UNIVERSITÄT AUGSBURG
Universitätsbibliothek

OPEN ACCESS
OPUS AUGSBURG

RGB-D/CT registration for preemptive patient pose alignment in longitudinal CT scans

Elisabeth Schiele, Melina Wördehoff, Josua A. Decker, Lukas Förner, Thomas Wendler

Angaben zur Veröffentlichung / Publication details:

Schiele, Elisabeth, Melina Wördehoff, Josua A. Decker, Lukas Förner, and Thomas Wendler. 2026. "RGB-D/CT registration for preemptive patient pose alignment in longitudinal CT scans." *Current Directions in Biomedical Engineering* 12 (1): 460–63. https://doi.org/10.1515/cdbme-2026-0217.

Nutzungsbedingungen / Terms of use:

CC BY 4.0

Dieses Dokument wird unter folgenden Bedingungen zur Verfügung gestellt: / This document is made available under these conditions:
CC-BY 4.0: Creative Commons: Namensnennung
Weitere Informationen finden Sie unter: / For more information see:
https://creativecommons.org/licenses/by/4.0/deed.de

Elisabeth Schiele*, Melina Wördehoff, Josua A. Decker, Lukas Förner, and Thomas Wendler

RGB-D/CT Registration for Preemptive Patient Pose Alignment in Longitudinal CT Scans

https://doi.org/10.1515/cdbme-2026-0217

Abstract: Registration of longitudinal Computed Tomography (CT) scans is critical for monitoring treatment responses, yet its precision is often compromised by inconsistencies in patient positioning between sessions. Historically, the medical imaging community has relied purely on post-hoc ("post-scan") image registration to correct these inconsistencies [1]. Whether utilizing rigid or deformable models, post-hoc registration is a reactive tool applied to data that is inherently flawed at acquisition. This work proposes a paradigm shift: addressing variability at its source through pre-scan patient pose alignment. By capturing the patient's current pose with an integrated RGB-D camera prior to the follow-up CT scan, discrepancies relative to a baseline CT scan can be registered and visualized. This paper outlines an articulated registration pipeline designed to overcome the domain gap between occluded RGB-D surface data and volumetric CT data. The torso serves as a rigid anchor to align the modalities, after which individual limbs undergo local point-to-plane Iterative Closest Point refinement. The spatial corrections are then projected onto the original 2D RGB image, providing intuitive visual guidance for clinical staff. Validated on an anthropomorphic phantom, our pipeline demonstrates highly stable torso anchoring (average centroid error over 10 iterations: 2.69 cm) and consistent limb alignment (mean surface error: 1.61–1.84 cm). This proof-of-concept establishes a foundation for proactive patient alignment in longitudinal imaging.

Keywords: Patient Pose Alignment, Longitudinal Registration, Multimodal Registration, Computed Tomography

***Corresponding author: Elisabeth Schiele,** Chair for Computer Aided Medical Procedures and Augmented Reality, Technical University of Munich, Germany, e-mail: elisabeth.schiele@tum.de
Lukas Förner, Thomas Wendler, Chair for Computer Aided Medical Procedures and Augmented Reality, Technical University of Munich, Germany
Melina Wördehoff, Josua A. Decker, Lukas Förner, Thomas Wendler, Department of Diagnostic and Interventional Radiology and Neuroradiology, University Hospital Augsburg, Germany
Melina Wördehoff, Lukas Förner, Thomas Wendler, Digital Medicine, University Hospital Augsburg, Augsburg, Germany
Melina Wördehoff, Lukas Förner, Thomas Wendler, Bavarian Center for Cancer Research (BZKF) Augsburg, Germany
Thomas Wendler, Center for Advanced Analytics and Predictive Sciences, University of Augsburg, Germany

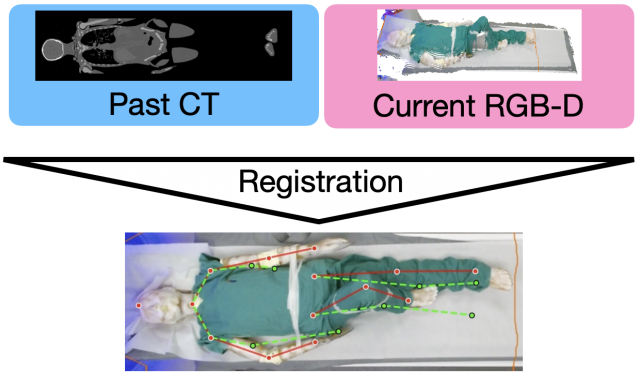


Fig. 1: From a prior CT scan and a current RGB-D image, the clinician is given a pose guidance to align the current pose with the past pose.

