Removing Highlighting in Paper Documents

Ricardo da Silva Barboza
DES – CTG – UFPE
Recife, PE, BRAZIL
rsbarboza@gmail.com

Rafael Dueire Lins
DES – CTG – UFPE
Recife, PE, BRAZIL
rdl@ufpe.br, rdl.ufpe@gmail.com

Victória de Souza Mattos
CEC – EST – UEA
Manaus – AM – BRASIL
vicky.mattos87@gmail.com

***Abstract* — Very often interested readers highlight documents with felt pens. Such marking may be seen as personal and physically "damaging" the original document, thus a recent taxonomy on noises on paper documents includes highlighting as a "physical noise". This paper addresses ways of filtering out highlighting in paper documents.**

Keywords* — *highlighting, paper documents, filtering.

I. Introduction

Very frequently readers make annotations in documents, for different reasons. In very seldom cases, such as the one that Fermat annotated the margins the Arithmetic of Diophantus of Alexandria, those annotations add little or nothing to the information of the document *per se*. Underlining was a kind of annotation in which the reader would somehow emphasize parts of a text for further reference.

In 1962, the modern felt-tip pen was invented in Japan by Yukio Horie. A *marker pen*, *marking pen*, *felt-tip pen*, or simply a *marker*, is a pen which has its own ink-source, and usually a tip made of a porous material, such as felt or nylon. Highlighters, such as the one used in this sentence, are permanent markers filled with transparent fluorescent ink used to cover texts, emphasizing such content. Highlighters are used to take notes in textbooks and every day becomes more popular. Many highlighters come in bright, often fluorescent colors, which glow under a black light. The most common color for highlighters is yellow, but they are also found in blue, green, orange, and magenta varieties. Red highlighters can be purchased along with a green translucent sheet used to hide the highlighted material. Some yellow highlighters may look greenish in color to the naked eye. Table 1 presents the most usual colors of highlighters and the components they affect of the original text.

Table 1. Component alteration due to highlighting

Highlight	Color	Components
🟨	Yellow	Blue
🟦	Blue	Red/Green
🟩	Green	Red/Blue
🟧	Orange	Green/Blue
🟪	Cyan	Red/Green
🟥	Magenta	Red/Green/Blue

Highlighting may be seen as personal annotation and physically "damaging" the original document, thus a recent taxonomy [1] on noises on paper documents includes highlighting as a "physical noise". Reference [1] says that the technical literature provides no solution for highlighting removal in document images. Real highlighting removal is far more complex than one may imagine at first glance, because the ink fades, sometimes non-uniformly, and interacts with the paper background.

II. Highlighting Removal in Images with Monochromatic Background

As one may observe in Table 1, most text highlighters commercially available affect one or more color components of the original document. The experiments performed showed that highlighting decreases the value of the intensity of the original color component, in relation to the unmarked areas. This observation was the starting point for the development of an algorithm for removing highlighting in monochromatic documents, shown in algorithm 1.

Algorithm 1 – Pseudo-code for the algorithm that removes highlighting that affects more than one primary RGB component in monochromatic documents.

```
For all pixels P(i) = (Red_i, Green_i, Blue_i) in the image do
    rg(i)=|Red_i – Green_i|;
    rb(i)=|Red_i – Blue_i|;
    gb(i)=|Green_i – Blue_i|;
    if ((rg(i) > limit) or (rb(i) > limit) or (gb(i) > limit)) then
      if (Red_i >= Blue_i) and (Red_i >= Green_i) then
         P(i) <- (Red_i, Red_i, Red_i);
      if (Green_i >= Red_i) and (Green_i >= Blue_i) then
         P(i) <- (Green_i, Green_i, Green_i);
      if (Blue_i >= Red_i) and (Blue_i >= Green_i) then
         P(i) <- (Blue_i, Blue_i, Blue_i);
```

In Algorithm 1 each pixel in the image is checked to see if any of the RGB components exceeds a preset limit. If so, all other components are exchanged for the value of the component with the greatest value. Figure 1 presents results of the Algorithm 1.

