WP3-DIGIDOC

Thesis: Extraction of metadata related to "image" and "structure" contained in old documents

Prepared by: Maroua MEHRI[1,2]

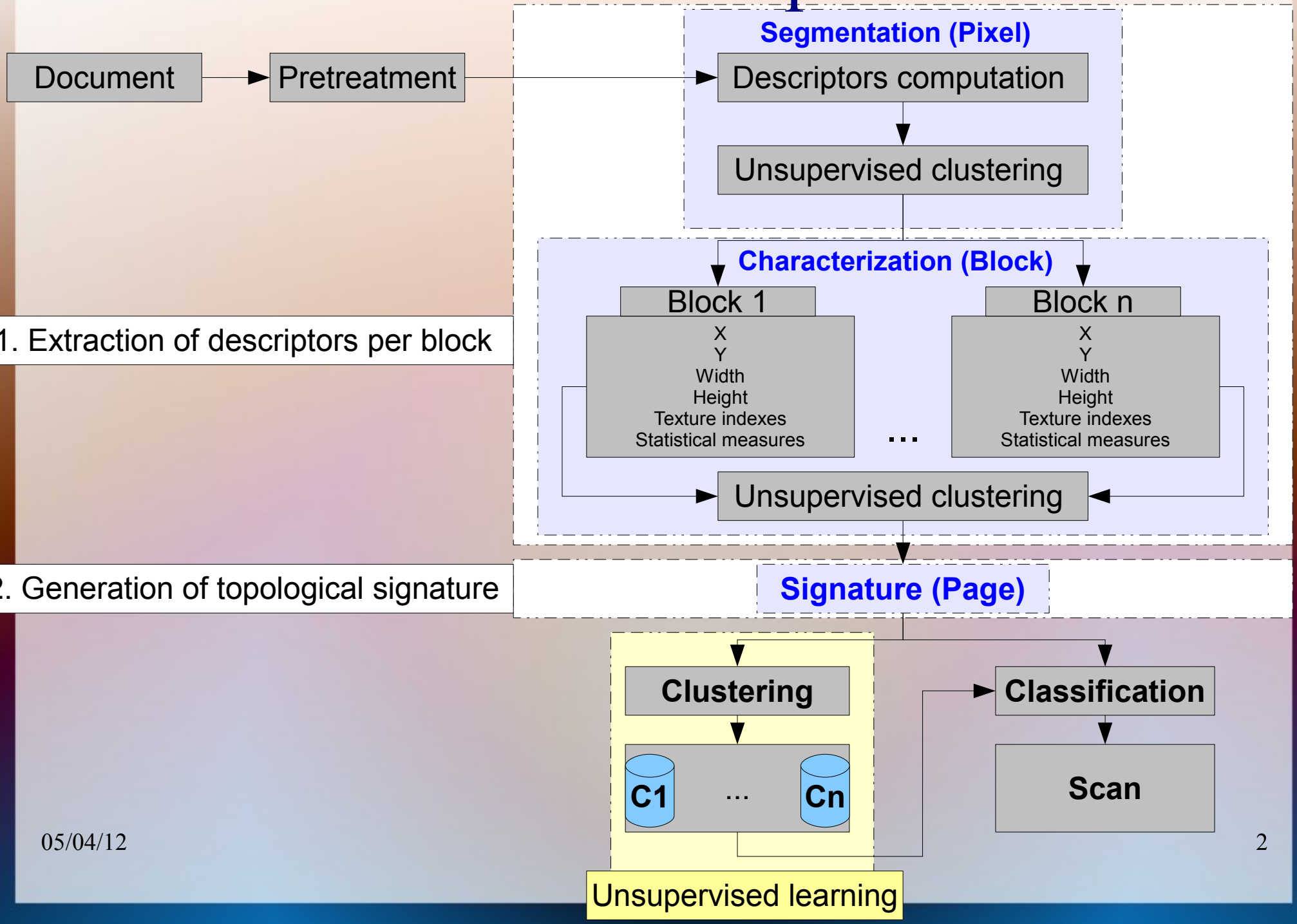
Supervised by: Petra GOMEZ-KRÄMER[1]