1 Introduction

Longitudinal CT scanning is essential for clinical diagnoses, particularly in evaluating tumor development and measuring bone mineral density [1, 2]. However, these analyses are highly sensitive to variations in patient posture. Minor postural changes cause fundamentally different settings for the scanner field of view and required dose distribution. Current post-hoc image registration techniques are computationally intensive, prone to error, and limited by the magnitude of the initial pose discrepancy. Rather than correcting positional inconsistencies after acquisition, this work proposes addressing the problem at its source: aligning the patient's pose *before* the follow-up scan. By capturing the patient's current position with an overhead RGB-D camera and registering it against a prior CT scan, the required postural corrections can be computed and visualized in real time. This proactive approach has the potential to reduce initial pose discrepancy, thereby improving both the raw data quality and the reliability of any subsequent post-hoc registration.

The key contributions of this work are:

1. We propose pre-scan patient pose alignment as a complement to conventional post-hoc registration and demonstrate its feasibility on phantom data.
2. We present a simple articulated registration pipeline that bridges the domain gap between RGB-D surface data and volumetric CT data using torso-based anchoring, requiring no training data or learned models.

Open Access. © 2026 The Author(s), published by De Gruyter. This work is licensed under the Creative Commons Attribution 4.0 License.



Fig. 2: Overview of the three phantom poses (Pose 1, Pose 2, Pose 3, left to right) with limb segmentation overlays.

2 Methodology

2.1 Data Acquisition and Preprocessing

Data was acquired using a photon-counting CT scanner (Siemens NAEOTOM Alpha [3]) equipped with a top-mounted Microsoft Azure Kinect RGB-D camera [4]. As patient pose differences are inherently three-dimensional, a depth-capable sensor is required; a conventional photograph captures no depth and cannot resolve out-of-plane positional differences. The system captures a calibrated RGB-D image immediately prior to each CT acquisition, providing paired surface and volumetric data of the same pose. Three such paired acquisitions were collected, each with a different phantom pose (Figure 2). To evaluate inter-pose registration, the CT scan from one acquisition is registered against the RGB-D image of a different acquisition, simulating the clinical scenario where a patient is in a different position for a follow-up scan. To comply with ethical standards and avoid unneeded radiation exposure for humans, the gold-standard CT Whole Body Phantom "PBU-60" [5] was utilized to simulate human anatomy in both RGB-D and CT data (Figure 4).

To register RGB-D and CT data, several preprocessing steps are necessary. The patient's skin is extracted from the volumetric CT scan using the Marching Cubes algorithm [6] at a threshold of -200 Hounsfield units (HU). From the Marching Cubes output the relevant components are manually selected and smaller artefacts are discarded. The CT mesh is then segmented into torso, upper/lower arms, and upper/lower legs using planar and region-of-interest cuts. This approach reliably captures the soft-tissue to air transition of the CT scan, as shown in the top image of Figure 4. Separately, the RGB-D data, which captures the clothed surface, is filtered based on the scanning bed height and an approximate human body threshold to isolate the target geometry. As off-the-shelf pose estimators (YOLOv8 [7, 8], Keypoint R-CNN [9]) failed to generalize to the supine clinical domain, the user interactively annotates the 2D skeleton onto the RGB image (14 keypoint clicks per scan, performed once per baseline), which is used to automatically generate semantic masks of upper/lower arms, upper/lower legs and the torso, as visualized in Figure 2. Through this preprocessing, the CT scan and RGB-D image are transformed into individual point clouds of these different anatomical regions.

