IMAGE RETRIEVAL IN DIGITAL LIBRARIES

A LARGE SCALE MULTICOLLECTION EXPERIMENTATION OF MACHINE LEARNING TECHNIQUES

Jean-Philippe MOREUX

Guillaume CHIRON (L3i, La Rochelle)
Outline

• Image Search in DLs
• ETL (Extract, Transform, Load) approach on World War 1 theme
• Machine Learning experimentation:
  • Image Genres Classification
  • Visual Recognition
• Image Retrieval PoC
• Conclusion

« L’Auto », photo lab, 1914
Our Users are Looking for Images

On gallica.bnf.fr:

• **63%** of the users consult the image collection, **85%** know its existence [2017 survey]

• **50% of the Top 500 user queries** contain named entities Person, Place, Historical Event [2016 analysis of 28M user queries]

• For these encyclopedic queries, giving users access to iconographic resources could be a **valuable service**

• But the Gallica image collection **only contains 1.2 M items**: silence, limited number of illustrations (only 140 results for "Georges Clemenceau«, 1910-1920)
DLs are full of Images!

- **1.2M pages** manually indexed and tagged as "image" (photos, engravings, maps...)
- **Huge reservoir of potential images** growing at a 20M digitized pages/year pace

To make these assets visible to users, we need automation:

- automatic recognition of images
- automatic description of images
For Printed Content, OCR can help

... to identify illustrations
And for other Materials?

- **Enlighted manuscripts, documents with no OCR**: image detection algorithms
- **Video**: each frame is an image

Google TensorFlow Object Detection API

Bayerische Staatsbibliothek Image-based Similarity Search, 43 M images indexed on morphological features
We Have Million of Images…

… but image retrieval is challenging…

• Content based image retrieval (CBIR) is still a scientific challenge
• Heritage images are often stored in data silos of various types (drawings, engravings, photos…) which may need specific CBIRs
• DLs catalogs don’t handle image metadata (size, color, quality, etc.) at the illustration granularity
CBIR: Other Issues to Keep in Mind

- **Different image retrieval use cases** must be considered:
  - Similarity search based on the selection of a source image
  - Content indexing with keywords

- **Various users needs**, from mining of pictures for social media reuse to scientific study of bindings

- **Usability**: DLs web apps have been designed for searching on catalog records and full text. They are **page based**
The page paradigm is an obstacle

Dix-sept dessins de George Barbier sur le *Cantique des Cantiques*, 1914

Classic page flip mode for browsing heritage documents
Newspapers have multiple illustrations per page and double page spread illustrations...particularly for newspapers.
Proof of Concept

- **Extract-Transform-Load** approach
- **On World War 1 materials**: still images, newspapers, magazines, monographs, posters (1910-1920)
- Enriched with **Machine Learning** techniques

![ETL approach diagram](image_url)
The Tool Bag

- **Standard tools and APIs**
- **Machine Learning**: Software as a Service (IBM Watson API) & pretrained models (Google TensorFlow)

<table>
<thead>
<tr>
<th>Extract</th>
<th>Transform</th>
<th>Load</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Gallica APIs</td>
<td>- Watson (IBM)</td>
<td>- BaseX</td>
</tr>
<tr>
<td>- OAI-PMH</td>
<td>- TensorFlow (Google)</td>
<td>- XQuery</td>
</tr>
<tr>
<td>- SRU</td>
<td>- IIIF</td>
<td>- IIIF</td>
</tr>
<tr>
<td></td>
<td>- Tesseract</td>
<td>- Mansory.js</td>
</tr>
</tbody>
</table>

➤ The glue: Perl and Python scripts
Extract

• All the available metadata from our data sources: catalog records, images, OCR, ToC
Extract: remarks

• This first step is worth the pain: it gives access to “invisible” illustrations to user! (invisible= deeply hide into the printed content)

• Challenges:
  • heterogeneity of formats, digitization practices and metadata available (e.g. image genres)
  • computationally intensive (but parallelizable)
  • noisy results for newspapers (≈50-70% of the illustrations are noise)

« Der Rosenkavalier » premiere in Dresden (Richard Strauss, Hugo von Hofmannsthal), L’Excelsior, 27/01/1911
Volumes

• ≈ **300k** (usable) illustrations (on ≈900k) illustrations extracted from 490k pages. Bibliographic selection (WW1) and samples of the newspapers collection

➤ **Just a scratch on the digital collections!**

• On the same time period, Gallica offers **490k** illustrations

➤ **Over the entire digital collection, we can expect hundreds of M of illustrations!**

• **Newspapers** are (really) generous…
  
  *(L’Excelsior: 90k illustrations, 3 ill./page)*

WW1 images database: sources of the images
Transform & Enrich

- **OCR around illustration** (if no text is available): Tesseract
- **Topic modeling**: semantic network, LDA (Latent Dirichlet Allocation)
- **Image genres classification**: TensorFlow/Inception-v3 model
- **Image content recognition**: Watson/Visual Recognition API
Image Genres Classification with TensorFlow

- **Machine learning approach based on Convolutional Neural Networks**: Google Inception-V3 model (1,000 classes, Top 5 error rate: 3.46%)

- **Retrained** (only the last layer, “transfer learning” approach) on our GT dataset of 12 classes, 7,750 img)

- **Evaluated** on a 1,950 images dataset

- **Retraining**: ≈ 3-4 hours

- **Labeling**: < 1s / image
Image Genres Classification with TensorFlow

- **Recall**: 0.90
- **Accuracy**: 0.90

Better performances can be obtained on less generic models (e.g. monographs only: recall=94%) or with full trained models (needs computing power).

