Journal of Universal Computer Science, vol. 15, no. 18 (2009), 3364-3375
submitted: 20/10/09, accepted: 23/12/09, appeared: 28/12/09 © J.UCS

Automatically Deciding if a Document was Scanned or Photographed

Gabriel Pereira e Silva, Rafael Dueire Lins, Brenno Miro
(Federal University of Pernambuco, Recife, Brazil
gfps@cin.ufpe.br, rdl@ufpe.br, brennope@hotmail.com)

Steven J. Simske
(HP Labs, Fort Collins, USA
steven.simske@hp.com)

Marcelo Thielo
(HP Labs, Porto Alegre, Brazil
marcelo.resende.thielo@hp.com)

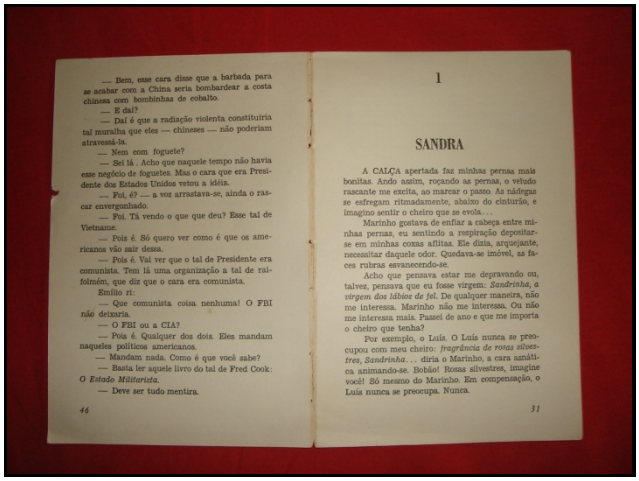
Abstract: Portable digital cameras are being used widely by students and professionals in different fields as a practical way to digitize documents. Tools such as PhotoDoc enable the batch processing of such documents, performing automatic border removal and perspective correction. A PhotoDoc processed document and a scanned one look very similar to the human eye if both are in true color. However, if one tries to automatically binarize a batch of documents digitized from portable cameras compared to scanners, they have different features. The knowledge of their source is fundamental for successful processing. This paper presents a classification strategy to distinguish between scanned and photographed documents. Over 16,000 documents were tested with a correct classification rate of over 99.96%.

Keywords: MPEG-7, content-based Multimedia Retrieval, Hypermedia systems, Web-based services, XML, Semantic Web, Multimedia
Categories: H.3.1, H.3.2, H.3.3, H.3.7, H.5.1

1 Introduction

Portable digital cameras are ubiquitous. Either in standalone versions, or incorporated in cell phones, the quality of the images has risen at a fast pace while their price has dropped drastically. Such pervasiveness has given rise to unforeseen applications such as using portable digital cameras for digitalizing documents by users of many different professional areas. For instance, students and professionals are taking photos of writing boards instead of taking notes; lawyers are taking photos of legal processes instead of going through a difficult bureaucratic path to take documents out of court to photocopy them, etc. This new research area [Doermann, 03] [Liang, 05] is evolving fast in many directions. People in general are non-specialized in image processing and claim for new algorithms, tools and processing environments to be able to provide simple and user-friendly ways of visualizing, printing, transcribing, compressing, storing and transmitting document images. Figure 1 presents an example of a document acquired with a portable digital camera. Reference [Lins, 07] points out some particular problems that arise in this document digitalization process:

the first 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 portable cameras have no fixed mechanical support, often there is some 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 arc, 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 non-uniform illumination.


— Bem, esse cara disse que a barbada para se acabar com a China seria bombardear a costa chinesa com bombinhas de cobalto.
— E daí?
— Daí é que a radiação violenta constituiria tal muralha que eles — chineses — não poderiam atravessá-la.
— Nem com foguete?
— Sei lá. Acho que naquele tempo não havia esse negócio de foguetes. Mas o cara que era Presidente dos Estados Unidos vetou a idéia.
— Foi, é? — a voz arrastava-se, ainda o rascar envergonhado.
— Foi. Tá vendo o que que deu? Esse tal de Vietname.
— Pois é. Só quero ver como é que os americanos vão sair dessa.
— Pois é. Vai ver que o tal de Presidente era comunista. Tem lá uma organização a tal de raifolmém, que diz que o cara era comunista.
Emílio ri:
— Que comunista coisa nenhuma! O FBI não deixaria.
— O FBI ou a CIA?
— Pois é. Qualquer dos dois. Eles mandam naqueles políticos americanos.
— Mandam nada. Como é que você sabe?
— Basta ler aquele livro do tal de Fred Cook: *O Estado Militarista*.
— Deve ser tudo mentira.

