Generation of Images of Historical Documents by Composition

Carlos A.B. Mello
Escola Politécnica
Universidade de Pernambuco
Recife-PE
Brazil
cabm@netpe.com.br

Rafael D. Lins
Departamento de Eletrônica e Sistemas
Universidade Federal de Pernambuco
Recife - PE
Brazil
rdl@ee.ufpe.br

ABSTRACT

This paper describes a system for efficient storage, indexing and network transmission of images of historical documents. The documents are first decomposed into their features such as paper texture, colours, typewritten parts, pictures, etc. Document retrieval forces the re-assembling of the document, synthetising an image visually close to the original document. The information needed to build the final image occupies, in average, 2 Kbytes performing a very efficient compression scheme.

General Terms

Algorithms, Documentation.

Keywords

Historical Documents, Segmentation, Texture, Synthesis.

1. INTRODUCTION

The work reported herein is part of the Nabuco Project[5][10] for preservation and broadcasting of letters and documents from Joaquim Nabuco[1]'s bequest. The file is composed of almost 6,500 documents from the end of the nineteenth century, totalizing more than 30,000 pages.

The Nabuco Project is developed by the Federal University of Pernambuco jointly with the Joaquim Nabuco Foundation (a social science research centre), both in Recife (Brazil). Documents are digitized in true colour with 200 dpi resolution and stored in JPEG [9] file format with 1% loss for preservation purposes. Even in this format each image of a document reaches, in average, 380 Kb.

A processing environment was envisaged to extract the basic features of documents, which allows for later image re-assembling. The extraction of these features is performed by the blocks presented in Figure 1 and it is explained in details in section 5.

In order to obtain satisfactory results, several new algorithms were developed within the scope of the project and are described herein.

Ink and paper segmentation is not always a simple task in this kind of image. In some documents, the ink has faded; some of the others were written on both sides of the paper and the ink transposed the document presenting back-to-front interference. A conversion into a monochromatic version of this kind of documents using a nearest colour threshold algorithm [3] does not achieve high quality results. Figure 2 presents a greyscale image from a document that is written on both sides. A conversion to monochromatic by a commercial software do not generate a good quality image. For this, a new entropy-based segmentation algorithm presented in [6][7] yielded better results.

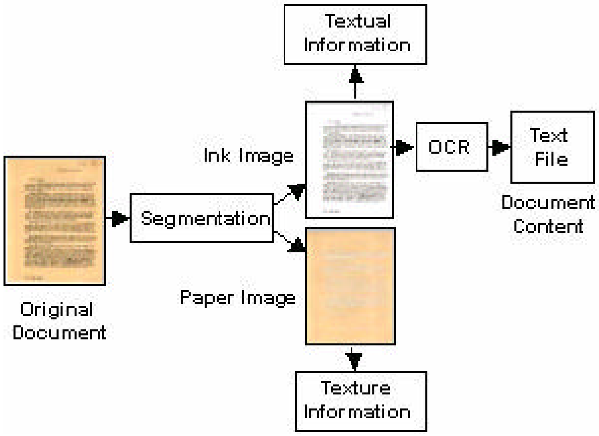
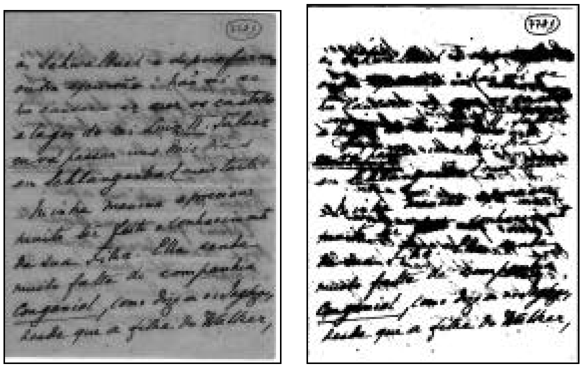


Figure 1. Block diagram for the data extraction of the paper and text image.

The documents digitized present two kinds of noise: one from the paper itself (such as humidity, marks, fungi, dirt, finger prints, etc.) and the other from the digitization process itself. Figure 3 presents an example of a digitization noise as image presents a "shadow" around the letters. This shadow is inserted during the digitization process and it may group together two or more consecutive characters. This brings another problem for this type of documents as it makes harder for recognition algorithms of OCR's (Optical Character Recognition) 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)

DocEng´02, November, 8, 2002, McLean, Virginia, USA.



a) b)

Figure 2. Sample document of the Nabuco's bequest where the paper is written on both sides. a) Original document in greyscale and b) its conversion to monochromatic using a nearest colour algorithm.

