Drop Caps Decomposition For Indexing - A New Letter Extraction Method

Mickaël Coustaty, Jean-Marc Ogier, Rudolf Pareti, Nicole Vincent

Drop Caps Decomposition For Indexing
A New Letter Extraction Method

Mickael Coustaty Jean-Marc Ogier
Imedoc Team - L3i Laboratory
Avenue Michel Crepeau
17042 La Rochelle, France
{mcoustat, jmogier}@univ-lr.fr

Rudolf Pareti Nicole Vincent
SIP Team - CRIP5 Laboratory
45, rue des Saints-Pres
75270 Paris Cedex 06, France
nicole.vincent math-info.univ-paris5.fr

Abstract

This paper presents a new method to extract shapes in drop caps and particularly the most important shape: Letter itself. This method relies on a combination of a Aujol and Chambolle algorithm threaded with a segmentation using a Zipf Law in a second step. This method can be enhanced as a three-step process: 1)Decomposition in layers 2)Segmentation using a Zipf Law 3)Selection of connected components to only emphasize the required information - letter itself.

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.

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of seven laboratories in France, the global objective of this project is to build a framework to derive benefit from historical documents.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 [8, 9] 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 [14] and [16] which used a Zipf law and a Wold decomposition to extract elements of drop caps.

1.1. 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 Figure. 1). An important step in the recognition process of the drop

Figure 1. Drop Caps Examples

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 in-

dexing 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. Aujol and Chambolle Algorithm to extract signatures

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 [12], in total variation minimization framework of L. Rudin, S. Osher and E. Fatemi [10], image decomposition into geometrical and oscillatory (i.e texture) components appears a 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. [15], Aujol et al. [1], [3], Aujol and Chambolle [2], Aujol and Ha Kang [4], Vese and Osher [13], [18], [17] and more recently Bresson and Chan [5] and Duval et al. [7] 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 [6] 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.

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

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

Specific Treatment Each layer will be seen as an image compound of uniform elements (first layer is only composed of shapes and the second is only composed of textures). In the case of the regularized layer, we model by a law of Zipf the distribution of patterns.

3. The Regularized Layer - Shapes

3.1. Introduction

In this section we are going to recall what the Zipf law is and show how it can be involved in case of images and especially particular images which are the drop caps. We observe a Zipf law mixture is concerned and we compare the nature of the mixtures according to the nature of the image.

3.2. Zipf Law

Zipf law [19] is an empirical law expressed fifty years ago which relies on a power law. The law states that in phenomena figured by a set of topologically organized symbols, the distribution of the occurrence numbers of n-tuples named patterns is organized in such a way that the frequencies of the patterns M_1, $M_2...M_n$, noted N_1, $N_2...N_n$, are in relation with the rank of these symbols when sorted with respect to their occurrence frequency. The following relation holds:

$$N_\sigma(i) = k * i^a$$

$N_\sigma(i)$ represents the occurrence number of the pattern with rank i, k and a are constants. This power law is characterized by the value of the exponent a. k is more linked to the length of the symbol sequence studied. The relation is not linear but a simple transform leads to a linear relation between the logarithm of N and the logarithm of the rank. Then, the value of exponent a can be easily estimated by the leading coefficient of the regression line approximating the experimental points of the 2D graph (log10(i), log10($N_\sigma(i)$)) with i varying from 1 to n. Further on, the graph is called Zipf graph and can be illustrated by a curve as can be seen in Figure 3. One way to achieve the approximation of the graph is to use the least square method on the experimental points. As points are not regularly spaced, the points of the graph are re-scaled along the horizontal axis.