46

1

SANDRA

A CALÇA apertada faz minhas pernas mais bonitas. Ando assim, roçando as pernas, o veludo riscante me excita, ao marcar o passo. As nádegas se esfregam ritmadamente, abaixo do cinturão, e imagino sentir o cheiro que se evola...

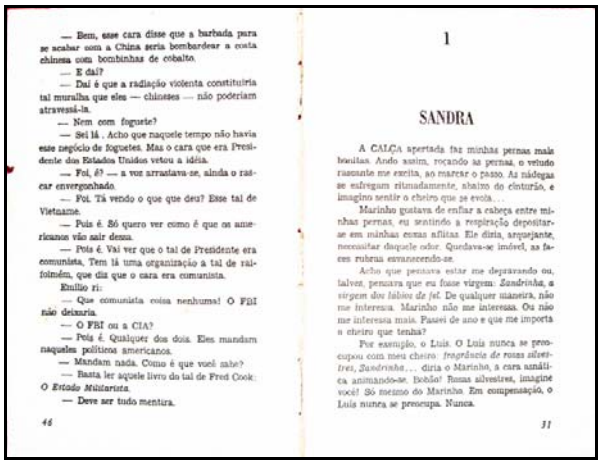
Marinho gostava de enfiar a cabeça entre minhas pernas, eu sentindo a respiração depositar-se em minhas coxas afilitas. Ele dizia, arquejante, necessitar daquele odor. Quedava-se imóvel, as faces rubras esvanecendo-se.

Acho que pensava estar me depravando ou, talvez, pensava que eu fosse virgem: *Sandrinha, a virgem dos lábios de fel*. De qualquer maneira, não me interessa. Marinho não me interessa. Ou não me interessa mais. Passei de ano e que me importa o cheiro que tenha?

Por exemplo, o Luís. O Luís nunca se preocupou com meu cheiro: *fragrância de rosas silvestres, Sandrinha*... diria o Marinho, a cara sanática animando-se. Bobão! Rosas silvestres, imagine você! Só mesmo do Marinho. Em compensação, o Luís nunca se preocupa. Nunca.

31

Figure 1: Example of a photo document


— Bem, esse cara disse que a barbada para se acabar com a China seria bombardear a costa chinesa com bombinhas de cobalto.
— E daí?
— Daí é que a radiação violenta constituiria tal muralha que eles — chineses — não poderiam atravessá-la.
— Nem com foguete?
— Sei lá. Acho que naquele tempo não havia esse negócio de foguetes. Mas o cara que era Presidente dos Estados Unidos vetou a idéia.
— Foi, é? — a voz arrastava-se, ainda o rascar envergonhado.
— Foi. Tá vendo o que que deu? Esse tal de Vietname.
— Pois é. Só quero ver como é que os americanos vão sair dessa.
— Pois é. Vai ver que o tal de Presidente era comunista. Tem lá uma organização a tal de raifolmém, que diz que o cara era comunista.
Emílio ri:
— Que comunista coisa nenhuma! O FBI não deixaria.
— O FBI ou a CIA?
— Pois é. Qualquer dos dois. Eles mandam naqueles políticos americanos.
— Mandam nada. Como é que você sabe?
— Basta ler aquele livro do tal de Fred Cook: *O Estado Militarista*.
— Deve ser tudo mentira.

46

1

SANDRA

A CALÇA apertada faz minhas pernas mais bonitas. Ando assim, roçando as pernas, o veludo riscante me excita, ao marcar o passo. As nádegas se esfregam ritmadamente, abaixo do cinturão, e imagino sentir o cheiro que se evola...

Marinho gostava de enfiar a cabeça entre minhas pernas, eu sentindo a respiração depositar-se em minhas coxas afilitas. Ele dizia, arquejante, necessitar daquele odor. Quedava-se imóvel, as faces rubras esvanecendo-se.

