Evaluation report

Dr. Philipp Aichinger, Feb. 13 2023

Doctoral thesis “Analysing Videokymograms Using Classical and Deep Learning Methods” by RNDr. Ales Zita

Following the invitation of the Faculty of Mathematics and Physics to evaluate the submitted Doctoral thesis by RNDr. Ales Zita, I am herewith submitting the requested report.

In the submitted thesis, a number of computer vision approaches of varying elaborateness were applied to a number of problems, i.e., coral detection and segmentation, automatic hand-drawn UI elements detection, the tracking of fast moving objects, and analyses of videokymograms of vocal folds.

Strengths of the presented work:

- Rigor visual evaluation of VKGs was performed for the purpose of training testing and validating automatic approaches to VKG analyses.
- Both classical and deep learning approaches were used
- The author has worked together with a number of different people and groups – only 1 paper with participation of the thesis supervisor indicated on the front page of the thesis- indicating scientific autonomy of the author

Overall, I have the impression that, through his publication list, the author - who appears to be enrolled in the PhD program since 2013 already - has managed to demonstrate his ability to integrate into scientific teams and contribute to their scientific outputs. The papers included in the thesis are from 2015, 2020 (3x), 2021, 2022, and 2023. The authors has one first authorship in a journal (paper 3), as well as three first and one last authorships in conference papers (papers 4 -7). However, what went a bit suboptimal in my opinion is that papers 4-6 do not really fit the overall topic that the student has chosen as a focus of the thesis. That may be only a matter of how the material (title and introductory chapters) is presented in the thesis, but in my opinion that may have deserved some more attention. Other weaknesses is that the introductory chapters are not really well elaborated, and the document is not self-contained. More detailed comments are listed below.

Detailed comments:

Introductory chapters:

In general, the introductory chapters appear to be a bit superficial and repetitive in some ways. Among the first chapters I would have expected a section regarding the state-of-the-art and a statement explaining in what specific ways the author claims to have advanced the state-of-the-art. The author does not really valuate or advertise his own contributions in the introductory chapters, but stays rather vague. (E.g, goal 1 is to “propose new approaches …” that can be more specific, goal 2 is to “gain expertise”. Thesis goals are no personal goals of what already existing knowledge the author wants to learn for himself but goals that advance the current state-of-the-art.)

Regarding the state-of-the-art, the author explains that not much deep learning based analyses of VKG were proposed in the past, which is true. However, in my opinion it would have been necessary to look a little more beyond one's own nose here, since VKG are very strongly related to high-speed videos (HSV), i.e., all VKG information (and more) is contained in the HSV. Deep neural networks do a great job in filtering redundancies, of which a lot are found in HSVs. Also, a google search for 'deep learning videokymography' currently returns approximately 11.600 result, and 227.000 for 'deep learning high-speed videolaryngoscopy'.

I find it laudable that the authors separates between "main contribution of the paper", and "main contribution of the author". However, while the former sections are full paragraphs, only bullet points / keywords describe author contributions. I think that this is a missed opportunity to explain in more detail the interfaces between the individual co-authors of the teams. (Things like "study participation" can be anything. Also, typically more substantial contributions are required for first or last authorship than for other authorships, such a difference is not reflected in sections 6.x.4.)

The term mucosal wave is not sufficiently explained before its first mention.

Heading 1.2 is "Anatomy", which would imply that explanations of "what is where (and how connected) in the body" would follow. Instead, it is rather the phonations process (physiology and function) what is described there. (Anatomy rather refers to the static layout of the body parts instead of its dynamics.)

Headings 1.4 is "Current trends in larynx visualization", but what is written there is rather a listing of the visualization methods invented during the last few decades. For example, current trends would include deep-learning based analyses, as well as 3D laryngoscopy.

One refers to a frame rate of 7200 fps in VKG, but I'd be wondering whether it wouldn't be more specific if one would refer to a 'line rate' of 7200 lines per second instead.

Figure 7: image optimization → image enhancement?

Section 2.3: "the author who invented the system". Not clear from the text who that was, the name and/or citation could have been included right next to that statement.

Section 6.1.3: The description of the contributions is not really clear to me here: "Length of the mucosal wave": Is that the lateral extend of the wave? "Left and right variability": Variability of what? "left and right skewing": of glottal pulses? Unfortunately, questions are not even answered in the full paper 1, but there other papers are cited. As a result, generally speaking, the thesis is not a self contained document it that sense. (See also comments on paper 1 below)

Section 6.1, 6.2, 6.7: Although crucial for the reported main result, it is a little bit unclear here what the human raters were asked to rate, and on which scale.

