DISSERTATION THESIS REVIEW

Applicant: Ing. Jan Zelený
Thesis title: Web page segmentation utilizing clustering techniques
Reviewer: Ing. Tomáš Kliegr, Ph.D.,
Faculty of informatics and statistics, University of Economics, Prague

Thesis summary

The thesis describes a new algorithm for segmenting web pages. Unlike existing approaches listed in the thesis, the proposed algorithm is based purely on visual cues. Evaluation included in the thesis shows that the proposed algorithm surpasses the VIPS algorithm, which the thesis designates as a state-of-the-art approach in this area, in a number of metrics.

Review structure

The Faculty of information technology asked me to reflect in particular the following criteria in the review:

1) Is the topic of the thesis relevant to the topical area of the dissertation and is it up-to-date with respect to current research.
2) Does the thesis contain original contribution in sufficient extent? What is the main contribution?
3) Was the core of the thesis published in sufficient extent?
4) Does it follow from the publication record of the applicant that he has scientific merit?
5) Other facts that would help to characterize the applicant.

These criteria are addressed in Part I of the review. In addition, I list several observations on the quality of the thesis in Part II. Part III contains questions I suggest the applicant answers during the defense as well as the overall recommendation.

Part I: Official Review Criteria

Ad 1) Is the topic of the thesis relevant to the topical area of the dissertation and is it up-to-date with respect to current research.

The applicant applies for PhD in the area of "Výpočetní technika a informatika" (Information technology and computer science"). The thesis topic clearly falls in this area.

Regarding the state-of-the-art character of the research. Development of vision-based segmentation algorithm is a current area of research (cf. e.g. Cormier et al, 2016). The review section of the thesis briefly covers some of the approaches proposed up to 2014. Nevertheless, it should be noted that the review of these approaches is very brief and is mostly limited to recounting related papers on page 18. Method [12], which is presented in greater detail, has been developed by the author's colleague and presented at a workshop in 2010. Other approaches published in leading journals or conferences are not presented in sufficient detail. The evaluation of the approach proposed by the author is made only with respect to the VIPS algorithm from 2003.

Overall, I have doubts that the presented thesis sufficiently considers current research in the target area. However, given that the core of the dissertation has been accepted for publication to a quality journal and the author demonstrated that he is aware of the need to delimit the proposed approach with respect to related research, I suggest that the thesis is evaluated positively, yet borderline, on this criterion.

Ad 2) Does the thesis contain original contribution in sufficient extent? What is the main contribution?

It follows from the statement of the author that the approach presented in the thesis was developed almost solely by the applicant. The only work that is stated to be done by someone else is a reference implementation of the VIPS algorithm used for evaluation. The main contribution is clearly the BCS + template clustering algorithmic framework presented on pages 33-61. While I am not an expert on page segmentation, I judge that this framework is sufficiently original. The framework composes of multiple algorithms, which seem to have been developed after several iterations and experimental evaluations, therefore I would judge that the extent of contribution in terms of proposed and implemented algorithms could be sufficient.

What I cannot answer is whether the proposed framework constitutes sufficient advancement over the state-of-the-art approaches, because of substantial deficiencies in the related work review as well as evaluations performed. Also, I cannot answer to what extent do the proposed algorithms differ from other state-of-the-art vision segmentation approaches, because a detailed comparison is not present in the thesis. Specifically, Cormier et al, 2016, a reference omitted from the thesis, presents a purely visual segmentation algorithm, while the applicant claims on page 73 that his method is the only purely visual segmentation algorithm.

Overall, considering that the core of the proposed approach has been published in a quality peer-reviewed publication [65] (cf. point 3 below), it can be assumed that the thesis constitutes sufficient advancement of the state-of-the-art.

Ad 3) Was the core of the thesis published in sufficient extent?

The most recent publication is from 2017, covers the BCS algorithm, which is the core of the dissertation. This appears in a quality journal published with Elsevier with 2015 impact factor of 1.397. I conclude that the core of the thesis was published in sufficient extent.

Ad 4) Does it follow from the publication record of the applicant that he has scientific merit?
The applicant's publication record is appropriate.

Applicant is the main author of five papers closely related to the topic of the thesis. Three of the submissions are indexed by DBLP and Scopus. These include two journals, one of them with impact factor, and WIMS 2013 conference.

It should be noted that according to the Scopus citation index, there are no citations of the applicant's work. Google scholar search also did not reveal any noteworthy citations. However, citation count is not one of the assessment criteria, and most applicant's papers are recent.

