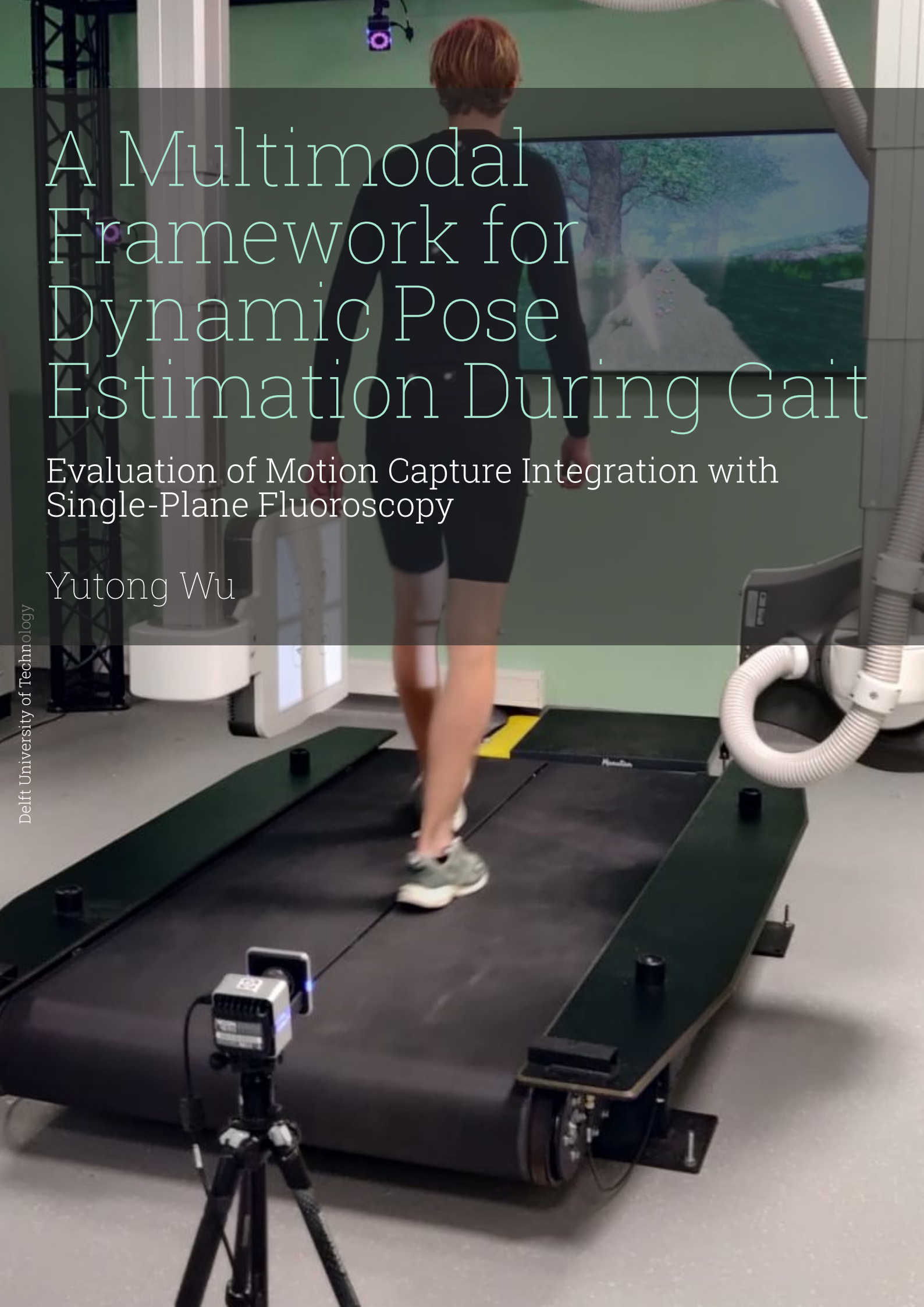


A Multimodal Framework for Dynamic Pose Estimation During Gait

Evaluation of Motion Capture Integration with Single-Plane Fluoroscopy

Yutong Wu

Delft University of Technology

A Multimodal Framework for Dynamic Pose Estimation During Gait

Evaluation of Motion Capture Integration with Single-Plane Fluoroscopy

by

Yutong Wu

Student Number: 5925754 Project Duration: December, 2024 – September, 2025
Faculty: Faculty of Mechanical Engineering, TU Delft
Thesis committee: Prof. dr. ir. J. Harlaar, TU Delft & Erasmus MC, supervisor, chair
Drs. N. Dur, Erasmus MC, daily supervisor
Dr. ir. J. Hirvasniemi, TU Delft & Erasmus MC, daily supervisor
Dr. ir. B.L. Kaptein, LUMC, external committee member

Cover: https://artrosegezond.nl/opening-artrose-onderzoekslab/

TU Delft

Preface

Studying for my Master's degree in Biomedical Engineering at Delft University of Technology has been a rewarding and memorable experience. I am sincerely grateful to the many people who have supported me along the way.

I would like to thank my daily supervisors, Niels Dur, Mariska Wesseling, and Jukka Hirvasniemi, for their guidance and encouragement throughout this project. I feel truly fortunate to have learned from their expertise. I am also grateful to Jaap Harlaar for his advice and feedback, and to Erin Macri and Wouter Schallig for their valuable input during the weekly meetings.

I am also thankful for the presence of someone special, and for the two little worlds that offered me calm and adventure whenever I needed them most. Above all, I owe my deepest thanks to my parents for their constant love and support, which gave me the strength to pursue my studies abroad and complete this journey.

Yutong Wu
Delft, September 2025

Summary

Accurate quantification of in vivo knee kinematics is essential for understanding joint health and disorders such as osteoarthritis. Optical motion capture (MoCap) is widely used in gait analysis, but its accuracy is compromised by soft tissue artefacts, limiting bone-level precision. Fluoroscopy provides direct skeletal visualization, yet it faces several limitations, including the high sensitivity of 2D–3D registration to initial pose estimation and the reliance on manual bone segmentation. Moreover, the potential of combining MoCap with fluoroscopy, particularly the role of MoCap in initializing registration for fluoroscopic knee tracking during walking—remains underexplored.

