

Synthetic data generation for automatic document layout analysis

SIFED Nancy - 6th June 2019

Mélodie Boillet

Christopher Kermorvant



Overview

- Document layout analysis
- Problematic and objectives
- Experimental data
- Reference system : dhSegment network
- Evaluation of *Yang et al.* generation algorithm
- Developpement of a new complex data generation method
- Conclusion

Document layout analysis

DLA = identification and categorization of the elements of a document according to their typology and relations

- First step of an automatic document understanding workflow
- Application example : locate the texts of a document to apply a text recognition algorithm
- Current state of the art models are based on Deep Learning and need large annotated training sets

Example of document and the result of its DLA :


Document

Development Update

WEST CHESTER TOWNE CENTRE

Four major projects are beginning to shape West Chester Towne Centre—the planned downtown area for West Chester Township.


The Square at Union Centre



Planned to be a gathering place for the township, The Square at Union Centre is nearing completion. The \$2.2 million project is aimed at creating a space similar to Cincinnati's Fountain Square to hold community events and concerts.

When completed, the park will include a large pond, plaza space, restroom facilities, and a central clock/bell tower. The clock tower was designed and built by The Verdin Company of Cincinnati and is to be the main focal point for the Towne Center.

The project is being paid for through a Tax Increment Financing (TIF) district established in the Union Centre area. The West Chester/Liberty Community Foundation has created an endowment fund for the park to be used for annual upkeep and operations.



Segmentation

section
Development Update

WEST CHESTER TOWNE CENTRE

section


text

Four major projects are beginning to shape West Chester Towne Centre—the planned downtown area for West Chester Township.

section

The Square at Union Centre

figure



text

Planned to be a gathering place for the township, The Square at Union Centre is nearing completion. The \$2.2 million project is aimed at creating a space similar to Cincinnati's Fountain Square to hold community events and concerts.


text

When completed, the park will include a large pond, plaza space, restroom facilities, and a central clock/bell tower. The clock tower was designed and built by The Verdin Company of Cincinnati and is to be the main focal point for the Towne Center.

text

The project is being paid for through a Tax Increment Financing (TIF) district established in the Union Centre area. The West Chester/Liberty Community Foundation has created an endowment fund for the park to be used for annual upkeep and operations.

figure

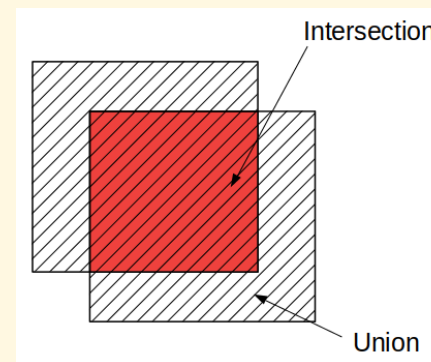


How to generate synthetic data for training deep learning DLA models ?

- Train a state-of-the-art deep learning DLA system on synthetic data
- Evaluate different document generation models

Test data and references

- DSSE-200 : database of 200 annotated real documents from magazines and academic papers created by [Xiao Yang et al. CVPR2017](#) :
"Learning to Extract Semantic Structure from Documents Using Multi-modal Fully Convolutional Neural Network"
- Reference result (deep network using image and text) : **73.30 %** of mean Intersection-over-Union (mIoU)

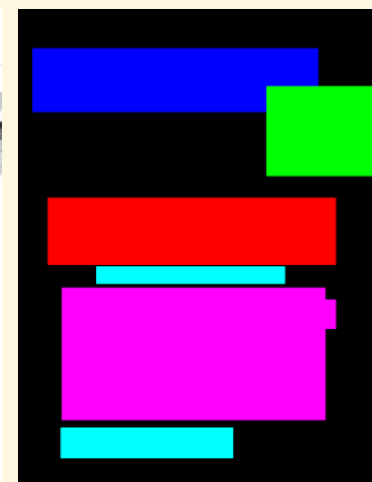
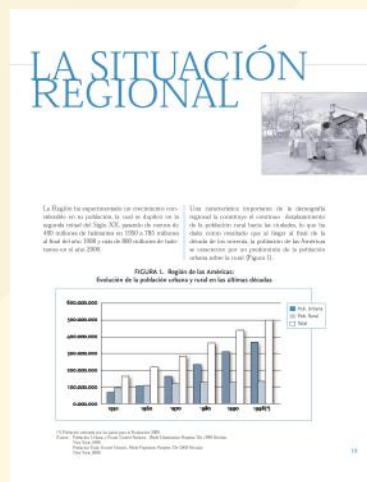
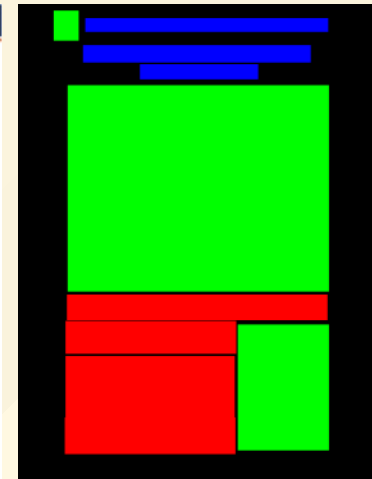


