From pixels to content: An overview of the main techniques used in DIA

Jean-Yves RAMEL
From Pixels to Contents

Introduction

Document Image Analysis
Machine Learning
Content based images Retrieval

Tools to automatically extract information inside the images / documents

Generation and utilization of meta data
From Pixels … to contents

Outline

- From pixels…
  - What is an image?
  - Image (pre-)processing

- … to Text
  - Transcription and Layout analysis
  - Segmentation and content extraction
  - An overview of Pattern Recognition

- … but also to non-Text
  - Content characterization and signatures
  - Content retrieval and spotting

- Back to meta-data?
  - From descriptive to perceptual meta-data
  - Is there adequate encoding formats?

- Conclusions and perspectives
From Pixels...
Images come from a grid of microscopic photosensitive cells called **PIXELS**

**Sampling**

- Assignment of a numerical value drawn from the received lighting energy / pixel (grid unit)
- Continuous value \((x_i, y_i)\) → Discrete value \((x_i, y_i)\) → Pixels
- The range of colors that each pixel can take
From Pixels…
What is an image?

Image Quantization

**Binary images:** \( I(i,j) = 0 \) black \ or \ \( I(i,j) = 1 \) white

**Gray level (8 bits/pixel) images:**
\( I(i,j) = 0 \ldots 255 \) from the lighter to the darker.

**Color images (24 bits/pixel):**
3 values of lighting intensity Red, Green, Blue
\( I_1(i,j) = 0 \ldots 255 \quad – \quad I_2(i,j) = 0 \ldots 255 \quad – \quad I_3(i,j) = 0 \ldots 255

Image Representation & Processing

Image = Array(s) of pixels = Matrices of values
1 pixel = A position inside the image \((i,j)\) \ + \ 1 color \ (1 to 3 values)

The values \( I(i,j) \) associated to each pixel \( s(i,j) \) represent their brightness intensity
From Pixels…
What is an image?

- **Sampling ➔ Image resolution**
  - Number of pixels per length unit
  - In dpi (dots per inches) or ppp (points par pouce)
  - When the resolution decreases, the precision decreases

- **VF Image processing**
  - Page A4 = 21x29.7 cm
  - 200 dpi : 1650 x 2340 pixels = 3 861 000 pixels
  - 300 dpi : 3500 x 2480 pixels = 8 680 000 pixels
  - 16M colors, 1 pixel = 3 octets ➔ 10 à 25 Mo/page !
  - **A trade-off between quality-quantity/time is mandatory**
  - Fidelity of the numerical version
  - Mass of storage size – Transmission / Processing time
From Pixels…

Why few pixels are so important?

- Isolated patterns can correspond to small numbers of pixels (from 20 to 50 pixels)!
- 20 to 40% of the useful information is situated on the boundaries of the shapes and is modified randomly by the digitization process.
- Loss of topological information can occur, leading to broken characters or touching characters.
- The human perception is not sensitive to the resolution.

OCR errors: procedure and interpretation of pattern that in spite of...
After the digitization, the images usually still have a lot of defaults:

- Curvature and skew due to scanning
- Noise on boundaries, dots, blur, ...

**From Pixels…**

**What is Image (pre-)processing?**

**Corriger l’image (rotation, wrapping)**
Curvature and skew correction is possible on text images.
From Pixels…

Image (pre-)processing

The problem is more complicated in case of heterogeneous content
The problem is more complicated in case of heterogeneous content.
... to text and layout
1-Offrant appres plusieurs rues sur ce faites
2-au lieu et temps acostume aly livre et
3- accenste la somme et quantite dessouz
4-estempte du voloir et consentement de
5-plusieurs des borgeys de la dite ville
6-de chastellion ______________IIIxxV flor(ins)
7-receu dudit franceys rotier pour
8-une albaleste dehue a cause de la
dite ferme dudit trezein avec la
9-somme et quantite dessus estempte ?
10-_________________Il flor(ins)
11-
12-receu de Alexandre ? escoffier al loup ?
13- dudit lieu de chatellion pour la
14-rense et ferme du commun de la dite
15-ville et communauite ? pour ung an commettre ?
16-albaleste nativite fi jehan baptiste lan
17-quatre cent et cinquante? A ly par le
18-pris dessouls esempt accense et hure ?
19- plus au plus offrant du voloyr et
20-et consentement de falseures des borgeys
21-de la dite ville de chastellion apres
22-plusieurs rues ....refaytes es lieus ?
First step: Image segmentation

- Transformation of the image (set of pixels) into patterns (regions of interest) of higher level (EoC)
- These EoC could be very simple (part of characters) or more sophisticated ones (paragraphs, illustrations, ...)
- EoC extraction: Background (white) / Foreground (black) separation
- Color Image $\rightarrow$ Grayscale $\rightarrow$ binarisation
From pixels ... to text

An overview of OCR mechanisms

Just to illustrate the difficulties...