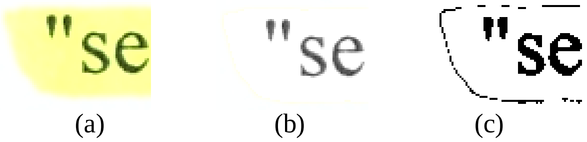
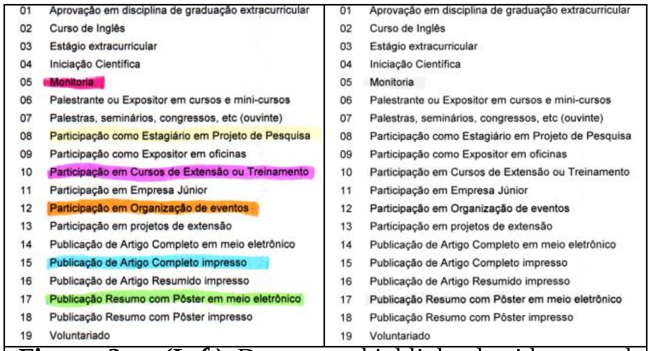


Figure 1 – (a) Image highlighted. (b) Image processed with Algorithm 1 (limit = 20). (c) Image 1b binarized and enhanced (threshold = 253) to visualization of the border effect.

The value of the limit influences the result of Algorithm 1 as high limit values tend to affect fewer pixels in the image, but yields a "border" effect in the edges of the highlighted areas (Figure 1b and 1c). To avoid the border effect the value of limit should be set to 4.

Figure 2 (left) shows pieces of text with parts highlighted with several colors. Processed with Algorithm 1 (limit = 4) to filter out the marking, the result may be found in the right part of the same figure. In all cases one may say that the result provided was satisfactory. The top highlighting (magenta) left some vestigial border, however.



01	Aprovação em disciplina de graduação extracurricular
02	Curso de Inglês
03	Estágio extracurricular
04	Iniciação Científica
05	Monitoria
06	Palestrante ou Expositor em cursos e mini-cursos
07	Palestras, seminários, congressos, etc (ouvinte)
08	Participação como Estagiário em Projeto de Pesquisa
09	Participação como Expositor em oficinas
10	Participação em Cursos de Extensão ou Treinamento
11	Participação em Empresa Júnior
12	Participação em Organização de eventos
13	Participação em projetos de extensão
14	Publicação de Artigo Completo em meio eletrônico
15	Publicação de Artigo Completo impresso
16	Publicação de Artigo Resumido impresso
17	Publicação Resumo com Pôster em meio eletrônico
18	Publicação Resumo com Pôster impresso
19	Voluntariado

01	Aprovação em disciplina de graduação extracurricular
02	Curso de Inglês
03	Estágio extracurricular
04	Iniciação Científica
05	Monitoria
06	Palestrante ou Expositor em cursos e mini-cursos
07	Palestras, seminários, congressos, etc (ouvinte)
08	Participação como Estagiário em Projeto de Pesquisa
09	Participação como Expositor em oficinas
10	Participação em Cursos de Extensão ou Treinamento
11	Participação em Empresa Júnior
12	Participação em Organização de eventos
13	Participação em projetos de extensão
14	Publicação de Artigo Completo em meio eletrônico
15	Publicação de Artigo Completo impresso
16	Publicação de Artigo Resumido impresso
17	Publicação Resumo com Pôster em meio eletrônico
18	Publicação Resumo com Pôster impresso
19	Voluntariado

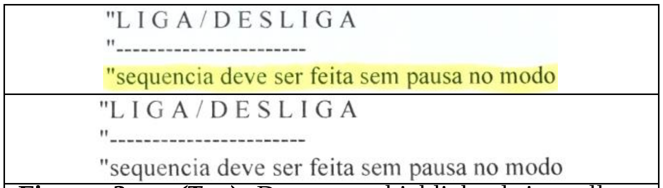
Figure 2 – ***(Left)*** Document highlighted with several colors. ***(Right)*** Highlight removed using Algorithm 1

Algorithm 1 may be optimized if the color of the highlighter is known *a priori*. For instance, if the marker used were yellow, the most widely used for highlighting, one knows that only the blue component is altered (as shown in Table 1), thus the comparison may test only the distance of the intensity from that component to the Red or Green component and copy, if the pixel is affected by the highlighting, the intensity of the component not affected to the Blue component. The optimized pseudo-code may be found in Algorithm 2. Although the component Red was used, exchanging it by the Green component will bring the same results in this case.