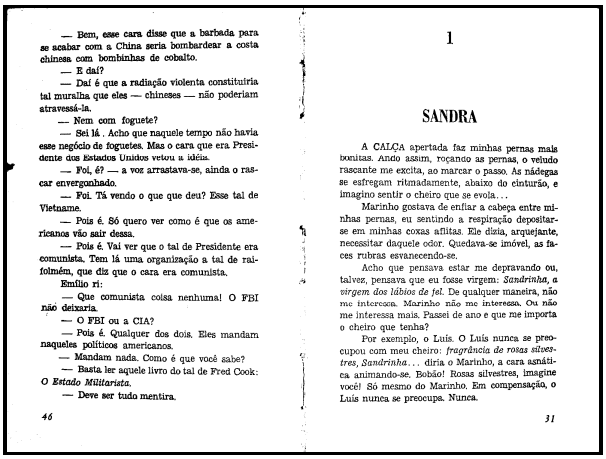
Acho que pensava estar me depravando ou, talvez, pensava que eu fosse virgem: *Sandrinha, a virgem dos lábios de fel*. De qualquer maneira, não me interessa. Marinho não me interessa. Ou não me interessa mais. Passei de ano e que me importa o cheiro que tenha?

Por exemplo, o Luís. O Luís nunca se preocupou com meu cheiro: *fragrância de rosas silvestres, Sandrinha*... diria o Marinho, a cara sanática animando-se. Bobão! Rosas silvestres, imagine você! Só mesmo do Marinho. Em compensação, o Luís nunca se preocupa. Nunca.

31

Figure 2: PhotoDoc processed photo document

Reference [Silva, 07] presents PhotoDoc, a freely available toolbox for processing document images acquired with portable digital cameras, which is implemented as a plugin in ImageJ [ImageJ, 09]. Figure 2 presents an example of a photo document processed with PhotoDoc, which is implemented as a Plugin in ImageJ [ImageJ, 09].



— Bem, esse cara disse que a barbada para se acabar com a China seria bombardear a costa chinesa com bombinhas de cobalto.
— E daí?
— Daí é que a radiação violenta constituiria tal muralha que eles — chineses — não poderiam atravessá-la.
— Nem com foguete?
— Sei lá. Acho que naquele tempo não havia esse negócio de foguetes. Mas o cara que era Presidente dos Estados Unidos teteu a idéia.
— Foi, é? — a voz arrastava-se, ainda o rascar envergonhado.
— Foi. Tá vendo o que que deu? Esse tal de Vietname.
— Pois é. Só quero ver como é que os americanos vão sair dessa.
— Pois é. Vai ver que o tal de Presidente era comunista. Tem lá uma organização a tal de ralfolmém, que diz que o cara era comunista.
Emílio ri:
— Que comunista coisa nenhuma! O FBI não deixaria.
— O FBI ou a CIA?
— Pois é. Qualquer dos dois. Eles mandam naqueles políticos americanos.
— Mandam nada. Como é que você sabe?
— Basta ler aquele livro do tal de Fred Cook: *O Estado Militarista.*
— Deve ser tudo mentira.

46

1

SANDRA

A CALÇA apertada faz minhas pernas mais bonitas. Ando assim, roçando as pernas, o veludo rascante me excita, ao marcar o passo. As nádegas se esfregam ritmadamente, abaixo do cinturão, e imagino sentir o cheiro que se evola...

Marinho gostava de enfiar a cabeça entre minhas pernas, eu sentindo a respiração depositar-se em minhas coxas aflitas. Ele dizia, arquejante, necessitar daquele odor. Quedava-se imóvel, as faces rubras esvanecendo-se.

Acho que pensava estar me depravando ou, talvez, pensava que eu fosse virgem: *Sandrinha, a virgem dos lábios de fel.* De qualquer maneira, não me interessa. Marinho não me interessa. Ou não me interessa mais. Passei de ano e que me importa o cheiro que tenha?

Por exemplo, o Luís. O Luís nunca se preocupou com meu cheiro: *fragrância de rosas silvestres, Sandrinha...* diria o Marinho, a cara asnática animando-se. Bobão! Rosas silvestres, imagine você! Só mesmo do Marinho. Em compensação, o Luís nunca se preocupa. Nunca.

