents as a practical and portable way to digitize documents saving time and the burden of having either to scan or photocopy documents. Figures 01 to 04 present different documents digitized using different camera models.

Figure 01 – Document image acquired with the camera of a LG cell phone KE-970 – (1600x1200 pixels) 409KB, without strobe flash

Figure 02 - Document image acquired with the camera of HP iPaq rx3700 (1280x960 pixels) 162KB, without strobe flash

This new use of portable digital cameras gave birth to a new research area [1][2] that is evolving fast in many different directions.

Figure 03 - Document image acquired with camera Sony DSC-P40 (4.1 Mpixels)

107

Figure 02. Scanned page of the Proceedings of CBDAR´07.

Table I presents the results of the total of errors found in the OCR transcription of all document images digitized with a Ricoh Affício 1075 flatbed scanner in 100, 200 and 300 dpi saved into four different file formats: bmp (uncompressed), jpg (1% losses), png (lossless), and tiff (uncompressed), using the software provided by the scanner manufacturer. Figure 02 shows an example of the test images used in this work.

TABLE I
CHARACTER ERRORS FOUND IN SCANNED DOCUMENT IMAGES

100 DPI	BMP	JPG	PNG	TIF
replacement	55590	63113	63646	66894
punctuation	3851	4174	4057	4977
missing	7907	8095	8852	8454
insertion	96078	95729	96126	96126
SIZE	498MB	21.6MB	75.2MB	498MB
200 DPI	**BMP**	**JPG**	**PNG**	**TIF**
replacement	49837	41194	39912	43270
punctuation	2575	2787	2591	2738
missing	4274	5565	5259	5345
insertion	71467	79298	81402	79904
SIZE	1.95GB	65.1MB	222MB	1.95GB
300 DPI	**BMP**	**JPG**	**PNG**	**TIF**
replacement	38529	58681	58742	58647
punctuation	2398	3376	3429	3397
missing	4478	5883	5940	6447
insertion	81909	91877	93803	92045
SIZE	4.42GB	118MB	395MB	4.42GB

Analyzing the data in Table I one may observe that different file formats yield varying error rate even amongst lossless ones (bmp, png, and tiff). One may possibly say that the best trade-off between space and OCR correct recognition rate is reached in the 200 dpi scanning saved in PNG. It is interesting to note that a higher resolution tended to drastically increase the insertion noise.

4. Unprocessed Image Transcription

The camera used in this work is a Sony Cyber-shot 7.2 Mega Pixels model DSC-W55, with lenses Carl-Zeiss Vario-Tessar 2.8-5.2/6.3-18,9. Figure 03 exemplifies the test image obtained in the “Planetarium” test bed. Documents were obtained in true-color, in 7.0 and 5.0 Mpixels, with and without the inbuilt camera strobe flash. The camera was set into “auto-focus” mode, i.e. the user leaves to the device the automatic setting of the focus. This is consistent with the expected knowledge of the end user. In the case of documents acquired with portable digital cameras with no mechanical support very often images have perspective distortion. Skewed images are unpleasant for human visualization, introduce extra difficulty in text reading, claim extra space for storage, degrade OCR performance, etc. this problem arises in almost all documents. This work used two strategies to measure perspective degradation the first one was using a device developed by the authors in which the position of the camera was set and in the second strategy the photo was taken “free hand”, without mechanical support. Figure 03 shows a document photo taken with the Planetarium.

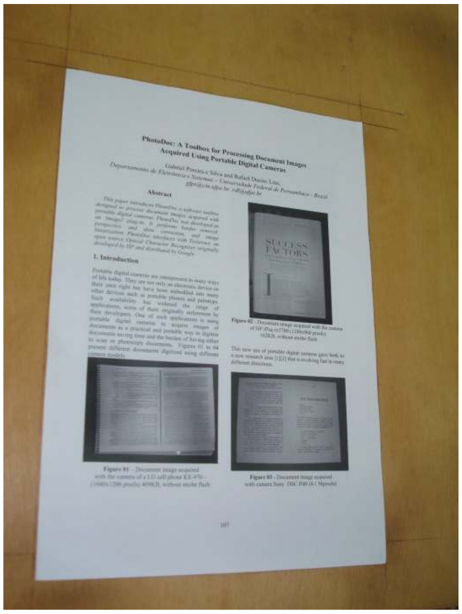


Figure 03. Photo taken in the Planetarium with 15^{o}S-15^{o}W.