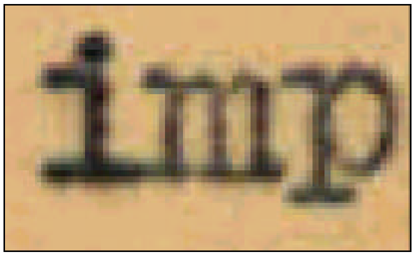


Figure 3. "Shadow" around characters due to the digitization process.

After segmentation, the paper and the ink images are processed separately for extracting their main features.

For type written documents, the ink part is processed by an OCR tool to generate an annotated text with entries to a fontset database.

Document re-assembling follows the scheme described in Figure 4 below. The texture information is used to build a blank sheet of paper visually similar to the original one.

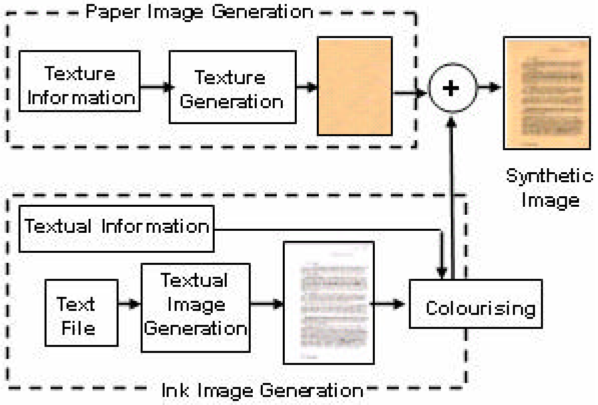


Figure 4. Scheme for generating the synthetic image of a historical document.

The "text file" (output of some OCR tool) generates a text image from an image database of characters. Some other data ("textual information") such as colour, hues, etc, extracted from the original file are used to colour the "ink image". The coloured ink image is added to the "blank sheet of paper" yielding the synthetic image.

The information needed to build the final image occupies less than 2 Kbytes.

The specification and analysis of a system that generates a synthetic image of typed documents is the main goal of this paper.

Next, each main step of the generation process is analysed beginning with a review of the entropy-based segmentation algorithm followed by the algorithm for texture and text synthesis and the complete system.

Other projects around the world are benig developed for historical documents such as the DEBORA Project (*Digital accEss to BOoks of the RenAissance*) [11].

2. THE ENTROPY-BASED SEGMENTATION ALGORITHM

The efficient segmentation between ink and paper is fundamental to the document processing environment presented herein. For this purpose, entropy-based segmentation algorithms presented the best results when applied to a set of more than 200 documents. A new algorithm was developed in [6][7] and it is reviewed herein.

For greyscale images, it is found the most frequent colour, *t*. As the environment works with images of letters and documents, it is reasonable to suppose that this colour belongs to the paper. The entropy [1] of the pixels below and above this value (*Hb* and *Hw*, respectively) is evaluated:

$$Hb = -\sum_{i=0}^{t} p[i]\log(p[i]) \qquad Hw = -\sum_{i=t+1}^{255} p[i]\log(p[i])$$

where the logarithmic basis is taken as the product of the dimensions of the image and *p[i]* is the probability of the colour *i* is present in the image. The entropy of the complete histogram, *H*, is evaluated as the sum of *Hw* and *Hb* and it defines two multiplicative factors, *mw* and *mb*, experimentally determined by:

- If H≤0.25, then mw=2 and mb=3;
- If 0.25<H<0.30, then mw=1 and mb=2.6;
- If H≥0.30, then mw=mb=0.8.

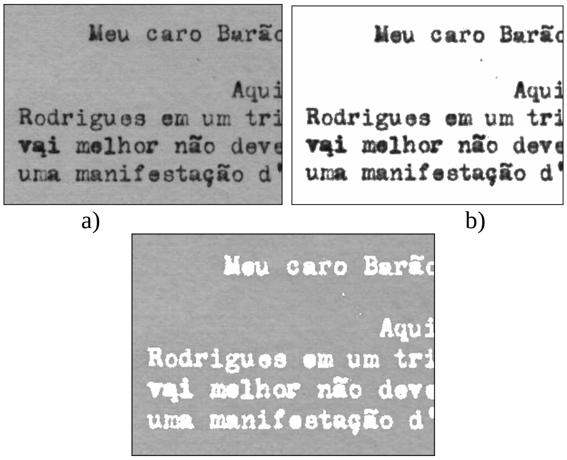
These values of mw and mb were applied to a set of 500 images achieving very satisfactory results.

