Received: 15 November 2020 | Revised: 15 March 2021 | Accepted: 28 June 2021

DOI: 10.1002/mp.15124

RESEARCH ARTICLE

MEDICAL PHYSICS

A new 2D-3D registration gold-standard dataset for the hip joint based on uncertainty modeling

Fabio D'Isidoro[1] | **Christophe Chênes**[2] | **Stephen J. Ferguson**[1] | **Jérôme Schmid**[2]

[1]Institute for Biomechanics, ETH Zürich, Zürich, Switzerland

[2]Geneva School of Health Sciences, HES-SO University of Applied Sciences and Arts of Western Switzerland, Geneva, Switzerland

Correspondence
Jérôme Schmid, Haute école de santé de Genève, 47 Avenue de Champel, 1206 Genève, Switzerland.
Email: jerome.schmid@hesge.ch

Funding information
EC | Seventh Framework Programme (FP7), Grant/Award Number: 310477; Kommission für Technologie und Innovation (CTI), Grant/Award Number: 25258.1 PFLS-LS

Abstract

Purpose: Estimation of the accuracy of 2D-3D registration is paramount for a correct evaluation of its outcome in both research and clinical studies. Publicly available datasets with standardized evaluation methodology are necessary for validation and comparison of 2D-3D registration techniques. Given the large use of 2D-3D registration in biomechanics, we introduced the first gold standard validation dataset for computed tomography (CT)-to-x-ray registration of the hip joint, based on fluoroscopic images with large rotation angles. As the ground truth computed with fiducial markers is affected by localization errors in the image datasets, we proposed a new methodology based on uncertainty propagation to estimate the accuracy of a gold standard dataset.

Methods: The gold standard dataset included a 3D CT scan of a female hip phantom and 19 2D fluoroscopic images acquired at different views and voltages. The ground truth transformations were estimated based on the corresponding pairs of extracted 2D and 3D fiducial locations. These were assumed to be corrupted by Gaussian noise, without any restrictions of isotropy. We devised the multiple projective points criterion (MPPC) that jointly optimizes the transformations and the noisy 3D fiducial locations for all views. The accuracy of the transformations obtained with the MPPC was assessed in both synthetic and real experiments using different formulations of the target registration error (TRE), including a novel formulation of the TRE (uTRE) derived from the uncertainty analysis of the MPPC.

Results: The proposed MPPC method was statistically more accurate compared to the validation methods for 2D-3D registration that did not optimize the 3D fiducial positions or wrongly assumed the isotropy of the noise. The reported results were comparable to previous published works of gold standard datasets. However, a formulation of the TRE commonly found in these gold standard datasets was found to significantly miscalculate the true TRE computed in synthetic experiments with known ground truths. In contrast, the uncertainty-based uTRE was statistically closer to the true TRE.

Conclusions: We proposed a new gold standard dataset for the validation of CT-to-X-ray registration of the hip joint. The gold standard transformations were

derived from a novel method modeling the uncertainty in extracted 2D and 3D fiducials. Results showed that considering possible noise anisotropy and including corrupted 3D fiducials in the optimization resulted in improved accuracy of the gold standard. A new uncertainty-based formulation of the TRE also appeared as a good alternative to the unknown true TRE that has been replaced in previous works by an alternative TRE not fully reflecting the gold standard accuracy.

KEYWORDS
2D-3D registration, CT-to-X-ray image registration, gold standard dataset, uncertainty propagation

1 | INTRODUCTION

The goal of 2D-3D registration is to find the spatial transformation that best aligns 3D imaging data with one or more 2D projection images, in the 3D physical space. Typically, the 3D volume consists of pre-intervention data such as computed tomography (CT) or magnetic resonance (MR) scans, while the 2D images are intra-intervention data such as a radiograph or fluoroscopic image. Orthopedic applications of 2D-3D registration include spine surgery, total hip replacement, orthopedic diagnostics, and kinematic analysis.[1] In spine surgery, the registration of single vertebrae is mostly used for pedicle screw placement and cement reinforcement.[2–4] For total hip replacement, the registration is used for intra-operative positioning of the femoral implant[5–7] and post-operative analysis of cup placement.[8–10] In orthopedic diagnostics, the 3D curvature of scoliotic spine[11] and the scoliotic rib cage were analyzed.[12] For kinematics analysis, 2D-3D registration between 3D models of the joint and fluoroscopic video sequences acquired during various activities of daily living was used to analyze the in vivo motion of the native knee[13–15] and hip,[16–18] as well as of the prosthetic knee[19–21] and hip.[22–24]

Evaluation of the accuracy of 2D-3D registration is paramount to determine the performance and limitations of proposed methods, and to clarify the potential clinical application and benefit compared to possible pre-existing methods.[25] Typically, registration accuracy is estimated by comparison to an accurate "gold-standard" registration method applied on a sample dataset that is representative of the specific application. Due to the large number of techniques proposed in literature, effective comparison between registration algorithms or evaluation of the accuracy of one registration technique for different applications is only possible with a standardized evaluation methodology and publicly available validation datasets.

To date, only few gold standard datasets are publicly available for the validation of 2D-3D registration for application in orthopedics. They include sets of CT, MR volumes, and x-ray images of human cadaveric spines[26,27] and of a fresh porcine cadaver head[28] and lungs[29] as well as a simulated dataset of CT and digitally reconstructed radiographs (DRRs) of human pelvis and vertebrae from the Visible Human Project.[30] Using synthetic DRR images provides an exact known ground truth but usually results in non-fully realistic x-ray images (e.g., absence of x-ray scattering or image noise). Some works proposed more realistic DRR images but were highly specific to body areas (e.g., chest[31]), although recent work leveraging deep learning advances look very promising (e.g., DeepDRR[32]). To the authors' knowledge, only one validation dataset of the hip joint[33] with real fluoroscopic images currently exists. Among the limitations of this dataset, we identified the limited rotation angles of the fluoroscopic images and most importantly the absence of quantitative assessment of its quality as targeted in this paper. Due to the large number of orthopedic research studies focused on the hip joint, the first aim of this work is to provide another gold-standard dataset to fill the present gap.

The second focus is the improvement of both the accuracy of a gold-standard dataset and of the method to estimate it. For most non-synthetic datasets,[26,28,34] the gold standard rigid transformations are retrieved with fiducial markers that are rigidly fixed to the phantom or patient. The location of these markers is extracted from both the 3D volume and 2D image datasets. The obtained corresponding 2D-3D pairs are used to compute the accurate spatial alignment of the 3D dataset in the calibrated coordinate system of the 2D image. However, such gold standard transformations are not guaranteed to be completely accurate due to the x-ray system calibration errors (e.g., inaccurate computation of the source-to-detector distance) and to the fiducial localization error (FLE), that is, the error in the extraction of the 3D and 2D locations of the fiducials. We propose an approach to compute the gold-standard transformations for 2D-3D registration which accounts for isotropic and anisotropic Gaussian FLEs in both 2D and 3D. This approach could also be used in the interventional context, where preoperative data are brought in correspondence with intraoperative information via 2D-3D transformations estimation.

Estimation of the accuracy of the ground truth is important as it defines the "uncertainty" of the gold-standard solution, against which the transformation obtained with a 2D-3D registration algorithm is evaluated. Most studies estimated the accuracy of their gold-standard dataset by computing the expected target registration error (TRE).[26,28,34] The TRE measures the displacement from their true position of registered target points not used as fiducials, which are typically chosen within the region of interest of the 3D dataset. The expected TRE was computed based on seminal works of Sibson[35] and Fitzpatrick et al.[36] under the assumption of isotropic, homogeneous, and independent Gaussian noise on the extracted location of the 3D fiducials.[37] In the present work, we investigate whether some of these conditions may not be always met and propose to use an alternative TRE computation grounded in uncertainty theory.

2 | MATERIALS AND METHODS

2.1 | Phantom preparation and image acquisition

We used a phantom including a female pelvis, proximal femurs, and lumbar spinal segments embedded in a resin substrate mimicking the radiological response of soft tissue. Metallic beads of 3 mm diameter (N = 21 fiducials) were rigidly attached to the outer surface of the phantom (Figure 1a). Fourteen retroreflective motion capture (MoCap) markers were additionally stuck to the surface.

A CT scan of the phantom with the fiducials and the MoCap markers was acquired with a Brilliance CT 64 scanner (Philips Medical Systems) at 140 kV (Figure 1b), which resulted after cropping in a 431 × 315 × 468 volume with a voxel size of 0.78 × 0.78 × 1.0 mm^3.

A video-fluoroscopy C-arm (BV Pulsera, Philips Medical Systems) was used to acquire a set of S = 19 2D images at different orientations of the phantom around the vertical axis of the lab (Figure 1c). For each view, the fluoroscope was operated at several different kV and mAs settings. The 2D fluoroscopic images have an image matrix size of 1000 × 1000 square pixels and a grayscale dynamic range of 12 bits.

A schematic overview of the measurements and of the variables involved in the computation of the gold-standard dataset is provided in Figure 2. The fiducial locations M_i (i: 1⋯N) were defined in the coordinate system of the CT scan CS_{ct}. For each view I ($I = 1 \cdots S$), the location of a 3D fiducial M_i is transformed to the coordinate system CS_I relative to the X-ray image I_I with the rigid transformation T^I, and subsequently projected onto the image plane I_I as the 2D fiducial position m_i^I:

$$m_i^I = P\, T^I M_i \tag{1}$$

where P is the 3 × 4 projection matrix and $T^I = [R^I \mid t^I]$ is the transformation from CS_{ct} to CS_I.