The aim of this study is to develop and evaluate a multimodal framework that integrates MoCap with single-plane fluoroscopy to improve automatic 2D–3D bone registration during dynamic gait analysis. Specifically, the framework addresses three challenges: (1) temporal synchronization and spatial calibration of the multimodal system, (2) MoCap-based initial pose estimation, and (3) deep learning–based automated femur segmentation.

A synchronized acquisition setup was implemented, simultaneously recording 100 Hz MoCap data and 15 Hz fluoroscopic images during treadmill walking. Sub-frame temporal alignment was verified through a pendulum experiment. Spatial calibration was achieved using a rigid marker box and solved via the Perspective-n-Point algorithm, while a dynamic correction procedure was introduced to update camera pose across trials using rigid markers attached to the X-ray source. For initialization, anatomical reference cubes were constructed from both MoCap markers and segmented MRI data, and rigid registration between these cubes enabled transformation into the fluoroscopy frame. The resulting initial pose was refined using a two-stage optimization algorithm and evaluated against a reference initialization across four trials. MoCap-based initialization produced anatomically plausible in-plane alignment (mean error < 8 mm, < 16°), but consistently showed a systematic depth offset of 30–40 mm.

To reduce reliance on manual annotation, a 2D nnU-Net segmentation model was trained on six manually annotated fluoroscopic images. Despite the limited dataset, the model demonstrated good anatomical plausibility, confirming its potential for automated workflows.

In conclusion, this thesis establishes a multimodal framework that combines MoCap-based initialization, dynamic camera calibration, and deep learning–based segmentation. The evaluation demonstrates both feasibility and limitations, providing a reproducible basis for quantitative analysis of dynamic knee kinematics and opening new avenues for more automated and personalized assessment of joint disorders.

Contents

1

Introduction

1.1. Background and Motivation

Osteoarthritis (OA) is a common chronic degenerative joint disease, with the knee joint being the most frequently affected site. It often leads to joint pain, restricted mobility, and a substantial reduction in quality of life [1, 2]. Numerous studies have demonstrated that abnormal mechanical loading and joint biomechanics play a critical role in both the onset and progression of OA. [3, 4] . Therefore, accurately characterizing the three-dimensional (3D) kinematics of the knee under physiological conditions is essential for understanding OA pathogenesis, evaluating treatment outcomes, and developing preventive strategies.

Various techniques have been developed to analyze joint motion, each with its own strengths and limitations. Conventional imaging modalities such as magnetic resonance imaging (MRI) and computed tomography (CT) provide high-resolution anatomical details but are typically restricted to static postures and are not suitable for dynamic activity analysis [5]. Even emerging dynamic CT imaging remains limited by temporal resolution, radiation dose, and restricted field of view, making it applicable only to small, controlled motions [6].

In parallel, non-invasive optical motion capture (MoCap) systems have become widely used in biomechanics research. MoCap offers high temporal resolution and the ability to capture large ranges of motion, enabling the collection of whole-body kinematic data during activities such as walking, and it remains a primary tool for gait experiments and movement analysis [7]. However, MoCap tracks reflective markers attached to the skin surface rather than the bones themselves, and is therefore susceptible to soft tissue artifact (STA), the relative motion between skin or soft tissue and the underlying skeleton. STA introduces systematic errors into kinematic measurements, with reported displacements of 0.8–14.9 mm and rotational errors of 1.6°–22.4° [8]. Such errors are unacceptable when bone-level accuracy is required [9]. Although various STA correction approaches have been proposed, such as optimized marker placement strategies or the incorporation of kinematic constraints, STA remains a major limitation in high-precision joint kinematic analysis [10, 11].

1.2. Fluoroscopy and Key Technical Limitations

Fluoroscopy is a medical imaging technique that continuously emits a low-dose X-ray beam through the body and projects the transmitted radiation onto a detector. This process produces a real-time sequence of two-dimensional (2D) images that visualize internal structures. Unlike MoCap, fluoroscopy can directly measure skeletal motion without being affected by STA [12]. However, fluoroscopic images represent only 2D projections of the underlying bones, to extract the actual 3D motion, these 2D images must be registered to subject-specific 3D bone models, a process known as 2D–3D registration. The core principle involves projecting the 3D bone model, under a candidate pose, onto the fluoroscopic image plane to generate a digitally reconstructed radiograph (DRR). A similarity metric is then used to evaluate the correspondence between the DRR and the acquired fluoroscopic image [13]. Through iterative optimization, the model parameters are adjusted to maximize similarity until convergence is

reached, at which point the estimated model pose is assumed to represent the in vivo position and orientation of the bone. Combined with high-resolution subject-specific CT or MRI models, this registration framework has been widely adopted to reconstruct in vivo joint kinematics during dynamic tasks [14].

Depending on the imaging configuration, fluoroscopy systems can be implemented as either single-plane or biplane. Single-plane fluoroscopy is more common due to its simpler setup, lower cost, and reduced radiation exposure. Nevertheless, the lack of depth information makes accurate 3D reconstruction more challenging because under a single view, bone poses at different depths may produce similar projection contours. Biplane fluoroscopy acquires images from two orientations and thereby reduces, but does not fully resolve, depth-related ambiguities. It also introduces greater system complexity, higher cost, and increased radiation dose [15]. Overall, 2D–3D registration provides a powerful framework for reconstructing skeletal kinematics from fluoroscopy during dynamic tasks. Yet, several well-recognized challenges remain unresolved. Two specific limitations most relevant to this study are discussed below.

