Project Proposal: Improved Generalization of Pelvis X-ray Landmark Detection

Benjamin Killeen (student)
killeen@jhu.edu

Cong Gao (mentor)
cgao11@jhu.edu

Mathias Unberath (mentor)
unberath@jhu.edu

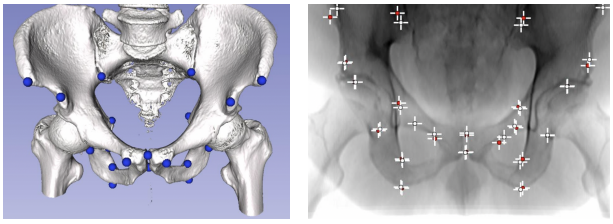
I. Background

Minimally invasive hip surgery is a desirable method for many patients. Although its benefits remain controversial with regard to pain management and recovery time, many patients strongly prefer a smaller incision to more traditional hip surgery [1]. Unfortunately, these cosmetic advantages translate to additional complexity for the surgeon. Minimally invasive hip surgery requires navigation in and manipulation of anatomical structures which are underneath unbroken skin and thus not reliably visible to the surgeon [2], [3]. At the same time, correctly aligning the cup and stem is crucial to the operation's success. In the past, this has been achieved by using the minimal incision as a "mobile window" for identifying anatomical landmarks, but this can result in unreliable outcomes [1].

Alternatively, fluoroscopic imaging provides intraoperative 2D visualization of the hip anatomy, but it presents its own set of challenges [2]. First and foremost, the mental interpretation of 2D X-ray images places an undesirable burden on the surgeon, at a time when her chief concern should be correctly aligning the hip. Computer-assisted tracking systems overcome the requirement for mental 2D/3D registration of the image with the anatomy [2]. Based on the fluoroscopic image of the hip, they can automatically track desired objects and display their poses in the context of a preoperative plan [2].

A. Significance

The systems we investigate here involve the registration of intraoperative 2D fluoroscopic images with a 3D preoperative model. Improper initialization of traditional registration algorithms, such as Iterative Closest Point (ICP) and its many variants, can lead to large registration errors. This is because of the numerous local minima which may exist in the cost optimization's function. A better "first guess" makes it much more likely that ICP converges on the actual minimum, a reliable registration. This first guess typically takes the form of human input, identifying anatomical landmarks in the hip anatomy, such as those shown in Fig. 1. Yet human input is undesirable for two reasons. First and foremost, the time required for human landmark annotation is not insignificant. Even a 4-5 second delay interrupts the surgical procedure, resulting a disjointed alignment process. Sub-second registration, on the other hand, would allow more continuous adjustment of the cup and stem.



(a) Simulated Model (b) Simulated X-ray

Fig. 1. Anatomical landmarks of the hip anatomy. (a) shows a simulated 3D model of the hip from [4]. (b) shows the simulated X-ray of that model, with the same anatomical landmarks. Images from [4].

B. Specific Aims

Prior work has shown that deep learning (DL) based techniques can identify anatomical landmarks in a fast and reliable manner, in order to initialize an ICP algorithm [2], [5]. Unlike ICP algorithms, deep neural networks (DNNs) learn generalizable features from labeled training data, and use them to interpret previously unseen images [6]. For example, [5] use simulated images to generate arbitrarily large training data with perfectly known ground truth anatomical landmarks. They show that a multi-stage DNN trained on these simulated images can generalize well to real-world images but are susceptible to scenarios not seen during training [5]. A surgical tool which occludes the image can severely compromise the DNN's ability to detect anatomical landmarks. Since the goal of automatic landmark detection is continuous, intraoperative feedback for the surgeon, it would be impractical for the surgeon to withdraw her tools during every registration. Thus, we aim to improve the generalization of DNNs from simulation to real-world scenarios which are not encountered during training.

There are many possible approaches for improving sim-to-real generalization. We propose using a novel patch-normalized convolution (PNC) layer, which constrains feature descriptors to a local region at every scale, described in Sec. III. Based on preliminary results, PNC shows an improved ability to generalize to unseen types of noise, especially additive noise patterns and contrast adjustments. We anticipate that DNNs which employ PNC will be particularly effective for occlusions by surgical tools due to the high contrast between these tools and typical intensities for an unoccluded X-ray.