Algorithm 2 – Filtering-out yellow highlighting in monochromatic documents.

```
For all pixels P(i) = (Red_i, Green_i, Blue_i) in the image do
    if (|Red_i – Blue_i|> limit) then
        P(i) <- (Red_i, Red_i, Red_i)
```

Applying Algorithm 2 to the image of Figure 3 (top) taking limit=4, the result is shown in the bottom part, where one may observe that the highlighting was suitably removed.



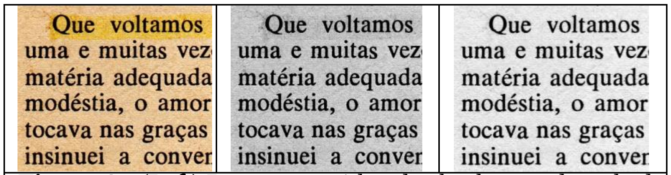
"LIGA/DESLIGA
"-----------------------
"sequencia deve ser feita sem pausa no modo

"LIGA/DESLIGA
"-----------------------
"sequencia deve ser feita sem pausa no modo

Figure 3 – ***(Top)*** Document highlighted in yellow. ***(Bottom)*** Highlight removed using Algorithm 2.

III. Highlighting Filtering in Documents with Color Background

Algorithms 1 and 2 work for monochromatic images digitized in true-color. This section focuses in yellow highlighting removal in documents with color background, such as the one shown in Figure 4 (left).



Que voltamos
uma e muitas vez
matéria adequada
modéstia, o amor
tocava nas graças
insinuei a conver

Que voltamos
uma e muitas vez
matéria adequada
modéstia, o amor
tocava nas graças
insinuei a conver

Que voltamos
uma e muitas vez
matéria adequada
modéstia, o amor
tocava nas graças
insinuei a conver

Figure 4 - ***(Left)*** Document with color background marked in yellow. ***(Center)*** Same document in gray scale. ***(Right)*** Same document with processed with Algorithm 1.

The direct application of Algorithm 1 to a color image produces an image in gray scale with the highlighting removed. Figure 4 shows an example of a document with color background converted into gray scale with a standard and Algorithm 1. One may observe that the image obtained from Algorithm 1 removed the shade of the marker. This section analyses the development of an algorithm capable of removing highlighting in color documents. In general, separating text and background in degraded documents is a complex task [2]. The first step in this direction is to analyze the effect of highlighting in gray scale images.

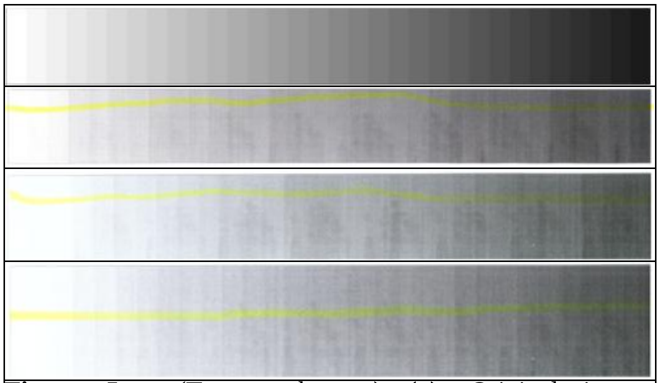


Figure 5 - (Top to bottom) **(a).** Original image. **(b).** Image with Pilot™ marker. **(c).** Image with Pointer™ marker. **(d).** Image with Zolben™ marker.

The image of Figure 5 presents a stripe of an image of 32 hues of gray from black to white. After printing, a line was made with highlighters of three different manufacturers.

Some statistical analysis was performed with samples of the different areas of the gray-level stripes above with and without highlighting. The analysis of the RGB components was plotted in the histograms shown in Figure 6.

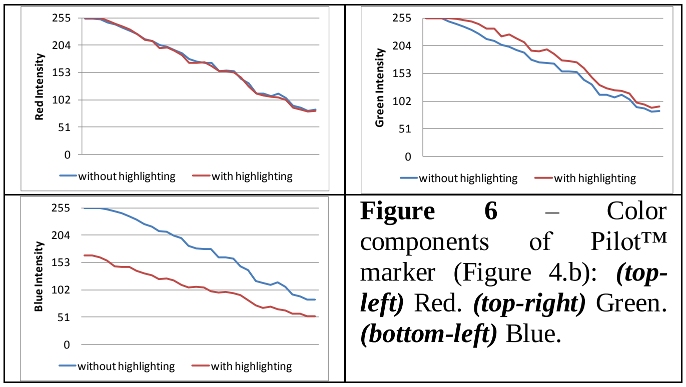


Figure 6 – Color components of Pilot™ marker (Figure 4.b): ***(top-left)*** Red. ***(top-right)*** Green. ***(bottom-left)*** Blue.

