Segmenting and Indexing Old Documents Using a Letter Extraction

Mickaël Coustaty, Sloven Dubois, Jean-Marc Ogier, Michel Menard

Segmenting and Indexing old documents using a letter extraction

Mickael Coustaty, Sloven Dubois, Jean-Marc Ogier, and Michel Menard

L3i Laboratory
Avenue Michel Crepeau, 17042 La Rochelle, France
E-mail: {mcoustat, sduboi01, jmogier, mmenard}@univ-lr.fr

Abstract. This paper presents a new method to extract areas of interest in drop caps and particularly the most important shape: Letter itself. This method relies on a combination of a Aujol and Chambolle algorithm and a Segmentation using a Zipf Law and can be enhanced as a three-step process: 1)Decomposition in layers 2)Segmentation using a Zipf Law 3)Selection of the connected components.

1 Introduction

With the improvement of printing technology since the 15th century, there is a huge amount of printed documents published and distributed. Since that time, books have been falling into decay and degrading. This means not only books themselves are disappearing, but also the knowledge of our ancestors. Therefore, there are a lot of attempts to keep, organize and restore ancient printed documents. With the improving digital technology, one of the preservation methods of these old documents is the digitization. However, digitized documents will be less beneficial without the ability to retrieve and extract the information from them which could be done by using techniques of document analysis and recognition. This paper presents a new method to improve old document images description using segmentation and characterization of letter inside.

1.1 NaviDoMass

NAVigation Into DOcuments MASSes is a french collaborative projec, financed by the National French Research Agency, with the challenge to index ancient documents. With the collaboration of seven laboratories in France, the global objective of this project is to build a framework to derive benefit from historical documents. It aims to preserve and provide public accessibility to this national heritage and is established on four principles: anywhere (global access), anyone (public and multilingual), anytime and any media (accessible through various channels such as world wide web, smartphone, etc.). The focus of NAVIDOMASS is on five studies: (1) user requirement, participative design and ground

truthing, (2) document layout analysis and structure based indexing, (3) information spotting, (4) structuring the feature space [HSO+08,JT08] and (5) interactive extraction and relevance feedback. As a part of NAVIDOMASS project, this paper focuses on the graphics part : graphics indexing and CBIR. However, the main interest of this study is based on specific graphics called drop caps, and on the extraction of shapes in drop caps and particularly on the most important shape : the letter itself. This work is inspired by [PV06] and [ULDO05] which used a Zipf law and a Wold decomposition to extract elements of drop caps.

1.2 Drop caps in details

The images of documents of the inheritance are heterogeneous and damaged by time. Drop caps (decorative capital letters also named drop caps or drop cap) belong to the images to index. These images are made up of two principal elements: the letter and the background. (See Fig. 1). An important step in the

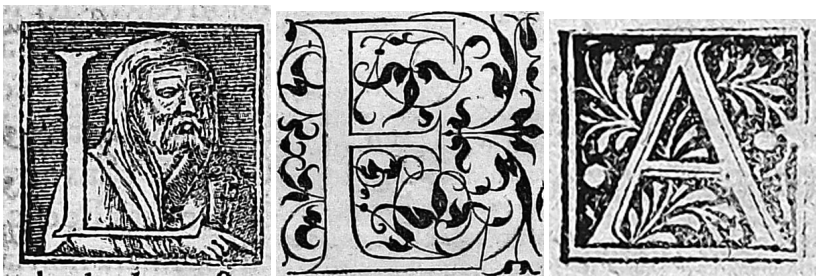


Fig. 1. Drop Caps Examples

recognition process of the drop caps consists in segmenting the letter and the elements of the background to characterize them using a signature. This signature will allow a simple and fast comparison for our indexing process of great masses of data. This paper presents in details the various stages of our method: 1) Simplification of the images using layers 2) Extraction of shapes from one of these layers 3) Selection of these shapes.