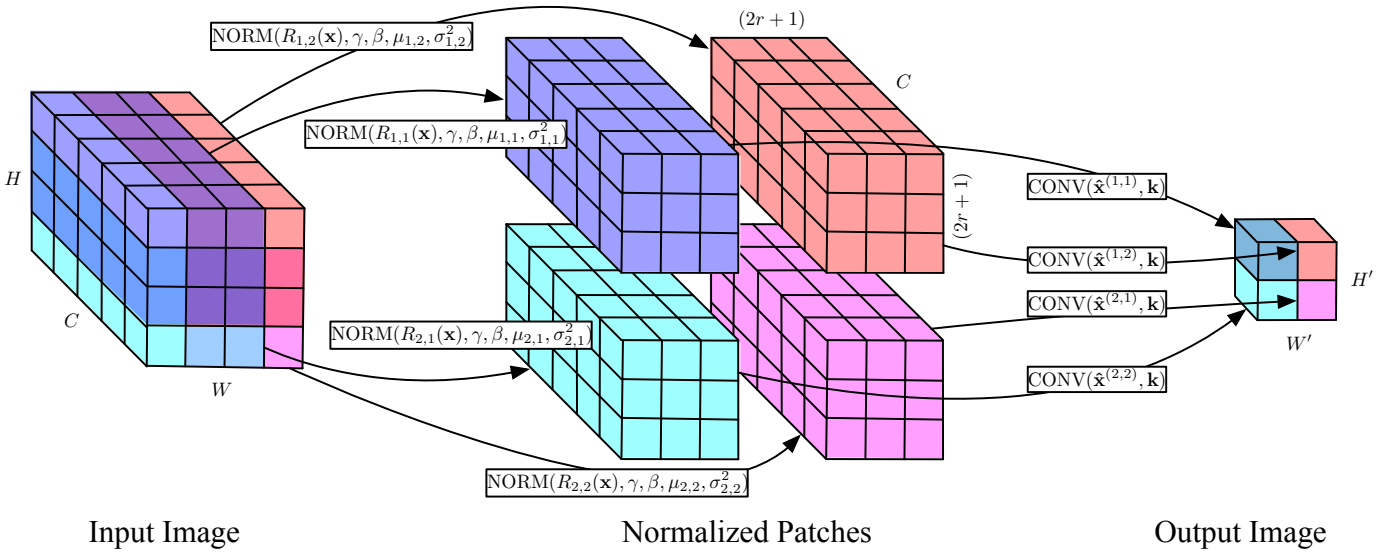


Fig. 2. Patch-normalized convolution consists of separate normalization parameters for each receptive field, or "patch." When a value is shared by two receptive fields, it must be functionally duplicated before being processed by the convolution.

II. Deliverables

See Table I.

III. Technical Approach

(The formulations in Sec. III-A, III-B, III-C are from a forthcoming paper by the authors. They are original work and are included here for completeness since they cannot yet be found in the literature. However, the focus of our proposal is the application of patch-normalized convolution, rather than its theoretical formulation.)

Much of the recent success in computer vision is due to the advent of the deep convolutional neural network, which has at its core the convolutional layer. We propose to apply DNN architectures based on prior work to anatomical landmark detection, incorporating a novel type of convolutional layer, the **Patch-normalized Convolution** (PNC). We hypothesis that the spatially local nature of the PNC layer as well as its robustness to noise will enable greater generalization to real X-ray data. We refer to [7] and [8] for a discussion of the U-Net and stage-based DNN architectures, respectively, which we will employ for landmark detection.

State-of-the-art DNNs, including the afforementioned U-Net and stage-based network, usually pair a convolutional layer with a normalization layer [9]. Here we review these concepts briefly in order to lay the groundwork for PNC, which combines a convolution with a kernel-dependent normalization. For a more detailed treatment of convolutional layers, we refer the reader to [9] and, for the normalizations we discuss, [10].