The "noisy illustrations" can be removed: cover & blank pages from portfolios; text, ornaments & ads from newspapers.
Image Genres Classification: Filtering

- Data mining raw OCR of newspapers can make you sick!

→ Full-scale test on a newspaper title (6,000 ill.): 98.3% of the noisy illustrations are identified
Image Genres Classification: Q&A

- **Better performances** can be obtained on less generic models (e.g. monographs only: F-measure=94%) or with full trained models.

- **Real life** use for newspapers?

A **98.3% filtering rate means**:
  - ≈900 noisy illustrations are **missed** on a 50,000 pages newspaper title
  - ≈900 valuable illustrations are **removed**… but these ones can be (quickly) checked by humans!

A **94% classification rate means**:
  - 6 illustrations are **missclassified** every 100, but far more less in real life, as we have (sometimes) genre metadata in our catalogs
  - Drawings or photos are classified as engravings, comics as drawing, etc. Not a big deal!

FULL-SCALE USE IS REALISTIC FOR DLs
CBIR: Introduction

- Historically, Content Based Image Retrieval (CBIR) systems were designed to:
  1. Extract **visual descriptors** from an image,
  2. Deduce a **signature** from it and…
  3. Search for similar images by minimizing the distances into the signatures space
CBIR: Introduction

- The **constraint** that CBIR systems can only be queried by a source image (or a sketch drawn by the user) has a negative impact on its **usability**.

- Now, deep learning techniques tend to overcome these limitations, in particular thanks to **visual recognition** of objects in images, which enables **textual queries**.

> IBM Watson Visual Recognition, Google TensorFlow Object Detection
Visual Recognition with IBM Watson

• Visual Recognition Service API
• Outputs pairs of class/confidence score
• Detects objects, persons, faces, colors…

```
"images": [
  {
    "classifiers": [
      {
        "classes": [
          {
            "class": "armored personnel carrier",
            "score": 0.568,
            "type_hierarchy": "/vehicle/wheeled vehicle/armored vehicle/armored personnel carrier"
          },
          {
            "class": "armored vehicle",
            "score": 0.576
          },
          {
            "class": "wheeled vehicle",
            "score": 0.705
          },
          {
            "class": "vehicle",
            "score": 0.706
          },
          {
            "class": "personnel carrier",
            "score": 0.541,
            "type_hierarchy": "/vehicle/wheeled vehicle/personnel carrier"
          },
          {
            "class": "fire engine",
            "score": 0.526,
            "type_hierarchy": "/vehicle/wheeled vehicle/truck/fire engine"
          },
          {
            "class": "truck",
            "score": 0.526
          },
          {
            "class": "structure",
            "score": 0.516
          },
          {
            "class": "Army Base",
            "score": 0.511,
            "type_hierarchy": "/defensive structure/Army Base"
          },
          {
            "class": "defensive structure",
            "score": 0.512
          },
          {
            "class": "gas pump",
            "score": 0.5,
            "type_hierarchy": "/mechanical device/pump/gas pump"
          },
          {
            "class": "pump",
            "score": 0.5
          },
          {
            "class": "mechanical device",
            "score": 0.501
          },
          {
            "class": "black color",
            "score": 0.905
          },
          {
            "class": "coal black color",
            "score": 0.691
          }
        ]
      }
    ]
  }...
```

« Les tanks de la bataille de Cambrai, la reine d'Angleterre écoute les explications données par un officiers anglais », 1917
Experimentation on Person Detection

- **Ground truth** of 2,200 images for Person detection. 600 images for Soldier detection.
  - “Person”: recall=60.5%, accuracy=98.4%
  - With a WW1 custom classifier: recall=65%
  - “Soldier”: recall=56%, accuracy=80.5%

- Modest rates but we’ve to keep in mind that Person or Soldier categories are **not available in catalog records**!
  - Keyword Search on the Soldier GT: "soldier" OR "military officer" OR "gunner" OR…: recall=21% (65 images)
  - Visual Recognition: recall=56% (172 images)
  - Mixed Search (text+visual): recall=70% (215 images)
Experimentation on Soldier Detection

<table>
<thead>
<tr>
<th>Recall</th>
<th>0%</th>
<th>20%</th>
<th>40%</th>
<th>60%</th>
<th>80%</th>
<th>100%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Text MD only</td>
<td>Visual Reco.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Mixed MD</td>
<td>+ custom classifier</td>
<td></td>
</tr>
</tbody>
</table>

