Extraction d'informations d'images de documents anciens

Information Extraction from Old Documents Images

Mickaël Coustaty[1] Jean-Marc Ogier[1] Rudolf Pareti[2] Nicole Vincent[2]
[1] Laboratoire d'Informatique, Images et Interactions - L3i - Equipe IMEDOC
Avenue Michel Crépeau
17042 LA ROCHELLE Cedex 1
{mcoustat, jmogier}@univ-lr.fr
[2] Laboratoire d'Informatique de Paris 5 - CRIP5 - Equipe SIP
45, rue des Saints-Pères
75270 Paris Cedex 06
nicole.vincent math-info.univ-paris5.fr

Résumé

Cet article propose une méthode d'extraction d'information à partir d'images de documents anciens. Elle repose sur un processus en trois étapes : 1) une décomposition en couches 2) une segmentation avec une loi de Zipf 3) une sélection des composantes connexes pour ne faire ressortir que l'information recherchée.

Mots Clef

Documents Anciens, Lettrines, Segmentation, Decomposition, Loi de Zipf

Abstract

This paper presents a new approach to information extraction in old document images. This approach relies on a three steps process : 1) a decomposition in layers 2) a segmentation with a Zipf law 3) a selection of connected components to bring out the information we are looking for.

Keywords

Old documents, Ornemental Letters, Segmentation, Decomposition, Zipf Law

1 Contexte/NaviDoMass

1.1 Enjeux

Le projet *NaviDoMass [2], soutenu par l'*Agence Nationale de la Recherche, réunit 7 laboratoires français autour du problème de reconnaissance et d'indexation de documents anciens. Ce projet vise à dématérialiser ces documents pour :

- les protéger
- permettre un accès depuis n'importe où
- permettre des consultations simultanées
- naviguer facilement et rapidement dans de grandes masses de données

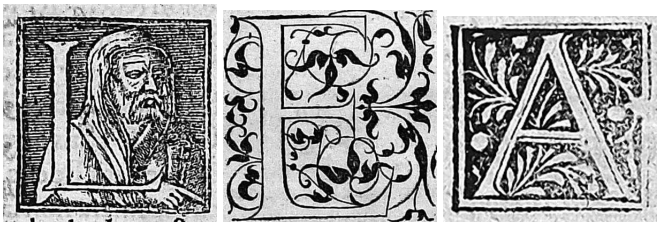

FIGURE 1 – Exemples de lettrines

1.2 Les lettrines en particulier

Les images de documents du patrimoine sont très hétérogènes et endommagées par le temps. Les lettrines (lettres majuscules ornementales) font partie des images à traiter et à indexer. Ces images sont composées de deux éléments principaux : la lettre et les motifs. (Voir Figure 1). L'une des étapes importante dans le processus de reconnaissance des lettrines consiste à segmenter la lettre et les éléments du motif pour les caractériser, extraire une signature. Cette signature permettra une comparaison simple et rapide pour notre processus d'indexation de grandes masses de données. La suite de l'article présente les différentes étapes de notre méthode : 1) Simplification des images à l'aide de couches 2) Extraction de formes à partir d'une de ces couches 3) Séléction de ces formes.

2 Décomposition en couches dans le but d'extraire une signature

2.1 Stratégie retenue

Les images de lettrines sont très complexes de par la masse d'information qu'elles renferment, il y a donc nécessité de les simplifier. Ces images sont principalement composées de traits qui rendent les méthodes de textures usuelles inadaptées. Nous avons donc utilisé l'approche développée

par Dubois et Lugiez [1] pour séparer l'image en plusieurs couches d'informations plus simples à traiter.

2.2 Les couches en détails

La décomposition utilisée repose sur la minimisation d'une fonctionnelle : $\inf F(u, v, w)$
où chaque paramètre de la fonctionnelle représente une des trois couches suivantes :

- La couche régularisée correspond aux zones de l'image qui présentent de faibles variations de niveau de gris. Elle nous permet de mettre en évidence les formes de l'image.
- La seconde couche quant à elle correspond à tout ce qui varie rapidement dans l'image. Dans notre cas, cette couche permet de faire ressortir les textures des lettrines.
- Une troisième et dernière couche permet de récupérer tout ce qui n'appartient pas aux deux premières. Ainsi, on retrouve tout ce qui correspond aux bruits, au problème de surimpression dû au vieillissement du papier, etc.

Traitement adéquats Chaque couche va ainsi pouvoir être traitée comme une image spécifique avec les traitements qui lui sont propres. Dans le cas de la couche régularisée, nous modélisons par une loi de Zipf la distribution des motifs.

3 Couche Régularisée - Formes

La couche régularisée obtenue après la décomposition contient toutes les formes de l'image. Une segmentation permet de sélectionner les plus intéressantes et d'extraire de l'information des lettrines.

3.1 Loi de Zipf

La loi définie empiriquement par *George Kingsley Zipf* se base sur la fréquence et le rang d'apparition des mots dans un texte. Cette loi peut-être appliquée sur l'image en prenant comme motifs des imagettes de l'image et en calculant leur fréquence et leur rang. Cette recherche va nous servir pour guider la segmentation.

3.2 Application à l'image - Extraction des motifs les plus fréquents

La méthode que nous avons utilisée pour guider la segmentation des formes de la couche régularisée a été développée par Pareti [3] et est composée de trois étapes :

- Simplification de l'image en appliquant un 3-means sur son histogramme des niveaux de gris afin de réduire le nombre de motifs
- Recherche des motifs de taille 3x3 dans l'image pour obtenir leur fréquence et leur rang
- Séparation des motifs en trois groupes en fonction de la loi d'évolution de la fréquence par rapport à leur rang

3.3 Simplification et recherche des motifs

Il existe un grand nombre de motifs 3x3 possibles dans une image en niveaux de gris. L'idée est donc d'appliquer un 3-means sur cette image pour la simplifier et réduire le nombre de motifs. Un simple dénombrement de chacun de ces motifs va permettre de connaître leur fréquence et leur rang sous forme d'une courbe de Zipf.

FIGURE 2 – Exemples de lettres extraites automatiquement à partir des lettrines

3.4 Segmentation fond/formes

A partir de la courbe de Zipf extraite de l'image, 3 droites sont calculées pour estimer ses 3 principaux paramètres des lois de Zipf qui interviennent. La première de ces droites, qui correspond aux motifs les plus présents, représente les formes de l'image (les zones uniformes). Elle nous permet donc de binariser l'image en séparant le fond des formes.

3.5 Extraction de la lettre

A partir des formes extraites, on recherche les différentes composantes connexes de l'image précédemment obtenue. Une sélection de ces composantes connexes va nous permettre d'obtenir la lettre puisqu'elle correspond à la plus grande composante connexe dont le centre de gravité est centré dans l'image et qui ne touche pas le bord de l'image. Quelques exemples d'extraction de lettres peuvent être observés dans la Figure 2. Les experimentations sont en cours et les résultats sont à venir.

Références

[1] 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 MCSŠ08 Final Program and Proceedings*, pages 394–398, 2008.

[2] NAVIgation into DOcuments MASSes. http ://l3iexp.univ-lr.fr/navidomass/, 2007-2010. Projet ANR de sauvegarde et d'indexation du patrimoine historique français.

[3] Rudolf Pareti and Nicole Vincent. Ancient initial letters indexing. In *ICPR '06*, pages 756–759, Washington, DC, USA, 2006. IEEE Computer Society.