Color Text Extraction from Camera-based Images
― the Impact of the Choice of the Clustering Distance ―

Céline Mancas-Thillou
Faculté Polytechnique de Mons, Belgium
celine.thillou@tcts.fpms.ac.be

Bernard Gosselin
Faculté Polytechnique de Mons, Belgium
bernard.gosselin@tcts.fpms.ac.be

Abstract

Character recognition has a continuous importance for several years and recently, new challenges appeared with camera-based pictures. This paper deals with text extraction for color natural scenes images. Many papers try to combine several color spaces or to choose the best one for a particular database. We show that the main problem is not in the choice of color spaces for generic text extraction but in the choice of clustering distances to handle all degradations present in this kind of images. Comparative results are given using a public database.

1. Introduction

Optical Character Recognition (OCR) from camera-based pictures is a quite recent research area where the challenge to get high recognition rates is still ongoing. For example, mobile reading systems for blind or visually impaired or mobile translation devices for foreign visitors are potential applications of OCR from camera-based pictures.

Nevertheless, camera-based images induce numerous degradations which are not present in scanner-based ones such as blur, uneven lighting, complex backgrounds… and in this paper, we shall deal with the issue of uneven lighting and complex backgrounds with images from the ICDAR 2003 database [1] as shown in Figure 1. Text extraction is the first step in the whole system from image to recognized characters where information can be lost in an irreversible way and incorrect extractions are today the main reasons of OCR performance deterioration. Our main challenge is to find solutions to handle various kinds of camera-based text areas in order to be robust enough and to improve one of the main current challenges in Document Image Analysis, as described by Baird [2]: "*One promising strategy for improving the performance of image understanding systems […] is, we believe, to aim for versatility first*."

First of all, this paper will describe previous work in color segmentation and text extraction. Section 3 will deal with the description of color spaces followed by the one of clustering distances. Section 5 will describe our text extraction system. Finally, in Section 6, results based on Precision and Recall will be detailed to assess the choice of several clustering distances before concluding and describing our future work.

Figure 1: Samples of our database.

2. Previous Work

Color text extraction is one of the fields of color segmentation. Nevertheless, it is needed not to make confusions between color quantization, color spotting, and color object segmentation such as text. For the first one, similarity measures between colors are done to group them and to reduce the number of colors. Objectives are quite different because no particular object is tracked. In another way, color spotting deals with the accuracy with which objects of a specific single color can be identified in a complex image. Hence it is convenient to work with phenomenal color spaces to match with human vision. For text extraction, no assumption is done on colors because the main task is to extract it independently from a particular color. Abadpour and Kasaei [3] use a PCA-based fast segmentation method for color images with the aim of spotting colors. Hence it is unsuitable for text extraction with strong uneven lighting presenting noisy text with colors changing gradually. Shin et al. [4] use ***RGB*** color space for skin detection and prove that for

this particular object no better color spaces than ***RGB*** can be found. Several papers deal with tracking in video sequences under varying illumination such as Stern and Efros [5], who work with several color spaces in a dynamic way based on properties of frames. Our main advantage compared to these methods is that characters are already detected in our images and is also the main information of the document.

Usually, for color thresholding text images, most of the papers convert the ***RGB*** image into a grey-level one and apply different algorithms either global or local. Those algorithms have proven their efficiency but they are not robust enough to handle any complex backgrounds and color can be used to get more information.

Wang et al. [6] try to combine both color and texture information to improve results. This technique works well for images similar to our database but computation time required is very high. Results are given under visual judgment. Garcia and Apostolidis [7] use a character enhancement based on several frames of video information and a K-means clustering. He obtains best non-quantified results with hue-saturation-value (***HSV***) color space. Our results based on a public database disagree with this last statement. Leydier et al. [8] use serialized unsupervised classification for color ancient documents with heavy defects and transparency. Two color spaces ***RGB*** and ***HSL*** are simultaneously chosen to handle several degradations but serialization needs parameters and manual cluster initialization which is against versatility. For Du et al. [9], color video frames are composed of three channels and an entropy-based thresholding is applied on each channel. Based on a between-class/within-class variance criterion, the three sub-images are merged to constitute a binary image. Results seem attractive but in our application, this algorithm performs poorly as explained in [10].

In most papers using color for text extraction, several color spaces based on experimental results are used. We shall show more the importance of the clustering distance than the color space. Our goal is to discriminate images to choose the best clustering distance.

3. Description of Color Spaces

Usually, for color segmentation, several color spaces are chosen depending on the application. Interestingly, no papers explained how to choose them for a given purpose, hence Abadpour and Kasaei [3] decided to build an own color space per image.