- Open-source document generation algorithm

Example of documents from DSSE-200 :

Documents

Annotations

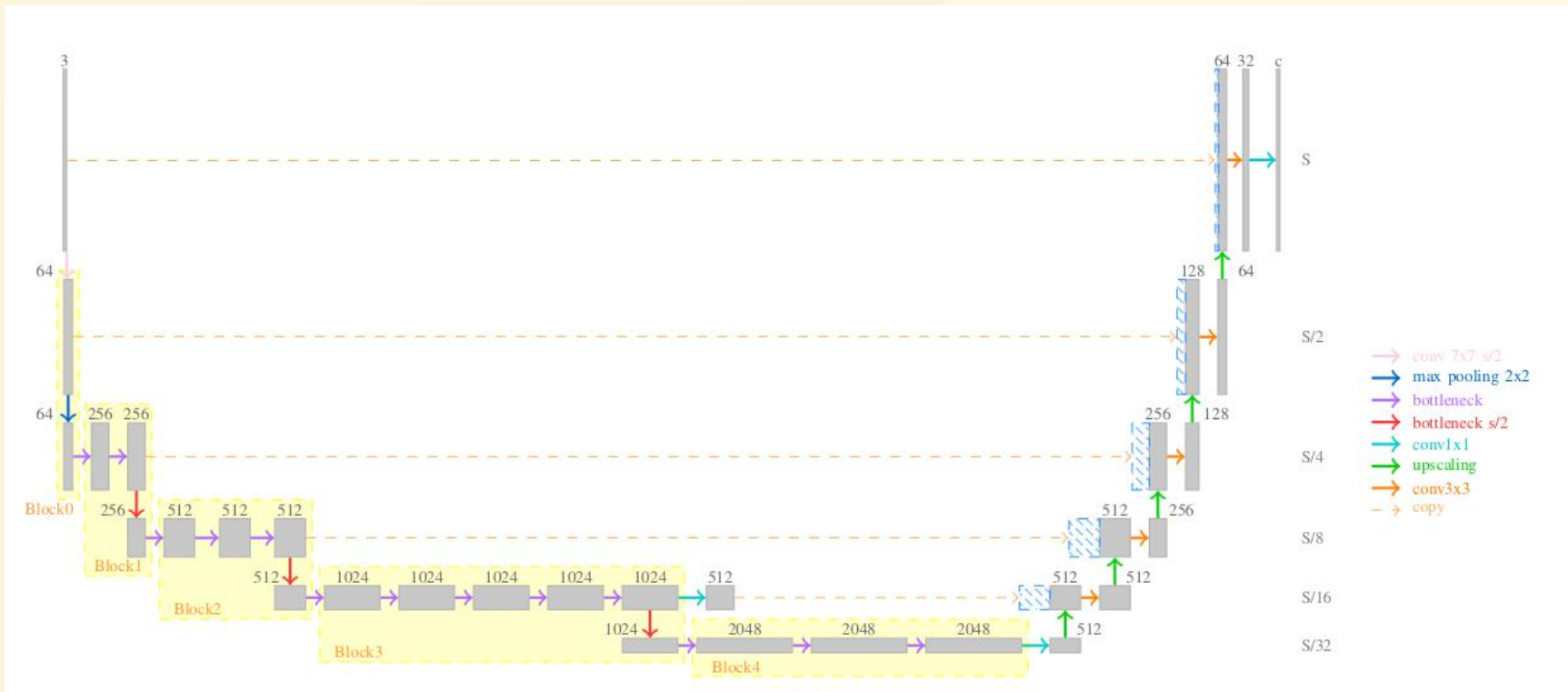


Reference system: dhSegment

- System developed by [Oliveira et al.](#) based on Deep Neural networks
- Page extraction, baseline extraction, DLA, etc...
- Pros :
 - Multitasking
 - Open-source
 - Competitive results on historical documents

dhSegment

- Input : document images
- Output : annotation image where each pixel is assigned a certain class



dhSegment architecture

Evaluation of *Yang et al.* generation algorithm

Experiment 1

- System : dhSegment
- Database :
 - Training and validation : synthetic data generated with Yang algorithm (*DSSE-augmYang*)
 - Test : idem