The image is then segmented with pixel *i* with colour *colour[i]* converted to *white* if:

(colour[i]/256) ≥ (mw.Hw + mb.Hb)

else it 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*. An example of the application of the algorithm may be found in Figure 5. An inversion of this condition builds an image where the pixels classified as ink are turned into white remaining only the paper texture (Figure 5.c).

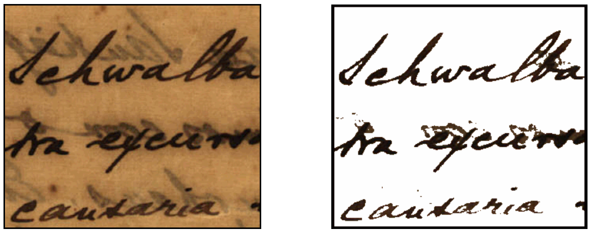
For true colour images, the algorithm works as above but now it is applied for each of the RGB (*Red, Green* and *Blue*) components. A pixel is classified as paper (thus turned into white) if the segmentation condition results true for, at least, one of the components R, G or B. Otherwise, its colour remains unchanged. Figure 6 shows an example of a true colour image with back-to-front interference and the result after the application of the algorithm.


Meu caro Barão
Aqui
Rodrigues em um tri
vai melhor não deve
uma manifestação d'

a) b)

c)

Figure 5. Example of the application of the new algorithm in a typed document.
a) Original image, b) segmented image of the ink and c) inversion of the segmentation condition generating an image of the paper.



a) b)

Figure 6. New entropy-based segmentation algorithm applied to a true colour image with back-to-front interference.
a) Original image and b) segmented version.

The use of the algorithm increased the answer of OCR tools more than 10%.

Both images produced by the algorithm are used in the system for automatic generation of documents, as detailed later. From now on, the segmented image of pixels that were classified as paper is called *paper image* and its negative (the pixels classified as ink) is called *ink image*.

Detailed information on the segmentation algorithm may be found in [6][7].

From the paper image information is acquired in means to re-build the complete sheet of paper without the text part as it is explained in the next section.

3. TEXTURE SYNTHESIS

Samples of the texture of 200 colour images were collected and analysed. From them, the mean, standard deviation and entropy of the histograms of RGB components were extracted. Figure 7 presents an example of a paper texture and its RGB histogram. It could be found that the histograms analysed are gaussian-like functions. This can be confirmed by the evaluation of the entropy as it is a measure of deformation from perfect gaussian curves [4].

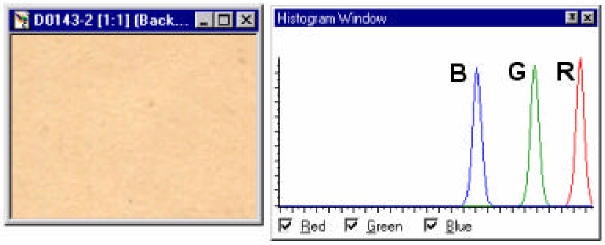


Figure 7. Sample texture and its histogram.

The mean and standard deviation values are used to specify the colour space of the image which has values between *mean – standard_deviation* and *mean + standard_deviation* for each of the three RGB components. However, not all colours generated in this space is part of the original texture. The *hue* is the feature used to define which colours must be in the texture. It can be evaluated by applying equation [2]:

$$hue = \cos^{-1}(((r-g)+(r-b))/2)/\sqrt{(r-g)^2+(r-b)*(g-b)}$$

where r, g and b are the values of the red, green and blue components for a given colour. For all 200 images analysed, it was observed that there is a predominant hue value (called *hue_max*). In some cases, this value is found in more than 40% of the colours of an image. A new image is then created where each pixel has its colour determined by a hue value, no longer by its RGB components alone. The *hue_max* of this new image is evaluated and this value is the factor that defines if a colour formed by some triple RGB can be accepted or not in the synthetic image. The most frequent hue value works as a boundary for the possible colours that can be generated in the paper image.

The entropy of the hues (*h_hue*) in the image is also evaluated. As before, the entropy is evaluated using as logarithmic basis the product of the dimensions of the image. The need for h_hue is explained below.

