2D/3D registration with a Statistical deformation model prior using deep learning

Jeroen Van Houtte
imec-Vision Lab
University of Antwerp
Antwerp, Belgium
jeroen.vanhoutte@uantwerp.be

Xiaoru Gao
Institute of Medical Robotics
Shanghai Jiao Tong University
Shanghai, China
xiaoru.gao@sjtu.edu.cn

Jan Sijbers
imec-Vision Lab
University of Antwerp
Antwerp, Belgium
jan.sijbers@uantwerp.be

Guoyan Zheng
Institute of Medical Robotics
Shanghai Jiao Tong University
Shanghai, China
guoyan.zheng@ieee.org

***Abstract*—Deep learning-based (DL) solutions are increasingly been adopted for 2D/3D registration as they can achieve faster 3D reconstructions from 2D radiographs compared to classical methods. This study proposes a novel semi-supervised DL-network for 2D-3D registration, in which an atlas is registered to two orthogonal radiographs. The deformation of the atlas is composed of an affine transformation and a local deformation constrained by a B-spline-based statistical deformation model. The validaton of the network on digitally reconstructed radiographs from 22 femur CT images shows that the atlas can accurately be registered.**

***Index Terms*—2D/3D registration, deep-learning, digitally reconstructed radiographs, statistical deformation model**

I. Introduction

The three-dimensional (3D) reconstruction of bones from two-dimensional (2D) radiographs is crucial in many biomedical engineering domains, such as kinematical studies, preoperative planning, implant design and post-operative evaluations [1], [2]. The reconstruction is known to be a degenerate, ill-posed problem, because of the limited number of projections. To resolve the ambiguity of the reconstruction, classical methods tackle the problem as a registration of a 3D atlas to the 2D projections. Statistical models have frequently been adopted to constrain the possible local deformations in a physical way [3], [4].

Much research in 2D/3D registration has recently turned to deep learning (DL) solutions to achieve real-time 3D reconstructions [5], being essential for intraoperative guidance and robotic-assisted surgeries [6], [7]. In contrast to the classical methods, current DL approaches often do not register an atlas to the radiographs, but directly decode the 3D image values from the encoded 2D image, ignoring the fundamental degenaracy of the problem.

Being composed of one or two 2D image encoders and a 3D decoder, these networks require a method to bridge between the different dimensionalities of the feature maps [8]–[10], which lacks any connection with the actual physical image generation process. Also, the combination of different projection directions is not physically well founded. The 3D registration to biplanar radiographs is often, by construction, limited to orthogonal radiographs [9], [10], because of the way in which both directional feature maps are combined.

Cone-beam projections from 3D volumes can actually be simulated by integrating the attenuation along a ray throughout the volume and has previously been integrated in neural networks [11], [12]. Gao et al. generalised the concept of spatial transformers to perspective projections, providing a much simpler and computational efficient way to simulate cone-beam projections [13].

In this paper, we propose a semi-supervised end-to-end neural network, which differs from the encoder-decoder architectures proposed in the current literature. Instead, our network estimates a registration field like 3D/3D registration networks [14]. The registration field, being learned from a 3D atlas image and two radiographs, warps the atlas image such that the forward projection of it matches the input radiographs. The deformation field is fully parameterised by an affine transformation and a B-spline-based statistical deformation model (SDM). To the authors' knowledge, this is the first study to present a DL-approach for 2D/3D registration with a B-spline-based SDM.

II. Methodology

A. B-spline-based statistical deformation model

The B-spline-based statistical deformation model is constructed from N_s training CT images, which were registered beforehand to an atlas image $V \in \mathbb{R}^{V_x \times V_y \times V_z}$ by a B-spline-based free-form deformation (FFD). The B-spline coefficients $\boldsymbol{C}$ are defined on a coarse regular lattice of B-spline control points with size $(L+3)\times(M+3)\times(N+3)$. The displacement field that brings the atlas into alignment with each training volume, is expressed as the 3D B-spline tensor product of 1D cubic B-spline coefficients $\boldsymbol{C}$:

$$T^{FFD}(\boldsymbol{C}) = \sum_{r=0}^{3}\sum_{s=0}^{3}\sum_{t=0}^{3} B_r(u)B_s(v)B_t(w)\boldsymbol{C}_{l+r,m+s,n+t}. \tag{1}$$

where B_i are B-spline basis functions. The indexes $-1 \leq l \leq (L+1)$, $-1 \leq m \leq (M+1)$, $-1 \leq n \leq (N+1)$ are the indexes of the grid control points, while u, v, w are the relative positions of the image space coordinate in the lattice. As the size of the control point lattice is much smaller than the size

