BVH

CESR Centre d'Études Supérieures de la Renaissance

Using Pattern Redundancy for Text Transcription and Retrieval

JY RAMEL, PP. ROY, N. RAGOT

Laboratoire Informatique (EA 2101)
Université Francois-Rabelais de Tours - France

UNIVERSITÉ FRANÇOIS - RABELAIS TOURS

LI Laboratoire d'Informatique EA 2101

POLYTECH TOURS

Overview

- Introduction
  - Context of the work
  - Challenges
- Proposed architecture
  - Learning from the user feedback and from pattern redundancy
  - User driven method with AGORA
- Pattern redundancy
  - Description
  - Pattern selection
- Using connected components
  - For text recognition (OCR)
  - For typography analysis
- Using glyphs
  - For word spotting
  - For text transcription (OCR)
- Conclusion

Introduction
Context of the work

- Collaboration with the CESR of Tours
  - □ The Humanistic Virtual Library (BVH in French)
  - □ A research center and library with rare books (Loire Valley)
  - □ From the Renaissance period (14th - 16th)
- A pluri-disciplinary collaboration
  Experts in DIA + Experts in old books + End-users
- Objectives: Deal with and manage specificity of old books
  - □ Fully automatic is impossible because of **variability**
  - □ Introduce more interaction into DIA systems
    ➔ user-driven method
  - □ Adaptation according to image contents (typography)
    ➔ not before but during the processing

Introduction
Challenges

- **Experiments with OCR realized by [AitMohand&Al2010]**
  - □ Segmentation in lines, words, characters is a problem

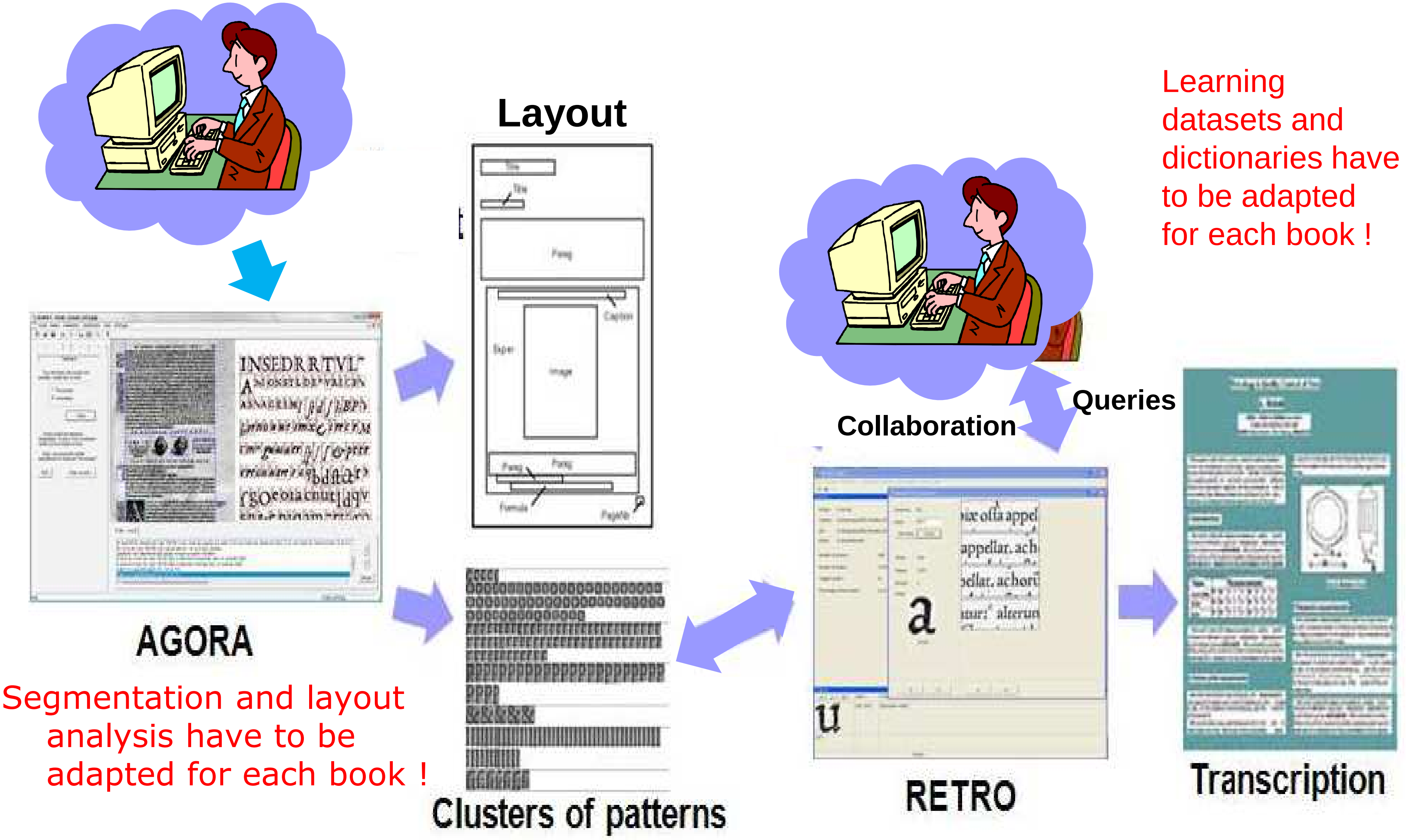
Books of the CESR	Omnipage classical segmentation	Omnipage with Ocropus segmentation
Recueil des antiquités Gauloises	89.82%	85.93%
Histoire de l'expédition chrestienne au royaume de Chine	86.48%	61.25%
Les treselegantes et copieuses annales	85.6%	73.92%
Les histoires de Diodore sicilien	90.19%	83.82%

  - □ Learning datasets (typography) and prior knowledge (dictionaries and linguistic aspects) are very important

Font	poly-font system	Adapted poly-font system	mono-font system
Average (30 fonts)	86.59	96.02	99.55
Berkeley Old – Berkeley Oldstyle	96.62	97.2	98.98
Banco – Banco	34.08	73.46	98.77
Mistral – Mistral	46.63	92.16	94.89
Fette Kanzlei – Fette Kanzlei	68.36	95.74	99.43

Proposed Architecture

Learning from the users and from the data



Layout

Learning datasets and dictionaries have to be adapted for each book !

Queries

Collaboration

AGORA

Segmentation and layout analysis have to be adapted for each book !

Clusters of patterns

RETRO

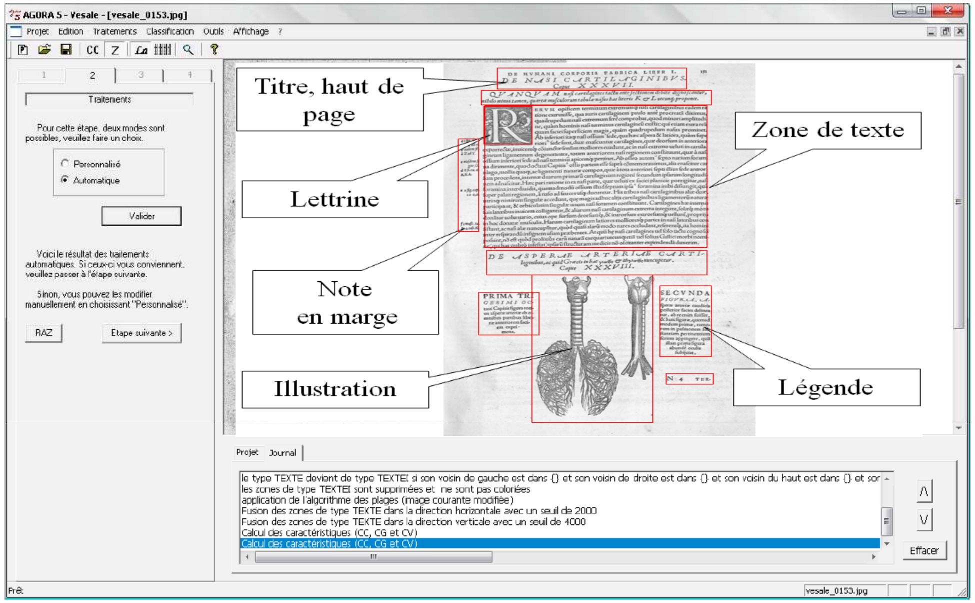
Transcription

System Architecture

User-driven analysis with AGORA

- **User-driven analysis**
- Extraction of specific elements of contents (dropcaps, …)
- Generate XML files describing the structure (similar to Alto)

  Lines, words and CC positions
- Used since 2004 (CESR)



Pour cette étape, deux modes sont possibles, veuillez faire un choix.

