SPATIAL AND COLOR SPACES COMBINATION FOR NATURAL SCENE TEXT EXTRACTION

Céline Mancas-Thillou and Bernard Gosselin

Faculté Polytechnique de Mons
Avenue Copernic 1, 7000 Mons, Belgium

ABSTRACT

Natural scene images brought new challenges for a few years and one of them is text understanding over images or videos. Text extraction which consists to segment textual foreground from the background succeeds using color information. Faced to the large diversity of text information in daily life and artistic ways of display, we are convinced that this only information is no more enough and we present a color segmentation algorithm using spatial information. Moreover, a new method is proposed in this paper to handle uneven lighting, blur and complex backgrounds which are inherent degradations to natural scene images. To merge text pixels together, complementary clustering distances are used to support simultaneously clear and well-contrasted images with complex and degraded images. Tests on a public database show finally efficiency of the whole proposed method.

1. INTRODUCTION

Visual information grows everyday and meets the increasing popularity of camera-based devices. Automatic processing is expected to handle this encounter, hence natural images brought new challenges to image processing with respect to color segmentation for compression and storage, text understanding for archival purpose, blind users or translation in foreign countries. Color segmentation could then be useful for color perception but also for object-driven segmentation such as text or faces. Until now, techniques used in both cases are quite similar handling uneven lighting to merge parts of a sky in a landscape picture for example. Nevertheless, small segmentation errors in color perception are not obstructing for the scene understanding but for object-driven segmentation, it could mislead to erroneous recognition, which is damageable for text for example. It is therefore of great importance to combine color and spatial information to correct and refine some segmentation errors.

After introducing some relative works in Section 2, we will present an overview of the whole proposed method in Section 3. Sections 4 and 5 will focus respectively on the use of color and spatial information. Finally, efficiency rates will be mentioned in Section 6 before concluding and presenting our future works in the last section.

2. STATE-OF-THE-ART

Several papers deal with color segmentation by using particular or hybrid color spaces as Abadpour and Kasei [1] who use a PCA-based fast segmentation method for color spotting. Garcia and Apostolidis [2] use a character enhancement based on several frames of video and a K-means clustering. They obtained best non-quantified results with hue-saturation-value color space.

Du et al. [3] use an entropy-based thresholding on each RGB color channel for several video frames and based on the silhouette criterion experienced in Section 6, the three subimages are merged to constitute a binary image. Results seem attractive but on natural scene images, the algorithm performs poorly.

Some papers exploit the combination of color and spatial information. Hase et al. [4] extract character strings from color documents by using quantification in La*b* histogram to get several binary candidates, then a multi-stage relaxation with a likelihood of a character string based on features of densities lines and spaces are performed to select character components. Chen et al. [5] merge text pixels together using a model-based clustering solved thanks to the expectation-maximization algorithm. In order to add spatial information, they use Markov random field (MRF) in video sequences. Tests have been processed on different kinds of images and results are satisfying but MRF is really computationally demanding.

In the next sections, we propose a method combining color and spatial information to extract text from various images with relatively fast algorithms.

3. OVERVIEW OF THE PROPOSED METHOD

For color segmentation, several color spaces are mainly used such as described in the previous section to circumvent lighting effects and so on. In this paper, we propose to use a single

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

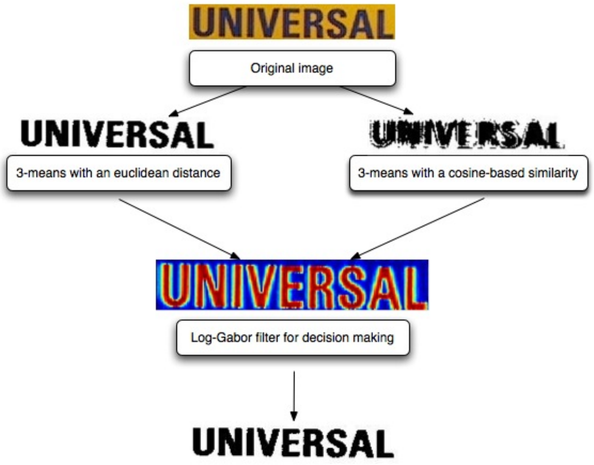
UNIVERSAL
Original image
UNIVERSAL
3-means with an euclidean distance
UNIVERSAL
3-means with a cosine-based similarity
UNIVERSAL
Log-Gabor filter for decision making
UNIVERSAL

