Automatically Discriminating between Digital and Scanned Photographs

Rafael Dueire Lins
Gabriel de França Pereira e Silva
Universidade Federal de Pernambuco
Recife, Brazil
{gabriel.psilva, rdl}@ufpe.br

Steven J. Simske
Hewlett-Packard Labs
Fort Collins, USA
steven.simske@hp.com

Abstract — True digital photos and the digital images of scanned photographs have very different properties. The illumination pattern and palette of the two kinds of images are different. Being able to distinguish between them is important, as each of these should be handled during printing with a class-specific pipeline of image transformation algorithms, and misclassification results in detrimental imaging effects. This paper presents an automatic classifier to discriminate between the two sources. The classifier proposed is fast enough to be embedded in the driver of any printing device today.

Keywords*- digital photo, analogical photos, printing.*

I. INTRODUCTION

Color perception goes beyond the psycho-physical phenomenon usually described in the literature. Cultural elements also influence the way people see printed images. One typical example of that is photo printing – the use of a color palette in place of a richer set of hues looks unpleasantly flat and pale. People expect photos to have sharp bright colors, rich in different hues, while keeping an overall rich color balance. Figure 1 presents an image which was obtained by scanning a printed analogical photograph (saved in JPEG) with a resolution of 600 dots per inch, while Figure 02 exhibits an image of a digital photograph. The difference between the two images is easily observable. The scanned photograph looks much "paler" than the digital photograph, but whoever scans and reprints a photograph expects it to look as "sharp and bright" as the digital one.

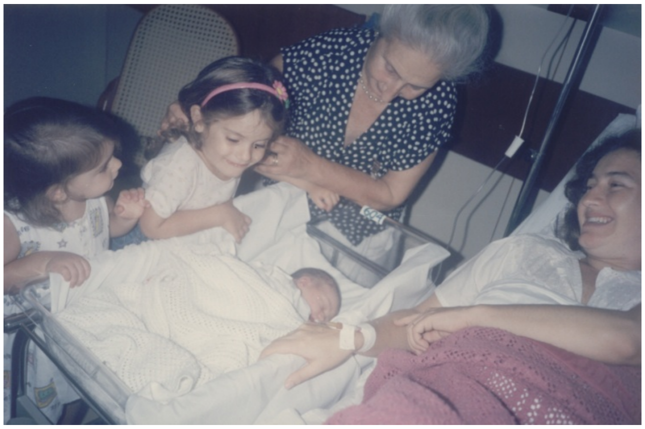

Figure 1. Scanned photo 600 dpi – 3.57 Mbytes – JPEG

Functional image classification is the assignment of different image types to separate classes to optimize their rendering for printing or another specific end task, and is an important area of research in the publishing and industries. To meet customer expectations, the printer needs to print each image with the correct color palette, balance and other image processing operations applied. To perform this task automatically in the absence of image metadata, the printer must perform accurate image classification based solely on the image raster information. This classification must be both accurate and fast due to the constraints of the printer embedded processor.

Image classification is used in all-in-one and multi-functional devices to differentially render images belonging to different clusters. In particular, document, photo and logo images require widely different imaging pipelines to optimize their appearance when copied or printed. Documents (text, tables), for example, require sharpening that would damage the appearance of photos and logos. Logos use a palette that would "posterize" photos. Photos, in turn, can be rendered with a lower resolution (but greater bit depth) than either documents or logos. Reference [6] presents an image classifier that replaced the previous one embedded in HP printers. The new classifier [6] largely outperformed the previous one [11] both in accuracy and time performance, an important feature for an algorithm to be embedded in low-cost, low-power consuming, fast printing devices. The present paper introduces a new classifier for photos: whether they are digital or from a printed (hardcopy) source which was later digitized.