The aim of the gold-standard dataset is to provide the set of ground truth transformations T^I for each view $I = 1 \cdots S$ based on the 3D-2D corresponding pairs

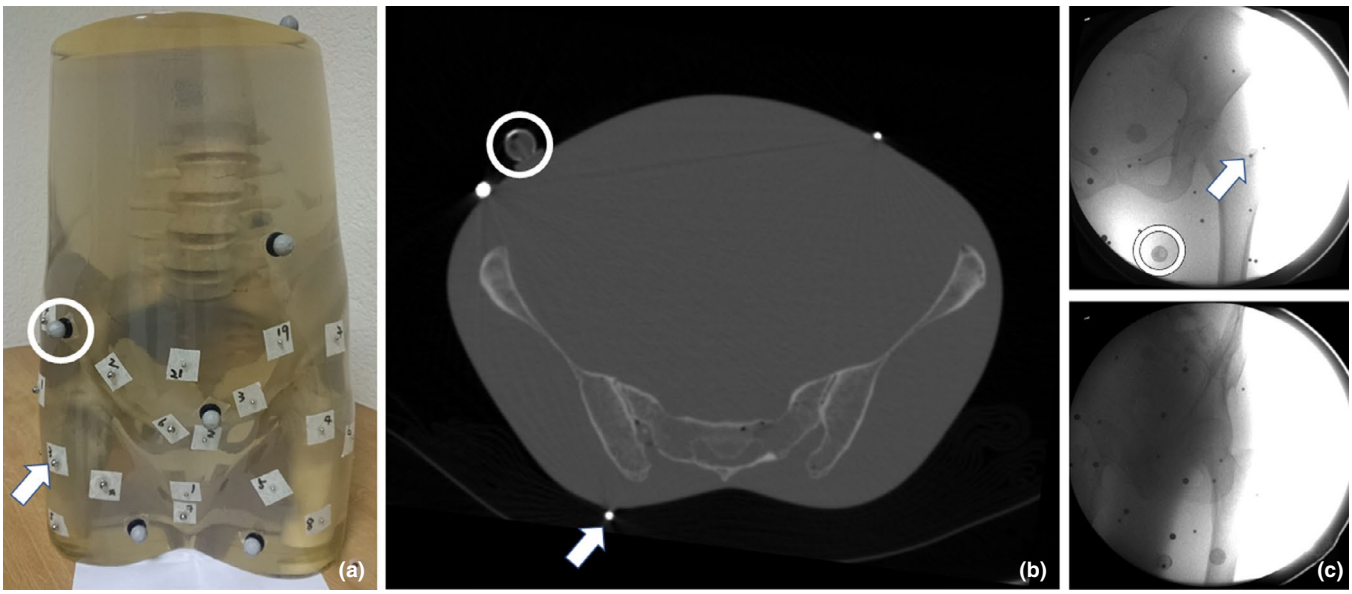

FIGURE 1 X-ray phantom of a female pelvis embedded into a material mimicking radiological response of soft tissue. (a) Example of motion capture (MoCap) marker (white circle) and metallic spherical fiducial (white arrow) stuck on the phantom surface, with other examples exemplified in (b) CT volume and (c) X-ray image acquired with different phantom orientations. For illustration purpose, depicted markers are not in correspondence between subfigures (a), (b), and (c)

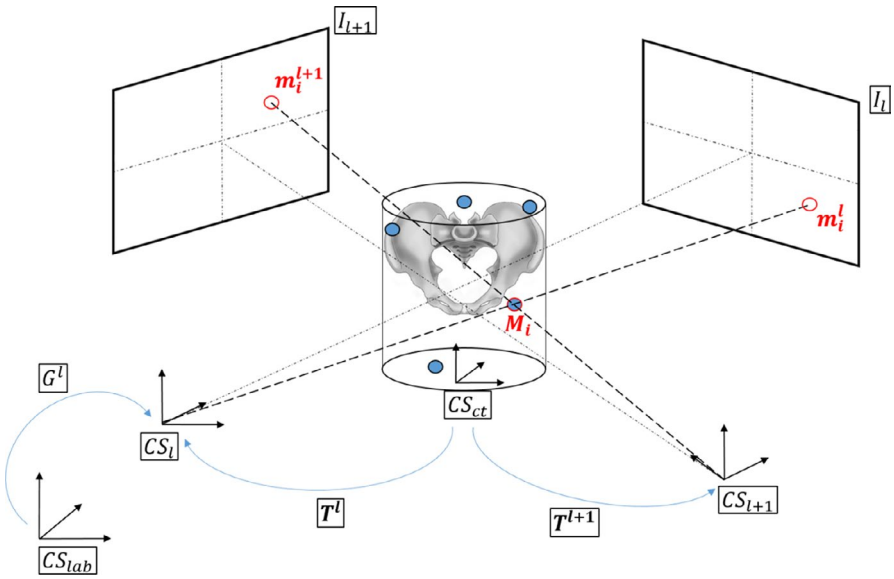

FIGURE 2 Schematic overview of the generation of the gold-standard dataset. M_i represents the 3D coordinates of the i-th fiducial in the coordinate system CS_{ct} of the CT scan, while m_i^l represents the pixel coordinates of the i-th fiducial in the image I^l at view l. T^l is the rigid transformation of the coordinate system of the phantom CS_{ct} relative to the X-ray coordinate system CS_l and represents the ground truth transformation. The equation describing the projection of M_i onto m_i^l is: $m_i^l = P\,T^l\,M_i$, where P is the intrinsic camera projection matrix relative to the X-ray imaging system. G^l is the rigid transformation of the lab coordinate system CS_{lab} relative to the X-ray coordinate system CS_l and is used to transform lab coordinates of the motion capture markers into corresponding coordinates in the CS_l in order to retrieve a coarse estimation of the ground truth T^{l*} from motion capture. In practice, the phantom was actually moved at each view l with respect to a static imaging system. Hence G^l is in fact the same for all views l

(m_i^l, M_i). The ground truth of the present gold-standard dataset considers that the 3D and 2D locations of the fiducials are affected by errors. In our work, we regroup the corrupted measured 2D and 3D positions $\tilde{m}_i^l$ and $\tilde{M}_i$ in a *measurement* vector $\chi = (\tilde{M}_1, \cdots, \tilde{M}_N, \tilde{m}_1^1, .. \tilde{m}_N^S)$. Ideal 3D positions M_i are similarly regrouped into the *model* vector M. We gather all the parameters of the unknown transformations T^l in a *transformation* vector $T = \left\{ T_1^1, \cdots, T_6^1, \cdots, T_1^l, \cdots, T_6^l, \cdots, T_6^S \right\}$, where subscripts 1 to 3 and subscripts 4 to 6 refer to the rotational and the translational parameters, respectively. We chose the rotation vector representation,[38] where the vector direction provides a rotation axis and its magnitude represents the rotation angle around this axis.

The fluoroscopic system was considered to be calibrated because the projection matrix P was estimated from a calibration procedure described in Appendix 1 (Supplementary Material). This work assumes that the re-projection error of 0.033 mm obtained from the calibration procedure is small enough, so that the propagation error originated from the calibration can be neglected in the uncertainty analysis.

2.2 | MoCap acquisition and processing

Optical MoCap was performed simultaneously to video fluoroscopy, in order to get a coarse estimate of the ground truth transforms based on motion capture and in order to automatically define the correspondences between 2D and 3D fiducial pairs. A VICON MX system (Oxford Metrics Group, UK), and 26 MX40 and T160 infrared cameras recorded at 100 Hz positions in the lab coordinate system CS_{lab} of the MoCap markers were attached to the phantom (Figure 1). The accuracy of 3D point computation of our MoCap setup is difficult to assess, as generally several factors, such as the number and coverage of cameras, impact the overall accuracy.[39] The impact of the MoCap setup accuracy will be investigated in the experiments validating our gold-standard dataset. For each view l, the rigid transformation $T^l_{Lab \leftarrow CT}$ of the phantom (CS_{ct}) relative to the lab coordinate system CS_{lab} was computed by 3D-3D registration between the positions of the optical markers measured in the lab and positions of the markers in the CT coordinate system CS_{ct}. The obtained transformation $T^l_{Lab \leftarrow CT}$ was converted into the coordinate system of the imaging system CS_l by applying the conversion matrix G^l relating coordinates in the lab with coordinates in the imaging system (Figure 2). In practice, the imaging system was static in CS_{lab} while the phantom was moved at each view l. Thus, $\forall l$, $G^l = G$ and G needed to be computed only once. Finally, a MoCap-based estimate of the ground truth transform T^{l*} was obtained for each view l as:

$$T^{l*} = G^l * T^l_{Lab \leftarrow CT}.$$

Interested readers can refer to Appendix 2 of Supplementary Material for further details.

2.3 | Fiducial positions measurement and correspondence

Our regularized deformable model framework[40] was used to automatically extract the centers of 3D fiducials $\tilde{M}_i$ and MoCap markers, referred to as "spherical objects," from the CT volume. For each fiducial/MoCap marker, a spherical mesh was deformed until it best matched the boundaries of the spherical object based on the alignment of intensity gradients and the mesh vertex normals. The centers of gravity of the resulting fitted spheres were set as the positions of the 3D fiducials/MoCap markers.

The pixel coordinates of the 2D fiducial centers $\tilde{m}_i^l$ from each fluoroscopic image I_l acquired at the kV value producing the best image contrast were retrieved by means of an in-house developed semi-automatic method. This algorithm relied on a blob detection algorithm provided by the open source computer vision library "OpenCV"[41] to interactively detect the 2D fiducial positions as centers of fitted ellipses to detected blobs.

Once we computed the positions of the 3D and 2D fiducials, 2D-3D correspondence was established in automatic fashion by exploiting the coarse estimate of the transform T^{l*} obtained from motion capture. Transformation T^{l*} was used to project the positions of the 3D fiducials $\tilde{M}_i$ to 2D positions $m_i^{l*} = P\,T^{l*}\,\tilde{M}_i$. Given a projected position m_i^{l*}, the closest measured 2D position $\tilde{m}_k^l$ was identified. If the Euclidean distance $\| m_i^{l*} - \tilde{m}_k^l \|$ was below the threshold of 5 mm, the point $\tilde{m}_k^l$ was flagged as visible in the image I_l and set in correspondence with the 3D fiducial $\tilde{M}_i$.

We first estimated the 3D fiducial extraction accuracy in synthetic experiments, in which an artificial noisy CT scan-like 3D volume was created including 20 spheres. We varied the volume characteristics (voxel size and isotropy, levels of additive Gaussian noise), sphere properties (radius and intensity), and initialization positions for the automatic segmentation. The resulting signed differences between expected and extracted centers of more than 24 000 spheres were: 0.018 ± 0.04, 0.014 ± 0.074, and 0.006 ± 0.10 mm in X-, Y-, and Z-directions, respectively, Z being the slice stacking direction. Then, we used a quality assurance (QA) phantom (Lucy 3D QA Phantom, Standard Imaging, Inc.) in an in vitro experiment. The QA phantom included twenty 2 mm diameter aluminum spheres spaced by 5 mm (manufacturing tolerance of 0.1 mm). We acquired a CT scan (120 kV, size 512 × 512 × 340, 0.31 × 0.31 × 0.5 mm^3 of voxel size, Philips Brilliance CT Big Bore model) of the QA phantom (Figure 3a) and extracted the centers of the segmented spheres (Figure 3b) in 400 trials in which we randomly varied the initial centers within the 20 spheres according to a normal distribution with $3\sigma = 0.5$ mm–mimicking a user click around the sphere centers. These were rigidly registered to the reference centers of the QA phantom, with resulting average errors of 1.76e-09 ± 0.16, 1.17e-09 ± 0.16, and −2.86e-09 ± 0.18 mm in X-, Y-, and Z-directions, respectively. By combining the signed differences of synthetic and Lucy phantom experiments altogether, we obtained average difference errors of 0.01 ± 0.09, 0.01 ± 0.10, and 0.004 ± 0.127 mm in X-, Y-, and Z-directions, respectively.