○ Personnalisé
● Automatique

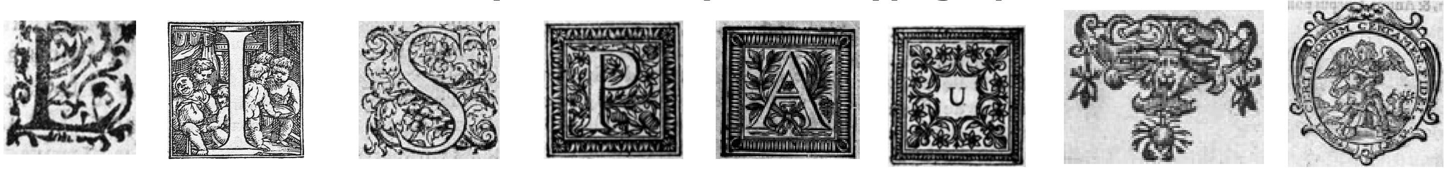
Voici le résultat des traitements automatiques. Si ceux-ci vous conviennent, veuillez passer à l'étape suivante.

Sinon, vous pouvez les modifier manuellement en choisissant "Personnalisé".

```
le type TEXTE devient de type TEXTEI si son voisin de gauche est dans {} et son voisin de droite est dans {} et son voisin du haut est dans {} et sor
les zones de type TEXTEI sont supprimées et  ne sont pas coloriées
application de l'algorithme des plages (image courante modifiée)
Fusion des zones de type TEXTE dans la direction horizontale avec un seuil de 2000
Fusion des zones de type TEXTE dans la direction verticale avec un seuil de 4000
Calcul des caractéristiques (CC, CG et CV)
Calcul des caractéristiques (CC, CG et CV)
```

Download : http://www.rfai.li.univ-tours.fr/pagesperso/ramel/fr/work1.html

- □ **Bases of ormemantal letters (+of 15000) and of typographical materials**



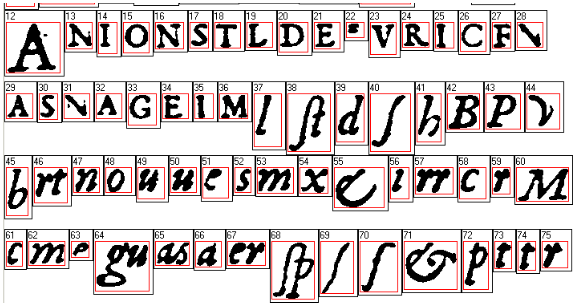
- □ **Bases of portraits (+ de 1500)**

BORYSTENITES BION
POETA ARATVS
CLAZOM. PHIL. ANAXAGORAS
D. BONICONTI
FRANCISCVS ZABARELLA

Voir sur http://www.bvh.univ-tours.fr

Pattern Redundancy in text
Description

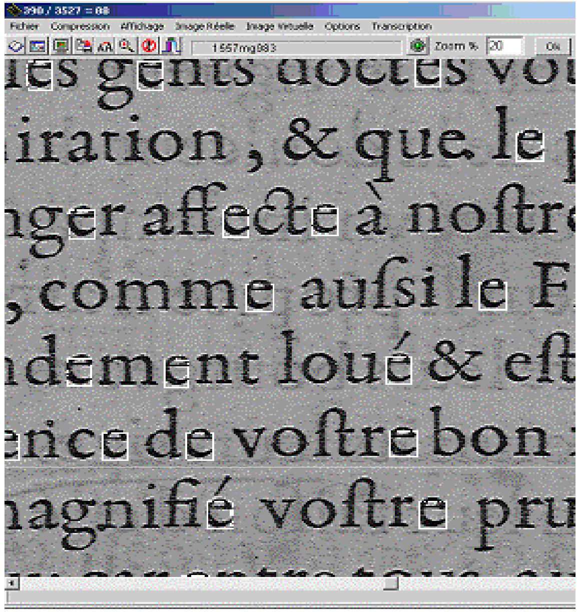
- Goal: Analyzing redundancy in images (text part for us)
  - A text, ancient or not, is made up of sequences of similar patterns
- Methods: Clustering of similar patterns to create groups (classes)
  - Comparison of patterns (matching techniques)
  - Without prior knowledge about the meaning of these patterns
- Constraints are that the techniques should:
  - Produce very homogeneous clusters ➔ Different patterns may not be blended into one cluster
  - Produce a minimal number of clusters
- What could be a pattern?
  - Connected components [Lebourgeois95]
  - Words [Kluzner&Al2009]
  - Others [Roy&Al2011]
  - Redundancy rate > 80 %



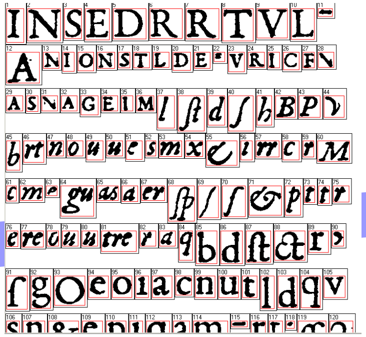
- Used first for compression in Debora project and DjVu Format

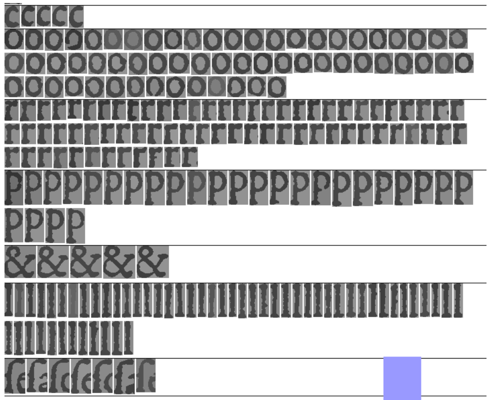
Using connected components

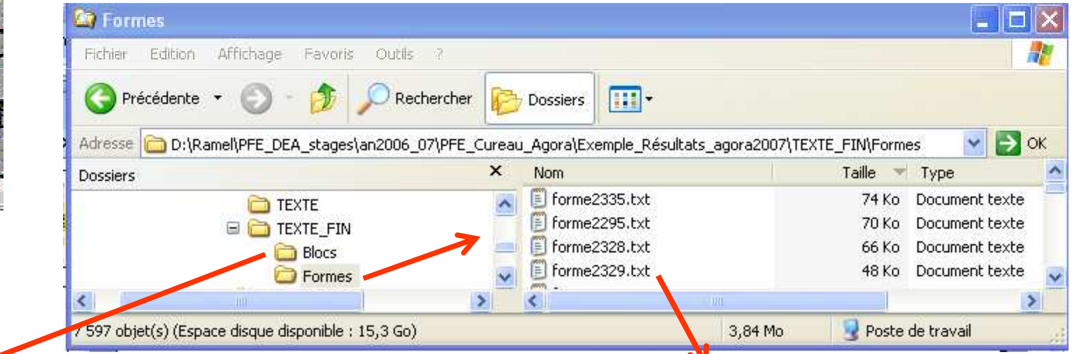
AGORA ouputs


390 / 3527 = 88
Fichier Compression Affichage Image Réelle Image Virtuelle Options Transcription
1557mg983
Zoom % 20
Ok

ies gents doctes vou
iration , & que le p
nger affecte à noſtr
, comme auſsi le F
ndement loué & eſt
ence de voſtre bon
nagnifié voſtre pru





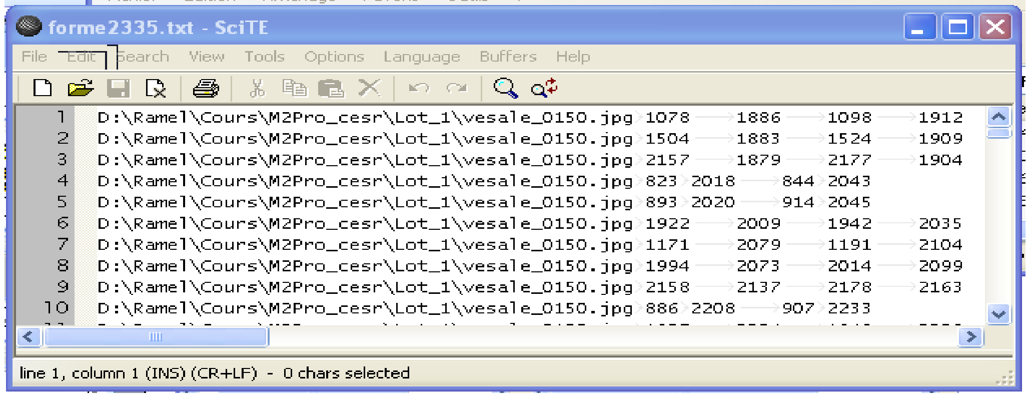


D:\Ramel\PFE_DEA_stages\an2006_07\PFE_Cureau_Agora\Exemple_Résultats_agora2007\TEXTE_FIN\Formes