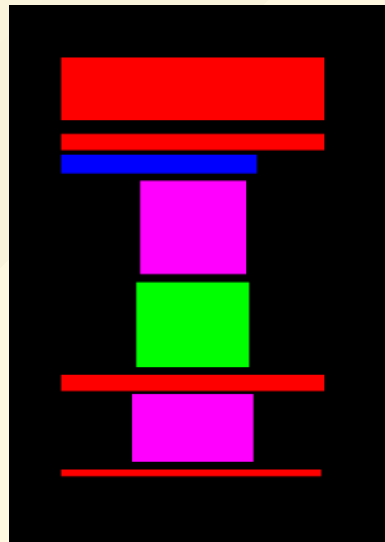
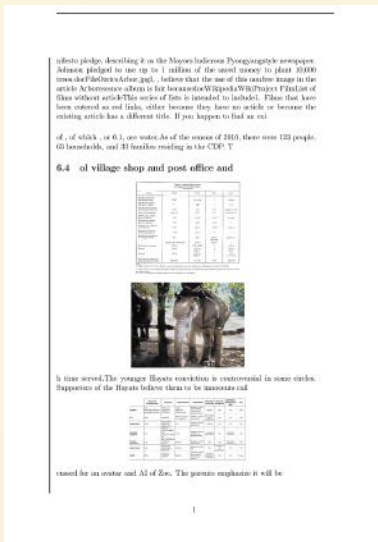
DSSE-augmYang

- 5000 synthetic data
- Data generated according to the generation algorithm of [Yang et al.](#)

Example of documents DSSE-augmYang :

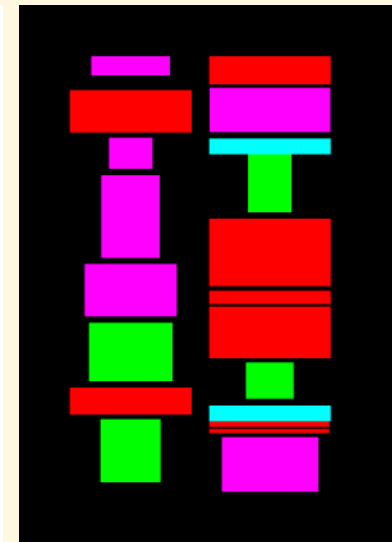
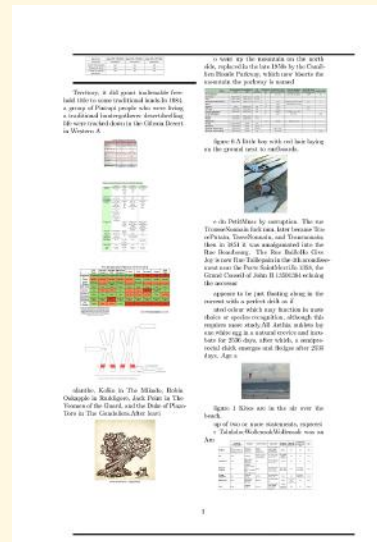
Document

Annotation



Document

Annotation



Results Experiment 1

	Experiment 1
<i>Training and Testing set</i>	DSSE-augmYang
<i>Splitting</i>	3840 / 550 / 1093
<i>mIoU</i>	79.5 %
<i>mIoU with P.P.</i>	96.3 %

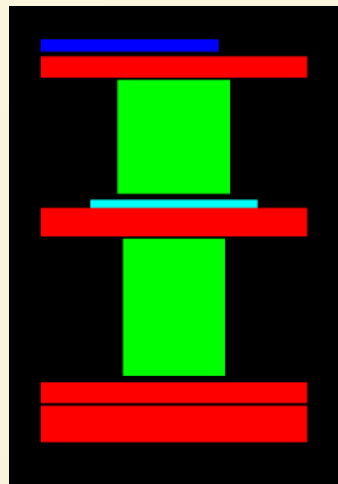
- Very good results
- Generated data allow for training and testing this dhSegment efficiently
- **but** not tested on real data : the test base is different from the *Yang et al.* (DSSE-200)

Example of documents, their annotations and predictions :

