

# Historical document image analysis: a structural approach based on texture

Presented by:

**Maroua Mehri<sup>a,b</sup>**

Supervised by:

**Pierre Héroux<sup>b</sup> , Petra Gomez-Krämer<sup>a</sup> and Rémy Mullot<sup>a</sup>**

<sup>a</sup>

L3i Laboratory, University of La Rochelle, La Rochelle, France

<sup>b</sup>

LITIS Laboratory, University of Rouen, Saint-Etienne-du-Rouvray, France

maroua.mehri@univ-lr.fr

Funded by: **ANR-DIGIDOC project**



# Outline

---

- Introduction
- Proposed approach : a structural signature based on texture
  - Digitized historical book page characterization
  - Digitized historical book page categorization
- Experiments
- Evaluation and Results
- Discussion
- Future Work

# Introduction – Context

---

- **Rapid growth of digital libraries worldwide**
  - access to large sets of cultural heritage documents
  - indexing tool of digitized resources
  - information retrieval in digital libraries
- **Historical document image analysis**
  - reliable historical document image (HDI) interpretation system
  - efficient and fast computer-aided indexing tool of HDIs
- **Objectives**
  - segment and characterize a HDI as easily, quickly and automatically as possible
  - find homogeneous regions (graphic and different font text regions)
  -  to characterize the HDI content and layout
  - propose a signature
  -  to categorize digitized historical book (DHB) pages according to several criteria (layout structure, graphical or typographical characteristics)

# Introduction – DIGIDOC (document image digitization with interactive description capability)

---

- design “smart” digitizers
- limit manual intervention
- perform easy and high quality digitization of HDIs
- assist the digitization operator to adjust the best set of parameters (e.g. resolution, lightening, color calibration),
- detect errors in the digitization process (e.g. blur, skewed or folded pages),
- provide appropriate assistance for document indexing (e.g. recognize page types or breaks in a sequence of pages), etc [LeBourgeois04]

[LeBourgeois04] LeBourgeois, F., Trinh, E., Allier, B., and Emptoz, H., "Document images analysis solutions for digital libraries", DIAL, 2004.

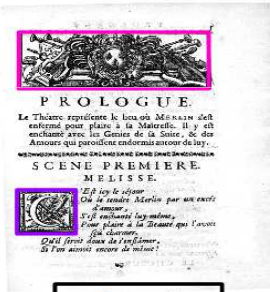
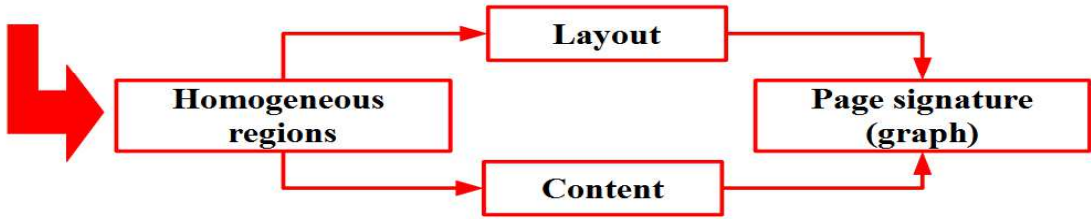
# Introduction – Goal

- identify similar groups of pixels sharing similar visual properties to characterize and compare the layout and content of HDIs by means of a graph-based signature

**Digitized historical book (DHB)**



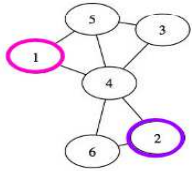
- Without any information about the document image layout
- Without knowledge of the document image content (typographical properties and graphical characteristics)



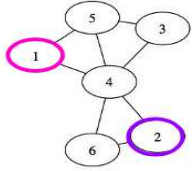
Page 1



Page 23



Signature 1



Signature 23

Similar

## Overview of the context and an example of signature-based applications

# Introduction – Historical document images



Some particularities of HDIs: superimposition of information layers (e.g. stamps, handwritten notes, noise, back-to-front interference) and page skew [Gallica]

[Gallica] French digital library Gallica: <http://gallica.bnf.fr>

# Introduction – Historical document images

---

- **Properties**
  - large variability of page layout, complex layouts, random alignment
  - use of multiple fonts and illustration styles
  - large variability of editorial style and logical structure
- **Life cycle**
  - noise and degradation caused by copying, scanning, or aging
  - superimposition of information layers (e.g. stamps, handwritten notes at the margins, noise, back-to-front interference)
- **Digitization**
  - page skew
  - capture defects (e.g. curvature, light), black borders, etc. [Coustaty11]

[Coustaty11] Coustaty, M., Raveaux, R., and Ogier, J. M., "Historical document analysis: A review of French projects and open issues", EURASIP, 2011.

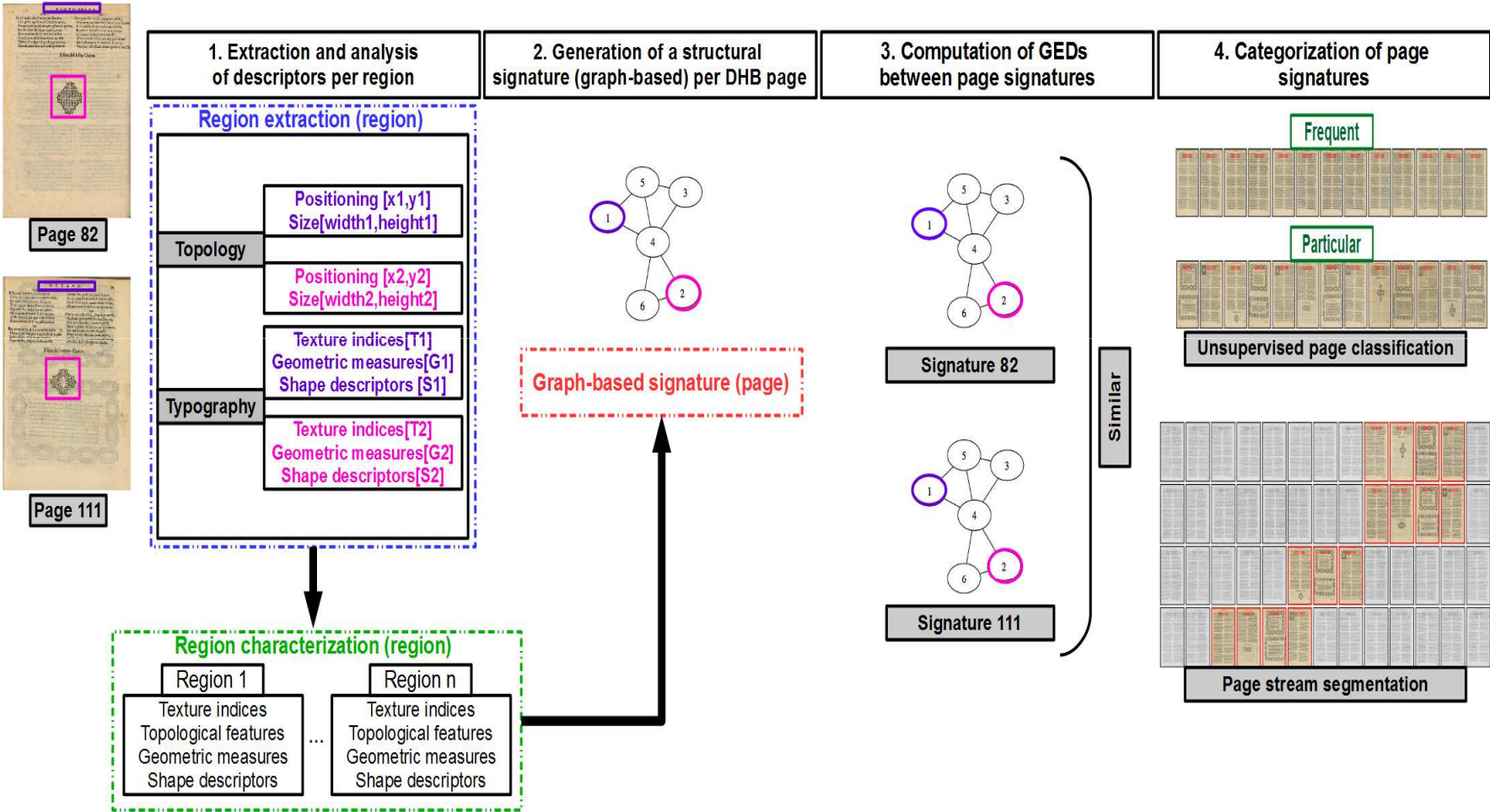
# Proposed approach

---

- a signature for DHB page characterization and categorization
- based on the use of texture and graph algorithms
- varying low-level features characterizing the HDI content (i.e. different text fonts, or graphic regions)
- structural information describing the HDI layout
- provide a rich and holistic description of the layout and content of the DHB pages
- applicable to a large variety of ancient books
- without hypothesis on the document layout and content



# Proposed approach



## Overview of the different steps of this work

# Proposed approach

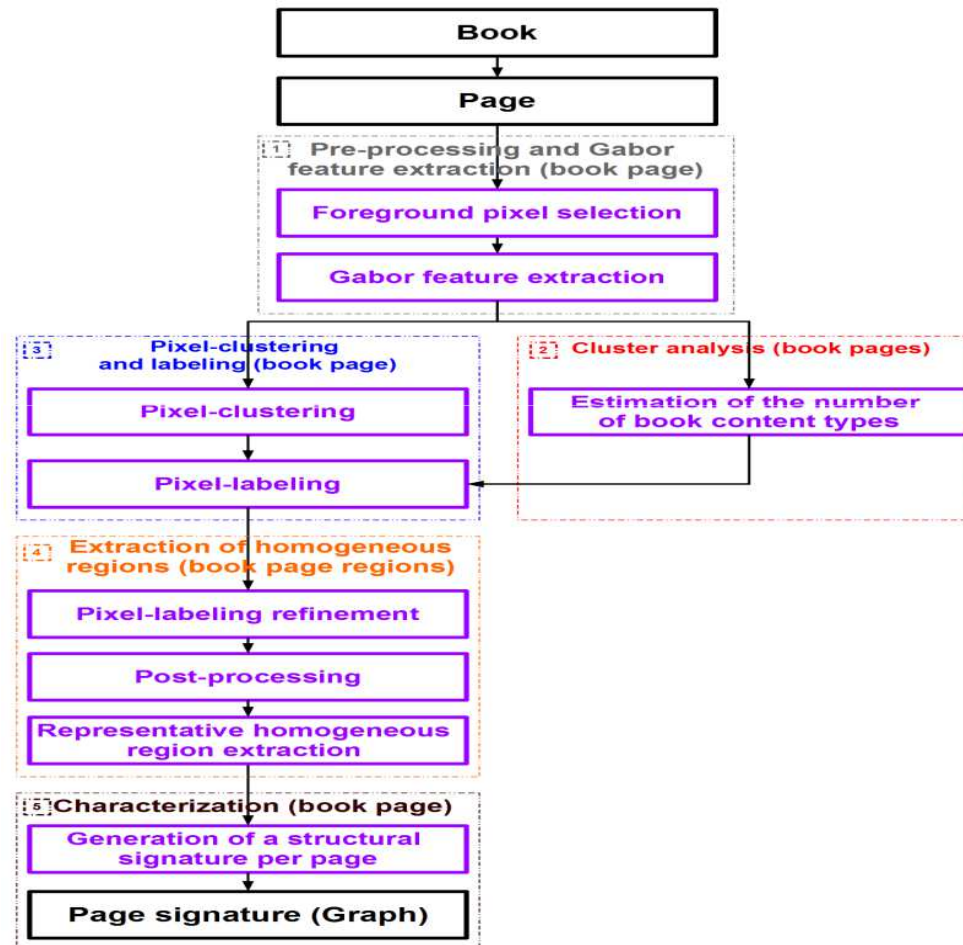
---

## 1. DHB page characterization

## 2. DHB page categorization

- a. Unsupervised DHB page classification
- b. DHB page stream segmentation

# DHB page characterization



Detailed schematic block diagram of the proposed approach used to generate the graph-based signature for DHB page characterization

# DHB page characterization

---

- Pre-processing and Gabor feature extraction (Step 1),
- Estimation of the number of DHB content types (Step 2),
- Pixel-clustering and labeling (Step 3),
- Pixel-labeling refinement (Step 4), [Mehri13, Mehri14, Mehri15]
- Post-processing (Step 5),
- Extraction of representative homogeneous regions (Step 6),
- Generation of a structural signature per page (Step 7).

