Multiresolution Approach Based on Adaptive Superpixels for Administrative Documents Segmentation into Color Layers

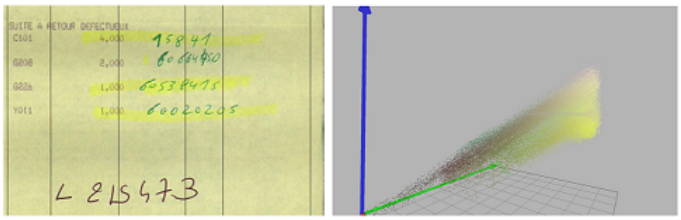
Elodie Carel, Jean-Christophe Burie, Vincent Courboulay, Jean-Marc Ogier, Vincent Poulain D 'Andecy

Multiresolution Approach Based on Adaptive Superpixels for Administrative Documents Segmentation into Color Layers

Elodie Carel*, Jean-Christophe Burie*, Vincent Courboulay*, Jean-Marc Ogier*, Vincent Poulain d'Andecy**
* L3i, University of La Rochelle, Avenue Michel Crépeau, 17042 Cedex, La Rochelle, France
Email:{elodie.carel, jean-christophe.burie, vincent.courboulay, jean-marc.ogier}@univ-lr.fr
** ITESOFT Headquarters: Parc d'Andron Le Séquoia, 30470 Aimargues, France
Email: Vincent.PoulaindAndecy@itesoft.com

***Abstract*—Administrative document images are usually processed in black and white what generates many problems due to the errors related to the binarization. Besides all semantic information provided by the color is lost. Document images have a rich and highly variable content. The presence of false colors and artefacts introduced by the scanning and the compression alter the segmentation of the regions. Problems arise when there is no correspondence between the point clouds which are detected in a color space and the real regions of an image. In order to help the segmentation, we propose the extraction of the main colors of an image as a set of binary layers. Due to the industrial context, our approach has to run unsupervised on a generic dataset of color administrative documents. The originality of this approach is the use of a multiresolution analysis to detect the number of colors automatically. At a low resolution, a set of local regions is obtained thanks to a SLIC-based approach which takes into account the structure of documents and which combines both colorimetric information and spatial information. Then, a merging stage is applied on each resolution separately based on the colors which have been extracted at a lower resolution. This contribution can both feed the traditional process and exploit colorimetric information.**

(a) Original images (on the left), binarized images (on the right)



(b) Original image (on the left), visualization in the RGB color space (on the right).

Fig. 1: Problems met by companies

I. Introduction

Industrial companies receive huge volumes of documents everyday. They are highly variable in terms of structure complexity, in terms of content, and in terms of colors (Fig. 6a). The traditional digitization process simplifies the images in order to respect the industrial constraints. Thus, images are usually processed in black and white which generates many problems due to the errors related to the binarization. Companies have to cope with the presence of false colors and artefacts introduced by the scanning and the compression, with overlapping elements, and with the loss of information when the segmentation fails (Fig. 1a). To reduce these problems, the application of post-processing requiring *a priori* knowledge and the tuning of parameters is often necessary. Furthermore, all semantic information provided by the color is lost. But color is often meaningful as for example in the case of highlighting (Fig.1b). Despite a growing interest in the document analysis community, there are still few works dealing with color information to improve the segmentation results. Most of them are dedicated to a specific dataset which contains documents sharing the same features. Here, we aim to process a generic dataset of color administrative documents with any kind of structure and any kind of color. It is not always possible to detect well separated point clouds in a given color space, and to have a correspondence with the real regions which are observable on an image (Fig. 1b).

We stand at the beginning of the digitization chain. Based on our industrial requirements, we propose the extraction of the main colors of the images as a set of binary layers. For a given layer, all pixels belonging to the same color appear black on a white background (Fig. 2). The purpose is to improve the segmentation results by using the color information. A segmentation into color layer images proposes an alternative to the simple foreground-background binarization which leads to some errors. Companies can still keep using the traditional process by feeding an OCR with the binary layers under the assumption that the textual layers have the same color. But, it is now possible to use the colorimetric information as extra semantic information about the components of the images. This approach could be used to describe a document by its main colors, or to process only a specific layer. We do not focus only on textual parts but also on other main elements such as regions, graphics and so on. The main difficulty is to segment a document without the help of models or any user interaction to add *a priori* knowledge. The novelty of this work is to detect automatically the number of colors of each document thanks to a multiresolution analysis. We also proposed to extend a SLIC-based approach to document images in order to combine both color and spatial information.