Sensitivity to Initial Pose Estimation

Despite significant progress in the field, initial pose estimation remains one of the central challenges in 2D–3D registration. Initial pose is a coarse approximation of the bone's spatial position and orientation before optimization begins. Since the registration process relies on a similarity-based objective function, which is often highly nonlinear and prone to local minima, the optimization must be initialized close to the true pose in order to increase the likelihood of converging to the correct solution [16].

In current clinical practice, it is common for operators to manually adjust the 3D bone model to achieve a rough alignment with the X-ray images, thereby providing an initial pose. However, such manual initialization increases both procedural complexity and time, and in intraoperative scenarios, it may also disrupt the surgical workflow [17, 18]

To overcome these limitations, researchers have proposed various automatic initialization strategies. One approach is to use external tracking devices, such as preoperative or intraoperative localizers, to provide a reasonable initial alignment [19, 13]. Although effective in surgical navigation, such approaches have not been widely applied in dynamic joint kinematics studies. Another approach relies on image information itself. For example, template-matching methods compare a set of pre-generated projection templates with the fluoroscopic image and select the most similar pose as the initialization [17, 20]. More recently, deep learning–based methods have emerged, in which neural networks are trained to directly or indirectly predict pose parameters [21, 22]. While these learning-driven approaches show great promise, they still suffer from high failure rates and limited generalizability in complex clinical environments. Moreover, most of these methods have been developed in intraoperative contexts for the hip or spine, whereas initialization strategies for dynamic knee analysis remain scarce. Thus, achieving robust and accurate automatic initialization remains an unsolved challenge.

Challenges in Bone Segmentation

Another critical factor affecting the accuracy of 2D–3D registration is the quality of bone contour extraction from fluoroscopic images. Many registration algorithms assume that the target bone contours can be reliably segmented from the X-ray image. However, fluoroscopic knee images often suffer from low contrast, overlapping anatomical structures, and image noise, which can blur bone boundaries and cause confusion with surrounding tissues [23]. These limitations make direct and accurate segmentation difficult.

In recent years, deep learning has emerged as the predominant approach for medical image segmentation, with convolutional neural networks (CNNs) consistently achieving state-of-the-art performance across diverse imaging modalities [24, 25]. Notably, the U-Net architecture introduced by Ronneberger et al. (2015) and the self-adapting nnU-Net framework developed by Isensee et al. (2021) [26] have established new benchmarks for segmentation accuracy in a wide range of medical imaging tasks.

1.3. Laboratory Setup and Research Questions

A multimodal measurement system has been established in our laboratory, integrating a single plane fluoroscopy system, a MoCap System, and a treadmill to enable synchronized acquisition of dynamic knee motion. This setup allows simultaneous recording of high-precision skeletal images and whole-body kinematics during activities such as walking and running, providing a robust foundation for the study of joint biomechanics. Previous studies have shown that combining optical tracking with biplane fluoroscopy enables direct quantification of tibiofemoral kinematics during dynamic daily activities [12]. Nevertheless, the practical implementation of such multimodal integration, particularly in the context of single-plane fluoroscopy, remains largely unexplored. MoCap offers the advantage of capturing whole-body motion patterns, while fluoroscopy provides bone-level imaging accuracy; together, they hold great potential to advance 2D–3D registration of dynamic knee motion.

Despite these capabilities, the current 2D–3D registration pipeline in our laboratory still faces the two challenges: the algorithm requires manual initialization of the knee pose, and bone segmentation in fluoroscopic images relies on manual annotation. These steps are time-consuming and limit both efficiency and scalability.

To address these limitations, the present study aims to enhance the automation of the registration workflow through two strategies. First, MoCap data are integrated into the pipeline by synchronizing the two systems, thereby providing an initial estimated pose for registration. The second strategy is to employ deep learning models for automatic bone segmentation in fluoroscopic images, thereby reducing reliance on manual annotation and improving efficiency.

Although recent studies have attempted to improve nnU-Net by incorporating attention mechanisms, additional loss functions, or hybrid architectures [27, 28, 29], the original framework has consistently demonstrated remarkable robustness. Therefore, this study deliberately adopts the unmodified 2D nnU-Net as a baseline, emphasizing practicality over architectural innovation. The focus is to evaluate whether a fully automated nnU-Net pipeline can provide reliable femur segmentation in fluoroscopic images with minimal user intervention. Given the limited size of the training dataset and the high anatomical consistency of the femur [30], we further hypothesize that nnU-Net's built-in data augmentation and adaptive configuration will be sufficient, making segmentation performance primarily dependent on the quality and representativeness of the training images rather than on model complexity.

Accordingly, the scope of this study focuses on the following three aspects:

1.**System Synchronization and Calibration:** What level of temporal and spatial synchronization accuracy can be achieved with this multimodal setup?

2.**Initial Pose Estimation:** How accurately can synchronized MoCap data provide initial pose estimates for each fluoroscopic frame during dynamic tasks?

3.**Bone Segmentation:** What is the feasibility of training a deep learning model to automatically segment bones in low-contrast fluoroscopic images using only a small set of manually annotated data?

1.4. Thesis Outline

The remainder of the thesis is organized as follows:

Chapter 2: Description of the experimental setup, data synchronization and calibration, initial pose estimation pipeline, and segmentation model training process.

Chapter 3: Presentation and analysis of experimental results, including calibration accuracy, registration performance, and segmentation outcomes.

Chapter 4: Discussion of the findings, limitations of the current approach, and possible directions for future improvement.

Chapter 5: Conclusions.

2

Methods

2.1. System Synchronization and Calibration

To ensure that MoCap data can be meaningfully integrated with fluoroscopic imaging, it is essential to establish both temporal and spatial synchronization between the two systems. This section outlines the multimodal acquisition setup, the strategies used for synchronization, and the calibration procedures for recovering camera geometry. Together, these steps provide the necessary foundation for reliable MoCap-based initial pose estimation in 2D–3D registration.