Nom	Taille	Type
forme2335.txt	74 Ko	Document texte
forme2295.txt	70 Ko	Document texte
forme2328.txt	66 Ko	Document texte
forme2329.txt	48 Ko	Document texte

597 objet(s) (Espace disque disponible : 15,3 Go) 3,84 Mo Poste de travail

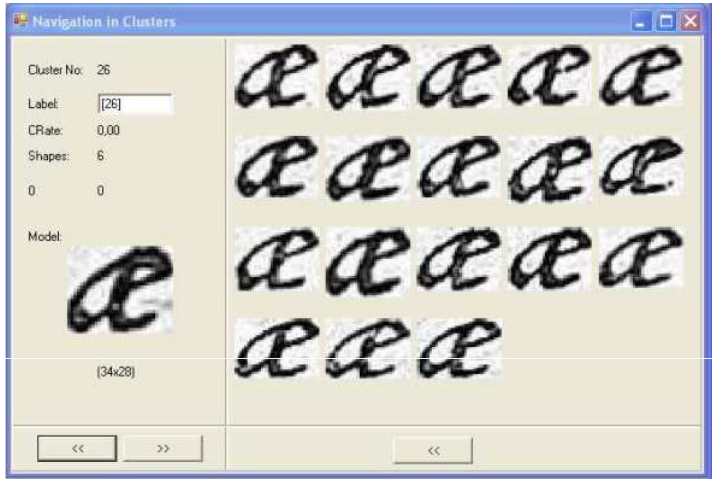
```
<bloc name="C:\PFE\ImagesCESR\Lot\vesale_0150.jpg"
  coord="798,107,2023,155">
  <ligne>
    <mot>
      <cc value="2">798,123,831,155</cc>
      <cc value="3">849,123,881,154</cc>
      <cc value="4">899,123,934,155</cc>
      <cc value="5">951,121,981,153</cc>
      <cc value="6">997,121,1024,153</cc>
      <cc value="7">1038,119,1069,151</cc>
    </mot>
  </ligne>
</bloc>
```


forme2335.txt - SciTE
File Edit Search View Tools Options Language Buffers Help
line 1, column 1 (INS) (CR+LF) - 0 chars selected

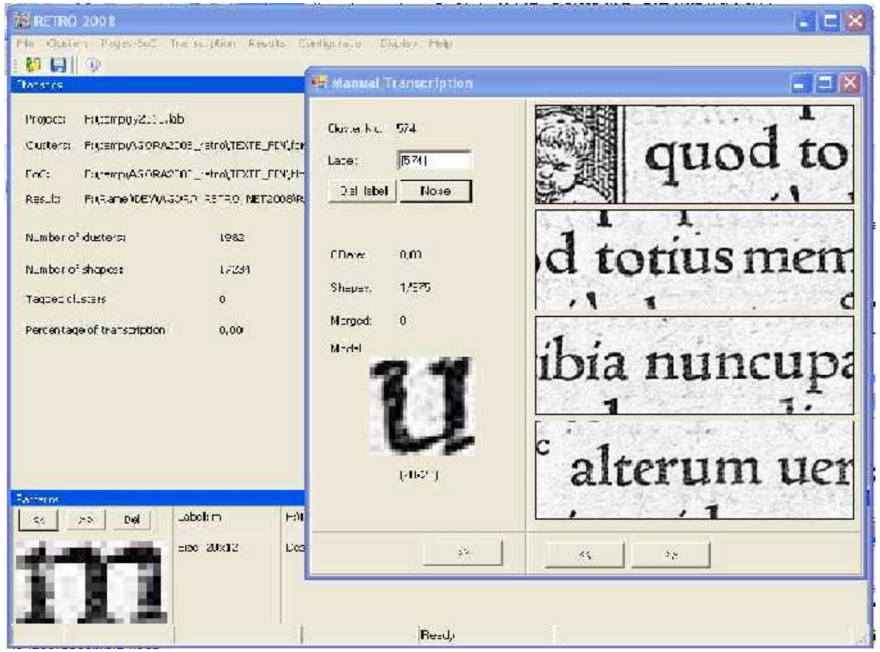
```
D:\Ramel\Cours\M2Pro_cesr\Lot_1\vesale_0150.jpg 1078    1886    1098    1912
D:\Ramel\Cours\M2Pro_cesr\Lot_1\vesale_0150.jpg 1504    1883    1524    1909
D:\Ramel\Cours\M2Pro_cesr\Lot_1\vesale_0150.jpg 2157    1879    2177    1904
D:\Ramel\Cours\M2Pro_cesr\Lot_1\vesale_0150.jpg 823 2018    844 2043
D:\Ramel\Cours\M2Pro_cesr\Lot_1\vesale_0150.jpg 893 2020    914 2045
D:\Ramel\Cours\M2Pro_cesr\Lot_1\vesale_0150.jpg 1922    2009    1942    2035
D:\Ramel\Cours\M2Pro_cesr\Lot_1\vesale_0150.jpg 1171    2079    1191    2104
D:\Ramel\Cours\M2Pro_cesr\Lot_1\vesale_0150.jpg 1994    2073    2014    2099
D:\Ramel\Cours\M2Pro_cesr\Lot_1\vesale_0150.jpg 2158    2137    2178    2163
D:\Ramel\Cours\M2Pro_cesr\Lot_1\vesale_0150.jpg 886 2208    907 2233
```

Using connected components
For Transcription

- **RETRO GUI – Computer Assisted Transcription (manual)**
  - □ For tagging the clusters using unicode
  - □ Cluster visualization
  - □ Characters (CCs) in context
  - □ Creation (selection) of new templates

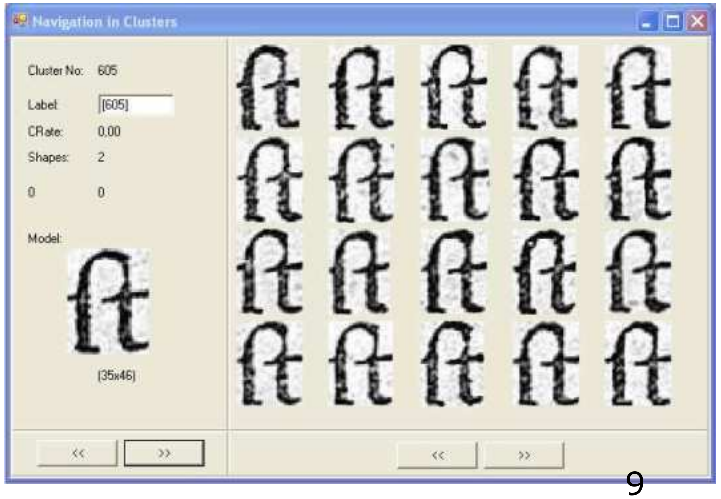


Cluster No: 26
Label: [26]
CRate: 0,00
Shapes: 6
0 0
Model:
(34x28)



Project: F:\temp\y2...lab
Clusters: F:\temp\AGORA2008_retro\TEXTE_FEN\for
FoC: F:\temp\AGORA2007_retro\TEXTE_FEN\H
Results: F:\Rame\DEV\AGORA_RETRO_NET2008\R
Number of clusters: 1982
Number of shapes: 17231
Tagged clusters: 0
Percentage of transcription: 0,00
Label: m
Size: 20x12

