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.

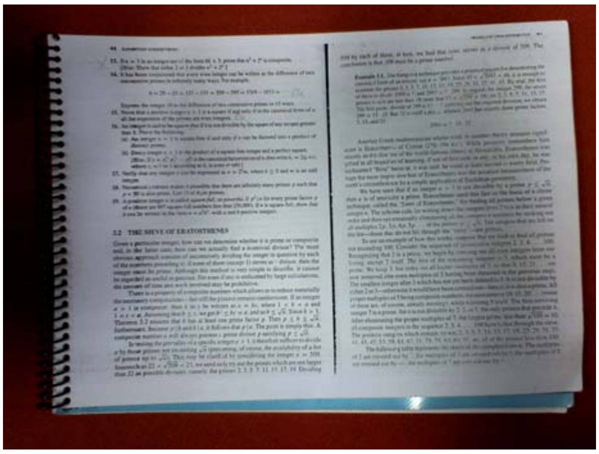
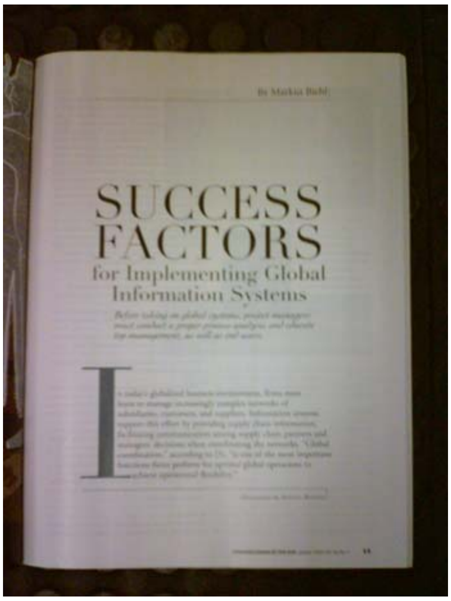


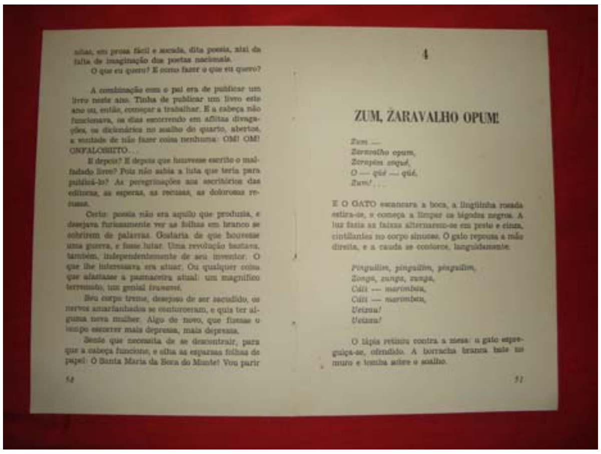
Figure 01 – Document image acquired with the camera of a LG cell phone KE-970 – (1600x1200 pixels) 409KB, without strobe flash



SUCCESS FACTORS
for Implementing Global Information Systems

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.



ZUM, ZARAVALHO OPUM!

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

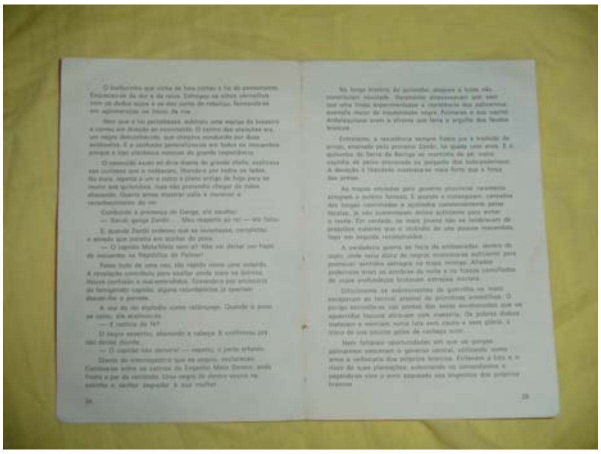


Figure 04 - Document image acquired with camera Sony DSC-P52 (3.2 Mpixels)

New algorithms, tools and processing environments are needed to provide users in general with simple ways of visualizing, printing, transcribing, compressing, storing and transmitting through networks such images. PhotoDoc is a processing environment developed to meet such needs.

The test documents used here were obtained in true-color, under different illumination conditions, with and without the inbuilt camera strobe flash, using a portable cell phone manufactured by LG KE-970 – (1600x1200 pixels) 409KB without strobe flash, a HP iPaq rx3700 (1280x960 pixels) 162KB without strobe flash, and two different models of portable cameras manufactured by Sony Corp. (models DSC-P52 and DSC-P40) of 3.2 and 4.1 Mega pixels, respectively. All cameras were set into "auto-focus" mode, i.e. the user leaves to the device the automatic setting of the focus.

Several specific problems arise in this digitalization process and must be addressed to provide a more readable document image, which also claims less toner to print, less storage space and consumes less bandwidth whenever transmitted through networks. The first of all is background removal as document photograph goes beyond the document size and incorporates parts of the area that surrounds it. The absence of mechanical support for taking the photo yields a non-frontal perspective that distorts and skews the document image. The distortion of the lenses of the cameras makes the perspective distortion not being a straight but a convex line, depending on the quality of the lens and the relative position of the camera and the document. Non-uniform illumination of the environment and strobe flash, whenever available in the device adds difficulties in image enhancement and binarization.

2. The PhotoDoc Environment

PhotoDoc was conceived as a device independent software tool to run on PCs. Whenever the user unloads his photos he will be able to run the tool prior to storing, printing or sending through networks the document images. The algorithms and the basic functionality of PhotoDoc may be incorporated to run on a device such as a PDA or even a camera itself. Such a possibility is not considered further herein. Due to implementation simplicity and portability the current version of PhotoDoc was implemented as an ImageJ [20] Plug-in. ImageJ is an open source image processing environment in Java developed by Wayne Rasband, is at the Research Services Branch, National Institute of Mental Health, Bethesda, Maryland, USA. Figure 05 shows a screen shot of PhotoDoc being activated from ImageJ.

