docExtractor: An off-the-shelf historical document element extraction

Tom Monnier and Mathieu Aubry
LIGM, Ecole des Ponts, Univ Gustave Eiffel, CNRS
Motivation

Massive digitalization

Need easy-to-use and efficient automatic tools for document analysis!
Motivation

Text detection for dataset A

Text detection for dataset B

Photo extraction for dataset C

Drawing extraction for dataset D

Dans son atelier ciréulent on se reposent plusieurs modèles mus, hommes et femmes.

Déliçieux compo(6,0),(995,992)(204,186),(516,816)(21,195),(208,815)(791,230),(990,821)(382,176),(685,824)(693,174),(990,821)(12,189),(220,812)
Motivation

Manual data annotation

Text detection for dataset A

Text detection for dataset B

Photo extraction for dataset C

Drawing extraction for dataset D

...
Motivation

Pb: Systems are problem-specific and require dedicated annotations
Our idea

Synthetic data with annotations

Generic

Text detection

Illustration extraction

Single system that only uses synthetic data for robust real data applications
Contributions

1. A synthetic document generation pipeline with fine GT
Contributions

1. A synthetic document generation pipeline with fine GT

2. A new system with impressive performances on unseen data
Contributions

1. A synthetic document generation pipeline with fine GT

2. A new system with impressive performances on unseen data

3. A new public dataset for illustration segmentation (IlluHisDoc)
Our approach

1. Synthetic document generator for line-level page segmentation

2. Improved U-Net like segmentation network with a ResNet backbone
Synthetic document generation pipeline

1

2

3
Synthetic document generation pipeline

1. Background composition

2

3
Synthetic document generation pipeline

1. Background composition

2. Content generation

3.
Synthetic document generation pipeline

1. Background composition
2. Content generation
3. Noise addition
Background composition

177 empty pages
Background composition

177 empty pages

1 Color
Background composition

177 empty pages

1. Color
2. Double page
177 empty pages

Color

1

Double page

2

Contextualization

3

15 context images
Content generation
Content generation

1. Draw layout
Content generation

1. Draw layout

2. Paste elements
Content generation

1. Draw layout

2. Paste elements
Element types

*image*: WikiArt dataset (8,1k)
Element types

**image:** WikiArt dataset (8,1k)

**glyph:** 91 decorated fonts + random letter
**Element types**

**image:** WikiArt dataset (8,1k)

**glyph:** 91 decorated fonts + random letter

**drawing:** synthetic drawings from Wikipedia images + augmentations

**text:** random Wikipedia content + 405 fonts + augmentations
Element types - *drawing*

Generation
Element types - *drawing*

1. Line detection

Generation
Element types - *drawing*

1. Line detection

2. Augmentations (contrast, color, blur)
Element types - *drawing*

1. **Line detection**
2. **Augmentations** (contrast, color, blur)

**Generation**

**Labeling = bounding shapes** (with closing operations)
Element types - *text*

405 fonts downloaded from the web

Random text from Wikipedia

5 layout constraints
*caption, floating-word, paragraph, table, title*
(line length, font size, position, rotation, …)
Element types - text

- 405 fonts downloaded from the web
- Random text from Wikipedia
- Hello World!
- 5 layout constraints:
  - caption, floating-word, paragraph, table, title
  - (line length, font size, position, rotation, ...)

32
Element types - *text*

405 fonts downloaded from the web

Random text from Wikipedia

5 layout constraints
  *caption*, *floating-word*, *paragraph*, *table*, *title*
  (line length, font size, position, rotation, ...)

Hello World!

Hello World!

你好，世界!

Hello World!

Hello World!

strike through, justification, ...
Element types - text

Hello World!
你好, 世界!
Hello World!
Hello World!
Hello World!

5 layout constraints
caption, floating-word, paragraph, table, title
(line length, font size, position, rotation, ...)

405 fonts downloaded from the web
Random text from Wikipedia

Labeling = x-height + border
Element types - text

Hello World!
你好，世界!
Hello World!

