Label Augmentation Method for Medical Landmark Detection in Hip Radiograph Images

Yehyun Suh[1,2] **Peter Chan**[3,4] **J. Ryan Martin**[4] **Daniel Moyer**[*1,2]
[1]Vanderbilt Univserity [2]Vanderbilt Institute for Surgery and Engineering
[3]University of Texas Southwestern Medical Center [4]Vanderbilt University Medical Center
{yehyun.suh, daniel.moyer}@vanderbilt.edu

Abstract

This work reports the empirical performance of an automated medical landmark detection method for predict clinical markers in hip radiograph images. Notably, the detection method was trained using a label-only augmentation scheme; our results indicate that this form of augmentation outperforms traditional data augmentation and produces highly sample efficient estimators. We train a generic U-Net-based architecture under a curriculum consisting of two phases: initially relaxing the landmarking task by enlarging the label points to regions, then gradually eroding these label regions back to the base task. We measure the benefits of this approach on six datasets of radiographs with gold-standard expert annotations. Link to code.

1 Introduction

Total Hip Arthroplasty (THA), also known as Total Hip Replacement, is a standard procedure to address hip pain by removing and replacing the damaged joint with artificial components [1]. Pre-surgical, intra-surgical, and post-surgical calculation of pelvic tilt, implant cup tilt, and evaluation of THA relies on medical landmarks on X-ray and Fluoroscope images of the patient's pelvis. In clinical practice, orthopedic clinicians manually select markers to make these assessments. Measurements of implant alignment during recovery and later regular use are vital as decision-making factors for potential future correction procedures, such as in the event of adverse implant positioning/alignment. Therefore, automating the process of medical landmark detection for orthopedic clinicians, especially in hip radiographs, is a target for computer vision in the medical field.

In this work, we extend a label augmentation method, initially constructed for Total Knee Arthroplasty (TKA) [2], to the THA case. We benchmark on both a knee dataset as well as five hip datasets with different imaging conditions. We show that our proposed method produces error distances on the order of $\sim$1-4 pixels, greatly outperforming baseline methods. Moreover, we show that traditional augmentation is actively harmful to this particular task due to the imaging protocol.

2 Method

Conventional augmentation on the training images, such as rotating, flipping, resizing without padding, or color jittering distorts the patterns or creates invalid medical images, making the label of the data to be no longer preserved post-transformation. This is because standardized positioning of radiographs is critical for maintaining consistency, reproducibility, and accuracy. Therefore, instead of augmenting the training images, we implemented a Label Augmentation method that enhances the performance of the U-Net [3] using the same basic architecture and gradient-based learning. To be specific, the images' labels are first iterated a predetermined number of times to dilate them. It is allowed for these dilated labels to overlap. The dilated labels are used to train the prediction network (a U-Net), and after the training stages are completed, labels gradually erode over time.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