The photos taken in the Planetarium are surrounded by part of the board in wood that serves as mechanical support to it. The photos were taken indoors with artificial illumination provided by fluorescent lights sufficiently high-up to light the document surface evenly.

Table II presents the results of the transcription of the documents under different inclinations and heights, with and without strobe flash. One may observe that the position of the camera in relation to the document causes a wide variation in results. As one expects, the greater the angle in relation to the frontal plane of the document, the larger the OCR transcription degradation. In general, the images acquired with 7.2 Mpixels yielded better results than with 5.0 Mpixels, but this was not the case when the perspective distortion happened in the West and South directions simultaneously (15^{o}E – 15^{o} S).

At the height of 37 cm parallel with the plane (0^{o}E – 0^{o}S) the inbuilt strobe flash of the camera yielded an uneven illumination not perceptible visually, but that degraded OCR response both at 5.0 and 7.2 Mpixels.

At the height of 45 cm the contrary phenomenon was observed and the flash yielded better OCR transcription results. When no mechanical support was used (Free-hand) the use of the strobe flash in 7.2 Mpixels provided the best results. These results are close to the ones obtained by using the height of 37 cm without any inclination angle. The experiments performed in reference [5] report a measure of the skew angle of the bottom line of the 50 documents analyzed was around 2° in each direction. That value was also observed in the experiments performed herein. One should remark, however that users in general tend to be less careful and tend to take photos with a higher perspective distortion. Experiments performed with several people that meet that profile showed that the perspective distortion does not exceed 10 degrees in each direction simultaneously and that the document image is seldom chopped off.

TABLE II
CHARACTER ERRORS FOUND IN CAMERA DOCUMENT IMAGES

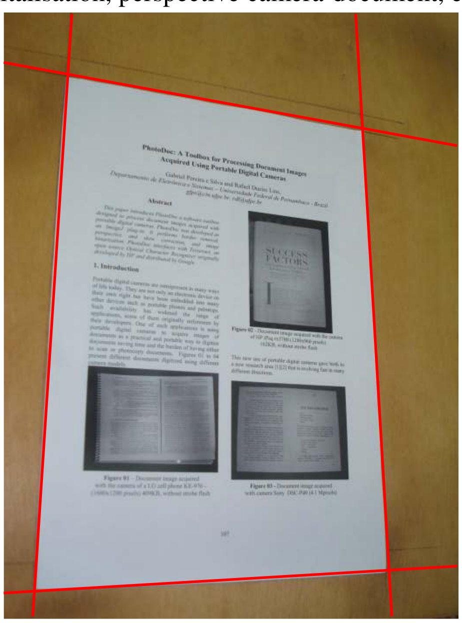
	5.0 Mpixels		7.2 Mpixels	
15° South	**+ flash**	**no flash**	**+ flash**	**no flash**
replacement	62349	58258	59900	62466
punctuation	3787	3694	3665	3769
missing	6381	6477	6010	6658
insertion	94282	93440	95943	94149
15° West	**+ flash**	**no flash**	**+ flash**	**no flash**
replacement	70138	72342	68136	69308
punctuation	5463	5494	5321	5763
missing	7685	6254	6223	6365
insertion	80154	85100	80498	80375
30° South	**+ flash**	**no flash**	**+ flash**	**no flash**
replacement	64499	68242	62004	57720
punctuation	3850	4513	3846	3935
missing	6154	6885	5963	6683
insertion	93420	88841	92783	86607
30° West	**+ flash**	**no flash**	**+ flash**	**no flash**
replacement	85854	89840	88910	85113
punctuation	8389	8655	8948	8819
missing	8429	8669	8041	9716
insertion	50630	53798	47855	49014
15° W - 15° S	**+ flash**	**no flash**	**+ flash**	**no flash**
replacement	82182	84307	86141	83036
punctuation	6781	7010	7210	7098
missing	8553	9455	9044	9167
insertion	69204	72238	71451	71213
height= 45 cm	**+ flash**	**no flash**	**+ flash**	**no flash**
replacement	60454	62047	63674	60961
punctuation	3651	3829	3706	3673
missing	7352	7486	7133	7973
insertion	93319	92799	94593	94748
height= 37 cm	**+ flash**	**no flash**	**+ flash**	**no flash**
replacement	71954	70470	63563	62380
punctuation	5537	5785	3552	3748
missing	8405	8045	7284	7253
insertion	79699	80515	95144	95153
Free hand	**+ flash**	**no flash**	**+ flash**	**no flash**
replacement	58891	62866	54996	63867
punctuation	5603	3911	5073	4304
missing	10033	7445	10474	9715
insertion	82741	93774	90232	90271