Color spaces are usually described in several categories: device-dependent/device-independent, uniform/non uniform, phenomenal/non phenomenal. A dynamic color space could be obtained by principal components analysis which decorrelates axis and is more perceptually uniform. However, Ruderman et al. [11] have shown that for natural image ensembles, the resulting axes have simple forms and interpretation, forming a new color space. The first principal component is achromatic while the second and third ones are yellow-blue and red-green color opponent pairs. This new color space is called $O_1O_2O_3$.

In this paper, we will detail all tests done with various color spaces, normalized or not and present our results for text extraction into natural scenes. The conversion between device-dependent color spaces such as ***RGB*** into device-independent ones such as ***XYZ*** assumes the CIE illuminant D50 as the white point, which is the default illuminant in the International Color Consortium.

First of all, it is suitable to define color spaces we tested in Table 1 to compare accurately results as several definitions could be found in the literature. In our experiments, some of them were reduced to two dimensions to remove brightness from the scene and some of them were combined to build a larger feature vector up to 5 dimensions in our experiments.

Table 1: Definitions of color spaces.

HSV	V=(R+G+B)/3, S=1-min(R,G,B)/V, H=arctan'($3^{1/2}$(G-B),2R-G-B)
XYZ	X=0.607R+0.174G+0.2B,Y=0.3R+0.587G+0.114B Z=0.066G+1.116B
Lab	L=25(100Y/Yo)$^{1/3}$, a=500[(X/Xo)$^{1/3}$-(Y/Yo)$^{1/3}$], b=200[(Y/Yo) $^{1/3}$-(Z/Zo)$^{1/3}$], [Xo, Yo, Zo] = white point
Lch	c=$(a^2+b^2)^{1/2}$,h=(arctan(b/a)+kπ/2)/(2π) if a$\neq$0 else h=0 k=0 if a>0 and b>0, k=1 if a>0 and b<0, k=2 if a<0 and b<0, k=3 if a<0 and b>0
$I_1I_2I_3$	I_1=(R+G+B)/3,I_2=(R-B)/2,I_3=(2G-R-B)/4
$O_1O_2O_3$	O_1=0.28X+0.72Y-0.107Z, O_2=-0.449X+0.29Y-0.077Z O_3=0.086X+0.59Y-0.501Z
YCbCr	Y=0.3R+0.587G+0.114B, Cb=-0.169R-0.331G+0.5B+128 Cr=0.5R-0.419G-0.081B+128

4. Description of Clustering Distances

Several clustering distances could also be defined depending on applications. Usually for color object segmentation, the Euclidean distance is chosen, as defined in Equation 1. Nevertheless, in the literature [12], other clustering distances are detailed such as Manhattan, Chebychev, Minkowski, Mahalanobis but also Canberra, Correlation or Cosine.

Moreover, it is useful to make a difference between clustering distances and similarity measures. Similarity means how two objects are similar and this measure

falls into the range [0,1] while distance computes a length and falls into the range [0,∞]. We tested all distances cited above but the shown results concern only two of them which were more relevant. In Section 6, we will detail results about computing the Euclidean distance (ED) and the Cosine similarity (CS) defined for two colors x and y and with dim, the dimension of the color space (equal to 3 for ***RGB***) as:

$$ED(x,y) = \sqrt{\sum_{i=1}^{dim}(x_i - y_i)^2} \quad , CS(x,y) = 1 - \frac{\sum_{i=1}^{dim} x_i y_i}{\sqrt{\sum_{i=1}^{dim} x_i^2 \times \sum_{i=1}^{dim} y_i^2}} \quad (1)$$

In Figure 2, an interpretation of CS is given to show how well it can handle uneven lighting and gradual noise in natural scenes for text extraction. When text is unevenly illuminated, the lighting spreads colors of a character on partial zones or on several letters. With ED, colors are considered as changing and can create a new cluster or be included in another one and degrade smoothness of characters. CS can handle this kind of situations as shown in the results. According to [13], ED and CS are very complementary as ED highlights intensity differences while CS enhances hue and saturation difference information. Hence, the advantages of this combination is fast computation as no complex transformation between different color spaces is required but also the ability to perform text extraction from the ***RGB*** image independently of highly colored or not text areas.

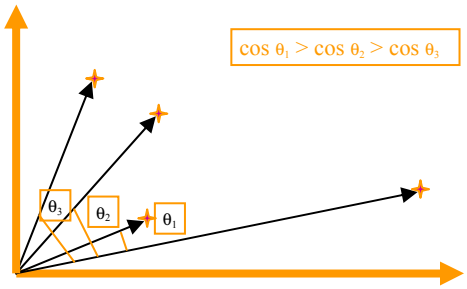

Figure 2: Interpretation of the Cosine similarity.

