Restoration of Old Document Images using Different Color Spaces
Restoration of Old Document Images

Ederson Marcos Sgarbi[1], Wellington Aparecido Della Mura[1], Nikolas Moya[2] Jacques Facon[3] and Horacio A. Legal Ayala[4]

[1]*Universidade Estadual do Norte do Paraná, Campus Luiz Meneghel, Bandeirantes, Paraná, Brazil*

[2]*UNICAMP - Universidade de Campinas, Campinas, SP, Brazil*

[3]*PPGIa, PUCPR-Pontifícia Universidade Católica do Paraná, Curitiba, Paraná, Brazil*

[4]*FPUNA, Universidad Nacional de Asunción, Campus de la UNA, San Lorenzo, Paraguay*

{*sgarbi, wellington*}*@uenp.edu.br, nikolasmoya@gmail.com, facon@ppgia.pucpr.br, hlegal@pol.una.py*

Keywords: Old Document, Restoration, Mathematical Morphology, *HSI*, *YCrCb*, *YIQ*.

Abstract: An obstacle in old document interpretation comes from the lack of image quality. Old documents frequently appear with digitization errors, uneven background, bleed-through effect. A new approach based on morphological color operators to restore the color text is presented. The morphological tools are based on three color spaces, *HSI* well known in morphological processes, *YCrCb* and *YIQ* rarely used in morphological procedures. Experimental results carried onto 100 old documents have proven that using *YCrCb* and *YIQ* is as effective as using *HSI* to recover ancient texts in uneven and foxed background images, without presenting problems in hue ordination.

1 INTRODUCTION

Uneven background, bleed-through effect, foxing marks, ink and paper alterations and digitization errors are common in old document images. To improve printed and handwritten text is a complex challenge.

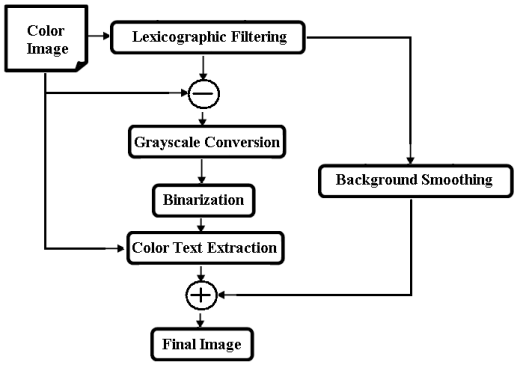


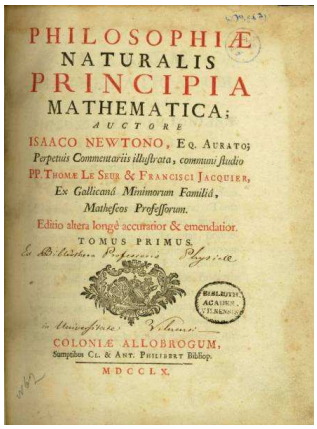
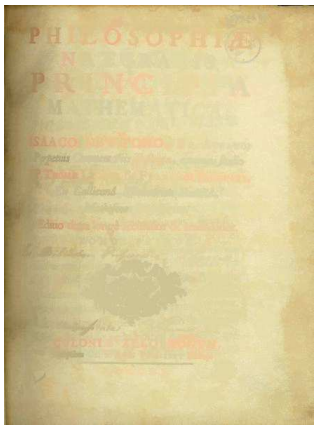
Figure 1: Restoration strategy flowchart of old document images

We can find in the literature some works about restoration of old documents. For instance, (Ramponi et al., 2005) have proposed a restoration strategy based on rational filter on *YCbCr* space. After converting the *RGB* image to *YCbCr*, the rational filter is then applied to Y luminance component to highlight the edges and smooth regions, generating the image Y_r. A specify tonality adjustment curve is used in Y_r to improve the contrast between the text and the background image. This result is multiplied by Y/Y_r to adjust the chrominance. An estimate of the background image Y_{bw} is obtained by Otsu 's thresholding method. The chrominance channels of background pixels are then modified, while the others remain unchanged. Finally the image in the *YCbCr* space is converted back to *RGB*.