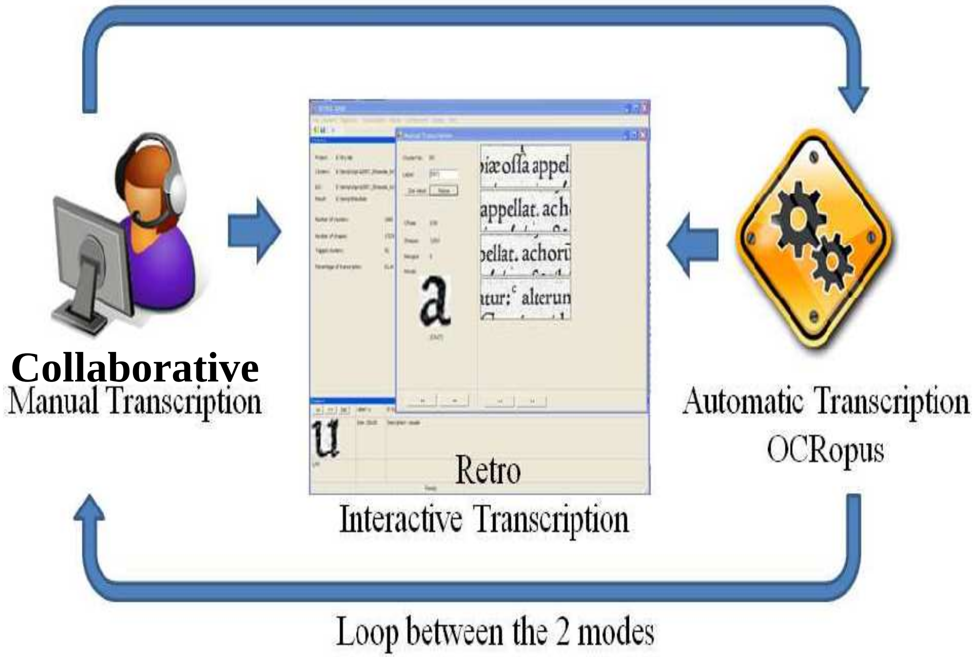
Cluster No: 574
Label: [574]
CRate: 0,00
Shapes: 1/375
Merged: 0
Model:



Cluster No: 605
Label: [605]
CRate: 0,00
Shapes: 2
0 0
Model:
(35x46)

Using connected components
For Transcription

- **Inside a loop !**
  - □ Cooperation between manual (collaborative users), automatic (OCR) and contextual (dictionaries) contributions
  - □ *Adaptive-system : From simplest to the more complicated*



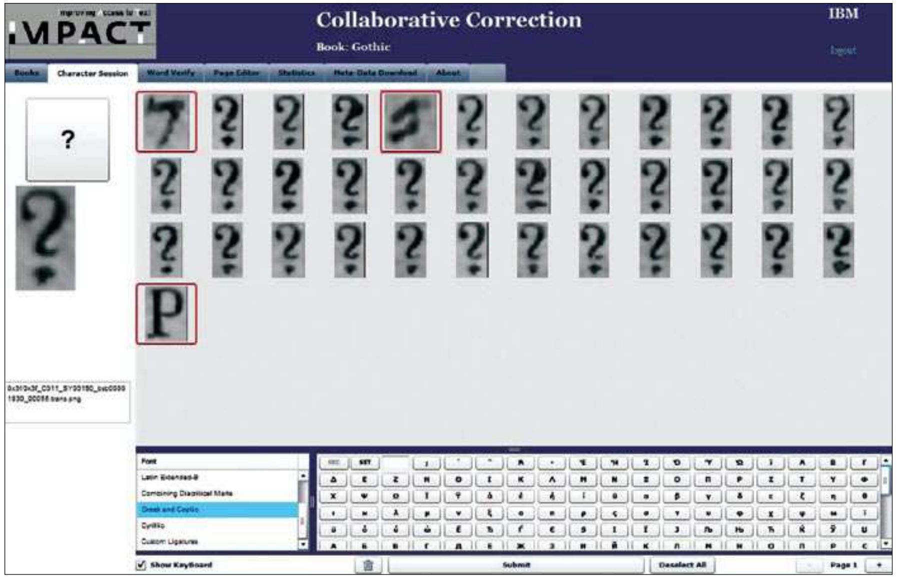
Collaborative
Manual Transcription

Retro
Interactive Transcription

Automatic Transcription
OCRopus

Loop between the 2 modes

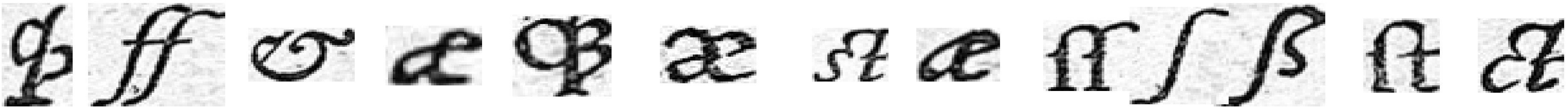
User collaboration [Neudeker&Al2010]



Using connected components
For Typographic analysis

- **Create a classification of the Early Modern fonts**
  - □ Relationship between words containing characters which have the same shape → typographical family and character style
  - □ Sorted newly created families to find the main typography class as well as minor typographies used for a precise logical meaning
  - □ Very small typographical families represent words which seldom occur in the text (text in graphics, titles, authors' names, etc.)

- **Study of aesthetic aspects of printing**
  - □ The thickness and the shape of printing types evolved greatly from the 15th to the mid-16th century
  - □ Extract and create new font packages from specific printing material (e.g. rare books printed with particular plug sets).

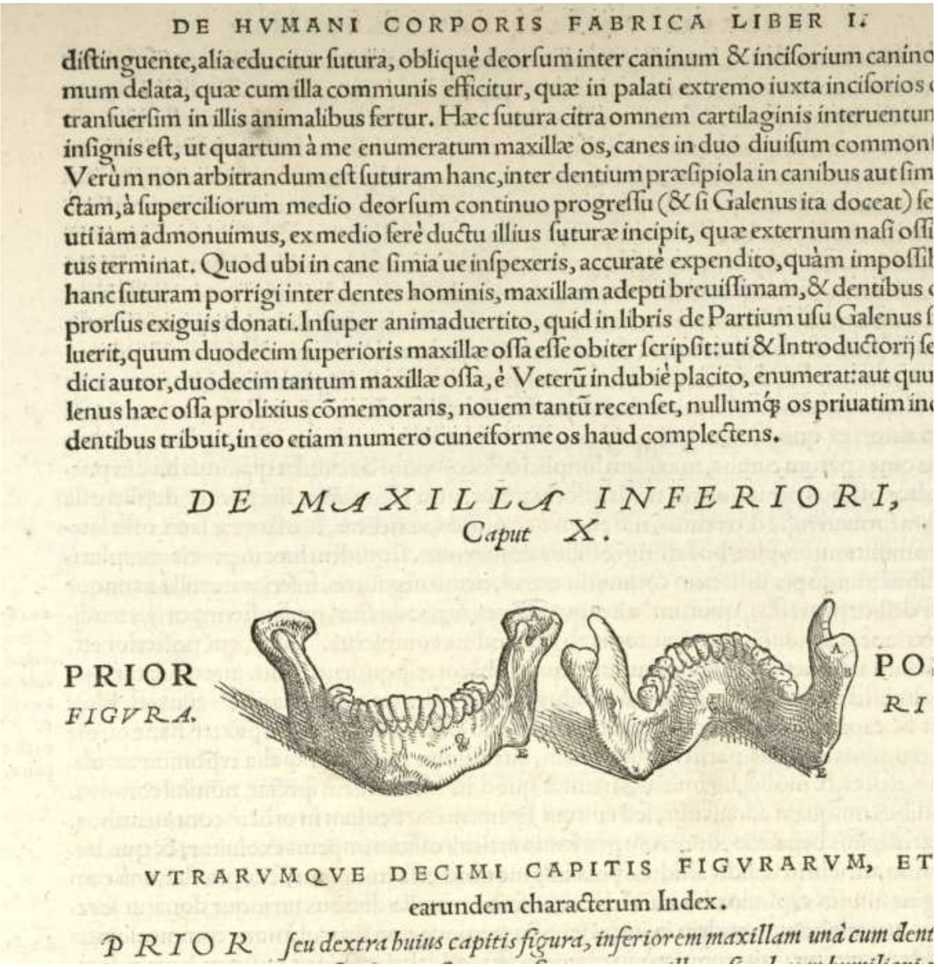


Using connected components
For Typographic analysis

- **Improving the OCR learning step (templates)**
- Dataset production
  - Based on the previous proposition ➔ typographic analysis
  - Produced fonts + model of distortion and degradation = adapted training sets
- Dynamic template selection (incremental learning)
  - Identification of specific fonts used inside the images
  - Automatic selection of specific OCR dedicated to that font (mono-font OCR)
  - Increase the potential performances of OCR engines
  - *Adaptive-systems able to learn from the data*

Using connected components
Experiments

- **Vésale – 1543**
- 150 pages in Latin
- 1.062.081 connected components
- Around 40.000 clusters
- The 200 largest classes correspond to 85% of the text
- 57% of the classes are composed of a single shape
- 90% of the classes are composed of less than 10 occurrences
- Ignoring these classes during transcription means to miss one character for 14 → more than one on each text line !!!