Pierre HEROUX[2]

Rémy MULLOT[1]

[1] L3I, University of La Rochelle, Avenue Michel Crépeau, 17042 La Rochelle, France
E-mails: maroua.mehri@univ-lr.fr, petra.gomez@univ-lr.fr, remy.mullot@univ-lr.fr
[2] LITIS, University of Rouen, Avenue de l'Université, 76800 Saint-Etienne-du-Rouvray, France
E-mail: pierre.heroux@univ-rouen.fr

Research topic



Texture attributes of autocorrelation

- **Angle: A_{ij}**
  - Corresponds to the main angle of the rose of directions
- **Autocorrelation intensity: I_{ij}**
  - Corresponds to the intensity of the autocorrelation function for the main orientation found (1st index)
- **Standard deviation: Std_{ij}**
  - Corresponds to the standard deviation of the intensities of the rose, expect for the orientation of maximal intensity
  - $Std_{ij} << 1$: The main orientation is significantly more prevalent than other orientations
  - $Std_{ij} >> 1$:The rose of directions is deformed and a large number of orientations are present in different proportions [Jou06, JRMV08]

Texture attributes of autocorrelation

$$\mathcal{D}^h(I)_0 = 0$$
$$\mathcal{D}^h(I)_n = \sum_{(x,y)\in\Omega} \|I(x,y) - T_h(I,n)(x,y)\|$$

With $T_h(I, \delta)$ is the translation of the gray-scale image I (defined on the plane Ω) of δ pixels along the horizontal axis.

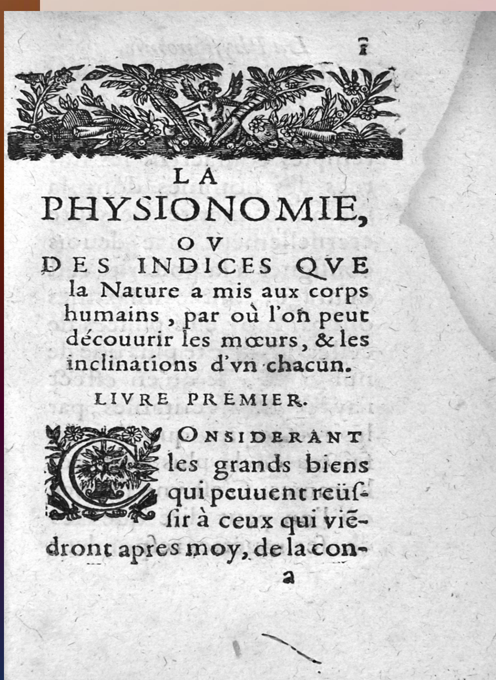
- **Mean stroke width: S_w**
  - Corresponds to an estimation of the mean stroke width of an image computed from the sequence $D^h(I)_n$ until $n = n_m$, when its growth rate becomes lower than 10%. Then, we obtain $S_w = n_m$
- **Mean Stroke height: S_h**
  - Corresponds to an estimation of the mean stroke height of an image computed from the sequence $D^v(I)_n$ until $n = n_m$, when its growth rate becomes lower than 10%. Then, we obtain $S_h = n_m$ [OLL11, OLL12]

Texture attributes computation

- Assigning a feature vector for each given number of pixels of the image using a sliding window
- Fixed-size window, positioned on the left corner of the image
- 5 texture attributes of autocorrelation are calculated
- The point at the center of the window is then marked by the 5 numerical values calculated
- Window is moving from a number of pixels and so on until going through the whole image

Database

- **25 images of old documents**
  - 16 images containing text and graphic areas (drop cap letters, ornaments...) including 7 images are snapshots
  - 8 images with only text zones (text containing many paragraphs characterized by different fonts size) including 7 images are snapshots
  - 1 image with only graphic block