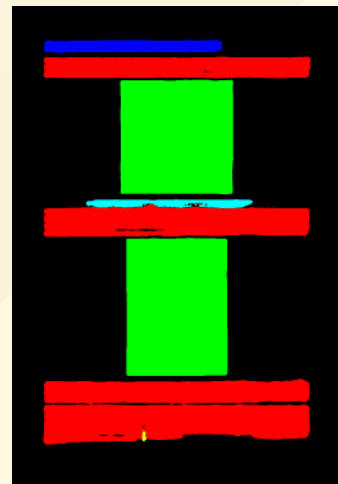
Documents



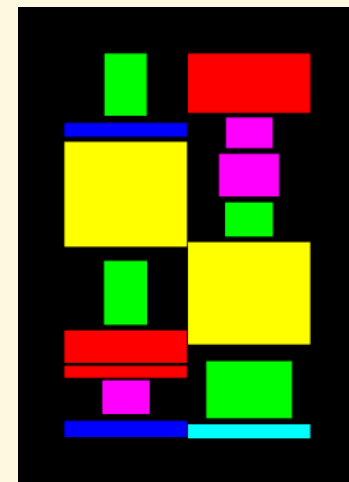
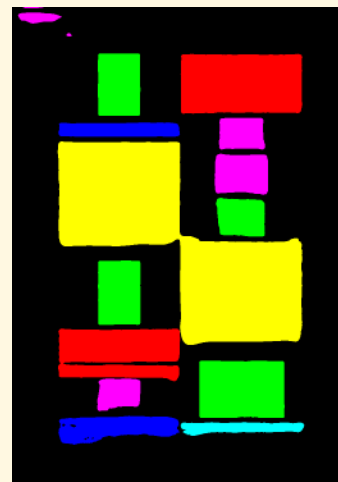
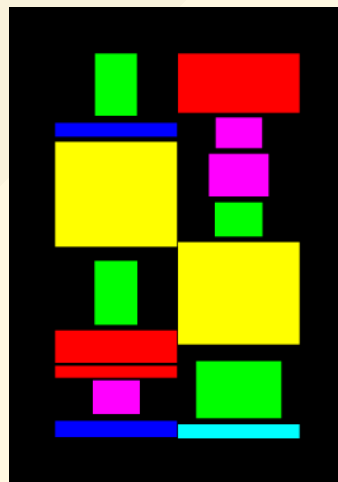
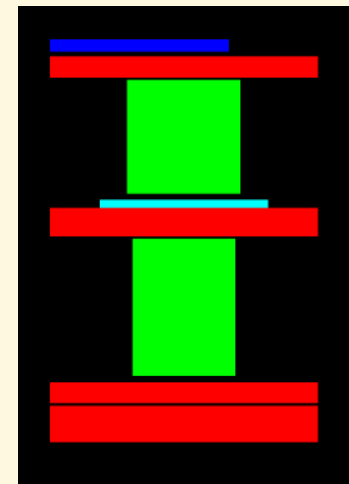
Annotations



Predictions



Predictions with post-proc



Experiment 2

- System : dhSegment
- Database :
 - Training and validation : DSSE-augmYang
 - Test : DSSE-200
- Goal : **73.30 %** mIoU

Results Experiment 2

	Experiment 1	Experiment 2
<i>Training set</i>	DSSE-augmYang	DSSE-augmYang
<i>Testing set</i>	DSSE-augmYang	DSSE-200
<i>Splitting</i>	3840 / 550 / 1093	2000 / 15 / 200
<i>mIoU</i>	79.5 %	22.3 %
<i>mIoU with P.P.</i>	96.3 %	46.6 %
<i>Yang et al. 2018</i>	73.3 %	

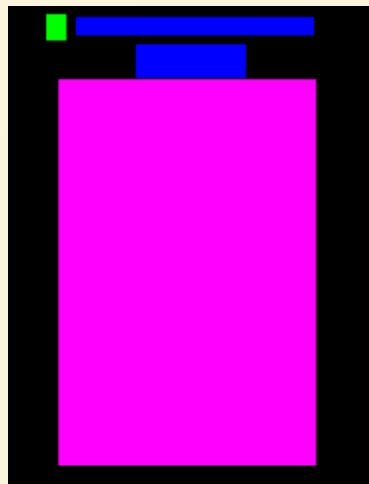
- Result not competitive but *Yang et al.* trained their model over ~130 000 documents.
- Training data less complex than testing data
- Goal : generate more complex documents

Example of documents, their annotations and predictions

Documents



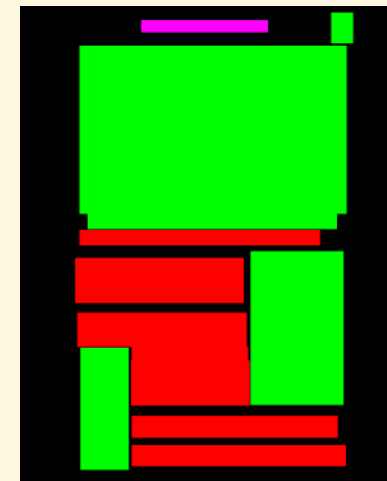
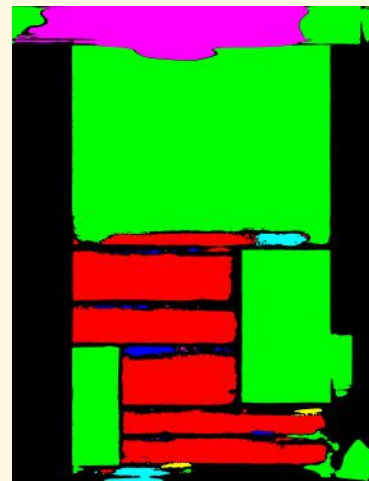
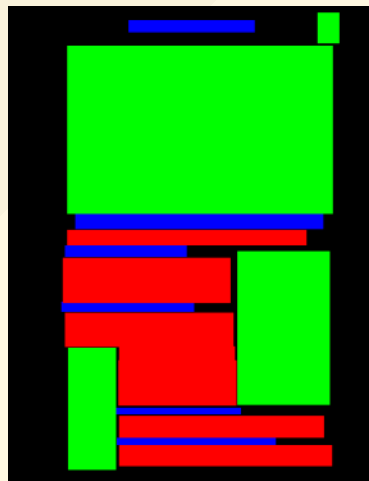
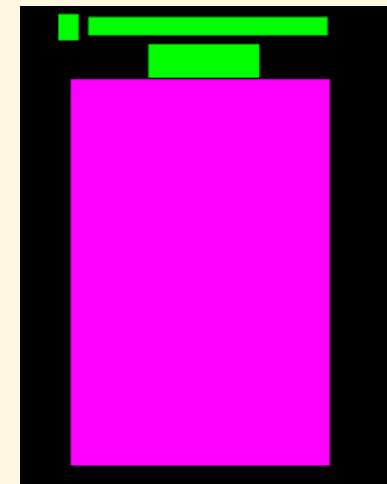
Annotations