*The work was partially supported by the Science and Technology Commission of Shanghai via project 20511105205 and the Research Foundation in Flanders (FWO-SB 1S63918N).

of the atlas image, the B-spline FFD gains speed compared to a regular FFD.

The B-spline-based SDM is computed as the singular value decomposition on the set of N_s B-spline coefficients. The SDM expresses any feasible B-spline coefficient vector as a linear combination of the eigenvectors $\boldsymbol{p_k}$ of the decomposition:

$$\boldsymbol{C}(\{\alpha_k\}) = \bar{\boldsymbol{C}} + \sum_{k=1}^{N_m} \alpha_k \sigma_k \boldsymbol{p_k}, \tag{2}$$

where σ_k are the associated singular values to the eigenvectors $\boldsymbol{p_k}$. The vector $\bar{\boldsymbol{C}}$ is the average B-spline coefficient vector. The weights $\{\alpha_k\}$ will act as the model parameters and $N_m \leq N_s - 1$ is the number of selected modes in the model. Each instance $\boldsymbol{C}(\{\alpha_k\})$ determines a forward FFD from the atlas to a floating image by (1).

B. Pseudo-inversion

The forward FFD of (1) can be used to warp a floating image backwards to the atlas image domain. For 2D/3D registration, however, this 3D floating image is unknown, and one needs to warp the atlas image backwards by the inverse of (1), which is computationally expensive to compute in case of a regular FFD. We therefore apply the pseudo-inversion algorithm on the B-spline coefficients themselves [4]. First, the forward displacement corresponding to the coefficients $\boldsymbol{C}$ is calculated by (1). Next, a fixed-point based inversion calculates the inverted displacement on only the control points [15]. Finally, the backward B-spline coefficients $\boldsymbol{C^{bck}}$ on the control points are recursively determined [16]. The 3D B-spline tensor product of (1) applied on those backward B-spline coefficients $T^{FFD}(\boldsymbol{C^{bck}})$ yields the displacement field that can warp the atlas to the floating image. Fore more details we refer the reader to [4].

C. Projective spatial transform

We use the projective spatial transformer (ProST) from [13] to simulate a 2D perspective projection image $\hat{I} \in \mathbb{R}^{S_x \times S_y}$ from a 3D volume $\hat{V}$. This method defines a fixed canonical grid $G \in \mathbb{R}^{S_x \times S_y \times K}$ of K sampling points, uniformly distributed along each ray connecting the source and each pixel of the 2D detector plane. Given a particular projection geometry, this canonical grid can be transformed by an affine transformation T_{geom} in order to represent the actual projection geometry. The 3D image volume can be interpolated at the transformed grid positions $T_{geom}(G)$. The cone-beam projection is then obtained by integrating along each ray, which is equivalent to a "parallel projection" of the interpolated volume:

$$\hat{I}^{(i,j)} = \sum_{k=1}^{K} (\hat{V} \circ (T_{geom}(G)))^{(i,j,k)}. \tag{3}$$

In contrast to [13], we define two fixed projection angles corresponding to lateral (LAT) and anterior-posterior (AP) projections. Instead of rotating the projection geometry, we apply the affine transformations in the image domain itself to solve the pose problem.

D. Registration network architecture

The registration network estimates a registration field Ψ that maps the atlas image V (with associated label map S) to the moving image space, such that the forward projection of the warped atlas, $V \circ \Psi$, matches the input DRR's I_i, with $i \in \{\text{AP}, \text{LAT}\}$. The registration field Ψ can be decomposed into an affine transformation T and a local deformation ϕ, which is constrained by the SDM. Both components are fully parameterised by respectively 7 affine parameters (rotation, translation and isotropic scaling) and N_m PC weights $\{\alpha_k\}$ of the SDM. The two sets of parameters are separately regressed by two sequential networks, depicted in Figure 1.

