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.

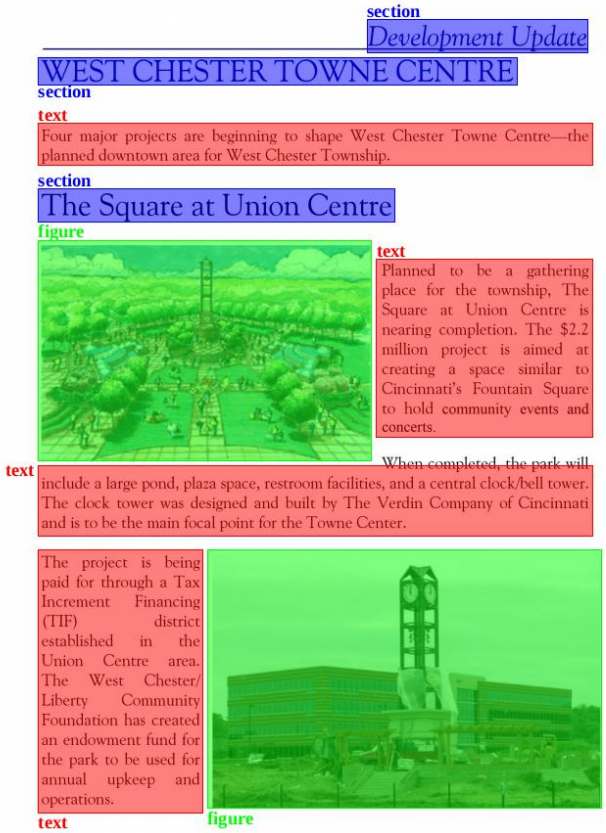
The Square at Union Centre

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

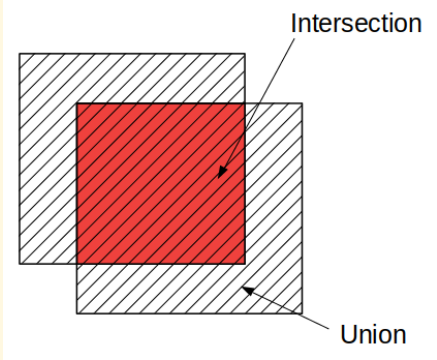


How to generate synthetic data for training deep learning DLA models ?

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

Test data and references

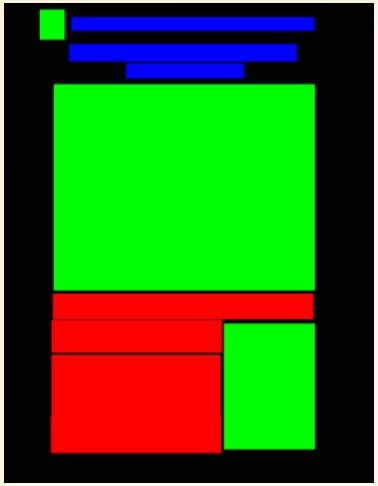
- DSSE-200 : database of 200 annotated real documents from magazines and academic papers created by Xiao Yang et al. CVPR2017 :
  "Learning to Extract Semantic Structure from Documents Using Multi-modal Fully Convolutional Neural Network"
- Reference result (deep network using image and text) : **73.30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

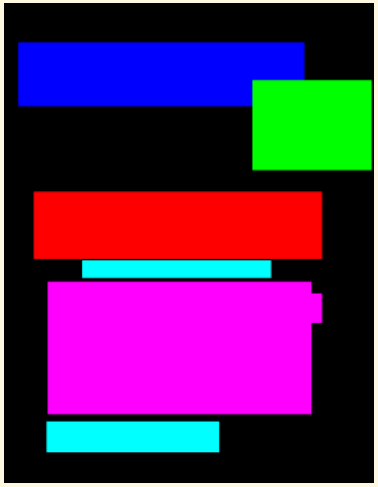


- Open-source document generation algorithm

Example of documents from DSSE-200 :