2 Simplification of images using layers

Decomposing an image into meaningful components appears as one of major aims in recent development in image processing. The first goal was image restoration and denoising; but following the ideas of Yves Meyer [Mey01], in total variation minimization framework of L. Rudin, S. Osher and E. Fatemi

[LSE92], image decomposition into geometrical and oscillatory (i.e texture) components appears an useful and very interesting way in computer vision and image analysis. There is a very large literature and also recent advances on image decomposition models, image regularization and texture extraction and modeling. So, we only cite, among many others, most recent works which appear most relevant and useful paper. In this way, reader can refer to the work of Stark et al. [SED05], Aujol et al. [AAFC05], [AGCO06], Aujol and Chambolle [AC05], Aujol and Ha Kang [AK06], Vese and Osher [OSV03], [VO04], [VO06] and more recently Bresson and Chan [BC07] and Duval et al. [DAV08] to cover the most recent and relevant advances.

2.1 The developped method

Images of drop caps are very complex and very rich images in terms of information and requires to be simplified. These images are mainly made up of lines, unsuitable for usual texture methods. We thus use an approach developed by Dubois and Lugiez [DLPM08] to separate original image in several layers of information, easier to process. This decomposition relies on minimization of a functional calculus F:

$$inf_{(u,v,w)\in X^3/f=u+v+w} F(u,v,w) =$$
$$\underbrace{J(u)}_{\substack{\text{Regularization}\\ \text{TV}}} + \underbrace{J^*\left(\frac{v}{\mu}\right)}_{\substack{\text{Texture}\\ \text{extraction}}} + \underbrace{B^*\left(\frac{w}{\delta}\right)}_{\substack{\text{Noise extraction by:}\\ \text{shrinkage}}} + \underbrace{\frac{1}{2\lambda}\|f-u-v-w\|_X^2}_{\text{Residual part}} \quad (1)$$

where each element of the functional represents a layer of information and corresponds to a type of information in the image. B^* can be seen as a wavelet soft threshold, J^* a computation of a gradient and J a linear computation between the original image and the two precedent elements. For deeper explanation about notations and each element, one can refer to [DLPM08].

2.2 Layers in details

We are aiming to catch the pure geometrical component in an image independently of texture and noise to extract shapes. So, we are studying here how to decompose images into three components:

– The Regularized Layer corresponds to the area of image which has low fluctuation of gray level. This layer permits to highlight geometry which corresponds to shapes in the image. In the following of this paper, we will name this layer the "*Shape Layer*".

– Oscillating Layer which corresponds to the oscillating element of the image. In our case, this layer highlights texture from drop caps and in the following of this paper, we will name this layer the "*Texture Layer*".
– Highly Oscillating Layer which corresponds to noise in image. in fact, this layer retrieves all that do not belong to the two first layers. So, we can find in this layer noise, text of background and problem of ageing. Our goal is to recognize old document images while being robust toward noise variations. That is why we will not use this layer in the next of this work.

An example of decomposition applied to the first image of Fig. 1 is given in Fig. 2.

(a) Shapes (b) Textures (c) Noise

Fig. 2. An example of drop cap decomposition using Aujol and Chambolle algorithm

3 Extraction of shapes

Regularized layer obtained by decomposition contain all shapes. In order to extract them and to select the most interesting, we used a *Zipf Law*. Zipf Law was empirically defined by *George Kingsley Zipf* and relies on the frequency and on the rank of appearance of words in a text. This law has been transposed on images by [PV06] by taking subimages as patterns and by calculating frequency and rank of these patterns. This method is a three steps process:

– Simplification of image applying a 3-means on gray level histogram to reduce number of patterns
– Seek for patterns of size three by three to obtain their frequency and their rank
– Classification of patterns in three classes according to the evolution law of the frequency compared to their rank

3.1 Simplification of images

Images provided by historians are composed of 256 gray levels. A huge amount of three by three patterns are possible (theoretically 256^9 different patterns). Indeed, if all patterns are represented only once, the model that is deduced from the pattern frequencies would not be reliable, the statistics would lose their significance. Then it is necessary to restrict the number of perceived patterns to give sense to the model.
To decrease the number of graylevels in the image, we have made use of k-means clustering algorithm [McQ67]. As images we are dealing with in this study are composed of three elements (background, foreground and motive), we decided to keep only three gray levels. Moreover, this reduction is made without loosing too much information.

