Graduate School, Uni Mittelstrasse, Mittelstrasse 43, CH-3012 Bern

Graduate School for Cellular and Biomedical Sciences

Thesis Assessment by External Co-referee

PhD candidate	Laurent Lejeune
Title of thesis	Annotating Medical Sequences in the Blink of an Eye: Segmention of Video and Volumetric Medical Sequences at Frame-rate using Sparse Point-wise Supervision
Institute/Division	Artificial Intelligence in Medical Imaging, ARTORG
Name of co-referee	Diana MATEUS
Academic degree	Professor
University/Department	Ecole Centrale Nantes / Laboratoire de Sciences du Numérique de Nantes
Town/Country	Nantes
Email address	Diana.mateus@ec-nantes.fr

We would like to ask you as co-referee to assess whether in your opinion the study meets the requirements for a doctoral thesis for the PhD program of the Graduate School for Cellular and Biomedical Sciences of the University of Bern.
We expect that at least one first author publication will be available upon completion of the doctoral thesis. In exceptional cases, the doctoral thesis can be completed without publication. In such a case, we ask you to check the quality and publishability of the data particularly thoroughly and to state explicitly whether a first author publication is to be expected from the available data.
The written dissertation should consist of a summary of the research work, a comprehensive introduction, a results section that can comprise copies of published and submitted manuscripts, an overall discussion including perspective and a CV with publication list.

Please do also grade the work with a mark (see grading scheme below) according to criteria listed on the following page. **The highest mark should be reserved for extraordinary work (top 10%).**

Your assessment and the grading will be forwarded to the doctoral candidate. Please use the box on last page for confidential remarks to the Graduate School.

PD Dr. Monica Schaller
Coordinator GCB
University of Bern
Uni Mitelstrasse
Mittelstrasse 43
CH-3012 Bern

Tel. +41 31 631 59 63
E-Mail: monica.schaller@gcb.unibe.ch
www.gcb.unibe.ch

Please send the review as signed PDF document to PD Dr. Monica Schaller (monica.schaller@gcb.unibe.ch) and Cc to Brigitte Schutte (brigitte.schutte@gcb.unibe.ch).

Please give a written assessment addressing the following topics:

- Accurate summary and correct citation of relevant literature in the introduction
- Definition of aim and hypothesis
- Novelty and importance of the findings
- Quality and interpretation of the data
- Suitability of used methods
- Overall discussion
- Clarity and comprehensibility of the text
- Quality and clarity of graphics and legends

(Text boxes will expand to accommodate full length of text.)

Assessment.

M. Lejeune dissertation addresses the problem of automatic segmentation of medical videos and volumes. The targeted scientific question is how to accelerate the collection of human expert annotations in order to build the large annotated databases required by supervised machine learning segmentation approaches. Several novel methodological formulations and software solutions are proposed discussed in the document. The dissertation is composed of eight chapters including an introduction, a background and literature review chapter, one chapter on the problem statement (protocol, dataset and software solutions), four methodological contributions each containing an experimental validation, a conclusion, and two appendices.

The **Introduction** discusses the recent shift in the medical image analysis field towards Deep Learning (DL) approaches, explaining how they differ from the more conventional machine learning methods, and emphasizing the critical role of expert annotations when training such high-capacity models. The collection of large databases is then presented as an important bottleneck in the medical image analysis field and the problem is illustrated through examples from different modalities and quantified in terms of the required annotation time. After the motivation, the chapter presents the central thesis statement, namely the proposition of an intuitive and fast annotation protocol relying on point-wise interactions and based on deep learning and multi-object tracking, capable of generalizing across datasets. The chapter concludes with a listing of the contributions and their organization in the document.

The general introduction smoothly guides the readers from the state of the art on medical segmentation to the identification of the annotation bottleneck, which defines the central targeted problem of the thesis. The main contributions were clearly identified and well-distributed into the chapters of the thesis. A short discussion on the fields involved but not developed (e.g. perception, gaze signal processing, human-machine interaction, etc) would be interesting to present the broader multidisciplinary nature of the problem addressed in this thesis.

