A NEW SEGMENTATION ALGORITHM FOR TRUE COLOUR IMAGES OF HISTORICAL DOCUMENTS

C. MELLO, R. LINS

Departamento de Eletrônica e Sistemas
UFPE - Brazil
{cabm,rdl}@cin.ufpe.br

ABSTRACT

It is presented herein a variation of a new entropy-based segmentation algorithm for filtering true colour images of historical documents. The algorithm is used to eliminate the interference of the paper in the ink yielding a better quality image. It works in typed and hand written documents just as in post cards.

1. INTRODUCTION

Image segmentation is one of the most important steps in many applications involving image processing as image analysis. In this work, we are interested in the application of the segmentation process to generate high quality monochromatic images of true colour documents. The coloured images are from the file of letters, documents and post cards of Joaquim Nabuco[1] held by the Nabuco's Foundation (a social science research institute in Recife-Brazil). The binarized images are used to minimize the storage space making easy to divulgate them thru some media as the Internet. The Nabuco's file is composed of documents from the end of the nineteen century. Because of its age, the paper is more suceptible to the wear and tear over time. This degradation yellows the sheet of paper and creates some noise that is perceptable to the digitization process. Even more, in some cases, the ink has faded. This is particularly important when the document is written in both sides of the paper. In some of these cases, the ink of one side interferes in the other.

In [9] we presented a new segmentation algorithm for greyscale images of the documents of Nabuco's bequest. Herein, we make a variation of this algorithm allowing the application of it to true colour images. The main objective here is the generation of better quality monochromatic images for storage purposes and use of Optical Character Recognition tools and the creation of paper texture for future projects. The segmentation process is used to identify the *object* and its background. The object can be the ink (or the paper) and the background can be the paper (or the ink). After this identification, each of these classes is used to produce two final images: one with the ink part and the other with the paper frequencies.

2. NABUCO PROJECT

The main objectives of the Nabuco Project [15] is the preservation and easy divulgation of a file of thousand of historical documents. For this purpose an environment is under development to acquire, process and storage the information of Joaquim Nabuco's bequest. In means to be used by any operator, this system must not require the specialized users.

The documents are digitized with 200 dpi (*dots per inch*), in *true colour* (16 million of colours - 24 bits) and they are being stored in JPEG file format [12] with 1% loss (an acceptable value considering the compression rate/quality relation). The monochromatic images are generated to for three main reasons: 1) to reduce their size allowing less storage space; 2) to improve the speed of network transmission over the Internet and 3) to be used by OCR tools.

Image processing commercial tools (such as Photoshop™ [13]) have a great variety of filters. However, the use of such softwares requires a specialized operator which is not interesting for the project. Even more, such tools did not provide images with good quality when applied to historical documents. All of this justifies the need for the development of new algorithms.

In the bequest, the are documents written in one side of the paper and in both sides. In this second case, two classes are identified:

- documents with no interference of one side on another
- documents with interference

The first class is the most common and simple to deal. The monochromatic image can be generated from the coloured one through the application of a threshold filter. A neighbourhood filter [2] can also be used to reduce the noise in the image.

The second class of documents is more difficult to work. A simple colour reduction by a straightforward threshold algorithm does not eliminate all the influence of the ink transposition from one side to the other.

A segmentation process can be used in a greyscale version of the original image. This, however, means the loss of 1/3 of the information of the document as we go from a 16 million colour table to a 256 grey levels. Although, with more information the system must have more variables to work with. As the documents are digitized in true colour, our segmentation algorithm works directly with these images generating monochromatic versions of the inputs.

A sample of a typed document with no interference can be seen in figure 1 and an example of of the application of a nearest colour algorithm to this image is presented in figure 2.

The low quality of the monochromatic image generated justifies the need of a new segmentation algorithm [3]. In special, we are also interested in improving the hit rates[2] of OCR tools.