2.2 Articulated Registration Framework

Because CT and RGB-D data do not share a common world coordinate system, and because CT fields of view (FOV) are often limited, global registration metrics are insufficient. We propose an anchored, articulated registration pipeline treating the human body as a set of locally rigid elements. The method consists of two steps:

1. **Torso Anchoring:** The torso, being the least articulated and most consistently visible structure, serves as the rigid point of reference. The CT torso mesh is converted to a point cloud via Poisson disk sampling [10], followed by Hidden Point Removal to simulate the 2.5D topology of the overhead RGB-D camera. Initial alignment is performed using Principal Component Analysis (PCA) [11] and centroid matching (Figure 3), followed by point-to-plane Iterative Closest Point (ICP) [12] refinement. The point-to-plane variation of ICP minimizes the distance from source points to the target's tangent plane, allowing surfaces to slide and reducing sensitivity to clothing-induced noise. The CT-torso mesh is sampled to 30,000 points for ICP and 10,000 points for PCA initialization.
2. **Limb-to-Limb Registration:** With the coordinate systems anchored via the global torso registration, the decoupled limbs are treated as independent rigid bodies. The rigid transformation obtained from torso registration, T_{global}, is first applied uniformly to all CT limb segments, placing them at the positions they would occupy if the patient matched the prior CT pose. The centroid of each transformed limb is recorded as $\mathbf{p}_{\text{ideal}}$. Each CT limb is first coarsely aligned by centroid and primary PCA axis, then the position is refined using local point-to-plane ICP. CT-limb meshes are sampled to 15,000 points for ICP and 5,000 points for PCA. All geometric primitives employed are established methods [6, 10–12]; our contribution is their composition into a torso-anchored articulated pipeline with 2D clinical guidance projection.

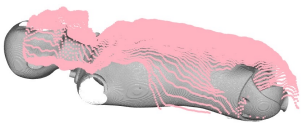


Fig. 3: Coarse torso alignment. The CT mesh (grey) is initialy aligned to the RGB-D torso point cloud (pink) using principal axis and centroid correspondence, prior to ICP refinement.

2.3 Visual Pose Guidance

The ultimate goal of the registration is to derive a 3D correction vector, defined as $\vec{v}_{\text{corr}} = [x, y, z]^T$, which represents the movement required to return the limb from its current registered position $\mathbf{p}_{\text{actual}}$ to the target position $\mathbf{p}_{\text{ideal}}$ derived from the baseline CT scan. To provide actionable clinical guidance, this vector is projected onto the high-resolution 2D RGB image. Using a pinhole camera model, the 2D pixel shifts $(\Delta u, \Delta v)$ are calculated alongside depth compensation:

$$\Delta u = \frac{x \cdot f_x}{Z_{\text{depth}}}, \quad \Delta v = \frac{y \cdot f_y}{Z_{\text{depth}}}. \tag{1}$$

The guidance overlay displays the current pose (red skeleton) and the target pose (green skeleton). Figure 4 shows the CT scan, RGB(-D) image, the result of the registration in 3D space and the output pose guidance. Here, the difference in pose is a raised knee in the current RGB-D image where the past CT scan had both legs lying flat on the scanning bed.

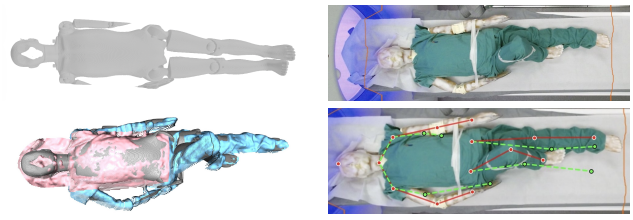


Fig. 4: Left to right, top to bottom: Surface extraction from CT Scan (1), RGB(-D) input (2), 3D alignment of (1) and (2), pose guidance to guide current RGB(-D) pose (2) to match prior CT pose (1).