The second chapter, **"Related works and background theory"**, starts by surveying conventional computer vision methods for segmentation from active contours, through graph-based approaches exploiting scribble annotations to conditional random fields. Then, two common techniques to overcome the lack of annotations are briefly discussed, namely active learning and crowd-sourcing. The chapter proceeds with a more complete review of weakly supervised approaches that focus on reducing the annotation complexity (e.g. through sparse or image-wise annotations). Focus is put on methods that establish a link between deep learning and graph-based approaches, as will be later the case in this dissertation. Next, the concepts of transfer learning, semi-supervised learning, and domain adaptation are introduced. The chapter concludes with a background section on random forest, deep convolutional networks, one multi path tracking approach and recursive Bayesian parameter estimation.

The chapter concisely summarizes the relevant state of the art from a broad field of study. The background for random forests, convolutional neural networks and multi-path tracking makes the thesis self-contained, providing a sufficient introduction for beginners. Although the state of the art sufficiently covers the closest trends, discussing the similarities and differences to some key closest related works in more details would highlight the novelty of the methodological propositions. Some of these discussions can be found inside the methodological chapters but a summary would have been welcomed here. The state of the art in multi-path tracking was not presented, only one method that serves as basis later was detailed, but its choice not justified. In this case, the key differences to the baseline method were clearly stated. Some details of this multi-path approach come clear only after the following chapters where the problem is formally introduced. The same applies to the recursive Bayesian Parameter estimation section. These parts could probably become appendices or a background section in a later chapter.

Chapter 3**, "Annotation Protocol, Datasets and Software Solution"**, presents the problem from the broader perspective of the user, clearly specifying requirements for the targeted annotation approach, and aiming at several improvements over current existing methods (e.g. automatic sequence unfolding, single point annotations, single pass over the sequence, generalization over different modalities and object types). The four type of sequences employed for the experimental validation are then introduced, including endoscopic videos, Computer Tomography (CT), optical microscopy and Magnetic Resonance (MR) sequences/volumes. Additional software requirements are then stipulated, which include the compatibility with an eye-gaze tracker. Finally, the chapter presents the developed web platform.

Although it is not so common to present software solutions as PhD contributions, here the role of the software is clearly determinant of the quality and usability of the proposed approach. The requirements specifications are quite ambitious and clearly translate the community's needs and challenges. The efforts to make the software publicly available and to push its development to a usable tool beyond the methodological advances are very

welcomed. It would be interesting to discuss if the chosen criteria impose any conditions on the type of volumes/sequences that can be treated, and/or the type of annotations that can be achieved.

Chapter 4, **"Expected exponential loss for gaze-based video and volume"** presents the first methodological contribution of the document. An introductory section motivates the use of eye gaze trackers and presents three closest methods using such technology. The problem of annotating sequences from gaze positions is formulated as that of semi-supervised learning. Practically the method is implemented with a traditional gradient boosting trained on superpixel overfeat features. The novelty of the proposition comes from a new Expected Exponential loss (ELL) function, generalizing the exponential loss to consider the ambiguity of unlabeled positive points. The loss extension translates to a weighting of the superpixels according to their similarity to the labeled regions. In practice, the weights are estimated through a Laplacian-based iterative label propagation algorithm, based on a distance and a gradient similarity, and run until a manual stopping criteria. Experiments on one sequence per database demonstrate a better performance of the EEL when compared against the baselines in terms of the Area Under the Curve (AUC) and F1 measures. The gazes of different experts were used to evaluate the variance of the segmentation results, and were also combined to mimic a crowdsourcing setup.

The formulation of the Expected exponential loss is sound and does effectively improve the results. It would have been interesting to report the average performance values over several sequences of each database and a comparison against more recent approach (e.g. a DL method trained on an equivalent amount of annotations). The protocol instructed experts to stare at the object only, the difficulties of processing gaze noise and the possibility of the participants non full-compliance were not addressed. Having several hyperparameters, I t would have been interesting to discuss their choice.