The values of the mean and standard deviation of the histograms of the RGB components, *hue_max* and *h_hue* are all the information needed to create the synthetic image of the texture of the paper. The dimensions of the image are also stored totalizing 40 bytes only.

The histogram of the synthetic image is now created using the mean and standard deviation of the histograms of the original image. The gaussian-like functions for the *RGB space* is expanded by the RGB values bounded by *mean* ± *factor**|*variance*| for each colour tone. The *factor* variable is defined by an interactive process where a gaussian function is created for the histogram using the mean and standard deviation of each of the RGB components. *factor* reaches its final value whenever the addition of the amplitudes of the histogram is either greater or equal to the number of pixels of the image or it reaches 20% of this number (in this case a correction is done to complete the number of pixels). The use of the variance instead of the standard deviation “stretches” the amplitude of the mean value of the gaussian distribution.

The entropy of the hues is used to determine how the distribution approaches to a gaussian one. A triple (r,g,b) from the *RGB space* can be in the final image if its hue value is between *hue_max – delta* and *hue_max + delta*, where if h_hue > 0.17, then delta = 10, else delta = 1. These values were found empirically and are called the *hue space*.

Parameter h_hue is also used to define how many colours of the *RGB space* will have the hue defined in the *hue space*. The entropy h_hue and the number of colours in *hue space* are related by a first order function:

number_colours_in_hue_space = 152.64*h_hue + 16.0089

Functions of higher orders were tested (second, third, fourth and fifth order) with no satisfactory results.

Whenever the number of colours in the *hue space* reaches its maximum, the rest of the image is filled with colours from the *RGB space* even in the case they do not belong to the *hue space*.

The colour of each pixel is determined by pseudo-random searches in the colour table until the maximum number of colours in the *hue space* is reached. After this the colours are selected from the *RGB space* with no restriction about its hue value. The synthesis of the sample texture of Figure 7 is presented in Figure 8.

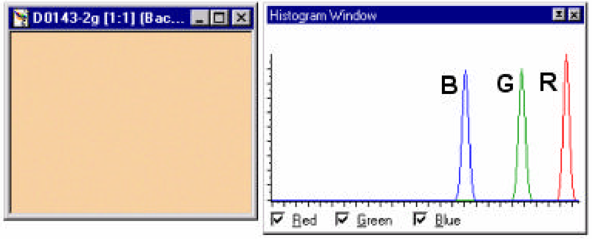


Figure 8. Automatic texture generation of sample texture of Figure 7.

The synthetic textures were analysed qualitatively by visual inspection and quantitatively by measuring the *Peak Signal-to-Noise Ratio* (PSNR) and *Analysis of Variance* (ANOVA). In both cases, a set of 200 textures were generated and compared achieving satisfactory results.

The ANOVA was applied for mean, standard deviation, skewness and kurtosis of the images histogram and for the same features of their Grey-Level Co-Occurrence Matrix (GLCM).

4. TEXT SYNTHESIS

The typed letters and documents of the bequest have very singular features. Most of the documents were written in the same typewriter, which did not allow for changes in the fontset. At the end of the 19th century, it was usual to writers to take their own typewriters along in travels, similarly to carrying a laptop today. Figure 9 below shows four samples of the *u* letter extracted from four different documents between 1882 and 1888. One can observe that there is a strong fading in the upper right corner of the character presented in the four samples, suggesting that they were typed by the same machine.

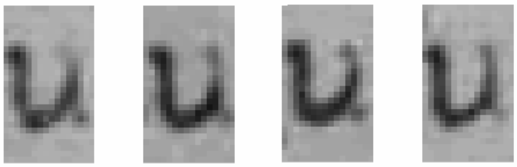


Figure 9. Samples of the *u* letter extracted from documents between 1882 and 1888.

To re-create the image of the text, a database of images of characters is formed as a fontset. Several samples of each character is collected from different greyscale ink images (images already segmented with the paper component removed). The number of samples varies from character to character according to their frequency in the language and the number of unmatching pixels. For example, the character "%" was found only once in the 200 documents analysed.

All the samples collected are used to create a unique fontset of images of each character. Each set of samples of a character has the same area and each pixel is compared looking for a majority vote or the mean of the pixels to define the colour of the final sample of the character.

