VerSe '19: Evaluation Report

Abstract

This report details the evaluation protocol adopted in ranking the performance of VerSe '19, the *Large Scale Vertebrae Segmentation Challenge* held in conjunction with MICCAI 2019, Shenzhen, China. The challenge consisted of two phases of evaluation, one with the predicted masks, and one with mask obtained from the submitted docker containers. The report also provides the performance measures of all the eligible teams on the adopted performance metrics in both the phases of the challenge.

1. Introduction

The challenge contains two tasks: vetrebrae labelling and vertebrae segmentation in CT scans. The ground for the labelling task are in a JSON file with 3D coordinates of the vertebrae present in the image. The ground truth for the segmentation task is a .nii.gz file in the same orientation and spacing as the image file. Note that 'partially-visible' vertebrae were not annotated, i.e: they do not have either a label a segmentation mask.

The challenge was held in two phases:

1. Test Phase 1: 40 test scans released publicly. Participants asked to submit the predicted *labels* and *masks*
2. Test Phase 2: 40 hidden/private test scans. Participants asked to submit docker containers. Initially, CPU-only containers were asked. However, due to their unrealistically high run-times, GPU containers were immediately asked for.

The challenge saw more than 250 data download requests as tracked on `deep-spine.de` At the end of Test Phase 1, 18 teams submitted their results on the online evaluation portal. Of these, 11 teams submitted their predictions to `verse@deep-spine.de` as per instructions, thus being eligible for phase 1. For phase 2, 10 of the 11 teams submitted the docker containers. These docker containers were run on both phase 1 and phase 2 of the test set. However, in this report we only report the phase 2 results of the docker evaluation. Therefore, we report two sets of evaluation numbers: one for the submitted prediction files (phase1) and one for the dockers run on files of the hidden phase 2 (docker-phase2).

1.1. Evaluation Metrics

We employ two metrics for each of the tasks: Identification Rate (Id.rate) and Localisation distance (d_{mean}) for the labelling task and Dice coefficient (Dice) and Hausdorff distance (HD) for the segmentation task. It is important to note that the Id.rate employed here is different from the one in [1]. Id.rate here is just the Recall metric defined in [2], i.e it is computed at a scan-level, while the Id.rate in [1] is computed at a dataset-level. The rationale for this switch being that labelling a partial scan is relatively more difficult than a full scan. Hence, a mis-label in a partial scan should have a higher effect on the overall performance. Also, note that the online evaluation portal lists one other metrics: Precision. This metric was not considered for ranking due to an inconsistency in the definition of the 'partially-visible' vertebrae. Several methods predicted a label for the partially visible vertebra while the ground truth did not have a label for them.

2. Results & Ranking

We report the results of *segmentation* and *labelling* tasks separately in Tables 1 and Table 2 respectively. Each task has two tables, one reporting the measures and one reporting the 'points' scored by each team. *Points* are earning by a team by being statistically better than the competitors. More specifically, every team is compared with every other team in the cohort to check if the team is statistically better than the other. Wilcoxon signed-rank test with a 'greater' or 'less' hypotheses (depending on the metric) testing was employed to test the significance of the difference. Following this, a p-value of 0.01 was employed to ascertain significance. For example, if team A is significantly better than team B with a $p < 0.001$, team A gets one point. Figs. 1 and 2 illustrate the p-value matrices and subsequent binarised point matrices across teams.

Table 1: **Segmentation**: Performance of various teams (in alphabetical order) in the segmentation task for the three sets of predictions. Note: brown (no docker submission), AlibabaDAMO and INIT (erroneous docker) have missing numbers.

Team	phase1		docker–phase2	
	Dice	HD	Dice	HD
AlibabaDAMO	82.70	11.22	–	–
brown	62.69	35.90	–	–
christian_payer	90.90	6.35	89.80	7.34
christoph	43.14	44.27	46.40	42.85
huyujin	84.66	12.79	81.82	29.44
iFLYTEK	93.01	6.39	83.74	11.67
INIT	71.88	24.59	–	–
LRDE	13.97	77.48	35.64	64.52
nlessmann	85.08	8.58	85.76	9.01
yangd05	76.74	14.09	67.06	28.76
ZIB	67.02	17.35	68.96	19.25

Team	phase1		docker–phase2	
	Dice	HD	Dice	HD
AlibabaDAMO	4	4	–	–
brown	1	1	–	–
christian_payer	8	8	5	5
christoph	1	2	0	1
huyujin	4	4	3	3
iFLYTEK	10	8	3	4
INIT	2	3	–	–
LRDE	0	1	0	0
nlessmann	4	5	3	5
yangd05	2	4	2	1
ZIB	1	3	2	2

Table 2: **Labelling**: Performance of various teams (in alphabetical order) in the labelling task for the three sets of predictions. Note: brown (no docker submission, no label predictions), AlibabaDAMO, INIT (erroneous docker), and huyujin (no label predictions) have missing numbers.

