Research Internship for Master 2 Students

Layout segmentation and classification of visual style elements in newspapers

- **Topic**: data science, data mining, machine learning
- **City and country**: Sophia-Antipolis, France
- **Team or project in the lab**: Centre de Recherche Inria Sophia Antipolis – Méditerranée, Biovision Lab
- **Name and mail of the advisors**: Hui-Yin Wu (hui-yin.wu@inria.fr), Pierre Kornprobst (pierre.kornprobst@inria.fr)
- Name and mail of the head of the department: Bruno Cessac (bruno.cessac@inria.fr)

General presentation of the topic

For the adaptation of print to digital journalism, a necessary step is the identification of the various visual elements in a newspaper page, with varied applications such as curation and accessibility. This is an extremely complex challenge due mainly to (1) the the large number visual (e.g., images, logos, ads) and textual (e.g., titles, headings, captions, cross-references), (2) the multi-heirarchical organization of elements (e.g., an article contains headings, images, columns, which contain paragraphs etc.), and (3) the different style constraints from one newspaper to another. To date, existing approaches are only able to identify a small subset of these elements, such as differentiating text from images, and do not address at all the question of hierarchy. A robust approach to segmenting and classifying newspaper elements thus requires incorporating knowledge on visual design and style into the algorithms for this task.

Objective of the internship

The objectives of this internship are thus to establish a workflow for segmenting and classifying text and visual elements of newspapers using the understanding of visual styles and layout. To achieve this goal, this project is composed of four main steps:

1. harnessing various tools including optical character recognition (OCR) and computer vision libraries like OpenCV to extract high-level visual features on the page, such as text, lines, and borders,
2. investigate machine learning approaches to instance segmentation and classification, notably Mask-RCNN, for complex document segmentation, and train and iteratively improve the model,
3. explore the question of multi-hierarchical layouts by investigating methods to recursively segment gradually more granular elements, and
4. as a final step, establish a baseline on which we can compare our approach to state-of-the-art results.

The work will be carried out on the basis of a pre-existing framework that includes annotation, segmentation, and classification. Training and testing will be conducted on a set of segemented and annotated newspapers.

Bibliographic references:

Almutairi, A., & Almashan, M. (2019, December). Instance segmentation of newspaper elements using mask R-CNN. In 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA) (pp. 1371-1375). IEEE.

Gautier C. and Gautier D. Design, typography etc.: A Handbook, Niggli, 2017.

Meier, B., Stadelmann, T., Stampfli, J., Arnold, M., & Cieliebak, M. (2017, November). Fully convolutional neural networks for newspaper article segmentation. In 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR) (Vol. 1, pp. 414-419). IEEE.

Prasad, A., Déjean, H., & Meunier, J. L. (2019, September). Versatile layout understanding via conjugate graph. In 2019 International Conference on Document Analysis and Recognition (ICDAR) (pp. 287-294). IEEE.

Wu, H. Y., & Kornprobst, P. (2019). Multilayered Analysis of Newspaper Structure and Design. [Research Report] RR-9281, UCA, Inria. 2019

Expected ability of the student

- familiarity with image processing and computer vision libraries,
- experience with Artificial Neural Networks (e.g., CNN, Mask-RCNN),
- formation in machine learning,
- good in Python programming
- some experience in C++ is a plus

Illustration

Example of result from our previous work (RR-9281).


Nameplate/Masthead
Image/Photo/Caption
Paragraph/
Column
Headline
Crossref
Image/
Photo


Masthead/
Nameplate
Crossref/
Column
Image / Article
L'exécutif
au rapport
AD
UNIQUE


Nameplate/Masthead/Skybox
Crossref
Article/Image/Photo
Caption/Image/Photo
Heading
Headline
Kinderastma
Headline
piekt in België
door verkeer
Crossref
Image/Photo
Lede
Paragraph/
Column
Image/
Infograph
Ad