2.1.1. Experimental Setup and Temporal Synchronization

Kinematic data were recorded at $100\,\mathrm{Hz}$ using a ten-camera optical motion capture system (Vicon, Oxford, UK). Participants walked on a 1×2 m dual-belt instrumented treadmill (M-Gait, Motek, Netherlands) during data acquisition. Fluoroscopic images were acquired at 15 Hz using a single-plane fluoroscopy system (Adora DRFi, Canon Medical Systems Europe, Netherlands).

To synchronize fluoroscopy with MoCap, an electrical pulse was sent from the X-ray tube to the MoCap system at each exposure, allowing the exact MoCap frame corresponding to every fluoroscopic image to be identified. The accuracy of this synchronization approach, however, remains to be validated.

2.1.2. Assessment of Temporal Synchronization

To assess the temporal synchronization, a pendulum experiment was conducted. A reflective MoCap marker was suspended from a fixed point and allowed to oscillate freely within the field of view of the fluoroscopic system. The motion of the marker was simultaneously captured by both the fluoroscopy and the MoCap systems.

In the fluoroscopy image sequence, the marker appeared as a circular object in each frame. Its vertical position was automatically extracted by applying the Hough Circle Transform (cv2.HoughCircles in Python OpenCV) to detect the circular contour and then computing the centroid. This centroid height was interpreted as the 2D vertical coordinate of the marker within that frame. An example frame is shown in Figure 2.1.

On the MoCap side, the 3D position of the same marker was recorded continuously at 100 Hz. The full trajectory was retained, and the vertical component (Z-axis) was extracted. In addition, the time points of the synchronization pulses received by the MoCap system were recorded and marked along this trajectory.

The vertical trajectories retrieved from both modalities were processed for comparison. The fluoroscopy trajectory was interpolated to match the MoCap sampling rate using cubic splines. Both were normalized to the range [0, 1] to eliminate amplitude differences. Two complementary analyses were performed. First, after setting the first pulses as time zero, the relative temporal differences between fluoroscopy frame timestamps and the corresponding MoCap pulse timestamps were calculated. The mean, standard deviation, and maximum error were reported. Second, a cross-correlation analysis was

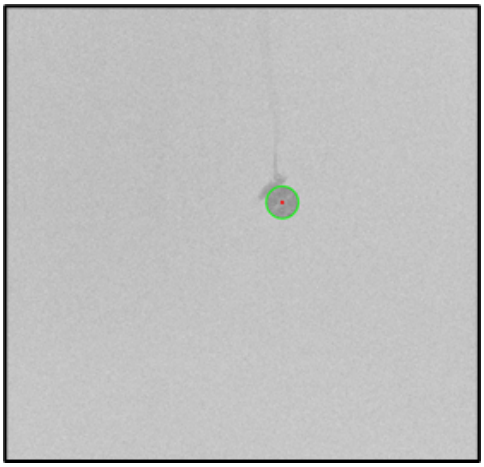


Figure 2.1: Example fluoroscopy frame showing the circular marker detection used for 2D vertical position extraction. The centroid of the detected region (red dot) was computed and used as the marker's vertical position.

conducted between the interpolated fluoroscopy trajectory and the MoCap trajectory. Prior to analysis, both signals were mean-centered. Correlation coefficients were computed as a function of temporal lag within ±50 ms, corresponding to ±5 samples at 100 Hz.

2.1.3. Spatial Synchronization: Intrinsic and Extrinsic Camera Calibration

The projection of 3D world coordinates onto a 2D fluoroscopic image is governed by the camera's intrinsic and extrinsic parameters. The intrinsic parameters define the internal geometry of the imaging system, including focal length and principal point. The extrinsic parameters represent the position and orientation of the x-ray camera with respect to the MoCap coordinate system.

Intrinsic Parameters. The intrinsic matrix $\mathbf{K}$ was computed based on recorded imaging geometry. Assuming square pixels and a centered principal point, the focal length in pixel units was calculated as

$$f_x = f_y = \frac{\text{SID}}{\text{pixel spacing}}.$$

Let (c_x, c_y) denote the image center. The intrinsic matrix was defined as:

$$\mathbf{K} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}.$$

The source-to-image distance (SID) was recorded individually for each trial, and the intrinsic matrix was updated accordingly.

Extrinsic Parameters The extrinsic parameters $(\mathbf{R}, \mathbf{t})$ define the rigid transformation from world coordinates (MoCap system) to camera coordinates:

$$\mathbf{P}_{\text{camera}} = \mathbf{R} \cdot \mathbf{P}_{\text{world}} + \mathbf{t}$$

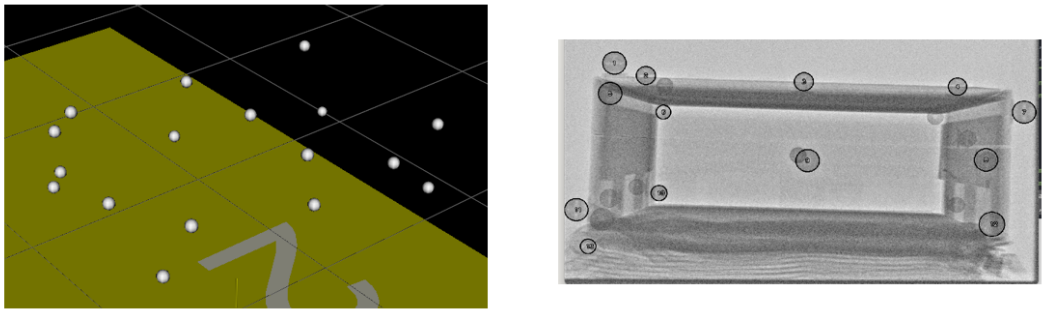
These were estimated through a calibration experiment using a rigid box equipped with reflective markers. The 3D positions of the markers were tracked in the MoCap system, while their corresponding 2D projections were annotated in fluoroscopy images (Figure 2.2).

