X-Ray to CT Rigid Registration Using Scene Coordinate Regression

Pragyan Shrestha[1](✉), Chun Xie[1], Hidehiko Shishido[1], Yuichi Yoshii[2], and Itaru Kitahara[1]

[1] University of Tsukuba, Tsukuba, Ibaraki, Japan
shrestha.pragyan@image.iit.tsukuba.ac.jp,
{xiechun,shishid,kitahara}@ccs.tsukuba.ac.jp
[2] Tokyo Medical University, Ami, Ibaraki, Japan
yyoshii@tokyo-med.ac.jp

Abstract. Intraoperative fluoroscopy is a frequently used modality in minimally invasive orthopedic surgeries. Aligning the intraoperatively acquired X-ray image with the preoperatively acquired 3D model of a computed tomography (CT) scan reduces the mental burden on surgeons induced by the overlapping anatomical structures in the acquired images. This paper proposes a fully automatic registration method that is robust to extreme viewpoints and does not require manual annotation of landmark points during training. It is based on a fully convolutional neural network (CNN) that regresses the scene coordinates for a given X-ray image. The scene coordinates are defined as the intersection of the back-projected rays from a pixel toward the 3D model. Training data for a patient-specific model were generated through a realistic simulation of a C-arm device using preoperative CT scans. In contrast, intraoperative registration was achieved by solving the perspective-n-point (PnP) problem with a random sample and consensus (RANSAC) algorithm. Experiments were conducted using a pelvic CT dataset that included several real fluoroscopic (X-ray) images with ground truth annotations. The proposed method achieved an average mean target registration error (mTRE) of 3.79+/1.67 mm in the 50th percentile of the simulated test dataset and projected mTRE of 9.65+/−4.07 mm in the 50th percentile of real fluoroscopic images for pelvis registration. The code is available at https://github.com/Pragyanstha/SCR-Registration.

Keywords: Registration · X-Ray Image · Scene Coordinates

1 Introduction

Image-guided navigation plays a crucial role in modern surgical procedures. In the field of orthopedics, many surgical procedures such as total hip arthro-

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43999-5_74.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14229, pp. 781–790, 2023.
https://doi.org/10.1007/978-3-031-43999-5_74

plasty, total knee arthroplasty, and pedicle screw injections utilize intraoperative fluoroscopy for surgical navigation [2,4,11]. Due to overlapping anatomical structures in X-ray images, it is often difficult to correctly identify and reason the 3D structure from solely the image. Therefore, registering an intraoperatively acquired X-ray image to the preoperatively acquired CT scan is crucial in performing such procedures [13,15,18,19]. The standard procedure for acquiring highly accurate registration involves embedding fiducial markers into the patient and acquiring a preoperative CT scan to obtain 2D-3D correspondences [6,10,17]. Inserting fiducial markers onto the body involves extra surgical costs and might not be viable for minimally invasive surgeries. To circumvent such issues with the feature-based method, an intensity-based optimization scheme for registration has been extensively studied [1,9]. Since the objective function is highly nonlinear for optimizing pose parameters, a good initialization is necessary for the method to converge in a global minimum. Therefore, it is usually accompanied by initial coarse registration using manual alignment of the 3D model to the image, interrupting the surgical flow. On the other hand, learning-based methods have proved to be efficient in solving the registration task. Existing learning-based methods can be broadly categorized into landmark estimation and direct pose regression. Landmark estimation methods aim to solve for pose using correspondences between 3D landmark annotations and its estimated 2D projection points [3,5,7], while methods based on pose regression estimate the global camera pose in a single inference [12]. Pose regressors are known to overfit training data and generalize poorly to unseen images [14]. This makes the landmark estimation methods stand out in terms of registration quality and generalization. However, there exist two main issues with landmark estimation methods: 1) Annotation cost of a sufficiently large number of landmarks in the CT image. 2) Failure to solve for the pose in extreme views where projected landmarks are not visible or the number of visible landmarks is small.