First a U-net with skip-connections, similar to [14], learns a 3D volumetric feature map $\hat{V} \in \mathbb{R}^{V_x \times V_y \times V_z \times N_f}$ from the 3D atlas image V, with $N_f = 16$ the number of features. The U-net consists of 4 encoder layers and 6 decoder layers with skip connections in between.

The resulting 3D feature map is projected by a ProST layer, along the AP and lateral direction. Note that the projected feature maps $\hat{I}_i$ still have the same number of features as the volumetric feature map $\hat{V}$. The input DRR's I_i are first convolved such that it has also the same number of features. The ProST output $\hat{I}_i$ and the convolution of the input DRR I_i are concatenated into a 2-channel 2D image and fed to a 2D encoder. Each projection direction i has its own encoder. Each of the four encoder levels consists of a strided convolution, a batch-normalisation layer and a Leaky-Relu activation. Each level reduces the spatial size of the feature map by a factor two and doubles the number of features. At each encoder level, the AP and lateral features (and the preceding combined features) are concatenated and convolved.

The accumulated 2D feature map at the last encoder level is flattened and fed to a dense layer which regresses the seven parameters of the affine transformation T between the floating image and the atlas. The bias and kernel weights of the dense layer are initialised by respectively zero and a narrow normal distribution, such that the initial affine transformation during training is close to identity.

The 3D feature map $\hat{V}$ is warped by the affine transformation T by a spatial transform layer [17]. The transformed 3D features are fed into a similar network as before in order to regress the N_m PC weights $\boldsymbol{\alpha}$, which determine the B-spline coefficients $\boldsymbol{C}$ through (2). The pseudo-inversion on $\boldsymbol{C}$ yields the backward B-spline coefficients which determine the backward B-spline-based deformation field ϕ through (1). The composition of the affine transformation T and the backward B-spline-based deformation ϕ is given by: $T \oplus \phi = T + \phi \circ T$.

E. Network loss function

The network loss-function, used to evaluate the registration quality during training, consists of a normalised cross-correlation (NCC) between the warped atlas and the ground-truth CT-image V_{gt}, and a Dice loss between the warped atlas label map and the ground-truth label map S_{gt}. Both metrics are