5 layout constraints
caption, floating-word, paragraph, table, title
(line length, font size, position, rotation, ...)

Labeling = x-height + border

ascent
ascender height
cap height
mean line
median
baseline
dx-height
descent
descender height
Element types - text

Hello World!
你好，世界!
Hello World!
Hello World!

405 fonts downloaded from the web
Random text from Wikipedia

5 layout constraints
caption, floating-word, paragraph, table, title
(line length, font size, position, rotation, ...)

Labeling = x-height + border

<table>
<thead>
<tr>
<th>ascent</th>
<th>ascender height</th>
</tr>
</thead>
<tbody>
<tr>
<td>cap height</td>
<td>median</td>
</tr>
<tr>
<td>mean line</td>
<td>x-height</td>
</tr>
<tr>
<td>descent</td>
<td>descender height</td>
</tr>
</tbody>
</table>

x-height generalizes better & borders increase perf

<table>
<thead>
<tr>
<th>text label</th>
<th>border label</th>
<th>Simple</th>
<th>Complex</th>
<th>cBAD2017</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>✓</td>
<td>0.663</td>
<td>0.719</td>
<td>0.637</td>
</tr>
<tr>
<td>baseline</td>
<td></td>
<td>0.714</td>
<td>0.771</td>
<td>0.678</td>
</tr>
<tr>
<td>x-height</td>
<td>✓</td>
<td>0.749</td>
<td>0.724</td>
<td>0.758</td>
</tr>
<tr>
<td>x-height</td>
<td></td>
<td>0.900</td>
<td>0.812</td>
<td>0.829</td>
</tr>
</tbody>
</table>
Noise addition
Noise addition

1. Gaussian blur
Noise addition

1. Gaussian blur

2. Structured noise (lines or random shapes)
Noise addition

1. Gaussian blur

2. Structured noise (lines or random shapes)

3. Bleed-through
SynDoc examples
SynDoc examples
SynDoc examples
SynDoc examples
SynDoc examples
SynDoc examples
SynDoc examples
SynDoc examples
SynDoc examples
SynDoc examples
Segmentation method

**Network** = U-Net with ResNet-18 backbone

* indicates max-pooling replacement with conv3x3
Segmentation method

Network = U-Net with ResNet-18 backbone

* indicates max-pooling replacement with conv3x3
Segmentation method

**Network** = U-Net with ResNet-18 backbone

* indicates max-pooling replacement with conv3x3

**Generic post-processing** = remove small regions using area threshold
Evaluations

Tasks

- **Baseline detection:**
  \[ \Rightarrow cBaD2017 \& cBaD2019 \text{ datasets} \]

- **Illustration segmentation:**
  \[ \Rightarrow \text{Mandragore, RASM2019 \& IlluHisDoc (new) with IoU} \]
IlluHisDoc dataset

IlluHisDoc = new dataset (400 imgs) for illustration segmentation

Printed documents (P)

Illuminated manuscripts (MSI)

Manuscripts with scientific diagrams (MSS)