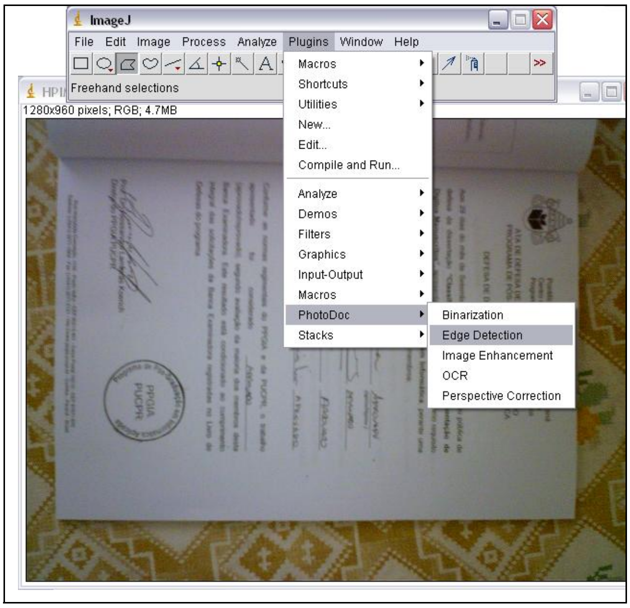


Figure 05 – PhotoDoc plugin in ImageJ

As one may observe in Figure 05, the present version of the PhotoDoc plug-in offers five different filters, which appear in alphabetical order:

- Binarization
- Edge Detection
- Image Enhancement
- OCR, and
- Perspective Correction and Crop

The fact that PhotoDoc is now in ImageJ also allows the user to experiment with the different filters and other plug-ins already present in ImageJ. It is important to stress, however, that ImageJ as an open code library allows the developer to extract from it only the needed functionality in such a way that the developer may provide to ordinary user a PhotoDoc interface that looks independent from ImageJ. At present, the authors of this paper consider such possibility premature. Such tool particularization seems to be more adequate is coupled with a particular device, which allows also a better fine tuning of the algorithms developed for and implemented in PhotoDoc. In what follows the PhotoDoc filter operations are described.

3. A New Border Detection Algorithm

The very first step to perform in processing a document image in PhotoDoc is to detect the actual physical limits of the original document [3]. The algorithm presented in reference [7] was developed based on images acquired by 3 and 4 Mpixel cameras. Unfortunately, its performance in lower resolution cameras has shown to be inadequate.

A new edge detection algorithm, based on ImageJ filters, was developed and is presented herein. This new algorithm behaved suitably on a wide range of images with different kinds of paper, including glossy ones. The new algorithm was obtained by composing existing filters in ImageJ. The steps of the new border removal algorithm are:

1. Process + Enhance Contrast.
2. Process + Find Edges.
3. Image Type 8 bits.
4. Image Adjust Threshold – Black and White value 122.
5. Process Binary + Erode.

At this point the resulting image provides well defined borders that allow finding the document edges. Figure 06 presents the result of applying the steps of the algorithm above to the document image presented in Figure 01, which appears in the top-left corner, until reaching the resulting image in the bottom-right one.

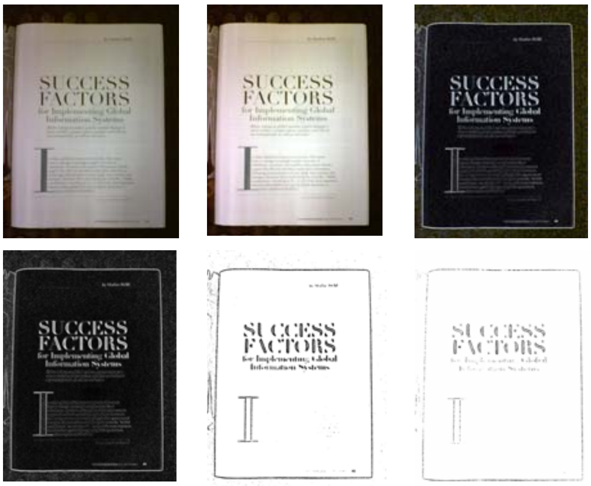


Figure 06 – Step by step filtering of image for border edge detection using ImageJ

The result of the activation of button "Edge Detection" on a document in PhotoDoc yields an image such as the one presented on Figure 07.

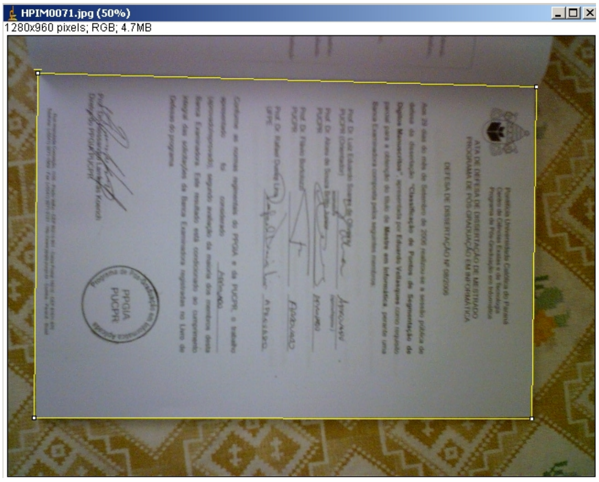


Figure 07 - Document image with edges (in yellow) automatically detected by PhotoDoc

Although the algorithm presented correctly detected edges for all the tested documents, the Edge Detection filter in PhotoDoc allows the user to adjust the edges by dragging along the four corners of the document image. This may also be of help whenever the document photo is not completely surrounded by a border of whenever strong uneven illumination causes edges not to be detected.

4. Perspective Correction and Crop

PhotoDoc incorporates a filter to correct the distortion and skew of the image introduced by the non-frontal position of the camera in relation to the document. A pinhole model for the camera was adopted [5, 6, 9] and provides a simple way to solve the problem at first. The experiments reported in [8] point at, narrowing edges and using bi-cubic interpolation as the rule-of-thumb to yield images more pleasant for the human reader and also with less transcription errors in OCR response. The difficulty inherent to such transformation is finding the four corners of the original image that will be taken as the basis for the correction. Edge or border detection, as explained above, is the first step the image should undergo. Once the document edges are determined as shown in Figure 07 the "Perspective Correction" filter in PhotoDoc may be called as shown in Figure 08.