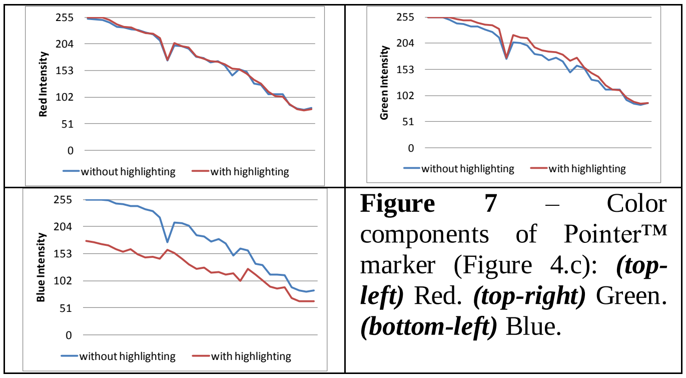


Figure 7 – Color components of Pointer™ marker (Figure 4.c): ***(top-left)*** Red. ***(top-right)*** Green. ***(bottom-left)*** Blue.

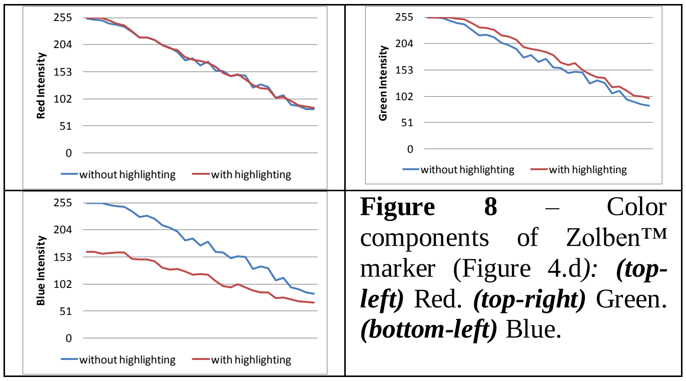


Figure 8 – Color components of Zolben™ marker (Figure 4.d*):* ***(top-left)*** Red. ***(top-right)*** Green. ***(bottom-left)*** Blue.

From the plotting from Figures 6 to 8 one may see that the red component suffered no alteration due to yellow highlighting. The green component suffered a slight attenuation. The blue component suffers a strong attenuation, overall in the lighter hues. This behavior is linked to the intrinsic use of highlighters that are more effective in lighter backgrounds.

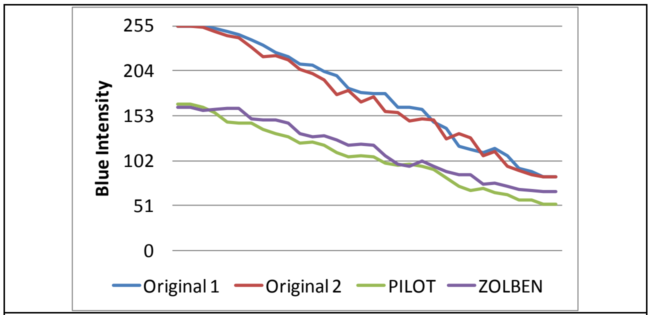


Figure 9 - Variation of the blue component for Pilot™ and Zolben™ markers in the image shown in Figure 3 (Left).

The analysis of Figure 9 allows one to say that the two markers affected the blue component similarly. This allows one to simulate the attenuation performed by the marker in the Blue component as:

$$g(x)=\frac{b-x}{b-a}f(a)+\frac{x-a}{b-a}f(b) \tag{1}$$

where:

b = 255 (white).

a = 0 (black).

$f(a)$ = 0 (the value of the intensity of the black pixels was kept unchanged after highlighting).

$f(b)$ = 165; (the average value of the blue component of the white parts after being highlighted in yellow with Pilot and Zolben markers).

x = New_Blue_hue

$g(x)$ = Old_Blue_hue

Thus, the formula for restoring the original intensity of the blue component of the highlighted area is:

New_Blue_hue = (Old_Blue_hue/165) * 255) (2)

Now, one needs to spot the yellow highlighted area for which the blue component will be corrected. In the case of the image shown in Figure 4 (left), it was split into three areas: the text, the unmarked background texture (paper), and the yellow highlighted area. The text area is formed by pixels with dark hues with an average intensity of less than 128, as shown in Figure 10. As the text is mainly formed by very dark pixels, the marker does not alter noticeably their intensity, as seen in the graph presented in Figure 9. This is exactly the role of highlighters, to color the background keeping the text legible.