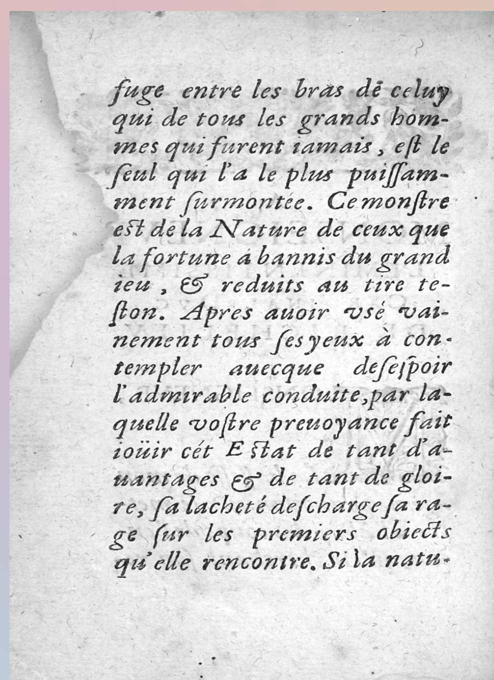


1

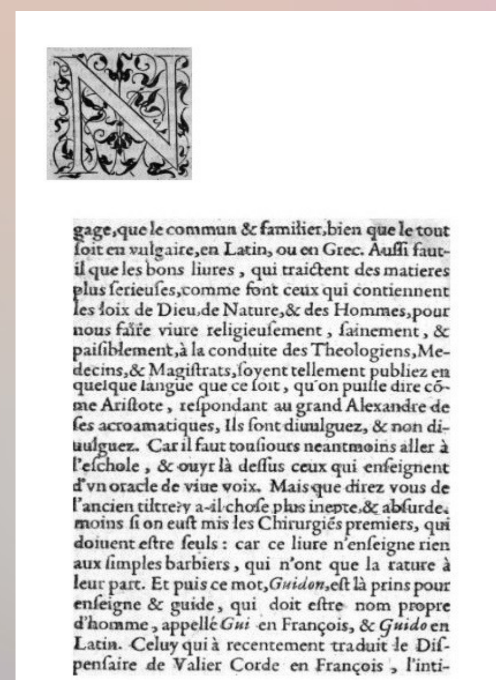
LA
PHYSIONOMIE,
OV
DES INDICES QVE
la Nature a mis aux corps
humains, par où l'on peut
découurir les mœurs, & les
inclinations d'vn chacun.

LIVRE PREMIER.

CONSIDERANT
les grands biens
qui peuuent reüſ-
ſir à ceux qui viẽ-
dront apres moy, de la con-
a