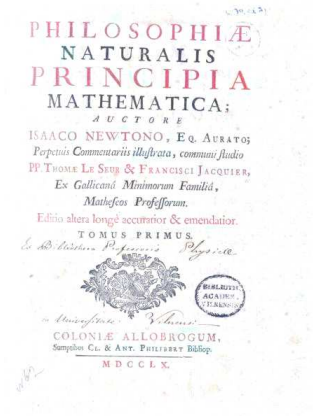
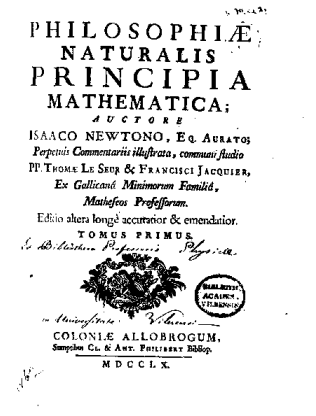
Drira et al (Drira et al., 2007) have proposed a restoration strategy of color document images mixing the standard Mean Shift algorithm with a modified one. First the *RGB* image is converted to $L^*u^*v^*$. A modified Mean Shift algorithm using the R-Nearest Neighbors Colors concept extracts different local maxima. The global maxima are then extracted by standard Mean Shift algorithm. And finally the pixels are classified to the closest previously extracted mode.

In this paper we propose a morphological methodology to improve the paper background appearance and to preserve the original characters by using a morphological lexicographic order applied to *HSI*, *YCrCb* and *YIQ* color spaces. And a restoration comparison is performed.

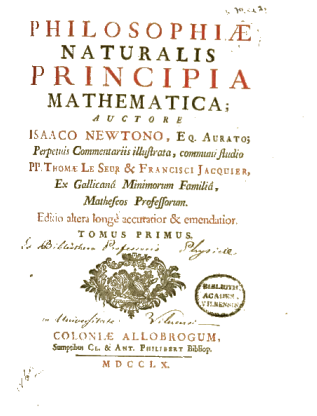
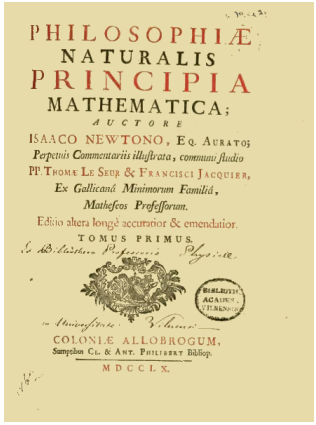
The rest of the paper is organized as follows. Sec-

a) Original image (c) $(I \rightarrow S \rightarrow H)$ Background Estimation

c) Tophat (d) Binarization

c) Colored Text Extraction (d) Background Estimation with Text Restoration

Figure 2: Restoration Strategy.

tion 2 describes the proposed approach based on morphological background estimation. Experimental results are discussed in Section 3.

2 METHODOLOGY

The document restoration strategy is composed of background estimation, original colored text extrac-

tion and image recovery. The image background approximation will be carried out by morphological operators based on a lexicographic order applied to *HSI*, *YCrCb* and *YIQ* color spaces. Figure 1 depicts the restoration strategy. And Figure 2 shows an example of the restoration strategy with *HSI* space.