DE HVMANI CORPORIS FABRICA LIBER I.

diſtinguente, alia educitur ſutura, obliquè deorſum inter caninum & inciſorium canino
mum delata, quæ cum illa communis efficitur, quæ in palati extremo iuxta inciſorios
tranſuerſim in illis animalibus fertur. Hæc ſutura citra omnem cartilaginis interuentun
inſignis eſt, ut quartum à me enumeratum maxillæ os, canes in duo diuiſum commont
Verùm non arbitrandum eſt ſuturam hanc, inter dentium præſipiola in canibus aut ſim
ctam, à ſuperciliorum medio deorſum continuo progreſſu (& ſi Galenus ita doceat) ſe
uti iam admonuimus, ex medio ferè ductu illius ſuturæ incipit, quæ externum naſi oſſi
tus terminat. Quod ubi in cane ſimia ue inſpexeris, accuratè expendito, quàm impoſſib
hanc ſuturam porrigi inter dentes hominis, maxillam adepti breuiſſimam, & dentibus
prorſus exiguis donati. Inſuper animaduertito, quid in libris de Partium uſu Galenus
luerit, quum duodecim ſuperioris maxillæ oſſa eſſe obiter ſcripſit: uti & Introductorij ſe
dici autor, duodecim tantum maxillæ oſſa, è Veterũ indubiè placito, enumerat: aut quu
lenus hæc oſſa prolixius cõmemorans, nouem tantũ recenſet, nullumꝗ os priuatim in
dentibus tribuit, in eo etiam numero cuneiforme os haud complectens.

DE MAXILLÆ INFERIORI,
Caput X.

PRIOR
FIGVRA.

PO
RI

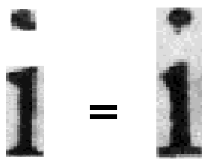
VTRARVMQVE DECIMI CAPITIS FIGVRARVM, ET
earundem characterum Index.

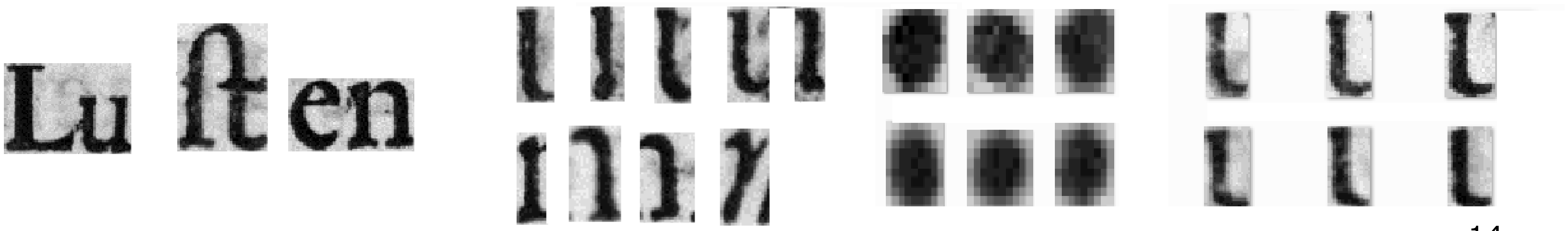
PRIOR ſeu dextra huius capitis figura, inferiorem maxillam unà cum dent

Using connected components (CC)
Discussions

- **Good points with CC redundancy**
  - □ CCs should correspond to characters
  - □ System can learn from the images and adapts itself to the used typographic materials
  - □ Cluster transcription or recognition instead of individual pattern recognition (collaborative, manual, contextual, automatic, …)

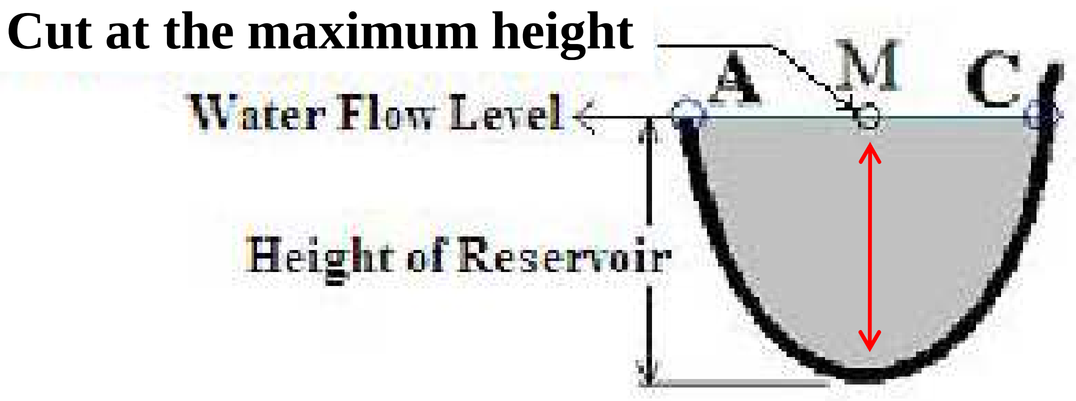
- **The segmentation problem is still remaining…**
  - □ Still require a segmentation step for characters, words and lines
  - □ Problem with touching and broken characters (CCs)
  - □ Problem with accents, punctuations, …



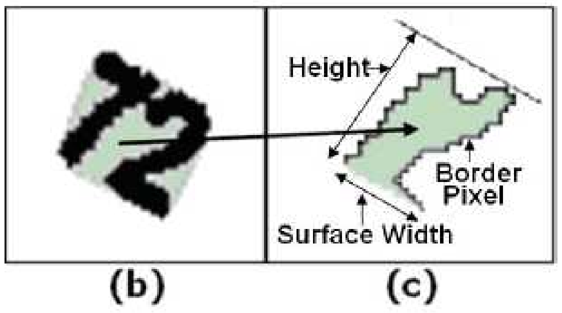


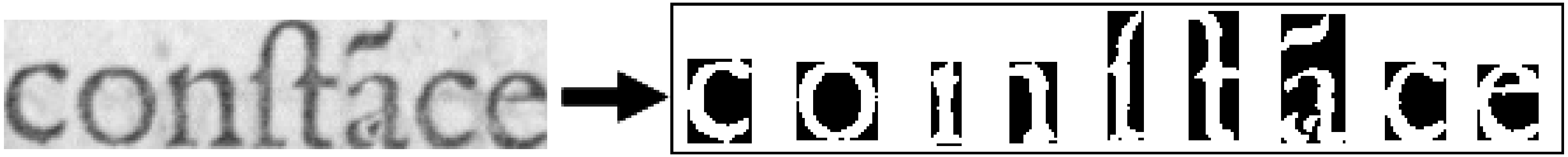
Using glyphs [Roy&Al2011]
For Word spotting

- **Overcoming of the segmentation problem**
  - □ Glyphs (parts of the connected components) instead of CCs
  - □ Don’t need a segmentation in words and in characters
  - □ Water reservoir method to split a CC into glyphs



Cut at the maximum height



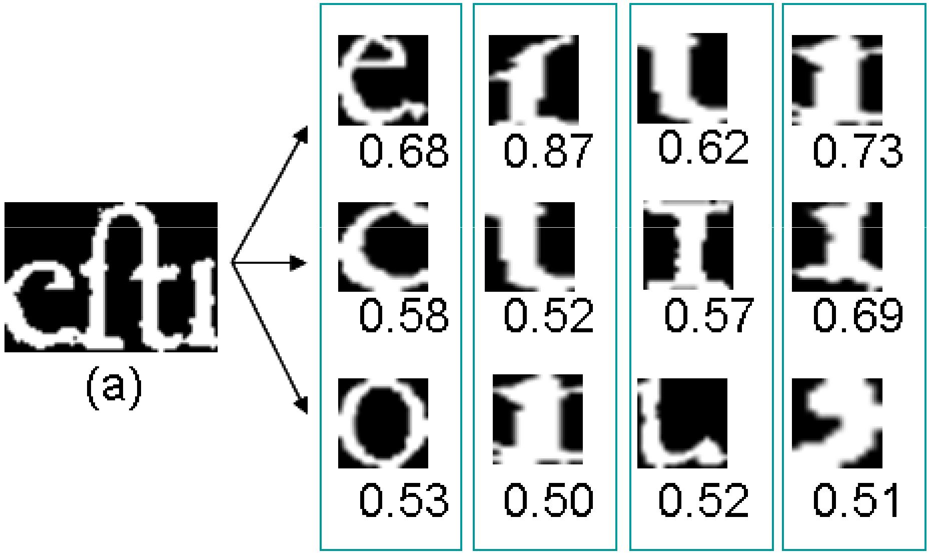


Using glyphs
For word spotting

- **Clustering of glyphs**
  - □ Similarity measure between glyphs
    - For comparison of 2 glyph images
    - Size normalization by 20x20
  - □ Creation of a codebook of frequent glyphs
    - Using a set of training images
    - Clustering of glyphs
    - Clusters selection (according to the size)
    - A glyph cluster is represented by a selected glyph (median)
  - □ Text blocks are encoded
    - Each text line is indexed by a string of codebook indexes
    - Glyph clasisication with the same similarity measure

