COMBINATION OF BINARIZATION AND CHARACTER SEGMENTATION USING COLOR INFORMATION

Céline Thillou and Bernard Gosselin

{celine.thillou,bernard.gosselin}@tcts.fpms.ac.be
Faculté Polytechnique de Mons, Avenue Copernic, 7000 Mons, Belgium
Tel: +32 65 37 47 17, Fax: +32 65 37 47 29

ABSTRACT

Character segmentation and recognition have been performed for several decades, especially typewritten characters from scanner. Commercial OCR softwares perform well on "clean" documents or need user to select the kind of documents. Recently, a new kind of images taken by a camera in a "real-world" environment appeared. It implies different strong degradations missing in scanner-based pictures and the presence of complex backgrounds. In order to segment text as properly as possible, a new method is proposed using color information in order to extract text as well as possible. In this paper, a focus is given on each chosen parameter with comparative results between different recent techniques using color information. Moreover an emphasis is placed on stroke analysis and character segmentation. The binarization method takes it into account in order to improve character segmentation and recognition afterwards.

Keywords : Binarization, Color Clustering, Wavelet, Character Segmentation, K-means

1. INTRODUCTION

Based on recent evolutions of technologies integrating digital cameras and powerful data processing in personal digital assistants, a new kind of images appeared with different degradations. As new needs were discovered with this imaging device such as help for the blind or visually impaired, it is important to understand these constraints in order to correct them as properly as possible.

Therefore this context implies a bunch of degradations, not present in classical scanner-based pictures, such as blur, perspective distortion, complex backgrounds, uneven lighting...In order to recognize characters in an entire OCR processing, words need to be binarized, segmented into characters before recognition.

First of all, text detection is performed on the whole image but is not dealt with this paper. Nevertheless, this point needs to be noticed because as text is already detected in a more or less closing bounding box, text becomes the more relevant information in the new image.

Thresholding, as the first step of OCR, is crucial and its success is preponderant for all other processings. Actually this is the first step where some information is lost after picture acquisition. Errors at this point are propagated all along the recognition system. The challenge to obtain a very robust binarization method is major.

This paper is organized as followed. Section 2 describes the state of the art of binarization techniques, in general and applied in camera or video-based images. Section 3 describes our binarization approach with all steps and a discussion is given in Section 4 to explain results of all compared methods. Finally, we address a conclusion and future work about this algorithm.

2. A STATE OF THE ART

Most existing binarization techniques are thresholding related and categorize into 6 groups according to [12]. These categories are histogram-based, clustering-based, entropy-based, object attribute-based, spatiality-based and locality-based methods. But every preliminary tests are always done between the two main groups : global (histogram-based) [6] and local or adaptive [5]. Global methods tempt to binarize the image with a single threshold. Among some most powerful global techniques, Otsu [8]'s algorithm can achieve high performance with simple backgrounds and without parameters to tune. By contrast, local methods change the threshold dynamically over the image according to local information. Meanwhile in our context, image processing systems need to process a large number of documents with different styles and without pre-specified parameters, which can be a failure for local methods such as the Sauvola [10] one.

Several tests were done according to Precision and Recall measures for these first steps after a denoising part and they are given in Section 4.

In [7], Liu and Srihari used Otsu's algorithm to obtain candidate thresholds. Then, texture features were measured from each thresholded image, based on which the best threshold was picked. Color information is not used and this technique fails when different colors with almost the same intensity are present. Seeger [11] created a new threshold technique for camera images, like in our context, by

computing a surface of background intensities and by performing adaptive thresholding for simple backgrounds.
For very complex color images, those methods are not sufficient and color information could be used to get more clues. In order for people to read text, in complex color images, color information is really significant, more than contrast between gray-level values.
Wang [16] tried 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 and no consideration on connectivity between components are presented and results are given under visual judgement. With other techniques, and some similar ones, our method fills these failures. Garcia [3] uses a character enhancement based on several frames of video information and uses a kmeans clustering as our method to binarize text information. His method is based on four clusters and combination of clusters to get as much text as possible is done on a bunch of criteria concerning characters properties. For him also, text areas are already detected. On the contrary, he does not take into account stroke analysis and results are worse for character segmentation. Moreover he obtained best non-quantified results with hue-saturation-value (HSV) color space. Our results based on a public database (Samples of words of robust reading competition ICDAR 2003 [9]) do not represent the same results.
The last algorithm we want to discuss in this section is the Du's one [1]. Color images are composed of three channels (red, green, blue) and entropy-based thresholding are applied on each gray-level channel. Based on a between-class/within-class variance criterion, the three subimages are merged to constitute a binary image. Results seem attractive but text areas are not already detected. As in our case, text information is the main one in the image, this algorithm does not give the same results. Experiments were done and are given in Section 4. Before explaining all these results, we will describe our binarization approach in the next section.