[Mehri13] Mehri, M., Héroux, P., Gomez-Krämer, P., Boucher, A., and Mullot, R., "A pixel labeling approach for historical digitized books", ICDAR, 2013.

[Mehri14] Mehri, M., Mhiri, M., Héroux, P., Gomez-Krämer, P., Majoub, M. A., and Mullot, R., "Performance evaluation and benchmarking of six texture-based feature sets for segmenting historical documents", ICPR, 2014.

[Mehri15] Mehri, M., Gomez-Krämer, P., Héroux, P., Boucher, A., and Mullot, R., "A texture-based pixel labeling approach for historical books", PAA, 2015.

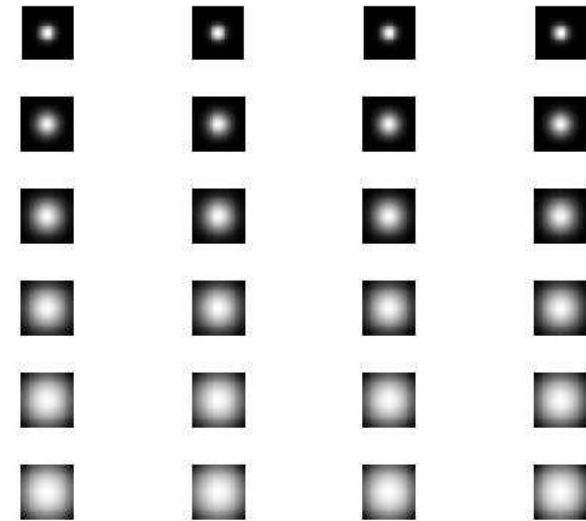
Step 1

# Pre-processing and Gabor feature extraction

# Multi-channel Gabor Filtering

---

- magnitude response of the output of Gabor filters
- space of Gabor filter is set as:  $\sigma = \sigma_x = \sigma_y$
- 6 spatial frequencies
  - $\{2\sqrt{2}, 4\sqrt{2}, 8\sqrt{2}, 16\sqrt{2}, 32\sqrt{2} \text{ and } 64\sqrt{2}\}$
- 4 orientations
  - $\{0, \pi/4, \pi/2 \text{ and } 3\pi/4\}$



**24 Gabor filters**

# Gabor feature extraction & Multi-scale Analysis

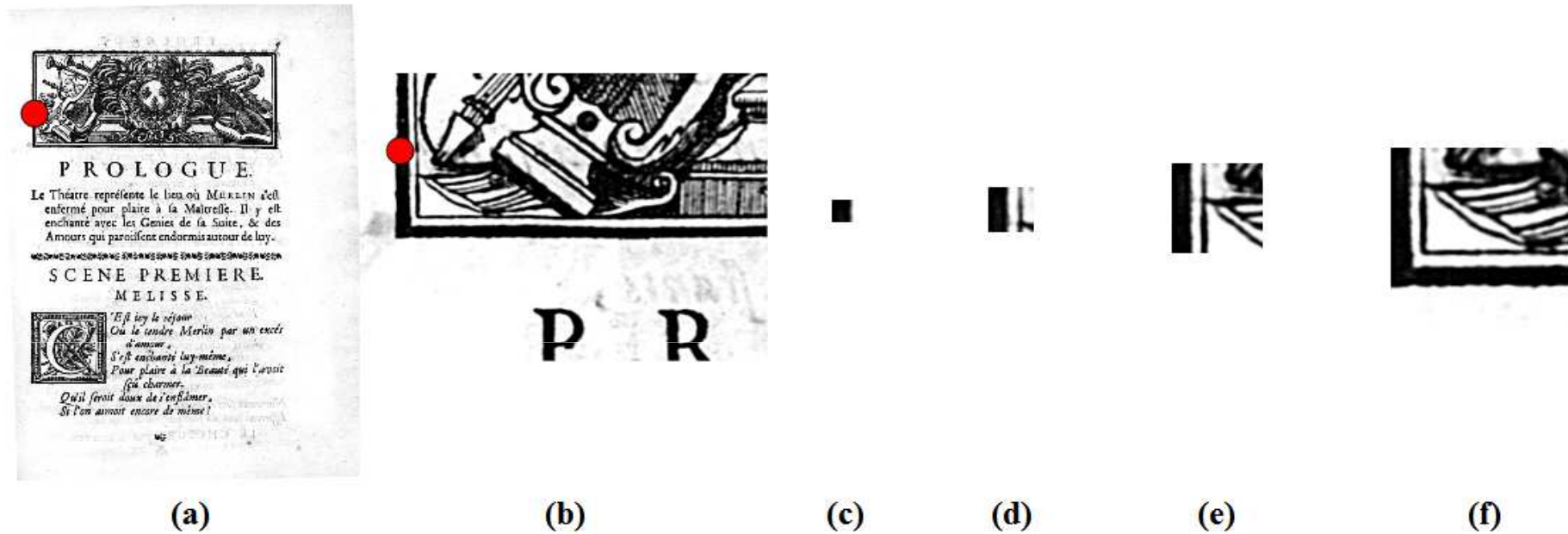


Figure (a) show the original image with a selected pixel position

Figure (b) zoom in selected region (red dot )

Figures (c), (d), (e) and (f) illustrate (16 x 16), (32 x 32), (64 x 64) and (128 x 128) windows

## 192-D feature vector:

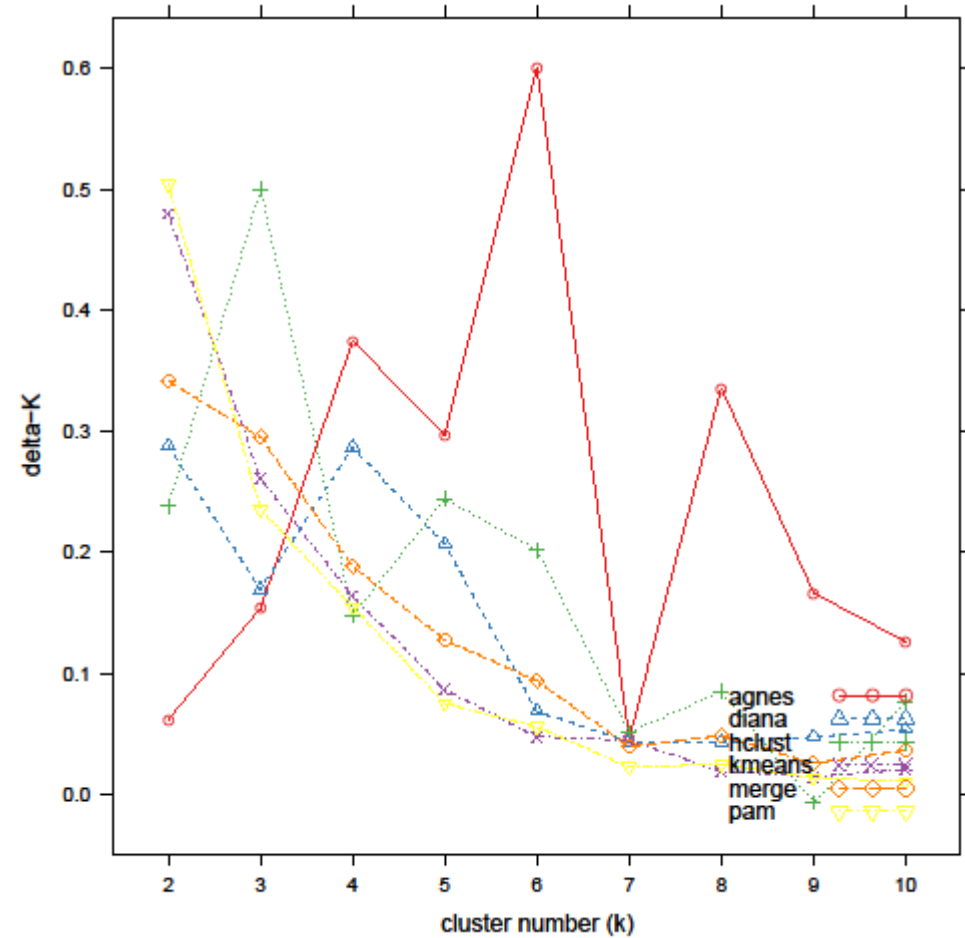
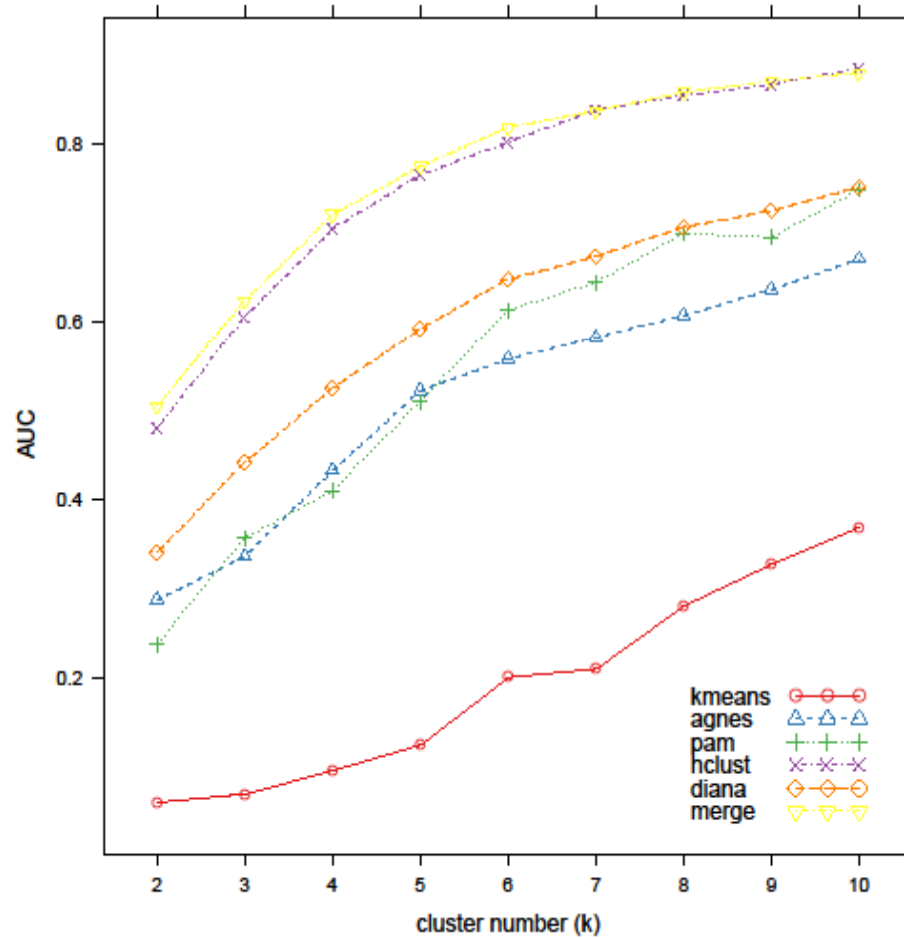
(6 spatial frequencies) X (4 different orientations) X (2 Gabor indices) X (4 sliding windows)