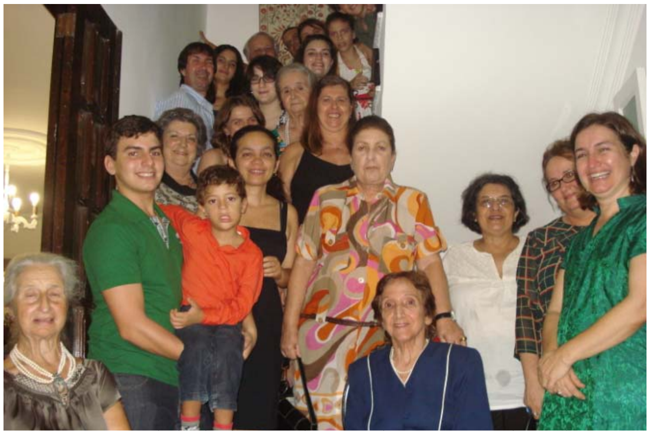

Figure 2. Digital photo taken with a Sony Cyber-shot 7.2 MPixels portable camera. 3.01 MBytes - JPEG

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.258

II. Motivation

Image clustering [2][3] has a long tradition in the database community for efficient information retrieval from image databases [4][8]. The classifier described in reference [6] is able to discriminate with over ninety percent accuracy between three clusters: documents, logo and photos. The document clusters include scanned, digitally generated (such as image files from pdf files), and photographed ones (processed through PhotoDoc, a software platform that removes borders, corrects perspective and skew, etc). The logo cluster includes color and monochromatic images. The photo cluster covers a wide range of images varying in palette (color, sepia, and monochromatic) and theme (landscapes, people, objects, and PhotoDoc unprocessed documents). Figure 3 presents examples of images in the three clusters.

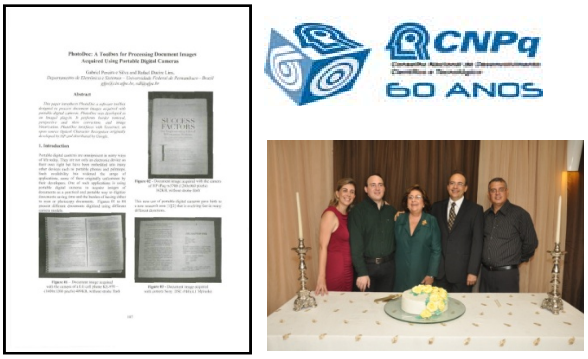
PhotoDoc: A Toolbox for Processing Document Images Acquired Using Portable Digital Cameras
CNPq
Conselho Nacional de Desenvolvimento Científico e Tecnológico
60 ANOS

Figure 03. Examples images of the different clusters discriminated by the classifier in [6]. Left-document, Right_T- logo, Right_B-photo

It is extremely difficult to an observer to distinguish between a scanned and PhotoDoc processed document image, until one tries to binarize them. The illumination pattern of the photographed document, although imperceptible to the naked eye, is non-uniform and the direct binarization using a global algorithm leads to some black areas as may be observed in Figure 04.

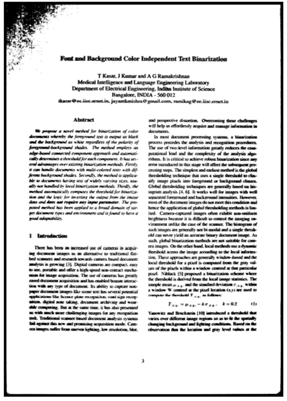
Font and Background Color Independent Text Binarization
T Kasar, J Kumar and A G Ramakrishnan

Figure 4. Binarization of a photo document using Otsu global algorithm [7].

Thus, for batch processing such images an image classifier to discriminate between the two sources is most desirable and this was the motivation for the work reported in [10], which also served as inspiration to the current one.

III. Methodology

The "Photo" cluster in [6] encompassed many different sorts of photos, which ranged from people (approximately 4,000), landscapes (about 3,700), objects (just under 400) and even documents (500) in different file formats (7,476 JPEG, 35 TIFF, and 457 BMP) and varied from true-color to grey scale ones. The resolution also varied widely from VGA (480x640 pixels) to 7.2 Mpixels. The photos were collected from family albums of the people linked to the authors to ones obtained from the Internet.