A. Convolutional Layer

For simplicity, we formulate the 2D convolutional layer over a single kernel, omitting the consideration of an additive bias. Let $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$ be a map of feature vectors, such as an image, where C is the number of input channels. In an RGB image, for example, $C = 3$, whereas for deeper convolutional layers, C may be much larger. Let $\mathbf{k} \in \mathbb{R}^{(2r+1)\times(2r+1)\times C}$ be a kernel with size $(2r+1)^2$. The stride s of the convolution is the interval between sample points in $\mathbf{x}$, and the padding p is the width of a border often added to $\mathbf{x}$ in order to preserve image dimensions [6] In general, both the kernel size and stride may have different values in the horizontal and vertical directions, but this is not often seen in practice.

The convolutional layer computes an output image $\mathbf{y} \in \mathbb{R}^{H' \times W'}$ with values

$$\text{CONV}(\mathbf{x}; \mathbf{k})_{i',j'} \equiv \sum_{c=1}^{C} \sum_{u=-r}^{r} \sum_{v=-r}^{r} x^{(c)}_{i+u,j+v} k^{(c)}_{r+u+1,r+v+1}, \tag{1}$$

where

$$i \equiv (i' \cdot s) - p + r, \qquad j \equiv (i' \cdot s) - p + r, \tag{2}$$

and H', W' are computed similarly. Often, $y_{i',j'}$ is referred to as an *output neuron* and $\mathbf{k}$ as a *feature filter*. This is because $y_{i',j'}$ is "activated" where the image $\mathbf{x}$ contains patterns similar to $\mathbf{k}$.

The *receptive field* of a neuron at any layer in the network is typically defined as the pixels in the initial input image that contribute to its calculation. For very deep layers in the network, the receptive field for a single neuron can encompass the entire input image, but for early layers it is much smaller. In particular, output neurons in the first convolutional layer have a receptive field determined by the kernel size, and this property holds when considering *immediate* receptive fields throughout the network. That is, the receptive field of a neuron $y_{i',j'}$ from its input feature tensor $\mathbf{x}$ consists of the components

$$\mathcal{R}_{i',j'}(\mathbf{x}) \equiv \{x^{(c)}_{i+u,j+v} | u, v \in [-r, r], c \in [1, C]\}, \tag{3}$$

	Algorithm	DNN for landmark detection.
	Implementation	PyTorch Implementation, Made Public on GitHub
Minimum	Validation	Anatomical landmark detection results on real data, matching prior work.
	Documentation	Inline code documentation.
	Presentation	Final written report, in-class presentation.
	Algorithm	DNN for landmark detection **using PNC**.
	Implementation	PyTorch implementation, made public on GitHub, **ready for academic use**.
Expected	Validation	Anatomical landmark detection results on real data, **exceeding prior work**.
	Documentation	**Organized and complete code** documentation.
	Presentation	Final written report, in-class presentation.
	Algorithm	DNN for landmark detection **using PNC**.
	Implementation	PyTorch implementation, made public on GitHub, ready for academic use.
Maximum	Validation	Anatomical landmark detection results on real data **with demonstrable generalization**.
	Documentation	Organized, complete code documentation, final report, **academic publication**.
	Presentation	Final written report, in-class presentation, **academic publication**.

TABLE I
DELIVERABLES

which are involved in the summation in (1). In Section III-C, we discuss the patch-normalized convolution, which utilizes $\mathcal{R}$.

B. Normalization Methods

A popular family of normalization methods involves learning an ideal distribution parameterized by γ and β. These methods compute image statistics to first normalize the inputs to a unit range, then rescale the inputs to match a distribution with mean β and variance γ;

$$\mathrm{NORM}(\mathbf{x};\gamma,\beta,\mu,\sigma^2) \equiv \gamma\frac{\mathbf{x}-\mu}{\sqrt{\sigma^2+\epsilon}}+\beta \tag{4}$$

where ϵ is a small positive number.

The calculation of image mean μ and variance σ^2 depends on the choice of method. These methods include batch normalization (BN) [11], instance normalization (IN) [12], layer normalization (LN) [13], and group normalization (GN) [10], with the most popular being batch normalization. To distinguish between these, let $\mathcal{S}_t(\mathbf{x})$ denote a set choice of elements in the input image, where t is an index. Each method utilizes a different definition of $\mathcal{S}$ to compute

$$\mu_t = \frac{1}{|\mathcal{S}_t(\mathbf{x})|}\sum_{x\in\mathcal{S}_t(\mathbf{x})} x, \quad \sigma_t^2 = \frac{1}{|\mathcal{S}_t(\mathbf{x})|}\sum_{x\in\mathcal{S}_t(\mathbf{x})}(x-\mu_t)^2. \tag{5}$$