Predictions



Predictions with post-proc

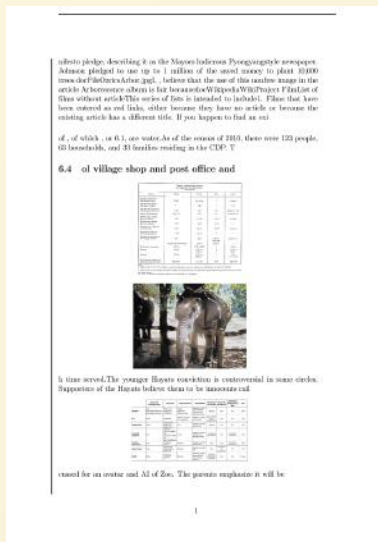


Developpement of a complex document generation method

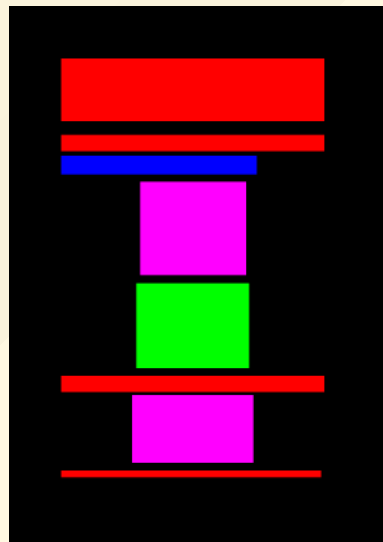
Problems with DSSE-augmYang

- Similar documents layouts : all documents have 1, 2 or 3 columns
- Too simple documents layouts : sequence of texts and figures within a column