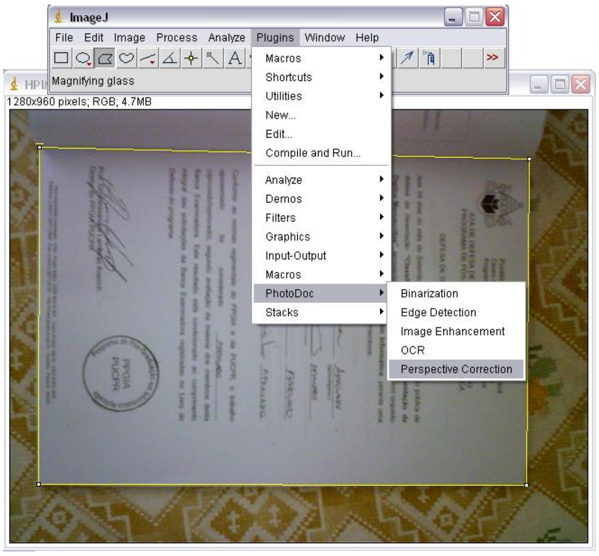
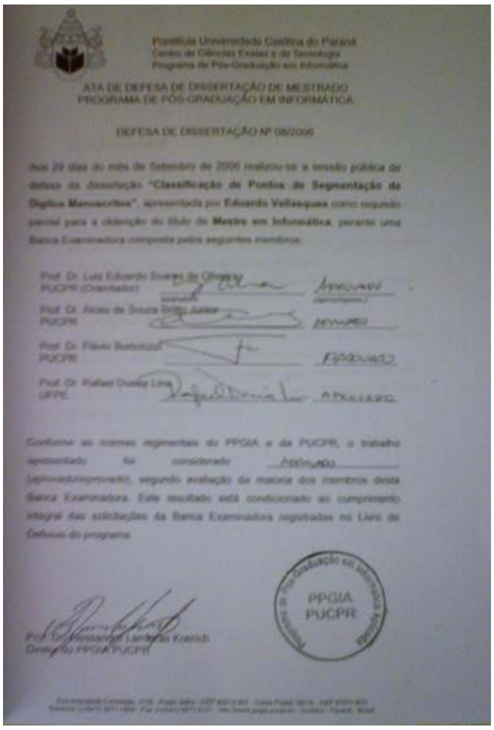


Figure 08 – Activating Perspective Correction in edge detected image from Figure 07 in PhotoDoc

PhotoDoc will automatically perform a perspective, skew, and orientation correction and then crop the resulting image, yielding a document as the one shown in Figure 09. One should observe that the cropped document whenever visualized in a screen display provides a far better image to the human reader, if printed saves toner and yields a more readable document, and whenever stored or transmitted through computer networks saves on average 25% of space [7].



Pontifícia Universidade Católica do Paraná
Centro de Ciências Exatas e de Tecnologia
Programa de Pós-Graduação em Informática

ATA DE DEFESA DE DISSERTAÇÃO DE MESTRADO
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

DEFESA DE DISSERTAÇÃO Nº 08/2006

Aos 29 dias do mês de Setembro de 2006 realizou-se a sessão pública de defesa da dissertação "**Classificação de Pontos de Segmentação de Dígitos Manuscritos**", apresentada por **Eduardo Vellasques** como requisito parcial para a obtenção do título de **Mestre em Informática**, perante uma Banca Examinadora composta pelos seguintes membros:

Prof. Dr. Luiz Eduardo Soares de Oliveira
PUCPR (Orientador) assinatura Aprovado
(aprovação/reprovação)

Prof. Dr. Alceu de Souza Britto Junior
PUCPR Aprovado

Prof. Dr. Flávio Bortolozzi
PUCPR Aprovado

Prof. Dr. Rafael Dueire Lins
UFPE Aprovado

Conforme as normas regimentais do PPGIA e da PUCPR, o trabalho apresentado foi considerado Aprovado (aprovado/reprovado), segundo avaliação da maioria dos membros desta Banca Examinadora. Este resultado está condicionado ao cumprimento integral das solicitações da Banca Examinadora registradas no Livro de Defesas do programa.

Prof. Dr. Alessandro Lameiras Koerich
Diretor do PPGIA/PUCPR

Programa de Pós-Graduação em Informática Aplicada
PPGIA
PUCPR

[illegible]

Figure 09 – Cropped document after perspective, skew and orientation correction by PhotoDoc

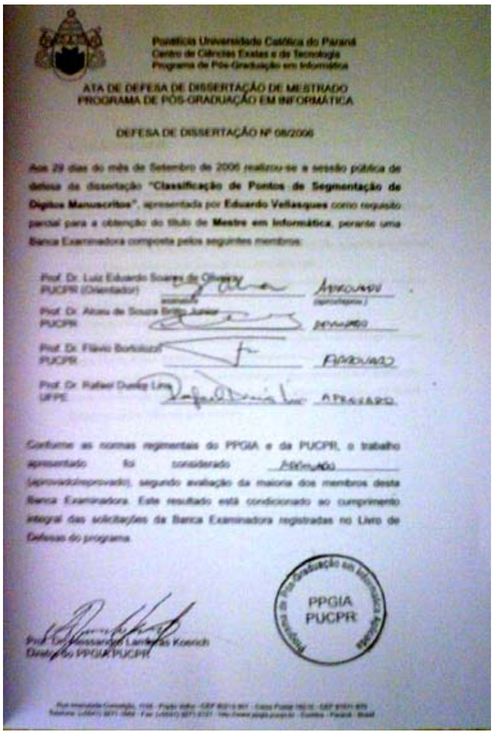
The perspective correction algorithm in PhotoDoc was implemented using the JAI (Java Advanced Imaging) library [23].