3 Results

3.1 Torso Alignment Accuracy

Torso alignment performance is evaluated against a rigid whole-body registration baseline across 10 iterations per pose. Since the CT scanner and RGB-D camera do not share a coordinate system, no absolute ground truth exists. Instead, a whole-body rigid registration using all available limbs provides a reference alignment. The centroid and orientation of the torso segment within this whole-body registration serve as the baseline for comparison. This setup isolates the bias introduced by the torso-only approach: when the full body is used, limb geometry pulls the registration toward a globally optimal fit, whereas the torso-only method must recover alignment from a single anatomical region. The reported errors thus quantify the spatial drift between these two strategies rather than deviation from an absolute ground truth and thus serve as a repeatability reference, not a clinical benchmark. Table 1 summarizes the results across the three poses.

Tab. 1: Torso Registration Consistency Analysis: Average Centroid and PCA Angle Error over 10 iterations.

Dataset	Centroid Error (cm)	PCA Angle Error (°)
Pose 1	2.58 ± 0.13	3.30 ± 0.14
Pose 2	3.44 ± 0.11	3.00 ± 0.13
Pose 3	2.07 ± 0.15	4.08 ± 0.16
Total Average	**2.69**	**3.46**

3.2 Articulated Registration and Guidance

Table 2 summarizes the surface registration performance across all three CT scans and RGB-D poses. The mean point-to-surface error remains consistently between 1.61 cm and 1.84 cm, with RMSE values between 2.24 cm and 2.54 cm, indicating stable global alignment. All configurations exhibit very low variability across ten repeated runs (standard deviations below 0.01 cm), demonstrating high repeatability. Maximum errors range from 8.25 cm to 12.55 cm. Scan 3, which includes the largest portion of the arms in its FOV, consistently shows the lowest maximum errors (below 10 cm); the outliers exceeding 12 cm in Scans 1 and 2 are attributable to FOV truncation artefacts rather than registration failure, as the CT limb geometry is incomplete at the scan boundary.

Tab. 2: Surface Registration Fit Accuracy (cm).

CT	Pose	Mean	RMSE	Max
1	1	1.61	2.24	12.55
	2	1.70	2.41	11.82
	3	1.80	2.51	11.33
2	1	1.70	2.25	11.88
	2	1.69	2.30	10.88
	3	1.82	2.46	11.23
3	1	1.84	2.54	9.98
	2	1.69	2.34	8.80
	3	1.84	2.44	8.25

Qualitatively, the final visual guidance successfully and intuitively indicated the required direction of positional adjustment for misaligned limbs, consistent with the system's goal of gross pose correction rather than sub-centimetre repositioning. This is visualized in the last image of Figure 4. For large articulations, such as moving arms from an overhead position to the torso, the system effectively communicates the broad necessary adjustment, but fails to give the exact goal position. Comparison between two identical poses shows consistent registration of the same pose in the output.

4 Discussion and Conclusion

This work establishes a proof-of-concept for proactive, pre-scan patient pose alignment using RGB-D to CT registration. The current system computes one-shot guidance from a single RGB-D capture; in deployment this naturally extends to an iterative loop, re-capturing after each adjustment until the residual correction falls below a defined threshold. Torso anchoring proves highly reliable across all evaluated configurations (Table 1), which can be attributed to the torso being the largest, least articulated body region, combined with the head resting stably on the scanner cushion thus providing consistent geometry in both modalities.

Limb registration poses greater challenges. Crucially, the primary source of error is not the registration algorithm itself, but the geometric mismatch between complete RGB-D limb surfaces and CT limbs truncated by the scanner's limited field of view. This is supported by the consistently lower maximum errors observed for Scan 3 (Table 2), which captures the largest portion of the arms. The use of the PBU-60 phantom introduces additional limitations: its rigid plastic joints restrict the range of achievable articulations compared to real human anatomy, its translucent skin requires surface spray and hospital clothing to produce usable depth data, and the domain gap to real patients remains unvalidated. Furthermore, standard pose estimation models (YOLOv8 [7, 8], Keypoint R-CNN [9]) fail to generalize to supine, occluded patients in the clinical scanning environment. Domain-specific in-bed pose estimators have been developed for hospital monitoring [14] and show that domain adaptation remains an active challenge [15], but none address the specific combination of clinical gowns and unconstrained patient positioning present in pre-scan patient positioning. Removing the current manual annotation step by addressing the domain gap, for instance through fine-tuning on clinical in-bed data and integrating parametric body models such as SMPL [13] is a key requirement for clinical deployment.