The same QA phantom was used to assess the 2D fiducial center extraction, which was performed on multiple DRRs of the QA phantom CT at various angles (Figure 3c), with a DRR spatial resolution of

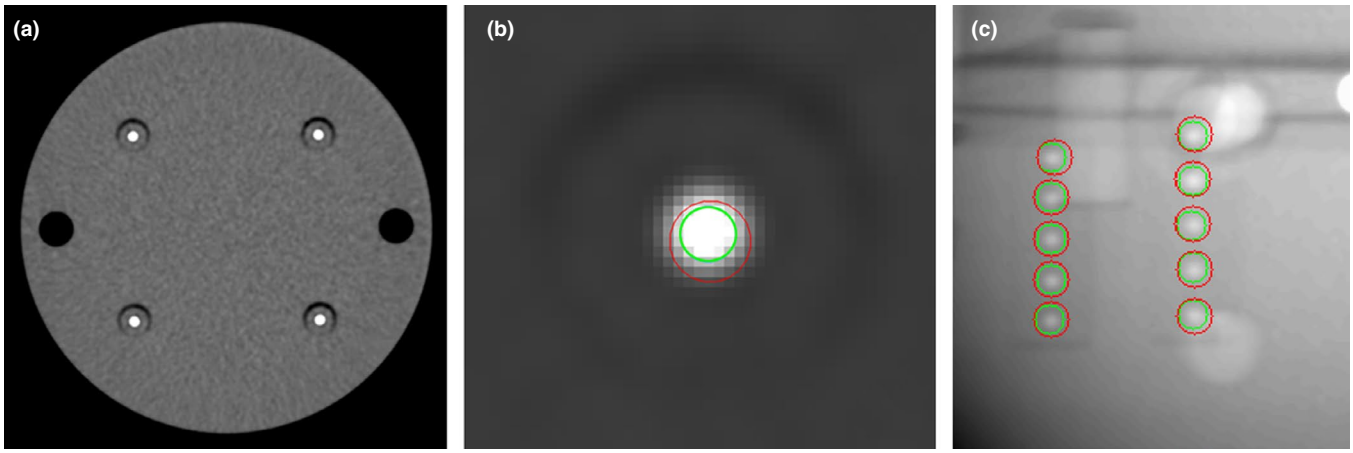

FIGURE 3 Metallic sphere detection of quality assurance phantom. a) CT scan of the phantom showing four 2 mm diameter aluminum spheres. b) Example of 3D sphere extraction based on regularized deformable models where the larger red circle is the initialized model and the smaller green circle is the final result. c) Example of 2D extraction where the reference locations (centers of larger red circles) are compared with the extracted locations (centers of smaller green circles)

0.29 × 0.29 mm^2. The reference 2D centers were obtained by projection of the 3D centers extracted using our 3D segmentation approach. Differences between extracted 2D centers and reference 2D positions were 0.15 ± 0.19 mm and −0.02 ± 0.19 mm in horizontal X- and vertical Y-directions, respectively.

2.4 | Multiple projective points criterion (MPPC)

Traditional Perspective-n-Point (PnP) algorithms[35,36] are commonly used to compute the ground truth transformations T_l from P and corresponding pairs (M_i, m_i^l) using Equation (1). However, standard PnP algorithms are not suited to account for inaccuracies in measured 2D fiducial positions. Different extensions of the PnP algorithm were proposed to address this issue, such as the CEEPnP[42] and ML-PnP[43] approaches. Alternatively, optimization approaches[26,44] were developed to reduce the impact of 2D inaccuracies by minimizing the 3D fiducial registration error (FRE), defined as the distance between 3D fiducials segmented from the CT scan and the 3D fiducials reconstructed by triangulation of the extracted 2D fiducial image positions. For a set of 2D image positions in multiple views corresponding to the same fiducial, the triangulation computes the 3D reconstructed point as the 3D closest point to back-projected lines passing through the 2D image positions.[26] Other studies[28] obtained the ground truth transformation by minimizing the 2D mean projection distance (mPD) between extracted 2D fiducial image positions and reprojected 3D fiducials.

However, most of these approaches continue to assume that the positions of the 3D markers of the model are perfectly known or that errors in their detection are negligible. These assumptions may become invalid when fiducials are manually placed on gold-standard phantoms for 2D-3D registration as proposed in the current work and previous studies.[26,28,34] In our work, we simultaneously optimized both the transformation parameters and the 3D markers location, similar to the work of Nicolau et al.[44] They defined the extended projective points criterion (EPPC) to determine the optimal T and the optimal M, hereafter referred to as $\hat{T}$ and $\hat{M}$, from a maximum likelihood estimator: $\widehat{T},\widehat{M} = \underset{T,M}{argmax} p(\chi \mid T, M)$, with p being the conditional probability density function.

Nicolau et al. considered that 2D and 3D fiducials were corrupted by zero-mean Gaussian isotropic noise parameterized by variances σ_{2D}^2 and σ_{3D}^2. In our work, we assume that positions of the 2D fiducials m_i^l and 3D fiducials M_i are identically and independently corrupted by additive zero-mean Gaussian noises with covariance matrices Σ_{2D}^l and Σ_{3D}. We can thus model both isotropic and anisotropic noises. Furthermore, in our case the transformations between the different X-ray CS_l are unknown so we have to optimize multiple transforms. Based on these assumptions, the conditional probability of our measurement vector χ is written as the product of independent probabilities[45]:

$$p(\chi|T,M) = \left(\prod_{l=1}^{S} \prod_{i=1}^{N} p\left(\tilde{m}_i^l | T, M\right)^{\epsilon_i^l} \right) * \prod_{i=1}^{N} p(\tilde{M}_i | T, M) \tag{2}$$

where $\epsilon_i^l = 1$ or 0 if 2D point $\tilde{m}_i^l$ is visible or not in image I_l. Taking the negative logarithm of $p(\chi|T,M)$, we aim at minimizing the proposed *multiple projective points criterion* (MPPC) f:

$$\widehat{T},\widehat{M} = \underset{T,M}{argmin}\, f(T,M;\chi) = \underset{T,M}{argmin}\,(f_{2D}(T,M;\chi) + f_{3D}(T,M;\chi)) \tag{3}$$

with subcriteria similar to squared Mahalanobis distances:

$$f_{2D}(T,M;\chi) = \sum_{l=1}^{S} \sum_{i=1}^{N} \epsilon_i^l \frac{1}{2} \left(m_i^l - \tilde{m}_i^l\right)^T \Theta_{2D}^l \left(m_i^l - \tilde{m}_i^l\right) \tag{4}$$

$$f_{3D}(T,M;\chi) = \frac{1}{2} \sum_{i=1}^{N} \left(M_i - \tilde{M}_i\right)^T \Theta_{3D} \left(M_i - \tilde{M}_i\right) \tag{5}$$

where Θ_{2D}^l and Θ_{3D} are the inverses of the 2D and 3D covariance matrices Σ_{2D}^l and Σ_{3D}. We point out that the proposed MPPC criterion is optimized for all views simultaneously.

To minimize the criterion f, we initialize unknown ideal fiducial positions M_i with the measured positions $\tilde{M}_i$ and the transformations T^l with the coarse transformations T^{l*} resulting from the MoCap analysis. Then the optimization of f is split into two sub-optimizations performed in an iterative interleaved manner until convergence[44]:

- M – optimization: at a given iteration i the current estimates of the transformations $\widehat{T^l}$ are considered as fixed and the positions M_i are optimized and
- T – optimization: in the next iteration $i+1$ the last estimates of positions $\widehat{M}_i$ are kept fixed while the transformations T^l are optimized.

In contrast to Nicolau et al.'s work,[44] we used the Levenberg–Marquardt (LM) optimization algorithm as subcriteria which are expressed as sums of squared residuals terms (see Appendix 3 in Supplementary Material).

2.5 | Accuracy of the gold standard transformations

The works of Sibson[35] and Fitzpatrick et al.[36] were used by most previous studies to compute the expected TRE

in order to estimate the accuracy of the gold-standard datasets. Assuming that the FLE of each fiducial M_i and of each corresponding transformed fiducial $T(M_i)$ is identically and independently distributed (i.i.d) as an isotropic zero-mean Gaussian distribution, by assuming a first-order approximation of the rotation component of the transformation T, Fitzpatrick et al.[36] proposed an estimation of the expected TRE at a target point E_i based on the expected FLE:

$$\langle TRE^2(E_i)\rangle = \frac{\langle FLE^2\rangle}{N}\left(1+\sum_{k=1}^{3}\frac{d_k}{g_k}\right) \tag{6}$$

where g_k is the root mean square (RMS) distance of the projections of the fiducials $M_i$to the kth principal axis of the fiducial configuration, d_k is the RMS distance of E_i projected to the kth principal axis, and $\langle\square\rangle$ indicates the expected value. Sibson[35] showed that under the same assumptions, the expected FLE can be retrieved from the expected FRE as:

$$\langle FLE^2\rangle = \frac{N}{N-2}\langle FRE^2\rangle \tag{7}$$

This TRE, hereafter referred to as *reconstructed* TRE (rTRE), has commonly be used to replace the "*true*" TRE (tTRE) that would be computed if true transformations were available.