This paper addresses these issues by introducing scene coordinates [16] to establish dense 2D-3D correspondences. Specifically, the proposed method regresses the scene coordinates of the CT-scan model from corresponding X-ray images. A rigid transformation that aligns the CT-scan model to the image is then calculated by solving the Perspective-n-point (PnP) problem with the Random sample and consensus (RANSAC) algorithm.

2 Method

2.1 Problem Formulation

The problem of 2D-3D registration can be formulated a finding the rigid transformation that transforms the 3D model defined in the anatomical or world coordinate system into the camera coordinate system. Specifically, given a CT-scan volume $V_{CT}(x_w)$ where x_w is defined in the world coordinate system, the registration problem is concerned with finding $T_w^c = [R|t]$ such that the following holds.

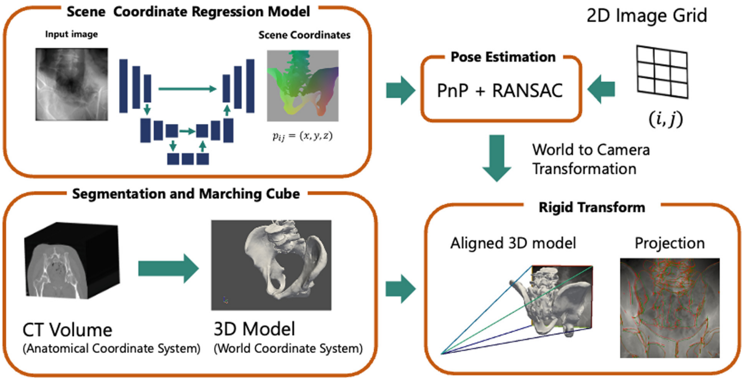

Fig. 1. An overview of the proposed method. Scene coordinates are regressed using a U-Net architecture given an X-ray image. With the obtained dense correspondences, PnP with RANSAC is run to get the transformation matrix that aligns the projection of the 3D model with the X-ray image in the camera coordinate system.

$$I = \Re\{V_{CT}(T_w^{c^{-1}} x_c); K\} \tag{1}$$

where $\Re\{\cdot\}$ is the X-ray transform that can be applied to volumes in the camera coordinate system given an intrinsic matrix K and I, the target X-ray image.

2.2 Registration

The proposed registration pipeline overview is shown in Fig. 1. The proposed method comprises the following four parts: first, the scene coordinates were regressed using a single-view X-ray image as input to the U-Net model; second, the PnP + RANSAC algorithm is used to solve for the pose of the captured X-ray system; third, the CT-scan volume was segmented to obtain a 3D model of the bone regions; and fourth, the computed rigid transformation from world coordinates to camera coordinates is used to generate projection overlay images.

Scene Coordinates. The scene coordinates are defined as the points of intersection between the camera's back-projected rays and the 3D model in a world coordinate system (i.e., only the first intersection and the last intersections are considered). The same concept was adapted for X-ray images and their underlying 3D models obtained from the CT-scans. Specifically, given an arbitrary point x_{ij} in the image plane, the scene coordinates X_{ij} satisfy the following conditions.

$$X_{ij} = [R^T| - t](dK^T x_{ij}) \tag{2}$$

where R and t are the rotation matrix and translation vector that maps points in the world coordinate system to the camera coordinate system, K is the intrinsic matrix, d is the depth, as seen from the camera, of the point X on the 3D model.

Uncertainty Estimation. The task of scene coordinate regression is to estimate these X_{ij} for every pixel ij, given an X-ray image I. However, the existence of X_{ij} is not guaranteed for all pixels because back-projected rays may not intersect the 3D model. One of the many ways to address such a case is to prepare a mask (i.e., 1 if the bone area, 0 otherwise) in advance so that only the pixels that lie inside the mask are estimated. As this approach requires an explicit method for estimating the mask image, an alternative approach was adopted in this study. Instead of estimating a single X_{ij}, the mean and variance of the scene coordinates are estimated. The non-intersecting scene coordinates were identified by applying thresholding to the estimated variance (i.e., points with high variance were considered non-existent scene coordinates and were filtered out). This approach assumes that the observed scene coordinates are corrupted with a zero mean, non-zero and non-constant variance, and isotropic Gaussian noise.

$$X_{ij} \sim N(u(I, x_{ij}), \sigma(I, x_{ij})) \tag{3}$$

where $u(I, x_{ij})$ and $\sigma(I, x_{ij})$ are the functions that produce the mean and standard deviation of the scene coordinates, respectively. This work represents these functions using a fully convolutional neural network.