- Most of the segmentation methods need a binarisation

Threshold selection?
From pixels ... to text
An overview of OCR mechanisms

Just to illustrate the difficulties...

- Most of the segmentation methods need a binarisation
- Global threshold ➔

- Local thresholds ➔

  Niblack: \( S = m + ks^2 \) avec \( k = -0.2 \)
  | \( m \): mean et \( s \): standard deviation
From pixels ... to text

An overview of OCR mechanisms

First step: Image segmentation / Connected components

- Then, we can try to group black pixels together to localize and recognize higher level Element of Content (EoC)
From pixels ... to text
An overview of OCR mechanisms

Next step: Layout analysis
- Connected Components ➔ Words ➔ Lines ➔ Paragraphs ➔ Page
- The results have to be saved in XML format (Alto, ...)
- Choosing how to organize the XML tree (physical / logical) is not so easy...
Next step: Layout analysis

Two kinds of structures have been identified by researchers in DIA:

- The logical structure — the generic one corresponding to a priori knowledge about the content of the document
- The physical structure — the analysed instance corresponding to the extracted EoC inside the image, each one associated to descriptive features (size, position, number of sub-patterns, ...)
- Layout analysis tries to recognize these 2 structures (EoC identification)
From pixels … to text
An overview of OCR mechanisms

Next step: Layout analysis

• The analysis / identification of the EoC is usually achieved based on a rule-based system defined through a grammar (static one) or defined interactively by the users.

Projet Européen Meta-e (http://meta-e.aib.uni-linz.ac.at/) ➔ First commercial system for automatic layout analysis Dutch books of XVIIIe

AGORA: an user-driven system for content extraction in historical printed books

Projet BVH - Paradiit (https://sites.google.com/site/paradiitproject/) ➔
From pixels … to text
An overview of OCR mechanisms

Next step: Pattern / EoC recognition (toward Machine Learning)

How computers can recognize objects?

- We need a large set of (labelled) examples similar to the patterns to be recognized ⇒ a training set
- We need a list of stable and discriminative features (shape, color, size,...) used to describe the patterns (labelled ones and unknown one)
Next step: Pattern / EoC recognition (toward Machine Learning)

How computers can recognize objects?

• When an unknown EoC arrives, we compute its features and compare it with the content of the training set (associated built models)

\[
D(A, ?) = \sqrt{(0 - 0)^2 + (3 - 10)^2 + (0 - 0)^2 + (4 - 5)^2 + \ldots + (3 - 2)^2}
\]

\[
D(A, ?) = 7,48 \text{ et } D(B, ?) = 19,05
\]

[2D Representation of the training set]
From pixels ... to text
An overview of OCR mechanisms

Deep Learning (Conv. Neural Net)
From pixels … to text
An overview of OCR mechanisms

- Why commercial OCR are not working well on historical documents?
  - Noises and degradations
  - Unusual layout
  - Unsuited training set

Fine Reader
From pixels … to text
An overview of OCR mechanisms

- **Why commercial OCR are not working well on historical documents?**
  - Lack of data, knowledge and experiences
    - Unusual fonts and characters ➔ training data needs to be created
    - Unusual languages ➔ Lexicons, dictionaries and language models need to be created

- Context often allows to modify our understanding of what is perceived by our senses
  - Until now, we tried to recognized EoC without using their context
  - The same EoC could be interpreted differently according to its surrounding context
  - Results of OCR are highly correlated to the adequacy of the used word dictionary

- Is there methods that need less a priori knowledge?
- Processing non-Textual parts can be good source of inspiration?

---

(chat! Silence! Le' chat dort)
... to non-text
From pixels ... to non-textual contents
Pictorial Content is also of high interest

- Ornamental letters (+ of 20000)
- Figures (+ de 1500)

From pixels ... to non-textual contents

An overview of Content Based Image Retrieval

**Perceptual meta data instead of classical meta data**

- Computation of signatures for all the images or even **sub-parts of the images (EoC)**
- Computation time not crucial
- **Signatures ➔ Visual Features ➔ Perceptual meta data**

![Diagram](image-url)

**Index database**

![Set of images to index](image-url)
From pixels ... to non-textual contents
An overview of Content Based Image Retrieval

Using images as query instead of words
From pixels ... to non-textual contents

An overview of Content Based Image Retrieval

It is again a question of features...