31

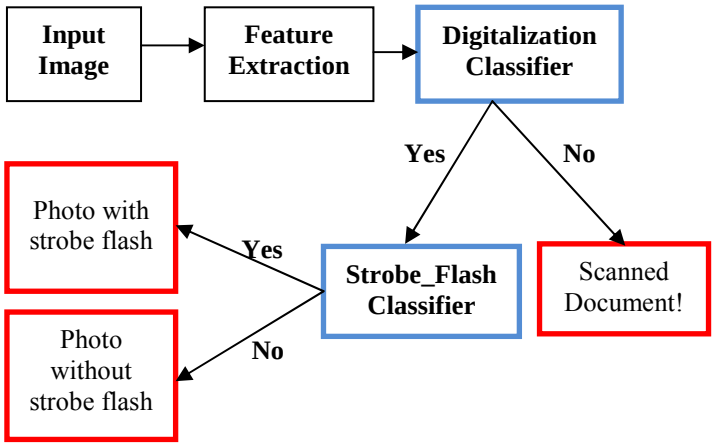
Figure 3: Binarization of a photo document using a global algorithm [Otsu, 79]

Illumination is less uniform for documents captured with digital cameras in comparison to scanned images. It may not be easy for a person to differentiate between a document processed using PhotoDoc and the same document captured with a scanner. Distinguishing between them is important in the case, for example, of image binarization. The irregular illumination in general tends to provide shaded black areas in the direct binarization of a photo document as shown in Figure 3. Color images such as the one in Figure 4, both scanned and photographed, are also present in the test set used here.

Once the digitalization device for a given image is known, one has valuable information about the nature of the possible noises present in the captured image. According to the taxonomy proposed in [Lins, 09a] the kinds of noises present in scanned documents are *physical* noises (considering an adequate scanner manipulation and its perfect functioning). Physical noises, such as stains, folding marks, annotations, etc. may be difficult to be removed, and some may even consider them as part of the document information. On the other hand, photographed documents, besides being passive of having the same noises as the scanned ones, may include the *digitalizatio*n noises that if known their existence is known a priori may be suitably removed. One of the few digitalization noises one finds in scanned documents is found in the digitalization of bound volumes in flatbed scanners, as the distance from the object to the flatbed caused a document warping. Also, in this case, it is of paramount importance the information of how the document image was obtained as different de-warping algorithms are used depending on the digitalization source.

Figure 4: PhotoDoc processed color photo document

This paper focuses on a classification strategy to distinguish, in a batch of documents, the scanned documents from documents acquired with portable digital cameras. Camera documents are further classified based on whether a strobe flash was used, as shown below:



The classification strategy depends on the following:

- The choice of the set of features to be extracted. The features selected must provide enough elements to distinguish between the clusters of interest. Feature extraction has also impact in classification time.

- The choice of the classifier. Some classifiers are able to perform better than some others depending on the nature and class of the problem, the representativeness of the features selected, etc.
- The quality and size of the training set used for the classifier. The training set must be carefully chosen to encompass the whole diversity of the universe of objects to be classified, with as less redundancy as possible.

This paper shows that the classifier presented in reference [Lins, 09] presents excellent performance for distinguishing between documents obtained from scanners and portable digital cameras with or without the strobe flash on. The results obtained are compared with the classification strategy in reference [Lins, 09a]. The new classifier not only reached a higher correct classification rate, but besides that, elapsed much less time for feature extraction and classification. The classifier presented herein was implemented using Weka [Witten, 05] [Weka, 09], an excellent, user friendly and open-source platform developed at the University of Waikato. The test set encompassed 17,781 documents of which only 3 documents were misclassified, yielding a correct classification rate of 99.98%.

2 Experiments Performed

The starting point for this work was collecting images that are representative of the two different clusters of interest: scanned and photographed documents. The photographed documents were split in two sub-clusters: images acquired with and without the strobe flash on.

The test set for the photo document cluster is formed by 9,573 documents acquired with a Sony Cybershot digital camera DSC-W55 in 5 and 7.2 Mpixels, with and without mechanical support, in-built strobe flash on and off. In the camera set there are also 404 photos taken with a portable camera Sony DSC-S40 and 60 photos from a cell phone LG Shine ME970, both without any mechanical support. All photo documents were processed with PhotoDoc [Silva, 07] a photo document processing tool that crops the framing border and corrects perspective and skew, should be classified as "document".

The 6,444 of the scanned documents were 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. Although the jpeg file format may be seen as unsuitable for such kind of image it is often used by people in general [Lins, 04]. In addition, 300 images were acquired with a scanner HP 5300c in 300 dpi, true color, stored in tiff (uncompressed) and 1000 jpeg images in different resolutions were collected from the Internet.