However, we found that the conditions to use such rTRE formulation are not met when using reconstructed 3D fiducials by multi-view triangulation of the 2D fiducials. In fact, their distribution was shown to be usually anisotropic,[46–48] and its Gaussianity may be valid only as a local approximation.[49] Furthermore, reconstructed fiducials will present heteroscedastic errors––characterized by inhomogeneous noise.[50] Despite alternative TRE computations were proposed to tackle this more complex noise models,[51,52] most of the works on gold-standard datasets[26,28,34] still used the original rTRE proposed by Fitzpatrick et al.[36]

We investigated an alternative formulation of the TRE for the MPPC method, which takes into account the propagated uncertainty of both 2D and 3D fiducial positions modeled as Gaussian and possibly anisotropic noise. Following Pennec and Thirion,[38] we can state that criterion f (3) will reach a well-defined local minimum $\left(\widehat{T},\widehat{M}\right)=\widehat{q}$ if and only if:

$$\begin{aligned}\Phi(T,M;\chi) &= \left(\frac{\partial f}{\partial q}(T,M;\chi)\right)^{\top}\Bigg|_{\widehat{q};\chi} = 0 \text{ and } H \\ &= \frac{\partial^2 f}{\partial q^2}(T,M;\chi)\Bigg|_{\widehat{q};\chi} \text{ is positive definite}\end{aligned} \tag{8}$$

We can consider the measurement vector χ as a random vector of mean $\overline{\chi}$ and covariance $\Sigma_{\chi\chi}=diag\left(\Sigma_{2D}^1,\cdots,\Sigma_{2D}^S,\Sigma_{3D}\cdots,\Sigma_{3D}\right)$. Using the implicit function theorem and a first-order Taylor series expansion[38] we get:

$$\Sigma_{\widehat{q}\widehat{q}} = H^{-1}\widehat{\left(\frac{\partial\Phi}{\partial\chi}\right)}\Sigma_{\chi\chi}\widehat{\left(\frac{\partial\Phi}{\partial\chi}\right)}^{\top}H^{-\top} \tag{9}$$

By definition, $\Sigma_{\widehat{q}\widehat{q}}=\begin{pmatrix}\Sigma_{\widehat{T}\widehat{T}} & \Sigma_{\widehat{T}\widehat{M}}\\ \Sigma_{\widehat{M}\widehat{T}} & \Sigma_{\widehat{M}\widehat{M}}\end{pmatrix}$, hence we could extract $\Sigma_{\widehat{T}\widehat{T}}$ from $\Sigma_{\widehat{q}\widehat{q}}$ (analytical details are provided in Appendix 4 of the Supplementary Material).

The estimation of the uncertainty of a target position E_i after application of the optimized transformation vector $\widehat{T}$, $\widehat{Y}_i=\widehat{T}E_i=\left(\widehat{T^1}E_i,\cdots,\widehat{T^S}E_i\right)^{\top}$ is obtained by uncertainty propagation:

$$\Sigma_{\widehat{Y}_i\widehat{Y}_i} = \frac{\partial(\widehat{T}E_i)}{\partial\widehat{T}}\Sigma_{\widehat{T}\widehat{T}}\frac{\partial\left(\widehat{T}E_i\right)^{\top}}{\partial\widehat{T}} \tag{10}$$

$$\frac{\partial(TE_i)}{\partial T} = \begin{pmatrix}\frac{\partial(T^1E_i)}{\partial T^1} & 0 & 0\\ 0 & \ddots & 0\\ 0 & 0 & \frac{\partial(T^SE_i)}{\partial T^S}\end{pmatrix}$$

where the expression of $\frac{\partial(T^lE_i)}{\partial T^l}$ is given by Pennec and Thirion.[38] From the same work, we derive an expression of the expected TRE for a target E_i transformed by the computed gold-standard $\widehat{T}$ as the expectation of the squared distance between true and estimated positions of the transformed target:

$$\langle TRE^2(E_i)\rangle = \frac{1}{S}\text{trace}\Sigma_{\widehat{Y}_i\widehat{Y}_i} \tag{11}$$

Given a number of F target points, we finally express the corresponding average TRE, coined hereafter as *uTRE*, as the following RMS:

$$uTRE\left(E_1,\ldots E_F\right)=\sqrt{\frac{1}{F}\sum_{i=1}^{F}\left\langle TRE^2\left(E_i\right)\right\rangle} \tag{12}$$

Compared to the standard rTRE, the uTRE is expected to better account for both noise distributions of 2D and 3D fiducials.

3 | RESULTS

The performance of the proposed MPPC criterion was compared against the iterative PnP algorithm

"solvePnP" of OpenCV,[41] referred to as iterative PnP (cvPnP) approach, in both synthetic and real experiments in presence of 2D and 3D noise. cvPnP minimizes the reprojection error with the LM algorithm, and contrary to the MPPC approach, it optimizes each view independently and does not explicitly model the 3D and 2D noises in the optimization. We used the different formulations of TRE as evaluation metrics for comparison between the proposed MPPC and the cvPnP algorithms:

- "Standard" reconstructed TRE (rTRE) (6),[26,28,34] relying on measured FRE and on the known FLE for synthetic experiments or the estimated FLE for real experiments. Reconstructed points were expressed in the coordinate system CS_r and computed as the closest points to back-projected lines.[26] In case of MPPC, we used the optimized fiducial positions $\hat{M}_i$ instead of the perturbed positions $\tilde{M}_i$ for computation of FRE. The estimated rigid transform between fiducial points and reconstructed points was based on standard least square error minimization.[53]
- Robust reconstructed TRE (*hTRE*)[52] designed to tackle the heteroscedastic and anisotropic errors of the reconstructed fiducials. In this case, we also used a robust rigid transform estimation technique[50] instead of the standard least square approach.
- The proposed uTRE (12) based on uncertainty derivation, only valid for our MPPC approach.
- In synthetic experiments, the true TRE (tTRE) computed as the RMS of the Euclidean distances between target points transformed by the ground truth Λ^l and by the tested transforms T^l:

$$tTRE(E_1, \ldots, E_F; T^1, \cdots, T^S) = \sqrt{\frac{1}{FS}\sum_{l=1}^{S}\sum_{i=1}^{F}\|\Lambda^l E_i - T^l E_i\|^2}$$

Both standard and robust rTRE rely on calculated Euclidean distances in the reconstructed coordinate system CS_r, while for uTRE and tTRE these distances are computed in the CS_l of each view. This discrepancy of CS between TRE formulations prevents the direct comparison of the TRE values. Assuming we know the transformations expressed from CS_r to CS_l, we can calculate additional TREs for these transformations and use the chain rule provided by West and Maurer[54] to get a comparable *"composite" reconstructedTRE*^{c}. The composite TRE can be derived for both standard ($rTRE^c$) and robust ($hTRE^c$) rTRE.

We tested the significance of the difference in paired observations using a paired two-sided t-test if the difference was normally distributed, or a paired two-sided Wilcoxon signed-rank test otherwise. Data normality was checked with a Shapiro–Wilk test. All tests used a confidence level at 99%.

3.1 | Synthetic evaluation of the multiple projective points criterion

We considered the MoCap transforms T^{l*} and the extracted fiducial positions from the CT volume as the ground truths. Both MPPC and cvPnP approaches were initialized with transformations computed using the ML-PnP algorithm of Urban et al.[43] We produced various FLE by perturbing the 3D positions of fiducials and of their 2D reprojections with different σ values of zero-mean Gaussian noises: $\sigma^2_{2D} = \{0.15, 0.29, 0.58, 0.87, 1.16, 1.45\}$ mm for 2D and $\sigma^2_{3D} = \{0.5, 1.0, 2.0\}$ mm for 3D noises. 2D covariance matrices were isotropic ($\Sigma_{2D} = diag(\sigma^2_{2D}, \sigma^2_{2D})$), while for 3D noise we considered both isotropic ($\Sigma_{3D} = diag(\sigma^2_{3D}, \sigma^2_{3D}, \sigma^2_{3D})$) and anisotropic ($\Sigma_{3D} = diag(\sigma^2_{3D}, \sigma^2_{3D}, 1.5\sigma^2_{3D})$) cases. For each configuration of 2D and 3D (anisotropic) noises, we randomly drew 100 samples from the respective distributions––leading to a total of 3600 experiments involving 19 views. Target points E_i were regularly sampled in a 9 × 9 × 9 grid around the hip bones ($F = 729$). Since ground truth transforms were known, we computed the *composite* standard and robust rTREs.

Results averaged over all trials and different noise levels are reported in Table 1. Based on the average values of TREs, we observed that tTRE was statistically different than the corresponding robust or standard composite TREs (p values <0.002), regardless of the chosen approach and of the 3D perturbation isotropy. The only exceptions without statistical difference were the robust composite TREs for the MPPC approach in the isotropic case with the highest level of 3D noise ($\sigma^2_{3D} = 2.0$). The rTRE^c obtained using the MPPC approach proved to be always statistically inferior to the rTRE^c obtained using the cvPnP approach, regardless of the noise levels and 3D noise isotropy. For the tTRE, average values suggested that the MPPC approach generally performed better than the cvPnP (e.g., 2.05 mm vs. 2.38 mm for the isotropic case) although statistical significance was not observed for $(\sigma_{2D}, \sigma_{3D})$ pairs with $\sigma^2_{3D} = 0.5$ mm and $\sigma^2_{2D} = \{0.15, 0.29, 0.58, 0.87\}$ mm in the isotropic case.

All TRE formulations were significantly higher when 3D fiducials were perturbed by anisotropic noise, except for the rTRE^c in the MPPC approach with $\left(\sigma^2_{2D} = 0.58, \sigma^2_{3D} = 1.0\right)$ for which statistical significance was not observed. For the MPPC approach and over all noise levels, the average uTRE was considerably closer to the average tTRE compared to the rTRE^c. When considering the effect of varying 2D and 3D noise levels and 3D noise isotropy (Table 2) tTRE and uTRE were statistically different for some noise configurations $(\sigma_{2D}, \sigma_{3D})$ with large (combined) noise levels. In those cases, the averaged uTRE generally overestimated the tTRE.

