HEIDELBERG UNIVERSITY HOSPITAL

Comparison of Evaluation Metrics for Landmark Detection in CMR Images

Sven Koehler[1,2] , Lalith Sharan[1,2] , Julian Kuhm[1] , Arman Ghanaat[1] , Jelizaveta Gordejeva[1] , Nike K. Simon[1] , Niko M. Grell[1] , Florian André[1] , Sandy Engelhardt [1,2]

[1] Department of Internal Medicine III, Heidelberg University Hospital, Heidelberg
[2] DZHK (German Centre for Cardiovascular Research), partner site Heidelberg/Mannheim

Data:

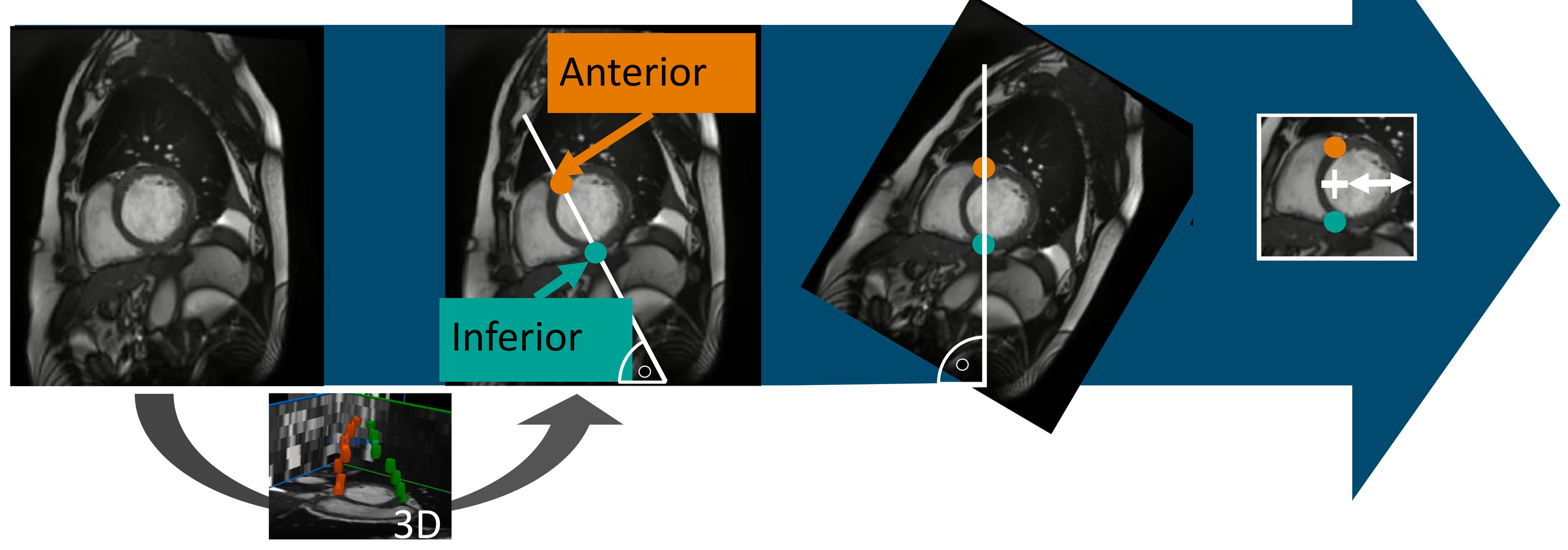
- Publicly available Automated Cardiac Diagnosis Challenge (ACDC) dataset [1]
- Short-axis cardiac magnetic resonance (CMR) images from 100 patients
- Normal & four pathological cases
- We extend this dataset by labels of the anterior and inferior RVIPs

What we do:

Right Ventricular Insertion Point (RVIP) detection

Our 1st target:

Rotational alignment & focus crop



Motivation:

- Initialization of a shape model
- Myocardial division into AHA segments
- Spatially aligned focus crop

Method:

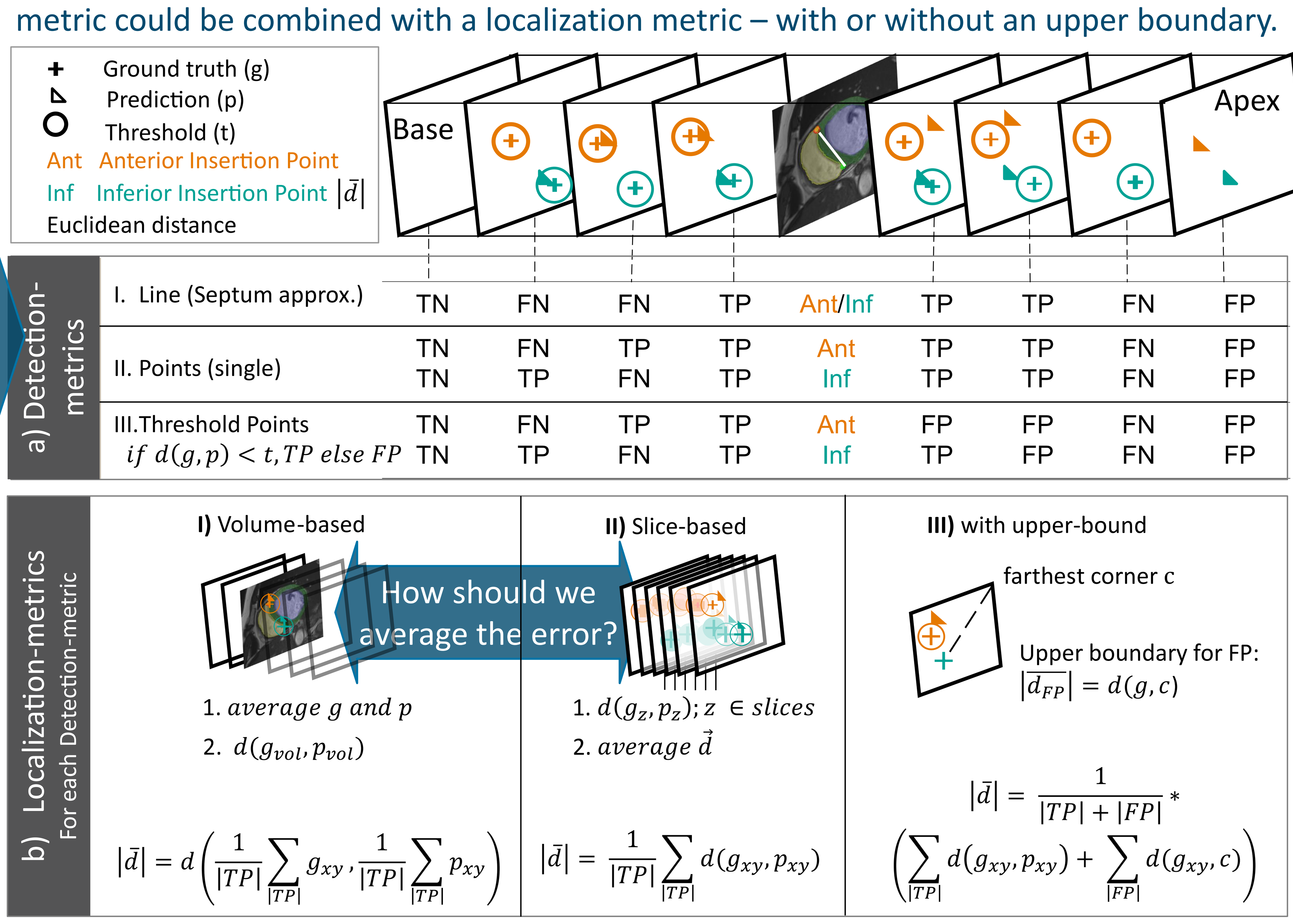
- Definition as slice-based segmentation task
- U- Net architecture
- Binary Cross-Entropy + Dice loss
- 4-fold cross-validation with respect to the pathologies as done by [2]
- Code and labels are on GitHub

Experiments:

- Baseline model (Base).
- Base + histogram matching (Var.1).
- Base + Gauss (σ = 2) GT mask smoothing (Var.2)
- Base + Gauss (σ = 4) GT mask smoothing (Var.3)

Fig. 1. Metric definition in terms of detection and localization for evaluation is not straight forward. Different definitions yield different results. (a) The total number of true positives (TP) and false positives (FP) are influential for accuracy measurements. (b) Each detection metric could be combined with a localization metric – with or without an upper boundary.

How to find a goal tailored TP/FP definition?



However, failure cases of evaluation metrics exist which influence model comparison!

Our 2nd target:

Which metrics are best to compare various CMR landmark detection methods?

How to handle FN cases for localization measure?

Tab. 1 Localization metric comparison for different experiments (Base: Baseline, Var.1 + hist. matching, Var.2 + Gauss σ = 2, Var.3: + σ = 4).