Chapter 5, "**Iterative multi-path tracking for video and volume segmentation with sparse point supervision,"** exposes the main methodological contribution of the thesis, based on a global optimization of the expert annotation propagation. Point-wise expert annotations are considered source superpixels, which "flow" through the frames/slides towards a global sink node, leading to a multi-source, single sink max-flow problem. The method considers an appearance model (based on a random forest trained on learned sequence-specific features) and a graph connecting all superpixels across the entire sequence. A cost is assigned to the different weights according to domain knowledge. In practice, the graph is segmented, by translating the Maximum a Posteriori (MAP) formulation of the network flow problem to an Integer Program (IP). The graph is first refined (pruned) with additional spatial and motion constraints, and the resultant IP solved with a K-shortest path algorithm. The approach is repeated iteratively using the initial predictions as pseudo labels. The approach was evaluated on several sequences of the 4 databases, showing significant improvements over concurrent methods. Further quantitative validations allow to further analyze the

components (e.g. the feature type) and behavior of the method under different supervising conditions (e.g. coverage or with outliers and missing 2D locations)

The chapter clearly positions the contribution w.r.t. the closest works among other graph-based and semi-supervised methods. The problem formulation is elegant, considering several modeling steps while maintaining the proposed approach general. The experimental validation compares against other methods and validates different aspects of the approach. The contribution was published in the Medical Image Analysis journal. It would have been interesting to compare against other learning methods that consider a temporal propagation.

Chapter 6, **"A deep embedding clustering approach, "** aims to overcome some limitations of the method above, where the superpixel feature description and appearance-similarity models are independent. The proposition assumes the object's appearance can be represented as a set of clusters in the feature space, and the distance to those clusters measures the similarity. A deep CNN is then trained following an existing deep clustering method in a self-supervised fashion. The trained model then replaces some of the likelihood terms in the original net-flow MAP formulation, which is then optimized as before. The experimental validation over the four datasets do not systematically prove the effectiveness of the approach. As discussed in the chapter's introduction and conclusion, there seem to be some sensible hyperparameters that make the optimization unstable.

I welcome the initiative to present negative results and the posterior analysis discussion. It is difficult to decide if the fault comes from the idea of simultaneous feature-similarity learning or only from its implementation with the Deep Embedded Clustering approach. It seems sound to do the joint learning, but many factors are still kept outside the learning loop, preventing the method from being fully end-to-end, and exploit this idea. A variational formulation may be better suited for the task. There are some repetitions in the introduction, base framework and dataset that could be omitted.

Chapter 7, "A Positive/Unlabeled Approach for the Segmentation of Medical Sequences using Point-Wise Supervision," presents a second more successful idea to integrate a CNN to learn the feature representation from the point-wise/gaze annotations. In this case, the importance shifts towards the loss function, modeled to handle the absence of negative annotations. The proposed loss relies on prior work on similar positive-unlabelled learning, defining a non-negative unbiased risk. The risk requires an estimate of the proportion of negative vs. positive class priors, self-learned here through a recursive Bayesian filter. In practice, A U-net-like architecture is trained with both the non-negative risk and Bayesian filter. After training, the method offers a stand-alone solution or may rely on the net-flow propagation approach to achieve regularization over time. Significant improvements were reported for the combined approach over the four datasets. The additional required input, an estimate of the class-prior proportion, is negligible regarding the achieved improvements he experimental validation is convincing and compares the state-of-the-art; it also analyzes the influence of the class-prior input in different configurations. As it experimentally

demonstrated, the class prior does not need to be accurate, given the Bayesian filter sufficiently adapts the estimate to each frame/slice.

The main contribution comes from the adaption of the non-negative risk formulation to the problem at hand and through the class-prior learning approach. The results show a compelling improvement over challenging problem. On the downside, the need for another block to predict the class priors makes the overall solution more imbricated with training steps and hyperparameters that accumulate. It would be interesting to see if a more direct learning of the class priors with the network is possible and if the spatio-temporal regularization cannot be achieved with a recurrent approach.

The conclusion chapter 8, summarizes the contributions and provides some interesting future directions.