Using these correspondences, the camera pose was estimated via the Perspective-*n*-Point (PnP) algorithm. The PnP problem refers to the task of recovering the position and orientation of a calibrated camera from n known 3D points in world coordinates and their corresponding 2D projections on the image plane. Mathematically, the problem is to find the rotation matrix $\mathbf{R}$ and translation vector $\mathbf{t}$ that minimize the reprojection error between the observed 2D points $\mathbf{p}_i$ and the projected points:

$$\mathbf{p}_i \sim \mathbf{K} \cdot (\mathbf{R} \cdot \mathbf{P}_i + \mathbf{t})$$

where $\mathbf{P}_i$ are the known 3D points in world coordinates, $\mathbf{K}$ is the intrinsic matrix, and $\sim$ denotes equality up to a scale factor in homogeneous coordinates.

In this work, OpenCV's `solvePnP` function was employed with the iterative Levenberg–Marquardt optimization method. Zero lens distortion was assumed, which is justified by the use of a flat detector panel instead of an image intensifier, minimizing geometric distortion. This yielded the extrinsic parameters describing the transformation from MoCap to camera coordinates.



(a) (b)

Figure 2.2: Box calibration experiment for estimating extrinsic parameters. 3D marker positions were recorded by the MoCap system, while corresponding 2D projections were identified in the fluoroscopy image.

2.1.4. Adjusting for Camera Movement

During experimental sessions, the physical configuration of the fluoroscopy system was not fixed: the X-ray source and image intensifier were frequently repositioned between trials to accommodate varying imaging angles or subject positioning. These changes led to variations in the extrinsic parameters of the imaging system, and occasionally in the intrinsic parameters when the SID was altered. Without a correction mechanism, such changes would necessitate performing a full camera calibration for each trial, which would substantially increase the experimental workload.

To address this limitation, a correction procedure was implemented that allowed the camera parameters to be updated based on the motion of a rigid marker cluster attached to the X-ray source and detector of the fluoroscopy system, thereby eliminating the need for repeated full calibration. This approach enabled flexible and efficient projection without sacrificing accuracy.

Rigid Transformation from Tube Markers

A cluster of reflective markers was rigidly affixed to the X-ray tube housing. These markers maintained a fixed spatial relationship with the imaging system and were tracked by the MoCap system during the initial calibration and throughout all subsequent trials.

Since the X-ray tube was repositioned as a rigid body between trials, the motion of the attached marker cluster reflected the spatial transformation of the imaging system. By comparing the marker positions recorded during each trial with those from the calibration session, the rigid-body transformation was computed. This transformation, denoted $\mathbf{T}_{cc}$, captured both translational and rotational changes in the camera pose in the MoCap coordinate system.

Updating the Camera Pose

The original extrinsic transformation $\mathbf{T}_{mc}$, mapping from the MoCap coordinate system to the camera coordinate system, was obtained during the initial calibration. To apply the correction, this transformation must first be inverted to express the camera pose relative to the MoCap frame:

$$\mathbf{T}_{cm} = \mathbf{T}_{mc}^{-1}$$

For each trial, the camera pose is then updated using the rigid transformation $\mathbf{T}_{cc}$ computed from the tube marker cluster:

$$\mathbf{T}_{cm}^{\text{new}} = \mathbf{T}_{cc} \cdot \mathbf{T}_{cm}$$

The updated MoCap-to-camera transformation is obtained by inverting the result:

$$\mathbf{T}_{mc}^{\text{new}} = (\mathbf{T}_{cm}^{\text{new}})^{-1}$$

From $\mathbf{T}_{mc}^{\text{new}}$, the updated extrinsic parameters $\mathbf{R}$ and $\mathbf{t}$ are extracted.

Intrinsic Matrix Update

In cases where the SID was adjusted, the focal length in pixel units was recalculated to reflect the new imaging geometry. The optical center (c_x, c_y) was assumed to remain constant, based on the fixed detector resolution and orientation.

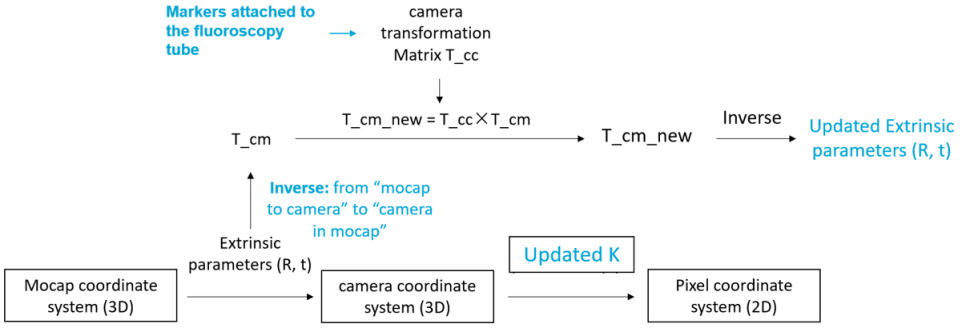


Figure 2.3: Schematic overview of the camera adjustment process. The original transformation $\mathbf{T}_{mc}$ is inverted to obtain $\mathbf{T}_{cm}$, which is updated using the rigid-body transformation $\mathbf{T}_{cc}$ estimated from the tube markers. The updated transformation is then inverted again to obtain $\mathbf{T}_{mc}^{\text{new}}$.

Verification of the Correction Approach

A schematic representation of the transformation and correction pipeline is shown in Figure 2.3.

To verify the accuracy of the correction procedure, the pendulum experiment conducted for temporal synchronization was used as a reference. The 3D marker positions recorded by the MoCap system were projected onto the fluoroscopic images using the updated intrinsic and extrinsic parameters. These projected positions were then compared with the corresponding 2D marker locations visible in the fluoroscopy frames. The mean and variance of the projection errors were calculated to evaluate the correction accuracy.