5. Image Enhancement

There is a paramount number of possibilities to improve the quality of document images depending of the features of the camera, such as its resolution, lens distortion, the quality of embedded algorithms, environment illumination, quality and color of the paper, etc. Devices have many different technical characteristics, thus it is impossible to find a general solution that would suit all of them, overall with their fast technological evolution pace. However, in the case of devices with embedded flash, if it was not used the resulting document photograph "looked" slightly blurred (out-of-focus). Most possibly, this is related with the fact that the diaphragm of the objective stayed open for much longer to compensate the loss in illumination. As no mechanical support was used to stabilize the camera, chances are that the photographer moved briefly during the shot. Some other times, a slight inclination of the camera in a clear environment may be enough to the luminosity sensor to assume that there is no need for the camera to provide extra illumination, canceling the flash activation. Thus, in order to minimize these factors one may recommend that document photos are:

1. Taken with the embedded flash of the camera set as "on", forcing its activation regardless of the luminosity of the environment.
2. Obtained indoors.

Several of the ImageJ image enhancement algorithms were tested. Non-uniform illumination brings a high degree of difficulty to the problem. A general algorithm that provided gains in all the images studied weakening the illumination problem was provided by the "Enhance Contrast" filter in ImageJ, as may be observed in the image presented in Figure 10.



Pontifícia Universidade Católica do Paraná
Centro de Ciências Exatas e de Tecnologia
Programa de Pós-Graduação em Informática

ATA DE DEFESA DE DISSERTAÇÃO DE MESTRADO
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

DEFESA DE DISSERTAÇÃO Nº 08/2006

Aos 29 dias do mês de Setembro de 2006 realizou-se a sessão pública de defesa da dissertação "Classificação de Pontos de Segmentação de Dígitos Manuscritos", apresentada por Eduardo Vellasques como requisito parcial para a obtenção do título de Mestre em Informática, perante uma Banca Examinadora composta pelos seguintes membros:

Prof. Dr. Luiz Eduardo Soares de Oliveira
PUCPR (Orientador) assinatura APROVADO (aprov/reprov.)

Prof. Dr. Alceu de Souza Britto Junior
PUCPR APROVADO

Prof. Dr. Flávio Bortolozzi
PUCPR APROVADO

Prof. Dr. Rafael Dueire Lins
UFPE APROVADO

Conforme as normas regimentais do PPGIA e da PUCPR, o trabalho apresentado foi considerado APROVADO (aprovado/reprovado), segundo avaliação da maioria dos membros desta Banca Examinadora. Este resultado está condicionado ao cumprimento integral das solicitações da Banca Examinadora registradas no Livro de Defesas do programa.

Prof. Dr. Alessandro Lameiras Koerich
Diretor do PPGIA/PUCPR

Programa de Pós-Graduação em Informática Aplicada
PPGIA
PUCPR

[illegible]
[illegible]

Figure 10 – PhotoDoc enhanced version of Figure 09.

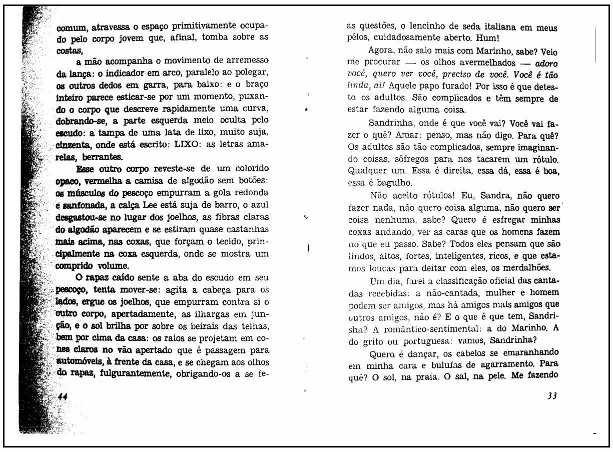
The "Enhance Contrast" filter in ImageJ performs histogram stretching. The *Saturated Pixels* value determines the number of pixels in the image that are allowed to become saturated. Increasing this value will increase contrast. This value should be greater than zero to prevent a few outlying pixels from causing the histogram stretch to not work as intended. For the case of PhotoDoc the best results were obtained with 0.5% of saturated pixels.

6. Document Binarization

Monochromatic images are the alternative of choice for most documents with no iconographic or artistic value saving storage space and bandwidth in network transmission. Most OCR tools pre-process their input images into grayscale or binary before character recognition. Reference [8] reports on the binarization of documents acquired with portable digital cameras. Fifty test images obtained with 3.2 and 4.1 Mpixel cameras (Sony DSC-P52 and DSC-P40, respectively) had their borders removed and were perspective and skew corrected before binarization, both globally and also splitting the images into 3, 6, 9 and 18 regions [16]. The following algorithms were tested:

1. Algorithm 1 – da Silva *et al.* [10];
2. Algorithm 2 – Mello *et al* [10];
3. Algorithm 3 – Pun [11];
4. Algorithm 4 – Kapur-Sahoo-Wong [12];
5. Algorithm 5 – Wu-Songde-Hanqing [13];
6. Algorithm 6 – Otsu [14];
7. Algorithm 7 – Yen-Chang-Chang [15];
8. Algorithm 8 – Johannsen-Bille [16].

According to [8], the global algorithm that yielded the best results both for visual inspection and in OCR response was the entropy based algorithm by da_Silva *et al* [10] (Figure 11). The best results obtained by applying the binarization algorithms in regions of the documents were provided by the algorithm by Kapur-Sahoo-Wong [12] with 18 regions (Figure 12).



comum, atravessa o espaço primitivamente ocupado pelo corpo jovem que, afinal, tomba sobre as costas,

a mão acompanha o movimento de arremesso da lança: o indicador em arco, paralelo ao polegar, os outros dedos em garra, para baixo: e o braço inteiro parece esticar-se por um momento, puxando o corpo que descreve rapidamente uma curva, dobrando-se, a parte esquerda meio oculta pelo escudo: a tampa de uma lata de lixo, muito suja, cinzenta, onde está escrito: LIXO: as letras amarelas, berrantes.

Esse outro corpo reveste-se de um colorido opaco, vermelha a camisa de algodão sem botões: os músculos do pescoço empurram a gola redonda e sanfonada, a calça Lee está suja de barro, o azul desgastou-se no lugar dos joelhos, as fibras claras do algodão aparecem e se estiram quase castanhas mais acima, nas coxas, que forçam o tecido, principalmente na coxa esquerda, onde se mostra um comprido volume.