In BN [11], for instance, the statistics are computed over multiple images, but separately for each image channel:

$$\mathcal{S}_{t=(i_t,j_t,c_t,n_t)}(\mathbf{x}) = \{x_{i,j}^{(c,n)} | c = c_t\}. \tag{6}$$

We refer to [10] for a full treatment of IN, LN, and GN. Here, the crucial point to note is that for each of these methods, $\mathcal{S}$ does not depend on the spatial indices i, j. In Sec. III-C, we describe a novel normalization technique which restricts the normalization to a local region.

C. Patch-normalized Convolution

PNC combines a novel normalization technique with a modified convolutional layer to compute image features based on local normalization. This is accomplished by computing the image statistics in Eq. 5 over the receptive field, *i.e.* $\mathcal{S} = \mathcal{R}_{i',j'}(\mathbf{x})$ (see also Eq. 3).

$$\mathcal{S}_{t=(i_t,j_t,c_t,n_t)}(\mathbf{x}) = \{x_{i,j}^{(c,n)} | c = c_t\}. \tag{7}$$

Note that for convolutions where $s < 2r$, receptive fields overlap, as shown in Fig. 2. For this reason, the initial normalization cannot be fully decoupled from the convolution.

Following our formulations above, the PNC layer computes

$$\mathrm{PNC}(\mathbf{x};\mathbf{k},\gamma,\beta)_{i',j'} \equiv \sum_{c=1}^{C}\sum_{u=-r}^{r}\sum_{v=-r}^{r}\left(\gamma\frac{x_{i+u,j+v}^{(c)}-\mu_{i,j}}{\sigma_{i,j}}+\beta\right)k_{r+u+1,r+v+1}^{(c)} \tag{8}$$

where $mu_{i,j}$ and $\sigma_{i,j}$ are computed over $\mathcal{R}_{i',j'}(\mathbf{x})$. Note that due to the duplication of memory, a naive implementation of Eq. 8 faces serious efficiency concerns. Fortunately, a reformulation of Eq. 8 allows for much more efficient memory utilization, taking advantage of box kernels to compute μ and σ^2. However, the details of this more efficient formulation are beyond the scope of this proposal.

IV. DEPENDENCIES

Our primary dependencies are simulated and real fluoroscopic images of the hip with anatomical landmarks. Simulated X-ray data has been used in an ongoing manner by Cong Gao. Fortunately, these are already resolved. The real X-ray data requires some formatting, for which Robb Grupp is an ongoing contact. Additionally, we are heavily reliant on advanced computational resources for experimentation and ablation studies of any proposed method. The MARCC compute cluster is a reliable high-compute system with multiple redundancies for high-capacity data storage. Alternatively, we have guaranteed access to two personal workstations with high-speed SSD primary drives and high-capacity HDD backup data drives. Any code, documentation, or statistical results are version-controlled and backed up using GitHub.

Recently, based on [8], we realized it might be of academic interest to evaluate our method's generalization ability to images which are occluded by surgical tools in a previously unseen manner. Although this is not core to our aim of improving sim-to-real generalization, it is nevertheless of interest. Therefore the effort to obtain real images with surgical tool occlusions is ongoing.

Table II summarizes all our dependencies.

Dependency	Solution	Alternative	Status
Anatomical Landmark Detection Software	`Generalizing_Pelvis_Landmark_Detection` Repository Access	NA	✓
DeepDRR Dataset of Simulated Fluoroscopic Images	Transfer from Cong Gao	NA	On Personal Workstation
Computational Resources (GPU)	MARCC Cluster Access	Personal Workstations (3x total GPUs)	Allocation Granted
Real X-ray Images for Testing	Robb Grupp	NA	On BIGSS Shared Drive
Real X-ray Images with Occlusions (new)	Authors of [5]	Mathias Unberath	IN PROGRESS
Efficient PNC PyTorch Implementation	Xingtong Liu	NA	✓

TABLE II
DEPENDENCIES

Milestone	Date	Status
Obtain simulated X-ray data from Cong Gao	02/15	✓
Obtain Real X-ray data from Robb Grupp	02/11	✓
Finalize simulation training pipeline	03/01	
Finalize Real X-ray validation pipeline	03/07	
Finalize DNN architecture/algorithm	03/21	
Finish ablation study	04/14	
Finish statistical analysis	04/21	
Presentation	05/05	
Final report	05/15	
Academic publication	TBD	

TABLE III
MILESTONES

V. Milestones and Status

See Table III.