## Step 2

# Estimation of the number of DHB content types



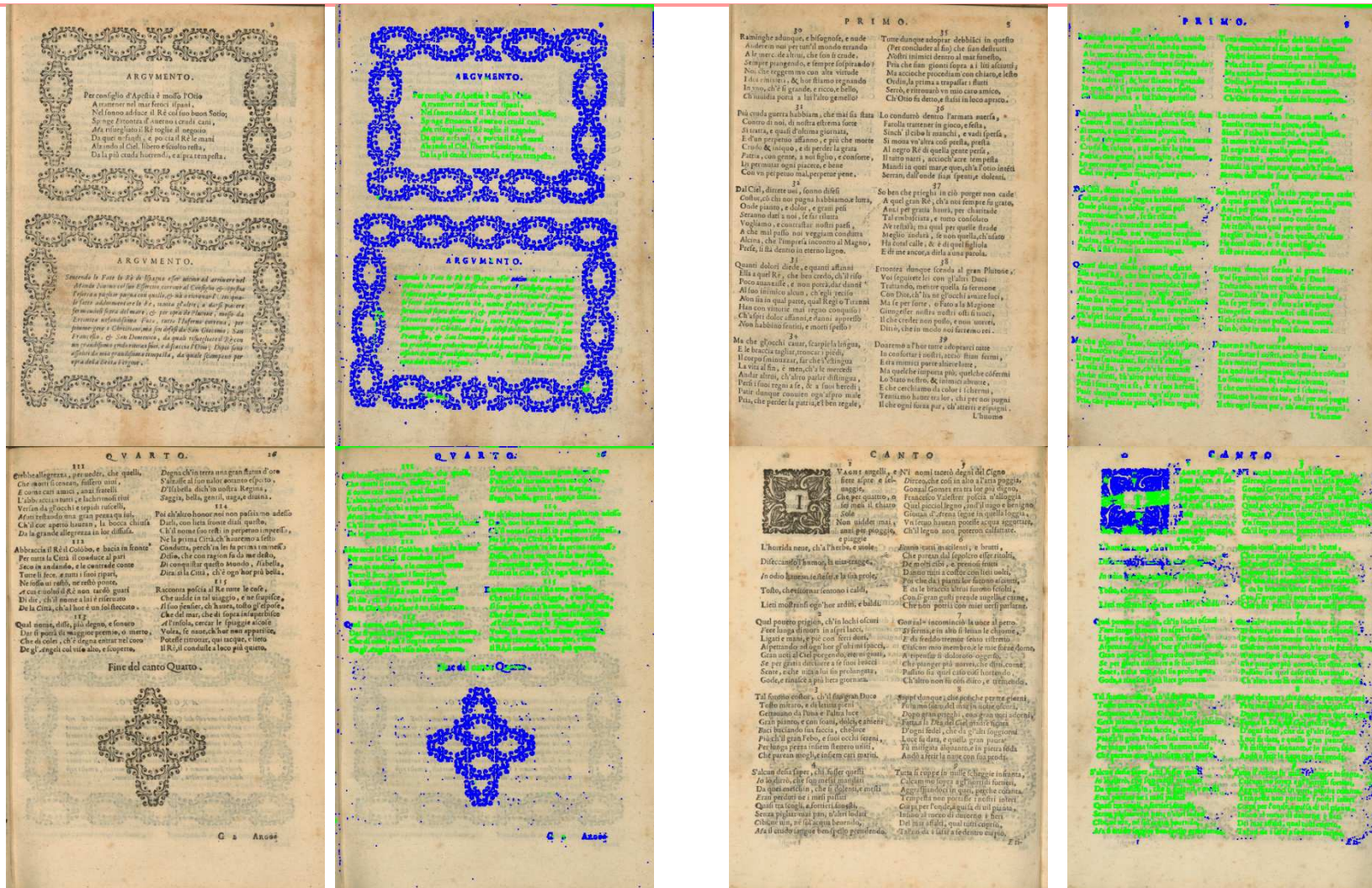
# Consensus Clustering



## Step 3

# Pixel-clustering and labeling

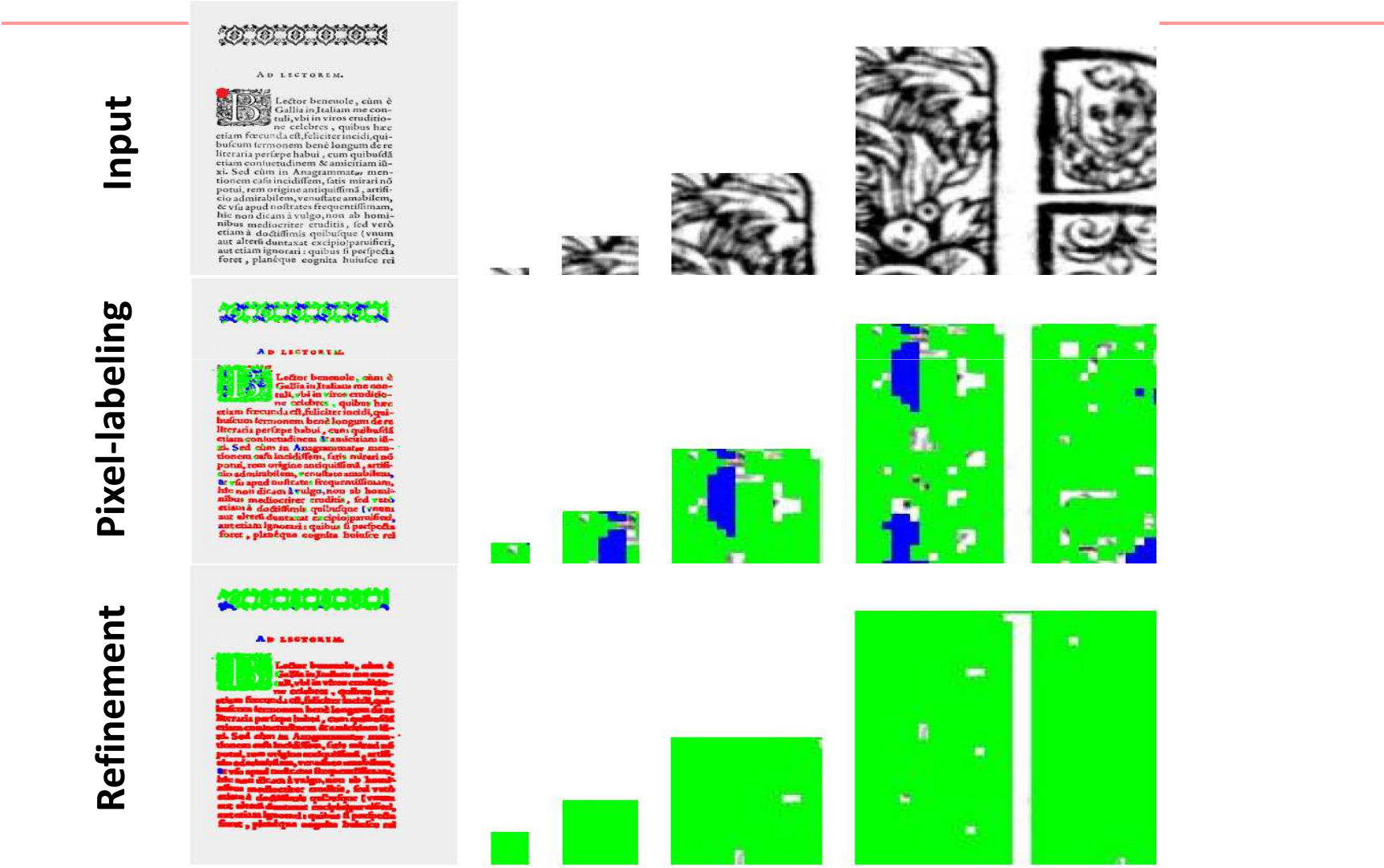
# Hierarchical Agglomerative Clustering (HAC) & Nearest Neighbor Search (NNS) algorithms