2.2. Initial Pose Estimation Using MoCap Data

2.2.1. Pose Representation and Coordinate Systems Definitions

The 6-DoF pose of the fluoroscopy camera relative to the 3D bone model is defined using a rigid-body transformation. This transformation is represented by a rotation matrix $\mathbf{R}$ and a translation vector $\mathbf{t}$, forming a homogeneous transformation matrix:

$$\mathbf{T}_{\text{bone}\leftarrow\text{cam}} = \begin{bmatrix} \mathbf{R} & \mathbf{R}\mathbf{t} \\ \mathbf{0}^\top & 1 \end{bmatrix}.$$

This convention follows the definition used in the `diffdrr` framework [31], where the transformation maps a point in the camera coordinate system to the bone coordinate system. The translation component $\mathbf{R} \cdot \mathbf{t}$ represents the position of the X-ray source (i.e., camera center) in the bone coordinate system.

By inverting the transformation, the bone coordinate system can be expressed in the camera frame:

$$\mathbf{T}_{\text{cam}\leftarrow\text{bone}} = \mathbf{T}_{\text{bone}\leftarrow\text{cam}}^{-1} = \begin{bmatrix} \mathbf{R}^\top & -\mathbf{t} \\ \mathbf{0}^\top & 1 \end{bmatrix}.$$

In this representation, $-\mathbf{t}$ corresponds to the position of the bone origin in the camera coordinate system, which is the translation input required for DRR rendering in `diffdrr`.

Bone Coordinate System. The bone coordinate system corresponds to the coordinate system of the 3D bone model, typically derived from CT or MRI imaging. This coordinate system is aligned to the RAS (Right–Anterior–Superior) orientation: the x-axis points to the right side of the patient, the y-axis points anteriorly, and the z-axis points superiorly. The origin is located at the geometric center of the volume.

In the `diffdrr` framework, this coordinate system is referred to as the world coordinate system. However, since it is defined by the input bone model and used throughout the pipeline to describe bone-related transformations, it is referred to here and in subsequent sections as the bone coordinate system.

Camera Coordinate System. The camera coordinate system follows the fluoroscopic imaging geometry. Its x-axis points in the horizontal direction within the image (in-plane horizontal), the z-axis points in the vertical direction within the image (in-plane vertical), and the y-axis is aligned with the X-ray beam direction from source to image receiver (out-of-plane), forming a right-handed coordinate system. The spatial relationship between the camera and bone coordinate systems is illustrated in Figure 2.4.

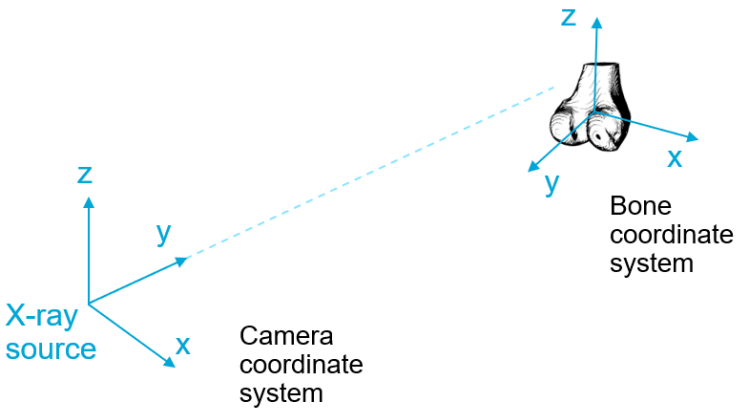


Figure 2.4: Relationship between the camera coordinate system and the bone coordinate system.

Interpretation of 6-DoF Parameters. The final 6-DoF pose of the bone relative to the camera is expressed using three Euler angles derived from $\mathbf{R}$ and a translation vector derived from $\mathbf{t}$. The Euler angles represent the camera orientation with respect to the bone coordinate system, while the translation vector corresponds to the position of the bone origin in the camera frame. These values are passed to `diffdrr` to render synthetic DRRs matching the observed fluoroscopy image.

2.2.2. Estimation Strategy Overview

The goal of the initial pose estimation is to determine the 6-DoF transformation from the camera coordinate system to the bone coordinate system, as required by the `diffdrr` framework for DRR rendering. This estimation is performed using synchronized MoCap data, without requiring additional manual alignment or 2D–3D optimization.

The estimation relies on establishing a geometric relationship between the MoCap system and the bone coordinate system. The marker configuration captured by the MoCap system represents the femur's pose at a given time, which enables the derivation of a rigid-body transformation from the MoCap coordinate system to the bone coordinate system, denoted as $\mathbf{T}_{\mathrm{MC}}^{\mathrm{bone}}$.

At the same time, the camera pose in the MoCap coordinate system $\mathbf{T}_{\mathrm{MC}}^{\mathrm{cam}}$ is known from the calibration step and updated using marker tracking if needed. By combining the two transformations, the pose of the bone in the camera frame can be computed as:

$$\mathbf{T}_{\mathrm{cam}}^{\mathrm{bone}} = (\mathbf{T}_{\mathrm{MC}}^{\mathrm{cam}})^{-1} \cdot \mathbf{T}_{\mathrm{MC}}^{\mathrm{bone}}.$$

This composite transformation $\mathbf{T}_{\text{cam}}^{\text{bone}}$ is expressed in the format required by the `diffdrr` framework, where the bone model is positioned relative to the camera coordinate system for DRR rendering.