VI. Schedule

See Table IV.

VII. Management Plan

Ongoing communication between the student, Benjamin Killeen, and the direct mentor, Cong Gao is facilitated by Slack and workspace proximity. Weekly meetings have been arranged to discuss progress among the student and both mentors, including Mathias Unberath. Version control is facilitated via GitHub.

Acknowledgments

Thanks to Philipp Nikutta and Xingtong Liu for their help implementing patch-normalized convolution.

References

[1] A. Malik and L. D. Dorr, "The Science of Minimally Invasive Total Hip Arthroplasty," *Clinical Orthopaedics and Related Research®*, vol. 463, pp. 74–84, Oct. 2007.

[2] R. Grupp, M. Unberath, C. Gao, R. Hegeman, R. Murphy, C. Alexander, Y. Otake, B. McArthur, M. Armand, and R. Taylor, "Automatic Annotation of Hip Anatomy in Fluoroscopy for Robust and Efficient 2D/3D Registration," *arXiv:1911.07042 [cs, eess]*, Nov. 2019.

[3] M. Woerner, E. Sendtner, R. Springorum, B. Craiovan, M. Worlicek, T. Renkawitz, J. Grifka, and M. Weber, "Visual intraoperative estimation of cup and stem position is not reliable in minimally invasive hip arthroplasty," *Acta Orthopaedica*, vol. 87, no. 3, pp. 225–230, May 2016.

[4] M. Unberath, J.-N. Zaech, C. Gao, B. Bier, F. Goldmann, S. C. Lee, J. Fotouhi, R. Taylor, M. Armand, and N. Navab, "Enabling machine learning in X-ray-based procedures via realistic simulation of image formation," *International Journal of Computer Assisted Radiology and Surgery*, vol. 14, no. 9, pp. 1517–1528, Sep. 2019.

[5] B. Bier, M. Unberath, J.-N. Zaech, J. Fotouhi, M. Armand, G. Osgood, N. Navab, and A. Maier, "X-ray-transform Invariant Anatomical Landmark Detection for Pelvic Trauma Surgery," *arXiv:1803.08608 [cs]*, Mar. 2018.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105.

[7] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, ser. Lecture Notes in Computer Science, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. Cham: Springer International Publishing, 2015, pp. 234–241.

[8] B. Bier, F. Goldmann, J.-N. Zaech, J. Fotouhi, R. Hegeman, R. Grupp, M. Armand, G. Osgood, N. Navab, A. Maier, and M. Unberath, "Learning to detect anatomical landmarks of the pelvis in X-rays from arbitrary views," *International Journal of Computer Assisted Radiology and Surgery*, vol. 14, no. 9, pp. 1463–1473, Sep. 2019.

[9] A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi, "A Survey of the Recent Architectures of Deep Convolutional Neural Networks," *arXiv:1901.06032 [cs]*, Feb. 2020.

[10] Y. Wu and K. He, "Group Normalization," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 3–19.

[11] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," *arXiv:1502.03167 [cs]*, Mar. 2015.

[12] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance Normalization: The Missing Ingredient for Fast Stylization," *arXiv:1607.08022 [cs]*, Nov. 2017.

[13] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer Normalization," *arXiv:1607.06450 [cs, stat]*, Jul. 2016.

	February			March				April				May		
Brainstorm Generalization Techniques	✓	✓	✓											
Obtain data access	✓	✓												
Obtain codebase access		✓												
Format Real X-ray Data			✓											
Ongoing Code Documentation			✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
Generate baseline results using U-Net [7]			✓	✓										
Test and Refine Generalization Algorithm				✓	✓	✓								
Perform Ablation Study on MARCC							✓	✓	✓					
Statistical Analysis of Results								✓	✓	✓				
Compile Presentation and Final Report										✓	✓	✓	✓	
Write Academic Publication													✓	→

TABLE IV
SCHEDULE