Each character image follows the rules:

- The upper and lower case letters, numbers and symbols are boxed in frames of 22x31 pixels. Some special symbols may use a larger box (as "%", "("and ")");
- Characters are stored in greyscale;
- Lower case letters are divided into three groups: wide (m and w), vertical (g, p, q, y and ç) and ordinary (the rest of the alphabet). According to this classification, the ordinary letters are layed in their frames with a 3-pixel left spacement and a 6-pixel bottom spacement; the wide letters have no extra space sideways; and the vertical letters have only a 1-pixel margin to the bottom;
- It was chosen to store accentuated letters (such as à, á, â, etc.) instead of only its accents for performing optimization;
- Any symbol found in the text file that is not present in the character image database is exchanged for an asterisk.

Examples of some characters may be found in Figure 10 next.

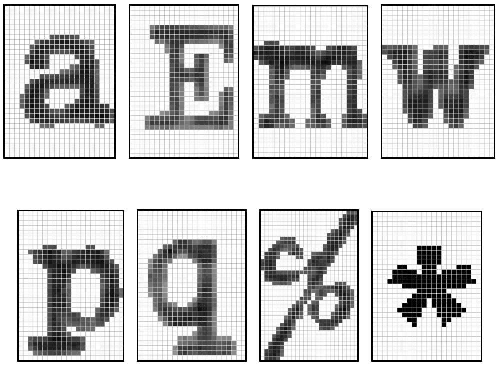


Figure 10. Sample character images extracted from typed documents of Nabuco's bequest.

Some parameters are also defined for the documents and taken as default:

- Left margin = 150 pixels
- Upper margin = 75 pixels;

- Blank space = 24 pixels;
- Space between lines = 70 pixels;
- Tabulation = 4 blank spaces.

The space between lines can have its value changed depending on the number of characters in the document. Some documents have almost the complete paper page typed and thus the space between lines is narrower (40 pixels now).

In order to have the synthesis of the document image, the ink image created by the segmentation algorithm is inserted in an OCR tool and the resulting text file in annotated ASCII code is stored. At the present only one fontset was defined, with one image per ASCII symbol. We forsee the generation of a much larger fontset in which several images may be associated with each ASCII symbol. Looking up to the annotated ASCII code for a document in the fontset database one may generate an image with the textual part of the document in greyscale. A system to colourise this image is explained below. The system presented herein uses Omnipage which, in previous tests [8], presented the best results for our application.

The system does not deal with recognition errors yet. The ink image is generated directly from the OCR output.

4.1 Colourising the Text

Similarly to the texture generation, a 40 byte long binary data file is created with the same information from the original true colour ink image as before: mean and standard deviation for each RGB component, the most frequent hue value and the entropy of hues.

The range of RGB colours is then defined by its minimum and maximum values as follows:

max = mean + factor1*std + factor2

and

min = mean - factor1*std - factor2

where mean and std are the mean and standard deviation of the tones (R, G or B) in the original image and factor1 and factor2 are two constants defined by the following rule:

if h_hue > 1 then factor1 = 1, factor2 = 0 and delta = 0,
else factor1 = 2, factor2 = 1 and delta = 2.

Constants factor1 and factor2 may expand the colour space, while delta allows for a variation on value of max_hue.

With the minimum and maximum values of R. G and B defined, one proceeds with the determination of the colours inside these ranges that generate hue_max; in fact, hue_max ± delta. These colours are stored on a table, called colour table. The delta value increases the range of acceptable hues by the system. It is defined as above, based on the entropy of the hues of the pixels (h_hue).

Selected the colours to be used in the final image, a new palette is created, called the hue-scale palette. For each entry of the colour table, one evaluates its corresponding grey value. As the difference between adjacent colours in the table is very small, there are few grey values. The grey palette is then replaced by another in which each grey value is exchanged by an entry of the colour table that may generate this grey value.

This technique modifies only a small part of the greyscale palette. To create a palette of similar colours, the other grey values are also changed based on the maximum value of the colour table that was inserted onto the grey palette. This is only to keep a visual uniformity of the palette; the colours not altered by the algorithm do not appear in the final colourised image.

As only the values of the grey level palette are altered the structure of the image itself is not changed, increasing the performance of the technique.

As the values of the standard deviation are very small for the ink images, they are added to 5, defining an offset.

So a synthetic ink image is generated from the OCR output and information about the original colours of the ink image.

5. THE COMPLETE SCHEME

As presented in Figure 1 and 4, the main purpose of this project is the synthesis of the images of historical documents. For this, the digitized image is first segmented and separated into the ink image and the paper image.