- 70%
Experimentation on Person Detection

- **Works on heritage documents:** engraving, drawing, printed photography, even on "difficult" ones
Experimentation on Person Detection

• But this experiment also exposes some limitations:

  • Generalization from a contemporary training corpus:
    \(\rightarrow\) anachronisms

  • Generalization from a limited training corpus:
    \(\rightarrow\) classification errors (1,000 classes is enough for encyclopedic search, not for the large spectrum of cultural heritage artifacts we cure)

  • Difficulty to handle complex scenes (multiclasses)
Experimentation on Face Detection

- The Watson API also performs **Face** and **Gender** detection:
  - “Face”: recall=30%, accuracy=99.8% / “Gender”: recall=22%
  - The mixed use of the two recognition APIs (Person and Face Detection) results in an improvement of the overall recall for **Person detection** from 60.5% to 65%
Experimentation on Face Detection

What are the use cases?

- **Digital Humanities:** gender studies, visual studies
- **Digital Mediation:**

- **Arts and DH:** Time Based Image Averaging
  - “Robots Reading Vogue”

“Gallica WW1 Portrait Gallery”

“Seb Przd, Time covers, 1923-2006”
Load (& Search)

- In a XML database (BaseX)
- Search with XQuery
- Display with IIIF
Image Retrieval: the Data Deluge

• The complexity of the search form and the large number of results it often leads to reveal that searching and browsing in image databases carries specific issues of usability and remains a research topic in its own right…
Retrieval: Encyclopedic Query on a Named Entity

- **Textual descriptors** (metadata and OCR) are used.

  “George Clemenceau” query: 140 ill. in Gallica/Images, >1,000 in the WW1 DB

Caricatures can be found with the “drawing” facet
Retrieval: Image Metadata Query

- **Image descriptors** are used.
  
  Search for **large illustrations**: maps, double spread page, posters, comics…

- Search for musical score covers with **red-dominant color**
Retrieval: Encyclopedic Query on Concept

- The **conceptual classes** extracted by the Watson API are used. Query on the superclass “vehicle” returns many instances of its subclasses (car, bicycle, airplane, airship, etc.)
Retrieval: Query on Concepts

• Search for visuals of a gun inside a bunker: 
  `class="bunker" AND class="gun"`

Textual metadata for this image is:
« Canon camouflé dans une casemate et soldat français »

« Casemate » is an aged synonym of bunker, blockhaus ➔ Classification
overcomes language dependant issues
Retrieval: Mixed Query

- **Conceptual classes, text** and **image MD** are used

Search for visuals relating to the urban destruction following the Battle of Verdun: `class=("street" OR "house") AND keyword="Verdun"`
Retrieval: Mixed Query

- Search for visuals of military vehicles used in French colonies
  class="wheeled vehicle" AND keyword=("sand" OR "dune")

« L'Aérosable », L'Aviation et l'automobilisme militaires : revue mensuelle des progrès scientifiques appliqués à la Défense nationale, 1914

(The image in the middle is a false positive)
Retrieval: Mixed Query

- Study of the *evolution of the French soldiers uniforms* during the conflict. The aim is to document the history of the famous *red trousers* worn until beginning of 1915.
- Based on two queries using the conceptual classes (“soldier”, “officer”, etc.), record metadata (date), and an image-based criteria (“color”).

*date > 01/01/1915*
Retrieval: Mixed Query

- Same use case: evolution of aeronautical techniques during the conflict

The illustrations provided by these queries could feed on averaging of images approaches, which increasingly escape the artistic sphere to address other subjects or other uses (e.g. automatic dating of photographs)
Opening the Data

The IIIF Presentation API provides a way to describe the illustrations in a document using Open Annotations attached to a layer (Canvas) in the IIIF manifest.

```json
{
  "@id": "http://wellcomelibrary.org/iiif/b28047345/annos/contentAsText/a31i0",
  "@type": "oa:Annotation",
  "motivation": "oa:classifying",
  "resource": {
    "@id": "dctypes:Image",
    "label": "Picture"
  },
  "on": "http://mylibrary.org/iiif/b28047345/canvas/c31#xywh=201,1768,2081,725"
}
```

All the iconographic resources can then be operated by machine (library-specific projects, data harvesting (Europeana), research, hacker/makers/social networks
Conclusion

• **Unified access to all illustrations** in an encyclopedic digital collection is an innovative service that meets a real **need**.

• It will foster the illustrations **reuse**

• It also opens new perspective for **researchers** (DH, visual studies)

• The maturity of **modern AI techniques** in image content processing makes possible their integration into the digital library toolbox.

• Their results, even imperfect, help to **make visible and searchable** the large quantities of illustrations of our collections.
Digital Humanities focus

• Today, the image is a new playground for DH researchers
• Tomorrow, **image datasets will be the daily life of researchers**
• **AI tools** will be free and trivialized
• Heritage libraries will be solicited for their **iconographic collections** (web archive, photo collections, newspapers and magazines, etc.) for **visual data mining**

**Note:** Datasets, scripts and code are available: https://altomator.github.io/Image_Retrieval/