ARGUS
MEDIA BASED INTELLIGENCE

PANOPTES – Automated Article Segmentation of Newspaper Pages for "Real Time Print Media Monitoring“

M. Arnold, M. Cieliebak, T. Stadelmann, J. Stampfli, and F. Uzdilli

Zurich University of Applied Sciences
zh aw School of Engineering
datalab
www.zhaw.ch/datalab

Overview

Partners

Who are we

ARGUS der Presse AG

- Switzerland's leading media monitoring and information provider
- Experience of more than 100 years

ZHAW Datalab

- Interdisciplinary research group at Zurich University of Applied Sciences
- Combining the knowledge of different fields related to machine learning

The Project

What do we do

Goal

- Real Time Print Media Monitoring
    - Extraction of relevant articles from newspaper pages
    - Delivering articles to customers

Problem

- Fully automated article segmentation
- Identification of article elements (e.g. title, subtitle, etc.)

Approach

Rule based

Segmentation based on hardcoded rules

Rule examples

- Each article must contain a title
- Titles define article’s width
- Articles are graphically separated by e.g. lines
- etc.

Pros

- Performance increases the more time is spent for finding rules
- Adding new rules is simple

Cons

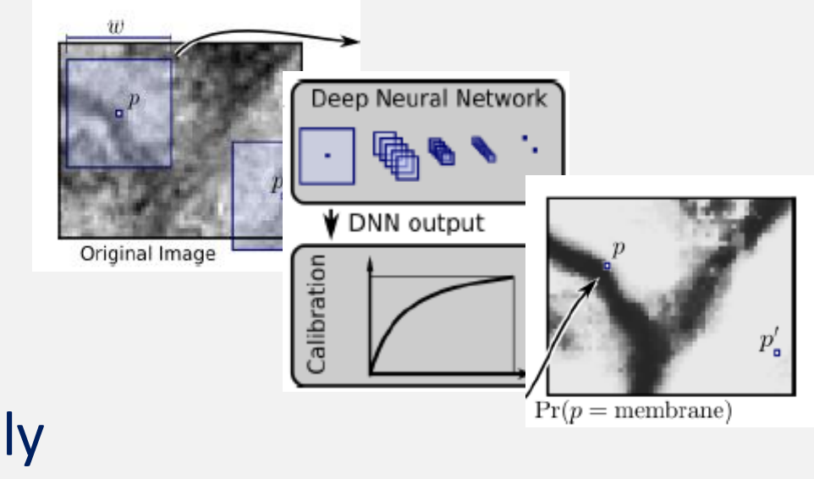
- Not every case can be covered
- Adaptation to new layouts is costly manual work

Image based

Segmentation based on visual features and deep learning

Approach

- Pixel classification (article/border) based on [1]