Sections 6.1-6.3, 6.7: Including figures / graphs for defining the parameters (humanly evaluated versus machine obtained) would have been strongly recommended.

Section 6.3: Statement regarding the availability / openness of the software would have been advisable.

It is here in section 6 that I first read about the aim of graphical segmentation, and comparing humanly assessed image parameters to parameters obtained by a machine. Why isn't that mentioned earlier already, e.g., in the list of aims?

Paper 1:

A strength of paper 1 is that agreement between humans and agreement between the computer and humans are compared for the purpose of demonstrating the computer's human level performance.

The thesis is not a self contained document: [The vibration parameters'] "detailed definition and discussion can be found in [9]", and "Manual evaluations [10]".

For segmentation, graph cuts from 2000 and 2008 were used in the thesis. Nowadays DNNs would be used for segmentation, but at the time of the publication of the paper (2015), graph cuts may have been ok.

For tracking the mucosal wave, authors mention "iterated masked cross-correlation", but it is not really clear how that works. The verbal explanation raises more questions than it answers. E.g., it says some kernel is iteratively updated, but I can only guess how the kernel is initialized and updated. No thorough evaluation of mucosal wave tracking is presented, only one example is shown (Figure 4).

Instead of explaining the iterated masked cross-correlation properly, the authors overemphasize an equation that relates to the coding of a variable mainly, presumably for the purpose of preparing statistical analysis.

Paper 2:

This is an award winning paper published 2020 (accepted in 2018) written by first author Kumar, last author Svec, and colleagues. In this paper, subjective visual evaluation of the sharpness of the lateral peaks is compared with objective measures OTQ and PQ showing strong correlations between subjective and objective measures on the level of human performance. While Woo and Metha used very similar objective measures in 1996 and 2011 using fixed amplitude thresholds, Kumar et al. varied these thresholds to maximize correlation between subjective and objective measures. Three raters independently evaluated sharpness of the lateral peaks in 45 kymograms on a 4 point scale for the left and right vocal folds separately.

Paper 3:

What can be considered extraordinary in paper 3 is the rigorous way of performing the validation, i.e., training and testing was done using three datasets, one of which contains as many as 13500 evaluations of ten evaluators annotating manually 50 VKGs of 50 patients and 200 VKGs of 40 healthy subjects by means of 9 vibratory features. Another dataset contains manual annotations of 834 keypoints, i.e., particular coordinates in the VKGs, by 6 evaluators.

A point of criticism may be that the discrimination thresholds of the machine obtained features were predetermined, but discrimination supposedly could have been further improved via tweaking these threshold.

Regarding the neighboring HSV analysis, all the information contained in a graphically segmented kymogram is also contained in the phonovibrogram (Döllinger et al. 2007), which has also been developed into a powerful software tool (approximately since 2009). In other words, analysis of segmented VKGs is pretty similar to analyzing a single line of the phonovibrogram.

Papers 4 and 5:

Paper 4 is a competition winning paper on coral detection and segmentation. The author explains that the individual methods (esp. Mask R-CNN, data augmentations) were not novel, but their joint use was “a substantial result”. Despite not focusing on this part of the thesis in my evaluation report, I believe that competitive result are reported in that paper. A similar observation may be made regarding paper 5 on automatic hand-drawn UI elements detection, but here I’d like to raise two small (potential) points of criticism: While it is laudable that the authors acknowledge for the unbalanced dataset and account for it by doing data augmentations especially in underrepresented classes, I would not call the data augmentation ‘synthesis’, since the new data is obtained by only altering the existing data in straight forward ways, i.e., no sophisticated generative approach is used as the term ‘synthesis’ would imply. Second, keywords that were raised in the last few years is few-shot learning or one-shot learning. In that sense, some additional methods may have been worthwhile trying.

Paper 6:

I also believe that paper 6 reports competitive results using ENet from 2019. A great deal has been made in paper 6 to generate realistically looking synthetic data.

One possible point of criticism relating to the suboptimal embedding of papers 4-6 into the overall topic of the thesis is that tracking of fast objects is only highly relevant to tracking vocal fold edges on the first sight. Paper 6 mainly works on non-high-framerate videos imposing the problem of heavily blurred objects that do not overlap in consecutive frames. That is not so in VKGs.

Paper 7:

Authors use MobileNetV2 to obtain lateral peak sharpness and length of the mucosal wave (on 3-point scales each). This is a first ever attempt to use a DNN for obtaining these features, so the results appear to be rather preliminary (accuracy between 0.5 and 0.61)

What is the 100% reference for the so-called mucosal wave length?