5. Our Text Extraction Method

5.1 Coarse pre-processing

An important problem for "real-world" pictures comes from a non-uniform illumination which introduces noise. This uneven lighting appears as wide noisy areas, so it is assumed to have a lower frequency spectrum than the one of characters. Based on a wavelet decomposition described in [14], the denoising is done with respect to no more degradation added. Considering properties of human vision, there is a large amount of redundancy in the 24-bit ***RGB*** representation of color images. We decided to represent each of the ***RGB*** channel with only 4 bits, which introduce very few perceptible visual degradation. Hence the dimensionality of the color space is 16*16*16 and it represents the maximum number of colors.

5.2 Color clustering

We use the K-means clustering with a fixed number of clusters equal to 3. The three dominant colors are extracted based on the color map of the picture and group into clusters iteratively updated by the K-means algorithm. Finally, each pixel in the image receives the value of the mean color vector of the cluster it has been assigned to. Three clusters are sufficient for the ICDAR 2003 database, which is large enough to be applicable on other camera-based images, when text areas are already detected.

5.3 Choice of a clustering distance

The choice of the clustering distance is computed based on a global histogram of the reduced color map for each dimension. In order to remain generic, no parameter is used but the general shape of these histograms is computed.

For example, the difference between two images, one handled with ED and one with CS, can be seen on the plotted shape of the color map. Actually, if the image is quite clean with strong differences between foreground and background, the histogram will be less noisy. On the contrary, for a complex background and an unevenly illuminated textual foreground, it will be noisy with several peaks. Features as the maximum of each peak, the spread of the histogram, the extrema for each dimension and the confusion between dimensions are provided to a linear discriminant analysis to take the right decision.

5.4 Eventual combination

The background color is selected very easily and efficiently as being the color with the biggest rate of occurrences on the image borders. Only two pictures left which correspond depending on the initial image to either two foreground pictures or one foreground picture and one noise picture. Here, the most probable useful picture is defined with a means of skeletization.

As the first thresholding corresponds in an approximate way to characters, a skeletization is used to get the color of centers of characters as in [10]. The Euclidean distance with both mean color pixel of the cluster and mean color of the skeleton is performed. The cluster with the smallest distance from the skeleton is considered as the cluster with the main textual information. The combination to be done is decided according to the distance between mean color values of the two remaining clusters. If distance is inferior to 0.5, colors are considered as similar and the second picture seems to be a foreground picture too.

On the database, this decision is valuable to 98.4% and no false alarm is detected. For the 1.6% remaining, some useful information is lost but the recognition is still possible as the first selected picture is the most relevant foreground one.

Figure 3: From left to right: initial image, clustering with ED and then with CS.

Figure 3 shows for the first row that best results are given with ED, for the second row with CS or for the last row, both give satisfying results.

6. Results and Discussion

For our experiments, we use the public database from the Robust Reading Competition of Words Recognition ICDAR 2003 [1], which is constituted by a total of 2266 words. As our algorithm of text extraction concerns the issue of uneven lighting and complex backgrounds but not blur and low resolution, we remove samples like in Figure 4, where text extraction needs other steps such as local gradient analysis and robust denoising. Hence, our database concerns 2073 words on the whole ICDAR 2003 database.

Figure 4: Samples of removed images from the database for results.

Results are presented in respect to Precision and Recall. These standard measures compare the performance of different algorithms and are defined as:

$$\text{Precision} = \frac{Correctly\ Detected\ Characters}{Totally\ Detected\ Characters} \quad (2)$$

$$\text{Recall} = \frac{Correctly\ Detected\ Characters}{Total\ Characters} \quad (3)$$

Correctly Detected Characters mean characters which are extracted without noise and without missing character parts. No ground truth is available.

In Table 2, results for each of the eleven tested color spaces, defined partly in Section 3, are computed with ED. A refined analysis is presented in Table 3 to show that not only one clustering distance has to be chosen but two with a dynamic choice to handle all kinds of situations.

Table 2: Results for all tested color spaces.

Color space	Precision	Recall
RGB	0.91	0.89
XYZ	0.89	0.59
XZ	0.42	0.10
HSV	0.69	0.6
HS	0.49	0.17
Lab	0.76	0.24
ab	0.87	0.42
RGBab	0.87	0.84
RGBch	0.86	0.83
Lch	0.84	0.26
ch	0.90	0.87
abch	0.66	0.48
$I_1I_2I_3$	0.67	0.58
$O_1O_2O_3$	0.66	0.32
YCbCr	0.76	0.34

Table 3: Results for the RGB color space and the impact of clustering distances.