The current study is far more restrictive and limited the test set of to only one theme – people. The starting point for this work was scanning a set of photos from a family photo album. All photos were printed in 10 x 15 cm on glossy paper without texture at a professional printing house. They were scanned with a 600 dots per inch resolution with an HP flatbed scanner model ScanJet 5300C. The photos were stored in JPEG file format with 1% loss, the standard used by portable digital cameras [5]. The choice of the resolution adopted was such as the size of the scanned photo was similar to the size of the photographed ones. The photos were taken with a Sony Cybershot digital camera DSC-W55 in 5 and 7.2 Mpixels and a Sony Cyber-shot DSC-T10 7.2 MPixels portable cameras. Table 1 presents some of the features of the test images.

JPG	Average Size (MB)	Variance (MB)	Total Number
Photo	3.18	0.10	241
Scanned	3.48	0.08	95

Table 1. Data of dimension of the test set and the size and variance of images (in JPEG)

The last cluster of images in the classifier in [6] is "Don´t Know" images (unassigned). These 529 images were included as to increase the possibility of misclassifications. They are images that appear in the "real world" and range widely in nature from biological images, to vector graphics (obtained by softwares such as Excell®, Powerpoint®, etc.), of which 202 are JPEG and 327 in BMP.

A. *The Classifier*

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 [11]. 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 in reference [6] assumed 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 (if RGB)
- 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 [7] algorithm. The data above are extracted for each image and placed in a vector of features.

The classifier "architecture" is made by cascading binary classifiers. The order they appear has an effect on the final classification accuracy. Figure 5 shows the way they are cascaded.

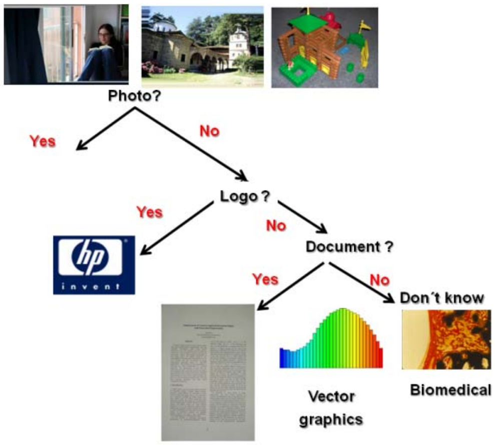

Figure 5. Cascaded binary classifier "architecture".

B. *Recognizing test images as Photos*

We wish to use the classifier presented here to refine the discriminator in [6], thus the first test performed was to submit the test images directly to that classifier and analyze to see if it was able to correctly recognize the images as photos. The test set was fed directly to the general (photo-document-logo) classifier without any training or tuning. The result obtained may be found in the confusion matrix presented in Table 2.

General	Photo	Logo	Doc	DK	Accuracy
Scanned Photos	91	0	4	0	95.7%
Digital Photos	268	3	5	5	95.3%
Total	359	3	9	5	95.3%

Table 2 – Confusion matrix of the general classifier with Scanned/Digital Photo test set

As one may observe from Table 2, the general classifier was able to correctly recognize 95.3 % of the photos in the Scanned/Digital test set as belonging to the cluster "Photo". Five of the Digital photos were misclassified as "Don´t Know" (DK).

C. *Sub-sampling*

Time performance is of paramount importance for embedded software such as an image classifier to run on printing devices. 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 files, 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!).