5. PhotoDoc Processing

As mentioned in the introduction of this paper, the motivation for research reported herein is to meet the needs of ordinary people such as students and professionals that acquire document images using portable digital cameras. As shown in Figure 03 such documents are framed by the place that serve as mechanical support the photo to be taken and as a perspective distortion. PhotoDoc is a simple processing environment developed with the aim to help those non-expert users that digitize documents with portable digital cameras. A brief overview of PhotoDoc processing capabilities are shown here as it was used to process all the documents that are also OCR transcribed. Experiments in PhotoDoc showed that lens distortion has negligible effects if compared with border removal and perspective and skew correction. Thus, it is ignored.

5.1 Border Removal

PhotoDoc performs automatic background border removal of images of documents obtained with portable digital cameras imposing as few restrictions as possible, because users tend to acquire those document images in non-ideal conditions of, illumination of the surface the document is placed on for digitalisation, perspective camera-document, etc.


PhotoDoc: A Toolbox for Processing Document Images Acquired Using Portable Digital Cameras

Abstract

1. Introduction

Figure 04. Image from Figure 03 showing perspective correction reference points and edges.

As may be observed in Figure 03, the document image is surrounded by a wooden background area of no value in terms of information. This area not only drops the quality of the resulting image for CRT screen visualization, but also consumes space for storage and large amounts of toner for printing, alters the segmentation algorithm of the OCR and thus affects the response obtained in the number of characters and words correctly transcribed, as shown later on in this paper. Several papers in the literature address this problem in different applications [9, 10, 11, 12]. Removing such frame manually is not practical due to the need of a specialized user and time consumed in the operation. The algorithm presented in reference [11] is used in PhotoDoc to automatically remove such border as an OCR pre-processing stage. It assumes that the background may be of any kind of colour or texture, provided that there is a colour difference of at least 32 levels between the image background and at least one of the RGB components of the most frequent colour of the document background (paper).

5.2 Perspective and skew correction

The freedom allowed in acquiring document images with portable digital cameras without mechanical support invariably leads to perspective distortion. Several algorithms in the literature address this problem [3, 14, 15, 16]. The correction of perspective distortion has border detection as a first step to find the polygon that margins the image and getting the four corner points that will serve as reference for the linear transformation. The image of the four corner points serve to crop the perspective corrected image and automatically performs skew correction. On the other hand, perspective distortion opens a number of alternatives which cause different effects in the quality of the image produced both in terms of visualization and OCR response. In general, the skew angle was small (less than 2^o), thus this means that the image tends to exhibit a trapezoidal shape. Two alternatives for correction arise: either to narrow the opening edges or to widen the closing edges. The latter alternative was discarded in PhotoDoc because the general trend is to disconnect contiguous areas, which has a serious degrading effect on OCR response. The interpolation methods applied in PhotoDoc is closest neighbor. The image obtained after perspective correction and cropping closely resembles the scanned one.

Table III presents the OCR response for documents, after PhotoDoc processing. Comparing the results obtained in Tables I, II, and III one may observe that Photodoc processing largely improved the OCR recognition rate of documents, yielding better OCR response than images scanned in 100 dpi resolution, in general. The only exceptions are in the case of very strong perspective distortion 30^o South and in the case of insertion errors when photos were taken at a height of 37 cm both with and without the strobe flash. Insertion errors are harder to be corrected with the help of dictionaries than the other errors.