➔ We speak about signatures
From pixels ... to non-textual contents
An overview of Content Based Image Retrieval

- From CBIR to Word spotting
From pixels ... to non-textual contents
An overview of Content Based Image Retrieval

- From CBIR to Word spotting

Query images

Results: retrieved images
From pixels ... to non-textual contents
An overview of Content Based Image Retrieval

- From CBIR to Word spotting
Is it just a question of meta-data?
The standard model

Descriptive meta-data + transcription

We have usually

- Descriptive meta-data in standard formats (MARC, EAD, Dublin Core, MODS, …)
  - Edited manually
  - “Semantical” information

- Text transcription associated to additional meta-data (TEI)
  - Semi-automatic transcription or manually edited
  - “Semantical” information

“Semantical” meta-data
The standard model
Perceptual meta-data

- It seems that **CBIR can help** to extract and save supplementary information about image content (EoC) without going to the semantical aspect (recognition)
  - Regions of interest
  - Visual features ➔ Perceptual **signatures and index**
  - Shapes, positions, colors, textures, … ➔ Numerical values (vectors)

+ “Perceptual” meta-data
Is it just a problem of Meta-data?
What about the encoding formats?

Structuration of data and file formats is a difficult problem
• **Data architects are needed**

• Some interesting **formats** linked with previous discussions
  o **METS** – Metadata Encoding and Transmission Standard
  o **ALTO** – Analyzed Layout and Text Object
  o **TEI** – Text Encoding Initiative
Is it just a problem of Meta-data?

ALTO for OCR

ALTO = Analyzed Layout and Text Object

- Standard XML
- Created in 2003 during METAe project
- Developed by Graz, Linz, Innsbruck universities
- Description of the content and the physical layout of one page
- Used by several OCR software
- Adapted and used by the BNF and other libraries
- Drawbacks: huge / static
Is it just a problem of Meta-data?

**TEI for transcription and enriched contents**

- **TEI = Text Encoding Initiative - Standard XML**
  - Content tagging and logical structure encoding (full document)
  - Used a lot by libraries
  - Too much « open »? → Quit « complex »
Is it just a problem of Meta-data? Link between meta data?

**METS – Metadata Encoding and Transmission Standard**

- Open XML Standard created in 2001 by the Digital Library Federation maintained by METS Editorial Board
- XSD-Schema

**Linking between multimedia objects**

- Complete Description of digitized content (images, texts, audio, sculptures, ...)
- Physical / logical structures
- Descriptive Meta data (DC, MODS, MARC, ...)
- ...

METS – First level Elements

- `metsHdr` METS document header (info author, software, ...)
- `dmdSec` descriptive metadata section (bibliographic notice)
- `amdSec` administrative metadata section (copyright, ...)
- `fileSec` file inventory section (file localisation)
- `structLink` structural map linking (link between structures)
- `structMap` structural maps (physical et logical)
METS – Physical Structure

METS allows to specify the locations of resource files ➔ File pointers METS (fpotr)

Possible link to the descriptive meta data (dmd Sec)
The Logical StructMap reflects the enrichment of the different logical blocks located in different pages that can be split by other logical blocks (like foot-notes, ....).
From Pixels to Contents

Conclusions

- Building tools for the valorisation of digitized historical content is a pluri-disciplinary task
  - Meta-data production ➔ Experts of the domains
  - Selection and verification of the data ➔ Experts + Data accuratist
  - Structuration of the data and system ➔ Data / system architect
  - Computer vision, Machine learning ➔ Data scientist

- Manual indexing is needed
  - Descriptive meta-data ➔ Semantical meta data
  - Standard formats for data encoding
  - Annotations could be seen as supplementary meta-data?

- Operational methods and tools are available
  - Acquisition devices
  - Automatic tools: low level image processing, OCR
  - Perceptual meta-data should be added : CBIR
From Pixels to contents
Conclusions - Perspectives

- The actual context: big data and heterogeneous collections
  - Connexion between data, mutual enrichment, interoperability
  - Introduction and management of additional knowledge
  - Facing the diversity of the types of contents and usages

- Quality of the interaction instead of only the quantity
  - Semantic Web: queries reformulation, smart crawlers, automatic categorisation
Thanks...

https://sites.google.com/site/paradiitproject/