Loss Function. A U-Net architecture was used to estimate the mean and standard deviation of the scene coordinates at every pixel in a given image. The loss function for the intersecting scene coordinates is derived from the maximum likelihood estimates using the likelihood X_{ij}. This can be expressed as follows:

$$Loss_{intersecting} = (\frac{(X_{ij} - u(I, x_{ij}))}{\sigma(I, x_{ij})})^2 + 2\log(\sigma(I, x_{ij})) \tag{4}$$

Because it is desirable to have a high variance for non-existent scene coordinates, the loss function for non-existent coordinates is designed as follows:

$$Loss_{non-existent} = \frac{1}{\sigma(I, x_{ij})} \tag{5}$$

2D-3D Registration. An iterative PnP implementation from OpenCV was run using RANSAC with maximum iteration of 1000 and reprojection error of 10px and 20px for the simulated and real X-ray images respectively. An example of a successful registration is shown in the left part of Fig. 2 below.

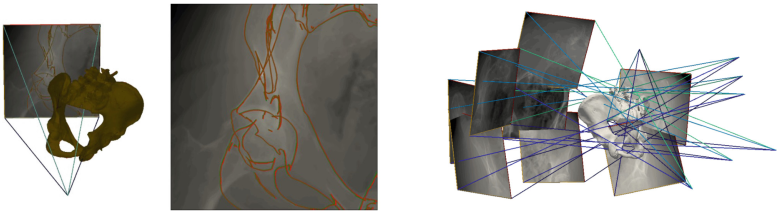

Fig. 2. An example of successful registration with the proposed method (left two images) and Randomly picked data samples in the test set (right). The X-ray image and model's gradient projection overlay (middle) and the model's pose in the camera coordinates system (left). The origin of the view frustum is the X-ray source position and the simulated X-ray images are placed in the detector plane for visualization (right).

3 Experiments and Results

3.1 Dataset

A dataset containing six annotated CT scans, each with several registered real X-ray images from [7] was used to properly evaluate the proposed method. The annotations included 14 landmarks and 7 segmentation labels. The CT scans were of the pelvic bones of cadaveric specimens. Because there were only a few real X-ray images, simulated X-ray images were generated from each CT-scans to train and test the model. In particular, DeepDRR [3] was used to simulate a Siemens Cios Fusion Mobile C-arm imaging device. Similar to [3], the left anterior oblique/right anterior oblique (LAO/RAO) views were sampled at angles of [45, 45] degrees with 1-degree intervals. A random offset was applied in each direction. The offset vector was sampled from a normal distribution with a mean of zero and standard deviations of 90 mm in the lateral direction, and 30 mm in the other two directions. Images intentionally included partially visible structures. A selection of randomly sampled images is displayed on the right side of Fig. 2. Ground-truth scene coordinates for each image were obtained by rendering the depth map of the 3D models and converting them to 3D coordinates using Eq. 2. In total, 8100 simulated X-rays were generated for each specimen. Among these, 5184 images were randomly assigned to the training set, 1296 to the validation set, and the remaining 1620 to the test set.

3.2 Implementation Details

The input image and output scene coordinates had a size of 512×512 pixels. The U-Net model had eight output channels, which consisted of three channels for scene coordinates and one channel for standard deviation, multiplied by two for the entry and exit points. Patient-specific models were trained individually for each dataset. Adam optimizer with a constant learning rate of 0.0001 and a batch size of 16 was used. Online data augmentation which includes random inversion, color jitter, and random erasing was applied. The scene coordinates were filtered using a log variance threshold of 0.0 for simulated images and −2.0 for real X-ray images.