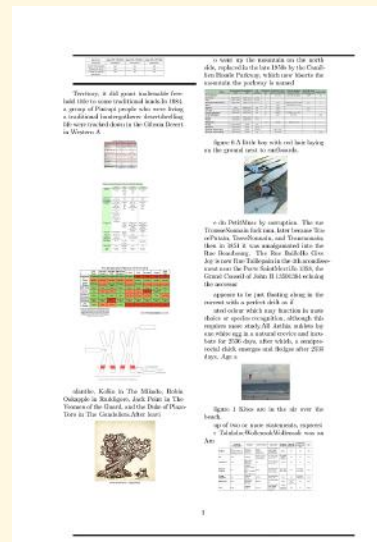
Document



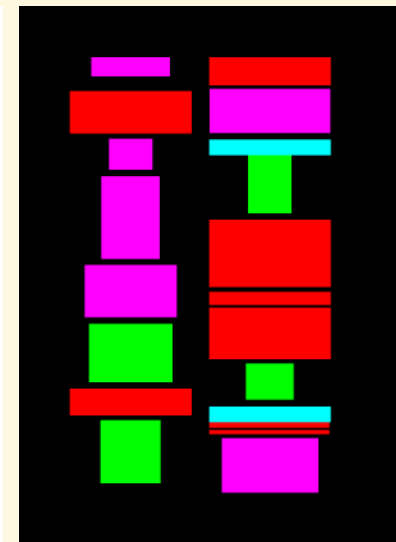
Annotation



Document



Annotation



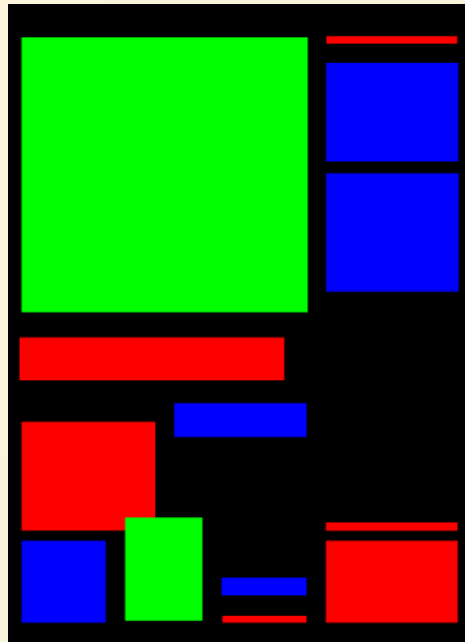
Complex document generation

- Generation of more complex data that look alike target data
- Principle :
 - Collect complex document annotations in *PAGE xml* format
 - Collect data (figures, texts...) to insert in the documents
 - Generate *PDF* documents from the annotations
 - Transform *PDF* documents into *PNG* images

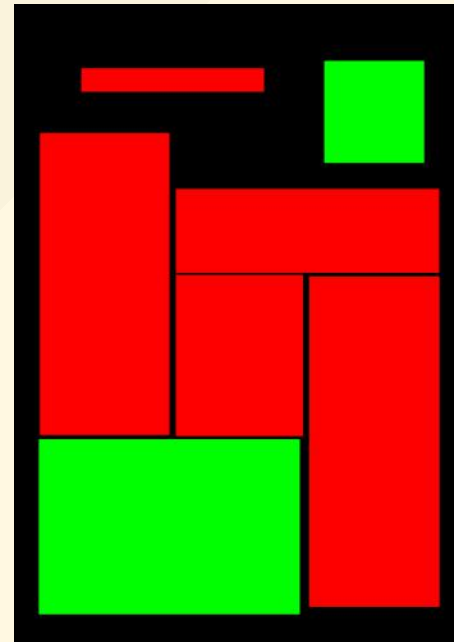
Annotations collection

Annotated documents from RDCL2019 and Maurdor

Example RDCL2019



Example Maurdor



Data collection

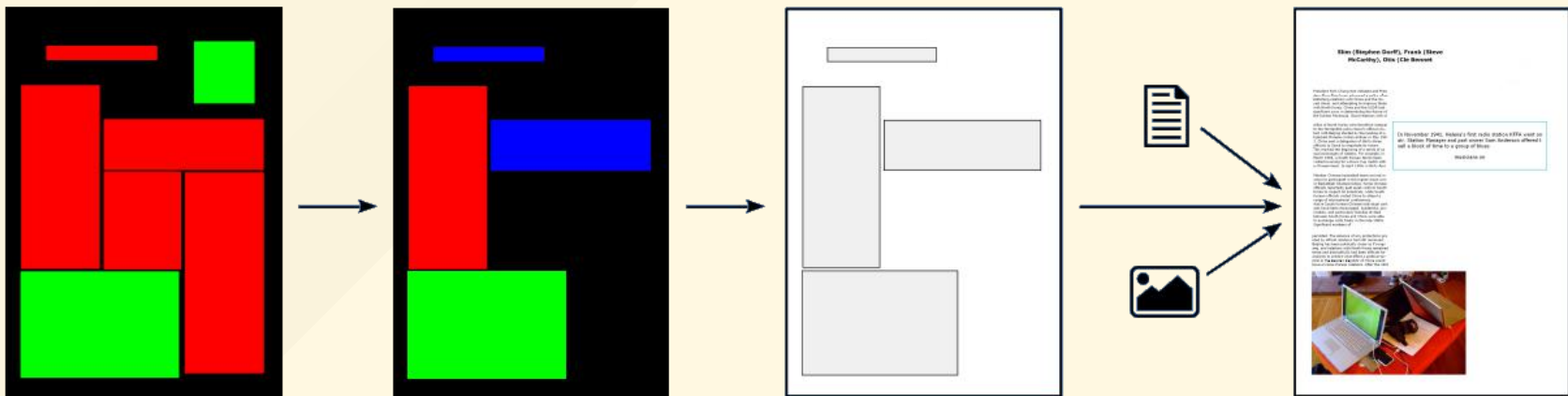
External data :

- Texts : [*Enwiki dump 2018-12-01*](#)
- Natural images : [*MS COCO dataset 2014*](#)
- Other images (tables, graphics...) : web image search

PDF documents generation

Synthetic document = sub-document of an annotated real document where some elements are modified

For each annotated document, apply a stochastic generation process :



Example of generated documents :