Step 4

## Pixel-labeling refinement

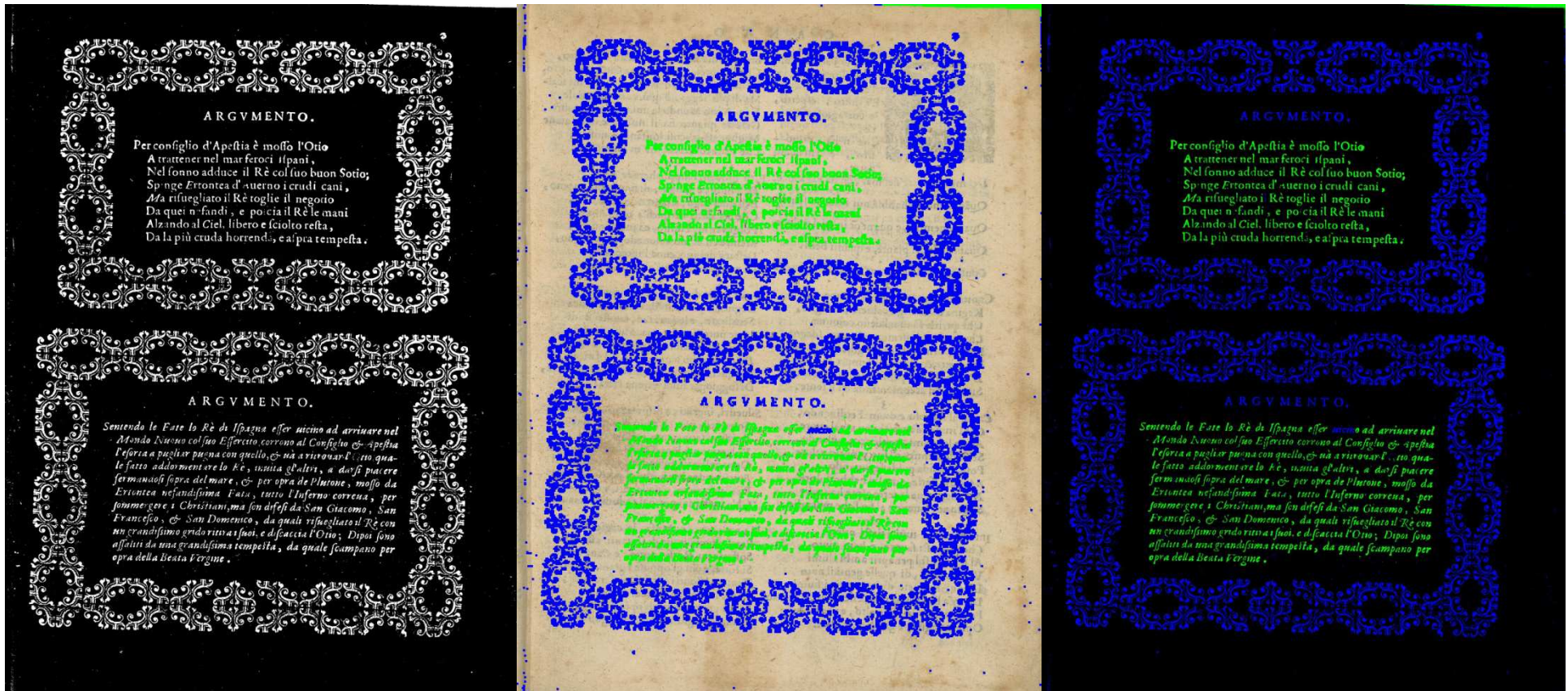
# Multi-Scale Majority Voting Technique



Step 5

Post-processing

# Connected component (CC) & Majority Voting (MV) analysis



Binarized image  
=> CC

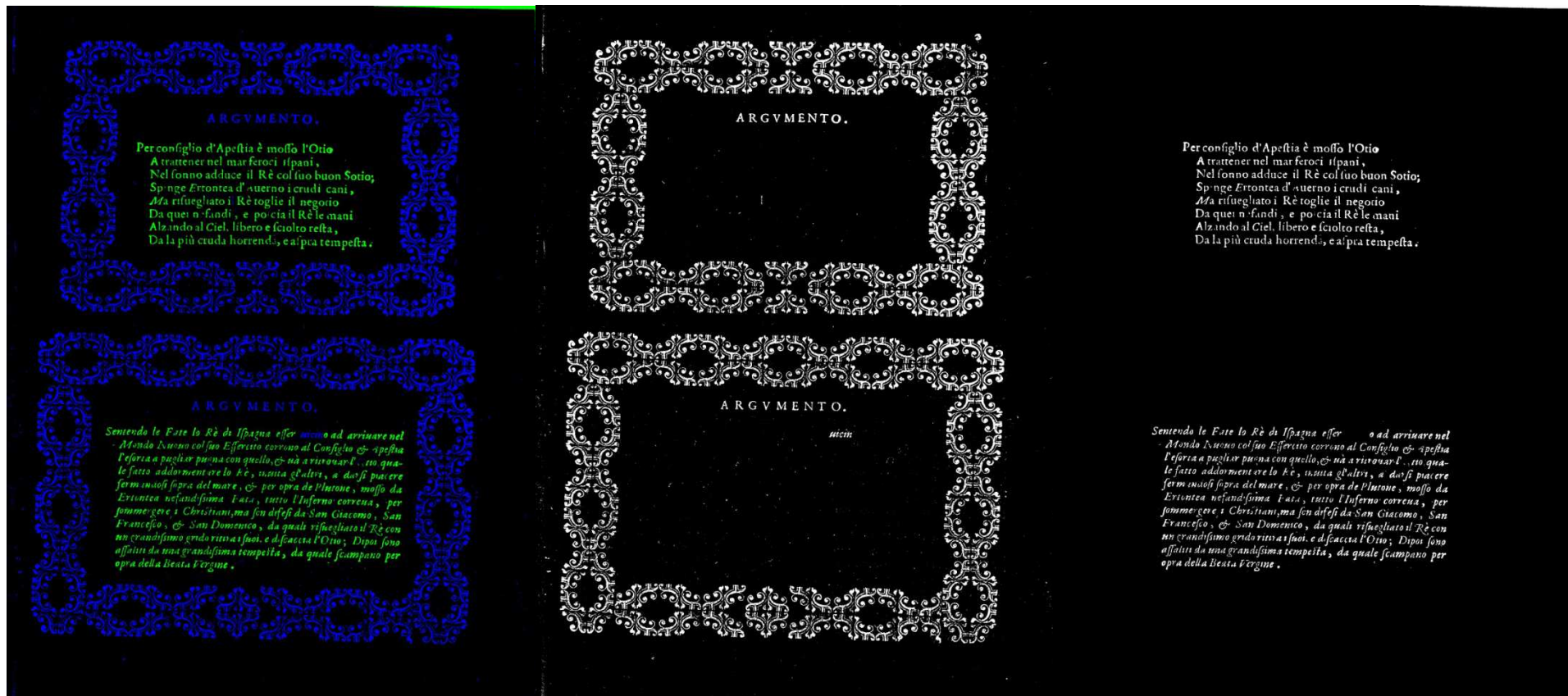
+

Refined pixel labeled image  
=> MV

=

Output image  
=> Labeled CCs

# CC & Color Layer Separation (CLS) analysis



Labeled CCs

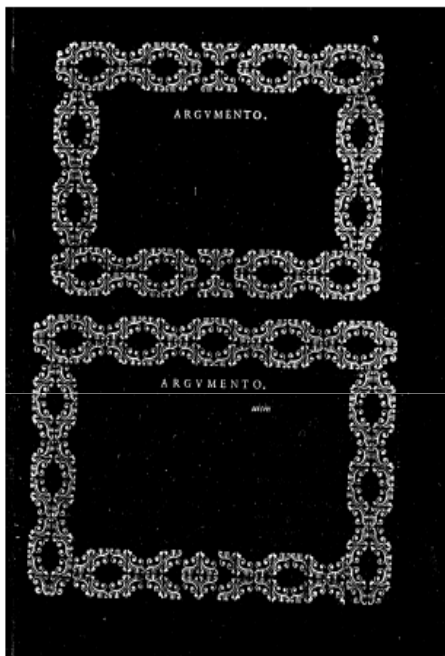
Blue CCs (graphics )

Green CCs (text)

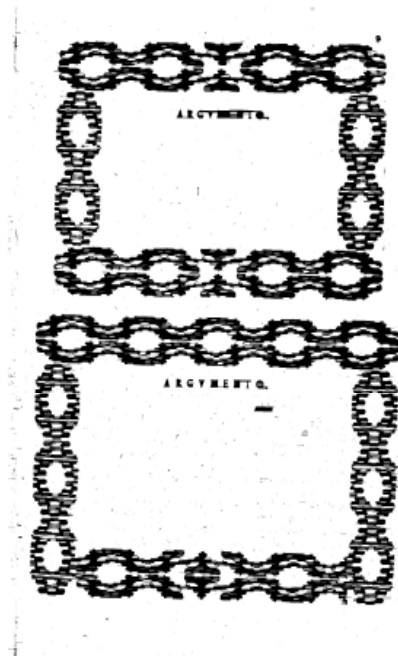


# Adaptive run-length smearing algorithm (ARLSA)

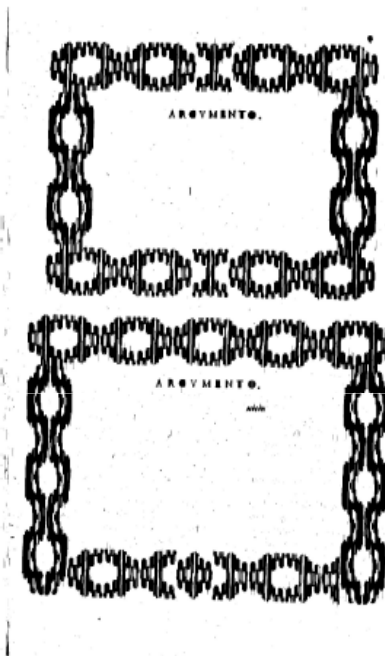
Blue CCs (graphics )



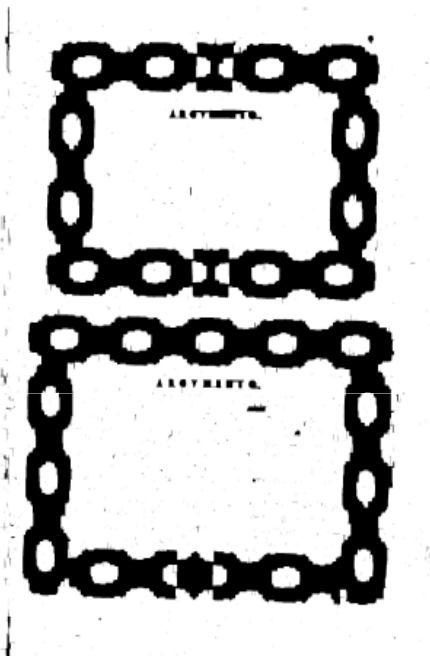
Blue CCs (graphics )



Horizontal direction **OR**



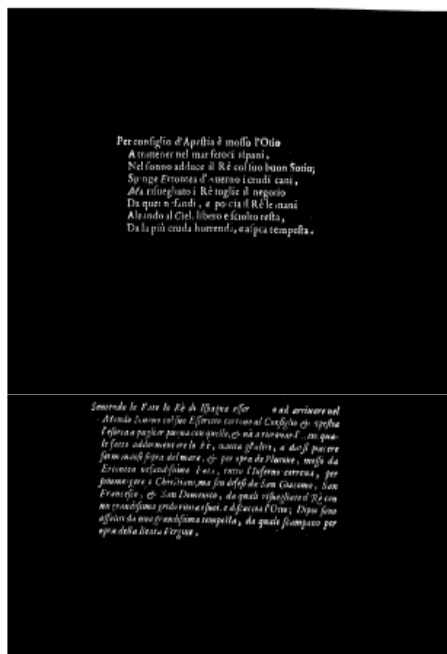
Vertical direction **=>**



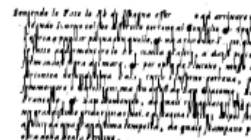
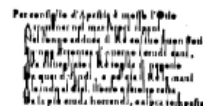
Resulting image of ARLSA on blue CCs

# Adaptive run-length smearing algorithm (ARLSA)

Green CCs (text )



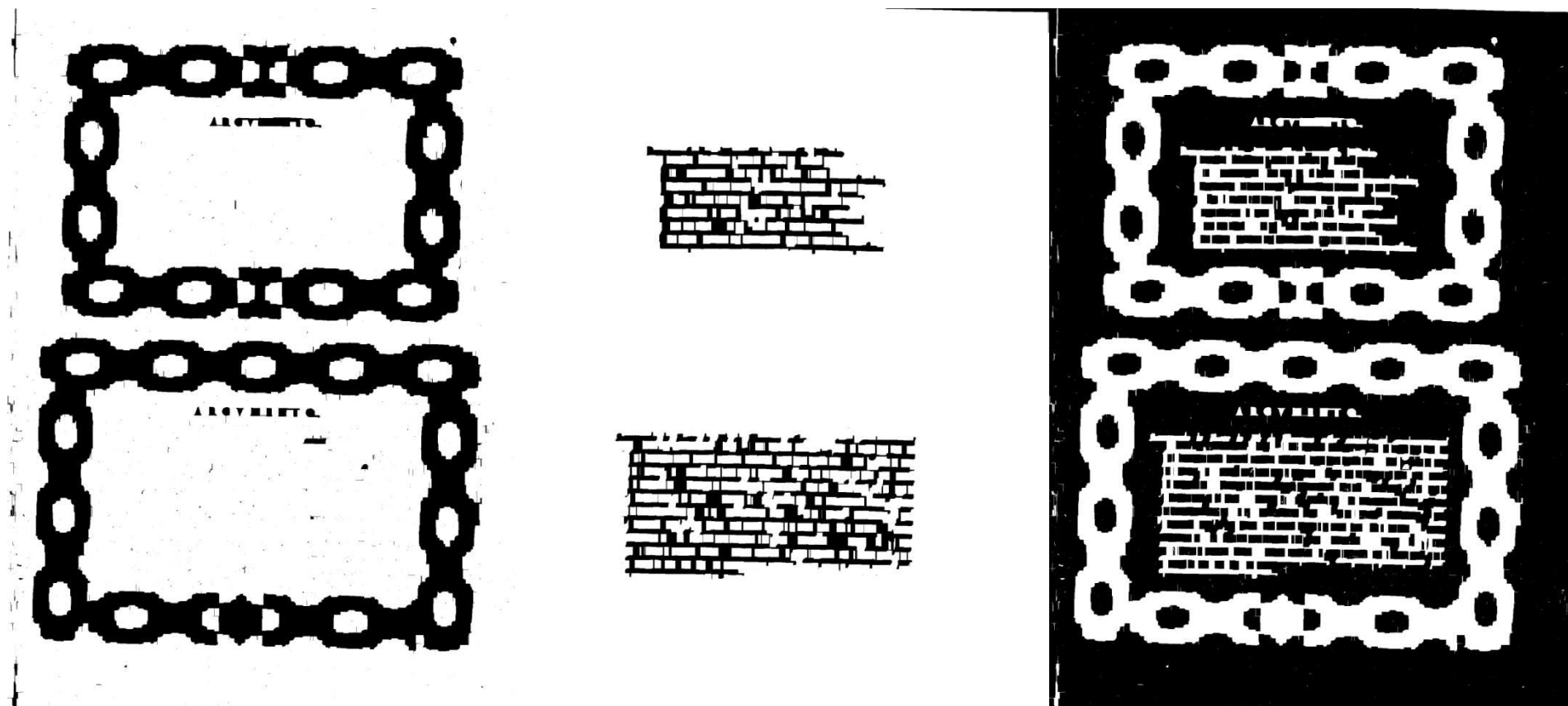
Green CCs (text )



Horizontal direction **OR** Vertical direction **=>** Resulting image of ARLSA on green CCs