Team	phase1		docker–phase2	
	Id.rate	d_{mean}	Id.rate	d_{mean}
AlibabaDAMO	89.82	7.39	–	–
brown	–	–	–	–
christian_payer	95.65	4.27	94.25	4.80
christoph	55.80	44.92	54.85	19.83
huyujin	–	–	–	–
iFLYTEK	96.94	4.43	86.73	7.13
INIT	84.02	12.40	–	–
LRDE	0.01	205.41	0.0	1000
nlessmann	89.86	14.12	90.42	7.04
yangd05	62.56	18.52	67.21	15.82
ZIB	71.63	11.09	73.32	13.61

Team	phase1		docker–phase2	
	Id.rate	d_{mean}	Id.rate	d_{mean}
AlibabaDAMO	3	5	–	–
brown	–	–	–	–
christian_payer	3	7	3	5
christoph	1	1	1	1
huyujin	–	–	–	–
iFLYTEK	5	7	2	4
INIT	2	3	–	–
LRDE	0	0	0	0
nlessmann	3	1	4	3
yangd05	1	1	1	1
ZIB	1	1	1	1

After the points have been computed and normalised to 1.0 (by dividing with the number of teams in the cohort), the final point count per-team was obtained as elaborated below. Finally, based on `final_points` as reported in Table 3, the final ranks were announced during the challenge with the winner being team **christian_payer**. Congratulations!

$$\texttt{final\_labelling\_points} = \frac{1}{3} \cdot \texttt{phase1\_labelling\_points} + \frac{2}{3} \cdot \texttt{phase2\_labelling\_points}$$

$$\texttt{final\_segmentation\_points} = \frac{1}{3} \cdot \texttt{phase1\_segmentation\_points} + \frac{2}{3} \cdot \texttt{phase2\_segmentation\_points}$$

$$\textbf{final\_points} = \frac{1}{3} \cdot \texttt{final\_labelling\_points} + \frac{2}{3} \cdot \texttt{final\_segmentation\_points}$$

References

[1] B. Glocker, J. Feulner, A. Criminisi, D. R. Haynor, E. Konukoglu, Automatic localization and identification of vertebrae in arbitrary field-of-view ct scans, in: Medical Image Computing and Computer-Assisted Intervention, Springer, 2012.

[2] A. Sekuboyina, M. Rempfler, J. Kukačka, G. Tetteh, A. Valentinitsch, J. S. Kirschke, B. H. Menze, Btrfly net: Vertebrae labelling with energy-based adversarial learning of local spine prior, in: Medical Image Computing and Computer Assisted Intervention, Springer, 2018.