3. OUR BINARIZATION APPROACH

A scheme of our proposed system is presented in Figure 1. Color information is only used after gray-scale denoising and coarse thresholding in order to consider only useful parts and to decrease the required time for color clustering with less pixels. Then a combination of results is either applied or not, according to a parameter of distance and this eventual combination is partial or total in order to take into account non-connectivity of characters.
All these different steps are detailed in the following sections : a coarse pre-processing to remove useless parts and to reduce the number of colors. Then color clustering is used to refine the initial thresholding, followed eventually by a partial or total combination of different clusters.

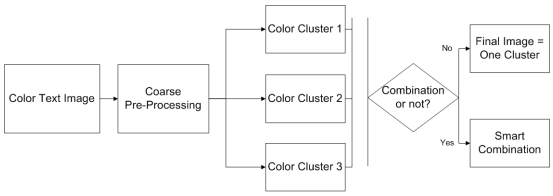

FIG. 1 – An overview of our binarization method

3.1. Coarse Pre-Processing

An important problem for thresholding methods and especially for "real-world" pictures comes from a non-uniform illumination which introduces noise. This uneven illumination appears as wide noisy areas, so the illumination noise 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.
In order to keep advantage of this denoising part, a zonal mask is computed thanks to a coarse thresholding and applied on the initial color image. This step is useful to consider only useful parts in the image and to decrease computation time [15].
Considering properties of human vision, there is a large amount of redundancy in the 24-bit RGB representation of color images. As in [16], we decided to represent each of the RGB channel with only 4 bits, what introduce few or no perceptible visual degradation. Therefore the dimensionality of the color space is 16*16*16 and it represents the maximum number of colors. But as all pixels are not different colors, the right number is much less.

3.2. Color Clustering

In [16], color clustering is done using Graph Theoretical clustering without giving the number of clusters because the picture was not pre-processed. Actually, preprocessing with an approximate thresholding does not lose any textual information in our database.
Thanks to this first step, we use the well-known K-means clustering with K=3. Nevertheless, this decision comes from a preliminary test to be sure the number of clusters is right and if it can be fix or dynamic as in [16]. In Figure 2, we plotted the overall mean error for our database for different values of K (from 3 to 5) to estimate the best K for K-means. Considering more than 5 clusters will be damageable for the computation time and decisions to take for eventual further combinations for clusters and considering less than 3 clusters does not enable to remove noise in a complex image.

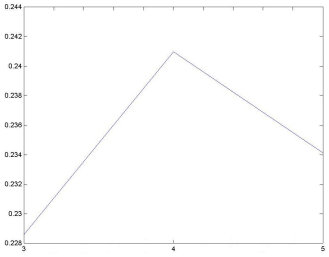

FIG. 2 – Graph of the mean overall error for 3 different values of K for our database : the best K for K-means algorithm is 3

FIG. 3 – Two different foreground clusters : in the first sample, foreground image and noise, in the second one, two foreground images

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 enough thanks to our pre-processing with the mask for our database.

3.3. Eventual Partial or Total Combination

The background color is selected very easily and efficiently (100 % in the ICDAR 2003 database) as being the color with the biggest rate of occurrences in the image edges. Only two pictures left which correspond depending on the initial image to either two foreground pictures or one foreground picture and one noise picture as shown in Figure 3.

In [16], combination is based on some texture features to remove inconvenient pictures and on a linear discriminant analysis with other criteria. Here, the most probable useful picture is defined with a means of skeletization. Actually, as the first thresholding corresponds in an approximative way to characters, a skeletization is used to get the color of centers of characters as in [14].

Euclidean distance D 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.

Combination to do is decided according to the distance D between mean color values of the two remaining clusters. If distance is inferior to 0.5, color are considered as similar and the second picture seems to be a foreground picture too. On the ICDAR 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.

Connected components on the first foreground picture are computed to get coordinates of their bounding box in order not to connect components with pixels to add in the combination. Only pixels which can be added will change the first foreground picture. On the contrary, some characters can be broken if they were broken in the first foreground picture. But, in this case, the correction will be facilitated by the fact that characters parts will be closer.

4. EXPERIMENTAL RESULTS AND DISCUSSION

The standard measures, Precision and Recall [4], were used to compare the performance of different methods and were defined on characters as :

$$\text{Precision} = \frac{\text{Correctly Detected Characters}}{\text{Totally Detected Characters}}$$

$$\text{Recall} = \frac{\text{Correctly Detected Characters}}{\text{Totally Characters}}$$

Our database is issued from the Robust Reading Competition Sample Words ICDAR 2003 [9]. We computed results for different hybrid algorithms : the first one (1) is the gray-level denoising followed by the global Otsu [8] thresholding, then (2) is the gray-level denoising followed by the adaptive Sauvola [10] algorithm with a window of 15 pixels. For algorithms with parameters to tune, we took the best parameter for the whole database after several tests. This is also to test the automatic way of an algorithm. With these first computations, the global method gives best results. It is mainly due to the fact that text is already detected and characters are the main information in the image.