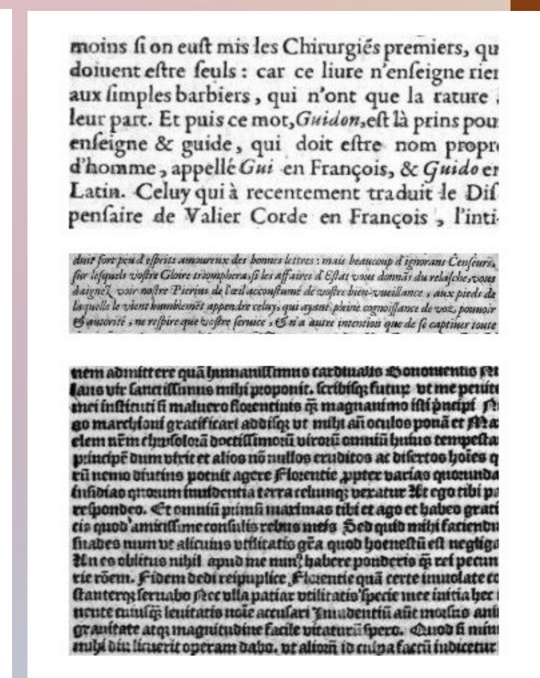


*fuge entre les bras de celuy
qui de tous les grands hom-
mes qui furent iamais, eſt le
ſeul qui l'a le plus puiſſam-
ment ſurmontée. Ce monſtre
eſt de la Nature de ceux que
la fortune a bannis du grand
ieu, & reduits au tire te-
ſton. Apres auoir vsé vai-
nement tous ſes yeux à con-
templer auecque deſeſpoir
l'admirable conduite, par la-
quelle voſtre preuoyance fait
ioüir cét Eſtat de tant d'a-
uantages & de tant de gloi-
re, ſa lacheté deſcharge ſa ra-
ge ſur les premiers obiects
qu'elle rencontre. Si la natu-*



N

gage, que le commun & familier, bien que le tout
ſoit en vulgaire, en Latin, ou en Grec. Auſſi faut-
il que les bons liures, qui traictent des matieres
plus ſerieuſes, comme ſont ceux qui contiennent
les loix de Dieu, de Nature, & des Hommes, pour
nous faire viure religieuſement, ſainement, &
paiſiblement, à la conduite des Theologiens, Me-
decins, & Magiſtrats, ſoyent tellement publiez en
quelque langue que ce ſoit, qu'on puiſſe dire cõ-
me Ariſtote, reſpondant au grand Alexandre de
ſes acroamatiques, Ils ſont diuulguez, & non di-
uulguez. Car il faut touſiours neantmoins aller à
l'eſchole, & ouyr là deſſus ceux qui enſeignent
d'vn oracle de viue voix. Mais que direz vous de
l'ancien tiltre? y a-il choſe plus inepte & abſurde,
moins ſi on euſt mis les Chirurgiẽs premiers, qui
doiuent eſtre ſeuls: car ce liure n'enſeigne rien
aux ſimples barbiers, qui n'ont que la rature à
leur part. Et puis ce mot, *Guidon*, eſt là prins pour
enſeigne & guide, qui doit eſtre nom propre
d'homme, appellé *Gui* en François, & *Guido* en
Latin. Celuy qui à recentement traduit le Diſ-
penſaire de Valier Corde en François, l'inti-



moins ſi on euſt mis les Chirurgiẽs premiers, qu
doiuent eſtre ſeuls: car ce liure n'enſeigne rier
aux ſimples barbiers, qui n'ont que la rature
leur part. Et puis ce mot, *Guidon*, eſt là prins pou
enſeigne & guide, qui doit eſtre nom propr
d'homme, appellé *Gui* en François, & *Guido* er
Latin. Celuy qui à recentement traduit le Diſ
penſaire de Valier Corde en François, l'inti-

*duit fort peu d'eſprits amoureux des bonnes lettres: mais beaucoup d'ignorans Cenſeurs,
ſur leſquels voſtre Gloire triomphera, ſi les affaires d'Eſtat vous donnãs du relaſche, vous
daignez voir noſtre Pierius de l'œil accouſtumé de voſtre bien-vueillance, aux pieds de
laquelle le vient humblemẽt appendre celuy, qui ayant pleine cognoiſſance de voz pouuoir
& authorité, ne reſpire que voſtre ſeruice, & n'a autre intention que de ſe captiuer toute*

nem admittere quã humaniſſimus cardinalis Bononienſis ...
lans vir ſanctiſſimus mihi proponit. ſcribiſq; futur ut me penit...
tuei inſtituti ſi maluero florentinis q̃ magnanimo iſti pncipi ...
go marchioni gratificari addiſq; ut mihi añ oculos ponã et Ma...
elem nrm chryſolorã doctiſſimorũ virorũ omniũ huius tempeſta...
principẽ dum viuit et alios nõ nullos eruditos ac diſertos hoies q...
cũ nemo diutius potuit agere Florentie pp̃ter varias quorunda...
inſidias quorum imprudentia terra celumq; vexatur At ego tibi p...
reſpondeo. Et omniũ primũ maximas tibi et ago et habeo grati...
eis quod amiciſſime conſulis rebus meis Sed quid mihi faciendu...
ſuades num ut alicuius vtilitatis grã quod honeſtũ eſt neglig...
Nã eo oblitus nihil apud me nunq̃ habere ponderis q̃ rei pecun...
rie rõem. Fidem dedi reipuplice Florentie quã certe inuiolate co...
ſtanterq; ſeruabo Nec ulla patiar utilitatis ſpecie mee initia hec ...
ncnte cuiuſq; leuitatis noie accuſari Imprudentiũ añt moriſ ani...
grauitate atq; magnitudine facile vitaturũ ſpero. Quod ſi min...
mihi diu licuerit operam dabo, ut aliorũ id culpa factũ iudicetur

