



HAL
open science

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

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

► To cite this version:

Maroua Mehri, Petra Gomez-Krämer, Pierre Héroux, Rémy Mullot. Extraction of metadata related to "image" and "structure" of old documents. Visual Recognition and Machine Learning Summer School (VRML), Jul 2012, Grenoble, France. hal-03167272

HAL Id: hal-03167272

<https://hal.science/hal-03167272>

Submitted on 11 Mar 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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



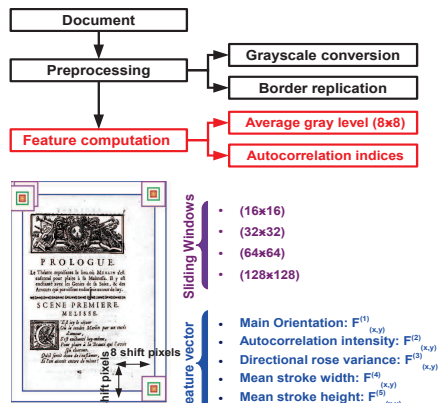
Our work is a part of the DIGIDOC project (Document Image diGitisation with Interactive DescriptiOn Capability). DIGIDOC is funded by ANR (French National Research Agency).

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

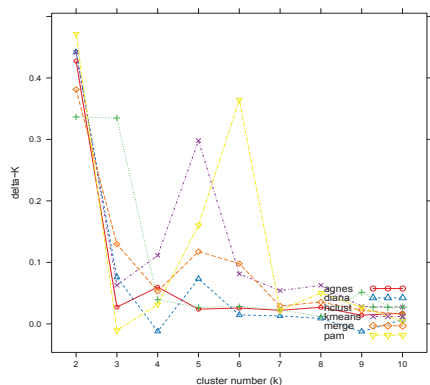
- Extraction of descriptors per block
- Generation of topological signature

TEXTURE FEATURES



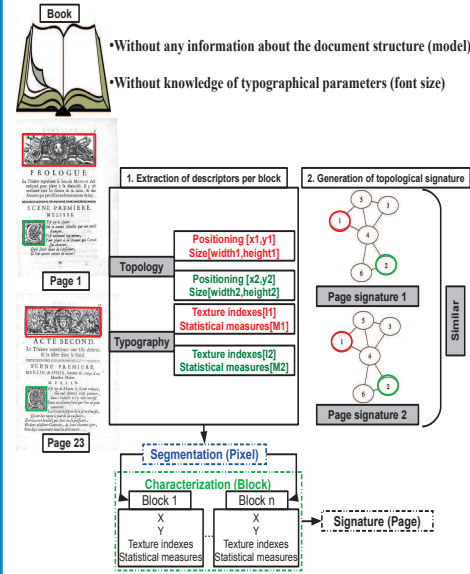
We propose a feature vector composed of texture indices, all based on the autocorrelation function and the directional rose^[1,2].

CONSENSUS CLUSTERING



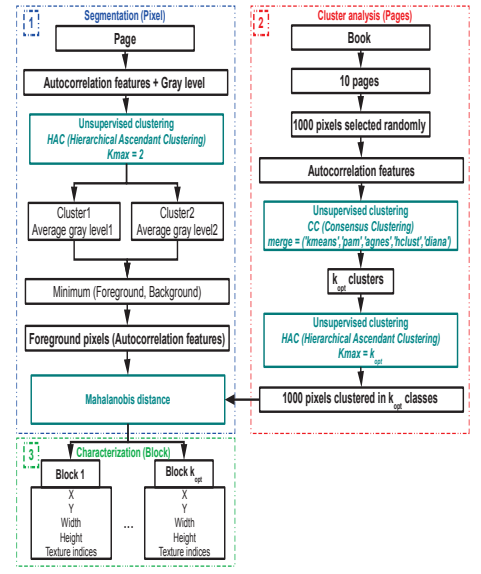
The optimal cluster number corresponds to the largest change in the area under the cumulative density curve Δk for the merge consensus matrix^[3].

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

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