Vision-based Web Page Segmentation

Bc. František Maštera, xmaste02@stud.fit.vutbr.cz

Supervisor: doc. Ing. Radek Burget, Ph.D.

Introduction

WPS (Web Page Segmentation) usage
- Data mining, document indexing
- Assistive technologies

Many WPS algorithms
- With **different approaches**, strengths & weaknesses

FitLayout framework
- Tools for **WPS evaluation** and further research
- Suite of already implemented WPS algorithms
- Goal: **Extend FitLayout** with another one

Integration & Pipeline

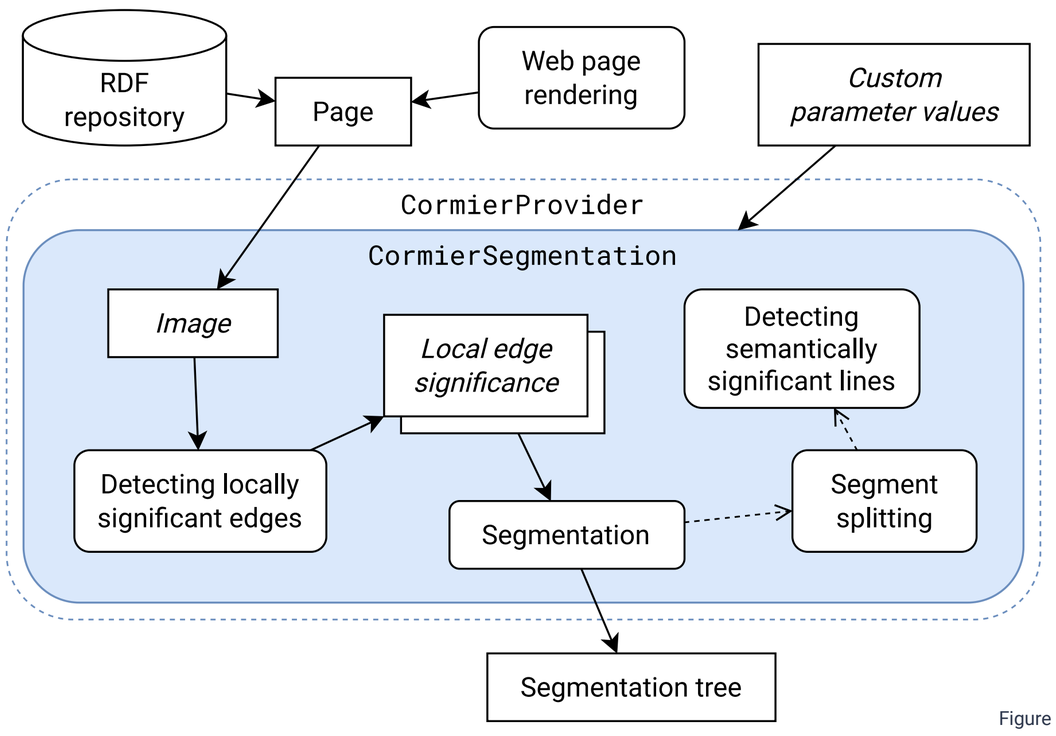


Figure 1

Locally Significant Edges

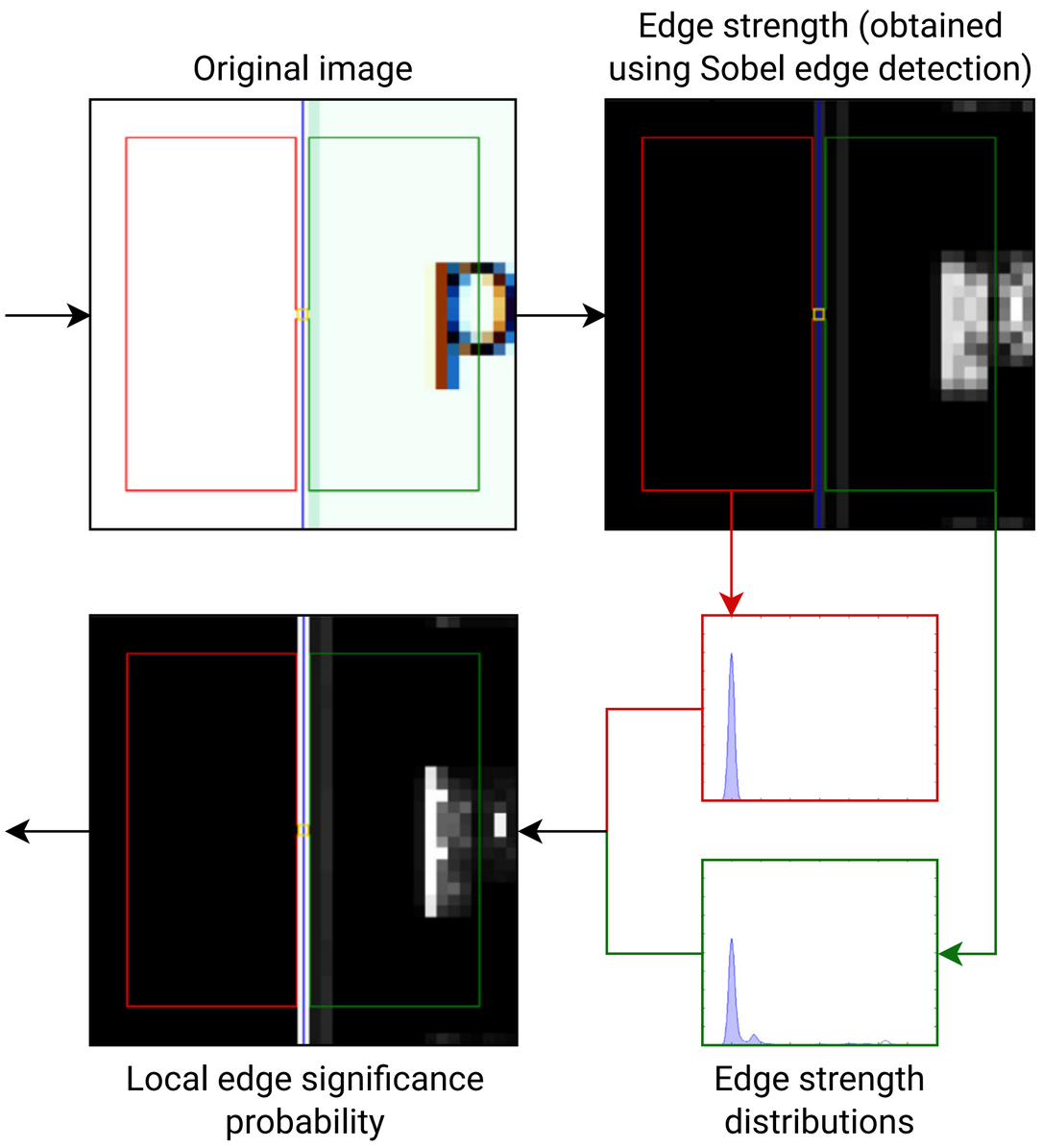


Figure 2[1]

