2011 International Conference on Document Analysis and Recognition

Chromatic / achromatic separation in noisy document images

Asma OUJI, Yann LEYDIER, Frank LEBOURGEOIS
Université de Lyon, CNRS
INSA-Lyon, LIRIS, UMR5205, 20 av. Albert Einstein
Villeurbanne, F-69621, France
E-mail: {asma.ouji, yann.leydier, frank.lebourgeois}@liris.cnrs.fr

***Abstract*—This paper presents a new method to split an image into chromatic and achromatic zones. The proposed algorithm is dedicated to document images. It is robust to the color noise introduced by scanners and image compression. It is also parameter-free since it automatically adapts to the image content.**

***Keywords*-color noise; chromatic /achromatic split; document image;**

I. Introduction

In document images, color pixels can not be treated in the same way as the gray-scale ones. Indeed, it would be meaningless to consider the hue in a gray-scale image. As most of document images consist of some color areas (*e.g.* figures) and some gray-scale ones (*e.g.* black and white text), it is often necessary to separate color pixels from gray-scale ones to process them differently. Such an issue seems to be obvious. However, digitized images are usually altered by color noise (see figure 1) introduced by scanners, especially when they are incorrectly set, and by image compression. This noise consists in chromatic pixels inside grey-scale areas.

We call a chromatic pixel any pixel having a defined hue (red, green, blue, yellow, *etc*). Otherwise, a pixel is called achromatic (shades of gray, including black and white).

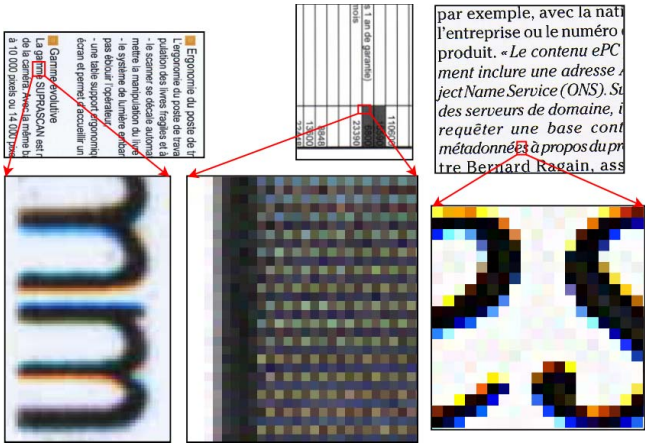

Figure 1. Samples of the color noise

The chromatic / achromatic split has several further applications. The most immediate one is to clean the image and get rid of the noise. Thus, our method's output would enhance quantization algorithms. We can also extract relevant features based on the chromatic / achromatic separation to classify document images [1]. Furthermore, it is a valuable preprocessing step for OCR (Optical Character Recognition), coding and image compression.

Distinguishing between true colors present in the original resource and noise colors introduced by the digitization process is a complicated task. Isolated chromatic zones can be noise as well as a small color object (e.g.: a bullet at the beginning of a line, such as in figure 1). Furthermore, some color noise, in low resolution and dithered images, makes it impossible to guess the original color in the source document and even to know whether the zone is supposed to be perceived as chromatic or achromatic. Figure 2, where the Nyquist-Shannon sampling theorem was not respected, shows an example of this kind of noise.

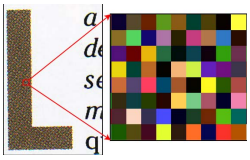

Figure 2. Uninterpretable multicolored dithering

Very few works dealing with chromatic / achromatic separation exist in the literature. In [2], the separation is simply carried out by thresholding the Saturation [3] channel. Such a method is not suited to noisy images since it does not consider the color noise. Karatzas *et al.* [4] perform the chromatic / achromatic separation in web images based on the human perceptual model. This method consists in thresholding the HSV colors based on human perception. However, web images do not include the same type of noise as digitized images. Thus, the color noise cannot be removed by such a method. Several researchers [5], [6] have been interested in distortions caused by digitization in gray-scale document images but color noise has not been handled yet.