TABLE 1 Results of different types of target reconstruction errors (true TRE (tTRE), reconstructed composite TRE (for both standard ($rTRE^c$) and robust ($hTRE^c$) approaches), and uncertainty-based TRE (uTRE)) from the synthetic experiments, averaged over 3600 trials with different 2D and 3D noise levels–3D noise having isotropic and anisotropic variants. An iterative PnP method (cvPnP) was tested against our method using the proposed multiple projective points criterion (MPPC)

Method (iso/anisotropic voxel size)		$rTRE^c$ [mm]	$hTRE^c$ [mm]	tTRE [mm]	uTRE [mm]
cvPnP	isotropic	2.52 ± 1.56	2.48 ± 1.58	2.38 ± 1.61	—
	anisotropic	2.80 ± 1.76	2.75 ± 1.78	2.67 ± 1.81	—
MPPC	isotropic	2.22 ± 1.25	1.99 ± 1.34	2.05 ± 1.30	2.19 ± 0.51
	anisotropic	2.39 ± 1.38	2.13 ± 1.48	2.23 ± 1.44	2.29 ± 0.54

3.2 | Validation of the MPPC-based gold-standard dataset

For the real experiments, we applied both the MPPC and the cvPnP approaches to compute the gold-standard transformations T^I of our dataset. Both algorithms were initialized with the transformations T^{I*} from MoCap. In addition, the MPPC was initialized with the measured 3D positions $\tilde{M}_i$. We set the FLE 2D and 3D covariance matrices Σ^I_{2D} and Σ_{3D} based on the variances of conducted experiments, both expressed in mm:

$$\Sigma_{3D} = \begin{pmatrix} 0.1^2 & 0 & 0 \\ 0 & 0.1^2 & 0 \\ 0 & 0 & 0.127^2 \end{pmatrix} \text{ and } \Sigma^I_{2D} = \begin{pmatrix} 0.19^2 & 0 \\ 0 & 0.19^2 \end{pmatrix}$$

The 3D covariance matrix modeled an anisotropic noise with larger variance in the Z-direction, which is common for medical imaging datasets with a lower resolution in the slice stacking direction in order to save acquisition time, improve signal-to-noise ratio, or reduce dose exposure. We set the 2D and 3D covariance matrices to have equal variances in X- and Y-directions because computed variances in the experiments were quasi-identical and it was reasonable to assume that noise would not be especially biased for any of the X- or Y-direction. In order to assess the accuracy of the MoCap setup, we also considered the initializations T^{I*} as the result of an approach to compute ground truth transformations, denoted as the "MoCap" method.

We tested the three approaches MPPC, cvPnP, and MoCap with different numbers of views: 2 (acquired in anteroposterior (AP) and quasilateral (LAT) positions), 9 (mimicking at best the angles of the work of Tomaževič et al.[26]), and all the 19 views. For the cvPnP and MoCap approaches, the number of views did not have any impact on the computation of the transformations T^I, but it will impact the results of the following evaluation metrics.

For comparison purposes with previous works, 2D metrics for the evaluation of the accuracy of the ground truths included the mean (mPD) and RMS (rmsPD) projection distance errors, as well as the standard rTRE since the robust variant was not used in these works. For the MPPC approach, we computed the values of the metrics using both measured positions $\tilde{M}_i$ and optimized positions $\hat{M}_i$ (denoted as the "non-noisy" case). The uncertainty-based TRE uTRE (12) was only computed for the non-noisy MPPC. It is worth noting that the formulations of the true TRE (tTRE) and of the composite TREs ($rTRE^c$ and $hTRE^c$) used in previous experiment could not be used as the ground truth transformations were unknown. For computation of the TREs, we defined 12 target points located at key anatomical landmarks such as the trochanters, hip joint centers, or the anterior superior iliac spines. Results are summarized in Table 3 and Figure 4.

When using non-optimized 3D fiducial positions $\tilde{M}_i$, values of mPD, rmsPD, and rTRE obtained with the MPPC method were similar to those from the cvPnP method, for all x-ray views. However, when the accuracy evaluation metrics were computed using the MPPC method with optimized 3D fiducial positions $\hat{M}_i$, the rTRE was statistically smaller than the rTREs of the best cvPnP and MoCap results using 19 views (rTRE = 0.11 and 0.34 mm), regardless of the number of views used for the MPPC, for which the rTRE ranged from 0.05 to 0.15 mm.

Increasing the number of x-ray views improved the rTRE, especially for the MPPC using optimized 3D fiducials for which rTRE decreased by 60% from 2 to 19 views. The MoCap method performed poorly compared to cvPnP and MPPC variants, with statistically higher evaluation metrics (both average and standard deviations), regardless of the number of views.

Like in the previous synthetic experiments, the computed uTRE was considerably higher than the rTRE, regardless the number of views.

4 | DISCUSSION

4.1 | A new gold-standard 2D-3D registration dataset for the hip joint

We proposed a public dataset for the validation of CT-to-x-ray 2D-3D registration of the hip joint that consists in 19 real fluoroscopic images of a female hip phantom, acquired at different x-ray voltages and different

TABLE 2 Comparison of the difference between true TRE (tTRE) and uncertainty-based TRE (uTRE) for our multiple projective points criterion (MPPC) method in synthetic experiments which included 3600 trials with varying 2D and 3D Gaussian noise levels (σ_{2D}, σ_{3D}): σ^2_{2D} *varied* from 0.15 to 1.45 mm and σ^2_{3D} from 0.5 to 2.0 mm (with isotropic and anisotropic variants of the 3D covariance matrix)

	tTRE—uTRE (isotropic) [mm]						tTRE—uTRE (anisotropic) [mm]					
	0.15	**0.29**	**0.58**	**0.87**	**1.16**	**1.45**	**0.15**	**0.29**	**0.58**	**0.87**	**1.16**	**1.45**
0.5	0.17	0.13	0.04	−0.04	−0.12	−0.21*	0.26	0.21	0.08	−0.08	−0.20	−0.33*
1.0	−0.00	−0.08	0.19	−0.27*	−0.34*	−0.40*	0.10	0.02	−0.12	−0.23	−0.31*	−0.40*
2.0	−0.06	−0.11	−0.19	−0.25	−0.31	−0.36*	0.18	0.11	−0.00	−0.10	−0.19	−0.27

Asterix* highlights a statistically significant difference.

phantom orientations with large rotations. This dataset is useful for standardized evaluation of the registration accuracy in orthopedic applications. Markelj et al.[30] generated a validation dataset including the human pelvis based on DRRs, which were however not fully realistic due to the absence of noise introduced by the imaging device and due to the discrete nature of the projected CT volumetric image. In the present work, the quality and the field of view of the fluoroscopic images were matched to those of typical in vivo acquisitions, such as fluoroscopy-based analyses of the hip joint during motion, for which the required voltage varies depending on the body mass index of the patient, the target hip is not always centered in the image and frequent overexposed areas present saturation of the pixel intensity. Moreover, the different poses of the phantom reproduce the varying irradiation angles used in the clinical practice depending on the instrumental setup (i.e., single-plane vs. dual-plane fluoroscopy), on the measured activity and subject, and on the limits in delivered radiation exposure.[55] The dependence of the performance of a registration algorithm on the x-ray voltage can also be investigated with the proposed dataset.

Still, virtual radiographs as proposed by Markelj et al.[30] are of interest since the ground truth transformations are exact, so we decided to also include synthetic radiographs to our dataset. We modified the DeepDRR approach[32] (e.g., modeling of the detector response to x-ray fluence, use of post-filtering such as adaptive histogram equalization) and used a higher resolution CT (dimensions 431 × 315 × 1418 and voxel size 0.78 × 0.78 × 0.33 mm^3) in order to produce better virtual radiographs. As shown in Figure 5, the DeepDRR approach generally produced quite convincing radiographs, but some artifacts were visible (e.g., vertical stripes, grainy areas) and the realism of the scattering effect (e.g., the sacrum is too visible in the virtual radiographs) or the overexposed effect could not be really reproduced.

Another novel aspect of the present work was the use of MoCap for automatic definition of the 3D-2D fiducial marker correspondences required for point-based registration to obtain the ground truth transformation. However, this technique requires motion capture equipment and may not be suited for intra-operative validations. This was the focus of the work from Madan et al.,[34] which proposed a method for fully automatic marker extraction and identification for point-based registration during endovascular image-guided interventions. While the accuracy of the transformations computed with MoCap was sufficient to establish point correspondence, it was considerably lower than the accuracy from the other tested methods. Hence, MoCap may not be suited to build an accurate gold-standard validation dataset. In fact, the accuracy of 3D point computation of MoCap systems

TABLE 3 Evaluation metrics of the accuracy of the ground truth transformations obtained with an iterative PnP method (cvPnP), with optical motion capture (MoCap), and with the proposed multiple projective points criterion (MPPC). Metrics for cvPnP and MoCap were computed based on 3D fiducial positions $\tilde{M}_i$ only, while metrics for MPPC were computed with both $\tilde{M}_i$ (MPPC) and the optimized 3D fiducial positions M_i (MPPC non-noisy)

Method (# views)		mPD [mm]	rmsPD [mm]	FRE [mm]	FLE [mm]	rTRE [mm]	uTRE [mm]
MoCap	2 views	0.76 ± 0.32	0.82	1.05	1.19	0.78	—
	9 views	0.97 ± 0.50	1.08	0.90	0.94	0.35	—
	19 views	0.93 ± 0.46	1.04	0.87	0.91	0.34	—
cvPnP	2 views	0.22 ± 0.11	0.25	0.27	0.31	0.20	—
	9 views	0.27 ± 0.12	0.30	0.28	0.30	0.11	—
	19 views	0.25 ± 0.12	0.28	0.27	0.28	0.11	—
MPPC	2 views	0.22 ± 0.11	0.25	0.26	0.30	0.20	—
	9 views	0.29 ± 0.14	0.32	0.31	0.33	0.12	—
	19 views	0.29 ± 0.14	0.32	0.32	0.34	0.12	—
MPCC (non-noisy)	2 views	0.15 ± 0.07	0.17	0.20	0.22	0.15	0.65
	9 views	0.14 ± 0.06	0.16	0.16	0.17	0.06	0.61
	19 views	0.12 ± 0.06	0.13	0.13	0.14	0.05	0.59

Abbreviations: FLE, fiducial localization error; FRE, fiducial registration error; mPD, mean reprojection distance; rmsPD, root mean square projection distance; rTRE, standard reconstructed target registration error; uTRE, target registration error based on uncertainty theory.