The ink image is then inserted in the OCR tool to generate the output text file. The mean, standard deviation, most frequent hue and entropy of the hue values is the evaluated from both ink and paper image and stored in two data files: one for the texture information and another to the textual information. This ends The codification part of the algorithm as presented before in figure 1.

In means to synthetise the complete document, each data file is processed separately to generate each part of the final image. The texture information is used to synthetise the image of the paper sheet using the algorithm presented in section 3.

With the OCR output, the textual image is build. Each ASCII character is converted to image according to the document parameters defined in the previous section for margins, spacement between words and lines, etc. The textual greyscale image is then colourised using the parameters of the ink data file.

The synthetic ink image is now superposed on the paper image, generating the final document image. Figures 11 and 12 present examples of the application of the complete algorithm.

The algorithm was applied to a set of 30 document images and the synthetic images were analysed using the ANOVA to evaluate the mean, standard deviation, skewness and kurtosis of their histograms. Only the skewness and kurtosis of the green and blue histograms presented differences statistically significant.

The PSNR was not analysed because the new image has different positions of the characters and the peak-to-noise ratio makes an comparison pixel-to-pixel, not producing a significant response in this case.

6. CONCLUSIONS

This paper presents a system for complete generation of a synthetic version of images of historical documents. An algorithm for automatic creation of texture is introduced and the settings for the building of the ink image from a text file produced by an OCR tool. The final synthetic image may be considered good both by qualitative and quantitative measures. The system was applied to a set of 30 document files and the synthetic images were analysed thru visual inspection and ANOVA.

The ANOVA method was applied to analyse the histogram of all three RGB components observing the mean, standard deviation, skewness and kurtosis.

In terms of compression, the results are very efficient as the original images stored in JPEG file format with 1% loss occupies

380 Kbytes in average and the text and binary data files occupies only less than 2 Kbytes.

7. REFERENCES

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

[2] D.H.Ballard e C.M. Brown. *Computer Vision*. Prentice-Hall, 1982.

[3] R.Gonzalez e P. Wintz. *Digital Image Processing*. Addison Wesley, 1995.

[4] J. N. Kapur, Measures of Information and their Applications, John Wiley and Sons, 1994.

[5] R.D.Lins, M.S. Guimarães Neto, L.R. França Neto and L.G. Rosa. *An Environment for Processing Images of Historical Documents*. Microprocessing & Microprogramming, pp. 111-121, North-Holland, January, 1995.

[6] C.A.B.Mello and R.D. Lins. *A New Segmentation Algorithm for True Colour Images of Historical Documents* (in portuguese), XVIII Simpósio Brasileiro de Telecomunicações, September, 2000, Brazil.

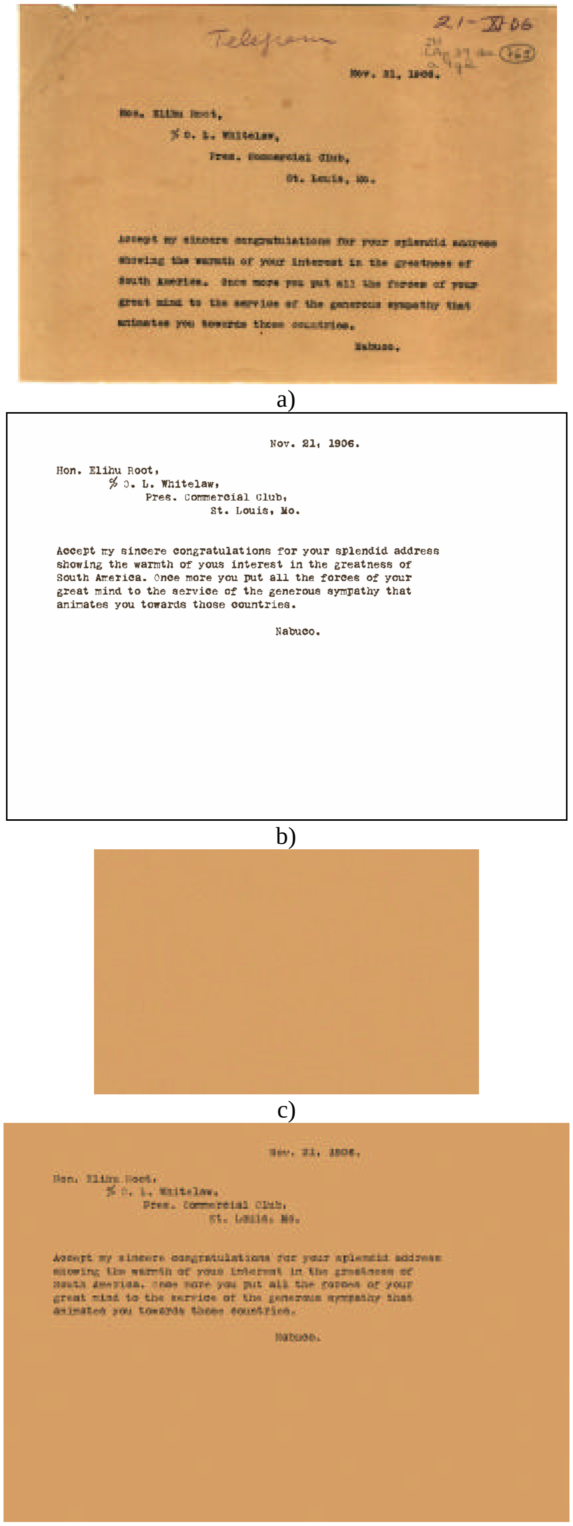
[7] C.A.B.Mello and R.D.Lins. *Image Segmentation of Historical Documents*, Visual 2000, August, 2000, Mexico.

