Visual perception of unitary elements for layout analysis of unconstrained documents in heterogeneous databases

ICFHR 2014
September 2, 2014

Baptiste Poirriez
Aurélie Lemaitre
Bertrand Coüasnon

Irisa / Insa – Intuidoc research group
couasnon@irisa.fr
Context : Heterogeneous Documents

- Document Layout Analysis
  - Unconstrained and Heterogeneous Documents
  - Heterogeneous Databases
  - No a priori on the document

- Maurdor Campaign dataset
  - Founded by the French Ministry of Defense
  - Publicly available in 2014

- 10 000 scanned documents, fully annotated
  - Heterogeneous documents
    - Forms, Tables, Business documents, Correspondences, Faxes
    - Diagrams, Drawings…
  - French (50%), English (25%), Arabic (25%), Mixed
  - Printed, Handwritten, Mixed
    - 1000 writers
Difficulties

- Unconstrained and Heterogeneous Documents
  - Methods usually need some homogeneity inside of the collection of documents
    - Stability of some blocks of text for example

- Heterogeneous Databases
  - Methods usually rely on document classification
  - New class for each new type of document

- Mixing both is complex
  - No a priori on the document
  - No a priori on the database

  - Only very general a priori on the document elements
Proposed Method : Strategy

- Analysis of Salient Elements
  - Use of Perceptive Vision Mechanisms
    ✦ Already used on homogeneous databases
  - Perceptive Vision Principle
    ✦ Some contents are salient for the human vision in a document
      • Often strongly structuring
    ✦ Combining several points of view
      • Prediction/verification mechanism
      • Layout is predicted in a global vision of a document and verified with details.
  - Multi level description of salient elements

- Primitives at different resolutions
  - Line segments
  - Connected components
  - Words recognized by OCR (Abbyy FineReader)
Proposed Method : Strategy

- Detection of orientation
  - Perceptive vision
  - Detection at low resolution of text lines (line segments)
  - Detection of the main direction of writings

- Iteratively find the most structuring and salient elements
  - 1: Tables, separators, boxes
  - 2: Latin printed blocks of text
  - 3: Remove of the detected salient elements : New Segmentation
  - 4: Handwritten and Arabic text lines
  - 5: Graphic regions

- Description of each salient element
  - General knowledge on each element
Proposed Method: Implementation with DMOS-P

- DMOS-P: a generic perceptive method
  - **EPF language**
    - Grammatical description of a document
    - Generation of an adapted recognition system
  - **Validated**
    - Wide range of kind of documents
    - More than 700,000 pages processed
    - On homogeneous documents

- Application of DMOS-P for Heterogeneous Documents
  - **Description with EPF of salient elements**
  - **New Segmentation during parsing**
  - **Combine different levels of perception**
    - Perceptive Layers
  - **Combine different salient elements**
1: Tables, separators, boxes

- Perceptive vision
  - Line segments at different resolution
  - Rulings detection

- General description of a table
  - 2 crossing rulings
  - Table frame on the cross
    - Virtual rulings
  - Set of rows
  - Set of columns
  - Recursive table detection inside each cell

- General description of
  - Boxes
  - Separators
2: Latin printed block of text

- Words detected by OCR (Abbyy Finereader)
  - On the complete document
    - Errors on mixed (printed/handwritten) documents
  - Added in a perceptive layer

- General description of Latin Printed block of text
  - Blocks built on words with a high level of confidence
  - Text lines and blocks
    - With constrains on size, alignments (left, right, centered), text columns
    - Very general knowledge on blocks of text
3: New Segmentation (Remove of Table Rulings, Separators, Latin Printed Text)
Je soussigné : DUPONT Théôse

Principes de l'effacement

Nom : DUPONT
Prénoms : Théôse
Date de naissance : 11.11.11
Lieu de naissance : Strasbourg
Demeurant à : Motheu
Rue : route du Rhin

Déclare acquérir, entrer en possession,
Je demande la délivrance du récépissé,
Je certifie sur l'honneur l'exactitude de

(1) A établir en deux exemplaires.
(2) Rayer les mentions inutiles.
4: Handwritten and Arabic Text lines

■ New Segmentation
  ◆ No more tables, rulings or Printed Latin text
  ◆ Simplify the description of handwritten and Arabic text lines

■ Perceptive Vision
  ◆ Text lines can be seen at low resolution as line segments
  ◆ Confirm the text line at high resolution with regular aligned connected components

■ Description of text lines detects
  ◆ Handwritten text lines: Latin and Arabic
  ◆ Printed text lines: Arabic and some remaining Latin (OCR Errors)

■ Blocks of text made of text lines
Example of Printed/Handwritten document

◆ Words from OCR
◆ Latin Printed Blocks
◆ Blocks after Handwritten text lines detection
5: Graphic Region

- Simple description of Graphics
  - Low resolution
  - Remaining connected components big enough

- Example

```
À l'occasion de ce nouvel an,
Je souhaite à tous les professeurs,
Chers Compagnons, une bonne et heureuse
année 2013.
Que ce nouvel an vous apporte
tant de joie et de progrès dans
vos professions.

J. Leclerc
```
Experimental Evaluation

- Maurdor Campaign, Module 1 (Layout Analysis task)
  - Training set: 6,000 documents
  - Dev set: 1,000 documents
  - Test set: 1,000 documents

- Metrics (see www.maurdor-campaign.org for details)
  - ZoneMap (deals with split and merge): lower better
  - Jaccard (pixel level labeling): higher better

- Results
  - 2nd on ZoneMap
  - Close to best on Jaccard

- Jaccard by class

<table>
<thead>
<tr>
<th>Participant</th>
<th>Text zone</th>
<th>Graphic zone</th>
<th>Table</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participant 1</td>
<td>0.552</td>
<td>0.394</td>
<td>0.363</td>
</tr>
<tr>
<td>Our method</td>
<td>0.553</td>
<td>0.402</td>
<td>0.307</td>
</tr>
<tr>
<td>Participant 2</td>
<td>0.307</td>
<td>0.176</td>
<td>0.174</td>
</tr>
</tbody>
</table>

It is still an open topic, and our system meets a lot of confusing cases. We mainly have difficulties when graphics zones and graphic zones, at pixel level.
Conclusion

■ DLA on Heterogeneous and Unconstrained Documents
  ◆ Difficult task, open problem
  ◆ General grammatical description of document elements
    ✦ Perceptive vision mechanism
  ◆ Iterative recognition of salient elements
  ◆ Ability to re-segment the document during the analysis
  ◆ New application of DMOS-P
    ✦ Up to now applied on homogeneous databases

■ Improvements
  ◆ Too much split and merge
    ✦ Need more homogeneous blocks
    ✦ Avoid confusion between graphics and text
  ◆ Introduce classifiers for adding local information
    ✦ Printed/handwritten detection
    ✦ Language detection
    ✦ Kind of graphics detection
  ◆ Mixing those classifiers with grammatical description