Twenty different sub sampling strategies were evaluated in [6]. The cascaded sub-sampling strategy consisted of removing more points from the larger image files and provided the best overall accuracy of any classification schema. The pseudo-code for the cascaded sub-sampler is shown below:

```
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 of the "cascaded" sub-sampler

Table 3 presents the results for the General classifier for the sub-sampled images in the Scanned/Digital Photo set.

S_sampled	Photo	Logo	Document	DK	Accuracy
Scanned Photos	90	0	4	1	94.7%
Digital Photos	268	3	4	6	95.3%
Total	359	3	8	7	95.0%

Table 3 – Confusion matrix of the general classifier with sub-sampled Scanned/Digital Photo test set

The data presented in the confusion matrix shown in Table 3 supports the conclusion that the sub-sampling procedure proposed marginal degrades the performance of the classifier. Later on, it will be shown that the performance gain largely compensates the accuracy degradation.

D. *Training and test sets*

To increase the difficulty of discriminating between scanned and digital photos, the images chosen are as similar in theme (people) and size as possible. All images are in color and stored in JPEG file format. Table 1 summarizes some of the features of the images in the test set.

The training set was carefully selected to guarantee the diversity of the images in the test set, having in mind that quality matters more than size. Table 4 presents the relative size of the training and test sets.

	Test	**Training**	**%**
Scanned Photos	**91**	**31**	**34.06**
Digital Photos	**281**	**60**	**21.35**
Total	**372**	**91**	**24.46**

Table 4 – Sizes of Training x Test sets

The Weka [12] classification strategy used was the Random Forests (number of trees equal to 10) [1].

E. *Results*

This section presents the results of clustering of the images in the Scanned/Digital photo test set after the classifier was specially trained for discriminating between scanned and digital photos.

Scanned/Digital	Scanned Photos	Digital Photos	Accuracy
Scanned Photos	95	0	100.00%
Digital Photos	1	280	99.99%
Total	96	280	99.99%

Table 4 – Confusion matrix of the Scanned/ Digital classifier with original images

Table 4 shows that the results obtained for the original images are extremely good with accuracy close to 100%.

Scanned/Digital	Scanned Photos	Digital Photos	Accuracy
Scanned Photos	95	0	100.00%
Digital Photos	1	280	99.99%
Total	96	280	99.99%

Table 5 – Confusion matrix of the Scanned/ Digital classifier with sub-sampled images

Image sub-sampling, as demonstrated by the data that are shown in Table 5, does not introduce any detrimental effects on image classification, and brings performance gains to the feature extraction phase of the classifier. In both cases, exactly one digital photo was misclassified as being a scanned photo either original or in the sub-sampled case. That artistic photo, shown in Figure 6, has a complex illumination pattern that makes difficult its correct classification.

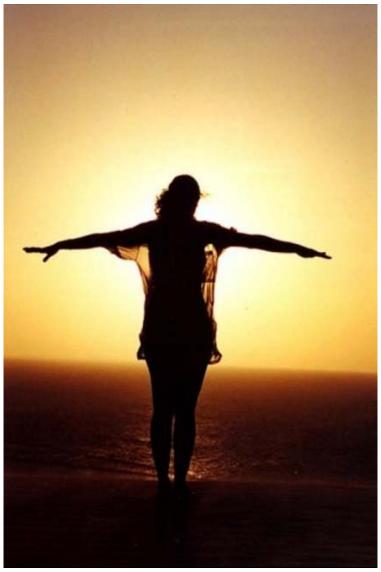

Figure 6. Misclassified digital photo

IV. FURTHER IMPROVEMENTS

Analyzing the results presented in Tables 2 and 3 one may observe that the general (photo-logo-document) classifier is less accurate than the scanned-digital photo classifier for the same test set. In particular, the number of images that were not classified and thus left in the Don´t_know (DK) set is not negligible. It is also important to note that the digital photo in Figure 6 when assigned by the general classifier was inserted in the Don´t_know set. If one observes the classifier architecture from Figure 5 and includes the new Scanned/Digital photo classifier, one may re-structure it to feed-back the Don´t_know cluster into the Scanned/Digital photo classifier, yielding the architecture shown in Figure 6.

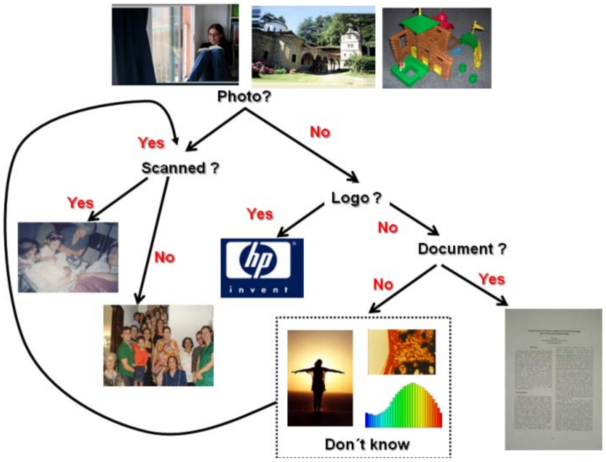

Figure 6. New classifier “architecture”

The new architecture proposed will raise the accuracy of the general classifier from 93.3% to 96.4%, as may be seen in Table 6 for the original images.

General	Photo	Logo	Document	DK	Accuracy