Manuscripts with drawings (MSD)
# Baseline detection results

## Results for cBAD2017 Dataset

<table>
<thead>
<tr>
<th>Method</th>
<th>Training set used</th>
<th>Simple Track</th>
<th>Complex Track</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P-val</td>
<td>R-val</td>
<td>F-val</td>
<td>P-val</td>
<td>R-val</td>
<td>F-val</td>
</tr>
<tr>
<td>Tesseract4</td>
<td>0.396</td>
<td>0.545</td>
<td>0.459</td>
<td>0.322</td>
<td>0.520</td>
<td>0.398</td>
</tr>
<tr>
<td><strong>Ours (off-the-shelf)</strong></td>
<td><strong>0.871</strong></td>
<td><strong>0.930</strong></td>
<td><strong>0.900</strong></td>
<td><strong>0.844</strong></td>
<td><strong>0.782</strong></td>
<td><strong>0.812</strong></td>
</tr>
<tr>
<td>LITIS [9, 12]</td>
<td>✓</td>
<td>0.780</td>
<td>0.836</td>
<td>0.807</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>IRISA [9]</td>
<td>✓</td>
<td>0.883</td>
<td>0.877</td>
<td>0.880</td>
<td>0.692</td>
<td>0.772</td>
</tr>
<tr>
<td>UPVLC [9]</td>
<td>✓</td>
<td>0.937</td>
<td>0.855</td>
<td>0.894</td>
<td>0.833</td>
<td>0.606</td>
</tr>
<tr>
<td>BYU [9]</td>
<td>✓</td>
<td>0.878</td>
<td>0.907</td>
<td>0.892</td>
<td>0.773</td>
<td>0.820</td>
</tr>
<tr>
<td>dhSegment [16]</td>
<td>✓</td>
<td>0.88</td>
<td>0.97</td>
<td>0.92</td>
<td>0.79</td>
<td><strong>0.95</strong></td>
</tr>
<tr>
<td>DMRZ [9, 13]</td>
<td>✓</td>
<td>0.973</td>
<td>0.970</td>
<td>0.971</td>
<td>0.854</td>
<td>0.863</td>
</tr>
<tr>
<td>Planet [17]</td>
<td>✓</td>
<td><strong>0.98</strong></td>
<td><strong>0.98</strong></td>
<td><strong>0.978</strong></td>
<td><strong>0.93</strong></td>
<td>0.92</td>
</tr>
<tr>
<td><strong>Ours (fine-tuned)</strong></td>
<td>✓</td>
<td><strong>0.948</strong></td>
<td><strong>0.978</strong></td>
<td><strong>0.963</strong></td>
<td><strong>0.883</strong></td>
<td><strong>0.947</strong></td>
</tr>
</tbody>
</table>

## Results for cBAD2019 Dataset

<table>
<thead>
<tr>
<th>Method</th>
<th>Training set used</th>
<th>P-val</th>
<th>R-val</th>
<th>F-val</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tesseract4</td>
<td>0.442</td>
<td>0.552</td>
<td>0.491</td>
<td></td>
</tr>
<tr>
<td><strong>Ours (off-the-shelf)</strong></td>
<td><strong>0.844</strong></td>
<td><strong>0.815</strong></td>
<td><strong>0.829</strong></td>
<td></td>
</tr>
<tr>
<td>Baseline [11]</td>
<td>✓</td>
<td>0.773</td>
<td>0.743</td>
<td>0.758</td>
</tr>
<tr>
<td>TJNU [11]</td>
<td>✓</td>
<td>0.852</td>
<td>0.885</td>
<td>0.868</td>
</tr>
<tr>
<td>UPVLC [11]</td>
<td>✓</td>
<td>0.911</td>
<td>0.902</td>
<td>0.907</td>
</tr>
<tr>
<td>DMRZ [11]</td>
<td>✓</td>
<td>0.925</td>
<td>0.905</td>
<td>0.915</td>
</tr>
<tr>
<td>Planet [11]</td>
<td>✓</td>
<td><strong>0.937</strong></td>
<td>0.926</td>
<td><strong>0.931</strong></td>
</tr>
<tr>
<td><strong>Ours (fine-tuned)</strong></td>
<td>✓</td>
<td>0.920</td>
<td><strong>0.931</strong></td>
<td>0.925</td>
</tr>
</tbody>
</table>