Ad 5) Other facts that would help to characterize the applicant.

From the provided brief professional CV of the applicant it follows that the applicant pursued the PhD in parallel to regular employment in the software industry, working in area not related to the thesis.

Part II: Additional Criteria

Quality of evaluation

Overall, I find the quality of the performed experimental evaluation sub-standard.

My main objections are:

- Dated VIPS algorithm is used as the only baseline without proper justification. On page 18 it is stated that there is number of approaches that improve VIPS accuracy. Why then compare only with VIPS?
- The description of how the gold standard was designed (3 volunteers using semi-automatic segmentation tool) is insufficient.
  - There are no instructions presented that would allow for replication of the gold standard generation process.
  - The raw data (webpages) from which the segmentation was performed are not available.
  - Details of the semi-automatic segmentation tool are not given, it is not discussed if its design could not benefit the proposed BCS approach at the expense of VIPS.
- The VIPS reference implementation was implemented by the applicant's student. It is unclear if runtime comparison is fair, as the student did not likely optimize the VIPS algorithm data structures and code to that extent that the applicant did for BCS. No information on whether both implementations were parallel/single core are given. No details of the hardware used to carry out the experiments are given.
- The evaluation methodology should be discussed in greater detail, including more references to other papers evaluating segmentation algorithms.

Given the objections listed above, I personally do not consider the results of the evaluation as sufficiently grounded. As stated elsewhere, this may not be an obstacle in accepting the thesis, since the core results have passed peer review in a quality journal. Also, considering the professional background of the applicant, I came to the conclusion that the applicant has not elaborated the Evaluation chapter sufficiently due to time constraints, rather than ignorance of scientific standards and research methodology.

Reproducibility of results

The thesis contains a DVD disc with source code and gold standard data for evaluation. This in principle allows for reproducibility of the results.

I would strongly encourage the author to make the implementation as well as the experimental data public if he has not yet done so. The quality of the documentation of the code as well as of the data should be, however, significantly improved in this case.

Quality of the text - content

- The introduction makes too abrupt transition from the general appeal of WWW to the topic of the thesis. It lacks justification of the application areas of the methods developed in the thesis.
- The explanation of the NCE method in sec 3.1.1 is not particularly clear. What is the relation between equation 3.1 and 3.2/3.3?
- Similarly, the VIPS algorithm is not sufficiently described in Section 3.1.2. While the author includes some figures to illustrate the algorithm, they are not commented enough. Another problem with the review is that it is only qualitative and lacks on any benchmarks. This

could, for example, be used to support the author's claim that VIPS has not been unanimously surpassed by any successor algorithms, which is made on pages 16 and 18.

- The reason why template detection methods are not relevant for the thesis given in section 3.2 is not sufficiently elaborated.
- In section 3.2.1 when explaining the SST method, the text does not explain how is p_i estimated, which is important for understanding the method.
- Section 3.2.4 seems as very important as it designates the algorithm to be further developed within the thesis. Unfortunately, this explanation is insufficiently short.
- The text misses concrete examples that would illustrate the differences in output between BCS and other approaches.

Quality of the text - language

- There is a typographic error in the definition of F on page 14 below 3.1.
- The author uses the term "tag" to mean "element" on several places in the document (p. 14).
- Small number of typos, such as "by by"
- Sparse use of abbreviations not suitable for formal text (it's instead of it is)
- Lack of capitalization (figure 3.4 instead of Figure 3.4)
- sometimes missing space between citation reference and the preceding text
- informational gain -> information gain (p. 26)

Part III: Questions and Conclusion

Questions for the applicant

Question 1. Compare the BCS algorithm to Cormier, 2016. Also compare the evaluation methodology that you used in the thesis to the evaluation methodology used in this paper.

Question 2. The thesis concludes that the proposed BCS approach is according to the benchmarks much faster than VIPS. However, the BCS approach is fully visual method, while VIPS does not fully render the page. Could you point at specific conceptual or algorithmic steps in VIPS that make it slower than BCS?

Conclusion

Despite the reservations expressed in the review, I consider the work done by the applicant of sufficient scope, quality and scientific merit for awarding a PhD degree in computer science.

I recommend the thesis to be approved.

June 5, 2017, Praha

Ing. Tomáš Kliegr, Ph.D.

Bibliography

Cormier, Michael, et al. "Purely vision-based segmentation of web pages for assistive technology." *Computer Vision and Image Understanding* 148 (2016): 46-66.