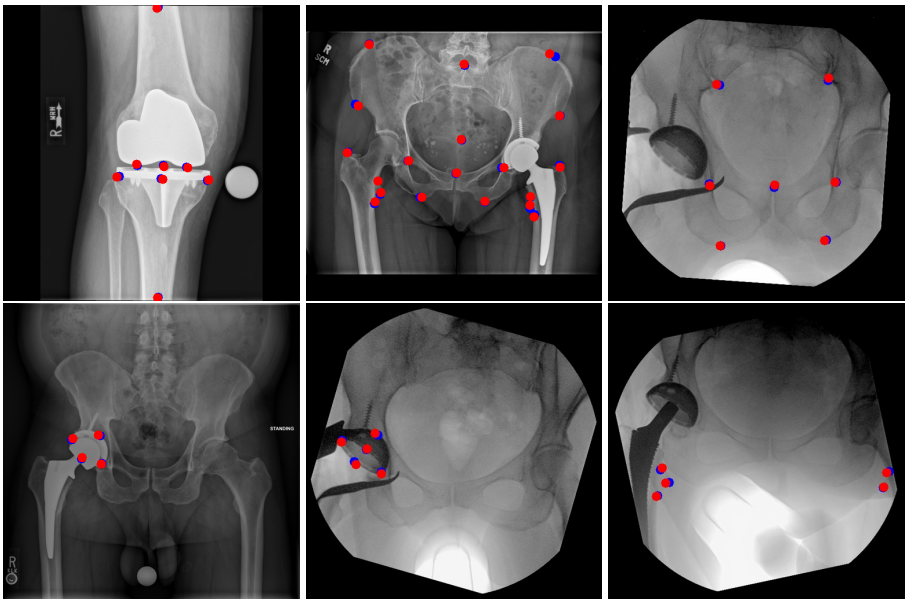

Figure 1: From left to right and top to bottom, each image represents an X-ray of the knee, an X-ray of the pelvis, a Fluoroscope of the pelvis, an X-ray of the pelvis with hip implant, a Fluorocope of the pelvis with hip implant, and a Fluoroscope of the pelvis with trochanter with predicted outputs (red) and ground truth labels (blue).

As training goes on, we increase or decrease the size of each label, which causes label imbalance. Re-weighting the loss function, which biases predictions away from degenerate solutions, is a frequent static imbalance solution. We create a dynamic re-weighting method for dynamic imbalances, where w is the label weight that would have been applied if we had not utilized the label augmentation and $\tilde{w}$ is the re-weighted w:

$$\tilde{w} = w \times \frac{\text{input image size} - (\text{number of dilated pixels} + \text{number of label pixels})}{(\text{number of dilated pixels} + \text{number of label pixels})} \tag{1}$$

By re-weighting labels by $\tilde{w}$, we keep the relative loss value of each label consistent across the learning curriculum.

3 Experiments & Results

To evaluate our model, we have used six datasets: an X-ray of the knee (216 images, 8 lables), an X-ray/Fluoroscope of the pelvis (329/286 images, 20/8 labels), an X-ray/Fluoroscope of the pelvis that consists implant (210/84 images, 4/5 labels), Fluoroscope of the pelvis that consists trochanter (159 images/ 6 lables). Internal datasets were annotated by one medical student and were divided 4:1 for the train:validation. Before resizing, all the images and the corresponding label masks were padded to have the same height and width. They were resized to a standard size (512×512) and normalized to [0,1].

Our model's encoder used pretrained weight from ImageNet [5] and used the Label Augmentation method and re-weighting scheme with pixel-wise cross-entropy loss to train the model. For the label augmentation, we have dilated the pixel for 65 iterations and eroded by 10 iterations every 50 epochs. As a validation metric, we utilize the Root Mean Squared Error (RMSE) of the pixel-wise distance between the most probable pixel (the maximal output logit value) and the groundtruth label position.

We compared the performance of our method to three different methods: a baseline U-Net, U-net with training set augmentation, and U-net with both training set augmentation and label augmentation. As shown in Table 1, in comparison to all baseline methods, our training method performs better: the mean RMSE across datasets has decreased from over 100 to less than 4. Also, we have done random rotation of a maximum of 20 degrees for the training set augmentation, and it did not work on both baseline U-net and U-net with label augmentation. We plot the results in Figure 1.

Table 1: Radio. stands for radiographs, Fluoro indicates fluoroscopes, TA is the training set augmentation, and SDR represents the Success Detection Rate.

Dataset	Validation Set (SDR(%))					
	Experiment	Mean RMSE	< 2	< 2.5	< 3	< 4
Knee Radio.	Baseline	124.82	0.29	0.29	0.58	0.58
	Baseline + TA	137.08	0	0	0	0
	Proposed + TA	5.26	39.53	54.36	63.95	68.60
	Proposed	**1.79**	**63.08**	**78.49**	**88.37**	**95.64**
Pelvis Radio.	Baseline	175.03	0	0	0	0
	Baseline + TA	180.77	0	0	0.08	0.15
	Proposed + TA	8.75	17.54	26.53	33.33	41.87
	Proposed	**3.98**	**38.15**	**52.08**	**63.38**	**74.15**
Pelvis Fluoro.	Baseline	129.70	0	0	0	0
	Baseline + TA	135.50	0	0	0	0
	Proposed + TA	8.74	15.27	22.18	27.29	36.45
	Proposed	**3.09**	**41.67**	**55.92**	**65.13**	**74.78**
Pelvis Radio. w Implant	Baseline	81.08	0	0	0	0
	Baseline + TA	151.79	0	0	0	0
	Proposed + TA	5.88	24.04	42.26	54.76	66.07
	Proposed	**1.82**	**64.35**	**80.10**	**86.65**	**93.52**
Pelvis Fluoro. w Implant	Baseline	154.86	0	0	0	0
	Baseline + TA	150.35	0	0	0	0
	Proposed + TA	5.46	24.14	45.98	56.32	64.37
	Proposed	**2.08**	**47.06**	**70.59**	**81.18**	**95.29**
Pelvis Fluoro. w Trochanter	Baseline	151.89	0	0	0.47	1.42
	Baseline + TA	166.79	0	0	0	0
	Proposed + TA	6.46	17.45	25.0	29.72	36.79
	Proposed	**3.55**	**34.90**	**43.75**	**54.17**	**66.67**