Documents Annotations



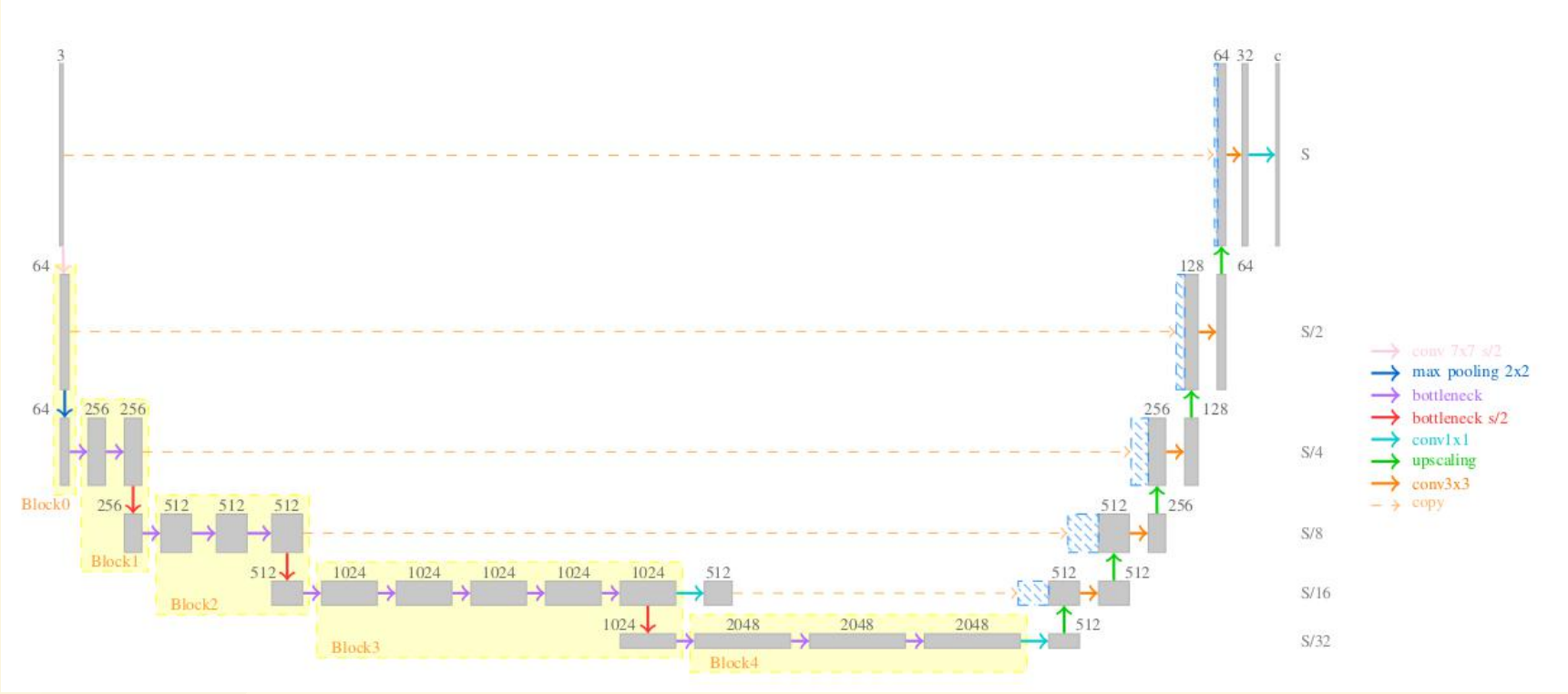


Reference system: dhSegment

- System developped 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

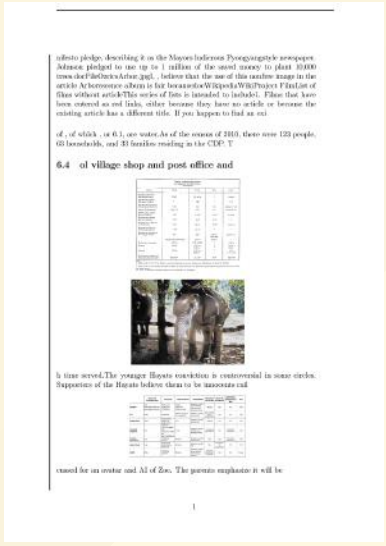
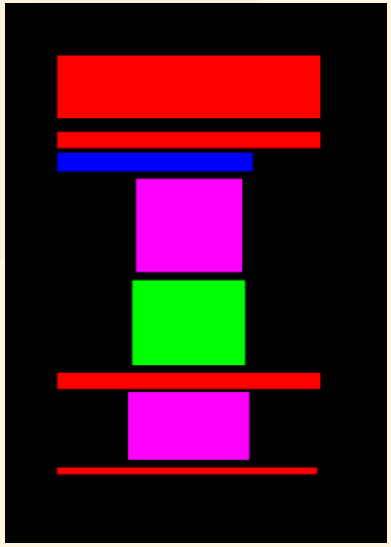