Using glyphs
For word spotting

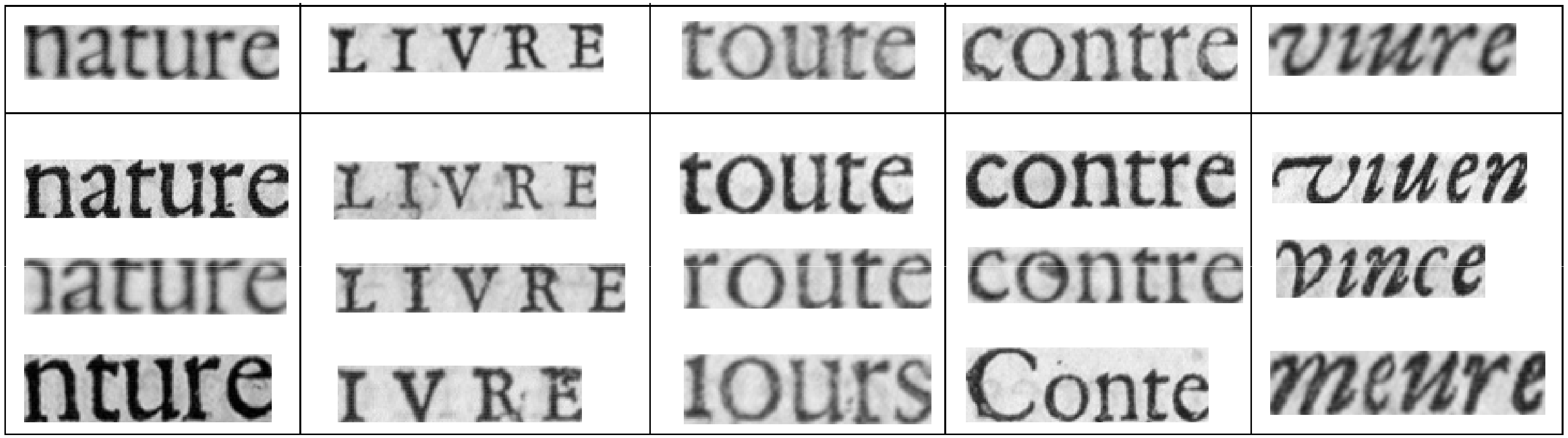
- **Word retrieval**
  - Glyph extraction and classification
  - Top *c=3* conservation (label and similarity measure)



  - Approximate string matching algorithm (DTW like)
  - Length of the strings *Query* and *Indexed* may be different
  - Finds all substrings of the *Query* that have at most k errors
  - Adapted to handle *'c'* choices for each glyph in the *Query*

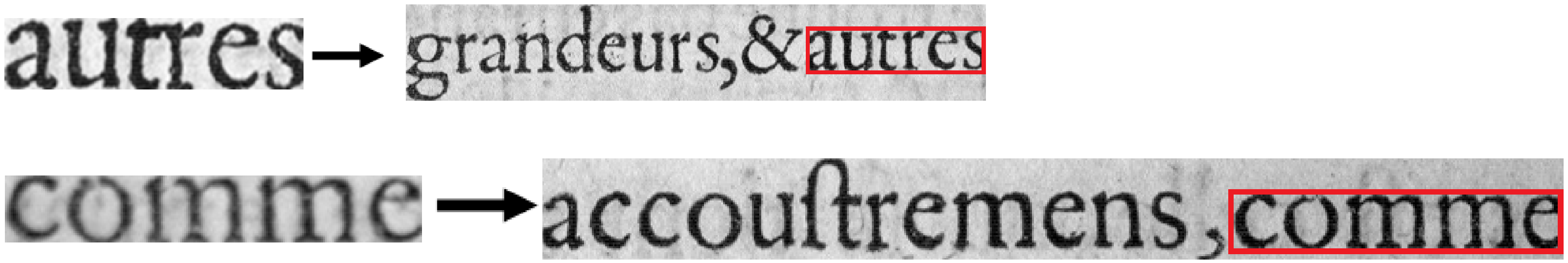
Using glyphs
For word spotting

- **Experiments**
  - □ Examples of Top 3 results for 5 queries



nature	LIVRE	toute	contre	viure
nature nature nture	LIVRE LIVRE IVRE	toute route iours	contre contre Conte	viuen vince meure

  - □ Examples of spotted regions

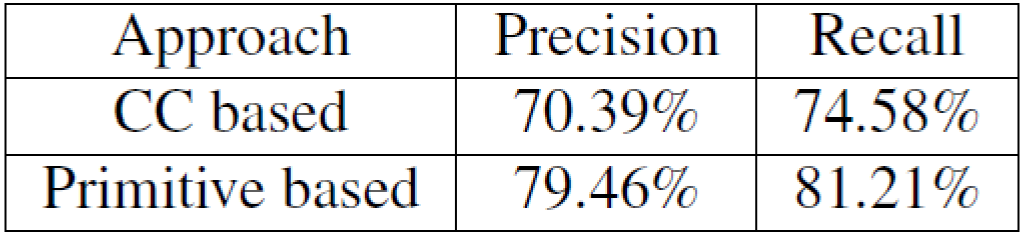


autres
grandeurs,&autres
comme
accouſtremens, comme

Using glyphs

For word spotting

- **Experiments**
  - □ 45 pages of a historical book written mostly in French
  - □ AGORA line segmentation ➔ 8675 word blocks
  - □ Codebook of glyph generated from 24 pages (training)
  - □ 57324 glyphs found in the training pages
  - □ Clustered in 183 representative glyphs
  - □ With connected components ➔ 326 representative clusters
  - □ Indexing = page processing
  - □ Results on 20 query word images



Approach	Precision	Recall
CC based	70.39%	74.58%
Primitive based	79.46%	81.21%

Conclusion

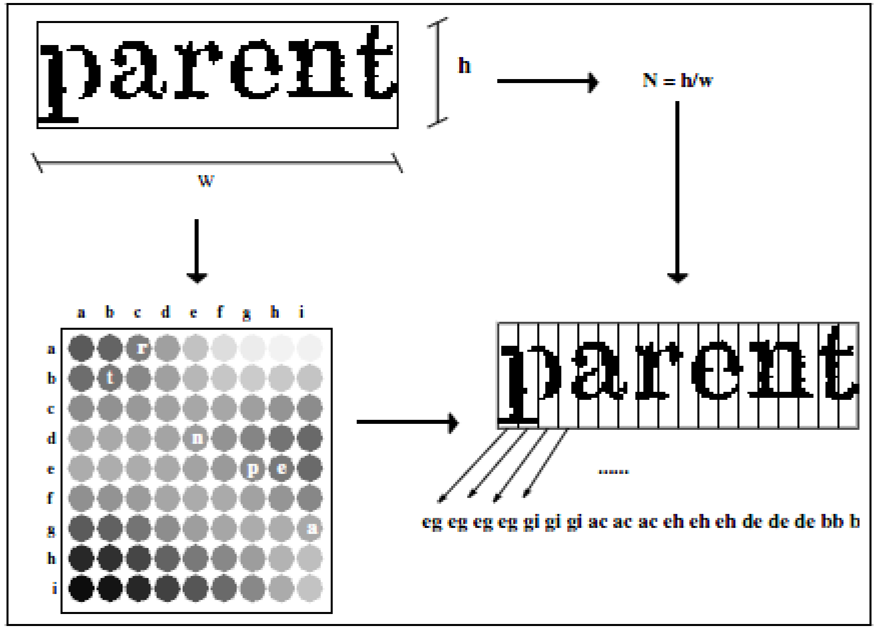
- Proposition of new methods
  - Learning from the images (using redundancy analysis)
  - Adaptive system, user-driven system
- Cluster transcription / recognition (collaborative, manual, contextual, automatic, …) instead of individual pattern recognition
- Segmentation of text in words and word in characters is a problem
  - Touching and broken characters/connected components
  - Accents, punctuations, …
- Using glyphs can be a solution
  - Done for Word spotting
  - To be studied for Transcription
- Work in progress…
  - Continue on using glyphs for text spotting and transcription
  - Google DH project (typography, lexicons, …) with CESR

Thanks

Questions ?