This paper proposes an algorithm that separates chromatic zones from achromatic ones resulting in a binary mask telling whether any pixel is chromatic or not. A statistical measure that estimates the stroke's thickness in the document

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.42

image is proposed. It allows the system to be completely automated and parameter-free.

The chromatic area detection is based on a new measure called pseudo-saturation and a series of morphological and smoothing operations.

All the steps are based on pixel-based approaches since, unlike structure-based algorithms, pixel-based methods do not require any document model or any *a priori* information on the document class.

II. Stroke thickness estimation

Most of the document image processing methods depend on a resolution related parameter (e.g.: width of a structuring element, size of a convolution matrix, *etc.*). To determine such values, it would be rather futile to lean on the digitization resolution. Indeed even if all documents were digitized in the same resolution (some providers proclaim that 300dpi is universally suitable!), each document has its own typographic characteristics. Modern letters and invoices are usually composed with 10 or 12 points fonts but advertisements, flyers and journals use fancy typography. Given a specific size, the strokes of a font may have variable thickness depending on the typeface appearance (see fig. 3).

Figure 3. Two fonts with the same size but with different stroke thickness.

In order to ensure that the algorithms presented in this paper will be free of a "resolution" parameter, we will base all our metrics on an estimation of the font's stroke thickness. We will restrain from binarizing the image because a high quality binarization is time-consuming.

A quick and easy way to estimate the strokes' width and height without binarizing the image is to compute the autocorrelation of the image along the horizontal and vertical axes. The estimation would obviously be more accurate if we determined the main orientation of the strokes but this would cost too much time. This is a statistical and global measure. Its purpose is not to describe precisely a font but to indicate the order of magnitude of the main text's strokes thickness.

Let $T_h(I, \delta)$ be the translation of the gray-scale image I (defined on the plane Ω) of δ pixels along the horizontal axis. We define the sequence $\left(\mathcal{D}^h(I)_n\right)_n$ with:

$$\begin{aligned} \mathcal{D}^h(I)_0 &= 0 \\ \mathcal{D}^h(I)_n &= \sum_{(x,y)\in\Omega} \|I(x,y) - T_h(I,n)(x,y)\| \end{aligned} \quad (1)$$

The sequence $\left(\mathcal{D}^h(I)_n\right)_n$ is asymptotic. To estimate the mean stroke width $\mathcal{S}_w$ of an image, we compute the sequence until $n = n_m$, when its growth rate becomes lower than 10%. Then, we obtain $\mathcal{S}_w = n_m$

The computation of the mean strokes' height $\mathcal{S}_h$ goes the same with the vertical translation $T_v(I, \delta)$ and the sequence $\left(\mathcal{D}^v(I)_n\right)_n$.

We tested this algorithm on various documents (see fig. 4) and we measured the strokes' width and height manually. The strokes' thickness range was from 2 to 10 pixels inclusive. The mean error between $\mathcal{S}_w$ and the measures was 1.25 pixels, which is quite good since the images are never binarized in the process. Due to prominence of vertical strokes in the writing, the mean error between $\mathcal{S}_h$ and the measures was 1.75 pixels.

Figure 4. Samples of the images used to test the strokes' thickness estimation.

The presence of pictures naturally affects the measurement but experiments have shown that the mean stroke thickness is never modified by more than one pixel. In an image with text in multiple fonts or sizes, the mean strokes' thickness is close to the thickness of the most represented font. Small text has many horizontal (resp. vertical) boundaries that respond well to the autocorrelation and limit the effects of the presence of text in large fonts (such as titles).

Since a stroke's edge in a gray-scale image is nearly always smoothed, it is impossible to precisely measure its thickness. Therefore, having a mean error lower than 2 pixels proves that our estimator is accurate enough.

In the following, as we do not know the orientation of the document images, we will use the estimator $\mathcal{S}_t = \max(\mathcal{S}_w, \mathcal{S}_h)$.