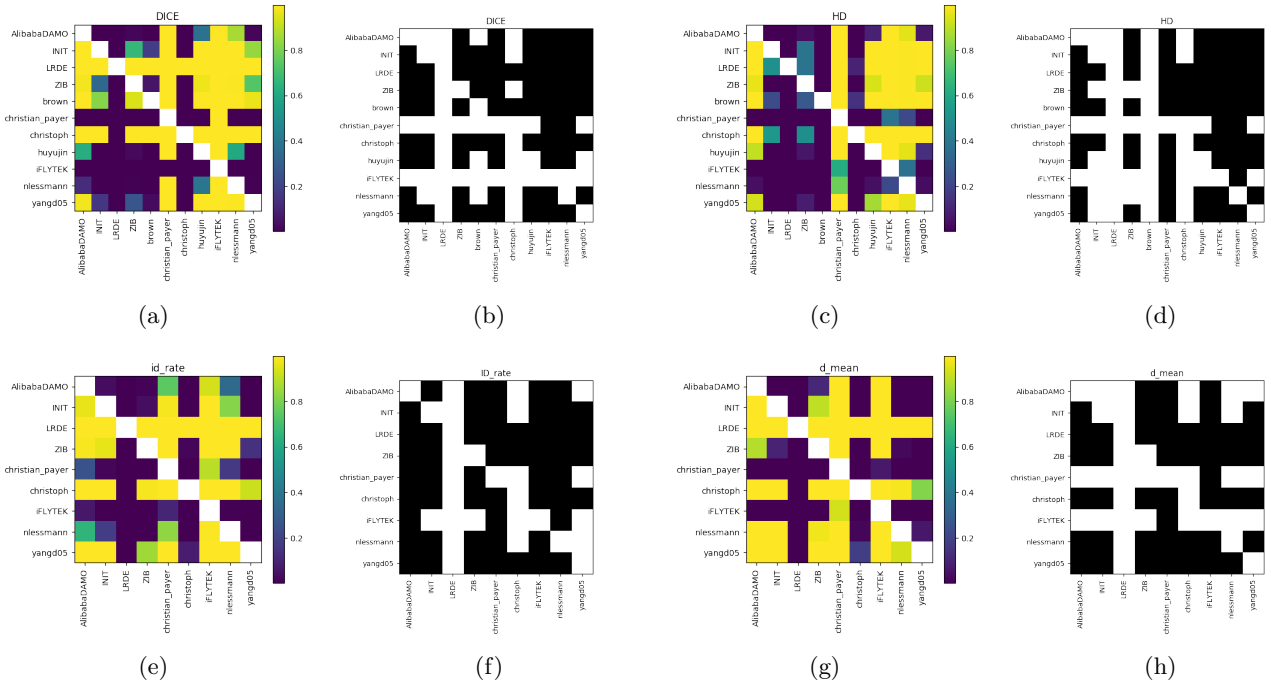

Figure 1: **Phase 1**: Illustrating the $p-$value matrices and their binarised versions for every metric used. Top and bottom rows correspond to the segmentation and labelling tasks.

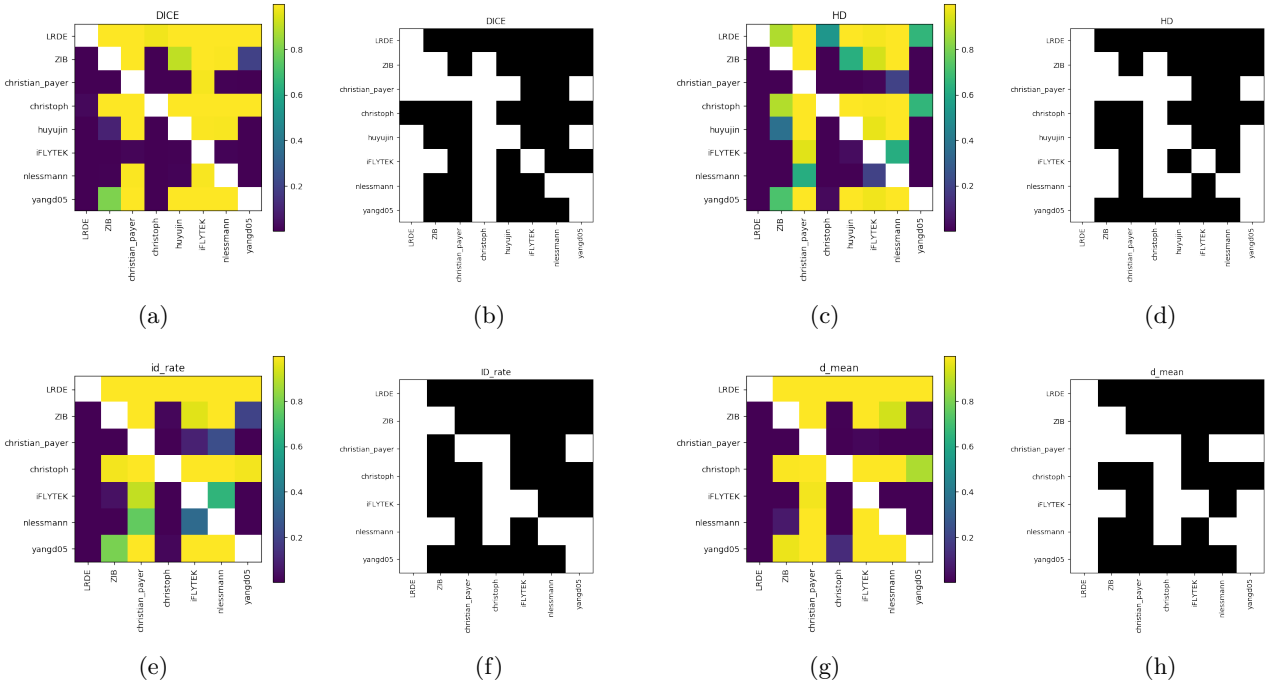

Figure 2: **Phase 2 (Docker)**: Illustrating the $p-$value matrices and their binarised versions for every metric used. Top and bottom rows correspond to the segmentation and labelling tasks.

Table 3: **Final normalised point count**: Table indicates the final points obtained by each team according to the evaluation protocol described in this article. Maximum point value by a team can be 1.0.

Team	**Final points**
christian_payer	0.691
iFLYTEK	0.597
nlessmann	0.496
huyujin	0.279
yangd05	0.216
ZIB	0.215
AlibabaDAMO	0.140
christoph	0.107
INIT	0.084
brown	0.022
LRDE	0.007