### Two insights
## Baseline detection results

### RESULTS FOR cBAD2017 DATASET

<table>
<thead>
<tr>
<th>Method</th>
<th>Training set used</th>
<th>Simple Track</th>
<th>Complex Track</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P-val</td>
<td>R-val</td>
<td>F-val</td>
</tr>
<tr>
<td>Tesseract4</td>
<td>0.396</td>
<td>0.545</td>
<td>0.459</td>
</tr>
<tr>
<td><strong>Ours (off-the-shelf)</strong></td>
<td><strong>0.871</strong></td>
<td><strong>0.930</strong></td>
<td><strong>0.900</strong></td>
</tr>
<tr>
<td>LITIS [9], [12]</td>
<td>✓</td>
<td>0.780</td>
<td>0.836</td>
</tr>
<tr>
<td>IRISA [9]</td>
<td>✓</td>
<td>0.883</td>
<td>0.877</td>
</tr>
<tr>
<td>UPVLC [9]</td>
<td>✓</td>
<td>0.937</td>
<td>0.855</td>
</tr>
<tr>
<td>BYU [9]</td>
<td>✓</td>
<td>0.878</td>
<td>0.907</td>
</tr>
<tr>
<td>dhSegment [16]</td>
<td>✓</td>
<td>0.88</td>
<td>0.97</td>
</tr>
<tr>
<td>DMRZ [9], [13]</td>
<td>✓</td>
<td>0.973</td>
<td>0.970</td>
</tr>
<tr>
<td>Planet [17]</td>
<td>✓</td>
<td><strong>0.98</strong></td>
<td><strong>0.98</strong></td>
</tr>
<tr>
<td><strong>Ours (fine-tuned)</strong></td>
<td>✓</td>
<td><strong>0.948</strong></td>
<td><strong>0.978</strong></td>
</tr>
</tbody>
</table>

### RESULTS FOR cBAD2019 DATASET

<table>
<thead>
<tr>
<th>Method</th>
<th>Training set used</th>
<th>P-val</th>
<th>R-val</th>
<th>F-val</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tesseract4</td>
<td></td>
<td>0.442</td>
<td>0.552</td>
<td>0.491</td>
</tr>
<tr>
<td><strong>Ours (off-the-shelf)</strong></td>
<td></td>
<td><strong>0.844</strong></td>
<td><strong>0.815</strong></td>
<td><strong>0.829</strong></td>
</tr>
<tr>
<td>Baseline [11]</td>
<td>✓</td>
<td>0.773</td>
<td>0.743</td>
<td>0.758</td>
</tr>
<tr>
<td>TJNU [11]</td>
<td>✓</td>
<td>0.852</td>
<td>0.885</td>
<td>0.868</td>
</tr>
<tr>
<td>UPVLC [11]</td>
<td>✓</td>
<td>0.911</td>
<td>0.902</td>
<td>0.907</td>
</tr>
<tr>
<td>DMRZ [11]</td>
<td>✓</td>
<td>0.925</td>
<td>0.905</td>
<td>0.915</td>
</tr>
<tr>
<td>Planet [11]</td>
<td>✓</td>
<td><strong>0.937</strong></td>
<td><strong>0.926</strong></td>
<td><strong>0.931</strong></td>
</tr>
<tr>
<td><strong>Ours (fine-tuned)</strong></td>
<td>✓</td>
<td>0.920</td>
<td><strong>0.931</strong></td>
<td>0.925</td>
</tr>
</tbody>
</table>