Overall, M. Lejeune's dissertation is clear and well-illustrated. It addresses a very challenging and timely topic, the scarcity of annotations, in medical image analysis and proposes several contributions. The first is a protocol and software to acquire medical annotations with a simple interaction (eye-gaze or single click). Then, building on this protocol, M. Lejeune proposes two important methodological contributions to propagate the annotations to full masks: a graph-based modeling of the problem and a deep learning approach, which can be both combined. The modeling is elegant and incorporates a significant amount of prior knowledge in the form of well-thought likelihoods. The deep learning approach, on the other side exploits the current self-supervised trend to bring as complement the feature learning capacity of CNNs. A substantial effort was made to design a method generalizing over different modalities instead of focusing on an ad-hoc solution. Some advances and results presented in the above chapters resulted in a Medical Image Analysis journal publication. In all, the dissertation demonstrates the candidate's broad fields of knowledge from conventional statistical modeling to deep learning, considering optimization techniques. It also shows his capacity to identify interesting questions and propose meaningful solutions. Finally, the initiative of making the software open-source software will be welcomed by the community. For the reasons above, I give a positive recommendation for the defense of M. Lejuene PhD.

Recommended changes for the final thesis version.

General

- In chapter 2, the background for the multi-path tracking approach and the Bayesian parameter estimation lack of clarity as they appear before the problem formulation. Consider a re-organization of these two sections as appendices, or make sure the problem and all variables are sufficiently described before.

- Revise repetitions across chapters (in particular in the introductions and datasets of chapters 6 and 7).
- The variable name “f” represents different things in different chapters, which can lead to confusions
- Refer to own work in first person to clarify and clarify when baselines are the approaches from previous chapters (instead than citing them only with a reference)

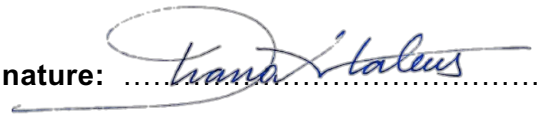
Other

- Page 3: “In their most primitive form” . Reformulate the sentence: features are still used and state of the art in many fields.
- The reliability and processing of eye-tracking readings are not mentioned. I would recommend a short discussion on how raw signals are processed to establish the sparse point annotations.
- In section 2.2. Background theory, please specify where the figures are taken from.
- Page 15 a chain of function->functionS
- Page 15 as in all machine learning setup-> as in every …
- Page 18, the alg ??? -> erroneous reference
- Page 20, each cell represent -> representS
- Page 20, maximum flow constraints -> constraint
- Page 21, what is I in Eq. 2.8?
- Page 21, the user-provided 2D locations were not yet introduced
- Page 22, \alpha and beta were not introduced
- Page 24, there seems to be a problem in Eq 2.22, one of the nominator terms and the denominator cancel each other.
- Page 24, models F and G were not introduced, same for B
- Page 29, what was the role of the other contributors (were they students?)
- Page 34, the github link seems broken
- Page 41, regarding the propagation algorithm:
    - Does the propagation affect only one frame at a time, is it only from one frame to next, or does it consider both?
    - it is not mentioned how the average gradient orientation were computed, are they taken from the intensity gradients? Are they related to the optical flow between frames?
    - how the sigma in Eq 4.4. parameters were set.
    - what does the notation Pm stand for?
    - How are the values of \epsilon set from the Pm’s
- Pages 41-42 mention optical flow, but this the purpose of the optical flow was not discussed in the text.
- Page 43, Fig 4.3. add the references to other comparing methods (Prob Estim?, EL?)
- Page 48 mentions that active learning, domain adaptation and crowd-sourcing only produce coarse segmentations, this is no longer the case
- Page 52 in this paper->chapter
- Page 91 bottom term of Eq 7.2 ?

UNIVERSITÄT
BERN

Mark: 5,5

Grading Scheme: 6.0 = excellent; 5.5 = very good; 5.0 = good; 4.5 = satisfactory; 4.0 = sufficient; below 4.0 = failed.

Date: March 25 2021. **Signature:** 

Confidential comments to the GCB.