Scanned Photos	91	0	4	0	95.7%
Digital Photos	272	3	5	1	96.7%
Total	363	3	9	1	96.4%

Table 6 – Confusion matrix of the **new architecture** general classifier with Scanned/Digital Photo test set.

Table 7 shows the results of the new architecture for the sub-sampled images increasing the overall accuracy.

S_sampled	Photo	Logo	Document	DK	Accuracy
Scanned Photos	91	0	4	0	95.7%
Digital Photos	273	3	4	1	97.1%
Total	364	3	8	1	96.7%

Table 7 – Confusion matrix of the **new architecture** general classifier with sub-sampled images.

V. Time Performance

Table 8 presents the feature extraction and classification times together with information about the language those procedures were implemented into. Besides classification accuracy per cluster, the average feature extraction and classification times are presented. Note that there is a difference in time scale between feature extraction and classification.

	Feature extraction		**Classification**	
	Time (s)	**Language**	**Time (ms)**	**Language**
Original	**0.4382**	**C++**	**0.10**	**C#**
Sub-sampled	**0.1502**	**C++**	**0.10**	**C#**

Table 8 – Feature extraction and classification times Processor Pentium IV 2.4GHZ 2GB RAM

The performance results presented shows that sub-sampling reduces the feature extraction time to one third of that needed for the original images. In some cases, as in the test data set here, sub-sampling also yielded an increase in accuracy.

VI. Discussion and Conclusions

Scanned and Digital photos have different features. For cultural reasons, people today are more acquainted with the way digital photos look with sharp bright colors than what one tends to get from scanning printer photos. The correct classification allows the printer to automatically meet the users´ expectations.

Weka [8], as in previous research [6], proved an excellent test bed for statistical analysis. The choice of the Random tree classifier [1] was made after performing several experiments with the large number of alternatives offered by Weka, although results did not vary widely. In the current case of the scanned/digital photo classifier, in opposition to the results of [6], the choice of the images in the training set was not of paramount importance to the performance of the classifier.

The new classifier "architecture" proposed here, besides improving the appearance of the printed output in the case of scanned and digital photos, also benefits the overall classification accuracy. It is important to note that the classifier "architecture" with feedback presented in this paper opens a new way of using binary classifiers for multiple classification.

VII. References

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

[2] H. Frigui and R. Krishnapuram. Clustering by competitive agglomeration. *P. Recognition*, 30(7), 2001.

[3] M. A. Hearst and J. O. Pedersen. Reexamining the Cluster Hypotesis: Scatter Gathet on Retrieval Results, SIGIR, 1996.

[4] S. Krishnamachari and M. Abdel-Mottaleb. Image Browsing using Hierarchical Clustering, IEEE Symposium on Computers and Communications, ISCC'99, July 99.

[5] R. D. Lins and D. S. A. Machado, A Comparative Study of File Formats for Image Storage and Trans., v13(1):175-183, Journal of Electronic Imaging, 2004.

[6] R. D. Lins, G. F. P. e Silva, B, S.J. Simske, J. Fan, M. Shaw, P. Sá, M. Thielo. Image Classification to Improve Printing Quality of Mixed-Type Documents, ICDAR 2009. IEEE Press, 2009. p.1106 - 1110.

[7] N. Otsu. "A threshold selection method from gray level histograms". IEEE Trans. Syst. Man Cybern. v(9):62-66, 1979.

[8] P. Scheunders. Comparison of Clustering Algorithms Applied to Color Image Quantization, Patt. Recog. Letters, v18(11-13):1379-1384, 1997.

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

[10] G. F. P. e Silva, R. D. Lins, B. Miro, S.J. Simske, M. Thielo, Automatically Deciding if a Document was Scanned or Photographed. Journal of Universal Computer Science., v.15, p.3364 - 3366, 2009.

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

[12] Weka 3: Data Mining Software in Java, website http://www.cs.waikato.ac.nz/ml/weka/.