2.1 Color Mathematical Morphology

The theory of Mathematical Morphology is to quantitatively describe geometric structures present in the image, and offers a wide range of tools (Soille, 2004). Motivated by the recent researches on extension of morphological operations to color images (Aptoula and S., 2009), (Hanbury, 2001),(Peters, 1997), we have decided to employ color mathematical morphology tools to achieve old document restoration in a flexible and fast manner. Based on complete lattices, the theory of mathematical morphology, requires an algebraic structure T (complete lattice) and an order relation "$\leq$". The lexicographic order (Aptoula and S., 2009) was chosen in our approach. Let's define the three channels as $C1$, $C2$ and $C3$. The maximum value between two three-component vectors $P(p_{C1}, p_{C2}, p_{C3})$ and $Q(q_{C1}, q_{C2}, q_{C3})$, also named as supremum $\vee$, is defined as follows:

$$\vee\{P,Q\} = \begin{cases} p_{C1} > q_{C1} \\ or \\ p_{C1} = q_{C1} \text{ and } p_{C2} > q_{C2} \\ or \\ p_{C1} = q_{C1} \text{ and } p_{C2} = q_{C2} \text{ and } p_{C3} > q_{C3} \end{cases}$$

Then, from this lexicographic order, morphological dilation δ of image f by using the structuring element B at the pixel x with respect to structuring element support set $E \subset \Re$ are defined as follows:

$$\delta^B(f(x)) = \vee\{f(y) + B(x-y) : y \in E\} \quad (1)$$

We can find in the literature many works using *HSC* color spaces (where $C = I$ or V or L) coupled with morphological tools (Ortiz et al., 2001) (Ortiz et al., 2002) (Tobar et al., 2007) . But it exists few works using *YCrCb* and *YIQ* color spaces (Popov, 2007). In our approach, we have decide to use and compare the potentialities of *HSI*, *YCrCb* and *YIQ* color spaces in morphological restoration of old documents.

The use of lexicographic order with *HSI* needs the use of a hue reference H_{ref} as follows ((Hanbury, 2001), (Peters, 1997)):

$$d(H_i, H_{ref}) = \begin{cases} |H_i - H_{ref}| \text{ if } |H_i - H_{ref}| < \pi \\ or \\ 2\pi - |H_i - H_{ref}| \text{ if } |H_i - H_{ref}| > \pi \end{cases}$$

2.2 Background Estimation

The background estimation is based on reconstruction function ρ from geodesic dilation $\delta_{...}$, i.e.,

$$\rho(f) = \lim_{n \to +\infty} \underbrace{\delta_g(\delta_g(...\delta_g(f)))}_{n} \text{ with } \delta_g(f) = \delta(f) \wedge g \quad (2)$$

where the color geodesic dilation $\delta_g(f)$ is based on restricting the color dilation of the marker image f to the mask g.

In our proposed approach, the background estimation based on color reconstruction is performed as follows:

- First the mask image g is inverted;
- Then the the marker image f is composed as follows:
  - The first row and column and the last row and column of the marker image are composed of g image pixels;
  - The rest of the pixels are defined as $H = S = I = 0$ or $Y = Cb = Cr = 0$ or $Y = I = Q = 0$, depending on the used color space.

2.3 Text Segmentation

Featuring an estimation of the background image with its peculiarities, the text can be segmented by subtracting the original image g from its estimated background $\delta_g(f)$. The colored text images are then converted to grayscale and segmented by the Johannsen 's binarization (Johannsen and Bille, 1982).

2.4 Old Document Image Restoration

To recover and reconstruct old document images, three steps as performed:

- First the binarized image is used to mask the original colored text in the original image. This step permits to feature an image with original color text;
- Then an averaged estimation of background is carried out from the morphological reconstruction image. This step permits to feature an image with its averaged background;
- Image Recovery Finally the Image Recovery of the old document image is carried out by adding the two images.

