Extraction of metadata related to ”image” and ”structure” of old documents

Maroua Mehri, Petra Gomez-Krämer, Pierre Héroux, Rémy Mullot

Extraction of metadata related to "image" and "structure" of old documents

Maroua MEHRI[a,b], Petra GOMEZ-KRÄMER[a], Pierre HÉROUX[b] and Rémy MULLOT[a]
{a}L3I, University of La Rochelle, France
{b}LITIS, University of Rouen, France
{**maroua.mehri**, petra.gomez, remy.mullot}@univ-lr.fr and pierre.heroux@univ-rouen.fr

Context

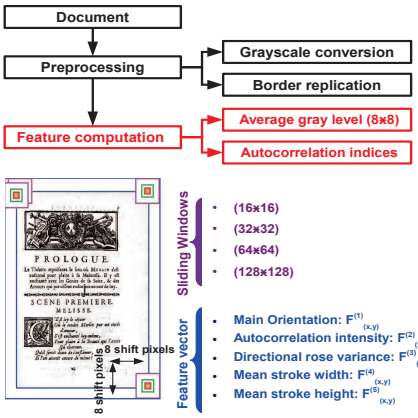
Our work is a part of the **DIGIDOC** project (Document Image diGitisation with Interactive DescriptiOn Capability). DIGIDOC is funded by **ANR** (French National Research Agency).

Objective

- Control the quality of old document image digitization
- Construct a computer-aided categorization tool of pages
- Characterize the document's content using intermediate level metadata
- Provide a similarity measure between pages

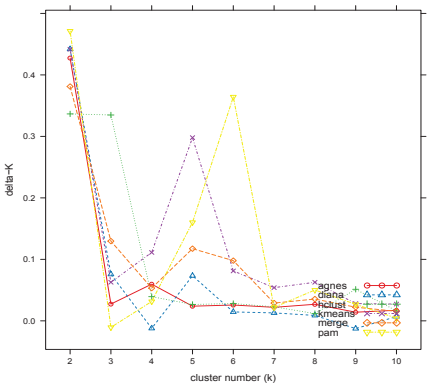
1. **Extraction of descriptors per block**
2. **Generation of topological signature**

Texture features



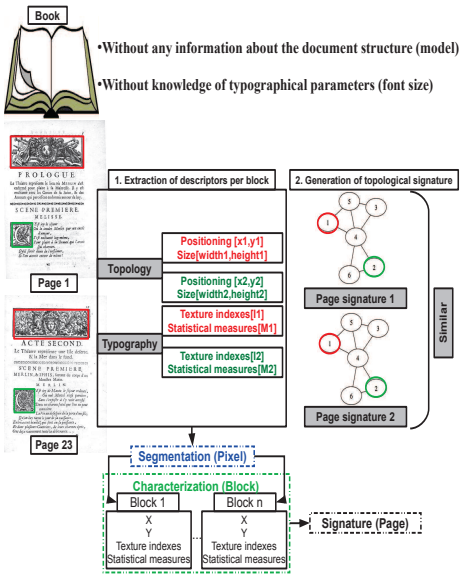
We propose a feature vector composed of texture indices, all based on the **autocorrelation function** and the **directional rose**[1,2].

Consensus clustering



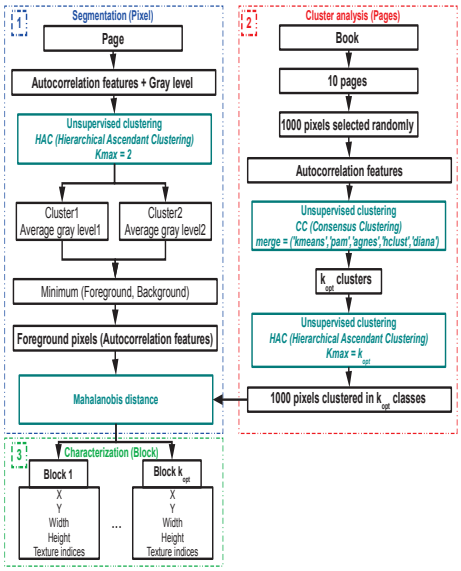
The optimal cluster number corresponds to the largest change in the area under the cumulative density curve Δk for the merge consensus matrix[3].

Overview



Our goal is to propose a set of metadata characterizing the **physical structure of pages** in terms of **homogeneous areas** and **topological relationships**.

Proposed approach



We use **non-parametric unsupervised clustering** for the computed features to determine the homogeneous regions defined by similar texture indices.

Results

Example of document segmentation. {**(a)**, **(d)**} are the original images of the same book, {**(b)**, **(e)**} are the results of the foreground pixel extraction using an adapted Hierarchical Ascendant Classification and {**(c)**, **(f)**} are the final results of clustering steps.

Future research

- Validation of autocorrelation indices
- Computation of frequency and statistical attributes and other texture features
- Definition of one or more signatures for each page (graph of homogeneous zones)

References

[1] Journet, N. *et al.* Document image characterization using a multiresolution analysis of the texture: application to old documents In *IJDAR'08*
[2] Ouji, A. *et al.* Chromatic/Achromatic Separation in Noisy Document Images In *ICDAR'11*
[3] Simpson, T. *et al.* Merged consensus clustering to assess and improve class discovery with microarray data In *BMC'10*