Fig. 2: Example of color layers, from the synthetic ground-truth dataset: input image (1st image on the left), and the output images (on the right).

Finally, the difficulties raised by the context are summarized below:

- industrial context,
- generic dataset,
- automatic detection of the number of colors,
- unsupervised segmentation,
- and duality between color segmentation and geometric segmentation.

The paper is organized as follows. Section 2 presents the state-of-the-art. Then, the different steps of the color layer extraction are described in section 3. Results are discussed in section 4. Finally, we conclude in section 5.

II. State-of-the-art

Document images are usually simplified due to the industrial context. [1] proposed a foreground-background separation for low quality color document images. After a connected component labeling, they locate dominant background areas which are used to extract foreground regions thanks to a local bicolor clustering. However, their approach is not applied on complex documents and color information is lost at the end of the process. Document images have a rich content which is highly variable in terms of structure complexity and in terms of number of colors. But since they are made to be easily readable by humans, we assume that they will contain only a few contrasted colors (often less than 10). Defining a relevant number of colors is difficult. The noise introduced by the digitization process also alters them. There are often shades of a same hue. Even if a region looks homogeneous for a human, variations and false colors can be noticed at a pixel-level. In the last few years, there has been a growing interest in the study of color as a source of information to help the segmentation. Color segmentation techniques can be classified into *spatially blind* and *spatially guided* approaches.

Histogram-based approaches are common spatially blind techniques which do not require any *a priori* knowledge. They are often used to perform a classification into two classes such as in [2] where document images are segmented into chromatic and achromatic areas. Clustering-based approaches group elements according to a similarity measure. Among them, the MeanShift is a popular segmentation procedure in document analysis. It aims at locating the maxima of a density function. [3] used it for a color reduction purpose. [4] proposed a fast optimized MeanShift for big document images. Both histogram-based and clustering-based approaches required well-separated classes. According to [5], there is not always a one-to-one correspondence between the regions in the image and the clusters of color points in the color space. The visualization of a document in a color space lets generally appear more strokes than easily separable point clouds.

To overcome this problem, some approaches combine both color and spatial features. [5] proposed a spatial-color compactness degree to study the likelihood that a given subset of points in the color space corresponds to an actual region in the image. [6] extracted text on color document images with a region growing-based approach. The main difficulty of these approaches is the selection of the seeds which have to be chosen carefully. They are often computationally expensive. The merging order can influence the results and they can lead to an over-segmentation of the image. Some other approaches voluntary over-segment the images into perceptually meaningful local regions known as *superpixels* (SP for short). [7] proposed a new SP algorithm : the *Simple Linear Iterative Clustering* (SLIC). SLIC extends a local *k*-means clustering approach. There is still a few works about the use of superpixels in document analysis. [8] used superpixels to extract drawing regions on historical documents. The advantage is that the extracted regions will be compact, rather homogenous and they will adhere well to the image boundaries. Working at a region-level is computationally less expensive than working at a pixel-level and it limits the effect of the noise on the result. This process can be used as a pre-processing to the final segmentation.

In segmentation, a merging process is usually performed by optimizing a cost function or until a specific condition is validated (e.g. number of iterations or heuristic rules). Finding a good stopping criterion is always challenging. Furthermore, image segmentation algorithms require precise tuning to work properly, and most of the time the tuning is adapted to the extraction of one particular kind of information. Inspired from visual perception principles, we consider that a multiresolution analysis can help the process. Indeed, an image at a full-resolution brings a lot of information which are difficult to be extracted by one algorithm. However, only the coarse structure will be perceived at a low-resolution. Thus, a multiresolution analysis can be used to perceive structures of different sizes. The tuning of one resolution can lead to the tuning of another resolution. In document analysis, it is commonly used for the extraction of text lines by analysing the connected components at different levels of resolution such as, for example, in [8].

In conclusion, a multiresolution analysis seems particularly interesting in our case for its ability to simulate the human perception. Main observable colors can be extracted at a low-resolution thanks to a segmentation approach. Based on this state-of-the-art, an over-segmentation into local regions seems more adaptive than a global segmentation since we aim at segmenting a generic dataset of documents with no *a priori* knowledge on their content. A spatially guided approach such as a SLIC seems the most appropriate to color document images due to the noise occurring on this kind of images. Locally, we can assume that colors contain enough contrast to allow the extraction of coherent regions thanks to a combination of spatial and color information. A SLIC procedure can produce coherent regions with respect to the geometric structure of the images. Finally, the segmentation can be refined by applying a merging step on each resolution separately based on colors which have been extracted at a lower resolution.