O rapaz caído sente a aba do escudo em seu pescoço, tenta mover-se: agita a cabeça para os lados, ergue os joelhos, que empurram contra si o outro corpo, apertadamente, as ilhargas em junção, e o sol brilha por sobre os beirais das telhas, bem por cima da casa: os raios se projetam em cones claros no vão apertado que é passagem para automóveis, à frente da casa, e se chegam aos olhos do rapaz, fulgurantemente, obrigando-os a se fe-

44

as questões, o lencinho de seda italiana em meus pêlos, cuidadosamente aberto. Hum!

Agora, não saio mais com Marinho, sabe? Veio me procurar — os olhos avermelhados — *adoro você, quero ver você, preciso de você. Você é tão linda, ai!* Aquele papo furado! Por isso é que detesto os adultos. São complicados e têm sempre de estar fazendo alguma coisa.

Sandrinha, onde é que você vai? Você vai fazer o quê? Amar: penso, mas não digo. Para quê? Os adultos são tão complicados, sempre imaginando coisas, sôfregos para nos tacarem um rótulo. Qualquer um. Essa é direita, essa dá, essa é boa, essa é bagulho.

Não aceito rótulos! Eu, Sandra, não quero fazer nada, não quero coisa alguma, não quero ser coisa nenhuma, sabe? Quero é esfregar minhas coxas andando, ver as caras que os homens fazem no que eu passo. Sabe? Todos eles pensam que são lindos, altos, fortes, inteligentes, ricos, e que estamos loucas para deitar com eles, os merdalhões.

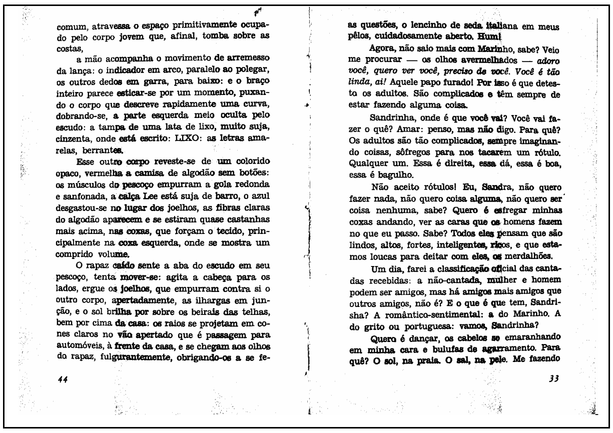
Um dia, farei a classificação oficial das cantadas recebidas: a não-cantada, mulher e homem podem ser amigos, mas há amigos mais amigos que outros amigos, não é? E o que é que tem, Sandrisha? A romântico-sentimental: a do Marinho. A do grito ou portuguesa: vamos, Sandrinha?

Quero é dançar, os cabelos se emaranhando em minha cara e bulufas de agarramento. Para quê? O sol, na praia. O sal, na pele. Me fazendo

33

Figure 11. Binarized with da Silva *et al.*(global)

Some new algorithms were tested herein including Sauvola [18], Niblack [17] and MROtsu [14]. The results obtained may be found in Figures 13 to 15. Unfortunately, the results obtained for binarizing images captured by the devices without embedded strobe flash were unsatisfactory. Further analysis and pre-processing must be studied for such images.


comum, atravessa o espaço primitivamente ocupado pelo corpo jovem que, afinal, tomba sobre as costas,

a mão acompanha o movimento de arremesso da lança: o indicador em arco, paralelo ao polegar, os outros dedos em garra, para baixo: e o braço inteiro parece esticar-se por um momento, puxando o corpo que descreve rapidamente uma curva, dobrando-se, a parte esquerda meio oculta pelo escudo: a tampa de uma lata de lixo, muito suja, cinzenta, onde está escrito: LIXO: as letras amarelas, berrantes.

Esse outro corpo reveste-se de um colorido opaco, vermelha a camisa de algodão sem botões: os músculos do pescoço empurram a gola redonda e sanfonada, a calça Lee está suja de barro, o azul desgastou-se no lugar dos joelhos, as fibras claras do algodão aparecem e se estiram quase castanhas mais acima, nas coxas, que forçam o tecido, principalmente na coxa esquerda, onde se mostra um comprido volume.

O rapaz caído sente a aba do escudo em seu pescoço, tenta mover-se: agita a cabeça para os lados, ergue os joelhos, que empurram contra si o outro corpo, apertadamente, as ilhargas em junção, e o sol brilha por sobre os beirais das telhas, bem por cima da casa: os raios se projetam em cones claros no vão apertado que é passagem para automóveis, à frente da casa, e se chegam aos olhos do rapaz, fulgurantemente, obrigando-os a se fe-

44

as questões, o lencinho de seda italiana em meus pêlos, cuidadosamente aberto. Hum!

Agora, não saio mais com Marinho, sabe? Veio me procurar — os olhos avermelhados — *adoro você, quero ver você, preciso de você. Você é tão linda, ai!* Aquele papo furado! Por isso é que detesto os adultos. São complicados e têm sempre de estar fazendo alguma coisa.

Sandrinha, onde é que você vai? Você vai fazer o quê? Amar: penso, mas não digo. Para quê? Os adultos são tão complicados, sempre imaginando coisas, sôfregos para nos tacarem um rótulo. Qualquer um. Essa é direita, essa dá, essa é boa, essa é bagulho.

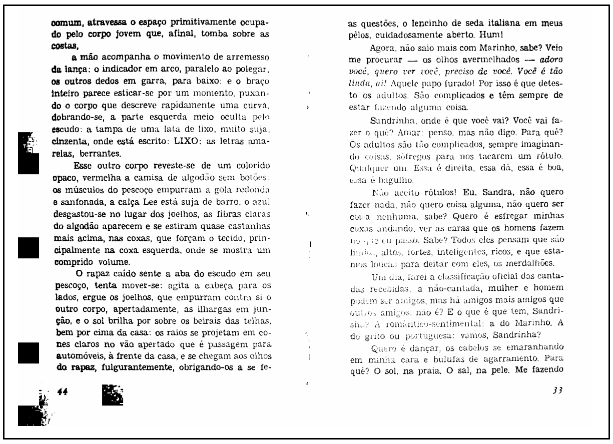
Não aceito rótulos! Eu, Sandra, não quero fazer nada, não quero coisa alguma, não quero ser coisa nenhuma, sabe? Quero é esfregar minhas coxas andando, ver as caras que os homens fazem no que eu passo. Sabe? Todos eles pensam que são lindos, altos, fortes, inteligentes, ricos, e que estamos loucas para deitar com eles, os merdalhões.