[1] Brazilian statesman, writer, and diplomat, one of the key figures in the campaign for freeing black slaves in Brazil, Brazilian ambassador to London (b.1861-d.1910)

[2] Number of characters correctly transcribed from image to text

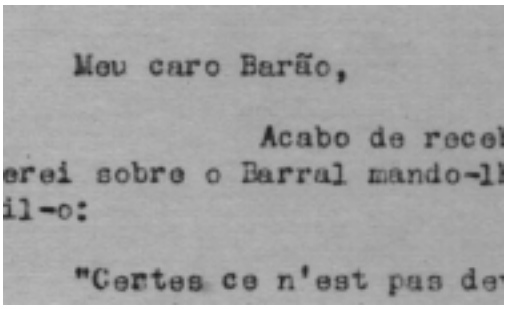

Fig 1. Zooming into true colour sample document from Nabuco's bequest

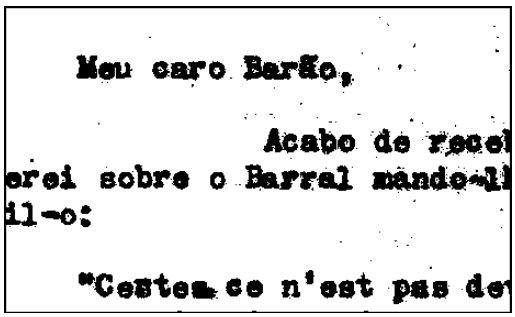

Fig 2. Monochromatic version of the sample document generated by nearest colour algorithm using Photoshop™

3. A NEW SEGMENTATION ALGORITHM

This segmentation algorithm was first proposed in [9] for greyscale images and now it is expanded to true colour ones. For greyscale images, in its first step, the algorithm scans the image in search for the most frequent colour of the histogram of the image. As we are working with images of typed documents and letters it is coherent to suppose that this frequency belongs to the background of the image (the paper). This frequency is used as a threshold value, *t*, to evaluate the entropy [1] of the black (*Hb*) and white (*Hw*) pixels. For black and white pixels we mean pixels with colours below and above *t*, respectively. The entropies are evaluated using the following equations:

$$Hb = -\sum_{i=0}^{t} p[i]\log(p[i]) \qquad \text{(Eq 1)}$$

$$Hw = -\sum_{i=t+1}^{255} p[i]\log(p[i]) \qquad \text{(Eq. 2)}$$

Where *p[i]* is the probability of the frequency *i* in the histogram. The logarithm is taken with basis X.Y, where X and Y are the dimensions of the complete image. As defined in [5], this change in the logarithmic basis does not changes the concept of entropy.

The entropy of the complete histogram, *H*, is also evaluated using the same logarithmic basis as before. This value is used to define two multiplicative factors, *mw* e *mb*, which values are determinated by the rules:

- If $0.25 < H < 0.30$, then $mw = 1$ and $mb = 2.6$
- If $H \leq 0.25$, then $mw = 2$ and $mb = 3$
- If $0.30 \leq H < 0.305$, then $mw = 1$ and $mb = 2$
- If $H \geq 0.305$, then $mw = mb = 0.8$.

The values of *mw* and *mb* were achieved experimentally after several tests. The segmentation is now performed in a new scan of the image. The pixel *i* is turned white if:

$$(\text{colour}[i]/256) \geq (mw*Hw + mb*Hb)$$

Otherwise, the colour of the pixel remains the same (generating a new greyscale image) or it is converted to black (to generate a monochromatic image). This is called the *segmentation condition*.

Another application of this algorithm works with an inversion of the segmentation condition. As before, the validation of the condition implies that the colour of the pixel is turned white. Otherwise, it remains the same. This creates an image with only the frequencies classified as paper.

3.1 Greyscale Images

As described before, the algorithm can be immediately applied to greyscale images. We can separate these images into two classes: document images (with and without interference) and post cards images.

The algorithm was used to generate images for two types of applications: 1) storage of good quality monochromatic images and 2) for optical character recognition. In the second application, we tried to get better responses from OCR's commercial tools applied to images segmented by the new algorithm. An example of the use of the algorithm in an image of Nabuco's bequest can be seen in figure 3 next.

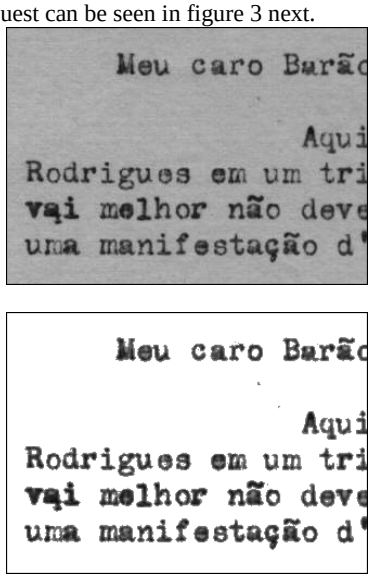

Fig 3. (top) Original greyscale image and (bottom) segmented image by the new algorithm

The new algorithm was also applied to images of post cards. In this case, our main objective is to achieve good quality monochomatic images. Figure 4 shows an example of this application. This example is used only to show the versatility of the algorithm. Other techniques (as variation of *dithering*) are being studied for this kind of image.