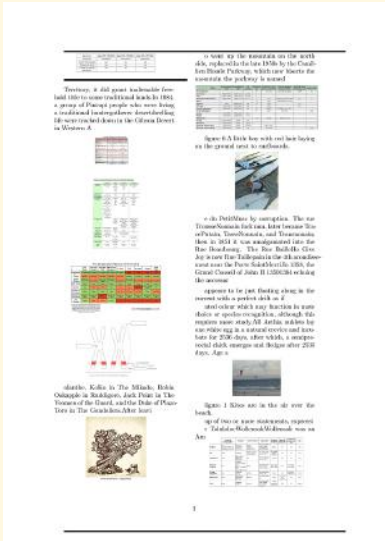
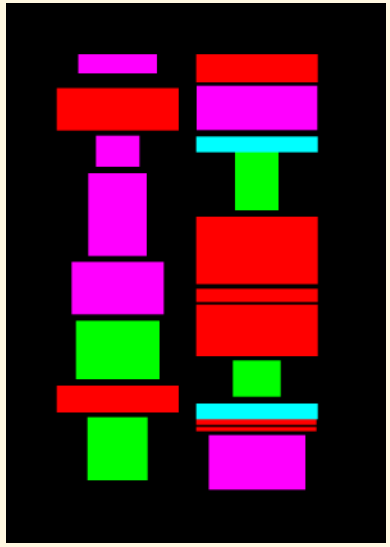
- System : dhSegment
- Database :
  - Training and validation : synthetic data generated with Yang algorithm (*DSSE-augmYang*)
  - Test : idem