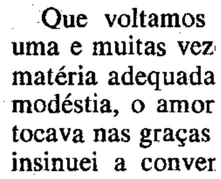

Que voltamos
uma e muitas vez
matéria adequada
modéstia, o amor
tocava nas graças
insinuei a conver

Figure 10 – Pixels of low hue intensity from Figure 4 (left), which are weakly affected by the yellow marker.

In order to find the highlighted area, the original texture shown in Figure 11 (left) was analyzed and it was observed that the image has a large quantity with the same red component and green of high intensity, as presented in the histogram of Figure 12. For all other pixels in the image, a routine similar to the one in Algorithm 1 was used. All pixels with intensity higher than 128 for which the blue component is 100 units away from the red and green components yields an image such as the one presented in Figure 11 (center). After filter the salt-and-pepper noise we have the final mask (Figure 11 right) that we will use to remove the highlighted area.

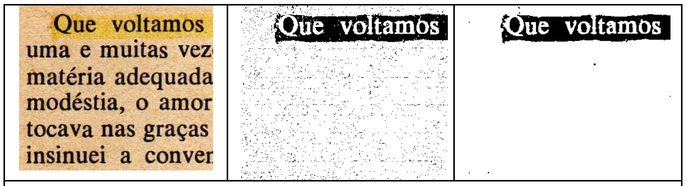

Que voltamos
uma e muitas vez
matéria adequada
modéstia, o amor
tocava nas graças
insinuei a conver
Que voltamos
Que voltamos

Figure 11 – ***(Left)*** Original document. ***(Center)*** Mask to identify the yellow highlighted pixels from Figure 4 (left). ***(Right)*** Final mask with salt-and-pepper noise filtered.

Figure 11 (center) shows some salt-and-pepper noise in the area off-highlighting, as well as in the marked area. After salt-and-pepper removal one obtains a mask with the pixels of the background of the highlighted area. Figure 12 presents the RGB-histogram for the highlighted background area and paper texture.

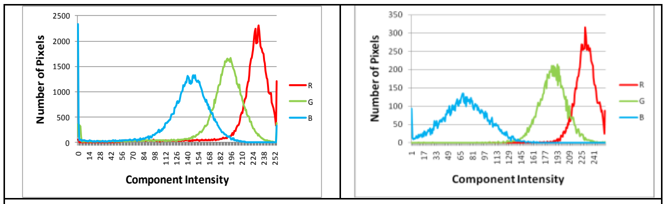


Figure 12 – ***(Left)*** RGB-histogram of the paper texture. ***(Right)*** RGB-histogram of the highlighted area.

The highlight removal algorithm works on the marked area (Figure 11 right) performing the operations described in Algorithm 3, which alters the blue components of each pixel in the yellow highlighted area applying formula 2.

Algorithm 3 – Filtering-out yellow highlighting in documents with colored background.

```
For all pixels P(i) = (Red_i, Green_i, Blue_i) in the yellow highlighted area do
        P(i) <- (Red_i, Green_i, (Blue_i/165) * 255)
```

The resulting image is shown in Figure 13, where one may observe that despite the fact that the effect of highlighting was made weaker, it is still clearly visible. As one may observe the histogram obtained for the highlighted area became closer to the one in Figure 12 (left) for the paper texture.

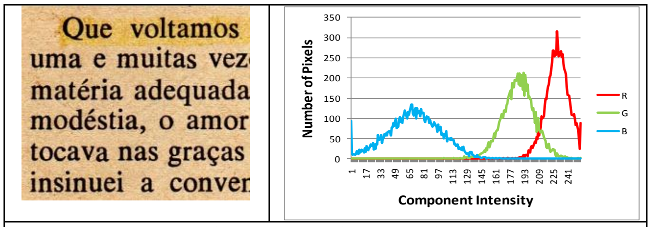


Figure 13 – ***(Left)*** Image after the first filtering step. ***(Right)*** RGB-histogram of the highlighted area of the image to the left.