TABLE III
CHARACTER ERRORS FOUND IN DOCUMENT IMAGES AFTER PHOTODOC PROCESSING

	5.0 Mpixels		7.2 Mpixels	
15° South	+ flash	no flash	+ flash	no flash
replacement	51435	59671	43163	41965
punctuation	2075	2964	2050	2948
missing	5085	5854	5927	5072
insertion	79835	76183	75844	80559
15° West	+ flash	no flash	+ flash	no flash
replacement	53025	51241	41716	42365
punctuation	2169	2432	2881	2880
missing	5744	5550	4713	4875
insertion	77884	75605	79517	78545
30° South	+ flash	no flash	+ flash	no flash
replacement	69041	76198	62129	62678
punctuation	3173	2953	3128	3245
missing	6039	6889	5882	5544
insertion	77543	78036	78078	71426
30° West	+ flash	no flash	+ flash	no flash
replacement	63709	62699	63509	65477
punctuation	3118	4096	4390	4499
missing	6820	7190	7728	8941
insertion	82663	92792	94419	93170
15° -15° S	+ flash	no flash	+ flash	no flash
replacement	63545	62588	62506	61783
punctuation	4397	4017	3888	3949
missing	7939	7118	7604	7835
insertion	95203	92656	96927	97765
height= 45 cm	+ flash	no flash	+ flash	no flash
replacement	50412	59689	40127	43567
punctuation	2048	2071	1862	1898
missing	7588	7386	7232	7795
insertion	71430	77838	84096	81263
height= 37 cm	+ flash	no flash	+ flash	no flash
replacement	52212	53935	41483	43019
punctuation	1671	1794	1559	1755
missing	7047	7617	6507	6583
insertion	95435	97526	97942	97685
Free hand	+ flash	no flash	+ flash	no flash
replacement	54269	60584	56880	71226
punctuation	4589	4823	5538	5597
missing	6225	6872	6946	5795
insertion	77711	84598	70237	92331

5. Conclusions

This paper provides a comparative analysis of the quality of documents acquired through 5.0 and 7.2 Mpixels Sony portable digital camera in comparison with their scanned version with three different resolutions (100, 200 and 300 dpi). A batch of 168 documents was studied totaling 479,154 characters. The quantitative analysis performed herein allows to conclude that portable digital cameras not only provide a simple way to digitalize documents to be read by humans, but the quality of documents allows means for image-to-text transcription using commercial OCRs. The OCR performance improves if the document is processed in an image processing environment such as PhotoDoc that removes the borders introduced during document photographing and is perspective and skew corrected.

Several challenges are faced to improve OCR performance. Illumination and compensating the effect of the embedded strobe flash are two of the most important ones as they pose difficulties to image binarization.

7. References

[1] D.Doermann, J.Liang, H. Li, "Progress in C.-Based Document Image Analysis," ICDAR’03, Vol(1): 606, 2003.

[2] J. Liang, D. Doermann and H. Li. Camera-Based Analysis of Text and Documents: A Survey. International Journal on Document Analysis and Recognition, 2005.

[3] R.D.Lins, A.R.Gomes e Silva and G.Pereira e Silva, Enhancing Document Images Acquired Using Portable Digital Cameras, ICIAR´07, LNCS, Springer-Verlag, 2007.

[4] G.Pereira e Silva and R.D.Lins. PhotoDoc: A Toolbox for Processing Document Images Acquired Using Portable Digital Cameras. CBDAR´2007, pp.107-114, 2007.

[5] R.D.Lins, *et al*. Assessing and Improving the Quality of Document Images Acquired with Portable Digital Cameras, ICDAR´2007, pp.569-573, IEEE Press, 2007.

[6] ABBYY FineReader Professional Edition 9 - http://www.abbyy.com/

[7] R.D.Lins and N.F.Alves. A New Technique for Assessing the Performance of OCRs. IADIS – Int. Conf. on Comp. Applications, IADIS Press, v. 1, p. 51-56, 2005.

[8] J. M. M. da Silva *et al.* Binarizing and Filtering Historical Documents with Back-to-Front Interference, ACM-SAC 2006, Nancy, April 2006.

[9] K.C.Fan, Y.K.Wang, T.R.Lay, Marginal noise removal of document images, Patt.Recognition. 35, 2593-2611, 2002.

[10] Lu S and C L Tan, Camera document restoration for OCR, CBDAR 2005/ICDAR 2005, Seoul, Korea.

[11] R. Gomes e Silva and R. D.Lins. Background Removal of Document Images Acquired Using Portable Digital Cameras. LNCS 3656, p.278-285, 2005.

[12] H.S.Baird, Document image defect models and their uses, ICDAR’93, Japan, IEEE Comp. Soc., pp. 62-67, 1993.

[13] L.G.Shapiro and G.C.Stockman, Computer Vision, March 2000. http://www.cse.msu.edu/~stockman/Book/book.html.

[14] L. Jagannathan and C. V. Jawahar, “Perspective correction methods for camera based document analysis,” pp. 148–154, CBDAR 2005, Seoul, Korea. 2005.

[15] P. Clark, M. Mirmehdi, “Recognizing Text in Real Scenes”, IJDAR, Vol. 4, No. 4, pp. 243-257, 2002.

[16] Clark, M. Mirmehdi, “On the Recovery of Oriented Docs. from Single Images”, CSTR-01-004, Bristol, 2001.