Um dia, farei a classificação oficial das cantadas recebidas: a não-cantada, mulher e homem podem ser amigos, mas há amigos mais amigos que outros amigos, não é? E o que é que tem, Sandrisha? A romântico-sentimental: a do Marinho. A do grito ou portuguesa: vamos, Sandrinha?

Quero é dançar, os cabelos se emaranhando em minha cara e bulufas de agarramento. Para quê? O sol, na praia. O sal, na pele. Me fazendo

33

Figure 12. Image binarized through the Kapur-Sahoo-Wong algorithm (18 regions).


comum, atravessa o espaço primitivamente ocupado pelo corpo jovem que, afinal, tomba sobre as costas,

a mão acompanha o movimento de arremesso da lança: o indicador em arco, paralelo ao polegar, os outros dedos em garra, para baixo: e o braço inteiro parece esticar-se por um momento, puxando o corpo que descreve rapidamente uma curva, dobrando-se, a parte esquerda meio oculta pelo escudo: a tampa de uma lata de lixo, muito suja, cinzenta, onde está escrito: LIXO: as letras amarelas, berrantes.

Esse outro corpo reveste-se de um colorido opaco, vermelha a camisa de algodão sem botões: os músculos do pescoço empurram a gola redonda e sanfonada, a calça Lee está suja de barro, o azul desgastou-se no lugar dos joelhos, as fibras claras do algodão aparecem e se estiram quase castanhas mais acima, nas coxas, que forçam o tecido, principalmente na coxa esquerda, onde se mostra um comprido volume.

O rapaz caído sente a aba do escudo em seu pescoço, tenta mover-se: agita a cabeça para os lados, ergue os joelhos, que empurram contra si o outro corpo, apertadamente, as ilhargas em junção, e o sol brilha por sobre os beirais das telhas, bem por cima da casa: os raios se projetam em cones claros no vão apertado que é passagem para automóveis, à frente da casa, e se chegam aos olhos do rapaz, fulgurantemente, obrigando-os a se fe-

44

as questões, o lencinho de seda italiana em meus pêlos, cuidadosamente aberto. Hum!

Agora, não saio mais com Marinho, sabe? Veio me procurar — os olhos avermelhados — *adoro você, quero ver você, preciso de você. Você é tão linda, ai!* Aquele papo furado! Por isso é que detesto os adultos. São complicados e têm sempre de estar fazendo alguma coisa.

Sandrinha, onde é que você vai? Você vai fazer o quê? Amar: penso, mas não digo. Para quê? Os adultos são tão complicados, sempre imaginando coisas, sôfregos para nos tacarem um rótulo. Qualquer um. Essa é direita, essa dá, essa é boa, essa é bagulho.

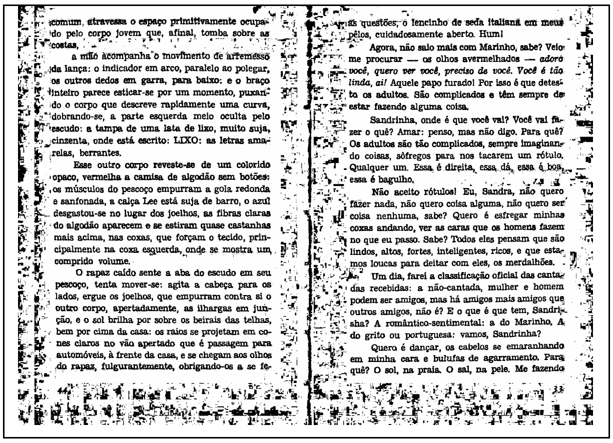
Não aceito rótulos! Eu, Sandra, não quero fazer nada, não quero coisa alguma, não quero ser coisa nenhuma, sabe? Quero é esfregar minhas coxas andando, ver as caras que os homens fazem no que eu passo. Sabe? Todos eles pensam que são lindos, altos, fortes, inteligentes, ricos, e que estamos loucas para deitar com eles, os merdalhões.

Um dia, farei a classificação oficial das cantadas recebidas: a não-cantada, mulher e homem podem ser amigos, mas há amigos mais amigos que outros amigos, não é? E o que é que tem, Sandrisha? A romântico-sentimental: a do Marinho. A do grito ou portuguesa: vamos, Sandrinha?

Quero é dançar, os cabelos se emaranhando em minha cara e bulufas de agarramento. Para quê? O sol, na praia. O sal, na pele. Me fazendo

33

Figure 13.Binarized with Sauvola (global).


comum, atravessa o espaço primitivamente ocupado pelo corpo jovem que, afinal, tomba sobre as costas,

a mão acompanha o movimento de arremesso da lança: o indicador em arco, paralelo ao polegar, os outros dedos em garra, para baixo: e o braço inteiro parece esticar-se por um momento, puxando o corpo que descreve rapidamente uma curva, dobrando-se, a parte esquerda meio oculta pelo escudo: a tampa de uma lata de lixo, muito suja, cinzenta, onde está escrito: LIXO: as letras amarelas, berrantes.

Esse outro corpo reveste-se de um colorido opaco, vermelha a camisa de algodão sem botões: os músculos do pescoço empurram a gola redonda e sanfonada, a calça Lee está suja de barro, o azul desgastou-se no lugar dos joelhos, as fibras claras do algodão aparecem e se estiram quase castanhas mais acima, nas coxas, que forçam o tecido, principalmente na coxa esquerda, onde se mostra um comprido volume.

O rapaz caído sente a aba do escudo em seu pescoço, tenta mover-se: agita a cabeça para os lados, ergue os joelhos, que empurram contra si o outro corpo, apertadamente, as ilhargas em junção, e o sol brilha por sobre os beirais das telhas, bem por cima da casa: os raios se projetam em cones claros no vão apertado que é passagem para automóveis, à frente da casa, e se chegam aos olhos do rapaz, fulgurantemente, obrigando-os a se fe-

as questões, o lencinho de seda italiana em meus pêlos, cuidadosamente aberto. Hum!