Table 1 shows the numbers of images per file format in the test set.

	JPG	PNG	TIFF	BMP	Total
Photo	10,037	***	***	***	10,037
Scanned	2,611	1,611	2,522	1,000	7,744
Total	12,648	1,611	2,522	1,000	17,821

Table 1: Images per file formats

2.1 Features Tested

The choice of the features to be extracted and tested is the key to the success and performance of the classification. Image entropy is often used as the key for classification [Simske, 05]. It has a large computational cost, however. Entropy calculation demands a scan in the image to calculate the relative frequency of a given color, for instance, which is than multiplied for its logarithm and added up. The classifier described in reference [Simske, 05] is based on the binary classification approach, and assumes a Gaussian distribution for each of the features. Its performance degrades in proportion to the non-Gaussian nature of the data. We designate this the entropy-based classifier, as the set of features chosen herein has entropy calculation as its key.

The work presented in reference [Lins, 09] proposes a new classification strategy that assumes that decreasing the gamut of an image, analyzed together with its grey scale and monochromatic equivalents would provide enough elements for a fast and efficient image classification. The features tested are:

- Palette (true-color/grayscale)
- Gamut
- Conversion into Grayscale
- Gamut in Grayscale (if RGB)
- Conversion into Binary (Otsu)
- Number of black pixels in binary image.
- (#Black_pixels/Total_#_pixels)*100%
- (Gamut/Palette)*100% (true-color/grayscale)

Image binarization is performed by using Otsu [Otsu, 79] algorithm. The data above are extracted for each image and placed in a vector of features. The classification strategy adopted herein follows the feature set proposed in reference [Lins, 09]. The training set used had size of about 8% of the test set and was selected from within the images of Sony Cybershot digital camera DSC-W55 in 5 and 7.2 Mpixel and the Ricoh Affício 1075 flatbed scanner in 100, 200 and 300 dpi. The images in the training set were not part of the test set. The entropy-based classifier [Simske, 05] was used to compare the results obtained. Both classifiers used the same training and test sets.

2.2 Sub-sampling

Image sub-sampling may be used as a way to reduce the time elapsed in feature extraction of images to be classified. The key points in image sub-sampling are:

1. The larger the image file, the richer in data redundancy; thus, if the redundant data are thrown away the efficiency both in feature-collection time and classification may rise.
2. The selection of points to be analyzed for feature collection should not be random. It should somehow provide a "reduced" version of the original image (although in some cases it may be distorted by unequal scaling!).

```
size = height*width
• If size ≤ 300,000  break;
• If 300,000< size ≤ 500,000:
     remove even lines or columns
                                   (whatever the larger);
• If 500,000 < size ≤ 700,000:
     remove even lines and columns;
• If 700,000 < size ≤ 900,000:
     remove 2 lines in every 3 lines and even columns,
                                   (if height>width)
     remove even lines and
          2 columns in every 3 columns, otherwise;
• If 900,000 < size  remove 2 lines and 2 columns
     in every 3 lines and columns;
```

Code for the "cascaded" sub-sampler

3 Results

The results of classification are presented in two steps. The group of results was obtained with 16,017 images digitized with the Sony Cybershot digital camera DSC-W55 in 5 and 7.2 Mpixel, and the Ricoh Afficio 1075 flatbed scanner in 100, 200 and 300 dpi. Several different classifiers implement in Weka [Witten, 05] [Weka, 09] were tested. Random forests provided the best classification results amongst the statistical classifiers. A Multi-layer Perceptron (MLP) neural classifier was also tested and the best results obtained for eight neurons on two layers.

The confusion matrices obtained by the classifiers that used the proposed set of features are shown in Table 2. The entry "Photo +sf" stands for the document images photographed with the strobe flash on, while "Photo -sf" denotes it off.

Table 2 points out that the Random Forest statistical classifier [Breiman, 01] with 10 trees presented the best classification results.