# Adaptive run-length smearing algorithm (ARLSA)

Merge with OR



Resulting image of ARLSA on blue CCs

OR

Resulting image of ARLSA on green CCs

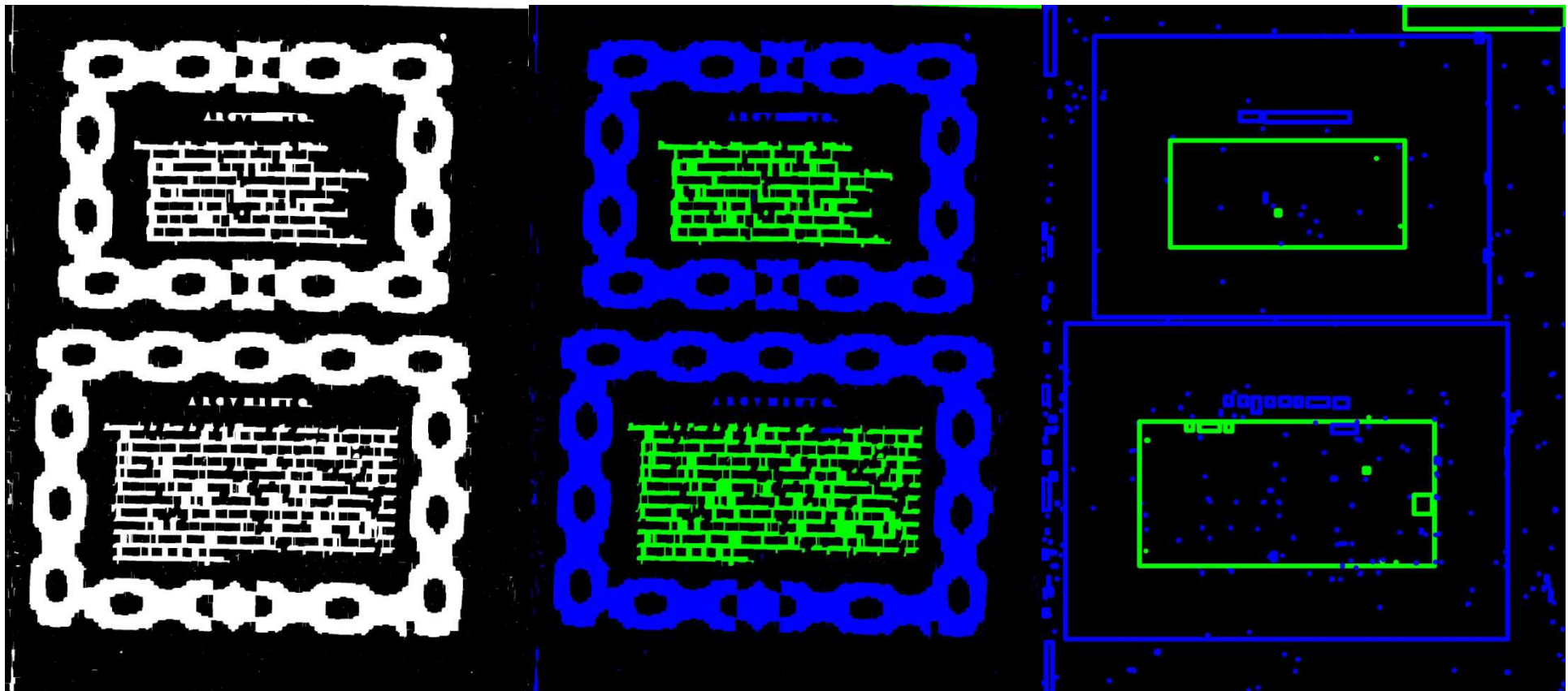
=>

Resulting image of ARLSA

## Step 6

# Extraction of representative homogeneous regions

# Extraction of Homogeneous regions



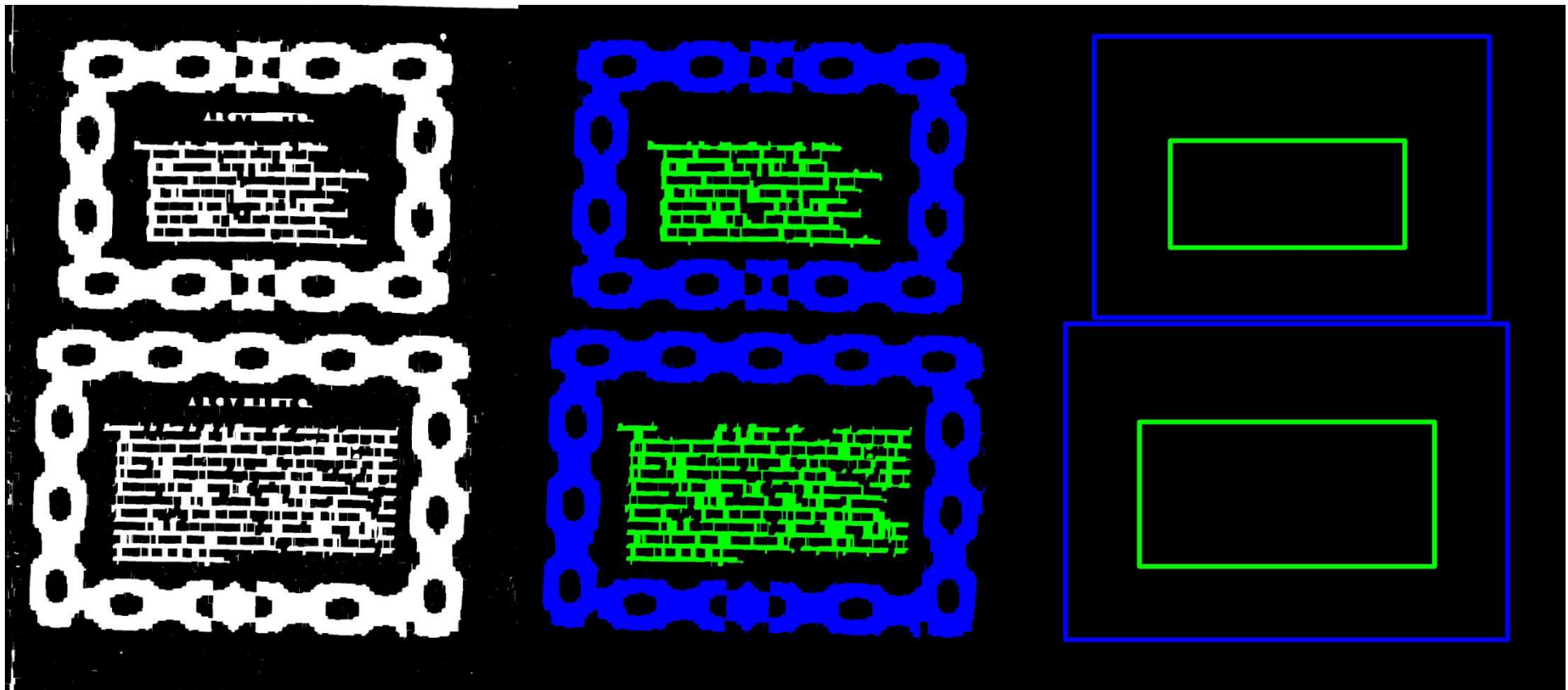
Resulting image of ARLSA  
=> Extracted CCs

+MV=>

Resulting image of MV  
=> Labeled CCs

Output image  
=> Extracted and labeled  
Homogeneous regions

# Extraction of Representative Homogeneous regions



Resulting image of ARLSA  
=> Extracted CCs

+MV=>

Resulting image of MV  
=> Labeled CCs

Extracted and labeled  
representative  
homogeneous regions

## Step 7

# Generation of a structural signature per page

# Graph-Based Signature – Approach

---

- vertex attributes
  - 192 Gabor attributes
  - 46 shape, geometric, and topological attributes
    - ✓ centroid position,
    - ✓ number of pixels,
    - ✓ gray-level average,
    - ✓ contour area and perimeter,
    - ✓ Hu, spatial, central, and central normalized moments, etc.
- edge attributes
  - absolute differences between the two extracted region centroids in the x- and y-axis
  - edge force



# Graph-Based Signature – Edge Force

---

- emphasize on the most representative, largest and spatially closest regions
- deduced from Newton's law of universal gravitation
- proportional to the number of pixels of the destination vertex of the graph
- inversely proportional to the square of the Euclidean distance between the two graph vertices

$$F_e^{s,d} = \frac{N_{G_v^d}}{(ED_{G_v^{s,d}})^2} \quad F_e^{s,d} \geq Th_e$$

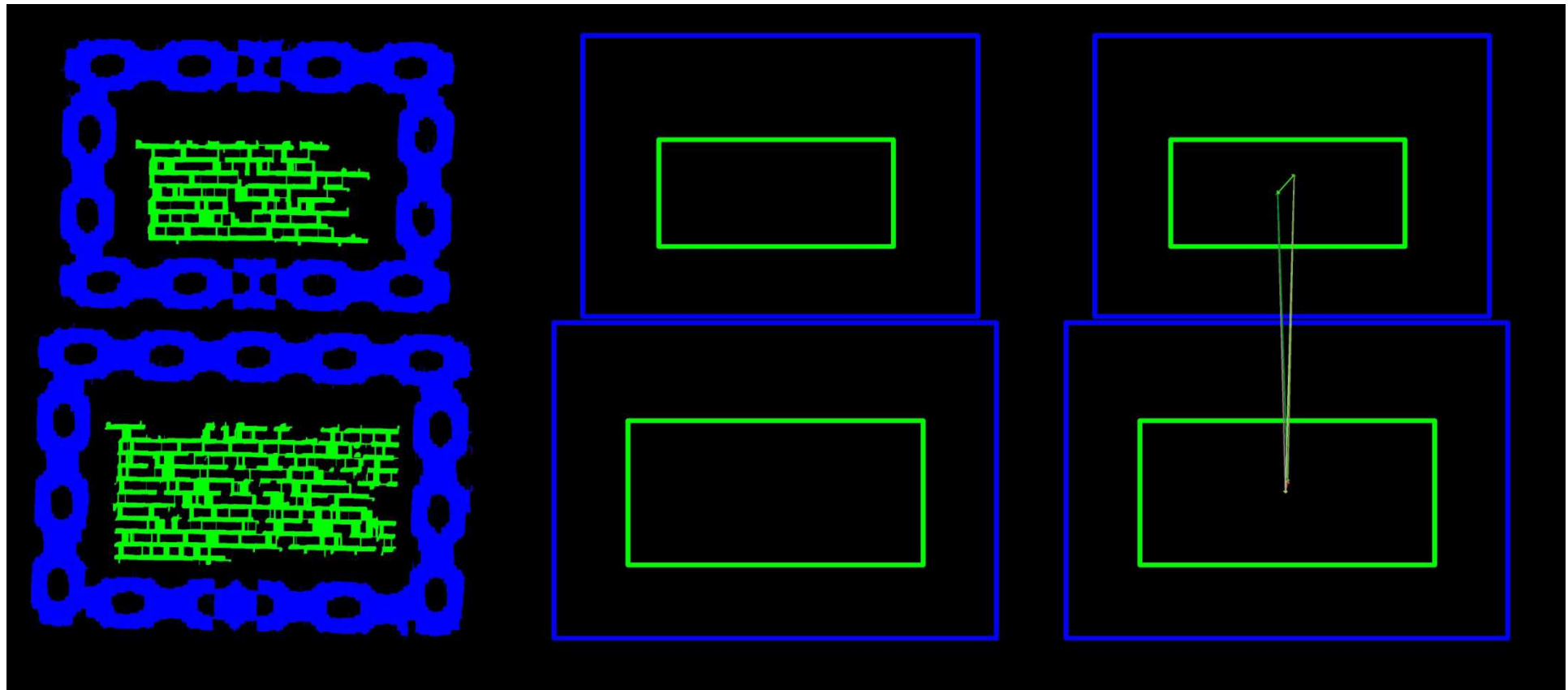
$N_{G_v^d}$ : number of pixels of the destination vertex ( $G_v^d$ ) of the built directed graph,

$ED_{G_v^{s,d}}$ : Euclidean distance between the two graph vertices: source ( $G_v^s$ ) and destination ( $G_v^d$ ),

$Th_e$ : edge force and threshold.