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

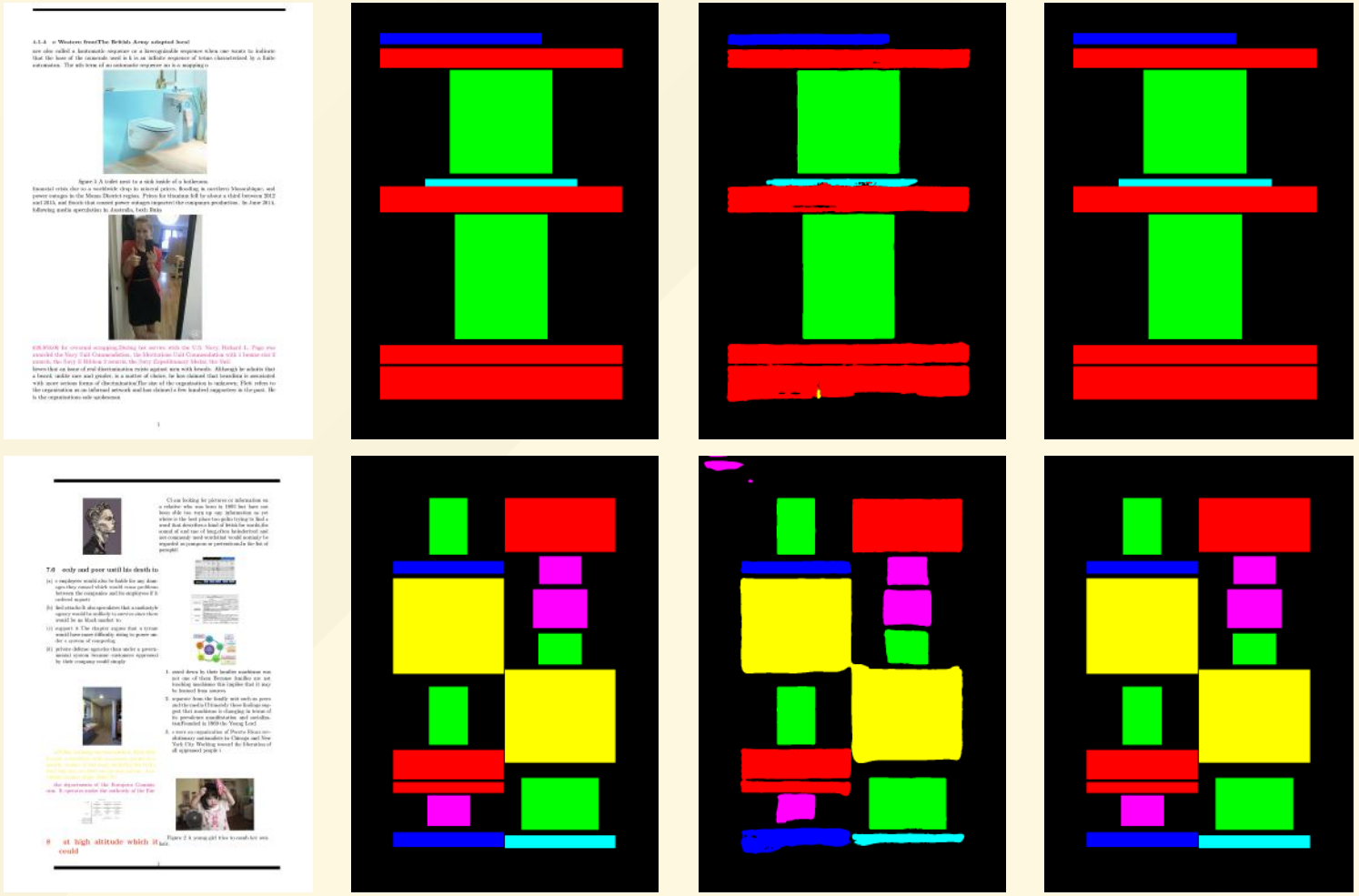
Results Experiment 1

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