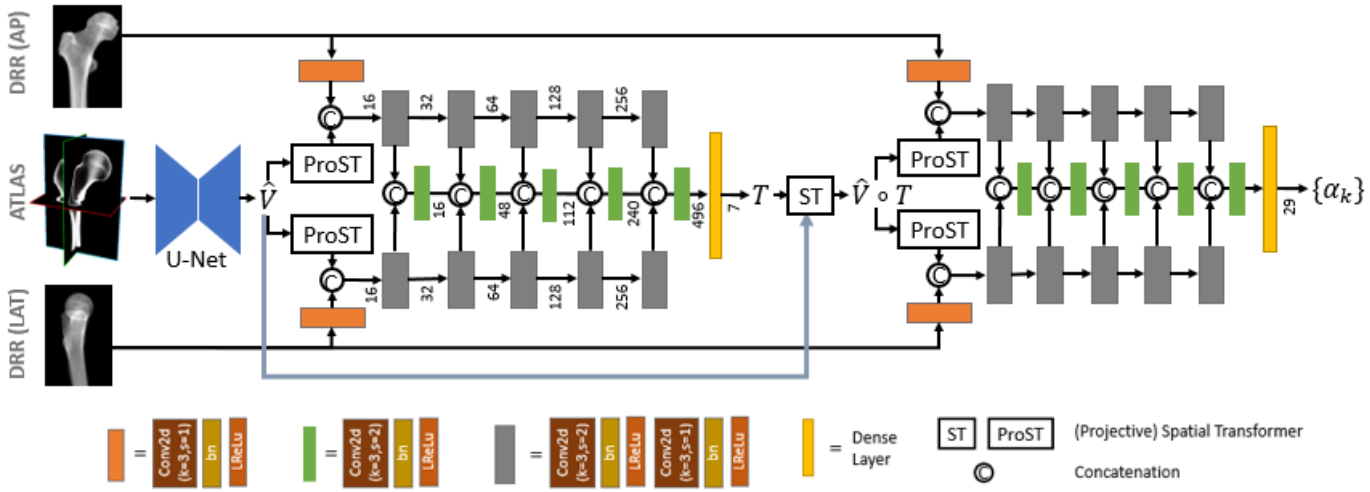

Fig. 1. Architecture of the end-to-end 2D/3D registration network. The network takes as input two 2D DRR's and a 3D atlas and estimates a deformation field which is parameterised by 7 affine parameters and 29 PC weights $\{\alpha_k\}$. Both parameter sets are separately regressed by two identical networks. For the first network we indicate the number of features at each level, which are identical for the second network.

evaluated after the affine registration and after the B-spline-based deformation:

$$\mathcal{L} = \gamma(NCC(V_{gt}, V \circ T) + NCC(V_{gt}, V \circ (T \oplus \phi))) \quad (4)$$
$$+ \delta(Dice(S_{gt}, S \circ T) + Dice(S_{gt}, S \circ (T \oplus \phi))) \quad (5)$$
$$+ \zeta \sum \alpha_k^2, \quad (6)$$

with $\gamma = 1.0$, $\delta = 0.1$ and $\zeta = 10^{-3}$ weighting factors to balance the different loss terms. The last term is the Mahalanobis distance and acts as a regularisation on the PC weights. It favors instances $\boldsymbol{C}$ of the SDM that are close to the average $\bar{\boldsymbol{C}}$.

III. Experiment

A. Dataset

The training dataset consists of 40 CT-images of naked cadaver femur bones. The validation dataset was acquired separately on different patients and constitutes of 22 CT images, from which the femur bone was masked. The SDM was built on the training dataset. Based on the compactness of the SDM, we have selected the first $N_m = 29$ variation modes from the SDM as they account for up to 99 % of the shape variability in the training data set. The other modes are regarded as noise and discarded from the set of degrees of freedom optimised by the registration network.

The training and validation datasets were augmented off-line by applying random affine transformations on the 3D CT-data, resulting in 1200 and 330 images respectively. From the transformed CT volumes near AP and lateral digitally reconstructed radiographs (DRR) were simulated with DeepDRR software [11]. The AP and lateral orientations of the femur were defined based on the femoral shaft and neck axis. Pose variations around the perfect AP/lateral view were allowed within a range of 30° internal/external rotation and within a range of 10° extension/flexion and abduction/adduction.

TABLE I
Average Validation Metrics

	Dice	ASSD (mm)
Initial	0.515 ± 0.083	8.48 ± 1.49
Affine	0.855 ± 0.038	2.16 ± 0.54
Affine + SDM	0.908 ± 0.018	1.29 ± 0.21

The volume size and voxel spacing of the CT volumes and of the atlas equal $(192 \times 128 \times 192)$ and $(0.66 \times 0.66 \times 1)\,\text{mm}^3$ respectively. The size and pixel spacing of the DRRs equal (141×213) and $(0.9 \times 0.9)\,\text{mm}^2$ respectively.

B. Results

The entire model, including the pseudo-inversion of the B-spline coefficients and the ProST layer, was implemented in Tensorflow. The network was trained by Adam optimizer for 50 epochs with a learning rate of 10^{-5}, on a NVIDIA Tesla V100 GPU.

The trained model was evaluated on the validation dataset in terms of the Dice metric and the average signed surface distance (ASSD). The average metric values are tabulated in Table I. Figure 2 shows two examples of the 2D/3D registration. Ground-truth and estimated surface models were created from the ground-truth label map and the warped atlas label map respectively. The unsigned distance between those surface models highlight the anatomical features, like the greater and lesser trochanter, as challenging parts to register accurately.