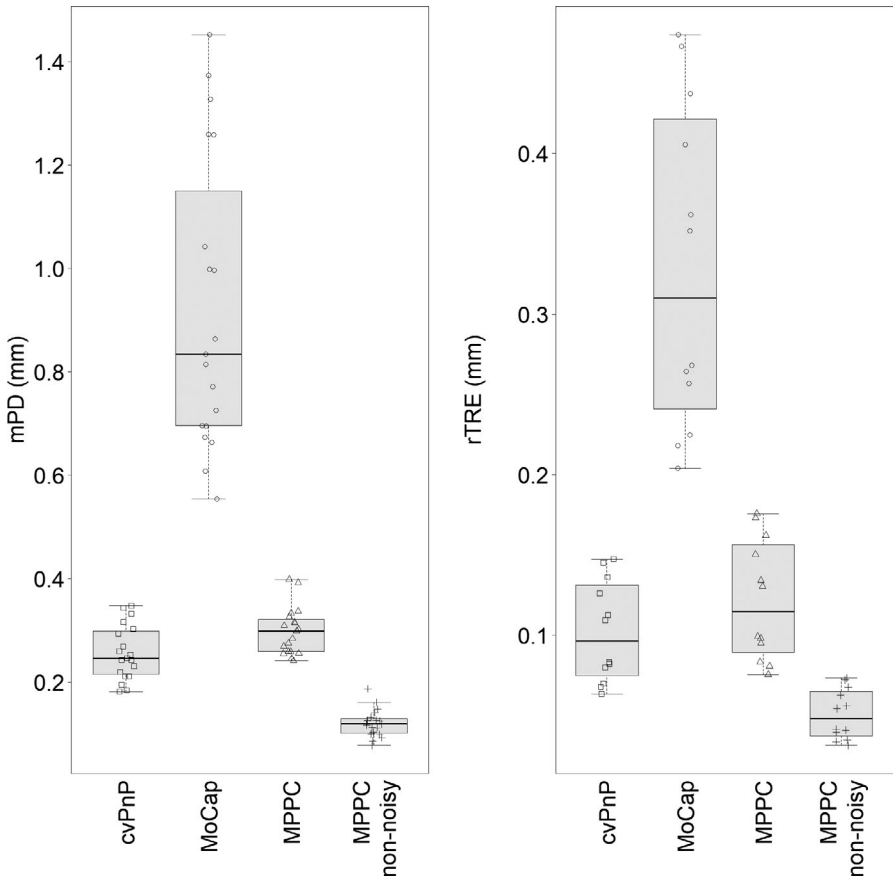

FIGURE 4 Box plot for evaluation metrics of the accuracy of the ground truth transformations retrieved with an iterative PnP method (cvPnP), with optical motion capture (MoCap), and with the multiple projective points criterion (MPPC) using measured 3D fiducial positions (MPPC) and optimized ones (MPPC non-noisy)

depends on several factors such as coverage, number and type of cameras, and static or dynamic setup.[39] For instance, previous works[39,56,57] reported a 95th percentile error (mean + 2 * std) ranging in [0.073, 6.7] mm.

4.2 | A validated dataset based on the multiple projective points criterion (MPPC)

The proposed MPPC method for computation of the gold-standard transformations a) modeled the noise of both

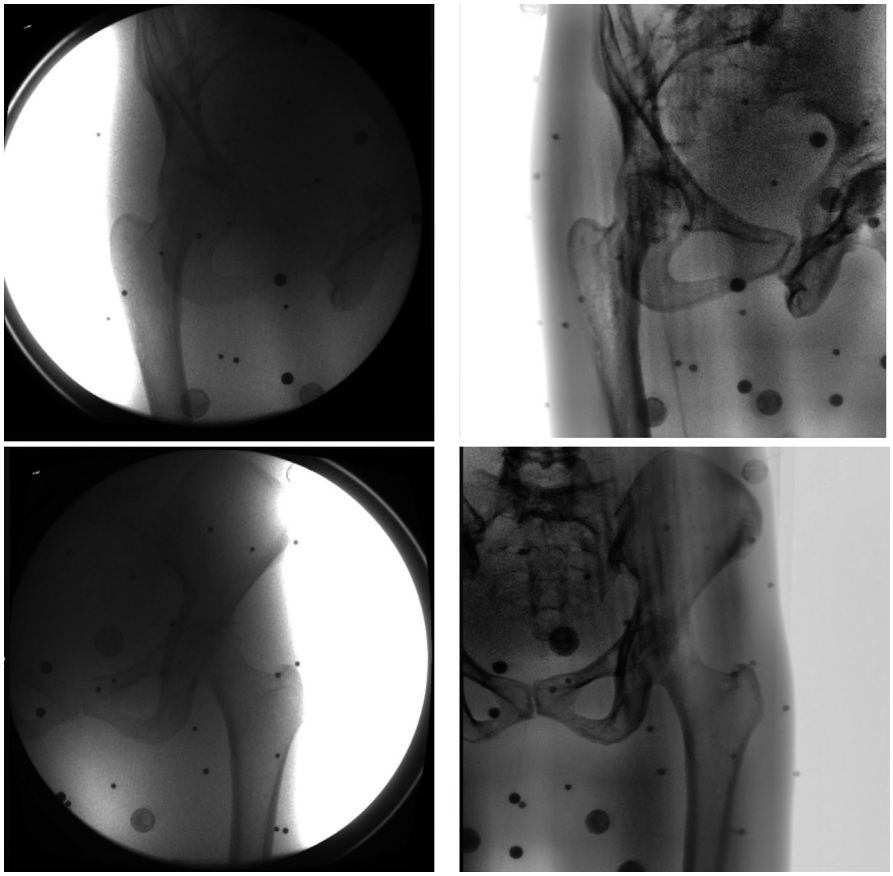

FIGURE 5 Comparison of fluoroscopic images (left column) versus synthetic DRR images (right column) generated with the DeepDRR[32] approach

corresponding 2D and 3D fiducials as identically and independently distributed zero-mean Gaussian noise, with a tunable degree of anisotropy and b) optimized the noisy 3D fiducial locations together with the transformations, including all X-ray views into a single optimization. In contrast, previous studies[26,28,34] did not model 3D fiducial errors and mainly optimized the unknown transformations, while assuming zero-mean Gaussian noise of 2D fiducial positions. Optimizing for the 3D fiducials together with the transformations had significant effects on the rTRE when evaluated with these optimized 3D fiducial locations. The rTRE was significantly smaller compared to the one evaluated from both MPPC and the iterative cvPnP method with measured 3D fiducial locations, regardless the number of x-ray views.

We observed that in both 2D and 3D experiments the computed error distributions did not follow a Gaussian distribution according to univariate (Shapiro–Wilk test) and multivariate (Mardia's test) normality tests.[58] However, the proposed MPPC performed better than other methods, despite the assumption of Gaussianity. The assumption of 2D zero-mean Gaussian distribution is commonly used for approaches minimizing the reprojection error in a least mean squares sense,[28] although it is often not formally verified.

As observed in previous works,[27,59] the accuracy metrics of the MPPC improved when increasing the number of x-ray views. For instance, an increase from 2 to 19 x-ray views slightly improved the mPD from 0.15 to 0.12 mm while the rTRE decreased from 0.15 to 0.05 mm.

Results obtained with our dataset were comparable and sometimes superior to previous published works. However, comparison with other fiducial-based validation datasets should be performed with caution due to the different types of fiducials, anatomy, target points as well as the quality of x-ray and volumetric images. Pawiro et al.[28] generated gold-standard CT-to-x-ray transformations from two views of a fresh porcine cadaver head, by minimizing the mPD in 2D. Their best values for FRE, rTRE, and mPD were 0.22, 0.17, and 0.51 mm, respectively, which are similar to the values of the present dataset for two views, but higher than those for 19 views. Tomazevic et al.[26] generated a CT-to-x-ray validation dataset for the lumbar spine by minimizing the FRE for nine views, and reported rTRE less than 0.26 mm, which is higher than the largest rTRE for our dataset using two views. Mitrovic et al.[60] and Madan et al.[34] produced the first gold-standard datasets based on pairs of clinical images, including 3D contrast-enhanced cone beam CT and 2D angiograms of 20 patients. Using only two quasi-orthogonal views, they both achieved better accuracy than the one herein reported for 19 views (FRE between 0.038 and 0.060 mm, rTRE between 0.033 and 0.056 for Mitrovic et al.[60]; FRE = 0.017 mm, rTRE smaller than 0.027 mm

for Madan et al.[34]). Their improved performance may originate from the better quality of the medical images, and to a possibly more accurate fiducial position extraction technique. Lastly, while Grupp et al.[33,61] proposed the first 2D-3D gold-standard dataset for the hip, no detailed information was given on quantitative performance of the registration, especially with the TRE. Similar to previously published gold-standard datasets, a limitation of our dataset is that it cannot encompass all anatomical and pathological variations across individuals. As a result, the performance of any algorithm validated with it cannot be deemed as representative of its general performance in clinical practice. However, as we previously mentioned, the purpose of 2D-3D datasets is to provide an objective way to benchmark 2D-3D algorithms.

4.3 | The need to account for data uncertainty

Our extraction of 2D and 3D fiducial positions generated fairly isotropic noise, except for the slice stacking direction of the CT scan which showed larger deviation values. Synthetic experiments showed an expected worsening of both reconstructed and true TREs when anisotropy affected the 3D positions of the fiducials. Hence, our analysis showed that the assumption of anisotropy shall be enforced, usually by considering the direction of the medical datasets with lower spatial resolution, commonly observed in clinical practices (e.g., CT[62,63] and MRI[64,65]). The proposed MPPC provides the mathematical framework to model uncertainty in both 2D and 3D fiducial locations, and we highlighted the superiority of the MPPC in considering noise to derive optimized transformations *and* 3D fiducial positions.

It can be argued that the smaller values of rTRE for the MPPC stems only from the optimization of the 3D fiducial positions and the use of these optimized values in the rTRE metric. In fact, when using the measured 3D fiducial positions instead of the optimized ones, the MPPC yielded accuracy metrics with our dataset that could not be statistically distinguished from those of the cvPnP approach. However, in our opinion the issue is that the accuracy metric of the rTRE may not be representative of the true accuracy of a gold-standard dataset.

In fact, synthetic experiments showed that the rTRE was significantly different than the true TRE. While the robust variant of the rTRE produced values closer to the true TRE than the standard rTRE by accounting for heteroscedastic errors and using a robust registration approach, statistical differences were still observed–with the only exception of the MPPC approach in the isotropic case with the highest level of 3D noise $\sigma_{3D}^2 = 2.0$. We suspect that this surprising result stems from the non-Gaussian distribution of the 3D reconstructed fiducials based on triangulation that may hinder the performance of the robust approaches designed with the assumption of Gaussianity. In contrast, the proposed TRE based on uncertainty analysis (uTRE) provided a better approximation of the true TRE in both isotropic and anisotropic synthetic cases, and it only overestimated the tTRE for large values of 3D and 2D noises, which should not be observed in practice. We think that observed differences between uTRE and true TRE in the synthetic experiments may be the result of approximations such as first-order Taylor series truncation. Since in the real experiments the uTRE was considerably different than the reconstructed rTRE, we suggest that the rTRE may not reflect the accuracy of a gold standard dataset and that reported rTREs may have to be considered with caution, especially when the error in fiducial extraction was not reported or was not negligible. While the uTRE provides more realistic results than the rTRE due to the inclusion of (anisotropic) Gaussian noise in the 3D fiducials as proven by the synthetic experiments, it may still not be fully realistic in our gold standard dataset due to the non-Gaussianity observed for both 2D and 3D fiducials.