To recover and reconstruct an old document image, three steps as performed: First the binarized image is used to mask the original colored text in the

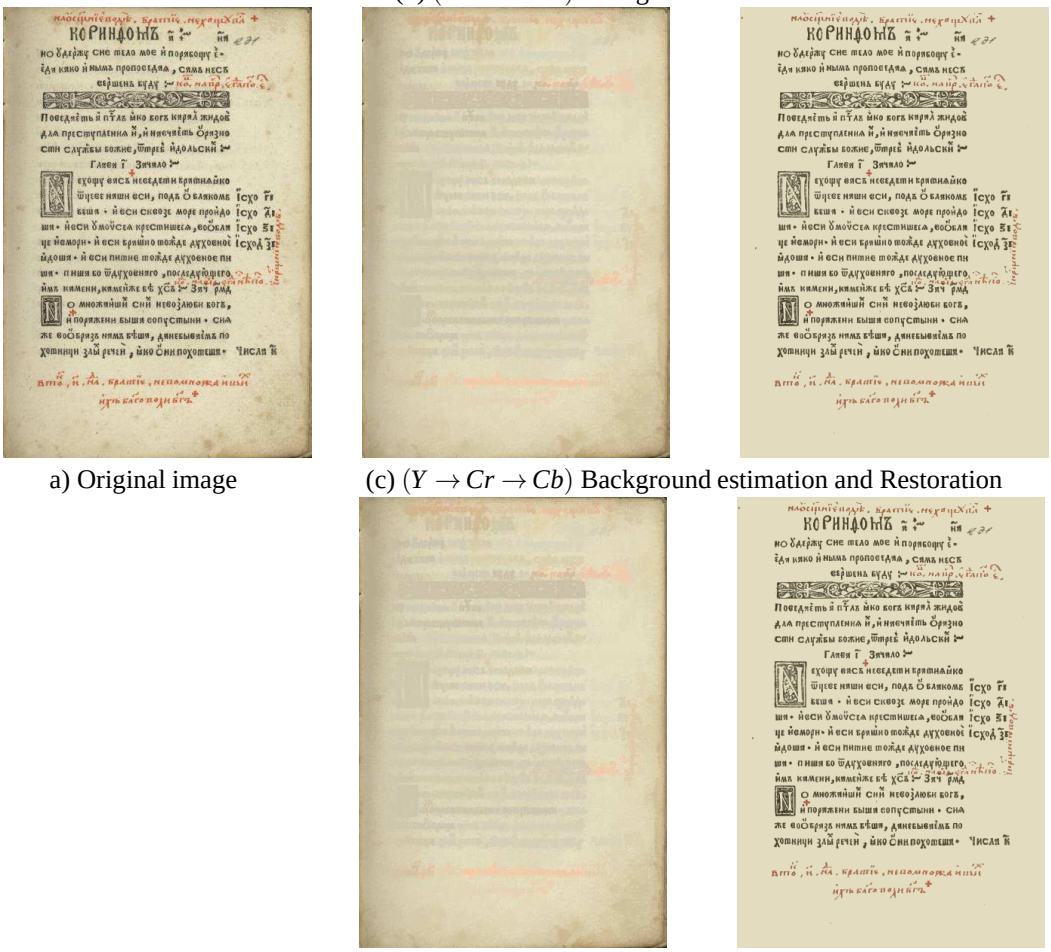
(b) $(I \rightarrow S \rightarrow H)$ Background estimation and Restoration

a) Original image

(c) $(Y \rightarrow Cr \rightarrow Cb)$ Background estimation and Restoration

(d) $(Y \rightarrow I \rightarrow Q)$ Background estimation and Restoration

Figure 3: Restoration Experiment.