IV. Discussion

This study presents an end-to-end DL-approach to 2D/3D registration, which differs from the typical encoder-decoder network architectures [5]. Instead of directly decoding the intensity values of a 3D volume without guarantees on feasibility

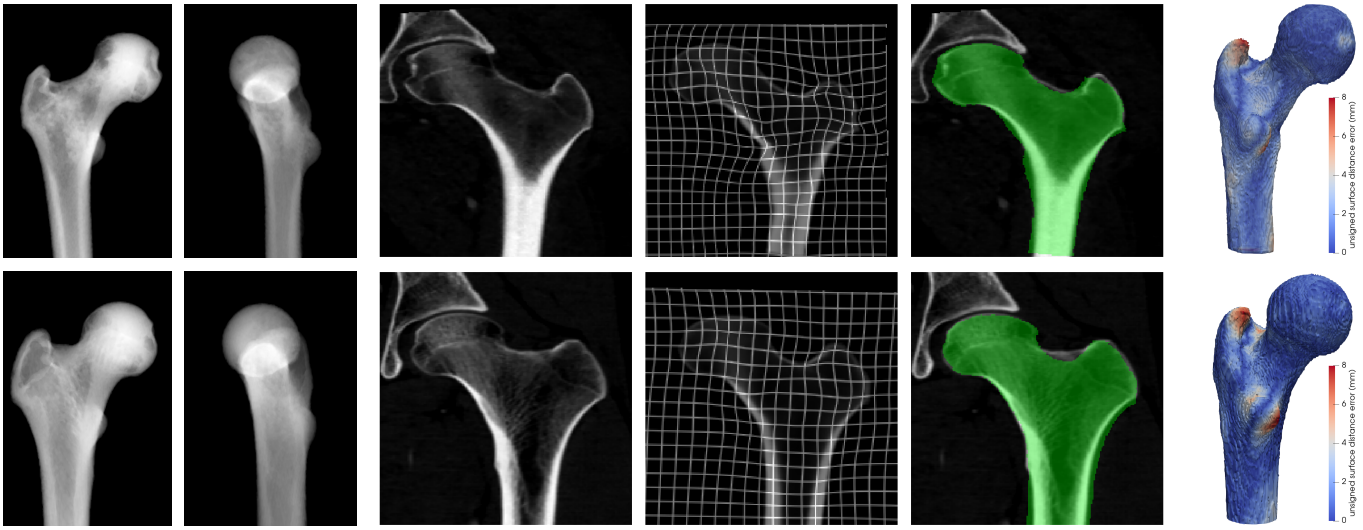

Fig. 2. Registration of the 3D atlas to orthogonal pairs of DRR's (left columns). The third to fifth column show the same coronal slice of the ground-truth CT-volume, the warped atlas volume together with the deformed grid and the warped label map on top of the ground-truth CT-volume. The last column shows the surface model generated from the deformed atlas segmentation map with the unsigned surface distance error represented by the color map.

and smoothness of the reconstruction, this model estimates a deformation field that warps the atlas image.

Although we used lateral and AP radiographs in this study, the network is not limited to this particular combination of projections, nor to orthogonal projections. The network can be trained for any combination of projection geometries, as long as the calibration is known beforehand. In the future we will investigate how re-training the network for each different projection geometry can be avoided.

The network as presented in this study is semi-supervised. The training of the network relies on the auxiliary ground-truth volumetric CT-volume and label map associated to the DRR. This type of data is not always available however. Future research could address unsupervised learning schemes for such cases.

As the network is trained on DRR's, the model might not generalise well yet to real experimental radiographs. This will be tackled in future work by augmenting the DRR appearance during training or by including style transfer prior to the network.

REFERENCES

[1] B. Postolka, R. List, B. Thelen, P. Schütz, W. R. Taylor, and G. Zheng, "Evaluation of an intensity-based algorithm for 2D/3D registration of natural knee videofluoroscopy data," *Med. Eng. Phys.*, vol. 77, pp. 107–113, 2020.

[2] G. Zheng, "Statistically deformable 2D/3D registration for estimating post-operative cup orientation from a single standard AP X-ray radiograph," *Ann. Biomed. Eng.*, vol. 38, no. 9, pp. 2910–2927, 2010.