Examples of database images

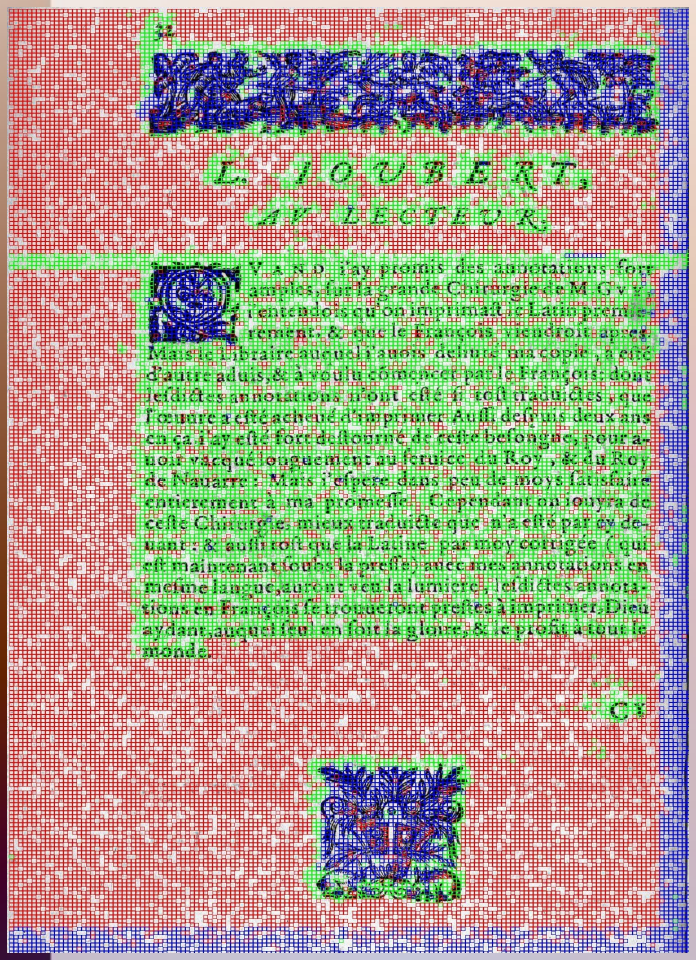
Constraints

- Effectiveness (segmentation and clustering performance)
- Efficiency (computing time)
- Number of clusters
- Large dataset for clustering
- Problem of out of memory

Parameters

- **Multi-resolution**
  - Keep the size of the original image and varying the size of the analysis window
- **Size of window**
  - [8, 16, 32,64]
- **Pixels number of window shift**
  - 8
- **Clustering method**
  - *HAC (Hierarchical Ascendant Clustering)*
  - *Consensus clustering*

Results



L. IOVBERT,
AV LECTEVR.

VAND i'ay promis des annotations fort amples, fur la grande Chirurgie de M. Guy, i'entendois qu'on imprimaſt le Latin premierement, & que le François viendroit apres. Mais le Libraire auquel i'auois donné ma copie, a eu d'autre aduis, & a voulu commencer par le François: dont eſdites annotations n'ont eſté icy traduites, que l'oeuure a eſté achené d'imprimer. Auſſi depuis deux ans en ça i'ay eſté fort deſtourné de cette beſongne, pour auoir vacqué longuement au ſeruice du Roy, & du Roy de Nauarre: Mais i'eſpere dans peu de moys la ſatisfaire entierement à ma promeſſe. Cependant on iouyra de ceſte Chirurgie mieux traduitte que n'a eſté par ey deuant: & auſſi toſt que la Latine par moy corrigée (qui eſt maintenant ſous la preſſe) auec mes annotations en meſme langue, aura veu la lumiere, leſdites annotations en François ſe trouueront preſtes à imprimer, Dieu aydant, auquel ſeul en ſoit la gloire, & le profit à tout le monde.