Although the result obtained after the first iteration of the algorithm not being completely satisfactory, one may observe that the RGB-histogram for the highlighted area obtained (Figure 13 right) has become much closer to the one of the RGB-histogram of the paper texture, shown in Figure 12 left. The resulting image is re-filtered with Algorithm 3, yielding the images and histograms obtained in Figures 14 to 16.

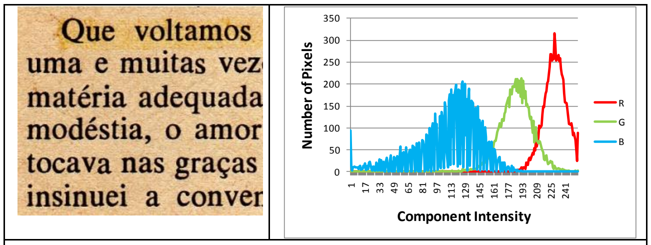


Figure 14 – ***(Left)*** Image after the second filtering step. ***(Right)*** RGB-histogram of the highlighted area of the image to the left.

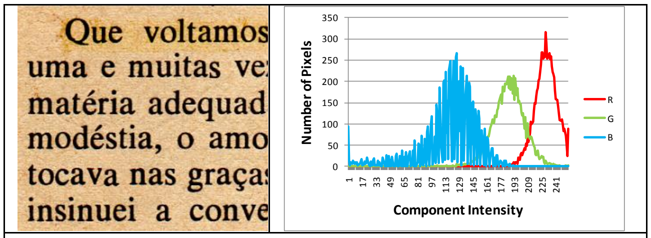


Figure 15 – ***(Left)*** Image after the third filtering step. ***(Right)*** RGB-histogram of the highlighted area of the image to the left.

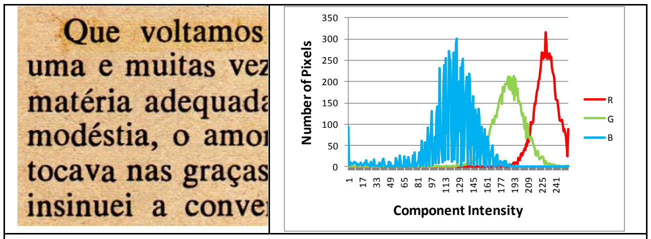


Figure 16 – ***(Left)*** Image after the fourth filtering step. ***(Right)*** RGB-histogram of the highlighted area of the image to the left.

As one may observe in Figure 16, after the fourth iteration of Algorithm 3 on the yellow highlighted area the marking becomes much weaker and almost imperceptible.

IV. Conclusions and Lines for Further Work

This paper presents three efficient algorithms for removing highlighting in textual documents. The first two address monochromatic documents and the third addresses the problem of yellow markers in color background.

The first algorithm covers markers of several colors, and the second is made more efficient for yellow highlighting, the most frequently used color.

The algorithm for color or aged paper background works iteratively. A new algorithm in which the highlighted pixels are replaced by other pixels (non-marked) copied from other parts of the paper background keeping the same entropy [3], along the lines suggested in references [4] and [5], is under development, and should provide an even better visual result and performance. Another alternative to be explored is the use of a noise classifier, such as the one described in [6], to select the highlighted areas for different colors of markers.

Acknowledgment

V.S.M. was sponsored by research grants received from Fundação de Amparo à Pesquisa do Estado do Amazonas – FAPEAM.

References

[1] Lins, R.D. A Taxonomy for Noise Detection in Images of Paper Documents - The Physical Noises. ICIAR 2009. LNCS v. 5627. p. 844-854, Springer Verlag, 2009.

[2] G. Leedham, S. Varma, A. Patankar, V. Govindaraju, *Separating text and background in degraded document images—a comparison of global thresholding techniques for multi-stage thresholding*, Proceedings of the Eighth International Workshop on Frontiers in Handwritten Recognition, pp. 244–249, 2002.

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

[4] C. A. B. Mello and R. D. Lins. *Image segmentation of historical documents*, Visual 2000, Mexico City, Mexico.

[5] C. A. B. Mello and R. D. Lins. *Generation of images of historical documents by composition*. ACM Document Engineering 2002, McLean, VA, USA.

[6] Lins, R.D, Silva, G.F.P., Banergee, S., Kuchibhotla, A. and Thielo, M. *Automatically Detecting and Classifying Noises in Document Images*, ACM-SAC`2010, ACM Press, v.1. p.33 – 39, March 2010.