2.2.3. Pose Estimation Pipeline

The goal is to compute the camera pose $T_{\text{cam}}^{\text{bone}}$, i.e. the homogeneous transformation representing the camera's position and orientation in the bone coordinate system. The pipeline consists of the following steps:

1. **Construct a bone-aligned cube in Σ_{MC}**: Identify MoCap markers placed near key anatomical landmarks (the medial epicondyle, the lateral epicondyle, and several along the thigh). Using the epicondylar axis (the vector between the epicondyle markers) and the thigh's longitudinal direction (derived from the shaft markers), construct a cube aligned with the femur. The pose of this cube represents the femur's orientation and position in the MoCap coordinate system.
2. **Construct a corresponding cube in Σ_{bone}**: Extract the corresponding medial and lateral epicondyle points on the segmented MRI bone model. Use the anatomical superior direction (assuming a standard RAS orientation) to approximate the femur's long axis, and construct a cube using the same logic as in the MoCap frame. The detailed procedure for constructing this cube will be described in the following section. This cube represents the pose of the same femur in the bone's coordinate system (derived from the MRI/CT volume).
3. **Compute the rigid bone transformation $T_{\text{MC}}^{\text{bone}}$**: Perform a rigid registration between the MoCap-defined cube and the Bone model-defined cube.
4. **Apply MoCap–camera calibration**: Utilize the pre-calibrated transformation $T_{\text{cam}}^{\text{MC}}$, which describes the camera's pose in the MoCap coordinate system.
5. **Compute the camera pose in the bone frame**: Transform the camera's pose into the bone coordinate system by chaining the above transformations. This gives the camera's position and orientation expressed in the bone coordinate system.
6. **Extract 6-DoF pose parameters**: Convert the resulting transformation $T_{\text{cam}}^{\text{bone}}$ into the six degrees-of-freedom parameters required by the 2D–3D rendering system (three orientation angles and three translation components, see details below).

Figure 2.5 gives an overview of the workflow of the proposed 2D–3D registration framework integrating MoCap and fluoroscopic imaging.

The following sections describe the implementation of each stage in the pipeline. The construction of the bone-aligned reference cubes in both the MoCap and bone coordinate systems is first introduced, followed by the estimation of the rigid transformation between them. This is then combined with MoCap–camera calibration to obtain the final camera pose in the bone coordinate system, from which the 6-DoF parameters are extracted.

2.2.4. Reference Cube Construction and Rigid Alignment

Bone Model-Based Reference Cube

To define the femur pose in the bone coordinate system, a reference cube is constructed based on the segmented MRI volume. The volume is preprocessed using `diffdrr` to follow the RAS (Right–Anterior–Superior) orientation, where the x-axis points rightward, the y-axis anteriorly, and the z-axis superiorly. The volume origin is placed at the geometric center.

As illustrated in Figure 2.6, medial and lateral epicondylar landmarks are identified within a narrow y-range ($y \in [-10, 10]$ mm) around the central sagittal plane of the distal femur, to confine the selection to the condylar region and avoid points located too far anteriorly or posteriorly. Within this band, the top 1% of points with the smallest and largest x-coordinates are averaged to obtain medial ($\mathbf{p}_1$) and lateral ($\mathbf{p}_2$) epicondyle points, respectively. The ± 10 mm range is an empirical choice but proved effective in focusing on the condylar area.

The primary axis of the reference cube is defined as:

$$\mathbf{u} = \frac{\mathbf{p}_2 - \mathbf{p}_1}{\|\mathbf{p}_2 - \mathbf{p}_1\|}$$

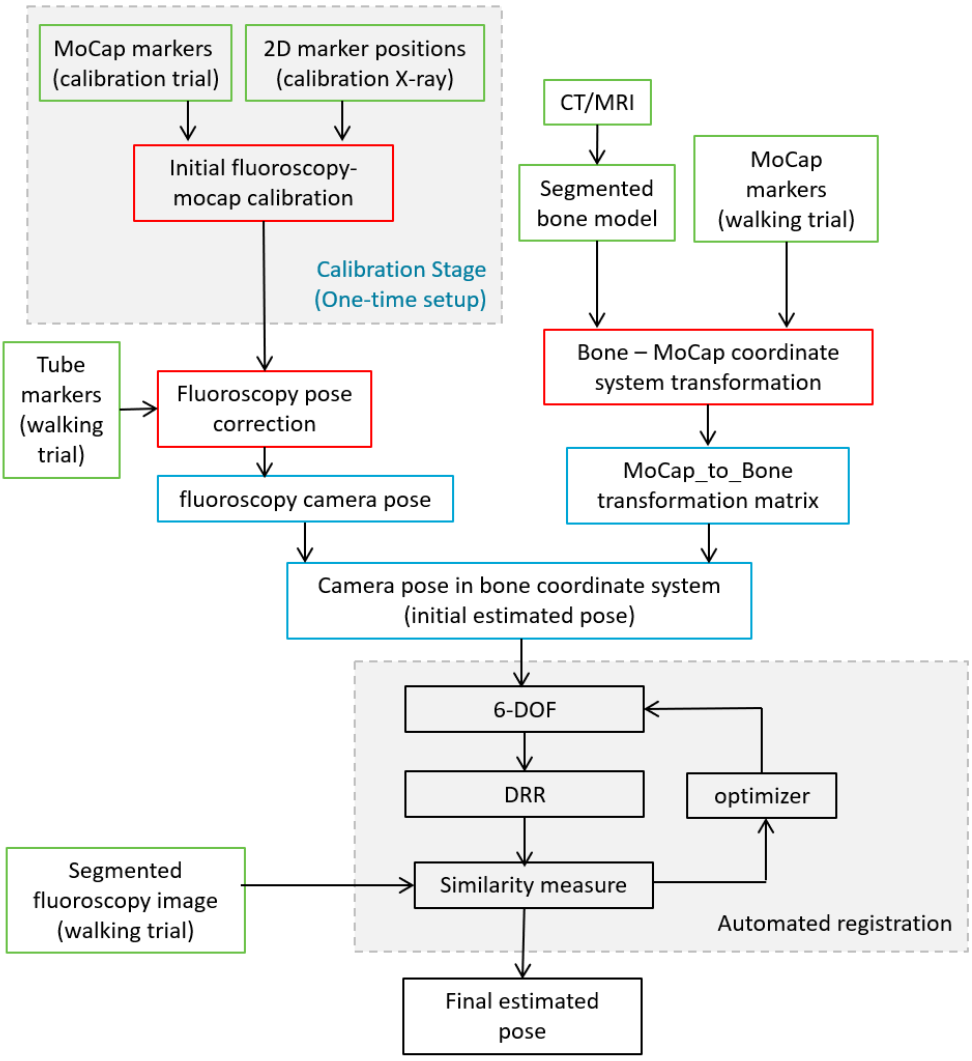


Figure 2.5: Workflow of the proposed registration framework combining motion capture and fluoroscopy. Color coding highlights different modules: green for inputs, red for calibration transformations, blue for pose estimation steps, and black for the final automatic registration.