(b) $(I \rightarrow S \rightarrow H)$ Background estimation and Restoration

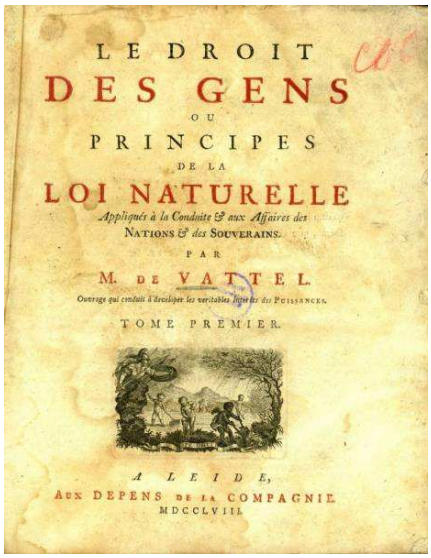
a) Original image

(c) $(Y \rightarrow Cr \rightarrow Cb)$ Background estimation and Restoration

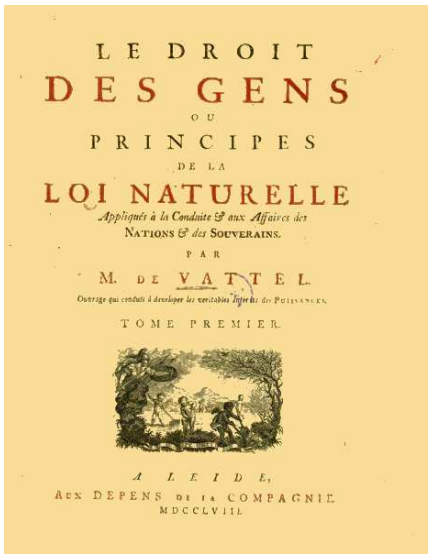
(d) $(Y \rightarrow I \rightarrow Q)$ Background estimation and Restoration

Figure 4: Restoration Experiment.

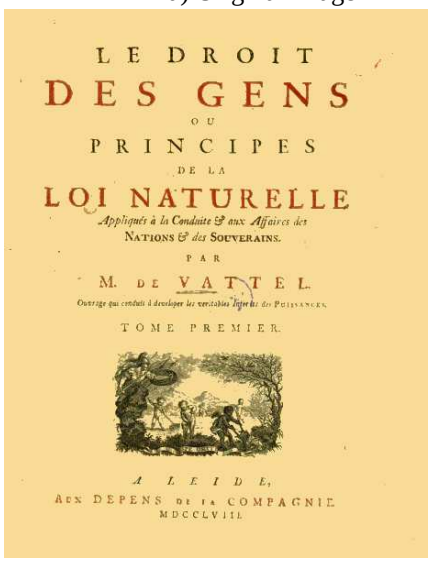
original image. This step permits to feature an image with original color text. Then an averaged estimation of background is carried out from the morphological reconstruction image. This step permits to feature an image with averaged background. Finally the reconstruction of the old document image is carried out by adding the two images.



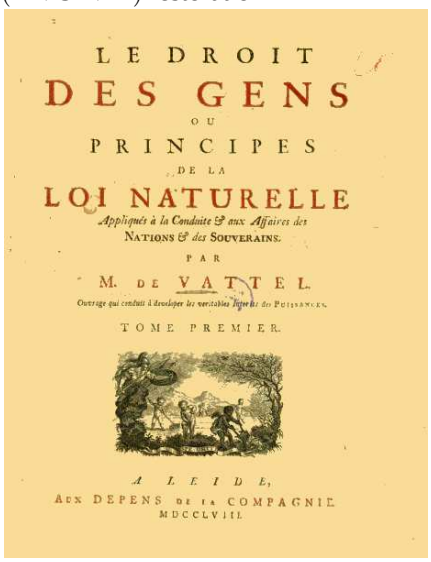
a) Original image



b) $(I \rightarrow S \rightarrow H)$ restoration



c) $(Y \rightarrow Cr \rightarrow Cb)$ restoration



d) $(Y \rightarrow I \rightarrow Q)$ restoration

Figure 5: Restoration Comparison

3 EXPERIMENTS

Since it is possible to apply the lexicographic ordination (equation 1) with different color spaces, experiments ware carried out to answer to these questions: are *HSI*, *YCrCb* and *YIQ* color spaces suitable to perform an efficient restoration of old document images? And can we conclude that one of these color spaces is more suitable?

To evaluate the approach to *HSI*, *YCrCb* and *YIQ* color spaces and to answer to these questions, a data base of 100 old documents has been created.