Detection-Strategy	Exp.	(i) Line $\lvert d\rvert_{Ant}\downarrow$	(i) Line $\lvert d\rvert_{Inf}\downarrow$	(i) Line $\Delta_{\alpha}\downarrow$	(ii) Points $\lvert d\rvert_{Ant}\downarrow$	(ii) Points $\lvert d\rvert_{Inf}\downarrow$	
(i) Volume-based	Base	5.92 ± 4.83	3.86 ± 5.32	3.80 ± 4.09	7.16 ± 6.88	5.79 ± 7.17	Var. 3
	Var.1	*5.58* ± 6.25	4.16 ± 5.75	4.60 ± 7.12	6.88 ± 7.71	4.86 ± 6.51	
	Var.2	6.26 ± 7.08	3.54 ± 3.83	4.13 ± 5.24	6.93 ± 8.06	5.40 ± 9.08	
	Var.3	5.86 ± 4.95	*3.33* ± 3.47	*3.67* ± 3.41	*6.67* ± 5.72	*4.17* ± 5.57	
(ii) Slice-based	Base	4.42 ± 5.66	3.96 ± 7.07	*2.70* ± 3.09	5.08 ± 9.04	3.89 ± 6.96	No clear winner!
	Var.1	*3.79* ± 7.20	3.02 ± 4.39	3.31 ± 5.77	*4.05* ± 7.67	3.00 ± 4.08	
	Var.2	3.88 ± 4.97	3.12 ± 7.10	3.63 ± 7.51	4.08 ± 5.47	3.51 ± 8.14	
	Var.3	4.42 ± 5.67	*2.48* ± 2.20	3.81 ± 9.00	4.58 ± 6.77	*2.71* ± 2.89	
(iii) Slice-based, ↑-bound	Base	50.33 ± 65.01	49.68 ± 65.98	30.85 ± 39.20	35.05 ± 46.46	29.93 ± 59.14	Var. 1
	Var.1	*37.07* ± 46.70	*36.83* ± 45.74	*24.25* ± 29.92	*29.62* ± 42.46	*14.80* ± 28.56	
	Var.2	48.53 ± 64.66	47.58 ± 63.89	30.79 ± 38.75	36.82 ± 57.30	27.39 ± 46.82	
	Var.3	55.05 ± 74.88	53.76 ± 75.66	34.66 ± 44.92	39.44 ± 61.48	34.57 ± 58.76	

What is the error? vs Which method is best?

Tab. 2 Detection metric comparison for different experiments (Base: Baseline, Var.1 + hist. matching, Var.2 + Gauss σ = 2, Var.3: + σ = 4).

Detection-Strategy / Exp.	(i) Line $TPR\uparrow$	(i) Line $PPV\uparrow$	(ii) Points $TPR_{Ant}\uparrow$	(ii) Points $TPR_{Inf}\uparrow$	(ii) Points $PPV_{Ant}\uparrow$	(ii) Points $PPV_{Inf}\uparrow$	
Base	0.84 ± 0.22	0.84 ± 0.22	0.89 ± 0.16	0.91 ± 0.19	0.80 ± 0.22	0.77 ± 0.24	Var. 1
Var.1	*0.88* ± 0.16	*0.85* ± 0.19	*0.91* ± 0.15	*0.96* ± 0.10	0.79 ± 0.21	0.79 ± 0.21	
Var.2	0.85 ± 0.21	*0.85* ± 0.23	0.88 ± 0.19	0.92 ± 0.16	*0.81* ± 0.23	*0.80* ± 0.22	
Var.3	0.82 ± 0.25	0.83 ± 0.26	0.88 ± 0.21	0.89 ± 0.20	0.79 ± 0.24	*0.80* ± 0.24	
Base & Thresh.	(iii)		0.88 ± 0.20	0.90 ± 0.22	0.76 ± 0.25	0.73 ± 0.26	Var. 1 or Var. 2
Var.1 & Thresh.	(iii)		*0.90* ± 0.18	*0.96* ± 0.10	0.77 ± 0.23	0.78 ± 0.21	
Var.2 & Thresh.	(iii)		0.88 ± 0.20	0.91 ± 0.18	*0.78* ± 0.24	*0.79* ± 0.23	
Var.3 & Thresh.	(iii)		0.86 ± 0.23	0.89 ± 0.21	0.77 ± 0.25	*0.79* ± 0.24	

Precision vs Recall
The definition of TP/ FN cases is crucial

Results

- Fast and robust method for spatial alignment of CMR
- Public extension of the ACDC dataset by RVIP labels
- Showcased some pitfalls of apparently simple metrics

For the comparison of detection-based methods, based on our results, we recommend:

1. Provide multiple views of your results. At least one detection- and one localization metric.
2. Especially for 3D data, the detection metric and image resolution needs to be reported.
3. Finally, the handling of False Negative (FN) cases within the localization-based metrics needs to be reported. Here, we welcome a general discussion on how to define transparent and fair upper boundaries for FN penalization.

Code and labels:
https://github.com/CardioAI/rvip_landmark_detection

Acknowledgement. The research was supported by Informatics for Life project funded by the Klaus Tschira Foundation and SDS@hd service by the MWK Baden Württemberg and the DFG through grant INST 35/1314-1 FUGG and INST 35/1503-1 FUGG.

[1] Bernard O et al. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: Is the Problem Solved? IEEE Trans Med Imaging 37.11 (2018), pp. 2514–2525.
[2] Koehler S et al. How well do U-Net-based segmentation trained on adult cardiac magnetic resonance imaging data generalise to rare congenital heart diseases for surgical planning? Med Imaging Ed. by Fei B, Linte CA. SPIE, 2020, p. 55.