3.2 Patterns research

Once the number of gray levels have been reduced, a simple count of each pattern permit to know their frequency and their rank. This step is essential to build the Zipf curve which represent the evolution law of the frequency compared to their rank. From this curve, three straight lines are computed to estimate three main parameters of Zipf laws that interfere. The splitting points are defined as the furthest points from the straight line linking the two extreme points of the curve. The first line, which correspond to the most frequent patterns, represent shapes of image (uniform areas). We have extracted pixels from each layer and we display them in an image. Figure. 3 show an example of a Zipf curve with its drop cap while Figure. 4 show an example of binarized images obtained with each straight line.

3.3 Shapes extration

Once shapes have been extracted, one can seek connected components of binarized image. When we observe all the connected components in Figure 5(b), we can see that the most important shapes have particular characteristics (based on size, location, center of mass and eccentricity). A selection of connected components in accordance with these parameters permit to obtain region of interest of drop caps. An example of extracted connected components can be seen in Figure 5(c). Finally, with an accurate selection on these parameters, the most important connected component for historian can be extracted: the letter. This one can be obtained by selecting the bigger connected component which center of mass is centered and which don't touch borders of image. An example of result can be seen in Figure 5.

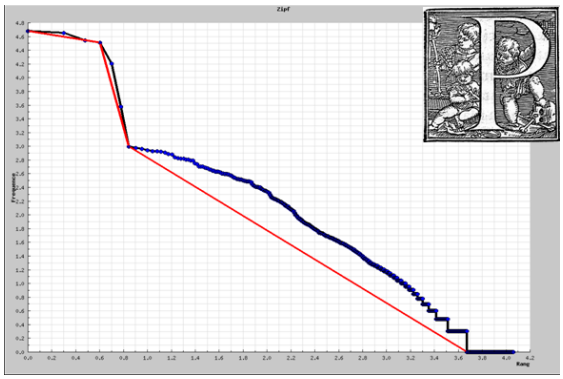
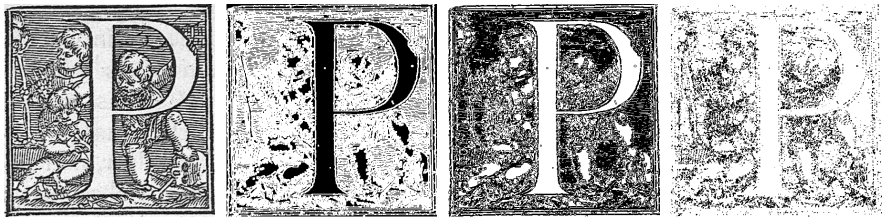


Fig. 3. Example of a Drop Cap and its Zipf plot where are indicated the different straight zones extracted



(a) Original image (b) First straight line (c) Second straight line (d) Third straight line

Fig. 4. Example of a Drop Cap and images corresponding to the straight lines of Zipf'curve

(a) Original image (b) Shapes of images (c) Three larger connected components (d) The letter extracted

Fig. 5. An example of treatment on a drop cap

4 Experimentations and validation

The evaluation of such a system is a fundamental point because it guarantees its usability by the users, and because it permits to have an objective regard

on the system. In the context of such a project, the implementation of an objective evaluation device is quite difficult, because of the variability of the user requirements: historian researchers, netsurfers, are likely to retrieve many different information which can be very different the ones from the others.
In the context of NAVIDOMASS project, and more specifically for this objective of drop caps indexing, we have decided to evaluate the quality of our system by considering the purpose of Letter Based Retrieval . This choice is motivated by the fact that many historians want to be able to retrieve drop caps in regard with this criterion. As a consequence, the evaluation of our system relies on the application of an OCR system at the issue of the letter segmentation. Considering these aspects, the classification rate is the main performance evaluation criterion of our system.
For the evaluation, we have used commercial OCR systems, as well as open source system. In order to implement the evaluation, we have used FineReader on the one hand, and Tesseract on the other hand. We have experimented the approach on an image database containing 4500 images. 1500 of these images were considered for the training set, while 3000 were considered for the tests. The results are summarized in the Table 1. As one can see the obtained results

	FineReader	Tesseract
Classification Rate	72,8%	67,9%

Table 1. Recognition rate of drop caps using two kinds of OCR

are still unsatifying, but very encouraging. We are working on the improvement of the processing chain, as one can see in the conclusion and perspective part. However, there is not such existing system dealing with this problem, and historians researchers are satisfied to use our system for the classification of their graphic images. The cases for which our system fails correspond to very difficult images, as one can see an example in Fig. 6.