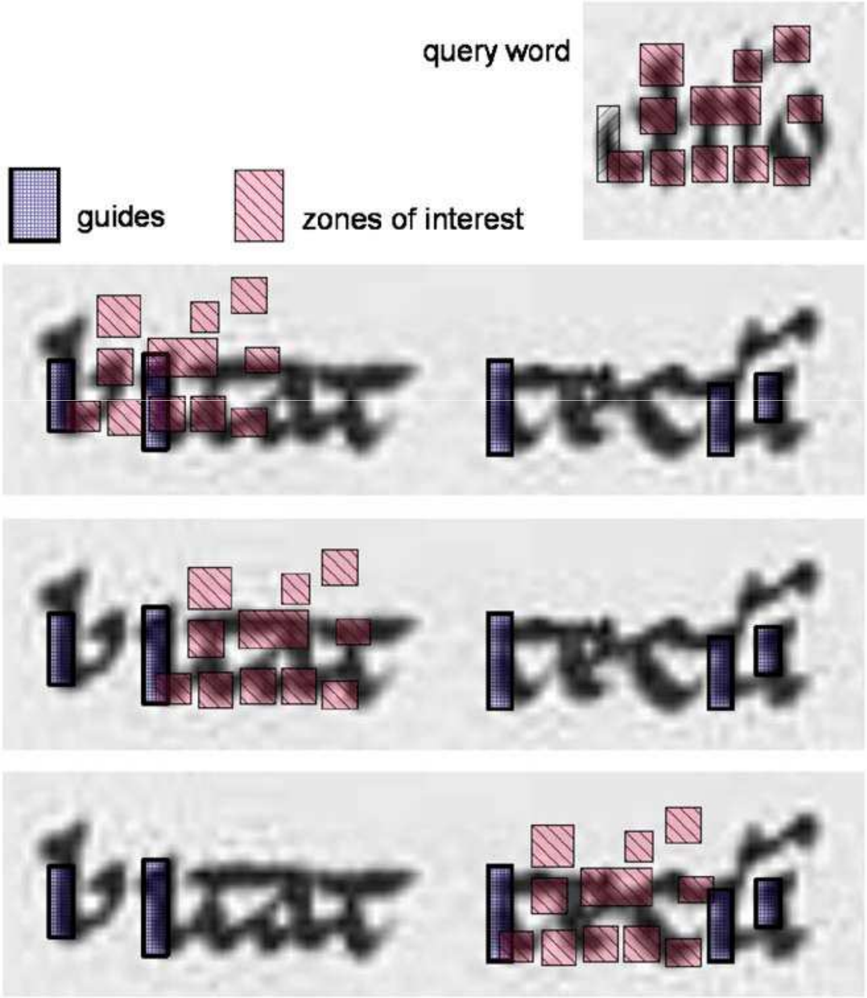
Context of the work

- {Marinia2003]
  - Character Objects = CO
  - Extraction and clustering of COs
  - The COs in the word are located
  - Each CO is labeled with the output neuron of the trained SOM.
  - The word image is partitioned into a fixed number of vertical slices.
  - Each slice gets the label of the CO with the largest overlap with it



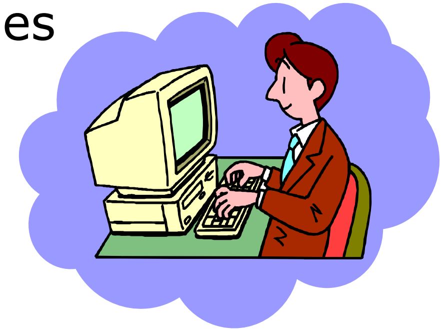
Context of the work

- □ Word Spotting [Leydier&Al2009]
- □ Guides and Zone of interests



Introduction
Our proposal

- **More interaction in DIA systems**
  - For adaptation according to each "book" specificities
  - For integration of the user needs



- Interactive analysis of images
  - □ Adaptation according to user objectives

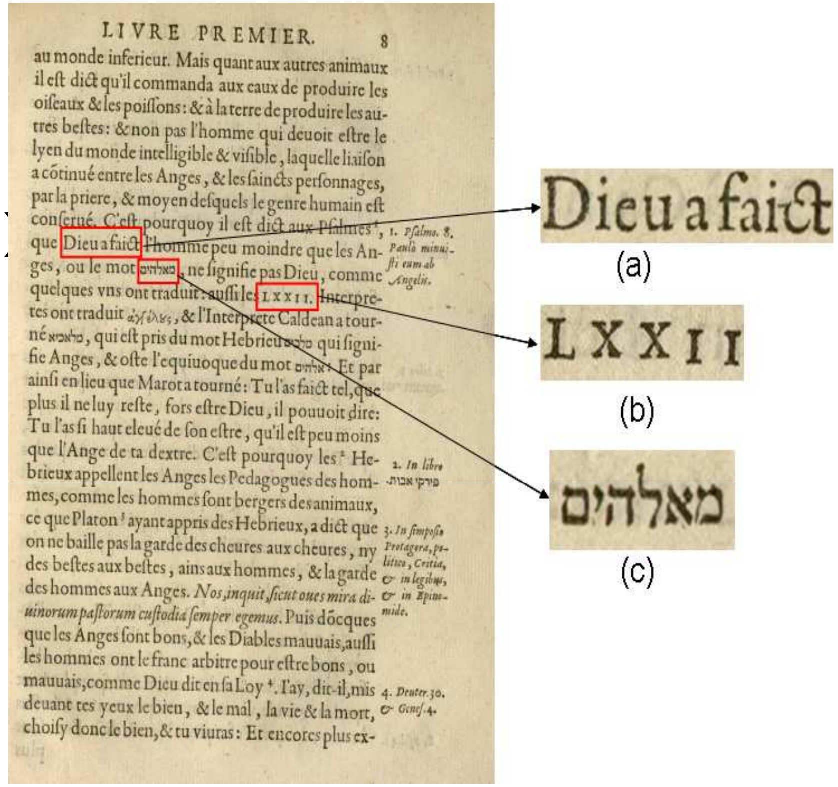
- Incremental analysis of images
  - □ Segmentation for recognition, recognition for segmentation
  - □ Solution: From the simplest to the more difficult

- Requirement
  - □ An adequate representation of the image content
  - □ Interoperability and compatibility capabilities between automatic and manual processing

Introduction

Our proposal

- **Why OCR software will never work on such books ?**
  - Linguistic aspects (Old French, Latin, …)
  - Typography
    - Materials (specific fonts)
    - Spacing (touching, broken, space)
  - For example
    - The “long s” characters often confused with the letter “f” by OCRs
    - The “ct” ligature used in European fonts before the 19th
  - **Pattern to recognize (words, characters or primitives)?**



LIVRE PREMIER. 8

au monde inferieur. Mais quant aux autres animaux il eſt dict qu'il commanda aux eaux de produire les oiſeaux & les poiſſons: & à la terre de produire les autres beſtes: & non pas l'homme qui deuoit eſtre le lyen du monde intelligible & viſible, laquelle liaiſon a cõtinué entre les Anges, & les ſaincts perſonnages, par la priere, & moyen deſquels le genre humain eſt conſerué. C'eſt pourquoy il eſt dict aux Pſalmes[1], que Dieu a faict l'homme peu moindre que les Anges, ou le mot אלהים, ne ſignifie pas Dieu, comme quelques vns ont traduit: auſſi les LXXII. Interpretes ont traduit ἀγγέλους, & l'Interprete Caldean a tourné מלאכיא, qui eſt pris du mot Hebrieu מלאך qui ſignifie Anges, & oſte l'equiuoque du mot אלהים. Et par ainſi en lieu que Marot a tourné: Tu l'as faict tel, que plus il ne luy reſte, fors eſtre Dieu, il pouuoit dire: Tu l'as ſi haut eleué de ſon eſtre, qu'il eſt peu moins que l'Ange de ta dextre. C'eſt pourquoy les[2] Hebrieux appellent les Anges les Pedagogues des hommes, comme les hommes ſont bergers des animaux, ce que Platon[3] ayant appris des Hebrieux, a dict que on ne baille pas la garde des cheures aux cheures, ny des beſtes aux beſtes, ains aux hommes, & la garde des hommes aux Anges. *Nos, inquit, ſicut oues mira diuinorum paſtorum cuſtodia ſemper egemus.* Puis dõcques que les Anges ſont bons, & les Diables mauuais, auſſi les hommes ont le franc arbitre pour eſtre bons, ou mauuais, comme Dieu dit en ſa Loy[4]. I'ay, dit-il, mis deuant tes yeux le bien, & le mal, la vie & la mort, choiſy donc le bien, & tu viuras: Et encores plus ex-

[1] *Pſalmo. 8. Paulò minuiſti eum ab Angelis.*

[2] *In libro* פרקי אבות.