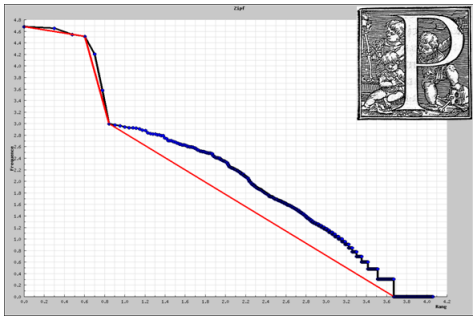


Figure 3. Example of a Drop Cap and its Zipf plot where are indicated the different straight zones extracted

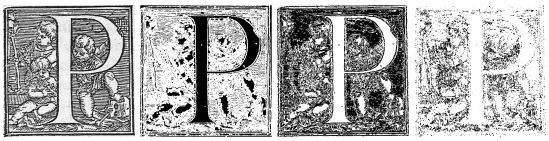
3.3. Image Application Layer Extraction

In this section, we point out some problems that may occur when images are concerned. In the case of the mono dimensional data, the masks used by linguists were limited to successive symbols. When images are concerned, the masks have to respect the topology of the 2D space the data is imbedded in. We have chosen to use 3x3 masks as a neighborhood of a pixel in the 2D space.
Then, the principle remains the same, the number of occurrences of each pattern is computed. Nevertheless as 256 symbols are used to code pixels, there would be theoretically 256^9 different patterns. This number is much larger than the number of pixels in an image. 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. For example a 640x480 image contains only 304964 patterns. Then it is necessary to restrict the number of perceived patterns to give sense to the model. The coding is decisive in the matter.
To decrease the number of patterns, the number of grey levels in the image can be decreased without loosing too much information. We have made use of k-means clustering algorithm [11]. As the images we are dealing with in this study are binary images, we decided to keep only 3 grey levels. This step is essential to build the Zipf curve associated with a drop cap. In the Figure. 3 we show an example of a Zipf curve.
As a matter of fact we see in Figure 3 that Zipf law does not hold. Indeed, the curve cannot be reasonably modeled by a straight line. Nevertheless, we can observe three different segments can be extracted to model the curve. Then, according to the frequency of the pattern, we can distinguish three sets of patterns for which Zipf law holds. Of course three different Zipf laws are involved. We can find an interpretation to the these sets. The first set, associated with the left part of the curve comprises the most frequent patterns, it is associated with texture patterns and with region zones. The two other sets are associated with patterns involved in the contours. You can see in the Figure 4 the example of four drop cap in which we have extracted the three layers. The image named first layer is composed with pixels, center of patterns involved in the first Zipf law, the layer 2 with pixels linked to the second power law and the layer 3 with pixels associated with the third Zipf law.
The method characterizes not only the image global appearance but also its structural composition, as shown in Figure 4. We can observe in the first layer mostly the letter itself but also large areas appearing in the background. The second layer is made of the thick contours and the third with the thin ones. The first layer seems really interesting to reach our goal.

4. Interpretation and Segmentation

4.1. Shapes Segmentation

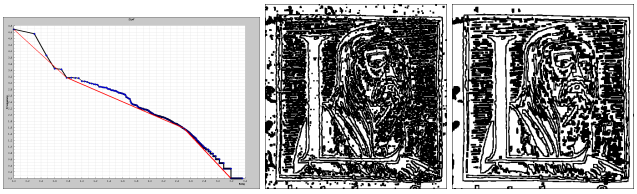
With the Zipf's curve extracted from the image (see Fig. 5(a)), three lines are computed to estimate the principal parameters (slopes) of the Zipf's law. The first of these straight lines, which corresponds to the most com-



(a) Image (b) 1_{st} layer (c) 2_{nd} layer (d) 3_{rd} layer

Figure 4. Example of layers extracted from Zipf Law decomposition

mon patterns, represents shapes in an image. We test this method on original image (see Fig. 5(b) and on the shapes'layer obtained by the Aujol and Chambolle algorithm (see Fig. 5(c)). It realizes a background-foreground segmentation by tagging elements of image which patterns are frequent. We can notice that using the shapes' layer of Aujol and Chambolle algorithm permits to reduce noise and to obtain bigger and better shapes in term of recognition.