[8] C.A.B.Mello and R.D.Lins. *A Comparative Study on Commercial OCR Tools*.Proc. of Vision Interface'99, Canada, May, 1999.

[9] K.Sayood. *Introduction to Data Compression*. Morgan Kauffman, 1996.

[10] Nabuco Project: http://www.cin.ufpe.br/~nabuco

[11] Debora Project. http://debora.enssib.fr



Telegram

21-XI-06

IN Cap 39 do a 49

765

Nov. 21, 1906.

Hon. Elihu Root,
℅ O. L. Whitelaw,
Pres. Commercial Club,
St. Louis, Mo.

Accept my sincere congratulations for your splendid address showing the warmth of your interest in the greatness of South America. Once more you put all the forces of your great mind to the service of the generous sympathy that animates you towards those countries.

Nabuco.

a)

Nov. 21, 1906.

Hon. Elihu Root,
℅ O. L. Whitelaw,
Pres. Commercial Club,
St. Louis, Mo.

Accept my sincere congratulations for your splendid address showing the warmth of yous interest in the greatness of South America. Once more you put all the forces of your great mind to the service of the generous sympathy that animates you towards those countries.

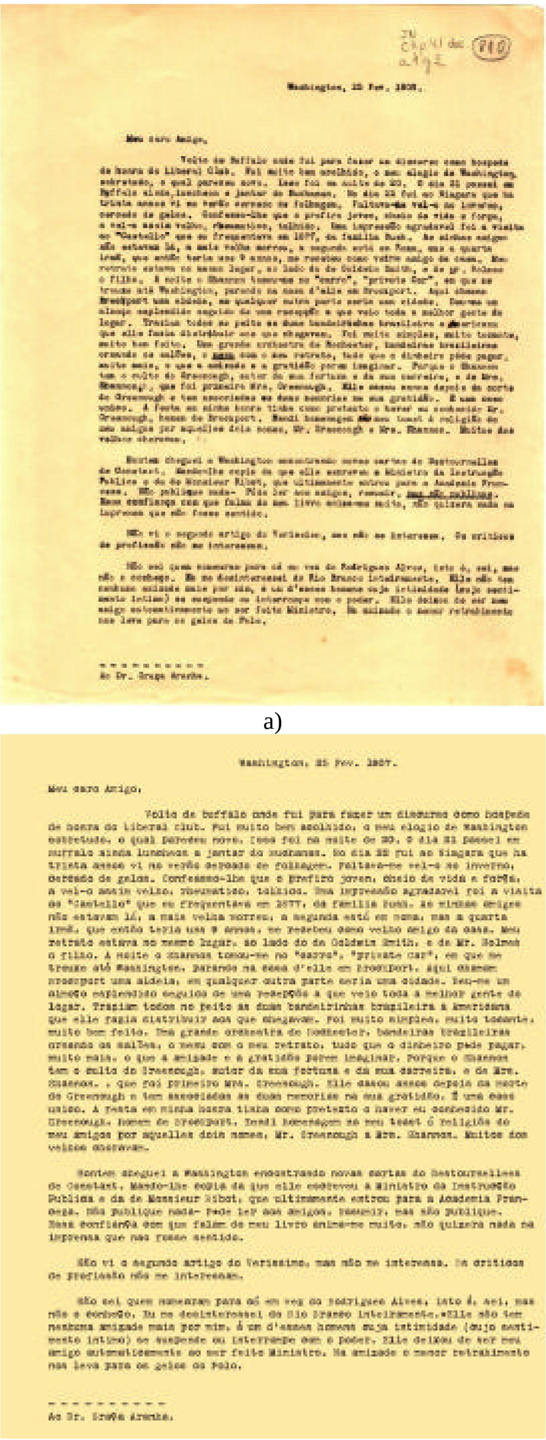
Nabuco.