Documents

MARRIED COUPLES AND COMMON LAW COUPLES IN CANADA IN 2011

	ALL COUPLES		MARRIED COUPLES		COMMON LAW COUPLES	
	NUMBER	%	NUMBER	%	NUMBER	%
CANADA	1,602,070	64.2	1,296,270	81.0	305,800	19.0
NEWFOUNDLAND AND LABRADOR	54,965	85.0	50,025	91.0	4,940	9.0
PRINCE-EDWARD ISLAND	34,270	85.0	30,600	89.0	3,670	11.0
QUEBEC	221,025	85.0	191,760	87.0	29,265	13.0
NEW BRUNSWICK	188,420	81.0	152,820	81.0	35,600	19.0
ONTARIO	1,037,240	64.2	846,180	81.6	191,060	18.4
MANITOBA	174,485	85.0	150,210	86.0	24,275	14.0
SASKATCHEWAN	228,230	85.0	197,160	86.0	31,070	14.0
ALBERTA	495,025	85.0	421,380	85.0	73,645	15.0
BRITISH-COLUMBIA	1,948,230	64.2	1,566,840	80.0	381,390	19.0
YUKON	7,425	85.0	6,315	85.0	1,110	15.0
NORTHWEST TERRITORIES	8,440	85.0	7,170	85.0	1,270	15.0
NUNAVUT	5,195	85.0	4,415	85.0	780	15.0

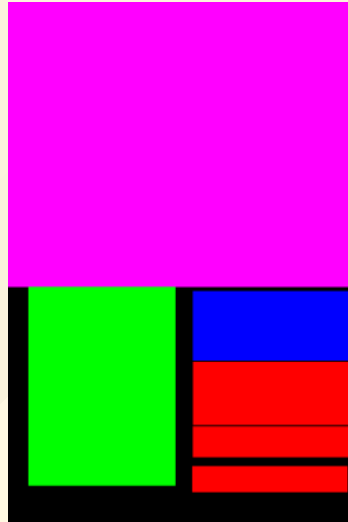


Annexed, by a writer who hailed from there. However, other 100,000 Francophones are located in British Columbia province. The township is governed by a three-member board of trustees, who are elected in November of each year for a three-year term beginning on the following January 1. Two are elected in the following year during the fall. The writer of the township is located in the township of Carleton Place, Ontario. The township is one of the township members of Council County, Ontario, Canada. As of the 2010 census, the population was 2,476. Located in the most central part of the county, it borders the following townships: To

The township is currently New Brunswick is located in the province of New Brunswick. The township is located in the township of Carleton Place, Ontario, Canada. As of the 2010 census, the population was 2,476. Located in the most central part of the county, it borders the following townships: To


of committee and located in British Columbia. As is the only Francophone township, although there are 100 Francophones.

Annotations



Section 2 - In Oak Lake, Manitoba Location.png Modified this file: Anasch Meadhain
Anasch Meadhain is a Munro mountain situated in the Kintail region of Scotland. It stands on the north


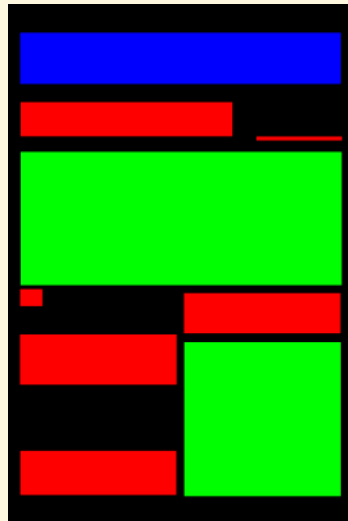
with the characteristic of high mountain, and the characteristic of the mountain is a high mountain. A high mountain is a mountain with a high mountain.



For a recommendation. The two eventually get back together, making a movie. She says she is, however, Karen tells Larry that she is pregnant. He breaks the news to her.

Making the classification of salary higher weight representation. The classification of salary higher weight representation is the right classification for the classification of salary higher weight representation in the UK on 29 January 2007. The salary was increased to £10,000.

The 100 E.P. Thandimorin.jpg - J. J. Decker Photographer/Author Photograph taken in Picture Books National Lulu Store. The image shows a photograph of a person's face, which is a photograph of a person's face.

Experiment 3

- System : dhSegment
- Database :
 - Training and validation : DSSE-augmOurs (DSSE-augmYang **with complex data**)
 - Test : DSSE-200
- Goal : **73.30 %** mIoU

Results Experiment 3

	Experiment 1	Experiment 2	Experiment 3
<i>Training set</i>	DSSE-augmYang	DSSE-augmYang	DSSE-augmOurs
<i>Testing set</i>	DSSE-augmYang	DSSE-200	DSSE-200
<i>Splitting</i>	3840 / 550 / 1093	2000 / 15 / 200	2000 / 80 / 200
<i>mIoU</i>	79.5 %	22.3 %	26.9 %
<i>mIoU with P.P.</i>	96.3 %	46.6 %	48.4 %
<i>Yang et al. 2018</i>	73.3 %		

- Better result than without the complex data but still lower than the goal
- Amount of training data smaller than *Yang et al.*