CY

LES RECEPTES.

Syrop de Guy digerant le phlegme. 619.18

Tablette à rendre sous humeurs. 626.16

Trisphaarmac. 674.3

Theriaque de sing especes. 247.

Trisphaarmac de Galen. 103.29

Trochises pour la douleur de dents auec chaleur. 547.20

Trochics pour la goutte rose. 492.17

Trochises pour la sirasne & inflammation. 512.18

Trochises alkarans. 980.36

Trochises d'esbdeles. 38.10

Trochisc. cahuicon. 681.5

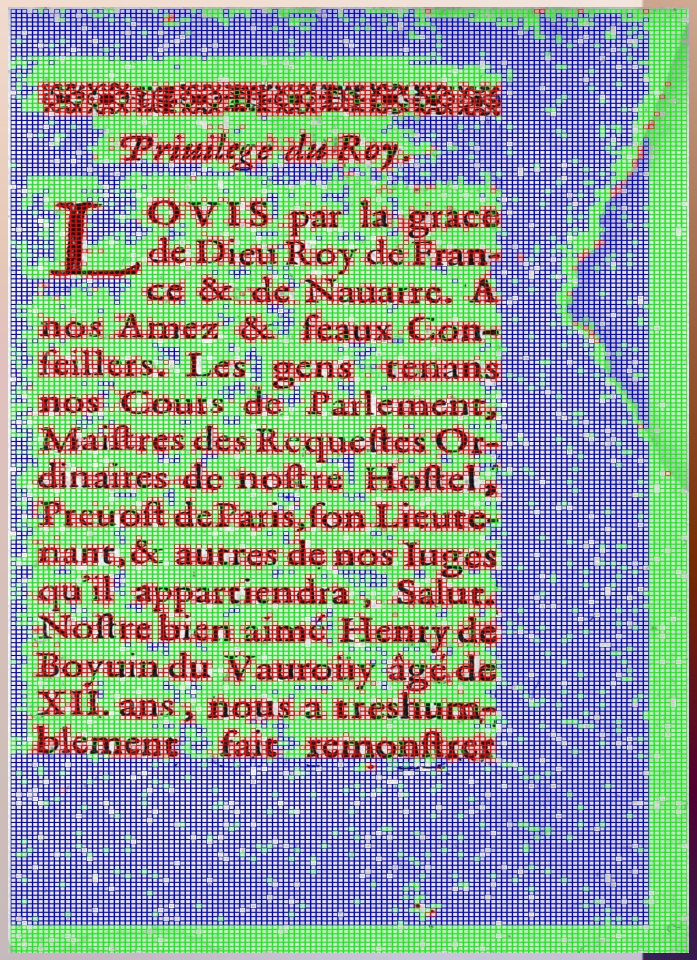
Trochises de berberis. 591.15

Trochisc de saulge. 680.16

Trochises de lacque. 29.35

Trochises narcotiques. 60.36

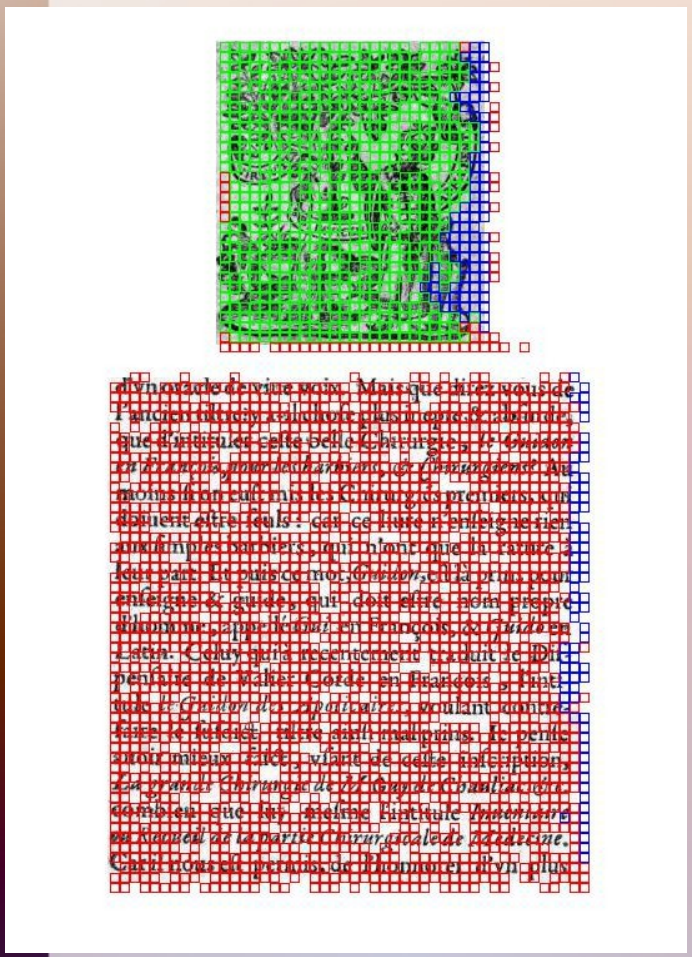
FIN.