### Two insights

1. Impressive results with synthetic data only!
## Baseline detection results

### RESULTS FOR cBAD2017 DATASET

<table>
<thead>
<tr>
<th>Method</th>
<th>Training set used</th>
<th>Simple Track</th>
<th></th>
<th>Complex Track</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P-val</td>
<td>R-val</td>
<td>F-val</td>
<td>P-val</td>
<td>R-val</td>
</tr>
<tr>
<td>Tesseract4</td>
<td>0.396</td>
<td>0.545</td>
<td>0.459</td>
<td>0.322</td>
<td>0.520</td>
</tr>
<tr>
<td>Ours (off-the-shelf)</td>
<td><strong>0.871</strong></td>
<td><strong>0.930</strong></td>
<td><strong>0.900</strong></td>
<td><strong>0.844</strong></td>
<td><strong>0.782</strong></td>
</tr>
<tr>
<td>LITIS [9], [12]</td>
<td>✓</td>
<td>0.780</td>
<td>0.836</td>
<td>0.807</td>
<td>-</td>
</tr>
<tr>
<td>IRISA [9]</td>
<td>✓</td>
<td>0.883</td>
<td>0.877</td>
<td>0.880</td>
<td>0.692</td>
</tr>
<tr>
<td>UPVLC [9]</td>
<td>✓</td>
<td>0.937</td>
<td>0.855</td>
<td>0.894</td>
<td>0.833</td>
</tr>
<tr>
<td>BYU [9]</td>
<td>✓</td>
<td>0.878</td>
<td>0.907</td>
<td>0.892</td>
<td>0.773</td>
</tr>
<tr>
<td>dhSegment [16]</td>
<td>✓</td>
<td>0.88</td>
<td>0.97</td>
<td>0.92</td>
<td>0.79</td>
</tr>
<tr>
<td>DMRZ [9], [13]</td>
<td>✓</td>
<td>0.973</td>
<td>0.970</td>
<td>0.971</td>
<td>0.854</td>
</tr>
<tr>
<td>Planet [17]</td>
<td>✓</td>
<td><strong>0.98</strong></td>
<td><strong>0.98</strong></td>
<td><strong>0.978</strong></td>
<td><strong>0.93</strong></td>
</tr>
<tr>
<td>Ours (fine-tuned)</td>
<td>✓</td>
<td>0.948</td>
<td>0.978</td>
<td>0.963</td>
<td>0.883</td>
</tr>
</tbody>
</table>

### RESULTS FOR cBAD2019 DATASET

<table>
<thead>
<tr>
<th>Method</th>
<th>Training set used</th>
<th>P-val</th>
<th>R-val</th>
<th>F-val</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tesseract4</td>
<td></td>
<td>0.442</td>
<td>0.552</td>
<td>0.491</td>
</tr>
<tr>
<td>Ours (off-the-shelf)</td>
<td></td>
<td><strong>0.844</strong></td>
<td><strong>0.815</strong></td>
<td><strong>0.829</strong></td>
</tr>
<tr>
<td>Baseline [11]</td>
<td>✓</td>
<td>0.773</td>
<td>0.743</td>
<td>0.758</td>
</tr>
<tr>
<td>TJNU [11]</td>
<td>✓</td>
<td>0.852</td>
<td>0.885</td>
<td>0.868</td>
</tr>
<tr>
<td>UPVLC [11]</td>
<td>✓</td>
<td>0.911</td>
<td>0.902</td>
<td>0.907</td>
</tr>
<tr>
<td>DMRZ [11]</td>
<td>✓</td>
<td>0.925</td>
<td>0.905</td>
<td>0.915</td>
</tr>
<tr>
<td>Planet [11]</td>
<td>✓</td>
<td><strong>0.937</strong></td>
<td>0.926</td>
<td><strong>0.931</strong></td>
</tr>
<tr>
<td>Ours (fine-tuned)</td>
<td>✓</td>
<td>0.920</td>
<td><strong>0.931</strong></td>
<td>0.925</td>
</tr>
</tbody>
</table>

### Two insights

1. Impressive results with synthetic data only!
2. Using training set leads to results on par with SotA
Data-efficient initialization for fine-tuning
**Data-efficient initialization for fine-tuning**

![Graph](image)

*SynDoc initialization* leads to **considerably better results**, with large gaps for less than 20 training samples.
## Illustration segmentation results (mIoU)

<table>
<thead>
<tr>
<th>Method</th>
<th>Training</th>
<th>Mandra.</th>
<th>RASM</th>
<th>IlluHisDoc</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>avg</td>
<td>P</td>
<td>MSS</td>
</tr>
<tr>
<td>Tesseract4</td>
<td></td>
<td>14.8</td>
<td>41.4</td>
<td>9.2</td>
</tr>
<tr>
<td>M-RCNN</td>
<td>PubLay.</td>
<td>11.5</td>
<td>34.3</td>
<td>3.1</td>
</tr>
<tr>
<td>Ours</td>
<td>PubLay.</td>
<td>24.3</td>
<td>57.7</td>
<td>16.9</td>
</tr>
<tr>
<td>M-RCNN</td>
<td>SynDoc</td>
<td>55.4</td>
<td>93.5</td>
<td>60.3</td>
</tr>
<tr>
<td>Ours</td>
<td>SynDoc</td>
<td>76.1</td>
<td>97.2</td>
<td>61.8</td>
</tr>
</tbody>
</table>