The new segmentation method was tested in a group of fifty greyscale images from Nabuco's bequest yielding very satisfactory results, under visual inspection, specially when used in images with interference of one side on the other. Others

techniques were tested [9] using Pun's [11], Kapur *et al.*'s [4] and Johannsen's [10] algorithms but the new algorithm achieved the best quality images. Futhermore, it has a differential particularity of being automatically applicable to the files without the need of a specialized operator which is one of the main objectives of the Nabuco's Project [3][15].

For optical character recognition, typed documents from Nabuco's file were segmented by the algorithm, improving the hit rate of OCR's tools as Omnipage 9.0 from Caere Corp [14] (which presented the best hit rates in previous tests amongst a set of six OCR commercial tools [8]). Its hit rate is of around 99% when applied to recent documents. This, however, is not what happens when we work with historical documents. For this kind of image the error rate achieves unnacpetable values.

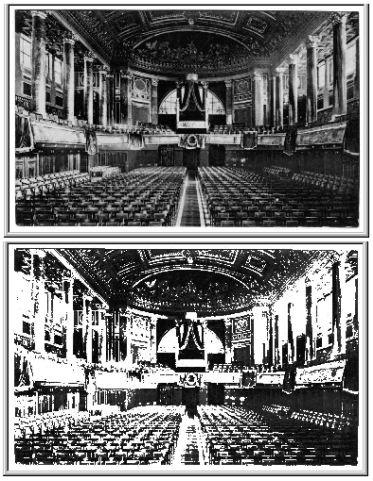

Fig 4. Application of the new algorithm in post cards

As said before, another application of the algorithm comes with the use of the segmented image without the frequencies classified as ink, inverting the segmentation condition. These images (as can be seen in figure 5 next) are being used to generate a texture database of papers from the beginning of the century. More details about this study can be seen at [7].

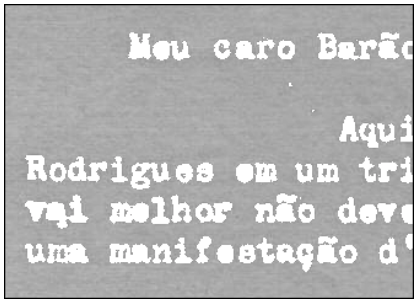

Fig 5. Image of figure 3-top segmented to take off the pixels classified as ink

3.2 True Colour Images

A variation of segmentation algorithm was developed to work with true colour. In this format each colour is formed by a combination of *red*, *green* and *blue* frequencies (the *RGB* standard). Each value of the hues are stored in a set of 8 bits totalizing 24 bits for any colour.

In this variation of the algorithm, all the previous calculations are evaluated to each frequency of R (*red*), G (*green*) and B (*blue*) that forms the colour. So when we had before the variables *H*, *Hw* and *Hb*, now we will have *Hr*, *Hwr* and *Hbr* (for the red component), *Hg*, *Hwg* and *Hbg* (for the green hue) and *Hb*, *Hwb* and *Hbb* (for the blue component). The same happens with the multiplicative factors mw and mb (now *mwr*, *mbr*, *mwg*, *mbg*, *mwb* and *mbb*). The same rules used before for the definition of these factors are applied here for each one of the frequencies RGB. In means to calculate *Hwf* and *Hbf* (with *f* = r, g or b) we use equations 1 and 2 with different cut off frequency depending of each hue.

A most significative variation in the original algorithm is done in the classification of the pixel as ink or paper. In the greyscale version this classification is based on the answer of the segmentation condition. With coloured images, however, we have three different hues to each colour. The same condition will be used for each hue but with another definition for its results. The pixel is classified as ink (so turned white) if the condition results *true* for, at least, one of the components R, G or B. Otherwise, if the condition is *false* for *all* three components, the pixel colour remains the same (thus classified as paper). Again, the inversion of the segmentation condition generates an image with only the colours classified as paper. One can see this technique applied to the sample document in figure 6.