While post-hoc registration methods may exceed our approach in accuracy, the proposed approach addresses an acquisition-time problem (FOV inclusion, dose distribution, and initial-misalignment magnitude) that post-hoc correction cannot. Future work will quantify net clinical benefit against established positioning protocols and a simple 2D-photo repositioning baseline, advancing this approach toward clinical deployment. Together, these results demonstrate the feasibility of proactive pose alignment as a training-free complement to conventional post-hoc registration in longitudinal CT.

Author Statement

Research funding: This research was partially funded by the Bavarian Center for Cancer Research as part of the Lighthouse "Local Therapies", as well as by the Bavarian Ministry of Economic Affairs, Regional Development and Energy (StMWi) under grant number DIK-2310-0004//DIK0556/02. **Conflict of interest**: Authors state no conflict of interest. **Informed consent**: Not applicable, as this study used phantom data only. **Ethical approval**: Not applicable, as this study used phantom data only and did not involve human subjects or animal experiments.

References

[1] Liu H, Durongbhan P, Davey CE, Stok KS. Image registration in longitudinal bone assessment using computed tomography. Current Osteoporosis Reports 2023;21:372–385.

[2] Hering A, Peisen F, Amaral T, Gatidis S, Eigentler T, Othman A, et al. Whole-Body Soft-Tissue Lesion Tracking and Segmentation in Longitudinal CT Imaging Studies. Proceedings of the Fourth Conference on Medical Imaging with Deep Learning 2021, PMLR 143:312-326.

[3] Shanblatt E, O'Doherty J, Petersilka M, Wolber P, Fung G, Ramirez-Giraldo J. NAEOTOM Alpha CT redefined. White Paper. 2023.

[4] Microsoft. Azure Kinect DK documentation. https://learn.microsoft.com/en-us/azure/kinect-dk/.

[5] Kyoto kagaku. CT Whole Body Phantom "PBU-60". https://www.kyotokagaku.com/products_data/ph2b_catalog_en.pdf.

[6] Cline H. Lorensen W. System and method for the display of surface structures contained within the interior region of a solid body. Patent US4710876A. 06/1987.

[7] Redmon J, Divvala S, Girshick R, Farhadi A. You Only Look Once: Unified, Real-Time Object Detection. Proceedings of Computer Vision and Pattern Recognition 2016.

[8] Jocher G, Qiu J, Chaurasia A. Ultralytics YOLO (Version 8.0.0) [software]. 2023. Available from: https://github.com/ultralytics/ultralytics

[9] He K, Gkioxari G, Dollár P, Girshick R. Mask R-CNN. Proceedings of the International Conference on Computer Vision (ICCV) 2017.

[10] Bridson R. Fast Poisson disk sampling in arbitrary dimensions. SIGGRAPH:Computer Graphics and Interactive Techniques 2007:22.

[11] Pearson K. On Lines and Planes of Closest Fit to Systems of Points in Space. Philosophical Magazine 1901;2:559-572.

[12] Low K. Linear Least-Squares Optimization for Point-to-Plane ICP Surface Registration. In: Technical Report TR04-004. Department of Computer Science, University of North Carolina at Chapel Hill. February 2004.

[13] Loper M, Mahmood N, Romero J, Pons-Moll G, Black MJ. SMPL: A Skinned Multi-Person Linear Model. ACM Transactions on Graphics 2015;34(6):248:1–248:16.

[14] Achilles F, Ichim AE, Coskun H, Tombari F, Noachtar S, Navab N. Patient MoCap: Human Pose Estimation Under Blanket Occlusion for Hospital Monitoring Applications. Medical Image Computing and Computer-Assisted Intervention (MICCAI) 2016:491–499.

[15] Bigalke A, Hansen L, Diesel J, Hennigs C, Rostalski P, Heinrich MP. Anatomy-guided domain adaptation for 3D in-bed human pose estimation. Medical Image Analysis 2023;89:102887.