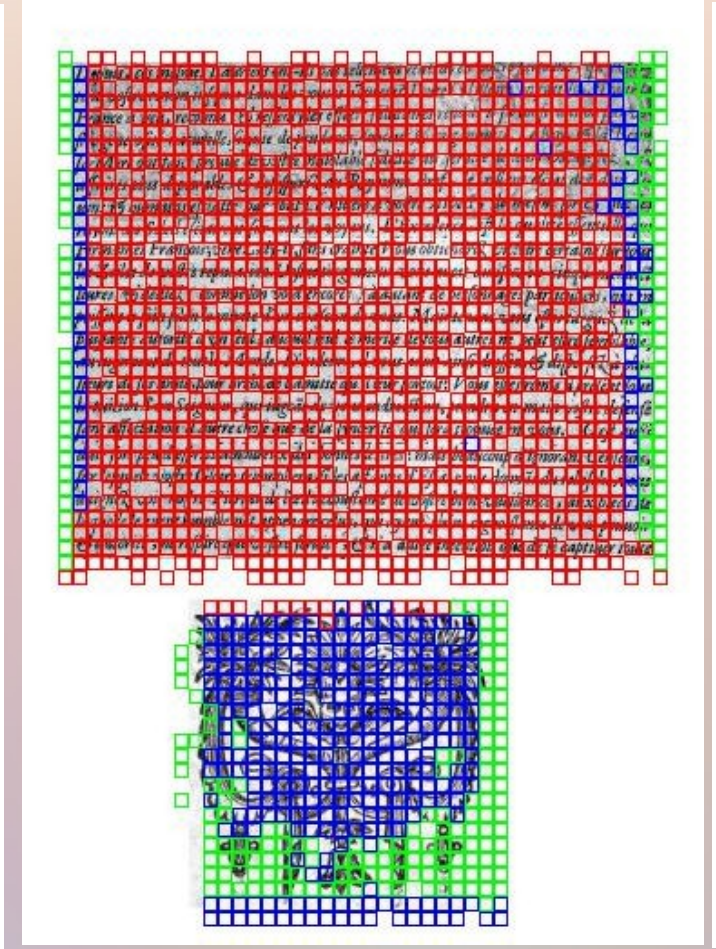


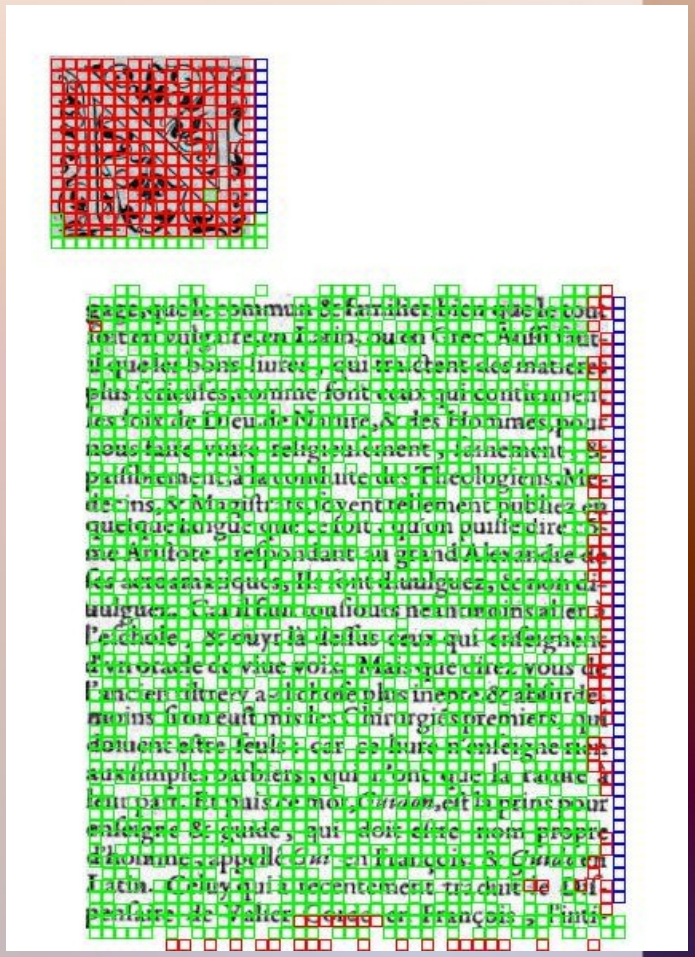
Priuilege du Roy.

LOVIS par la grace de Dieu Roy de France & de Nauarre. A nos Amez & feaux Conſeillers. Les gens tenans nos Cours de Parlement, Maiſtres des Requeſtes Ordinaires de noſtre Hoſtel, Preuoſt de Paris, ſon Lieutenant, & autres de nos Iuges qu'il appartiendra, Salut. Noſtre bien aimé Henry de Boyuin du Vauroüy âgé de XII. ans, nous a treshumblement fait remonſtrer

Results







Future works

- **Consensus clustering and validation of autocorrelation descriptors**
- **Frequency attributes**
  - *First feature:* Characterizes the transitions between paper and ink
  - *Second feature:* Characterizes the white spaces separating the collateral elements
- **Statistical attributes**
  - *Grey Level Co-occurrence Matrix*
  - *Entropy*
  - *Homogeneity degree*
  - *Connection degree*
- **Texture attributes**
  - *Wavelets*

References

[Jou06] Journet, N. Analyse d'images de documents anciens : une approche texture. PhD thesis, L3i, 2006.

[JRMV08] Document Image Characterization Using a Multiresolution Analysis of the Texture: Application to Old Documents. N. Journet, J.Y Ramel, R. Mullot, V. Eglin. International Journal of Document Analysis and Recognition (IJDAR 2008).

[OLL11] Chromatic / achromatic separation in noisy document images. A. Ouji, Y. Leydier, F Lebourgeois. International Conference on Document Analysis and Recognition (ICDAR 2011), Beijing, China. 2011.

[OLL12] Extraction de texte à base de segmentation colorimétrique dans les images de presse. A. Ouji, Y. Leydier, F Lebourgeois. CIFED, Bordeaux, France. 2012.

Thank you for your attention