Final Segmentation

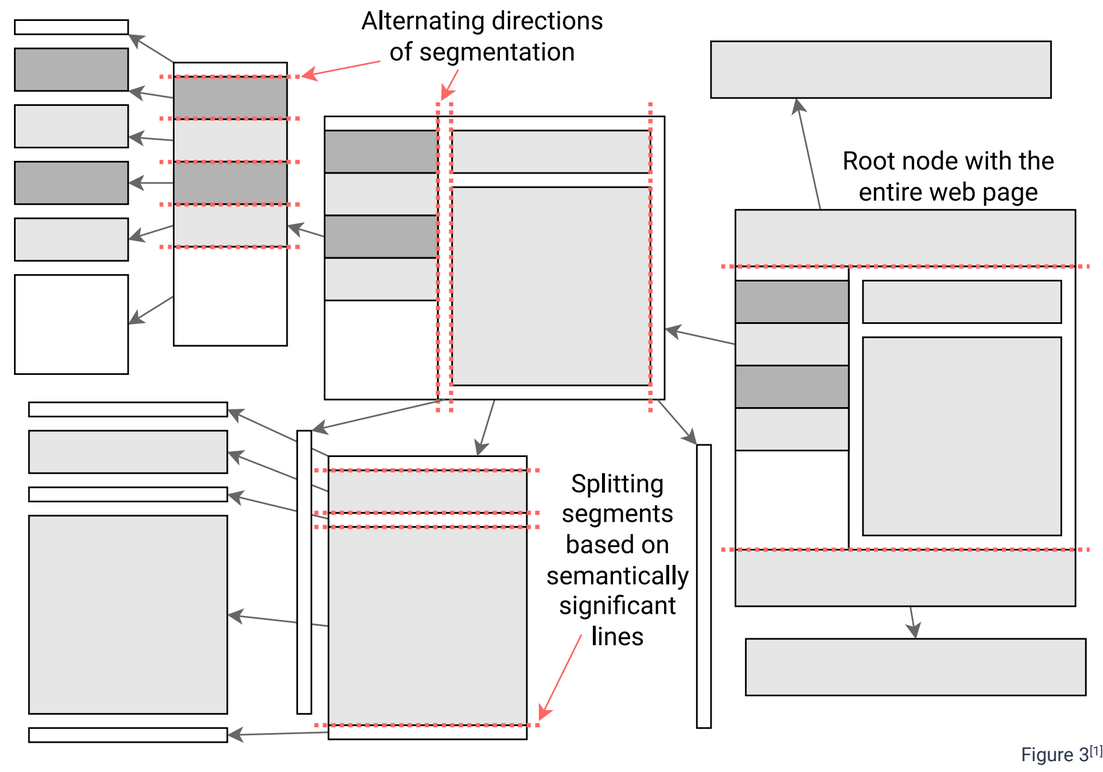
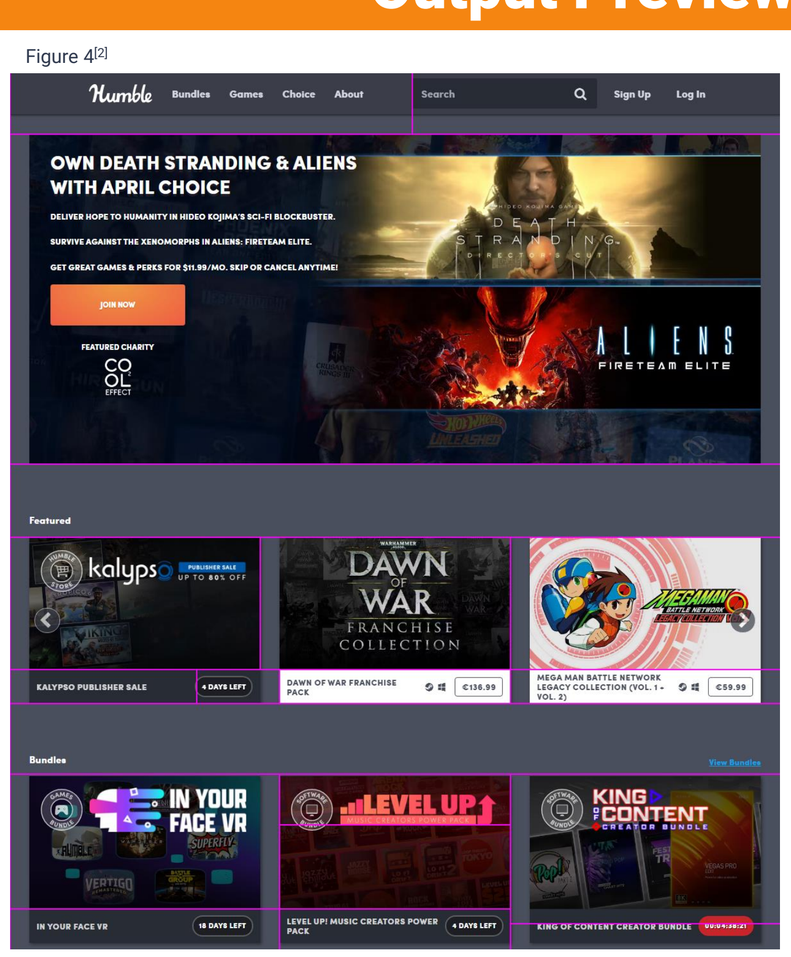