III. COLOR LAYER EXTRACTION

In this section, the segmentation of document images into global color layers will be described. We need our system to run unsupervised and to be adaptive since we do not know the number of colors for each document.

A. The multiresolution analysis

A full-resolution image often contains too much details to be analysed by algorithms applied at a unique resolution. This is particularly true in our case because we aim at dealing with complex document images which can contain any kind of structure. A multiresolution analysis allows us to perceive elements of different sizes. The idea is to simulate the human perception. By far, only the global structures are perceived. Close, we can focus on small elements. To follow this idea, we have built a Gaussian pyramid. At the lowest resolution, we apply an over-segmentation which combines spatial and color information. According to the state-of-the-art, this approach is more robust to noise and seems more adaptive than a global segmentation because it can produce coherent regions with respect to the local geometric structure of the images. It will be presented in the part B. The work is done in the L*a*b color space which is a perceptual color space for which the relevance of the euclidean distance for the computation of the distance between two colors has been recognized. The terms color and regions will be employed for a cluster of pixels sharing close colorimetric features. After the segmentation, the local regions are merged into global color layers and the segmentation is refined thanks to a multiresolution analysis as summarized in Fig. 3.

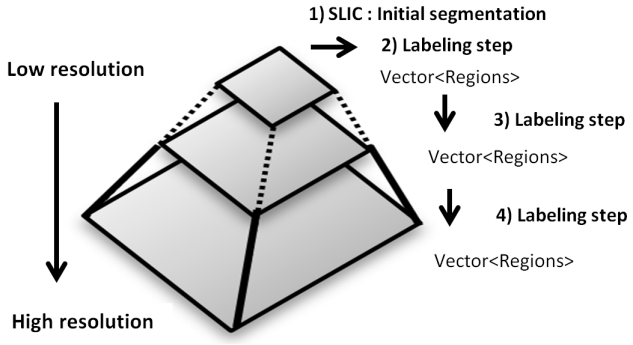


Fig. 3: Multiresolution analysis used for the merging stage.

Thus, the segmentation is applied only on the lowest resolution and the labeling stage is the same for all the resolution levels. It is decomposed into two sub-steps: a merging step which aims to clean the results of the segmentation, and the labeling step at a pixel-level which aims to refine the segmentation.

1) The merging procedure: This step merges very close colors, removes too small areas, and removes color clusters when their pixels are too sparsely scattered through the image. First, in order to reduce noise by removing too small areas, we study the distribution of the color of each pixel inside a 3X3 window centered on this pixel. If only a few occurrences of this color are found, we look for a background color and a foreground color assuming that only two colors will appear within this small window. We assume that the background color is the color with the highest distribution value and the foreground color is the one having the highest color distance with the detected background. Pixels are then reassigned. Then, to detect regions which are too sparsely scattered through the image and which correspond to noise, we threshold the connectedness degree defined in [5] in order to keep only strongly connected regions.

2) The labeling stage: Here, we aim to refine the segmentation. For a given resolution, the idea is to use the color labels obtained at a lower resolution and to check if the label had been well assigned according to the information provided by the L*a*b image of the current resolution. The advantage is that we can have a more global vision of the main colors. Moreover, some parts of the image can be not properly detected at a low resolution and could be corrected if we have an idea of the colors which are observable within a close area. That is why we go through the image and for each pixel, we check, within a search region, whether there is a closer color than the one associated to the label of the pixel according to the measure of the euclidean distance. Assuming that the orientation of most of the elements of documents is horizontal or vertical (e.g column of labels, graphics), the search window is the line and the column centered on the pixel. To speed up the process, we consider only one pixel over N. At a low resolution where we perceive color at a global level, we consider a whole line and a whole column. At high resolution, we aim at refining the segmentation and we reduce the size of the search window in order to avoid introducing noise.