After having decided that color information could be useful, we tested the Du's algorithm (3) thanks to the code published on line [2]. We give results with the GRE (Global Relative Entropy) color thresholding which was the algorithm which worked best for our database. Results are

quite low and it can be explained by the fact that it is normally applied on a whole image without text detection.
Then we compared two algorithms : (4) with denoising, our binarization approach on RGB color space and (5) on HSV color space. Best results are given by the algorithm (4) on RGB color space over all algorithms tested here. Usually, based on visual clues, HSV color space contains more relevant information but in the process of K-means algorithm with the Euclidean distance, the RGB color map gives more pertinent results than the HSV color map.
An important point to note is that not only the result of ratios have meaning in these tests but also the ratios themselves. That means that different and more or less characters are each time considered and it could be interested to combine different algorithms to take advantage of each of them. All results are described in Table 1.

TAB. 1 – Comparison of several algorithms : the description of the numbers of algorithms are given in Section 4

	Precision	Recall
(1)	$440/558 = 0.789$	$440/805 = 0.547$
(2)	$345/487 = 0.708$	$345/805 = 0.429$
(3)	$362/463 = 0.781$	$362/805 = 0.450$
(4)	$451/586 = 0.770$	$451/805 = 0.560$
(5)	$390/519 = 0.751$	$390/805 = 0.484$

5. CONCLUSION AND FUTURE WORK

In this paper, we have presented a new binarization method for "real-world" camera-based pictures. Color information is not used from the beginning in order to reduce computation time and to use the color information at a more convenient step.
Moreover a smart combination is done between clusters to get as much information as possible with a compromise with the number of connected components in order to improve character segmentation and recognition.
Numerous experiments were done to justify each parameter and each choice taken in this algorithm. Moreover comparisons were done with other recent techniques using color information.
A way to discriminate backgrounds between clean and noisy ones is currently under investigation to further decrease the computation time and to get smoother results in the case of clean backgrounds.

6. ACKNOWLEDGEMENTS

This work is part of the Sypole project [13] and is funded by Ministère de la Région wallonne in Belgium.

7. REFERENCES

[1] Y.Du, C.Chang, P.Thouin : Unsupervised approach to color video thresholding, Proceedings of SPIE Optical Imaging, vol.43, n.2, (2004) 282–289

[2] Du homepage : Retrieved October 20, 2004 from http ://userpages.umbc.edu/ ydu1/

[3] C.Garcia and X.Apostolidis : Text detection and segmentation in complex color images, Proceedings of ICASSP 2000, vol.4, (2000) 2326–2330

[4] M.Junker and R.Hoch : On the eEvaluation of document analysis components by recall, precision, and accuracy, Proceedings of ICDAR 1999, (1999) 713–716

[5] J.Kittler, J.Illingworth : Threshold selection based on a simple image statistic, CVGIP, n.30, (1985) 125–147

[6] G.Leedham, C.Yan, K.Takru, J.Tan and L.Mian : Comparison of some thresholding algorithms for text/background segmentation in difficult document images, Proceedings of ICDAR 2003, (2003)

[7] Y.Liu and S.N.Srihari : Document image binarization based on texture features, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.19, n.5, (1997) 540–544

[8] N.Otsu : A thresholding selection method from gray-level histogram, IEEE Transactions on Systems, Man, and Cybernetics, n.9 (1979) 62–66

[9] Robust Reading Competition Database : Retrieved October 20, 2004 from http ://algoval.essex.ac.uk/icdar/RobustWord.html

[10] J.Sauvola and M.Pietikinen, Adaptive document image binarization, Pattern Recognition, vol.33, (2000) 225–236

[11] M.Seeger and C.Dance : Binarising camera images for OCR, ICDAR 2001, (2001) 54–59

[12] M.Sezgin and B.Sankur : Survey over image thresholding techniques and quantitative performance evaluation, Journal of Electronic Imaging, vol.13, n.1, (2004) 146–165

[13] Sypole project : Retrieved October 20, 2004 from http ://tcts.fpms.ac.be/projects/sypole/sypole.html

[14] C.Thillou and B.Gosselin : Robust thresholding based on wavelets and thinning algorithms for degraded camera images, Proceedings of ACIVS 2004, (2004)

[15] C.Thillou and B.Gosselin : Segmentation-based binarization for color degraded images, Proceedings of ICCVG 2004, (2004)

[16] B.Wang, X-F.Li, F.Liu and F-Q.Hu : Color text image binarization based on binary texture analysis, Proceedings of ICASSP 2004, (2004) 585–588