Classifier		Photo +sf	Photo -sf	Scanned	Accuracy %
Random Forest 5-trees	Photo +sf	4029	0	0	100
	Photo -sf	4	5534	6	99.81962
	Scanned	0	0	6444	100
Random Forest 10-trees	Photo +sf	4,029	0	0	100
	Photo -sf	4	5,537	3	99.8737
	Scanned	0	0	6,444	100
Random Forest 15-trees	Photo +sf	4029	0	0	100
	Photo -sf	7	5535	2	99.83766
	Scanned	0	0	6444	100
Random Forest 20-trees	Photo +sf	4029	0	0	100
	Photo -sf	8	5534	2	99.81962
	Scanned	0	0	6444	100
Random Forest 100-trees	Photo +sf	4029	0	0	100
	Photo -sf	7	5535	2	99.83766
	Scanned	0	0	6444	100
MLP	Photo +sf	4029	0	0	100
	Photo -sf	13	5531	0	99.76551
	Scanned	0	1	6443	99.98448
RBF	Photo +sf	3975	54	0	98.65972
	Photo -sf	47	5497	0	99.15224
	Scanned	0	5	6439	99.92241

Table 2: Confusion matrix of the proposed classifier with 16,017 original images

Table 3 shows the results obtained for the same set of classifiers trained and tested with sub-sampled images.

Classifier		Photo +sf	Photo -sf	Scanned	Accuracy %
Random Forest 5-trees	**Photo +sf**	4029	0	0	100
	Photo -sf	4	5534	6	99.81962
	Scanned	4029	0	0	100
Random Forest 10-trees	**Photo +sf**	2	5525	17	99.65729
	Photo -sf	0	0	6444	100
	Scanned	4,029	0	0	100
Random Forest 15-trees	**Photo +sf**	0	5,540	4	99.9278
	Photo -sf	0	0	6,444	100
	Scanned	4029	0	0	100
Random Forest 20-trees	**Photo +sf**	2	5539	3	99.90981
	Photo -sf	0	0	6444	100
	Scanned	4029	0	0	100
Random Forest 100-trees	**Photo +sf**	2	5539	3	99.90981
	Photo -sf	0	0	6444	100
	Scanned	4029	0	0	100
MLP	**Photo +sf**	3	5539	2	99.90981
	Photo -sf	0	0	6444	100
	Scanned	4029	0	0	100
RBF	**Photo +sf**	3971	58	0	98.56043
	Photo -sf	48	5496	0	99.13419
	Scanned	0	5	6439	99.92240

Table 3: Confusion matrix of the proposed classifiers with 16,017 subsampled images

Using sub-sampling, the relative performance of the classifiers was stable. Again, Random-forests with 10 trees provided the best results. Curiously, sub-sampling, besides speeding-up the feature extraction time, increased correct classification rate. One important point worth noting is that the misclassified documents, when binarized using a global algorithm, performed satisfactorily. Having the strobe flash off may resemble a scanned document, provided there is enough uniform illumination from the environment. Then, the misclassification errors in this case do not cause serious problems to the overall process.

Now, the entropy-based set of features for classification proposed by reference [Simske, 05] was tested on the original data and the results obtained are presented on Table 4.

Proposed Classifier	Photo +sf	Photo -sf	Scanned	Accuracy
Photo +sf	3402	272	355	84.4378 %
Photo -sf	71	4466	1007	80.5555 %
Scanned	32	152	6260	97.1446 %

Table 4: Confusion matrix of the entropy-based classifier with original images

The results obtained for entropy based classifier with subsampled images are shown on Table 5.

Proposed Classifier	Photo +sf	Photo -sf	Scanned	Accuracy
Photo +sf	3402	270	357	84.4378 %
Photo -sf	69	4562	913	82.2871 %
Scanned	24	158	6262	97.1756 %

Table 5: Confusion matrix of the entropy-based classifier with subsampled images

The comparison between the entropy-based and the new one proposed here shows that the new one is about 10% better than the previous one.

The classification of the 404 photos taken with a portable camera Sony DSC-S40 and 60 photos from a cell phone LG Shine ME970, both without any mechanical support, and the images obtained with scanner HP 5300c and the images collected from the Internet did not bring any misclassification at all.

4 Time Performance

Table 6 presents the feature extraction and classification times along with the programming language used for implementation. Besides classification accuracy per cluster, the average feature extraction and classification times are presented. One

should also remark that there is a difference in time scale between feature extraction and classification.