5 Conclusions

This paper presents a new method to extract informations in drop caps. It relies on a combination of two decomposition. The first one simplifies image to only extract shapes of original image while the second one, a Zipf Law' decomposition, realize a background-foreground segmentation. From this segmentation, a selection of shapes segmented permit to extract some interesting shapes and particularly the letter itself. The first experimentations are encouraging and we are actually working on improvements of this global process.

(a) Original lettrine (b) Letter extracted

Fig. 6. An example of very difficult letter extraction

References

[AAFC05] J. F. Aujol, G. Aubert, L. Blanc Feraud, and A. Chambolle. Image decomposition into a bounded variation component and an oscillating component. *Journal of Mathematical Imaging and Vision*, 22(1):71–88, January 2005.

[AC05] Jean-François Aujol and Antonin Chambolle. Dual norms and image decomposition models. *International Journal of Computer Vision*, 63(1):85–104, 2005.

[AGCO06] Jean-François Aujol, Guy Gilboa, Tony Chan, and Stanley Osher. Structure-texture image decomposition - modeling, algorithms, and parameter selection. *International Journal of Computer Vision*, 67(1):111–136, 2006.

[AK06] Jean-François Aujol and Sung Ha Kang. Color image decomposition and restoration. *J. Visual Communication and Image Representation*, 17(4):916–928, 2006.

[BC07] X. Bresson and T. Chan. Fast minimization of the vectorial total variation norm and applications to color image processing. In *SIAM Journal on Imaging Sciences (SIIMS)*, submitted 2007.

[DAV08] Vincent Duval, Jean-Franois Aujol, and Luminita Vese. A projected gradient algorithm for color image decomposition. Technical report, CMLA Preprint 2008-21, 2008.

[DLPM08] S. Dubois, M. Lugiez, R. Péteri, and M. Ménard. Adding a noise component to a color decomposition model for improving color texture extraction. *CGIV 2008 and MCS08 Final Program and Proceedings*, pages 394–398, 2008.

[HSO+08] H.Chouaib, S.Tabbone, O.Ramos, F. Cloppet, and N.Vincent. Feature selection combining genetic algorithm and adaboost classifiers. In *ICPR'08*, Florida, 2008.

[JT08] Salim Jouili and Salvatore Tabbone. Applications des graphes en traitement d'images. In *ROGICS'08*, pages 434–442, Mahdia Tunisia, 2008. University of Ottawa, Canada and University of Sfax, Tunisia.

[LSE92] L.Rudin, S.Osher, and E.Fatemi. Nonlinear total variation based noise removal. *Physica D*, 60:259–269, 1992.

[McQ67] J. B. McQueen. Some methods for classification and analysis of multivariate observations. In University of California Press, editor, *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297, Berkeley, 1967.

[Mey01] Y. Meyer. *Oscillating patterns in image processing and nonlinear evolution equations*. The fifteenth dean jacqueline B. Lewis Memorial Lectures, 2001.

[OSV03] S. J. Osher, A. SoIe, and L. A. Vese. Image decomposition, image restoration, and texture modeling using total variation minimization and the H-1 norm. In *International Conference on Image Processing*, pages I: 689–692, 2003.

[PV06] Rudolf Pareti and Nicole Vincent. Ancient initial letters indexing. In *ICPR '06: Proceedings of the 18th International Conference on Pattern Recognition*, pages 756–759, Washington, DC, USA, 2006. IEEE Computer Society.

[SED05] J. L. Starck, M. Elad, and D. L. Donoho. Image decomposition via the combination of sparse representations and a variational approach. *IEEE Trans. Image Processing*, 14(10):1570–1582, October 2005.

[ULDO05] Surapong Uttama, Pierre Loonis, Mathieu Delalandre, and Jean-Marc Ogier. Segmentation and retrieval of ancient graphic documents. In *GREC*, pages 88–98, 2005.

[VO04] L. A. Vese and S. J. Osher. Image denoising and decomposition with total variation minimization and oscillatory functions. *Journal of Mathematical Imaging and Vision*, 20(1-2):7–18, January 2004.

[VO06] Luminita A. Vese and Stanley Osher. Color texture modeling and color image decomposition in a variational-PDE approach. In *SYNASC*, pages 103–110. IEEE Computer Society, 2006.