A limitation of the proposed evaluation methodology was that the proposed MPPC method did not take into account of the propagation of possible errors in the estimated intrinsic camera parameters. We could take inspiration from previous works jointly minimizing the intrinsic and extrinsic parameters,[26,28,34] but the derivation of a new criterion based on an uncertainty analysis will be more much complex. Furthermore, noises on 2D and 3D points were assumed Gaussian but were not measured as such and were also considered without bias. Indeed, our 3D extraction technique applied to the synthetic datasets yielded an average bias of 0.013 mm, which we considered negligible with respect to the average voxel size of 0.7 mm. The 2D technique had however a higher bias of 0.15 mm compared to a 0.29 mm pixel size, which should be further improved with a better designed 2D extraction technique. More complex noise distribution (e.g., Gaussian noise with bias studied by Moghari and Abolmaesumi[66]) could be hence investigated in future research.

5 | CONCLUSIONS

We proposed the first publicly available dataset for the standardized validation of 2D-3D registration of the hip joint based on real fluoroscopic images presenting large rotation angles. Our dataset is a perfect complement to the recently released public dataset of the hip joint[33] in which fluoroscopic images presented slight rotations–hindering the study of multiview 2D-3D reconstruction. In addition to the new anatomical target, the present paper introduces novel aspects in both

computation of the gold-standard transformations and the evaluation of their accuracy based on uncertainty analysis. We presented approaches to extract the positions of 2D and 3D fiducials from x-ray and volumetric images. The uncertainty in the measured 2D and 3D fiducials was modeled as independently and identically distributed zero-mean isotropic and anisotropic Gaussian noises. This uncertainty was used to derive a new iterative PnP criterion (MPPC) that computes the ground truth transformations by optimizing the noisy 3D fiducial positions as well. The proposed MPPC exhibited good performance in both synthetic and real experiments. Furthermore, a new target reconstruction error (uTRE) was formulated, which included the uncertainty in the extraction of the 2D and 3D fiducials and anisotropy. Failing at including such uncertainties may provide incorrect estimation of the accuracy of a gold standard dataset. We demonstrated the utility of MPPC algorithm for the estimation and assessment of 2D-3D transformations for gold-standard datasets. The proposed algorithm could also be used intraoperatively to put into correspondence pre- and intra-operative data–while obtaining an estimation of the resulting uncertainty.

ACKNOWLEDGMENTS

This work was supported by the European Union Seventh Framework Programme FP7-NMP-2012 LARGE-6, “LifeLongJoints” (grant number 310477), as well as the InnoSuisse project “MyPlanner” (grant number 25258.1 PFLS-LS).

The authors express their sincere gratitude to Yabin Wu for his help in the measurement, and to Steffen Urban for discussions on the ML-PnP approach.

CONFLICT OF INTEREST

The authors have no conflict of interest to report.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in the repository “2D-3D registration gold-standard dataset for the hip joint based on uncertainty modeling” at https://doi.org/10.26037/yaret a:2kw5s2jadbbv3ngtkzsovihtnu.

REFERENCES

1. Markelj P, Tomaževič D, Likar B, Pernuš F. A review of 3D/2D registration methods for image-guided interventions. *Med Image Anal*. 2012;16(3):642-661. https://doi.org/10.1016/j.media.2010.03.005
2. Penney GP, Batchelor PG, Hill DLG, Hawkes DJ, Weese J. Validation of a two- to three-dimensional registration algorithm for aligning preoperative CT images and intraoperative fluoroscopy images. *Med Phys*. 2001;28(6):1024-1032. https://doi.org/10.1118/1.1373400
3. Tomazevic D, Likar B, Pernus F. 3-D/2-D registration by integrating 2-D information in 3-D. *IEEE Trans Med Imaging*. 2006;25(1):17-27. https://doi.org/10.1109/TMI.2005.859715
4. Markelj P, Tomazevic D, Pernus F, Likar B. Robust gradient-based 3-D/2-D registration of CT and MR to X-ray images. *IEEE Trans Med Imaging*. 2008;27(12):1704-1714. https://doi.org/10.1109/TMI.2008.923984
5. Guéziec A, Wu K, Kalvin A, Williamson B, Kazanzides P, Van Vorhis R. Providing visual information to validate 2-D to 3-D registration. *Med Image Anal*. 2000;4(4):357-374. https://doi.org/10.1016/S1361-8415(00)00029-3
6. Sundaram R, Cohen D, Barton-Hanson N. Tibial plateau fracture following gracilis-semitendinosus anterior cruciate ligament reconstruction: the tibial tunnel stress-riser. *Knee*. 2006;13(3):238-240.
7. Dong X, Gonzalez Ballester M, Zheng G. Automatic Extraction of Femur Contours from Calibrated Fluoroscopic Images. In: *2007 IEEE Workshop on Applications of Computer Vision (WACV '07)*. IEEE; 2007:55. https://doi.org/10.1109/WACV.2007.15
8. Ellis RE, Peters TM, eds. *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2003: 6th International Conference, Montréal, Canada, November 15-18, 2003. Proceedings*. Vol 2879. Springer; 2003. https://doi.org/10.1007/b93811
9. Penney GP, Edwards PJ, Hipwell JH, Slomczykowski M, Revie I, Hawkes DJ. Postoperative calculation of acetabular cup position using 2-D–3-D registration. *IEEE Trans Biomed Eng*. 2007;54(7):1342-1348. https://doi.org/10.1109/TBME.2007.890737
10. LaRose D, Cassenti L, Jaramaz B, Moody J, Kanade T, DiGioia A. Post-operative measurement of acetabular cup position using X-ray/CT registration. In: Delp SL, DiGoia AM, Jaramaz B, eds. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2000*. Vol 1935. Lecture Notes in Computer Science. Springer; 2000:1104-1113. https://doi.org/10.1007/978-3-540-40899-4_115
11. Benameur S, Mignotte M, Labelle H, DeGuise JA. A hierarchical statistical modeling approach for the unsupervised 3-D biplanar reconstruction of the scoliotic spine. *IEEE Trans Biomed Eng*. 2005;52(12):2041-2057. https://doi.org/10.1109/TBME.2005.857665
12. Benameur S, Mignotte M, Destrempes F, DeGuise JA. Three-dimensional biplanar reconstruction of scoliotic rib cage using the estimation of a mixture of probabilistic prior models. *IEEE Trans Biomed Eng*. 2005;52(10):1713-1728. https://doi.org/10.1109/TBME.2005.855717
13. Dennis DA, Mahfouz MR, Komistek RD, Hoff W. In vivo determination of normal and anterior cruciate ligament-deficient knee kinematics. *J Biomech*. 2005;38(2):241-253. https://doi.org/10.1016/j.jbiomech.2004.02.042
14. Tang T. Accurate assessment of patellar tracking using fiducial and intensity-based fluoroscopic techniques. *Med Image Anal*. 2004;8(3):343-351. https://doi.org/10.1016/j.media.2004.06.011
15. You B-M, Siy P, Anderst W, Tashman S. In vivo measurement of 3-D skeletal kinematics from sequences of biplane radiographs: application to knee kinematics. *IEEE Trans Med Imaging*. 2001;20(6):514-525. https://doi.org/10.1109/42.929617
16. Fiorentino NM, Atkins PR, Kutschke MJ, Goebel JM, Foreman KB, Anderson AE. Soft tissue artifact causes significant errors in the calculation of joint angles and range of motion at the hip. *Gait Posture*. 2017;55:184-190. https://doi.org/10.1016/j.gaitpost.2017.03.033
17. Sato T, Tanino H, Nishida Y, Ito H, Matsuno T, Banks SA. Dynamic femoral head translations in dysplastic hips. *Clin Biomech*. 2017;46:40-45. https://doi.org/10.1016/j.clinbiomech.2017.05.003
18. Ward TR, Hussain MM, Pickering M, et al. Validation of a method to measure three-dimensional hip joint kinematics in subjects with femoroacetabular impingement. *HIP Int*. Published online October 17, 2019:112070001988354. https://doi.org/10.1177/1120700019883548