	Feature extraction		Classification	
	Time (s)	Language	Time (ms)	Language
Original	**0.4174**	**C++**	**0.12**	**C#**
Subsampled	**0.1470**	**C++**	**0.12**	**C#**
Original	**0.4174**	**C++**	**0.10**	**C++**
Subsampled	**0.1470**	**C++**	**0.10**	**C++**
Entropy Or.	**1.4576**	**C#**	**6.13**	**C#**
Entropy Ss.	**0.497**	**C#**	**6.13**	**C#**

Table 6: Feature extraction and classification times

Table 6 shows that the set of features used for image classification based on image palette conversion outperforms the entropy-based classifier by a factor of four for feature extraction and by a factor of fifty for image classification. ("Entropy Or." stands for the Entropy-based classifier [Simske, 05] with the original images, while "Entropy Ss." corresponds to the Entropy-based classifier with subsampled images).

The figures of the relative performance of the classifiers for the proposed set of features varying the number of trees and the MLP implemented in Weka (Java) are shown on Table 7.

Proposed Classifier	Java -Time (ms)	C++ - Time (ms)
Random Forest 5-trees	5.4	3.7
Random Forest 10-trees	6.1	5.0
Random Forest 15-trees	6.7	5.1
Random Forest 20-trees	7.9	6.4
Random Forest 100-trees	9.5	6.9
MLP	6.8	***

Table 07: Classification times in Random Forest [Breiman, 01]

One may observe that the Random-forests classifier reaches the best trade-off classification and time efficiency.

5 Conclusions

Weka [Witten, 05] [Weka, 09] has shown to be an excellent test bed for statistical analysis. The choice for a Random tree classifier was made after performing several experiments with the large number of alternatives offered by Weka, although results did not vary widely. Amongst them a preliminary comparison between the new statistical classifier proposed here and a MLP neural classifier provided worse results (around 94% of accuracy).

The choice of the images in the training set is of paramount importance to the performance of the classifier. They must be representative of the whole universe of images in a cluster.

The classification scheme presented in this paper increased the correct classification rate by more than 10%. This automatic classification allows distinguishing scanned from photographed document images yielding better ways to suitably process document images.

Acknowledgements

Research presented herein was partly sponsored by CNPq - Conselho Nacional de Desenvolvimento Científico e Tecnológico, and HP - UFPE Project TechDoc sponsored by MCT, both of the Brazilian Government.

References

[Breiman, 01] Breiman, L.: Random Forests, Machine Learning, 45(1), pp. 5-32, 2001.

[Doermann, 03] Doermann D. and Liang J., Li H.: Progress in Camera-Based Document Image Analysis, ICDAR'03, Volume (1): 606, 2003.

[Silva, 07] Silva, G.P and Lins, R.D.: PhotoDoc: A Toolbox for Processing Document Images Acquired Using Portable Digital Cameras. CBDAR 2007, pp.107-114, 2007.

[Liang, 05] Liang, J., Doermann, D., and Li, H.: Camera-Based Analysis of Text and Documents: A Survey. International Journal on Document Analysis and Recognition, 2005.

[Lins, 04] Lins, R.D. and Machado, D.S.A.: A Comparative Study of File Formats for Image Storage and Transmission, vol. 13(1):175-183, Journal of Electronic Imaging, 2004.

[Lins, 07] Lins, R.D., Gomes, A.R. e Silva and Silva, G.P.: Enhancing Document Images Acquired Using Portable Digital Cameras, ICIAR´07, LNCS 4633, pp. 1229-1241, Springer-Verlag, 2007.

[Lins, 09] Lins, R.D., Silva, G.P, Simske, S.J., Fan,J., Shaw, M.S., Sá, P., Thiello,M.R.: Image Classification to Improve Printing Quality of Mixed-Type Documents. ICDAR 2009, IAPR Press, Barcelona, 2009.

[Lins, 09a] Lins, R.D.: A taxonomy for noises in paper documents – the physical noises, LNCS 5624, pp. 844-854, Springer Verlag, 2009.

[Otsu, 79] Otsu, N.: A threshold selection method from gray level histograms. IEEETrans.Syst.Man Cybern. Vol. (9):62-66, 1979.

[Simske, 05] Simske, S.J.: Low-resolution photo/drawing classification: metrics, method and archiving optimization, Proceedings IEEE ICIP, IEEE, Genoa, Italy, pp. 534-537, 2005.

[Witten, 05] Witten, I.H. , Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques (2nd Edition) -- Morgan Kaufmann, June 2005. ISBN 0-12-088407-0.

[ImageJ, 09] ImageJ http://rsb.info.nih.gov/ij/; last visited 09.02.2010.

[Weka, 09] Weka 3: Data Mining Software in Java, website http://www.cs.waikato.ac.nz/ml/weka/; last visited 09.02.2010.