B. The segmentation stage

In this section, we will describe the segmentation procedure applied at the lowest resolution level. We have chosen to apply the SLIC segmentation proposed in [7] because of its ability to produce compact local regions which adhere well to the image boundaries. By default the only parameter of the original algorithm is *k*, the desired number of approximately equally-sized local regions. SLIC is an adaptation of k-means. During the initialization step, k SPs are sampled on a regular grid. Let S be the step of the grid interval. Each SP in position $[x_i, y_i]$ is associated with a vector $C_i = [l_i, a_i, b_i, x_i, y_i]^T$. To speed up the algorithm, the clustering limits the search to a region 2S x 2S around the SP center (Fig.4). A weighted distance combining a color distance with a spatial distance is measured between any pixel and its neighboring SPs. Then, each pixel is associated with the nearest cluster center whose search region overlaps its location. SPs centers are the mean $[l, a, b, x, y]^T$ vector of all their pixels. The process is then iterated. SLIC requires setting the initial number of SPs and they will be approximately equally-sized. If the step value is too big, a lot of small elements such as characters will be missed because their color will not be taken into account during the sampling. On the other hand, document images contain a lot of big and homogeneous regions where small SPs are useless. In a generic dataset, it is difficult to tune properly the number of initial superpixels because of the variations in terms of content between two documents.
In this paper, we propose to make use of the structure of document images and to adapt the initialization of the original SLIC to the content of the structure of document images. The originality is that the centers of the SPs are not distributed

homogeneously anymore during the initialization stage. Elements which are likely part of the foreground are separated from those which belong likely to the background. Then, we use the two masks to initialize the centers as for the original SLIC but with different values for the grid interval: a big one for the background mask, a small one for the foreground part. Therefore, the search window of the SP centers initialized on the background mask will be bigger than the search window of the SP centers initialized on the foreground mask. Finally, the clustering is performed as the original SLIC. In order to reduce the number of final regions, all SPs having the same color (i.e. the same center value) are merged. At the end of the process, homogeneous regions where the color variability is low will have bigger SPs than in gradient areas.

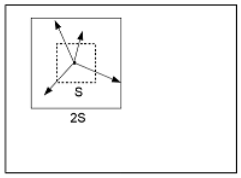


Fig. 4: SLIC searches a limited region [7].

IV. Results and Discussion

To suit the issues companies have to deal with, we have selected 150 real documents with both handwritten, and printed text, of different sizes, with different fonts. Background regions can be complex with texture or gradient areas. They contain a variable number of colors. In order to evaluate the performance of our approach 2000 synthetic images have been automatically generated. The color value is known at a pixel-level. In this work, we assumed that the images have a main background color and a main foreground color. Two templates have been defined so far to get semi-structured images which look like documents : half of the documents in the dataset are letter documents, and the other half leaflet documents containing only text on background (Fig. 5). Several documents of a same family were generated with a part with fix content, and another one with random content. The elements which are randomly initialized are : the colors, the text, the size, and the shape of the elements. For each synthetic image from the original dataset, the generated ground-truth is a set of binary layer images. For a given layer, all the pixels belonging to the same color appear black on a white background. The ground-truth is available for download at this address: http://navidomass.univ-lr.fr/ColorSegmentationGT/.

In order to simulate the real noise on our synthetic dataset, Gaussian noise has been added to the images. Then, they have been saved with a high jpeg compression to suit the constraints met by our industrial partner in its digitization flow. All experiments have been conducted on this noisy dataset on three levels of resolution. For comparison we have also applied the approach using only the full-resolution. Fig. 6b and Fig. 6c show the results of the segmentation applied to a real document. Visual examinations show the relevance of a SLIC-based approach for document analysis. The adaptive version takes into account the structure of document images to produce SPs of different sizes which adhere well to the edges. Table I gives the number of colors for each step of the process. Their final number will depend on the threshold values which are applied during the process. Since global colors are detected at a low-resolution where the effect of the noise is less important, we can notice that the multiresolution-based approach produces fewer layers corresponding to noise such as transition areas between foreground and background. What is interesting is that we get almost the final number of colors after the labeling step at the low-resolution level. As the low-resolution image is much smaller than the full-resolution one, its segmentation will be faster. The multiresolution analysis allow us to reduce the final number of colors by refining the segmentation: from a global point of view to a local one. At a given resolution, the labeling process is even able to correct some non-detected areas when the corresponding color appears in the search window. It has also been proved necessary to build properly small elements such as textual regions.

Fig. 5: Synthetic images.

	Multiresolution	Full-resolution
After SLIC segmentation	141	175
low-resolution: labeling	23	-
medium-resolution: labeling	20	-
Full-resolution: labeling	16	25

TABLE I: Number of labels extracted for each step of the process for the real image, Fig.6a.