(a) Example of Zipf's Law extracted from an image (b) Elements of first slope from original image (c) Elements of first slops from shapes' layer of Meyer

Figure 5. Examples of automatically extracted shapes from drop caps

4.2. Letter Extraction

Once all the shapes have been extracted, we will seek the various connected components of the image. Each connected component corresponds to a particular shape like the head, a shoulder for images containing humans, flowers, or the most important of them : the letter. A selection of these connected components, based on criteria of size, position of the center of mass and distance to the edges, will enable us to obtain the letter. Indeed, letter corresponds to the largest connected component whose center of mass is centered in the image and which does not touch the edge of the image. That can be explained by the fact that letter is in the middle, is one of the biggest shapes and surrounded by the background. Some examples of extraction of letters can be observed in Figure 6.

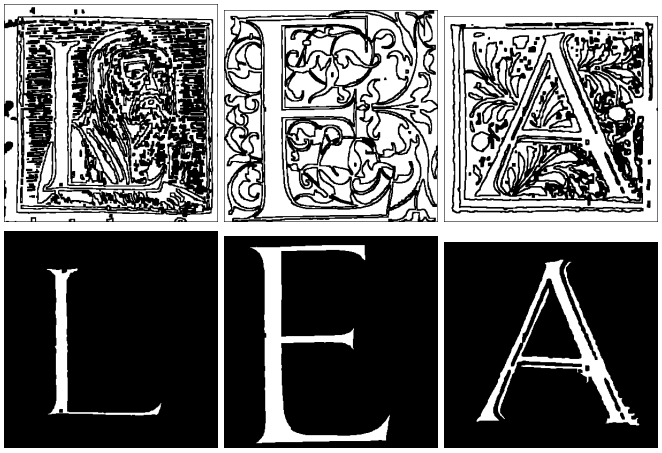


Figure 6. Examples of automatically extracted letters from drop caps

5. 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 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

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

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

our system fails correspond to very difficult images, as one can see an example in Figure7.

(a) Original lettrine (b) Letter extracted

Figure 7. An example of very difficult letter extraction

6. Conclusions

This paper presents a new method to extract letter in drop caps. It relies on a combination of two decompositions. The first one, an Aujol and Chambolle algorithm, 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 permits to extract the letter itself. The first experimentations to recognize letter using two different OCR are promising and will be improved in futur works. To improve the processing chain, we will try to ameliorate the criterion of connected-component selection. The better the connected component will be selected, the better the letter will be extracted then the better will be the recognition rate.

References

[1] J. F. Aujol, G. Aubert, L. B. 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, Jan. 2005.

[2] J.-F. Aujol and A. Chambolle. Dual norms and image decomposition models. *International Journal of Computer Vision*, 63(1):85–104, 2005.

[3] J.-F. Aujol, G. Gilboa, T. Chan, and S. Osher. Structure-texture image decomposition - modeling, algorithms, and parameter selection. *International Journal of Computer Vision*, 67(1):111–136, 2006.

[4] J.-F. Aujol and S. H. Kang. Color image decomposition and restoration. *J. Visual Communication and Image Representation*, 17(4):916–928, 2006.

[5] 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.

[6] 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.

[7] V. Duval, J.-F. Aujol, and L. Vese. A projected gradient algorithm for color image decomposition. Technical report, CMLA Preprint 2008-21, 2008.

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

[9] S. Jouili and S. 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.

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

[11] J. B. McQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*.

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

[13] 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.

[14] R. Pareti and N. 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.

[15] 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, Oct. 2005.

[16] S. Uttama, P. Loonis, M. Delalandre, and J.-M. Ogier. Segmentation and retrieval of ancient graphic documents. In *GREC*, pages 88–98, 2005.

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

[18] 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, Jan. 2004.

[19] G. Zipf. *Human Behavior and the Principle of Least Effort*. Hafner Pub. Co, 1949.