Fig. 1. Overview of the proposed algorithm combining color and spatial information.

color space, the RGB one, but two color clustering distances to merge colors with their metamers, colors which are different due to some degradations (reflecting materials, uneven lighting,...). For non-degraded images, the Euclidean distance D_{eucl} has proven its efficiency over all kinds of clustering distances. Based on recent studies [6] and color properties and behaviours on different materials [7], reflecting or not, the matching of different colors could be handled using a cosine-based similarity, S_{cos}. Hence, we use both of them to handle a large diversity of images and to build an algorithm as versatile as possible. Following this first step, a particular segmentation has to be chosen and we use spatial information to take the right decision. A smart way to combine color or gray-level variation with spatial information is to use Gabor-based filter. For this purpose and in the context of natural scene images, we have chosen to use Log-Gabor filters as explained in Figure 1 to get our final text cluster, which will be then fed into an OCR algorithm.

4. USE OF COLOR INFORMATION

In order to segment similar colors together, we use an unsupervised segmentation algorithm with a fixed number of clusters. In this paper, the focus is done on how natural scene text can be extracted to increase recognition results; we consider here only already detected text areas. As areas are constrained, we use a 3-means clustering. The identification of clusters are a textual foreground, a background and a noisy cluster which consists either in noise in complex images or in edges of characters, which are always slightly different, in clear images.

The 3-means algorithm is performed for both clustering distances, D_{eucl} and S_{cos} described in the subsection 4.1. The background color is selected very easily and efficiently as being the color with the biggest rate of occurrences on the image borders.

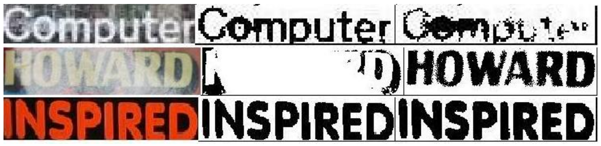

Fig. 2. Left: initial color images, middle: extraction done by D_{eucl}, right: extraction done by S_{cos}.

A new measure M is then proposed in this paper to find the most textual foreground cluster over the two remaining clusters. Based on properties of connected components of each cluster, spatial information is already added at this point to find the main textual cluster. M is based on the larger regularity of connected components of text than the one of noise and background and is defined as below:

$$M = \sum_{i}^{N} |area_i - \frac{1}{N}(\sum_{i}^{N} area_i)|$$

where N is the number of connected components and $area_i$, the area of the component i. This measure enables to compute the variation in candidate areas. The main textual cluster is identified as the one having the smallest M. If the last cluster belongs to text, both clusters need to be merged. A new computation of M is done considering the merging of both clusters. If M decreases, the merge is processed.

Figure 2 displays examples where D_{eucl} performs better than S_{cos} (top), the inverse result (middle) and a last example (bottom) where both clustering distances perform quite similarly.

4.1. Complementarity of Euclidean distance and Cosine-based Similarity

Both clustering distances enable to handle a large number of degradations in a complementary way. Images presenting a strong contrast between text information and background give usually better results with D_{eucl}. In the opposite direction where images are corrupted by uneven lighting or shiny surfaces, S_{cos} performs better. Due to the material, the angle of acquisition and the lighting, colors of text vary gradually and can present strong differences inside a character. As initially, colors inside the text were similar, we may face to metamers, which can be handled with a cosine-based similarity because the angle of a color and its metamer remains small compared to the Euclidean distance between both clusters.