III. Chromatic / achromatic decomposition

This section aims to create a binary mask where each entry indicates whether the corresponding pixel is chromatic or not.

The algorithm encompasses three main steps: the computation of the pseudo-saturation image S^*, the coarse mask $\mathcal{M}_C$ generation and the final mask $\mathcal{M}_F$ inference.

A. Computation of the pseudo-saturation

Chromatic / achromatic separation is usually achieved by thresholding the saturation channel [2]: chromatic content is generally highly saturated whereas achromatic pixels have low values of saturation. However, the standard saturation of dark pixels is disruptive since its computation involves division by lightness which is nearly null. For instance, figure 5.b shows black (achromatic) zones having greater values of saturation than the green (chromatic) background.

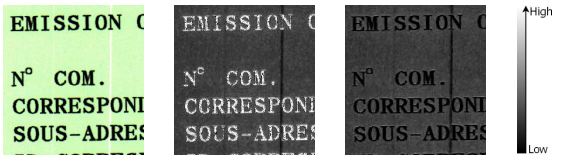

Figure 5. a. Color image, b. Saturation image, c. Pseudo-saturation image, d. Scale

For this reason, we introduce a new pseudo-saturation measure, which is defined below, instead of the traditional saturation. Let a color image I be defined in the RGB representation system by:

$$\begin{aligned} I: \quad & \Omega \rightarrow N^3 \\ & p \mapsto (R_p, G_p, B_p) \end{aligned} \tag{2}$$

The pseudo-saturation S^* of I is defined by:

$$\begin{aligned} S^*(I): \quad & \Omega \rightarrow N \\ & p \mapsto \max(|R_p - G_p|, |R_p - B_p|, |G_p - B_p|) \end{aligned} \tag{3}$$

The pseudo-saturation is valid on dark pixels (as well as light ones) since its computation does not imply any division (see Figure 5.c).

B. Coarse detection of chromatic zones

Whatever the saturation formula, we cannot get rid of the color noise directly. Figure 6 represents the pseudo-saturation output of a sample color image (blue handwritten text on the top, black printed text below). It shows that it is impossible to find an appropriate threshold that removes the noise and detects chromatic regions at the same time.

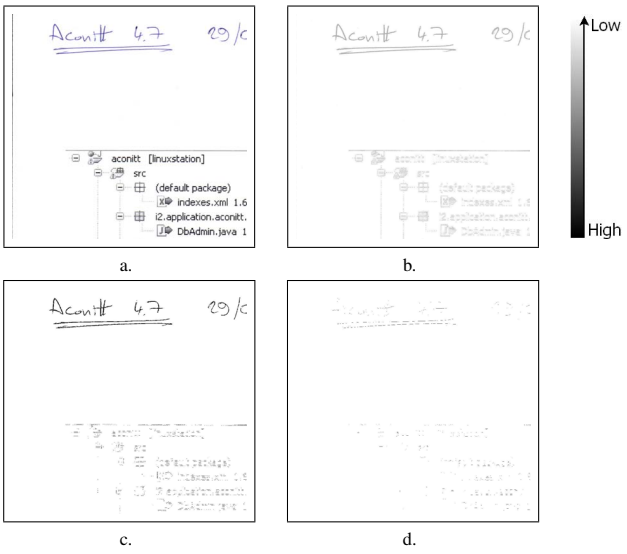

Figure 6. a. I, b. $S^*(I)$, c. $S^*(I)$ thresholded at 25%, d. $S^*(I)$ thresholded at 40%

To get rid of the color noise, we operate a coarse detection of the chromatic areas. Then, we refine the results by localizing those zones accurately. Such a decomposition has a practical purpose: it is possible to stop at the end of the first step if we just want to know whether an image includes chromatic (or achromatic) regions, which is time saving.

The output at this level will be the coarse mask $\mathcal{M}_C$. At this stage, we aim to remove the color noise, even if the localization is not precise.

