# Information Extraction from Old Images of Documents for Indexing

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

# Information Extraction from Old Images of Documents for Indexing

Mickael Coustaty and Jean-Marc Ogier

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

**Abstract.** This paper present 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.

**Key words:** Zipf Law, Decomposition, Connected Components, Information Extraction

## 1 Context

The french collaborative project *NaviDoMass*, financed by the National French Research Agency, challenge to index, to preserve and to provide public accessibility to ancient documents. The main interest of this study is based on specific graphics called drop caps, and on the extraction of shapes in them. This work is inspired by [PV06] and [ULDO05] which used a Zpif law and a Wold decomposition to extract elements of drop caps.

**Fig. 1.** An example of drop cap

*Drop caps* are heterogenous, decayed by time and composed of two main elements: the letter and the background (See Figure 1). They are complex to recognize due to the variability of representations, the mix of foreground and background and the degradations of ink and paper. This article present the differents steps of our method: 1) Simplification of images using layers 2) Extraction of shapes 3) Selection of extracted shapes.

## 2 Our method

*Simplification using layers to extract signatures* Drop Caps images are principally composed of strokes which make usual textures algorithms unsuitable. We thus used an approach developped by Dubois et Lugiez [DLPM08] to separate three layers of informations easier to treat.

*Layers in details* Decomposition used relies on the minimization of a functionnal: $F(U, V, W)$ where each parameter represent one of the three following layers:

- U Regularized layer which contain areas of image with low variations of graylevel. It permit to highlight shapes of an image
- V Second layer contain elements that oscillate quickly. In our case, this layer permit tot highlight textures of Drop Caps
- W The last layer permit to retrieve all that does not belong to the two first. Thus, one finds all that corresponds to noise, problem of overprinting, etc

*Adapted treatment* Each layer can be seen as a specific image with special caracteristics and treatments. On the regularized layer, a segmentation using a Zipf Law permit to segment shapes regardless of the color of background and foreground.

### 2.1 Regularized layer - Shapes

Regularized layer obtained by decomposition contain all shapes. An extraction of them permit to select the most interesting and to extracct information from drop caps. *Zipf Law* was empiricallydefined by *George Kingsley Zipf*. This law is based on frquency 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 composed of three steps:

- Simplification of image apllying 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

### 2.2 Simplification, patterns research and Letter extraction

*Simplification and pattern research:* A huge amount of three by three patterns are possible in an image of graylevel. We thus apply a 3-means on image's histogram of graylevel to simplify and to reduce the number of possible patterns. A simple count of each pattern permit to know their frequency and their rank.

*Zipf curve calculation:* From the frequencies and the ranks, a Zipf curve is calculated according to 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 first one, which correspond to the most frequent patterns, represent shapes of image (uniform areas). We thus binarize image by separating shapes from background.

*Shapes extration:* Once shapes have been extracted, one can seek connected components of binarized image. A selection of these connected components, based on size, location, center of mass and exccentricity, permit to obtain region of interest of drop caps. The most important connected component, fro historian, is 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. Experimentations are in progress and an example of result can be seen in Figure 1.

## 3 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 in progress and we will try to

## References

[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.

[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.

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