b)

c)

d)

Figure 11. Application of the generation scheme on the image of the document d0765.
a) Original image, b) Ink image created after synthetis and colourising processes, c) Created texture and d) Final synthetic image.



IN Cap XI do a 49 I

810

Washington, 25 Fev. 1907.

Meu caro Amigo,

Volto de Buffalo onde fui para fazer um discurso como hospede de honra do Liberal Club. Fui muito bem acolhido, o meu elogio de Washington sobretudo, o qual pareceu novo. Isso foi na noite de 20. O dia 21 passei em Buffalo ainda luncheon e jantar do Buchanan. No dia 22 fui ao Niagara que ha trinta annos vi na verão despido de folhagem. Faltava-me vel-o no inverno, cercado de gelos. Confesso-lhe que o prefiro joven, cheio de vida e força, a vel-o assim velho, rheumatico, tolhido. Uma impressão agradavel foi a visita ao "Castello" que eu frequentava em 1877, da familia Rush. As minhas amigas não estavam lá, a mais velha morreu, a segunda está em Roma, mas a quarta irmã, que então teria uns 8 annos, me recebeu como velho amigo da casa. Meu retrato estava no mesmo lugar, ao lado do de Goldwin Smith, e de Mr. Holmes o filho. A noite o Shannon tomou-me no "carro", "private car", em que me trouxe até Washington, parando na casa d'elle em Brockport. Aqui chamam Brockport uma aldeia, em qualquer outra parte seria uma cidade. Deu-me um almoço esplendido seguido de uma recepção a que veio toda a melhor gente do logar. Traziam todos no peito as duas bandeirinhas brazileira e americana que elle fazia distribuir aos que chegavam. Foi muito simples, muito tocante, muito bem feito. Uma grande orchestra de Rochester, bandeiras brazileiras orneando os salões, o menu com o meu retrato, tudo que o dinheiro pode pagar, muito mais, o que a amizade e a gratidão podem imaginar. Porque o Shannon tem o culto do Greenough, autor da sua fortuna e da sua carreira, e de Mrs. Shannon, que foi primeiro Mrs. Greenough. Elle casou-se depois da morte de Greenough e tem associadas as duas memorias na sua gratidão. É uma casa unica. A festa em minha honra tinha como pretexto o haver eu conhecido Mr. Greenough, homem de Brockport. Rendi homenagem no meu toast á religião do meu amigo por aquelles dois nomes, Mr. Greenough e Mrs. Shannon. Muitos dos velhos choravam.

Hontem cheguei a Washington encontrando novas cartas do Destournelles de Constant. Mando-lhe copia da que elle escreveu ao Ministro da Instrucção Publica e da de Monsieur Ribot, que ultimamente entrou para a Academia Franceza. Não publique nada. Pode ler aos amigos, resumir, mas não publique. Essa confiança com que fallam de meu livro animou-me muito, não quizera nada na imprensa que não fosse sentido.

Não vi o segundo artigo do Verissimo, mas não me interessa. Os criticos de profissão não me interessam.

Não sei quem nomearam para cá em vez do Rodrigues Alves, isto é, sei, mas não o conheço. Eu me desinteressei do Rio Branco inteiramente. Elle não tem nenhuma amizade mais por mim, é um d'esses homens cuja intimidade (cujo sentimento intimo) se suspende ou interrompe com o poder. Elle deixou de ser meu amigo automaticamente ao ser feito Ministro. Na amizade o menor retrahimento nos leva para os gelos do Polo.

- - - - - - - - - - -
Ao Dr. Graça Aranha.

a)

b)

Figure 12. Application of the algorithm on the image of the document D0810-1.
a) Original image and b) its synthetic version.