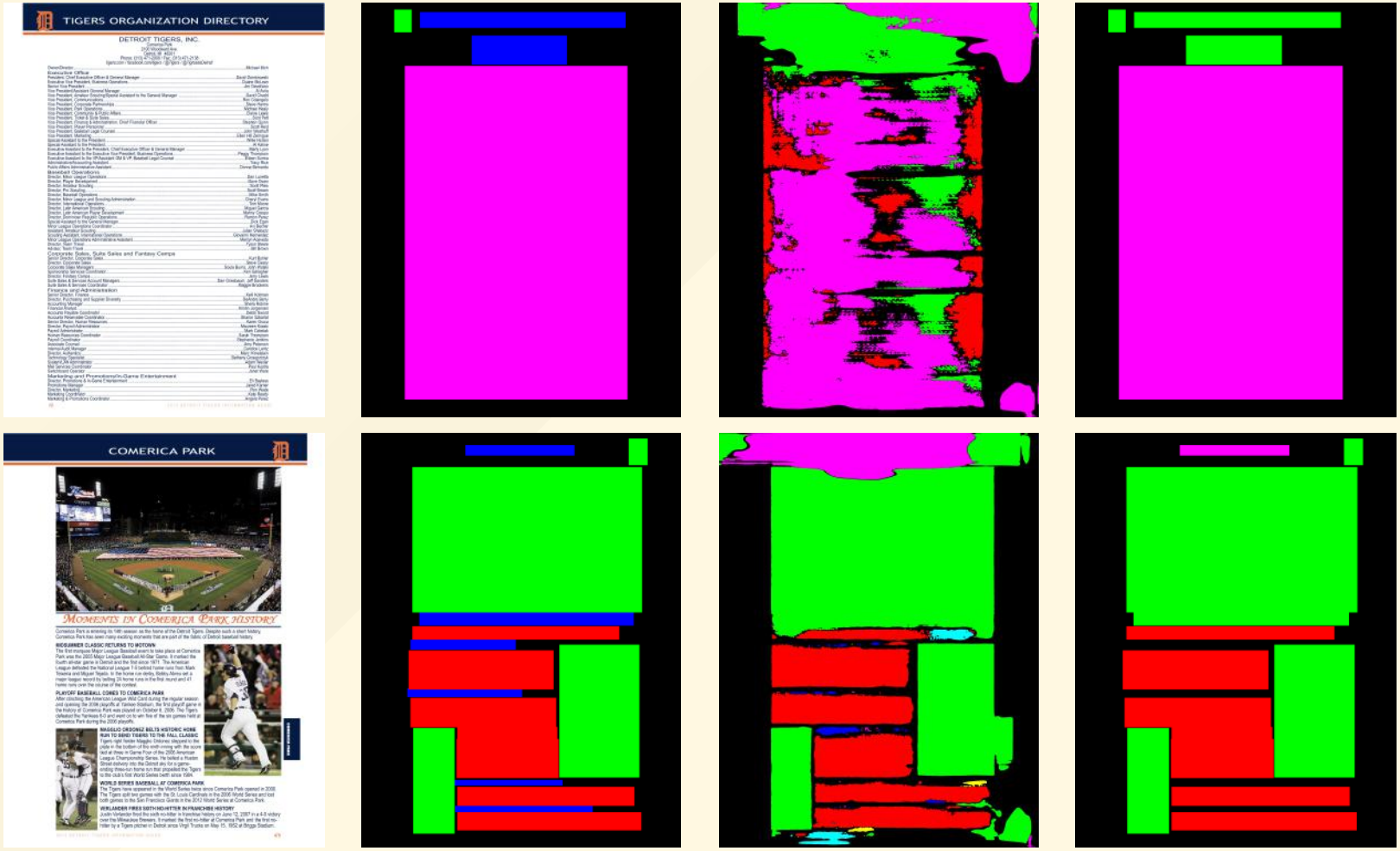
Results Experiment 2

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