Agora, não saio mais com Marinho, sabe? Veio me procurar — os olhos avermelhados — *adoro você, quero ver você, preciso de você. Você é tão linda, ai!* Aquele papo furado! Por isso é que detesto os adultos. São complicados e têm sempre de estar fazendo alguma coisa.

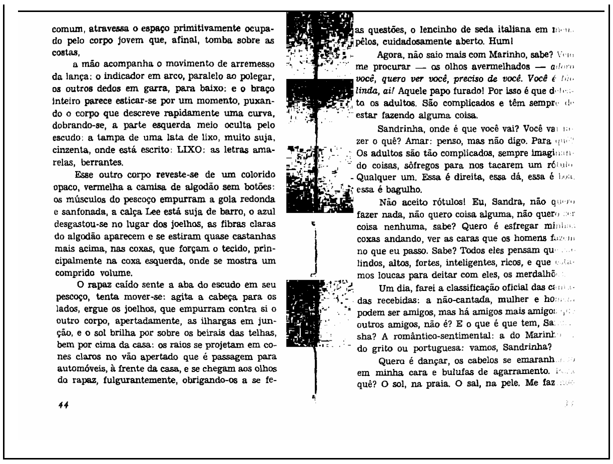
Sandrinha, onde é que você vai? Você vai fazer o quê? Amar: penso, mas não digo. Para quê? Os adultos são tão complicados, sempre imaginando coisas, sôfregos para nos tacarem um rótulo. Qualquer um. Essa é direita, essa dá, essa é boa, essa é bagulho.

Não aceito rótulos! Eu, Sandra, não quero fazer nada, não quero coisa alguma, não quero ser coisa nenhuma, sabe? Quero é esfregar minhas coxas andando, ver as caras que os homens fazem no que eu passo. Sabe? Todos eles pensam que são lindos, altos, fortes, inteligentes, ricos, e que estamos loucas para deitar com eles, os merdalhões.

Um dia, farei a classificação oficial das cantadas recebidas: a não-cantada, mulher e homem podem ser amigos, mas há amigos mais amigos que outros amigos, não é? E o que é que tem, Sandrisha? A romântico-sentimental: a do Marinho. A do grito ou portuguesa: vamos, Sandrinha?

Quero é dançar, os cabelos se emaranhando em minha cara e bulufas de agarramento. Para quê? O sol, na praia. O sal, na pele. Me fazendo

Figure 14. Binarized with Niblack (local, window_size = 50, k = -0.02)


comum, atravessa o espaço primitivamente ocupado pelo corpo jovem que, afinal, tomba sobre as costas,

a mão acompanha o movimento de arremesso da lança: o indicador em arco, paralelo ao polegar, os outros dedos em garra, para baixo: e o braço inteiro parece esticar-se por um momento, puxando o corpo que descreve rapidamente uma curva, dobrando-se, a parte esquerda meio oculta pelo escudo: a tampa de uma lata de lixo, muito suja, cinzenta, onde está escrito: LIXO: as letras amarelas, berrantes.

Esse outro corpo reveste-se de um colorido opaco, vermelha a camisa de algodão sem botões: os músculos do pescoço empurram a gola redonda e sanfonada, a calça Lee está suja de barro, o azul desgastou-se no lugar dos joelhos, as fibras claras do algodão aparecem e se estiram quase castanhas mais acima, nas coxas, que forçam o tecido, principalmente na coxa esquerda, onde se mostra um comprido volume.

O rapaz caído sente a aba do escudo em seu pescoço, tenta mover-se: agita a cabeça para os lados, ergue os joelhos, que empurram contra si o outro corpo, apertadamente, as ilhargas em junção, e o sol brilha por sobre os beirais das telhas, bem por cima da casa: os raios se projetam em cones claros no vão apertado que é passagem para automóveis, à frente da casa, e se chegam aos olhos do rapaz, fulgurantemente, obrigando-os a se fe-

44

as questões, o lencinho de seda italiana em meus pêlos, cuidadosamente aberto. Hum!

Agora, não saio mais com Marinho, sabe? Veio me procurar — os olhos avermelhados — *adoro você, quero ver você, preciso de você. Você é tão linda, ai!* Aquele papo furado! Por isso é que detesto os adultos. São complicados e têm sempre de estar fazendo alguma coisa.

Sandrinha, onde é que você vai? Você vai fazer o quê? Amar: penso, mas não digo. Para quê? Os adultos são tão complicados, sempre imaginando coisas, sôfregos para nos tacarem um rótulo. Qualquer um. Essa é direita, essa dá, essa é boa, essa é bagulho.

Não aceito rótulos! Eu, Sandra, não quero fazer nada, não quero coisa alguma, não quero ser coisa nenhuma, sabe? Quero é esfregar minhas coxas andando, ver as caras que os homens fazem no que eu passo. Sabe? Todos eles pensam que são lindos, altos, fortes, inteligentes, ricos, e que estamos loucas para deitar com eles, os merdalhões.

Um dia, farei a classificação oficial das cantadas recebidas: a não-cantada, mulher e homem podem ser amigos, mas há amigos mais amigos que outros amigos, não é? E o que é que tem, Sandrisha? A romântico-sentimental: a do Marinho. A do grito ou portuguesa: vamos, Sandrinha?

Quero é dançar, os cabelos se emaranhando em minha cara e bulufas de agarramento. Para quê? O sol, na praia. O sal, na pele. Me fazendo

33

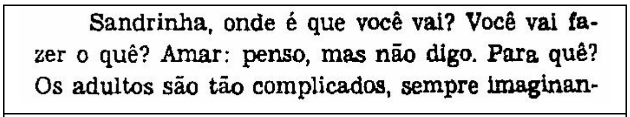
Figure 15.Binarized with MROtsu

PhotoDoc provides to the user two different ways of performing image binarization. The first one is "Automatic" and the second one opens a menu with all the algorithms above that may work either in global or image segmented mode with 2, 4, 8 and 16 regions. Local algorithm such as Niblack allows the user to define the parameters. The current implementation of "Binarization + Automatic" is set to run the algorithm by da Silva *et al.*(global) [10]. Later implementations may encompass a statistical analysis of images to choose the most suitable filter to a given image.