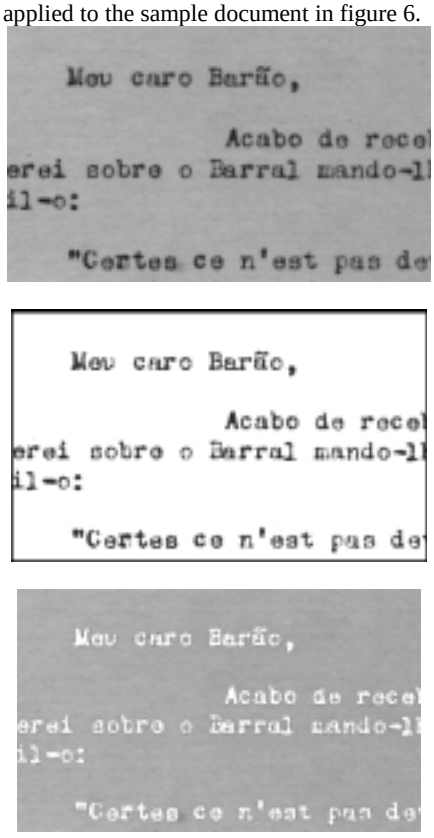

Fig 6. (top) Original true colour image, (center) segmented image and (bottom) segmented image with the background frequencies

In the previous example shown in figure 6, the segmentation algorithm were applied to the entire image. This can be changed decreasing the area of actuation of the scheme to a set of *N* lines of the image. This means that the original image with dimensions *x.y* is now divided into a group of subimages each one with dimension *N.x*. Each subimage is called a *region*. When the algorithm is applied to the complete image, we say that *N=y*. The algorithm is then applied to each region. Figure 7 shows an example of the application of the algorithm in the complete image and with *N = 10* (this means that the calculation is done in each group of 10 lines). As can be seen in figure 7-bottom, the segmented image still presents vestiges of the ink of the other side of the paper. In means to achieve better quality images, the algorithm can be applied in cascade. The algorithm was applied line-by-line (*N = 1*) three times to the image presented in figure 7-top left and its results can be seen in figure 8.

Fig 7. (top left) Original image, (top right) segmented image considering the complete image and (bottom) segmented image working with *N = 10*

Fig 8. Application of the new algorithm in cascade to the complete image (*N = y*) of figure 7-top (top-left) by the first time, (top-right) after two times and (bottom) after the third time

4. CONCLUSIONS

This variation on the new segmentation algorithm presented in [9] not only achieves better results than other segmentation algorithms for true colour images but also it validates even more the original scheme. It can reach satisfactory results if applied to the complete image or to small regions of it reducing (even eliminating) the influence of the background in the foregorund in the cases where the document is written in both sides.

For OCR tools, the new algorithm generated better quality images and consequently better hit rates.

Coloured texture of the paper of the documents were also created using the images without the colours classified as ink.

5. REFERENCES

[1] N.Abramson. *Information Theory and Coding*. McGraw-Hill, 1963.

[2] R.Gonzalez and P.Wintz. *Digital Image Processing*. Addison Wesley, 1987.

[3] L.R.França Neto, C.A.B.Mello and R.D. Lins. *Filtering Techniques for Digital Images of Historical Documents*. XV Brazilian Symposium of Telecommunications, Recife, Brazil,September, 1997.

[4] J.N.Kapur *et al*. *A New Method for Gray-Level Picture Thresholding using the Entropy of the Histogram*. C. Vision, Graphics and Image processing, 29(3), March, 1985.

[5] S.Kullback. *Information Theory and Statistics*.Dover Publications, Inc. 1997.

[6] R.D.Lins *et al*. *An Environment for Processing Images of Historical Documents*. Microprocessing & Microprogramming, pp.111-121, N.Holland, January, 1995.

[7] C.A.B.Mello & R.D.Lins. *Generating Paper Texture with Statistical Moments.* IEEE Interntional Conference on Acoustic, Speech and Signal Processing. Istanbul. Turkey. June, 2000.

[8] C.A.B.Mello & R.D.Lins. *A Comparative Study on OCR Tools*. Vision Interface 99, pp. 224-232, Quebéc, Canada, May, 1999.

[9] C.A.B.Mello & R.D.Lins. *Image Segmentation of Historical Documents* (in portuguese). Proceedings of the XVII SBT, pp.700-704, Brazil, September, 1999.

[10] J.R.Parker. *Algorithm for Image Processing and Computer Vision*. John Wiley, 1997.

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

[12] K.Sayood. *Introduction to Data Compression*. Morgan Kauffman Publishers, Inc., 1996.

[13] *Adobe Systems Inc*. URL: http://www.adobe.com

[14] *Caere Inc.* URL: http://www.caere.com

[15] *Nabuco Project.* URL: www.di.ufpe.br/~nabuco