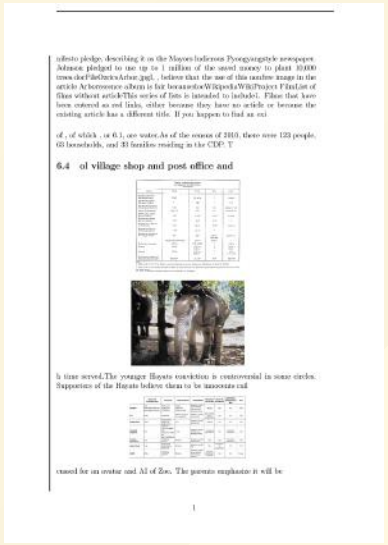
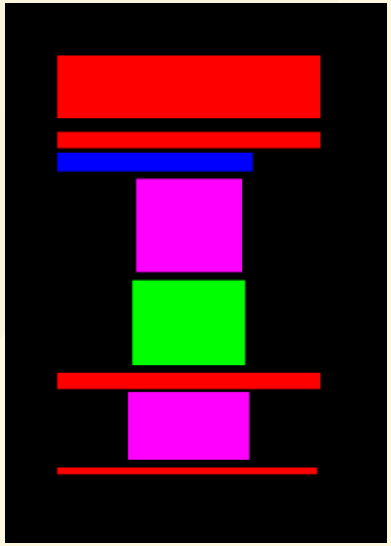
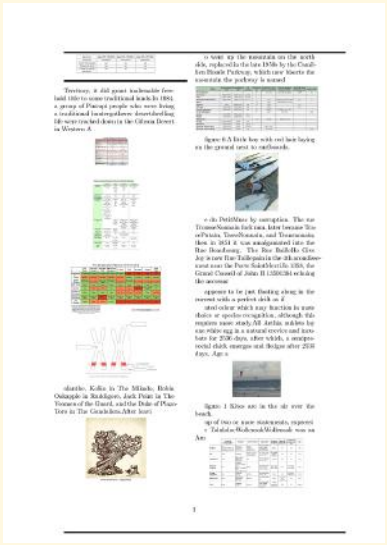
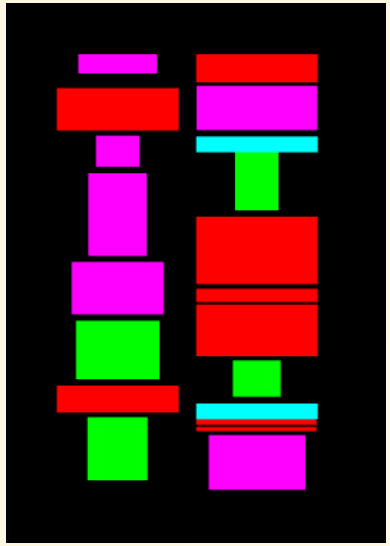