[3] *In ſimpoſio Protagora, politico, Critia, & in legibus, & in Epinomide.*

[4] *Deuter. 30. & Geneſ. 4.*

Introduction
Our proposal

- **Experiments (on synthetic data)**
  - □ Significant improvement when modifying the learning set of the OCR according to fonts present in documents

- **Experiments (on books from 17th - 18th)**
  - □ Significant improvement when replacing the default learning set by template characters from Human or Garalde font families:
    - Numerous ligatures between characters
    - Special characters used during the Renaissance period

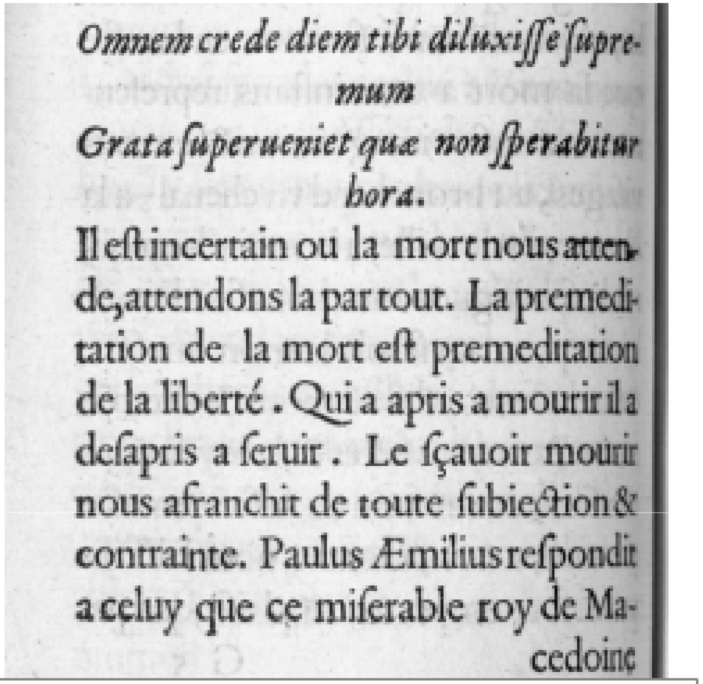
Results provided by [AitMohand&Al2010]

Using connected components

Experiments

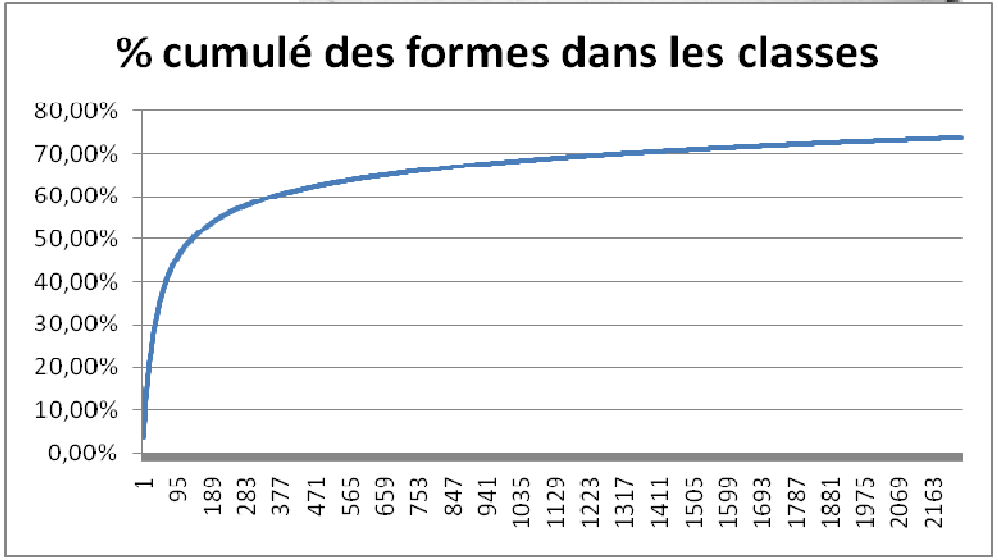
- ***Montaigne - 1557***
- 119 pages – 3260 text blocks
- 125 744 connected components (pseudo characters)
- 29 943 clusters
- 25 classes = 25% of the text
- 136 classes = 50%
- 4 000 classes = 75%
- 20 000 classes = 90%
- 79% of the classes are composed of a single shape
- The biggest class = 3%

- 1,2% of the shapes are put in the wrong cluster



Omnem crede diem tibi diluxiſſe ſupremum
Grata ſuperueniet quæ non ſperabitur hora.

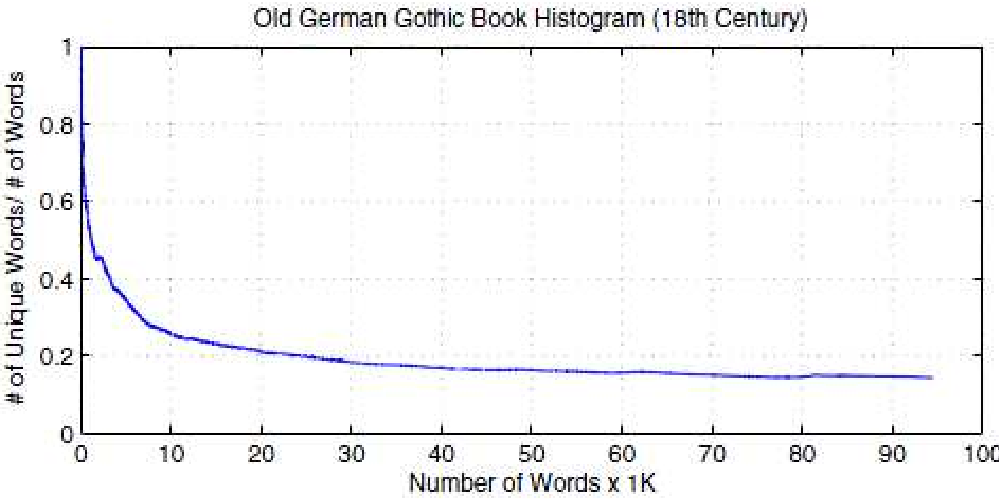
Il eſt incertain ou la mort nous attende, attendons la par tout. La premeditation de la mort eſt premeditation de la liberté. Qui a apris a mourir il a deſapris a ſeruir. Le ſçauoir mourir nous afranchit de toute ſubiection & contrainte. Paulus Æmilius reſpondit a celuy que ce miſerable roy de Macedoine

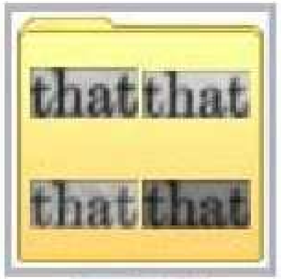


Pattern Redundancy
Pattern selection

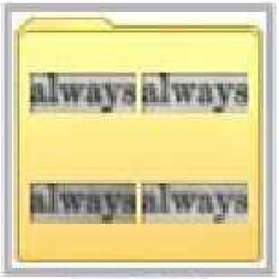
- **Using Words as Patterns** [Kluzner&Al2009]
  - □ Converge after about 30000 words - limit of 85% for 40000 words
  - □ Experiments : 101 scanned pages from a book printed in 18th century

	Reco. Rate	Subst. Rate
Commercial OCR	82.5%	1.85%
Commercial OCR after addition of Adaptive OCR	86.6%	1.7%

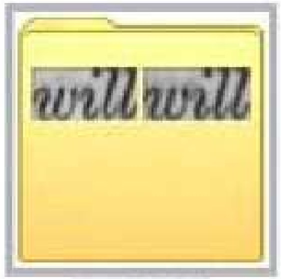




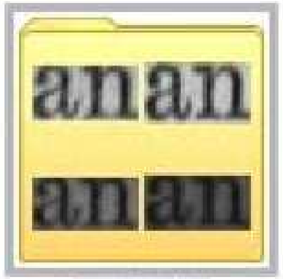
word_105



word_107



word_114



word_116

Pattern Redundancy
Pattern selection

- **Using connected components as patterns**
- The first and simplest way to realize such analysis
- Redundancy rate starts around 75% when using a single page
- Redundancy can reach up to 95% when processing an entire book (modern)
- This rate depends largely on the quality of printing

Number of pages	1	2	3	4	5	6	7	8
Total # of clusters of binary patterns	555	915	1,209	1,485	1,678	1,870	2,083	2,262
Total # of characters	2,327	4,245	6,681	8,681	11,159	13,589	16,141	18,028
Redundancy rate	76%	78%	81%	82%	84%	86%	87%	88%

- Redundancy rates slightly upwards of 80% when documents present high typographical variabilities of character style, size, and font.