Finally, we measured the precision and the recall of the resulting segmentation according to the ground-truth. Results are summarized in Table II. The precision value for the multiresolution-based approach can seem contradictory to the previous conclusions. But our evaluation does not penalize the tested approach when it has found too many layers compared to the ground-truth. It simply compared the ground-truth layer to the most likely resulting layer. The analysis of the cases where the precision were low concludes that the comparison methods leads to an over-segmentation in most of the cases: sometimes twice as many layers found compared to the multiresolution-based approach which generally extracts a number of layers close to the ground-truth. The recall values confirms this conclusion. The multiresolution-based approach

	Multiresolution	Full-resolution
Precision	0.85	0.92
Recall	0.97	0.94

TABLE II: Performance of the color layer extraction applied on the synthetic dataset.

can lead to a under-segmentation which explains the difference between the precision of the two methods. In our case, results are satisfying for our issues. Indeed, we prefer an under-segmentation, especially when it involves close colors, where coherent geometric regions can be extracted, rather an over-segmentation which produces a lot of noisy layers difficult to be analysed.

V. Conclusion

The binarization of complex color documents causes segmentation problems. In this paper, we proposed the extraction of the main colors observable on a document image as a set of binary layers which can feed the traditional process. The aim is to benefit from the color information to improve the segmentation process. The originality of our approach is the use of a multiresolution analysis to detect the number of colors automatically. The segmentation approach combines both colorimetric information and spatial information and takes into account the structure of document images. Due to the industrial context, our approach runs unsupervised on a dataset of documents containing any structures, and any colors. Results show that a multiresolution-based approach is relevant in the context of the segmentation of this kind of images. In this paper, we have been focused on the segmentation step. At the end of the process, the colorimetric information brought by each layer is a numerical value which make the query not intuitive for humans. As a future work, we plan to associate semantic information to color layers with a color naming approach.

References

[1] U. Garain, T. Paquet, and L. Heutte, "On foreground-background separation in low quality color document images," in *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, vol. 2005, 2005, pp. 585–589.

[2] A. Ouji, Y. Leydier, and F. Lebourgeois, "Chromatic / Achromatic Separation in Noisy Document Images," *2011 International Conference on Document Analysis and Recognition*, pp. 167–171, 2011.

[3] N. Nikolaou and N. Papamarkos, "Color reduction for complex document images," *International Journal of Imaging Systems and Technology*, vol. 19, no. 1, pp. 14–26, 2009.

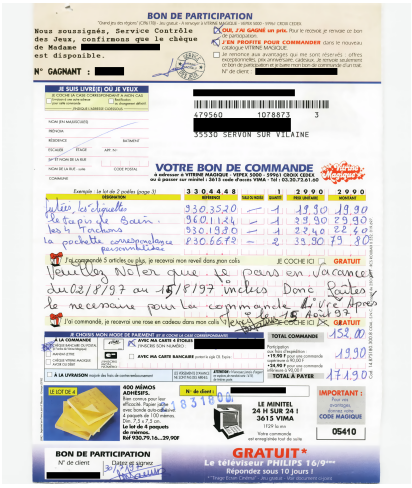
[4] F. Lebourgeois, F. Drira, D. Gaceb, and J. Duong, "Fast Integral MeanShift : Application to Color Segmentation of Document Images," in *Twelfth International Conference on Document Analysis and Recognition (ICDAR 2013)*, IEEE, Ed., 2013, pp. 52–56.

[5] L. Macaire, N. Vandenbroucke, and J. G. Postaire, "Color image segmentation by analysis of subset connectedness and color homogeneity properties," *Computer Vision and Image Understanding*, vol. 102, pp. 105–116, 2006.

[6] P. K. Loo and C. L. Tan, "Adaptive Region Growing Color Segmentation for Text Using Irregular Pyramid." in *Document Analysis Systems*, ser. Lecture Notes in Computer Science, S. Marinai and A. Dengel, Eds., vol. 3163. Springer, 2004, pp. 264–275.

[7] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, pp. 2274–2281, 2012.

[8] R. Cohen, A. Asi, K. Kedem, J. El-Sana, and I. Dinstein, "Robust Text and Drawing Segmentation Algorithm for Historical Documents," in *Proceedings of the 2Nd International Workshop on Historical Document Imaging and Processing*, ser. HIP '13. New York, NY, USA: ACM, 2013, pp. 110–117.



(a) Original image

(b) Low-resolution segmentation (in false colors)

(c) Final segmentation (in false colors)

Fig. 6: Results of the multiresolution segmentation. For readibility, we make the results appear on one image with false colors. A set of binary color layers can be created from it.