Developpement of a complex document generation method

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

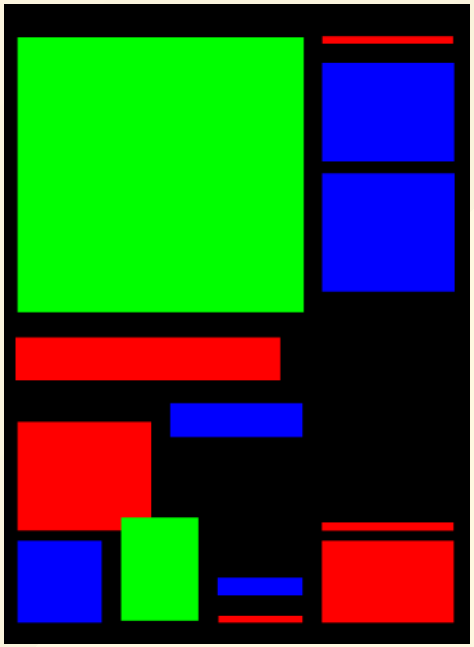
Complex document generation

- 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

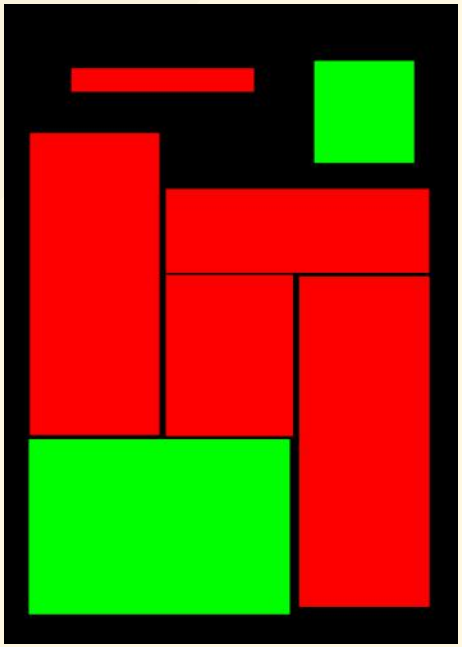
Annotations collection

Annotated documents from RDCL2019 and Maurdor

Example RDCL2019



Example Maurdor



Data collection

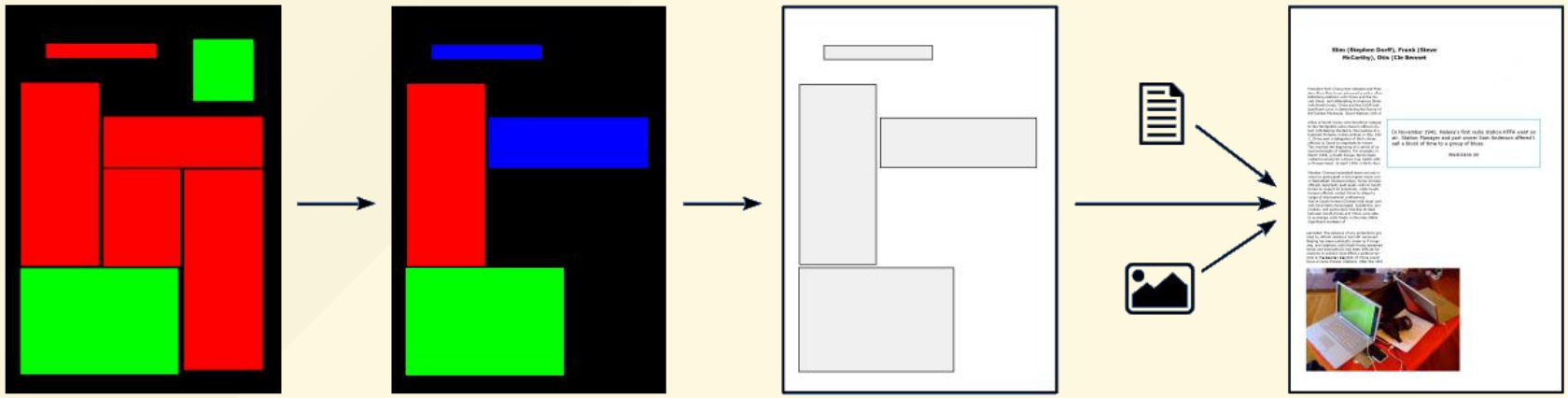
External data :

- 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 Annotations

Experiment 3

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

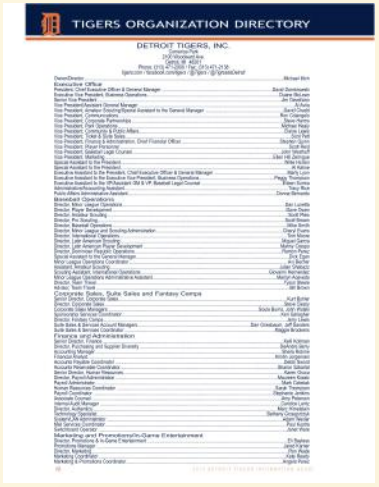
Results Experiment 3

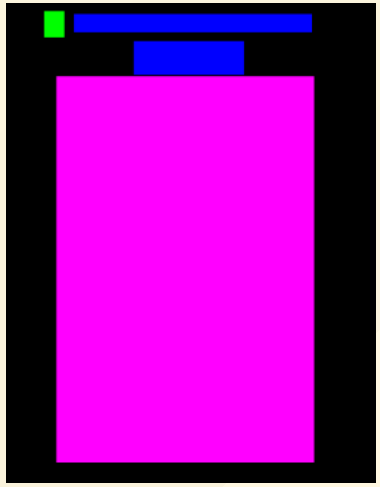
	Experiment 1	Experiment 2	Experiment 3
Training set	DSSE-augmYang	DSSE-augmYang	DSSE-augmOurs
Testing set	DSSE-augmYang	DSSE-200	DSSE-200
Splitting	3840 / 550 / 1093	2000 / 15 / 200	2000 / 80 / 200
mIoU	79.5 %	**22.3 %**	**26.9 %**
mIoU with P.P.	96.3 %	**46.6 %**	**48.4 %**
Yang et al. 2018	**73.3 %**		