Table 2: Label 1 to 5 stands for Teardrops, most inferior aspect of Ischium, medial most point on Lesser Trochanter, superior most point on Greater Trochanter, and center of Sacrococcygeal Juntion.

Model	RMSE(mm)					
	Mean	Label 1	Label 2	Label 3	Label 4	Label 5
Muford et al. [4]	**3.3**	2.7	**3.1**	**2.1**	3.0	**5.6**
Ours	3.60	**2.19**	3.43	2.13	**2.06**	8.17

We also compared our results with Mulford et al. [4] on pelvis X-ray as shown in Table 2. Despite our model having one-third of the dataset size, we have outperformed Mulford et al. [4] in some labels where they had 1000 images for training and validation. Also, considering that our dataset had 55 images that did not have the center of the Sacrococcygeal Junction (Label 5) and only one annotator, our model has shown a decent performance.

4 Conclusion & Discussion

Our study presents a promising approach to automated medical landmark detection in hip radiographs using the Label Augmentation method combined with dynamic re-weighting. We have shown that the conventional augmentation on medical dataset decreases the performance of the model and have shown the potential to be used in general landmark detection in medical imaging. However, our results were based on the validation set not test set and our model has yet to show outstanding performance on some of the tasks in datasets such as on X-ray of the pelvis. We hypothesize that this may be due to the inconsistency in data labeling since we had only one annotator compared to Mulford et al. [4], where they had two annotators and selected the medial points from each annotator. In our future works, we plan to collect prospective test dataset and experiment inter-rater reliability to test the performance of our model to determine the noise ceiling of prediction accuracy and implement this work in diverse fields of medicine.

Potential Negative Societal Impact

Since the model is trained on data collected from actual patients, if this data is not representative of the entire population, the automated systems may be biased against certain groups of people, leading to disparities in healthcare access and outcomes.

Acknowledgement

This work was funded in part by NSF 2321684 and the Wellcome LEAP Multi-Channel Psych program, as well as a seed grant from the Vanderbilt Institute for Surgery and Engineering (VISE).

References

[1] M. Anger, T. Valovska, H. Beloeil, P. Lirk, G. P. Joshi, M. Van de Velde, J. Raeder, and the PROSPECT Working Group* and the European Society of Regional Anaesthesia and Pain Therapy . Prospect guideline for total hip arthroplasty: a systematic review and procedure-specific postoperative pain management recommendations. *Anaesthesia*, 76(8):1082–1097, 2021. doi: https://doi.org/10.1111/anae.15498. URL https://associationofanaesthetists-publications.onlinelibrary.wiley.com/doi/abs/10.1111/anae.15498.

[2] Yehyun Suh, Aleksander Mika, J. Ryan Martin, and Daniel Moyer. Dilation-erosion methods for radiograph annotation in total knee replacement. In *Medical Imaging with Deep Learning, short paper track*, 2023. URL https://openreview.net/forum?id=bVC9bi_-t7Y.

[3] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015. URL http://arxiv.org/abs/1505.04597.

[4] Kellen L. Mulford, Quinn J. Johnson, Tala Mujahed, Bardia Khosravi, Pouria Rouzrokh, John P. Mickley, Michael J. Taunton, and Cody C. Wyles. A deep learning tool for automated landmark annotation on hip and pelvis radiographs. *The Journal of Arthroplasty*, 2023. ISSN 0883-5403. doi: https://doi.org/10.1016/j.arth.2023.05.036. URL https://www.sciencedirect.com/science/article/pii/S0883540323005600.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.