3.3 Baselines and Evaluation Metrics

The proposed method was compared with two other baseline methods: PoseNet [8] and DFLNet [7]. PoseNet was implemented using ResNet-50 as the backbone for the feature extractor and was trained using geometric loss. DFLNet uses the same architecture as the proposed method however the last layer regresses 14 heatmaps of the landmarks instead of scene coordinates. Note that the original study's segmentation layer and the gradient-based optimization phase were omitted for architectural comparison. Each baseline was trained in a patient-specific manner following the proposed method. The mean target registration error (mTRE) and Gross Failure Rate (GFR), were used as the evaluation metrics for comparison with the baselines. mTRE is defined in 6, where X_k is the position of the ground truth landmark $\hat{X}_k$ after applying the predicted transformation. The GFR is the ratio of failed cases, defined as the registration results with an mTRE greater than 10 mm. Because we could only obtain the projection of the ground truth landmarks and not the ground truth transformation matrix for the real X-ray images, the projected mTRE (proj. mTRE) was used for the evaluation. It is similar to mTRE except the X_k and $\hat{X}_k$ represent the projected coordinates of the landmarks, in the detector plane (i.e., the pixel coordinates are scaled according to the detector size to match the units).

$$\text{mTRE} = \frac{1}{N}\sum_{k=1}^{k=N} \|X_k - \hat{X}_k\|_2 \tag{6}$$

3.4 Registration Results

Simulated X-Ray Images. Table 1 shows the mTRE for the 25th, 50th, and 95th percentiles of the total test sample size and the GFR. The proposed method could retain a GFR below 20% for most of the specimens, whereas PoseNet and DFLNet failed to register with more than 20% GFR in most cases. This is because the network in PoseNet cannot reason about the spatial structure or its local relation to the image patches. For DFLNet, this is inevitable because of the visibility issue of the landmark points, mostly located in the pubic region of the pelvis. Comparing the mTRE of each specimen with that of each method, the proposed method achieved an mTRE of 7.98 mm even in the 95th percentile of Specimen 2. DFLNet achieved the lowest mTRE of 0.98 mm in the 25th percentile of Specimen 4. This illustrates the highly accurate registration of landmark estimation methods. However, with extreme or partial views such as the one shown in Fig. 3, the method cannot estimate the correct pose parameter because of incorrect landmark localization or insufficiently visible landmarks. Please refer to the supplemental material for the registration overlay results of the different specimens using the proposed method.

Real X-Ray Images. Table 2 lists the mTRE values calculated for the projected image points (abbreviated as proj. mTRE) for PoseNet and the proposed method, respectively. DFLNet did not adapt to real X-ray images, therefore, it was omitted from the table. Because our dataset consisted mostly of images with partially visible hips, only a few landmarks were visible in each image. This causes the DFLNet to overfit to the partially visible landmark distribution, whereas our proposed model mitigates this issue by learning the general structure (i.e., every surface point that is visible). The proposed method estimates good transformations (that is proj. mTRE approximately 10 mm in the 50th percentile). In contrast, the proj. mTRE for PoseNet is significantly higher. This suggests that PoseNet overfitted the training data despite applying domain randomization. This result agrees with previous reports [14] addressing this issue. A visualization of the overlays is presented in the Supplemental Material.

Table 1. The mean target registration errors each in 25th, 50th and 95th percentile of the simulated test dataset. All models are trained individually on the 6 specimens shown below. The proposed method outperforms other methods regarding 50th percentile mTRE and GFR in most specimens.

Specimen	PoseNet				DFLNet				Ours			
	mTRE[mm]↓			GFR[%]↓	mTRE[mm]↓			GFR[%]↓	mTRE[mm]↓			GFR[%]↓
	25th	50th	95th		25th	50th	95th		25th	50th	95th	
#1	5.75	8.37	24.81	38.53	3.36	212.59	680.58	62.27	1.37	**2.50**	9.80	**4.87**
#2	6.97	10.23	25.95	51.63	1.98	7.04	656.41	46.15	1.15	**2.14**	7.98	**2.54**
#3	5.42	7.86	23.34	35.35	1.03	**2.51**	583.63	28.65	1.67	3.05	12.25	**8.56**
#4	4.67	6.46	16.91	18.77	0.98	**2.30**	558.20	23.59	1.76	3.38	19.19	**12.54**
#5	4.81	6.52	18.98	22.43	1.51	**4.28**	767.56	37.47	3.09	5.30	17.32	**18.85**
#6	4.06	**5.85**	18.42	**22.69**	2.26	139.96	15321.19	58.72	3.80	6.37	18.25	23.18
mean	5.28	7.55	21.40	31.57	1.85	61.45	3094.60	42.81	2.14	**3.79**	14.13	**11.76**
std	1.02	1.62	3.77	12.58	0.90	91.88	5990.24	15.76	1.06	1.67	4.75	8.05