We begin by reducing the image size using a Gaussian re-sampling. The scale reduction smoothes the image; thus, it eliminates some noise. Furthermore, processing a reduced image is faster than handling the full-sized one. The scale reduction factor is fixed to $\mathcal{S}_t$ (*cf.*section II). Such a value achieves a reduction in the color noise without destroying the small chromatic zones (such as text).

As the saturation noise is located next to black text pixels, we apply a morphological color closing [7] of the darker elements. This replaces the remaining chromatic noise with regular text pixels without removing the truly chromatic zones. Since the image size has already been normalized, the size of the closing structural element can be common to all images and is fixed to a low value, 3×3, to avoid information loss. Furthermore, the saturation noise is connected to the character boundaries [5]. Baird [8] asserts that noise is an inverse exponential curve sharply increasing from the edge of the contour. Therefore, a closing with a structural element of radius 1, *i.e.* a 3×3 matrix, is enough to remove the noise. We call the resulting image I^r.

The pseudo-saturation measure is computed over I^r (see Figure 8.b). The image $S^*(I^r)$ is then thresholded to create the coarse mask $\mathcal{M}_C$. Only the pseudo-saturation values above the threshold correspond to chromatic areas. The pseudo-saturation threshold estimation is based on $S^*(I^r)$'s histogram, mainly its first peak. Entirely chromatic images have very few low values of pseudo-saturation (Figure 7.a). Thus, if the first mode's position is significantly greater than zero the image is judged to be wholly chromatic and the threshold is set to the minimal value. Similarly, if all the histogram's peaks are close to zero, the image is considered to be entirely achromatic and the threshold is set to 100% (Figure 7.b). If the two above conditions are not met, the pseudo-saturation threshold is given by the position of the local minimum following the first peak (Figure 7.c).

The mask $\mathcal{M}_C$ displayed in figure 8.c shows that the color noise is successfully removed. The localization of the chromatic zones will be refined in the next section.

C. Accurate chromatic / achromatic split

Provided $\mathcal{M}_C$, we can now precisely extract the shapes and create $\mathcal{M}_F$. This can only be done by using the full scale image. Therefore, we will combine $\mathcal{M}_C$ (Figure 8.c) with a new mask called $\mathcal{M}_A$ given by thresholding the full-sized pseudo-saturation image (Figure 6.d). The thresholding method is the same as explained in section III-B.

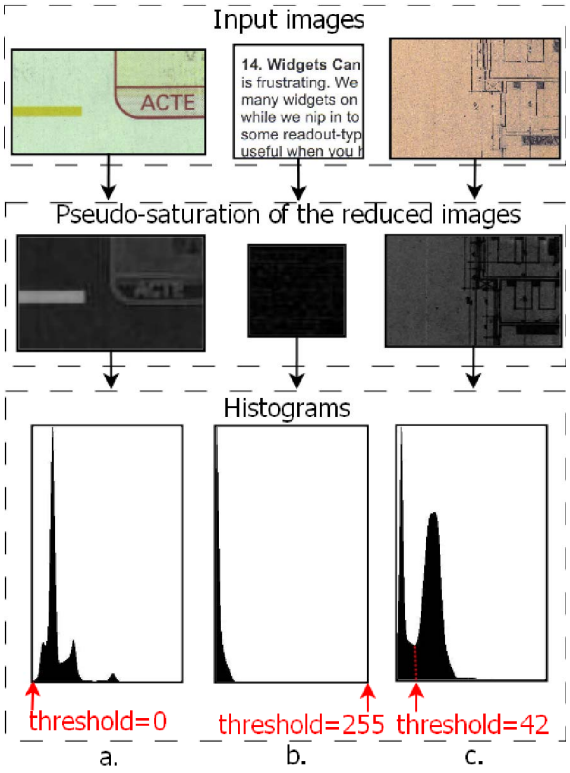

Figure 7. Thresholding of a. a chromatic, b. an achromatic and c. a mixed images

Figure 8. Manuscript blue line atop black printed text. a. I, b. I^r, c. $\mathcal{M}_C$