Figure 3[1]

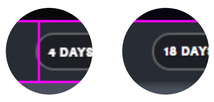
Output Preview

Figure 4[2]

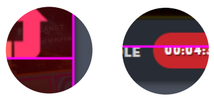


Hierarchical structure
- Main sections:
  - Navigation bar, main section, featured items, ...
- Separate items
- Item sections:
  - Banner, details

Imperfections
- Inconsistencies

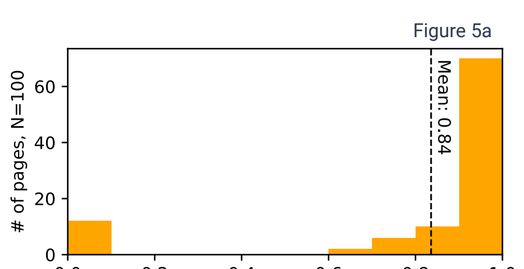


- Poor segmentations



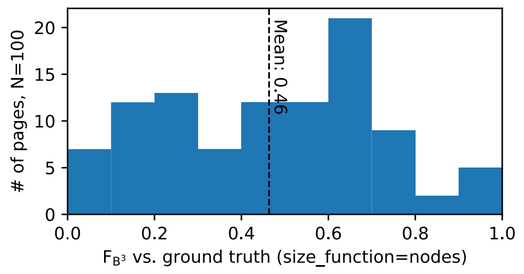
Testing & Evaluation

Figure 5a



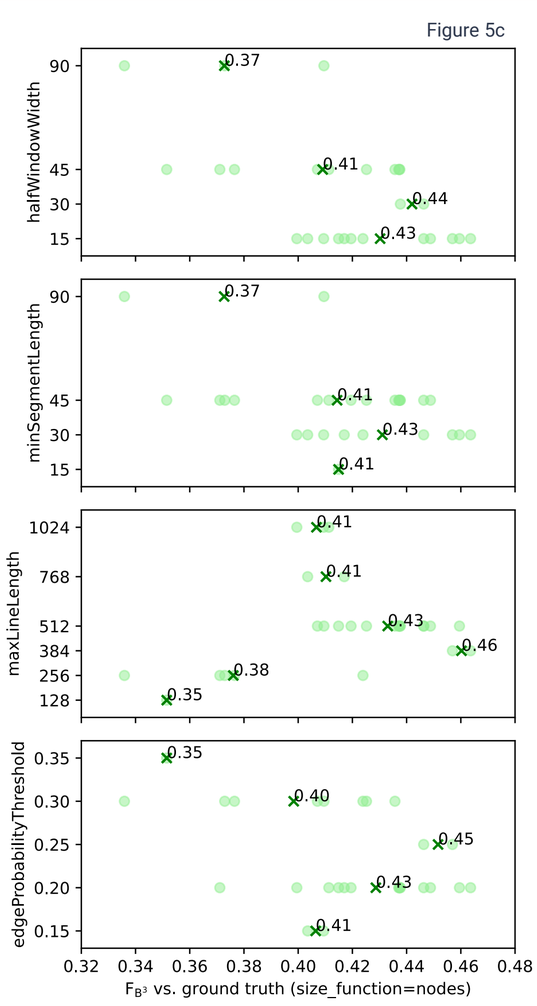
Testing against public impl.
- Most results meet expectations
- Exceptions with differences

Figure 5b



Parameter evaluation
- Affects granularity, speed, ...
- Optimal values for most pages
- Improve the results by **23.13%**

Figure 5c



BRNO UNIVERSITY OF TECHNOLOGY — FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2023

[1] Michael Cormier (2018). Computer Vision on Web Pages: A Study of Man-Made Images. UWSpace. http://hdl.handle.net/10012/13523
[2] https://www.humblebundle.com/