### Two insights

---

Illustration segmentation results (mIoU)

<table>
<thead>
<tr>
<th>Method</th>
<th>Training</th>
<th>Mandra.</th>
<th>RASM</th>
<th>IlluHisDoc</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>avg</td>
</tr>
<tr>
<td>Tesseract4</td>
<td></td>
<td>17.2</td>
<td>6.0</td>
<td>14.8</td>
</tr>
<tr>
<td>M-RCNN</td>
<td>PubLay.</td>
<td>9.8</td>
<td>4.2</td>
<td>11.5</td>
</tr>
<tr>
<td><strong>Ours</strong></td>
<td>PubLay.</td>
<td><strong>18.3</strong></td>
<td><strong>14.8</strong></td>
<td><strong>24.3</strong></td>
</tr>
<tr>
<td>M-RCNN</td>
<td>SynDoc</td>
<td>72.3</td>
<td>36.9</td>
<td>55.4</td>
</tr>
<tr>
<td><strong>Ours</strong></td>
<td>SynDoc</td>
<td><strong>86.6</strong></td>
<td><strong>71.0</strong></td>
<td><strong>76.1</strong></td>
</tr>
</tbody>
</table>

Two insights

1. Our segmentation approach provides much better results than detection-based method.

Illustration segmentation results (mIoU)

<table>
<thead>
<tr>
<th>Method</th>
<th>Training</th>
<th>Mandra</th>
<th>RASM</th>
<th>IlluHisDoc avg</th>
<th>P</th>
<th>MSS</th>
<th>MSI</th>
<th>MSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tesseract4</td>
<td></td>
<td>17.2</td>
<td>6.0</td>
<td>14.8</td>
<td>41.4</td>
<td>9.2</td>
<td>2.0</td>
<td>6.5</td>
</tr>
<tr>
<td>M-RCNN</td>
<td>PubLay</td>
<td>9.8</td>
<td>4.2</td>
<td>11.5</td>
<td>34.3</td>
<td>3.1</td>
<td>1.5</td>
<td>7.2</td>
</tr>
<tr>
<td>Ours</td>
<td>PubLay</td>
<td>18.3</td>
<td>14.8</td>
<td>24.3</td>
<td>57.7</td>
<td>16.9</td>
<td>5.3</td>
<td>17.2</td>
</tr>
<tr>
<td>M-RCNN</td>
<td>SynDoc</td>
<td></td>
<td></td>
<td>72.3</td>
<td>93.5</td>
<td>60.3</td>
<td>40.3</td>
<td>27.6</td>
</tr>
<tr>
<td>Ours</td>
<td>SynDoc</td>
<td>86.6</td>
<td>71.0</td>
<td>76.1</td>
<td>97.2</td>
<td>61.8</td>
<td>76.8</td>
<td>68.5</td>
</tr>
</tbody>
</table>

Two insights

1. Our segmentation approach provides much better results than detection-based method
2. Training on SynDoc provides much better generalization than PubLayNet

Illustration segmentation results - example
Illustration segmentation results - example
Web application - demo

https://enherit.paris.inria.fr
BnF. Département des Manuscrits. Grec 2736
Future work - failure cases

Improve synthetic generation with **more elements** and **more advanced augmentations**!
A single and ready-to-use system that leverages synthetic data to efficiently extract text lines and illustrations from real historical documents!
Conclusion

docExtractor

A single and ready-to-use system that leverages synthetic data to efficiently extract text lines and illustrations from real historical documents!

Project webpage:  http://imagine.enpc.fr/~monniert/docExtractor/
Code:              https://github.com/monniert/docExtractor
Web application:   https://enherit.paris.inria.fr