The most intuitive combination method may be the logical intersection of $\mathcal{M}_C$ and $\mathcal{M}_A$. However, such an operation fails to perform on images with a chromatic background (see Figure 9). To overcome this problem, we propose to compute

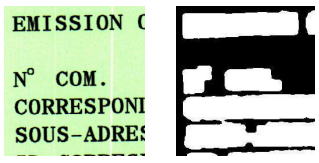

Figure 9. a. Green background image, b. Mask obtained by logical intersection of $\mathcal{M}_C$ and $\mathcal{M}_A$: approximately the same as $\mathcal{M}_C$

the final mask as the logical intersection of $\mathcal{M}_C$'s bounding boxes [9] (in the original scale) and $\mathcal{M}_A$. The red boxes in Figure 10.a represent the bounding boxes extracted from $\mathcal{M}_C$; Figure 10.b displays the mask $\mathcal{M}_A$ and Figure 10.c $\mathcal{M}_F$.

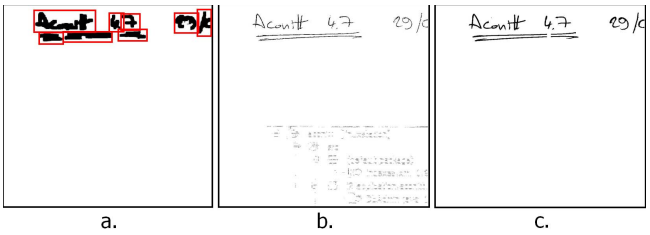

Figure 10. a. $\mathcal{M}_C$, b. $\mathcal{M}_A$, c. $\mathcal{M}_F$

D. Results

The algorithm has been tested on a database composed of a large variety of digitized documents (magazines, newspapers, map images, manuscripts, *etc.*). As we do not have the electronic sources of images (and even then, the registration step would have been problematic) we had to label the 320 images manually. The testing database includes approximately 30% images that are exceedingly noisy (such as the ones in Figures 1 and 11).

The results will be displayed in terms of precision P and recall R. The precision is defined by the intersection area of the ground truth and the segmentation method chromatic zones divided by the area of the segmentation method's zones. The recall is the ratio between the intersection area and the ground truth area.

$$R = \frac{\bigcap area}{ground{-}truth\ area} \qquad P = \frac{\bigcap area}{detected\ area} \tag{4}$$

Our method's results are compared with the ones given by thresholding the saturation (in [2] the threshold is fixed to 20%).

Table I shows that our method reaches an almost perfect precision value. Indeed, it is designed to remove the color noise and thereby the false detections. Both of the presented methods reach good recall values; *i.e.* almost all the chromatic areas are correctly detected.

Table I
P/R RESULTS

Method	R	P
Kim 2009 [2]	93.26	70.03
Ouji 2011	91.54	99.88

Our engine has managed to remove all the color noise in Figure 1. Figure 11 shows the final mask of another noisy image.

The shape extraction is very accurate (see Figure 12.d). Figure 12.b displays the final film of a sample image. The zones that may appear grey in the film (because of the scaling in the current document) correspond to a thin color dithering, which is correctly extracted (see Figure 12.c).