# Graph-Based Signature – Results (1/2)

---

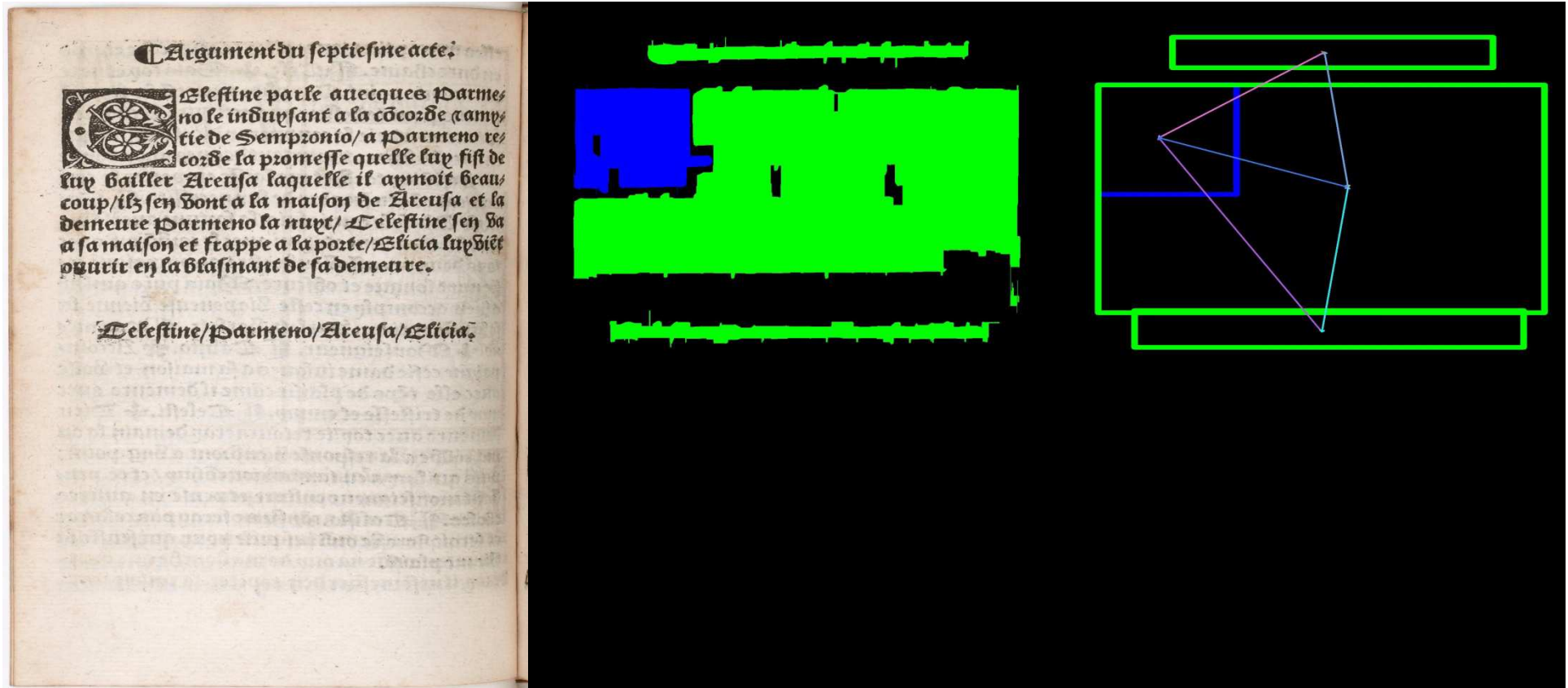


Labeled CCs

Extracted and labeled  
representative homogeneous  
regions

Graph-based signature

# Graph-Based Signature – Results (2/2)



Input image

Labeled CCs

Graph-based signature

# Proposed approach

---

## 1. DHB page characterization

## 2. DHB page categorization

- a. Unsupervised DHB page classification
- b. DHB page stream segmentation

# DHB page categorization – GED

---

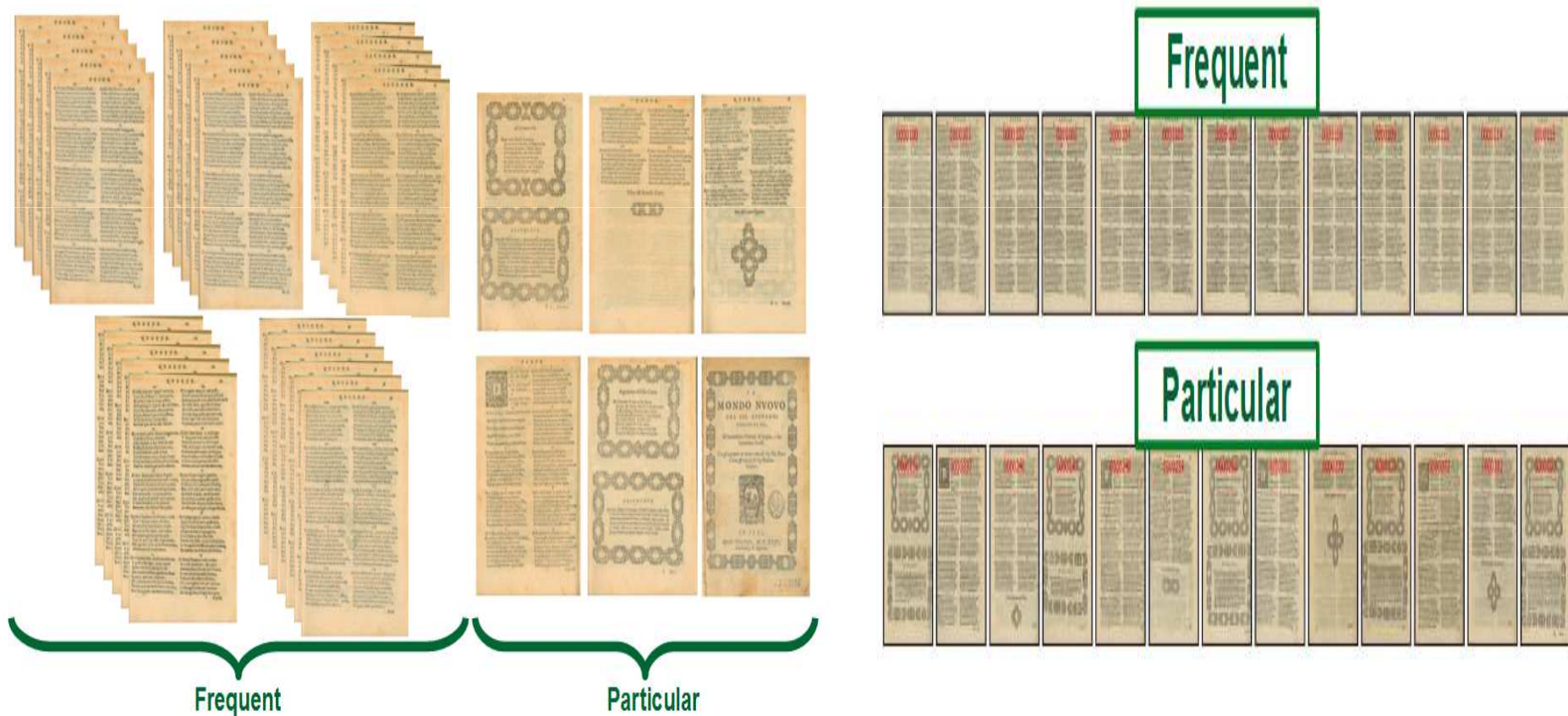
- measure the (dis)similarity between the obtained graph-based DHB page signatures
- computation of the minimum-cost sequence (e.g. insertion, deletion and substitution of vertices or edges) [Bunke07]
- This work
  - an optimized binary linear programming [GEM++]
  - a statistical analysis of the feature variations to determine weight of each feature
  - computational complexity of the GED is reduced (i.e. up to 11 vertices in the obtained graphs )
  - compute a distance matrix, whose elements represent the dissimilarity between the compared graphs

[Bunke07] Bunke, H. and Riesen, K. "Towards the unification of structural and statistical pattern recognition", PRL, 2012.

[GEM++] <http://litis-ilpiso.univ-rouen.fr/ILPiso/gem++.html>

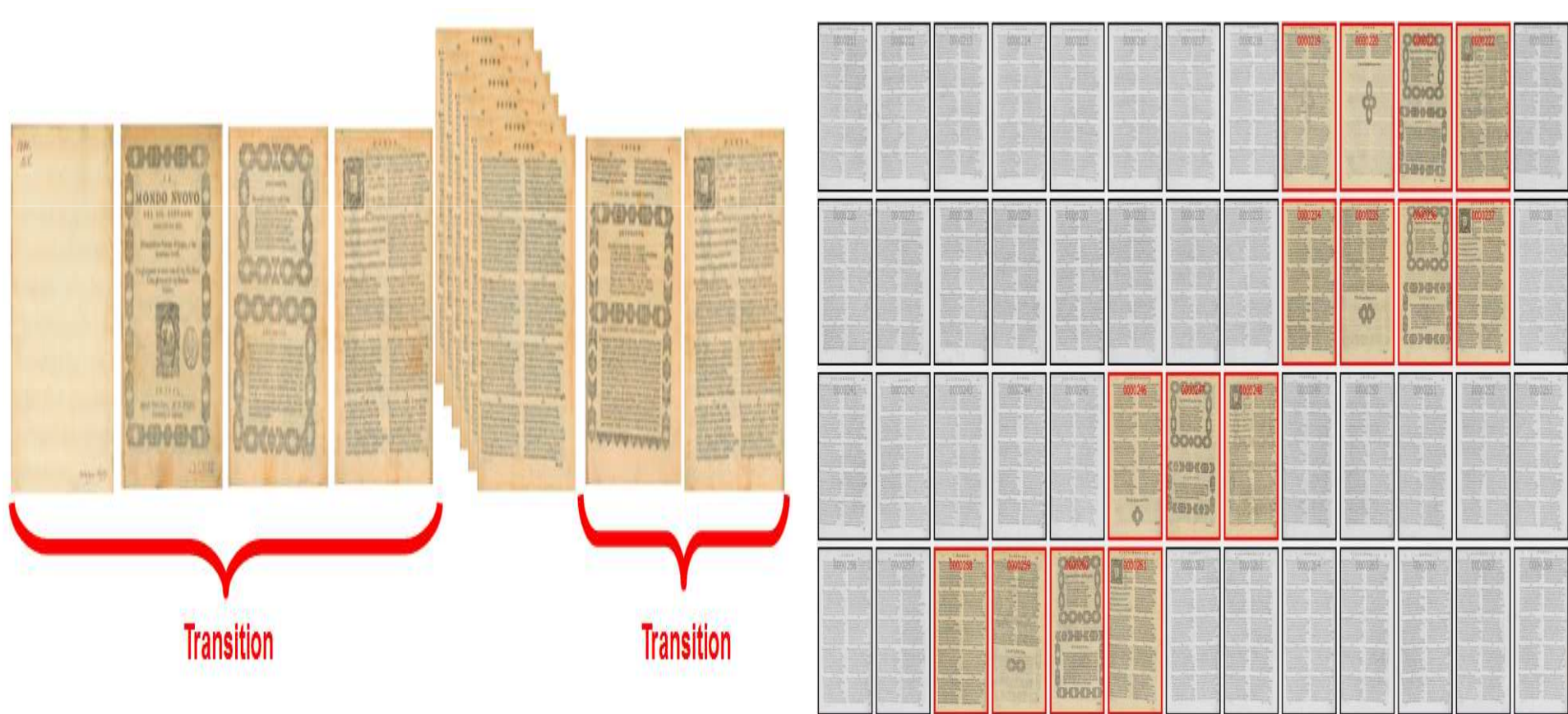
# Unsupervised DHB page classification

- Unsupervised DHB page classification to group or gather similar layout and/or content DHB pages



# DHB page stream segmentation

- DHB page stream segmentation to generate automatically a table of content/summary of the analyzed DHB



# Experiments – Corpus

---

- a printed monograph which is dated 1596, titled “Il mondo nuovo, del sig. Giov. Giorgini da Jesi” and written in Italian, which is composed of 322 ground-truthed one-page color HDIs [Gallica]