Table 2. The projected mean target registration errors for real X-ray images. The proposed method achieved significantly low registration errors compared to PoseNet, implying that it generalizes well to unseen data and domains.

Specimen	Number of Images	PoseNet			Ours		
		proj. mTRE [mm]↓			proj. mTRE [mm]↓		
		25th	50th	95th	25th	50th	95th
#1	111	43.64	49.11	64.23	5.45	**8.02**	55.87
#2	24	19.42	27.18	43.68	2.74	**3.32**	6.48
#3	104	31.18	38.97	66.06	7.60	**11.85**	162.83
#4	24	35.52	38.37	57.07	11.12	**15.52**	92.34
#5	48	38.97	46.60	69.06	6.09	**9.07**	21.91
#6	55	34.51	37.13	47.72	7.18	**10.14**	20.78
mean		33.87	39.56	57.97	6.70	**9.65**	60.04
std		8.25	7.77	10.37	2.77	4.07	59.15

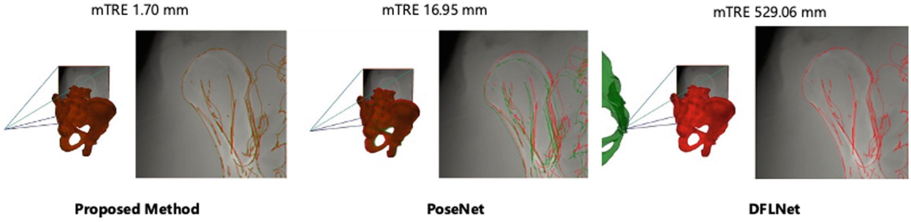

Fig. 3. An example case illustrating an extreme partial viewpoint. The proposed method successfully registers the image with 1.70 mm mTRE, while PoseNet struggles with 16.95 mm mTRE. Since there is an insufficient number (less than 4) of visible landmarks, the DFLNet hallucinates landmarks providing incorrect 2D-3D correspondences, leading to a large mTRE.

4 Limitations

As the proposed method was designed to give initial estimates of the pose parameters, a further refinement step using an intensity-based optimization method would be required to obtain clinically relevant registration accuracy. Although the proposed method provided a good initial estimate, the average runtime for the entire pipeline was 1.75 s which was approximately two orders of magnitude greater than that of PoseNet, which had an average runtime of 0.06 s. This is because RANSAC must determine a good pose from a dense set of correspondences. This issue can be addressed by heuristically selecting a good variance threshold per image that filters out bad correspondences.

5 Conclusion

This paper presented a scene coordinate regression-based approach for the X-ray to CT-scan model registration problem. Experiments with simulated and real X-ray images showed that the proposed method performed well even under partially visible structures and extreme view angles, compared with direct pose regression and landmark estimation methods. Testing the model trained solely on simulated X-ray images, on real X-ray images did not result in catastrophic failure. Instead, the results were positive for instantiating further refinement steps.

Acknowledgement. This work was partially supported by a grant from JSPS KAKENHI grant number JP23K08618. This study (in part) used the computational resources for Cygnus provided by the Multidisciplinary Cooperative Research Program at the Center for Computational Sciences, University of Tsukuba, Japan.