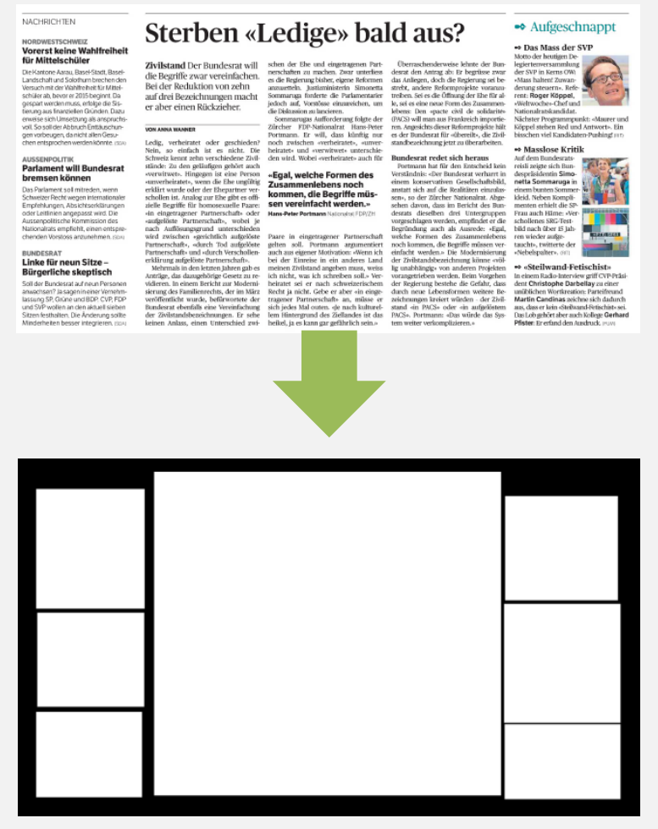


Pros

- Rules can be learned implicitly
- New layouts can be adapted automatically

Cons

- Success factors on new data and problems are unknown
- Training requires a huge amount of data

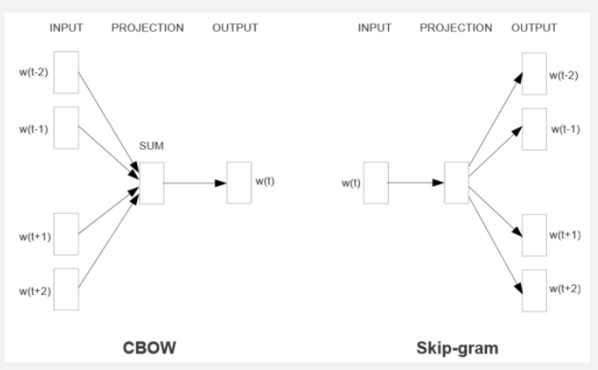


Text based

Segmentation based on textual features and neural nets

Approach

- Text block clustering (semantic distance) based on [2]



Pros

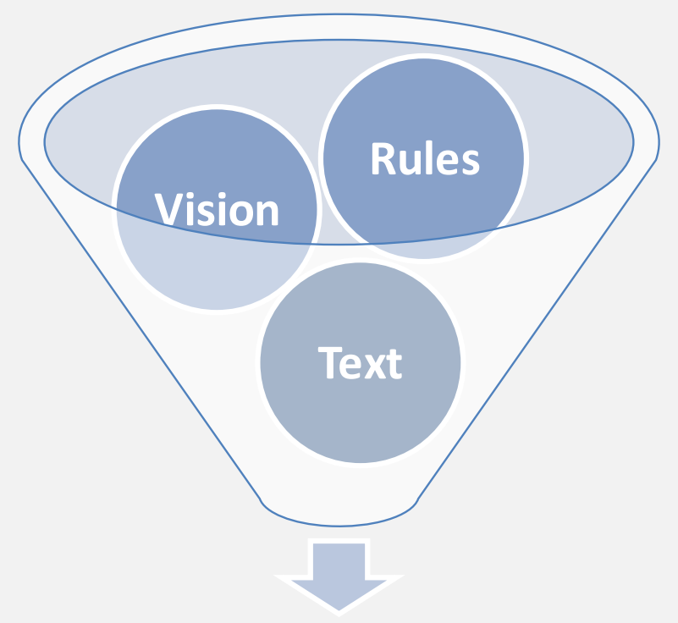
- Rules can be learned implicitly
- Not layout dependent

Cons

- Only text can be processed

Combination

Combination of rules, visual and textual features



Final segmentation

Result

References

[1] D. C. Ciresan, A. Giusti, L. M. Gambardella, and J. Schmidhuber. *Deep neural networks segment neuronal membranes in electron microscopy images.* In *NIPS*, pages 2852–2860, 2012.
[2] T. Mikolov, K. Chen, G. Corrado, and J. Dean. *Efficient Estimation of Word Representations in Vector Space.* In Proceedings of Workshop at *ICLR*, 2013.

This project was funded by CTI under project number 17719.1 PFES-ES

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Federal Department of Economic Affairs,
Education and Research EAER

Commission for Technology and Innovation CTI
Innovation Promotion Agency

swiss group for artificial intelligence
and cognitive science