[3] C. J. F. Reyneke, M. Lüthi, V. Burdin, T. S. Douglas, T. Vetter, and T. E. Mutsvangwa, "Review of 2-D/3-D reconstruction using statistical shape and intensity models and X-ray image synthesis: toward a unified framework," *IEEE Rev. Biomed. Eng.*, vol. 12, pp. 269–286, 2018.

[4] W. Yu, M. Tannast, and G. Zheng, "Non-rigid free-form 2D–3D registration using a B-spline-based statistical deformation model," *Pattern recognition*, vol. 63, pp. 689–699, 2017.

[5] A. Yuniarti and N. Suciati, "A review of deep learning techniques for 3D reconstruction of 2D images," in *2019 12th Int. Conf. Inf. Commun. Technol. Syst. (ICTS)*. IEEE, 2019, pp. 327–331.

[6] C. Gao, R. B. Grupp, M. Unberath, R. H. Taylor, and M. Armand, "Fiducial-free 2D/3D registration of the proximal femur for robot-assisted femoroplasty," in *Medical Imaging 2020: Image-Guided Procedures, Robotic Interventions, and Modeling*, vol. 11315. International Society for Optics and Photonics, 2020, p. 113151C.

[7] P. L. Sousa, P. K. Sculco, D. J. Mayman, S. A. Jerabek, M. P. Ast, and B. P. Chalmers, "Robots in the operating room during hip and knee arthroplasty," *Curr. Rev. Musculoskelet. Med.*, vol. 13, no. 3, pp. 309–317, 2020.

[8] P. Henzler, V. Rasche, T. Ropinski, and T. Ritschel, "Single-image tomography: 3D volumes from 2D cranial X-rays," in *Computer Graphics Forum*, vol. 37, no. 2. Wiley Online Library, 2018, pp. 377–388.

[9] Y. Kasten, D. Doktofsky, and I. Kovler, "End-to-end convolutional neural network for 3D reconstruction of knee bones from bi-planar X-ray images," in *International Workshop on Machine Learning for Medical Image Reconstruction*. Springer, 2020, pp. 123–133.

[10] X. Ying, H. Guo, K. Ma, J. Wu, Z. Weng, and Y. Zheng, "X2CT-GAN: reconstructing CT from biplanar X-rays with generative adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 619–10 628.

[11] M. Unberath, J.-N. Zaech, S. C. Lee, B. Bier, J. Fotouhi, M. Armand, and N. Navab, "DeepDRR–a catalyst for machine learning in fluoroscopy-guided procedures," in *Int. Conf. Med. Im. Comp. Comp.-Assist. Interv. (MICCAI)*. Springer, 2018, pp. 98–106.

[12] C. Syben, M. Michen, B. Stimpel, S. Seitz, S. Ploner, and A. K. Maier, "PYRO-NN: Python reconstruction operators in neural networks," *Medical physics*, vol. 46, no. 11, pp. 5110–5115, 2019.

[13] C. Gao, X. Liu, W. Gu, B. Killeen, M. Armand, R. Taylor, and M. Unberath, "Generalizing spatial transformers to projective geometry with applications to 2D/3D registration," in *Int. Conf. Med. Im. Comp. Comp.-Assist. Interv. (MICCAI)*. Springer, 2020, pp. 329–339.

[14] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, and A. V. Dalca, "Voxelmorph: a learning framework for deformable medical image registration," *IEEE Trans. Med. Imag.*, vol. 38, no. 8, pp. 1788–1800, 2019.

[15] M. Chen, W. Lu, Q. Chen, K. J. Ruchala, and G. H. Olivera, "A simple fixed-point approach to invert a deformation field," *Medical physics*, vol. 35, no. 1, pp. 81–88, 2008.

[16] A. Tristán and J. I. Arribas, "A fast B-spline pseudo-inversion algorithm for consistent image registration," in *International Conference on Computer Analysis of Images and Patterns*. Springer, 2007, pp. 768–775.

[17] A. V. Dalca, G. Balakrishnan, J. Guttag, and M. R. Sabuncu, "Unsupervised learning for fast probabilistic diffeomorphic registration," in *Int. Conf. Med. Im. Comp. Comp.-Assist. Interv. (MICCAI)*. Springer, 2018, pp. 729–738.