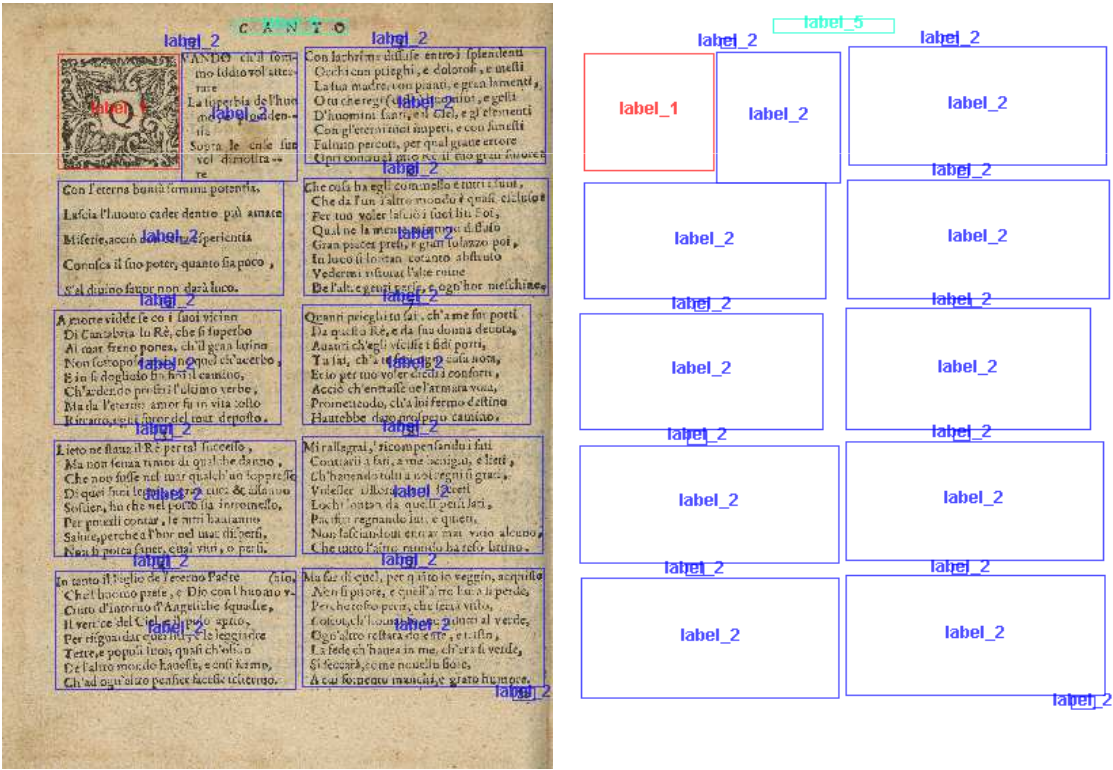


[Gallica] <http://gallica.bnf.fr/ark:/12148/bpt6k132294p/f5.planchecontact.r=.langFR>



# Experiments – Ground truth

- Ground-truthing Environment for Document Images [GEDI]
- By assigning predefined content types to rectangular regions in a document image, GEDI generates an XML schema representing:
  - location on the page
  - height
  - width
  - label



[GEDI]: <http://gedigroundtruth.sourceforge.net/>

# Evaluation – Accuracy metrics

---

- Evaluation of the different steps of the proposed algorithm
  - **Per-pixel classification accuracy (CA)**<sup>[Baird07]</sup>
- Evaluation of the homogeneous region extraction step
  - **Jaccard ( $J_{AR}$ )**: overlap ratio between the number of foreground pixels defined in the two areas:  $B_{gt}$  and  $B_r$  <sup>[Brunessaux14]</sup>

where

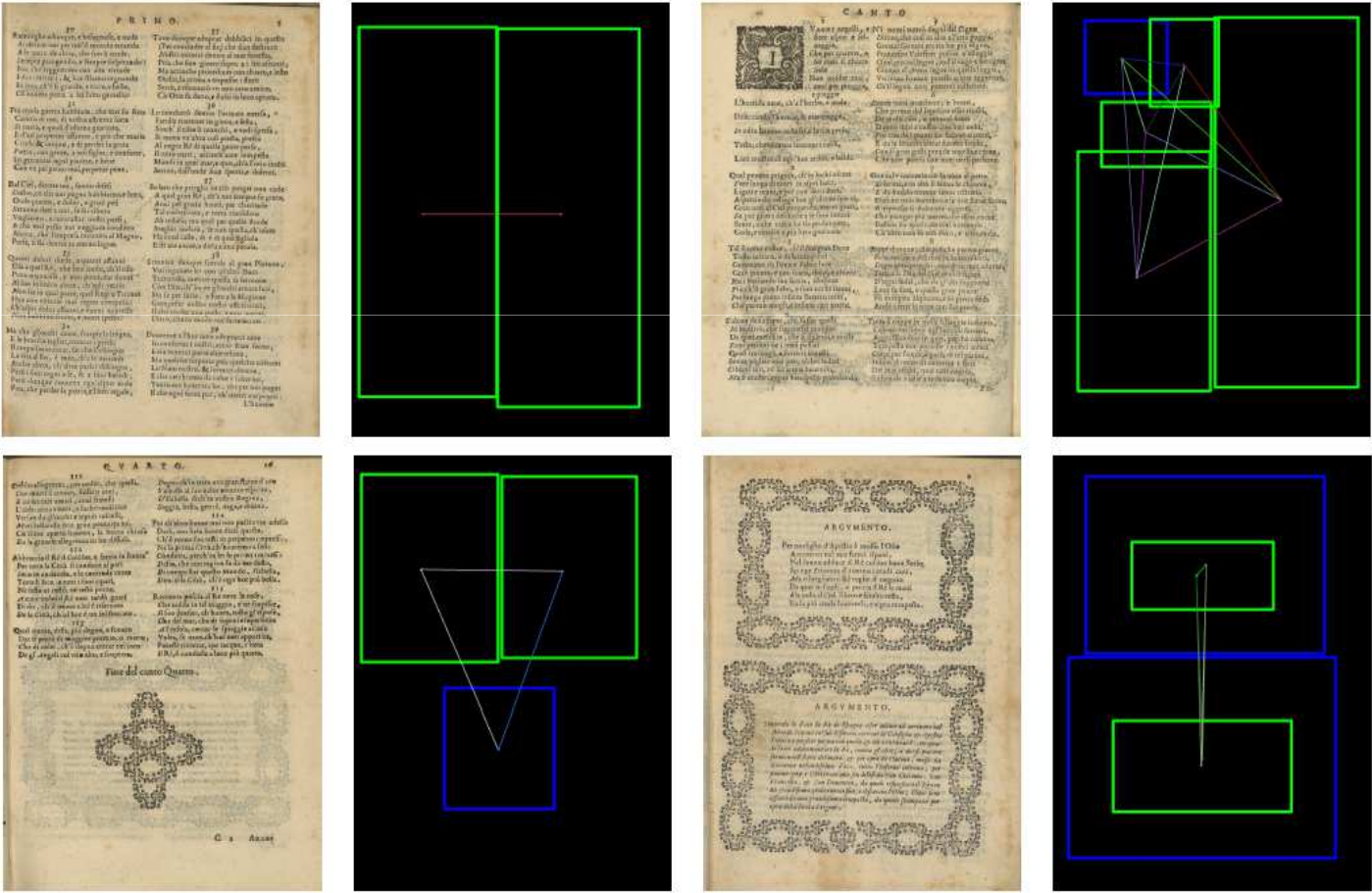
$B_{gt}$  : bounding box of the ground-truth ,  
 $B_r$  : bounding box of the result block.

[Baird07] Baird, H. S., Moll, M. A., An, C., and Casey, M. R., "Document image content inventories", DRR, 2007.

[Brunessaux14] Brunessaux, S., Giroux, P., Grilheres, B., Manta, M., Bodin, M., Choukri, K., Galibert, O., and Kahn, J. "The Maudor project: Improving automatic processing of digital documents", DAS, 2014.

# Evaluation – Results (1/5)

- DHB page characterization



# Evaluation – Results (2/5)

---

- DHB page characterization
  - **Step 3:** Pixel-clustering and labeling
  - **Step 4:** Pixel-labeling refinement
  - **Step 5:** Post-processing
  - **Step 6:** Extraction of representative homogeneous regions

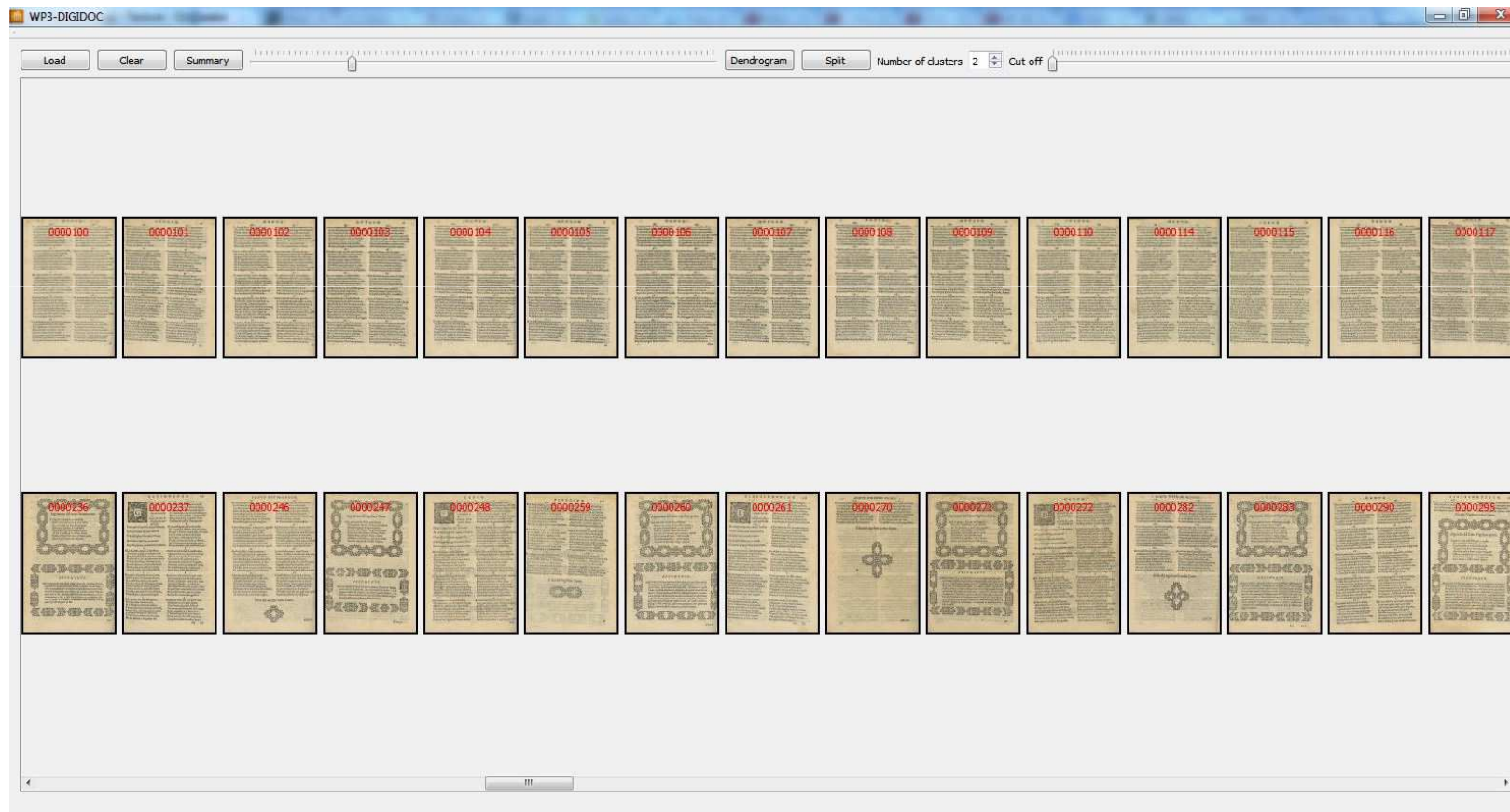
		<i>Step 3</i>	<i>Step 4</i>	<i>Step 5</i>
<i>CA</i>	$\mu$	0.977	0.983	0.987
	$\sigma$	0.066	0.085	0.076