Several cosine-based similarities have been designed and can be found in a large review [8]. After tests on natural scene images and especially on the database described in Section 6, we chose the following similarity S_{cos}:

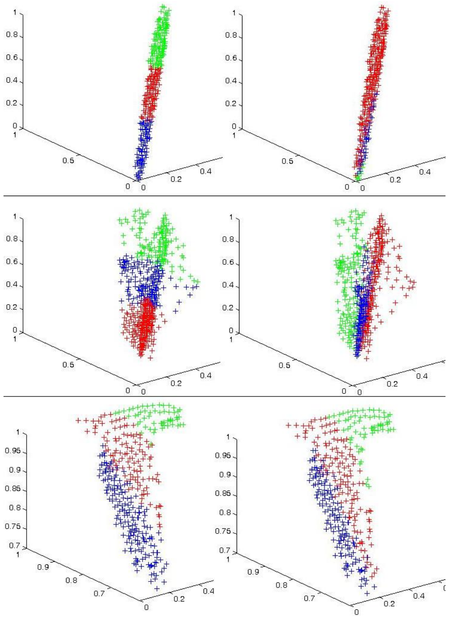

Fig. 3. (R-G-B) view of the clustering results done by D_{eucl} (left) and by S_{cos} (right) on samples of Figure 2.

$$S_{cos} = 1 - (\frac{x_i.x_j}{\|x_i\|.\|x_j\|})(1 - \frac{\|x_i\| - \|x_j\|}{max(\|x_i\|, \|x_j\|)})$$

This similarity presents a more compact support and performs better in natural scene text images [6].

The complementarity of D_{eucl} and S_{cos} can also be observed after clustering. In Figure 3, we show the cluster definition done by D_{eucl} and S_{cos} for the same examples of Figure 2. From the RGB color space, D_{eucl} separates pixels in the (R-G-B) view mostly in a horizontal way with clusters presenting quite same volumes while S_{cos} does the same operation in a more vertical way with clusters presenting different sizes. These observations are quite logical due to the definition of each distance but really show a complementarity of these two distances depending on colors in the image.

5. USE OF SPATIAL INFORMATION

In order to segment characters properly, we need to have simultaneously spatial information to locate characters in the image and frequency information to use illumination variation to detect character edges. Gabor-based filters could be a choice to address this problem. In this paper, we opt for Log-Gabor filters proposed by Field [9] and having an extended tail in high frequencies as required for natural scene images.

Log-Gabor filters in frequency domain can be defined in polar coordinates by $H(f, \theta) = H_f \times H_\theta$ where H_f is the radial component and H_θ the angular one:

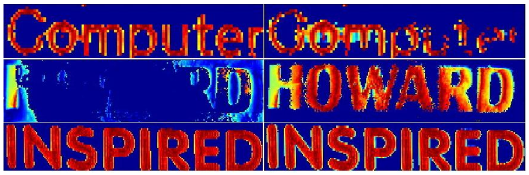

Fig. 4. Log-Gabor filters results on the three examples of Figure 2. Left: M_{eucl}, right: M_{cos}.

$$H(f, \theta) = \exp\{\frac{-[\ln(f/f_0)]^2}{2[\ln(\sigma_f/f_0)]^2}\} \times \exp\{\frac{-(\theta - \theta_0)^2}{2\sigma_\theta^2}\}$$

with f_0, the central frequency, θ_0, the filter direction, σ_f, which defines the radial bandwidth B in octaves with $B = 2\sqrt{2/\ln 2} \times |\ln(\sigma_f/f_0)|$ and σ_θ, which defines the angular bandwidth $\Delta\Omega = 2\sigma_\theta\sqrt{(}2\ln 2)$.

As we are looking for approximative vertical separation between characters, globally horizontal aligned, we use only two directions for the filter, the horizontal and the vertical one. Hence, for each directional filter, we got a fixed angular bandwidth of $\Delta\Omega = \Pi/2$. Log-Gabor filters are not really strict with directions and defining only two directions enables to handle italic or misaligned characters. For strongly misaligned ones, the number of directions can be simply increased to handle this additional degradation. f_0 is chosen as the inverse of the mean width of characters defined by the clustering result. We finally chose a bandwidth of 0.3 octaves to get a narrow filter to take into consideration more characters than background.

Figure 4 shows the result of the same three examples multiplied by the mask of each segmentation performed previously, M_{eucl} and M_{cos}; Log-Gabor filters induce globally higher values for characters than for background. Hence, in order to choose efficiently which clustering distance is better to handle text extraction, we perform an average of pixel values inside each mask. The segmentation which has the highest average is chosen as the final segmentation, as it represents mostly characters.