To apply the lexicographic ordination, the order of channels has to be done first. Following the assessment of (Ortiz et al., 2001) and (Ortiz et al., 2002) where the authors have declared that around 95% of the processed pixels are sorted by the luminance, we have decided to use the ordination with luminance in the first place as followed, $I \rightarrow S \rightarrow H$ and $Y \rightarrow Cr \rightarrow Cb$ and $Y \rightarrow I \rightarrow Q$.
By testing the three lexicographic ordinations to 100 old document images, we could observed that (Figures 3, 4, 5):

- *HSI*, *YCrCb* and *YIQ* color spaces are suitable to

perform efficient old document restorations;

- The obtained results are similar.
- The use of *YCrCb* and *YIQ* color spaces is faster and more simple than using *HSI* where a hue reference H_{ref} (equation 2) is necessary.

Figures 3, 4, 5 show that, without using prior knowledge about the old document images, the proposed approach is efficient in removing the heterogeneous and foxed background, in restoring the background and in recovering in a realistic manner the old document images.

4 CONCLUSIONS

A new approach based on morphological color operators to restore old document images has been proposed. By comparing the background estimation and restoration, based on morphological tools onto *HSI*, *YCrCb* and *YIQ* color spaces, we can conclude that they are suitable to restore old document images and despite the rare use in morphological processes, it is faster and more advantageous to apply *YCrCb* and *YIQ* color spaces than *HSI* where a hue reference H_{ref} (equation 2) is necessary. It is possible to conclude that the luminance channel should have lexicographic precedence compared to the other channels.

REFERENCES

Aptoula, E. and S., L. (2009). Multivariate mathematical morphology applied to color image analysis. In *Chapter 10: Multivariate Image Processing , 2009*. Christophe Collet, Jocelyn Chanussot, Kacem Chehdi.

Drira, F., Lebourgeois, F., and Emptoz, H. (2007). A coupled mean shift-anisotropic diffusion approach for document image segmentation and restoration. In *ICDAR, The 9th International Conference on Document Analysis and Recognition*, pages 814–818. ICDAR.

Hanbury, A. (2001). Lexicographical order in the hsl colour space. In *TechnicalReport N-04/01/MM Centre de Morphologie Mathmatique*. ÉCole Des Mines de Paris, 2001.

Johannsen, G. and Bille, J. (1982). A threshold selection method using information measures. In *Proceedings, 6th Int. Conf. Pattern Recognition*. Proceedings, Munich, Germany, pp. 140-143.

Ortiz, F., Torres, F., Angulo, J., and Puente, S. (2001). Comparative study of vectorial morphological operations in different color spaces. In *Proc. Of Intelligent Robots and Computer Vision XX: Algorithms, Techniques, and Active Vision*. SPIE vol.4572 pp.259-268.

Ortiz, F., Torres, F., De Juan, E., and Cuenca, N. (2002). Colour mathematical morphology for neural image analysis. In *Real-Time Imaging*. Vol. 8 pp. 455-465.

Peters, A. (1997). Mathematical morphology for angle valued images. In *Proceedings of SPIE Non-Linear Image Processing VIII*. Proceedings SPIE Vol. 3026, p.84-94.

Popov, A. T. (2007). Fuzzy mathematical morphology and its applications to colour image processing. In *Plzen*. Czech Republic: Union Agency, Science Press.

Ramponi, G., Stando, F., Russo, W., Plusi, S., and Paolo, M. (2005). Digital automated restoration of manuscripts and antique printed books. In *Proceeding of EVA 2005, Electronic Imaging and the Visual Arts - Firenze, Italia pp. 764-767*. Proceeding of EVA.

Soille, P. (2004). Morphological image analysis: Principles and applications. In *Springer, 2004*. 2nd Edition.

Tobar, M. C., Platero, C., Gonzalez, P. M., and Asensio, G. (2007). Mathematical morphology in the hsi colour space. In *Pattern Recognition and Image Analysis*. Lecture Notes in Computer Science, vol 4478, pp. 467-474.