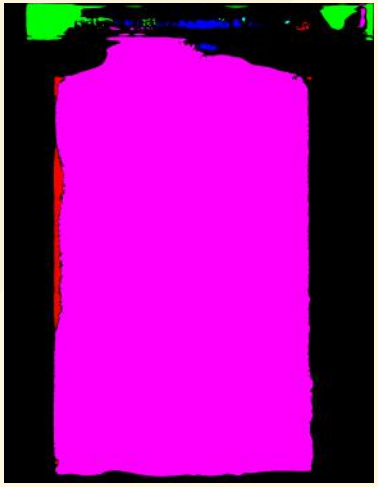
- Better result than without the complex data but still lower than the goal
- Amount of training data smaller than *Yang et al.*

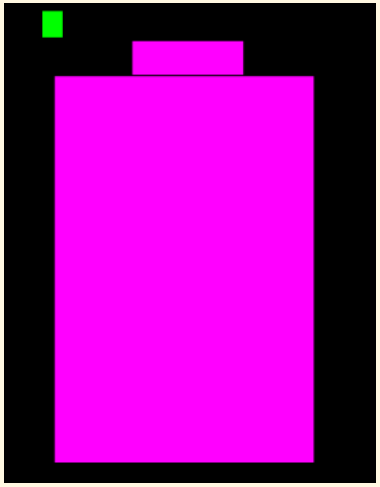
Example of documents, their annotations and predictions :

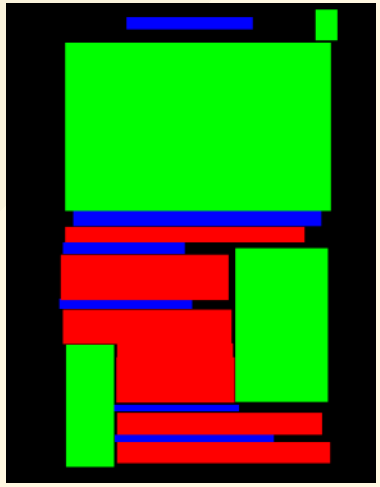
Documents | Annotations | Predictions | Predictions with post-proc

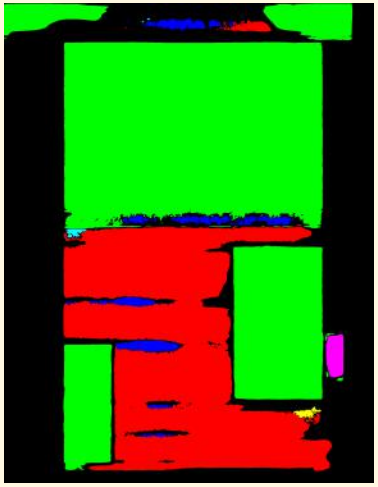


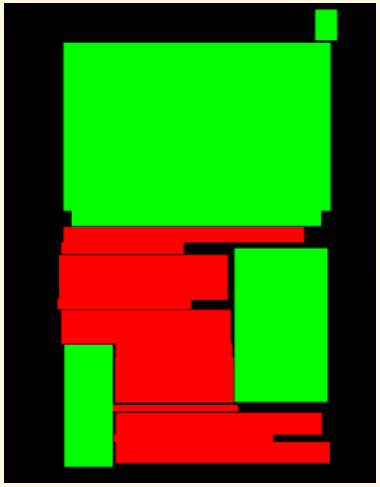












Conclusion

- 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.teklia.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.