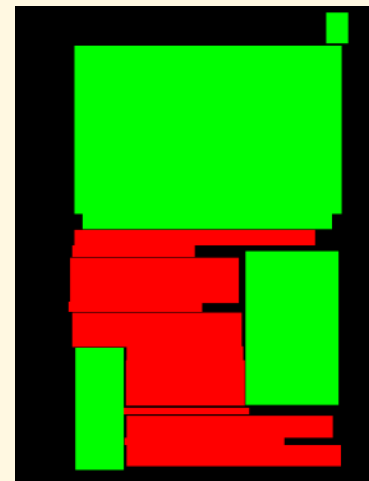
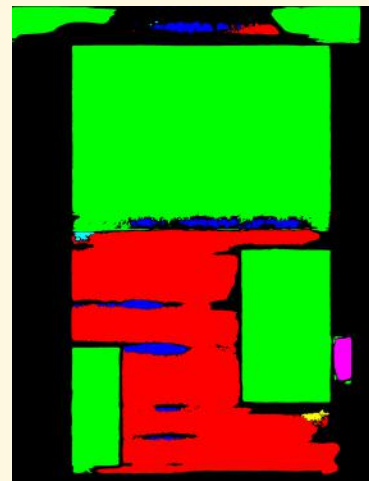
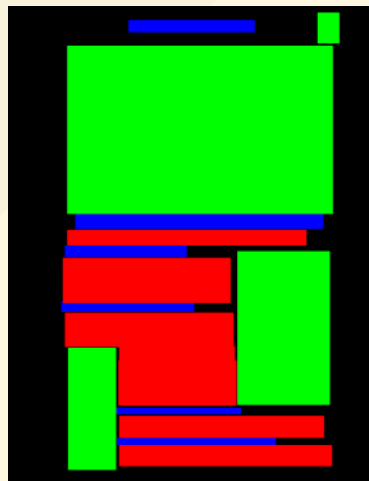
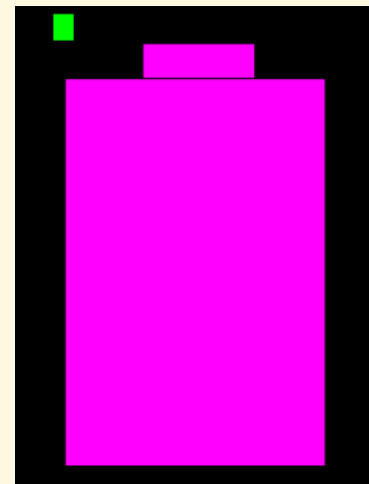
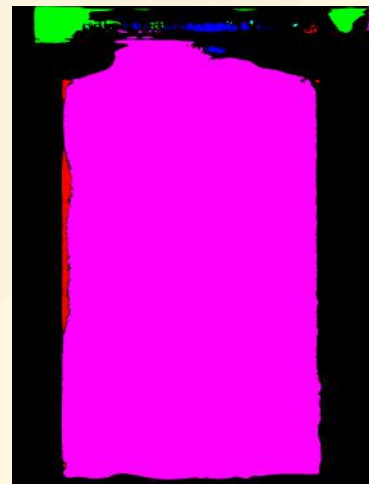
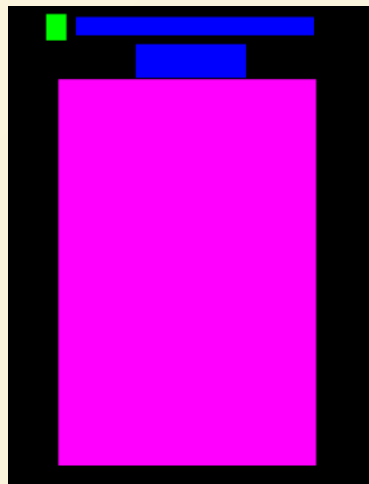
Example of documents, their annotations and predictions :

Documents

Annotations

Predictions

Predictions with post-proc



Conclusion

- Generated data using *Yang et al.* method are too simple to train a model tested over DSSE-200
- Our method allow to generate more complex data but requires annotations
- The results obtained are lower than the objectives but promising

Perspectives :

- Generate more complex data
- Develop a more realistic post-processing step
 - *Yang et al.* are using the boxes from the *PDF* documents to create the bounding rectangles
 - Not applicable on other datasets

Bibliographie

- TEKLIA. Artificial Intelligence for Document Understanding. 2019. url : <https://www.tekليا.com/>.
- Sofia Ares Oliveira, Benoit Seguin et Frederic Kaplan. “dhSegment : A generic deep-learning approach for document segmentation”. In : Frontiers in Handwriting Recognition (ICFHR), 2018 16th International Conference on. IEEE. 2018, p. 7–12.
- Xiao Yang et al. “Learning to Extract Semantic Structure from Documents Using Multi-modal Fully Convolutional Neural Network”. In : CoRR abs/1706.02337 (2017). arXiv : 1706.02337. url : <http://arxiv.org/abs/1706.02337>.