References

1. Aouadi, S., Sarry, L.: Accurate and precise 2D–3D registration based on X-ray intensity. Comput. Vis. Image Underst. **110**(1), 134–151 (2008)
2. Belei, P., et al.: Fluoroscopic navigation system for hip surface replacement. Comput. Aided Surg. **12**(3), 160–167 (2007)
3. Bier, B., et al.: X-ray-transform invariant anatomical landmark detection for pelvic trauma surgery (2018)
4. Bradley, M.P., Benson, J.R., Muir, J.M.: Accuracy of acetabular component positioning using computer-assisted navigation in direct anterior total hip arthroplasty. Cureus **11**(4), e4478 (2019)
5. Esteban, J., Grimm, M., Unberath, M., Zahnd, G., Navab, N.: Towards fully automatic X-ray to CT registration. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11769, pp. 631–639. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32226-7_70
6. George, A.K., Sonmez, M., Lederman, R.J., Faranesh, A.Z.: Robust automatic rigid registration of MRI and X-ray using external fiducial markers for XFM-guided interventional procedures. Med. Phys. **38**(1), 125–141 (2011)
7. Grupp, R.B., et al.: Automatic annotation of hip anatomy in fluoroscopy for robust and efficient 2D/3D registration. Int. J. Comput. Assist. Radiol. Surg. **15**(5), 759–769 (2020)
8. Kendall, A., Grimes, M., Cipolla, R.: PoseNet: a convolutional network for real-time 6-DOF camera relocalization. In: 2015 IEEE International Conference on Computer Vision (ICCV), pp. 2938–2946 (2015)
9. Livyatan, H., Yaniv, Z., Joskowicz, L.: Gradient-based 2-D/3-D rigid registration of fluoroscopic X-ray to CT. IEEE Trans. Med. Imaging **22**(11), 1395–1406 (2003)
10. Maurer, C.R., Jr., Fitzpatrick, J.M., Wang, M.Y., Galloway, R.L., Jr., Maciunas, R.J., Allen, G.S.: Registration of head volume images using implantable fiducial markers. IEEE Trans. Med. Imaging **16**(4), 447–462 (1997)
11. Merloz, P., et al.: Fluoroscopy-based navigation system in spine surgery. Proc. Inst. Mech. Eng. H **221**(7), 813–820 (2007)
12. Miao, S., Jane Wang, Z., Liao, R.: Real-time 2D/3D registration via CNN regression (2015)
13. Reichert, J.C., Hofer, A., Matziolis, G., Wassilew, G.I.: Intraoperative fluoroscopy allows the reliable assessment of deformity correction during periacetabular osteotomy. J. Clin. Med. Res. **11**(16) (2022)
14. Sattler, T., Zhou, Q., Pollefeys, M., Leal-Taixé, L.: Understanding the limitations of CNN-Based absolute camera pose regression. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3297–3307 (2019)
15. Selles, C.A., Beerekamp, M.S.H., Leenhouts, P.A., Segers, M.J.M., Goslings, J.C., Schep, N.W.L.: EF3X study group: the value of intraoperative 3-dimensional fluoroscopy in the treatment of distal radius fractures: a randomized clinical trial. J. Hand Surg. Am. **45**(3), 189–195 (2020)
16. Shotton, J., Glocker, B., Zach, C., Izadi, S., Criminisi, A., Fitzgibbon, A.: Scene coordinate regression forests for camera relocalization in RGB-D images. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2930–2937 (2013)
17. Tang, T.S.Y., Ellis, R.E., Fichtinger, G.: Fiducial registration from a single X-ray image: a new technique for fluoroscopic guidance and radiotherapy. In: Delp, S.L., DiGoia, A.M., Jaramaz, B. (eds.) MICCAI 2000. LNCS, vol. 1935, pp. 502–511. Springer, Heidelberg (2000). https://doi.org/10.1007/978-3-540-40899-4_51

18. Woerner, M., et al.: Visual intraoperative estimation of cup and stem position is not reliable in minimally invasive hip arthroplasty. Acta Orthop. **87**(3), 225–230 (2016)
19. Wylie, J.D., Ross, J.A., Erickson, J.A., Anderson, M.B., Peters, C.L.: Operative fluoroscopic correction is reliable and correlates with postoperative radiographic correction in periacetabular osteotomy. Clin. Orthop. Relat. Res. **475**(4), 1100–1106 (2017)