	ED	CS	Both chosen manually
Precision	0.91	0.94	0.97
Recall	0.89	0.39	0.95

The ***RGB*** and the ***ch*** color space are presented as the ones with the best results. Nevertheless, tests with both clustering distances (ED and CS) show more relevant results. We can see that almost all images of this database can be handled with both clustering distances using the ***RGB*** color space. The ***ch*** color space brought no improvements when using several clustering distances. Nevertheless it is already polar coordinates of the ***ab*** ones: therefore CS is not so relevant. The use of the combination of both color spaces ***RGB*** and ***ch*** enables to consider 2.3% more

images while the use of both clustering distances enables 6.7% more. This computation brings the efficiency of the combination between these two clustering distances as more characters in more images could be extracted. Hence, the conclusion is that no transformation of color space is needed and the use of two different clustering distances is enough to handle various complex camera-based images.

Results can seem surprising compared to all studies done on color spaces but our results match very accurately with the ones of Shin et al. [4] and Wesolkowski and Jernigan [13] where very detailed results are also described. For text extraction, the problem is different than the one of colors spotting for example and the analysis done by computers on images taken by perfect sensors could not be compared with the same techniques as the ones wishing to reproduce the human visual system.

For this database, our algorithm for the choice of the color clustering distance works in 90.3% of cases with very simple and fast features. A refinement will be needed and is currently under progress. This rate is computed by considering good a choice when all characters could be extracted. For clean images with uniform backgrounds, sometimes both distances could be considered as good as shown in the third row in Figure 3; hence no error is counted.

7. Conclusion and Future Work

In this paper, we have presented a new text extraction method for natural scene camera-based pictures, showing the relevance of the use of two clustering distances in the original color space instead of choosing different color spaces experimentally.

Moreover, with this study, as no transformation of color spaces needs to be applied, computational time is drastically reduced and no dynamic choice between several color spaces is useful reducing also much the processing time.

The main problem associated with this method is that a special algorithm will be needed to deal with achromatic image regions or homogeneous chromatic regions with several characters like in Figure 4. Nevertheless as human visual system does for this kind of images, a local analysis of gradients is needed with a very high-level processing such as statistics and linguistic information to reach high recognition rates.

Our main future work is to refine the choice of color clustering distances with other complementary algorithms such as the analysis of the 3D graph of the ***RGB*** color space to highlight other differences. Another extension could be brought to enable character extraction without preliminary text detection; hence the number of clusters should be chosen dynamically.

Acknowledgements

This work is part of the project Sypole and is funded by Ministère de la Region wallonne in Belgium.

References

[1] Robust Reading Competition: Retrieved May 31, 2005 from http://algoval.essex.ac.uk/icdar/RobustWord.html.

[2] H.Baird, D.Lopresti, B.Davison, and W.Pottenger, "Robust document image understanding technologies", *Proc. of ACM HDP Workshop*, USA, 2004, pp. 9-14.

[3] A.Abadpour and S.Kasaei, "A new parametric linear adaptive color space and its implementation", *Proc. Of Annual CSICC*, Iran, 2004, pp. 125-132.

[4] M.Shin, K.Chang, and L.Tsap, "Does color space transformation make any difference on skin detection", *IEEE Workshop on ACV*, USA, 2002, pp. 275-279.

[5] H.Stern and B.Efros, "Adaptive color space switching for tracking under varying illumination", *Image and Vision Computing*, vol.23, n.3, 2005, pp. 353-364.

[6] B.Wang, X-F.Li, F.Liu, and F-Q.Hu, "Color text image binarization based on binary texture analysis", *Proc. of ICASSP*, 2004, pp. 585-588.

[7] C.Garcia and X.Apostolidis, "Text detection and segmentation in complex color images", *Proc. of ICASSP*, vol.4, 2000, pp. 2326-2330.

[8] Y.Leydier, F.Le Bourgeois, and H.Emptoz, "Serialized unsupervised classifier for adaptive color image segmentation: application to digitized ancient manuscripts", *Proc. of ICPR*, 2004, pp. 494-497.

[9] Y.Du, C.Chang, and P.Thouin, "Unsupervised approach to color video thresholding", *Proc. of SPIE Optical Imaging*, vol.43, n.2, 2004, pp. 282-289.

[10] C. Thillou and B. Gosselin, "Combination of binarization and character segmentation using color information", *Proc. of IEEE ISSPIT*, 2004, pp. 107-110.

[11] D.Ruderman, T.Cronin, and C.Chiao, "Statistics of cone responses to natural images: implications for visual coding", *Journal Opt. Soc. Am. A*, vol.15, n.8, 1998, pp. 2036-2045.

[12] G.Sharma, "Digital Color Imaging", *by CRC Press*, 2003.

[13] S. Wesolkowski and E.Jernigan, "Color edge detection in rgb using jointly euclidean distance and vector angle", Vision Interface, Canada, 1999, pp. 9-16.

[14] C. Thillou and B. Gosselin, "Robust thresholding based on wavelets and thinning algorithms for degraded camera images", *Proc. of ACIVS*, Belgium, 2004, pp. 231-235.