19. Banks SA, Hodge WA. Accurate measurement of three-dimensional knee replacement kinematics using single-plane fluoroscopy. *IEEE Trans Biomed Eng*. 1996;43(6):638-649. https://doi.org/10.1109/10.495283
20. Mahfouz MR, Hoff WA, Komistek RD, Dennis DA. A robust method for registration of three-dimensional knee implant models to two-dimensional fluoroscopy images. *IEEE Trans Med Imaging*. 2003;22(12):1561-1574. https://doi.org/10.1109/TMI.2003.820027
21. Taylor WR, Schütz P, Bergmann G, et al. A comprehensive assessment of the musculoskeletal system: the CAMS-Knee data set. *J Biomech*. 2017;65:32-39. https://doi.org/10.1016/j.jbiomech.2017.09.022
22. Lombardi AV, Mallory TH, Dennis DA, Komistek RD, Fada RA, Northcut EJ. An in vivo determination of total hip arthroplasty pistoning during activity. *J Arthroplasty*. 2000;15(6):702-709. https://doi.org/10.1054/arth.2000.6637
23. Glaser D, Dennis DA, Komistek RD, Miner TM. In vivo comparison of hip mechanics for minimally invasive versus traditional total hip arthroplasty. *Clin Biomech*. 2008;23(2):127-134. https://doi.org/10.1016/j.clinbiomech.2007.09.015
24. Tsai T-Y, Li J-S, Wang S, Scarborough D, Kwon Y-M. In-vivo 6 degrees-of-freedom kinematics of metal-on-polyethylene total hip arthroplasty during gait. *J Biomech*. 2014;47(7):1572-1576. https://doi.org/10.1016/j.jbiomech.2014.03.012
25. Jannin P, Fitzpatrick JM, Hawkes DJ, Pennec X, Shahidl R, Vannier MW. Validation of medical image processing in image-guided therapy. *IEEE Trans Med Imaging*. 2002;21(12):1445-1449. https://doi.org/10.1109/TMI.2002.806568
26. Tomaževič D, Likar B, Pernuš F. Gold standard" data for evaluation and comparison of 3D/2D registration methods. *Computer Aided Surgery*. 2004;9(4):137-144. https://doi.org/10.3109/10929080500097687
27. van de Kraats EB, Penney GP, Tomazevic D, van Walsum T, Niessen WJ. Standardized evaluation methodology for 2-D-3-D registration. *IEEE Trans Med Imaging*. 2005;24(9):1177-1189. https://doi.org/10.1109/TMI.2005.853240
28. Pawiro SA, Markelj P, Pernuš F, et al. Validation for 2D/3D registration I: a new gold standard data set: Validation for 2D/3D registration. Part I. *Med Phys*. 2011;38(3):1481-1490. https://doi.org/10.1118/1.3553402
29. Xia W, Jin Q, Ni C, Wang Y, Gao X. Thorax x-ray and CT interventional dataset for nonrigid 2D/3D image registration evaluation. *Med Phys*. 2018;45(11):5343-5351.
30. Markelj P, Likar B, Pernuš F. Standardized evaluation methodology for 3D/2D registration based on the Visible Human data set: standardized evaluation methodology for 3D/2D registration. *Med Phys*. 2010;37(9):4643-4647. https://doi.org/10.1118/1.3476414
31. Moore CS, Liney GP, Beavis AW, Saunderson JR. A method to produce and validate a digitally reconstructed radiograph-based computer simulation for optimisation of chest radiographs acquired with a computed radiography imaging system. *BJR*. 2011;84(1006):890-902. https://doi.org/10.1259/bjr/30125639
32. Unberath M, Zaech J-N, Lee SC, et al. DeepDRR – a catalyst for machine learning in fluoroscopy-guided procedures. In: Frangi AF, Schnabel JA, Davatzikos C, Alberola-López C, Fichtinger G, eds. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*. Vol 11073. Lecture Notes in Computer Science. Springer International Publishing; 2018:98-106. https://doi.org/10.1007/978-3-030-00937-3_12
33. Grupp RB, Unberath M, Gao C, et al. Automatic annotation of hip anatomy in fluoroscopy for robust and efficient 2D/3D registration. *Int J CARS*. 2020;15(5):759-769. https://doi.org/10.1007/s11548-020-02162-7
34. Madan H, Pernuš F, Likar B, Špiclin Ž. A framework for automatic creation of gold-standard rigid 3D–2D registration datasets. *Int J CARS*. 2017;12(2):263-275. https://doi.org/10.1007/s11548-016-1482-4
35. Sibson R, Bowyer A, Osmond C. Studies in the robustness of multidimensional scaling: euclidean models and simulation studies. *J Stat Comput Simul*. 1981;13(3–4):273-296. https://doi.org/10.1080/00949658108810502
36. Fitzpatrick JM, West JB, Maurer CR. Predicting error in rigid-body point-based registration. *IEEE Trans Med Imaging*. 1998;17(5):694-702. https://doi.org/10.1109/42.736021
37. Moghari MH, Ma B, Abolmaesumi P. A theoretical comparison of different target registration error estimators. In: Metaxas D, Axel L, Fichtinger G, Székely G, eds. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2008*. Springer; 2008:1032-1040.
38. Pennec X, Thirion J-P. A framework for uncertainty and validation of 3-d registration methods based on points and frames. *Int J Comput Vision*. 1997;25(3):203-229. https://doi.org/10.1023/A:1007976002485
39. van der Kruk E, Reijne MM. Accuracy of human motion capture systems for sport applications; state-of-the-art review. *Eur J Sport Sci*. 2018;18(6):806-819. https://doi.org/10.1080/17461391.2018.1463397
40. Schmid J, Kim J, Magnenat-Thalmann N. Robust statistical shape models for MRI bone segmentation in presence of small field of view. *Med Image Anal*. 2011;15(1):155-168. https://doi.org/10.1016/j.media.2010.09.001
41. Bradski G. The opencv library. *Dr Dobb's J Software Tools*. 2000;25:120-125.
42. Ferraz Colomina L, Binefa X, Moreno-Noguer F. Leveraging feature uncertainty in the pnp problem. In: *Proceedings of the BMVC 2014 British Machine Vision Conference*. BMVA Press; 2014:1-13.
43. Urban S, Leitloff J, Hinz S. MLPnP - a real-time maximum likelihood solution to the perspective-N-point problem. In: *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences*. Vol 3. Copernicus Publications; 2016:131-138.
44. Nicolau S, Pennec X, Soler L, Ayache N. Evaluation of a new 3D/2D registration criterion for liver radio-frequencies guided by augmented reality. In: Ayache N, Delingette H, eds. *Surgery Simulation and Soft Tissue Modeling*. Vol 2673. Lecture Notes in Computer Science. Springer; 2003:270-283. https://doi.org/10.1007/3-540-45015-7_26.
45. Nicolau S, Pennec X, Soler L, Ayache N. *Validation of a New 3D/2D Registration Criterion Including Error Prediction. Application to Image Guided Radio-Frequency Ablation of the Liver Tumors*. INRIA; 2003. https://hal.inria.fr/inria-00071585.
46. Rumpler M, Irschara A, Bischof H. Multi-view stereo: redundancy benefits for 3D reconstruction. In: *35th Workshop of the Austrian Association for Pattern Recognition*. Vol 4. OAGM; 2011.
47. Freundlich C, Zavlanos M, Mordohai P. Exact bias correction and covariance estimation for stereo vision. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE; 2015:3296-3304. https://doi.org/10.1109/CVPR.2015.7298950
48. Beder C, Steffen R. Determining an initial image pair for fixing the scale of a 3D reconstruction from an image sequence. In: Franke K, Müller K-R, Nickolay B, Schäfer R, eds. *Pattern Recognition*. Vol 4174. Lecture Notes in Computer Science. Springer; 2006:657-666. https://doi.org/10.1007/11861898_66
49. Olague G, Mohr R. Optimal camera placement for accurate reconstruction. *Pattern Recogn*. 2002;35(4):927-944. https://doi.org/10.1016/S0031-3203(01)00076-0
50. Matei B, Meer P. Optimal rigid motion estimation and performance evaluation with bootstrap. In: *Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149)*. IEEE Comput. Soc; 1999:339-345. https://doi.org/10.1109/CVPR.1999.786961

51. Ma B, Moghari MH, Ellis RE, Abolmaesumi P. Estimation of optimal fiducial target registration error in the presence of heteroscedastic noise. *IEEE Trans Med Imaging*. 2010;29(3):708-723. https://doi.org/10.1109/TMI.2009.2034296
52. Danilchenko A, Fitzpatrick JM. General approach to first-order error prediction in rigid point registration. *IEEE Trans Med Imaging*. 2011;30(3):679-693. https://doi.org/10.1109/TMI.2010.2091513
53. Umeyama S. Least-squares estimation of transformation parameters between two point patterns. *IEEE Trans Pattern Anal Machine Intell*. 1991;13(4):376-380. https://doi.org/10.1109/34.88573
54. West JB, Maurer CR Jr. Extension of target registration error theory to the composition of transforms. In: Sonka M, Fitzpatrick JM, eds. *Proceedings SPIE Medical Imaging 2002: Image Processing*. Vol. 4684. International Society for Optics and Photonics; 2002:574–580. https://doi.org/10.1117/12.467200
55. D'Isidoro F, Eschle P, Zumbrunn T, Sommer C, Scheidegger S, Ferguson SJ. Determining 3D kinematics of the hip using video fluoroscopy: guidelines for balancing radiation dose and registration accuracy. *J Arthroplasty*. 2017;32(10):3213-3218. https://doi.org/10.1016/j.arth.2017.05.036
56. Eichelberger P, Ferraro M, Minder U, et al. Analysis of accuracy in optical motion capture – a protocol for laboratory setup evaluation. *J Biomech*. 2016;49(10):2085-2088. https://doi.org/10.1016/j.jbiomech.2016.05.007
57. Merriaux P, Dupuis Y, Boutteau R, Vasseur P, Savatier X. A study of Vicon system positioning performance. *Sensors*. 2017;17(7):1591. https://doi.org/10.3390/s17071591
58. Korkmaz S, Goksuluk D, Zararsiz G. MVN: an R package for assessing multivariate normality. *R Journal*. 2014;6:151-162.
59. Tomaževič D, Likar B, Pernuš F. 3D/2D image registration: the impact of x-ray views and their number. In: Ayache N, Ourselin S, Maeder A, eds. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2007*. Vol 4791. Lecture Notes in Computer Science. Springer; 2007:450-457. https://doi.org/10.1007/978-3-540-75757-3_55
60. Mitrovic U, Spiclin Z, Likar B, Pernus F. 3D–2D registration of cerebral angiograms: a method and evaluation on clinical images. *IEEE Trans Med Imaging*. 2013;32(8):1550-1563. https://doi.org/10.1109/TMI.2013.2259844
61. Grupp RB, Hegeman RA, Murphy RJ, et al. Pose estimation of periacetabular osteotomy fragments with intraoperative x-ray navigation. *IEEE Trans Biomed Eng*. 2020;67(2):441-452. https://doi.org/10.1109/TBME.2019.2915165
62. Dalrymple NC, Prasad SR, El-Merhi FM, Chintapalli KN. Price of isotropy in multidetector CT. *Radiographics*. 2007;27(1):49-62.
63. Shafiq-ul-Hassan M, Zhang GG, Latifi K, et al. Intrinsic dependencies of CT radiomic features on voxel size and number of gray levels. *Med Phys*. 2017;44(3):1050-1062.
64. Van Hecke W, Leemans A, De Backer S, Jeurissen B, Parizel PM, Sijbers J. Comparing isotropic and anisotropic smoothing for voxel-based DTI analyses: a simulation study. *Hum Brain Mapp*. 2010;31(1):98-114.
65. Mulder MJ, Keuken MC, Bazin P-L, Alkemade A, Forstmann BU. Size and shape matter: the impact of voxel geometry on the identification of small nuclei. Bergsland N, ed. *PLoS ONE*. 2019;14(4):e0215382. https://doi.org/10.1371/journal.pone.0215382
66. Moghari MH, Abolmaesumi P. Understanding the effect of bias in fiducial localization error on point-based rigid-body registration. *IEEE Trans Med Imaging*. 2010;29(10):1730-1738. https://doi.org/10.1109/TMI.2010.2051559

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: D'Isidoro F, Chênes C, Ferguson SJ, Schmid J. A new 2D-3D registration gold-standard dataset for the hip joint based on uncertainty modeling. *Med Phys*. 2021; 48:5991–6006. https://doi.org/10.1002/mp.15124