		<i>Step 6</i>
<i>J<sub>AR</sub></i>	$\mu$	0.952
	$\sigma$	0.174

## Unsupervised page classification

# Evaluation – Results (3/5)

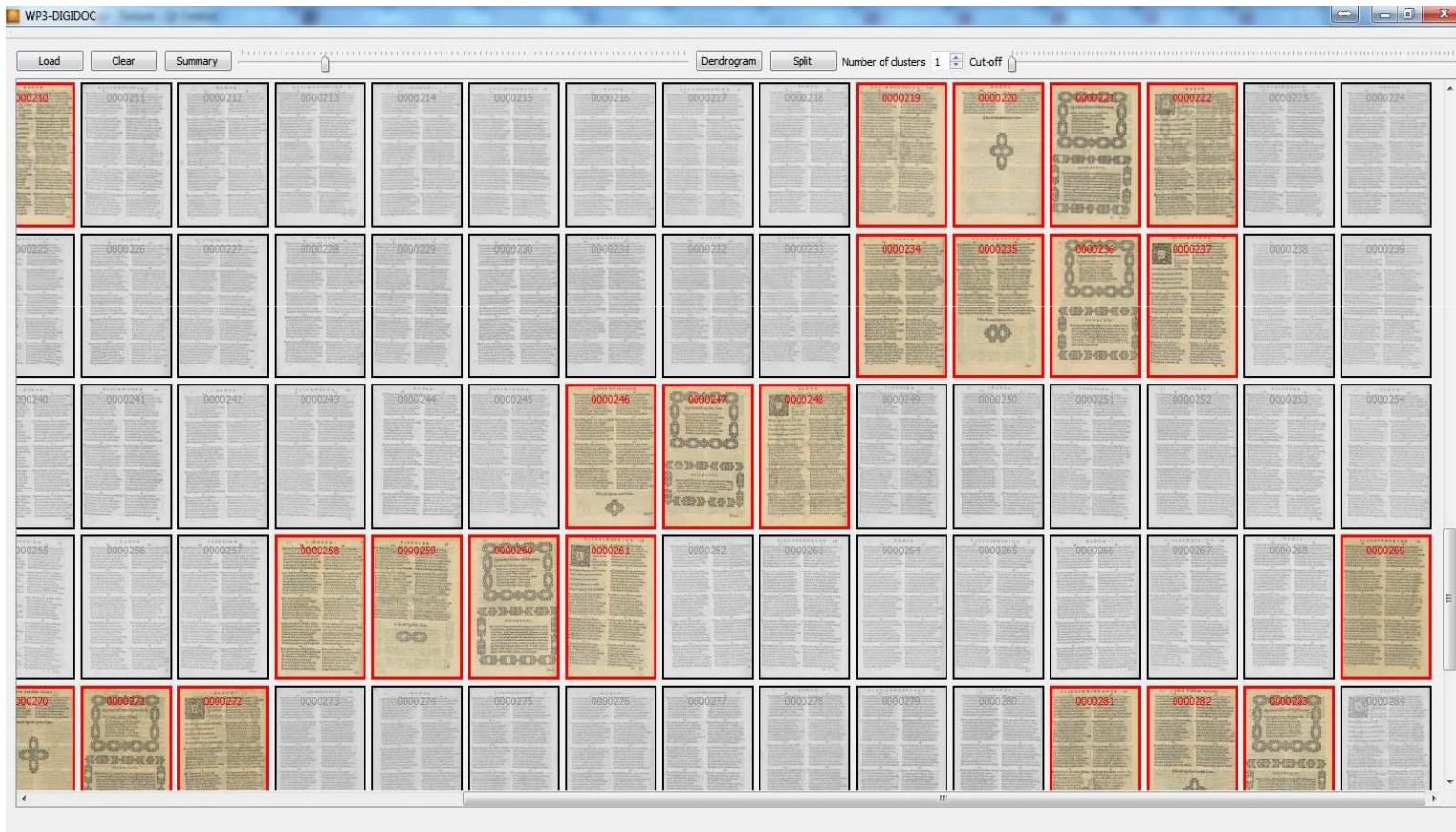
- DHB page categorization



## Unsupervised page classification

# Evaluation – Results (4/5)

- DHB page categorization

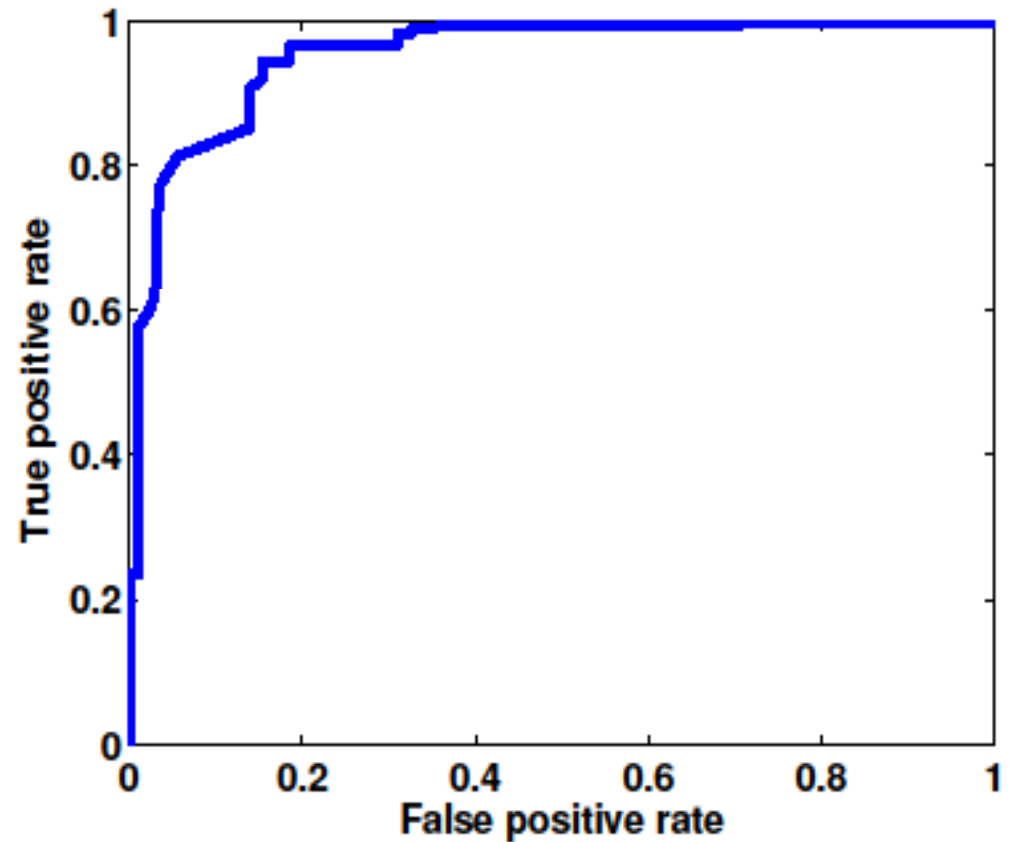
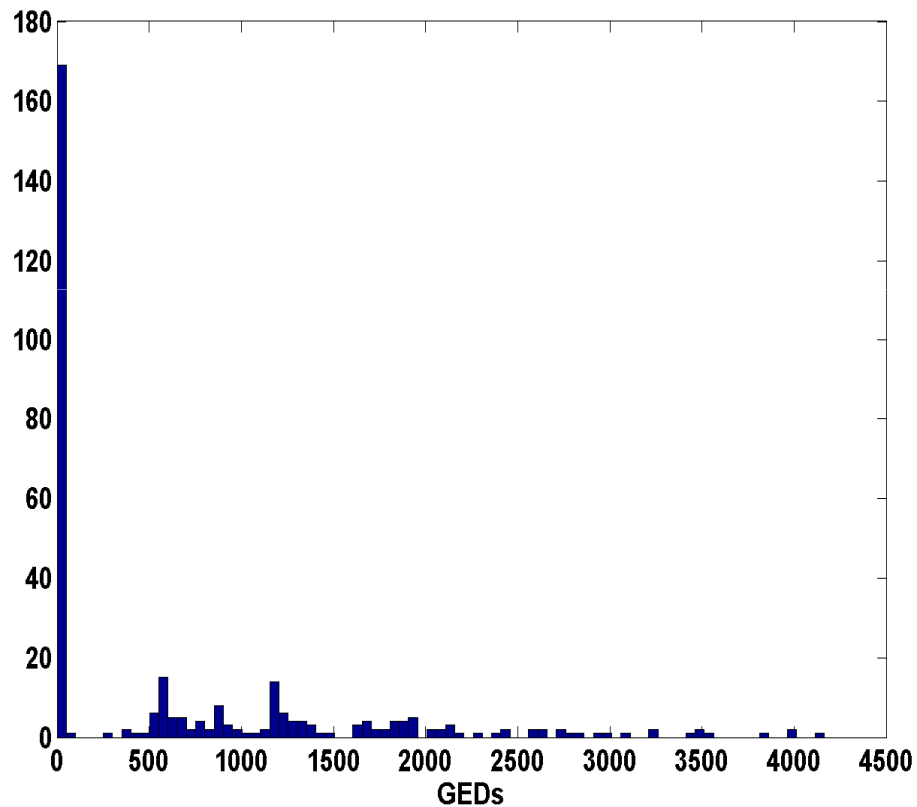


## Page stream segmentation

# Evaluation – Results (5/5)

---

- DHB page categorization



## Page stream segmentation

# Discussion

---

- good discrimination of foreground layers
  - particularly between text and graphics
- simple GUI tool for characterization and categorization of DHB pages
- numerous signature-based applications for managing effectively a corpus or collections of books
- 90% classification accuracy is noted
- table of contents of the DHB by detecting different or dissimilar pages



# Future Work (1/2)

---

- use the proposed signature on a larger databases with more variable content
- find pages in a DHB or HDI corpus (contain a particular content component or a group of patterns)
- retrieve similar pages in a HDI corpus query tool
- detect the scanning failure
- investigate a finer book page classification

## Future Work (2/2)

---

- analyze the impact of different feature weighting schemes in the cost of the GED operations
- assess other state of the art graph dissimilarity techniques
- improve the designed GUI tool for characterization and categorization of DHB pages
- ...

# Thanks

**Maroua Mehri:** [maroua.mehri@univ-lr.fr](mailto:maroua.mehri@univ-lr.fr)

University of La Rochelle - France

# Appendices

# Introduction – State of the art

---

- Contemporary document image analysis
  - textual content (e.g. OCR) [Google, “smartFIX”, “A2IA Document Reader”]
  - interest point detection (e.g. BoW, BoVW) [Bouguelia13, Augereau14]
- State-of-the-art
  - pixel-based algorithms to extract regions [Baird07,An07,An10]
  - texture features to analyze pages with overlapping layers [Okun99,Kise14]
  - poor performance of textual content based approaches due to many particularities of HDIs [Grana14]

[Bouguelia13] Bouguelia, M. R., Belaid, Y., and Belaid A., "A stream-based semi-supervised active learning approach for document classification", ICDAR, 2013.

[Augereau14] Augereau, O., Journet, N., Vialard A., and Domenger J. P., "Improving classification of an industrial document image database by combining visual and textual features", DAS, 2014.

[Baird07] Baird, H. S. and Moll, M. A., "Document content inventory and retrieval", ICDAR, 2007.

[An07] An, C., Baird, H. S., and Xiu, P., "Iterated document content classification", ICDAR, 2007.

[An10] An, C., Yin, D., and Baird, H. S., "Document segmentation using pixel-accurate ground truth", ICPR, 2010.

[Okun99] Okun, O. and Pietikäinen, M., "A survey of texture-based methods for document layout analysis", TAMV, 1999.

[Kise14] Kise, K., "Page segmentation techniques in document analysis", Document Image Processing and Recognition, 2014.

[Grana14] Grana, C., Serra, G., Manfredi, M., Coppi, D., and Cucchiara, R., "Layout analysis and content enrichment of digitized books", MTA, 2014.