6. EFFICIENCY RATES ON A PUBLIC DATABASE

The combination of color and spatial information has to be assessed to measure the impact of the taken decision in the whole algorithm. The sample word ICDAR 2003 database $Sample$ [10], which is public, is used to compare results with other papers. $Sample$ includes 171 natural scene words with different degradations such as complex backgrounds, strong uneven lighting, blur, low resolution and so on. Until now, no papers use this database for this task and no other databases are mentioned to compare results.

Two tests have been done to evaluate the efficiency of the whole algorithm. First of all, in one of our previous tests [6], the use of several clustering distances is compared with a single distance-based clustering using only D_{eucl}, which works in most cases and the improvement is 6.3% in the number of well-extracted words, showing the efficiency of the simultaneous use of both clustering distances.

Secondly, to know which segmentation is right, we use visual judgement for segmentations presenting different results, as one image is hardly readable and for cases where both segmentations give a similar result, we use a home-made OCR, after separation into individual characters. It is important to note that an OCR algorithm is not sufficient to choose the best segmentation because in many cases, a separation into individual components is required and adds therefore confusion and computation time. In order to assess the use of spatial information to choose between the two distances, the silhouette Sil [11] is computed as described below and is a measure of how well clusters are separated:

$$Sil = \frac{min(mean_{between}(i,k)) - mean_{within}(i)}{max(mean_{within}(i), min(mean_{between}(i,k)))}$$

where $mean_{within}(i)$ is the average distance from the i^{th} point to the other points in its own cluster and $mean_{between}(i,k)$ is the average distance from the i^{th} point to points in another cluster k. The average distance is defined either by D_{eucl} or S_{cos}.

It can be logical to think that best text extraction results present the best separation between clusters. It is not necessary the case because this method chooses the right segmentation in 77.7% images and our proposed method using spatial information in 93.2% which is an improvement of 19.9 %.

Both tests show the efficiency of our method compared to variations in particular steps. Comparisons with the huge number of binarizations techniques are not relevant to mention as they performed poorly on color complex images and even more on natural scene images.

7. CONCLUSION AND FUTURE WORK

A novel text extraction method has been proposed in this paper combining color information and spatial information. The color information uses two different but complementary clustering distances to handle colors and their (slight) differences, created by several degradations such as reflecting materials and uneven lighting. The spatial information exploits the result of Log-Gabor filters, designed to support natural scene images. In the Result Section 6, the combination has been assessed and the efficiency of the use of spatial information to choose the right segmentation was shown.

Thanks to this whole method, more images could now be handled to a recognition purpose but one kind of images can not be handled yet. Images with very low-resolution and poor contrast such as in Figure 5 can not be segmented with unsupervised methods without interpolation or enhancement. One of our future works is also to take into consideration these difficult images with pre-processing algorithms on top of the novel method explained in this paper.

Fig. 5. Typical example which requires pre-processing to denoise and increase initial resolution.

8. REFERENCES

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

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

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

[4] H.Hase, T.Shinokawa, M.Yoneda, C.Y.Suen, **Character string extraction from color documents**, Pattern Recognition, 34:1349-1365, 2001.

[5] D.Chen, J-M.Odobez, H.Bourlard, **Text detection and recognition in images and video frames**, Pattern Recognition, 37:594-608, 2004.

[6] C.Mancas-Thillou, B.Gosselin, **Color text extraction from camera-based images-the impact of the choice of the clustering distance-**, Proc. of Int. Conf. on Document Analysis and Recognition, pp. 312-316, 2005.

[7] G.Wyszecki, W.S.Stiles, **Color science: concepts and method, quantitative data and formulae**, Second Edition, Wiley Editions, 1982.

[8] M.Hild, **Color similarity measures for efficient color classification**, Jour. of Imaging Sc. and Tech., pp. 529-547, 2004.

[9] D.J.Field, **Relations between the statistics of natural images and the response properties of cortical cells**, Jour. of the Optical Society of America, pp. 2379-2394, 1987.

[10] Robust Reading Competition: Retrieved April 28, 2006 from http://algoval.essex.ac.uk/icdar/RobustWord.html

[11] L.Kaufman, P.J.Rousseeuw, **Finding groups in data: an introduction to cluster analysis**, Wiley Editions, 1990.