7. OCR

Optical Character Recognition is one of the key functionalities in any modern document processing environment [4]. Even when the recognition output is poor it may provide enough information for document indexing end keyword search. Reference [8] assesses the quality of the transcription of document images acquired with portable digital cameras with Omnipage Professional Edition version 15.0 [21]. It shows that the performance of the transcription of 50 of such documents obtained with the same models of Sony cameras used herein is close to the performance of the 100 dpi scanned counterparts.

ImageJ allows the call of executable code from within. The Tesseract [22] is an OCR Engine developed at HP Labs between 1985 and 1995. It was one of the top 3 engines in the 1995 UNLV Accuracy test. Since then, it has had little work done on it, but it is probably one of the most accurate open source OCR engines available today. The source code reads a binary, grey or color image and output text. PhotoDoc OCR whenever chosen activates the Tesseract OCR Engine. Preliminary tests such as the one made by submitting the image in the top part of Figure 15 provided as output the text in the bottom part of Figure 15.



Sandrinha, onde é que você vai? Você vai fa-
zer o quê? Amar: penso, mas não digo. Para quê?
Os adultos são tão complicados, sempre imaginan-

Sandrinha, onde E que voce vai? Voce val fa-
zer o que? Amar: penso, mas nao digo. Pars. que?
Os adultos sin tao complicados, sempre imaginan-

Figure 15 – Top: segment of textual image
Bottom: Transcribed text by Tesseract OCR

One should remark that better transcription could possibly be obtained if a Portuguese dictionary were used in conjunction with the Tesseract OCR. According to the measure of OCR performance presented in [19] the transcription above reached the figures presented on Table I.

TABLE I
ORIGINAL CHARACTER AND WORD ERRORS FOUND IN IMAGES

Character	**Replacement**	1	**Word**	**Errors**	9
	Punctuation	1		**Exclusions**	0
	G. Accents	8			
	Missing	0			
	Insertion	0			

From Table I one may see that the transcription errors were simple to be corrected as neither words nor characters are missing and there is no character insertion in the text. Besides that, word errors appeared in isolation and no word presented more than one character error in it. Most errors were due to the absence of Graphical Accents.

Conclusions and Lines for Further Works

PhotoDoc is a user friendly tool for processing document images acquired using portable digital cameras. Its current version was developed as a plug-in in ImageJ an open source portable Java library. PhotoDoc runs on users' PC and is device and manufacturer independent, working suitably from low end images acquired using cameras embedded in cell phones and palmtops to better models such as medium range, 3 and 4 Mpixels cameras, or even the state-of-the-art 6 Mpixels devices.

Several lines may be followed to provide further improvements to PhotoDoc filters. Most possibly, the most important of them is studying ways of compensating uneven illumination in documents. In the case of cameras with embedded strobe flash this may be simpler because the light source emanates from a specific point. In the case of devices that have no embedded flash illumination is provided by the environment and may come from different sources in position and power, increasing the complexity of its compensation. On the OCR front, much may be done ranging from providing dictionary help to Tesseract to performing post-processing in the textual output. This feature is already part of the newest (Jul/2007) version of Tesseract. Other open source OCR´s may also be analyzed, tested and incorporated to PhotoDoc. Preliminary testing with the open-source OCR engine OCRopus [24] showed that it was outperformed by the Tesseract OCR. For conclusive results, further testing is needed.

The PhotoDoc code is freely available at: http://www.telematica.ee.ufpe.br/sources/PhotoDoc

Acknowledgements

The work reported herein was partly sponsored by CNPq – Conselho Nacional de Desenvolvimento Científico e Tecnológico, Brazilian Government.

References

[1] D.Doermann, J.Liang, H. Li, "Progress in Camera-Based Document Image Analysis," ICDAR'03, V(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] K.C.Fan, Y.K.Wang, T.R.Lay, Marginal noise removal of document images, Patt.Recognition. 35, 2593-2611, 2002.

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

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

[6] L. Jagannathan and C. V. Jawahar, "Perspective correction methods for camera based document analysis," pp. 148–154, CBDAR 2005/ICDAR 2005, Seoul, Korea. 2005.

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

[8] R.D.Lins, A.R.Gomes e Silva and G.Pereira e Silva, Assessing and Improving the Quality of Document Images Acquired with Portable Digital Cameras, ICDAR´07, Curitiba, Brasil, 2007.

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

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

[11] T. Pun, Entropic Thresholding, A New Approach, C. Graphics and Image Processing, 16(3), 1981.

[12] J. N. Kapur, P. K. Sahoo and A. K. C. Wong. A New Method for Gray-Level Picture Thresholding using the Entropy of the Histogram, Computer Vision, Graphics and Image Processing, 29(3), 1985.

[13] L. U. Wu, M. A. Songde, and L. U. Hanqing, An effective entropic thresholding for ultrasonic imaging, ICPR'98: Intl. Conf. Patt. Recog., pp. 1522–1524 (1998).

[14] M.R .Gupta, N.P. Jacobson and E.K. Garcia. OCR binarization and image pre-processing for searching historical documents. Pattern Recognition 40 (2007): 389-397 (2007).

[15] J. C. Yen,*et al.* A new criterion for automatic multilevel thresholding. IEEE T. Image Process. IP-4, 370–378 (1995).

[16] G. Johannsen and J. Bille. A threshold selection method using information measures. ICPR'82: 140–143 (1982).

[17] W.Niblack, "An Introduction to Image Processing" pp.115-116, Prentice-Hall,Englewood Cliffs,NJ(1986).

[18] J. Sauvola, M. Pietikainen, Adaptive document image binarization, Pattern Recognition 33 (2) (2000) 225–236.

[19] 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.

[20] ImageJ http://rsb.info.nih.gov/ij/

[21] Nuance Corp. http://www.nuance.com/omnipage/professional

[22] Tesseract http://code.google.com/p/tesseract-ocr/

[23] JAI (Java Advanced Imaging). https://jai.dev.java.net.

[24] OCRopus .http://www.ocropus.org.