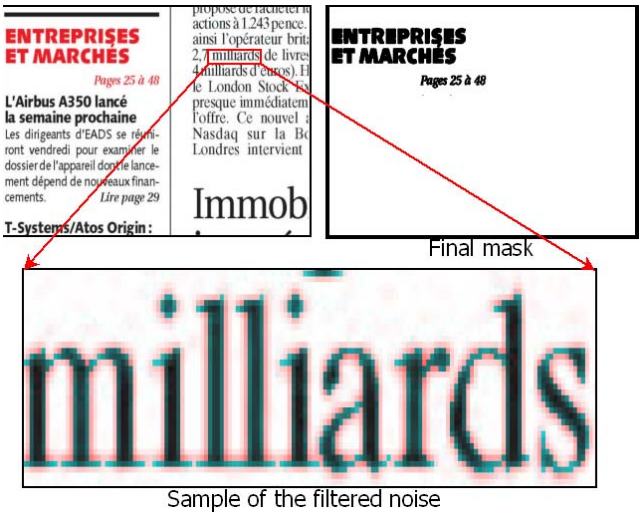

Figure 11. Sample result

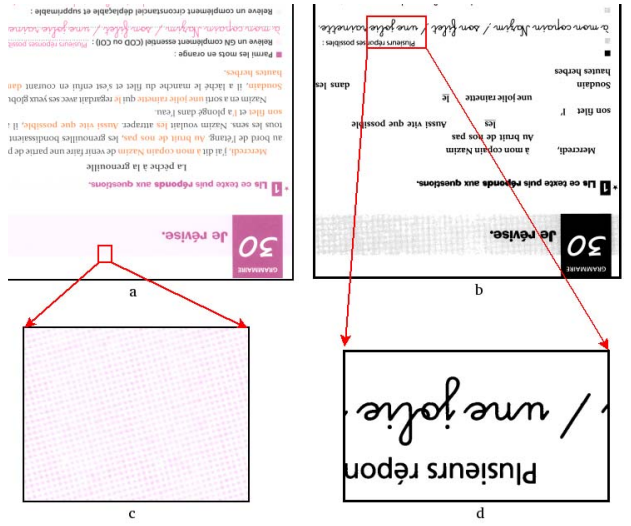

Figure 12. a. Sample image, b. $\mathcal{M}_F$, c. Chromatic dithering, d. Chromatic text.

IV. Conclusion

This paper presents a generic method to remove the color noise so that the chromatic / achromatic split is achieved successfully. The proposed approach is pixel based and independent of the document image's type; it does not require any *a priori* information. It is also parameter-free owing to the statistical measure of the strokes thickness. A new measure of pseudo-saturation has been introduced to detect chromatic areas efficiently. Gaussian scaling and morphological closing have been used to get rid of the color noise.

We held tests on a representative set of images coming from various documents. The proposed method reaches very satisfying results, especially very high precision values. Comparisons with a baseline method confirm our method's effectiveness.

References

[1] A. Ouji, Y. Leydier, and F. LeBourgeois, "Advertisement detection in digitized press images," in *IEEE International Conference on Multimedia & Expo*, 2011.

[2] J. H. Kim, D. K. Shin, and Y. S. Moon, "Color transfer in images based on separation of chromatic and achromatic colors," in *MIRAGE '09*. Berlin, Heidelberg: Springer-Verlag, 2009, pp. 285–296.

[3] R. Gonzalez and R. E. Woods, *Digital Image Processing*. Prentice Hall Press, 2002.

[4] D. Karatzas and A. Antonacopoulos, "Colour text segmentation in web images based on human perception," *Image Vision Comput.*, vol. 25, no. 5, pp. 564–577, 2007.

[5] Y. Li, D. Lopresti, G. Nagy, and A. Tomkins, "Validation of image defect models for optical character recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, pp. 99–108, 1996.

[6] P. Sarkar, G. Nagy, J. Zhou, and D. Lopresti, "Spatial sampling of printed patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, pp. 344–351, 1998.

[7] J. Serra, *Image Analysis and Mathematical Morphology*. Orlando, FL, USA: Academic Press, Inc., 1983.

[8] H. S. Baird, "Document image defect models," in *Document image analysis*, L. O'Gorman and R. Kasturi, Eds. Los Alamitos, CA, USA: IEEE Computer Society Press, 1995, ch. Document image defect models, pp. 315–325.

[9] L. He, Y. Chao, K. Suzuki, and K. Wu, "Fast connected-component labeling," *Pattern Recogn.*, vol. 42, pp. 1977–1987, September 2009.