To define an orthogonal frame, the superior anatomical axis $\mathbf{h} = [0, 0, 1]^\top$ is projected orthogonally to $\mathbf{u}$ to obtain:

$$\mathbf{v} = \frac{\mathbf{h} - (\mathbf{h}^\top \mathbf{u})\,\mathbf{u}}{\|\mathbf{h} - (\mathbf{h}^\top \mathbf{u})\,\mathbf{u}\|}, \quad \mathbf{w} = -\mathbf{u} \times \mathbf{v}$$

Let $L = \|\mathbf{p}_2 - \mathbf{p}_1\|$. The scaled axes are defined as:

$$\mathbf{U} = L \cdot \mathbf{u}, \quad \mathbf{V} = L \cdot \mathbf{v}, \quad \mathbf{W} = L \cdot \mathbf{w}$$

The eight vertices of the cube are generated by linear combinations of these vectors originating from $\mathbf{p}_1$. This cube represents the femur's orientation and position in the bone coordinate system.

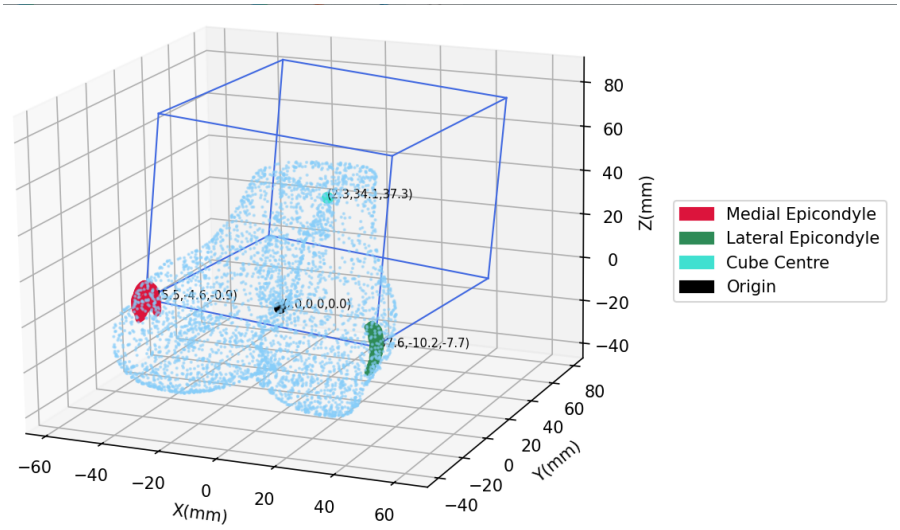


Figure 2.6: Reference cube constructed in the bone coordinate system using epicondylar landmarks.

MoCap-Based Reference Cube

The MoCap reference cube is constructed using five optical markers placed on the thigh: two near the knee joint (LKNM and LKNE for the medial and lateral epicondyles of the left leg) and three distributed along the femoral shaft (LTHI, LTHAD, and LTHAP). These markers are placed following the CGM2.5 protocol, as illustrated in Figure 2.7 [32].

The medial ($\mathbf{q}_1$) and lateral ($\mathbf{q}_2$) epicondyle markers define the primary axis of the cube. The femoral longitudinal direction was estimated using two alternative strategies: (1) the vector connecting LTHAD and LTHAP, and (2) a line perpendicular to the epicondylar axis passing through the lateral thigh marker (LTHI). Both definitions are theoretically valid for approximating the femoral long axis, but rely on different marker sets. Since it is not evident a priori which definition provides a more reliable alignment, both approaches were tested and compared.

Using these two axes, an orthogonal frame is constructed following the same procedure as described for the bone-based reference cube (i.e., orthogonal projection and cross product). However, unlike the anatomical landmarks extracted from the CT/MRI volume, MoCap markers are mounted on the skin surface. As a result, the distance $\|\mathbf{q}_2 - \mathbf{q}_1\|$ between epicondyle markers on the skin is systematically larger than the corresponding inter-epicondylar distance in the bone coordinate system.

To ensure consistency during rigid registration and avoid scale mismatches, the MoCap reference cube is rescaled: its origin is set at the midpoint between $\mathbf{q}_1$ and $\mathbf{q}_2$, and its axis length is adjusted to match the primary axis length of the MRI-derived bone cube ($L = \|\mathbf{p}_2 - \mathbf{p}_1\|$). The resulting cube preserves the orientation structure of the MoCap definition but is normalized to the same scale as the bone cube.

Rigid Alignment and Pose Transfer

After both cubes are constructed, a rigid-body transformation $T_{\text{MC}}^{\text{bone}}$is estimated via point-based registration. This transformation aligns the MoCap-based cube to the cube in the bone coordinate system and encodes the femur's pose observed in MoCap, expressed in the bone frame.

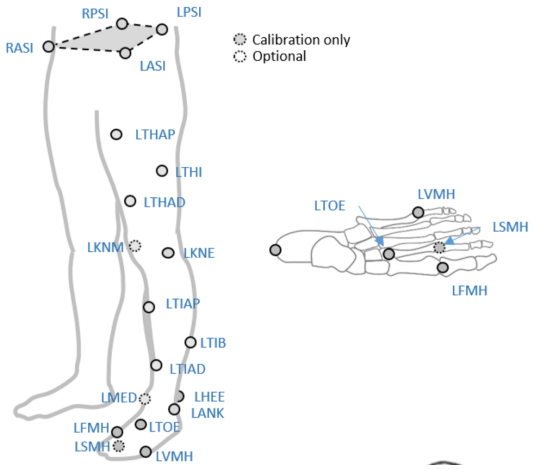


Figure 2.7: MoCap marker setup used to construct the reference cube.

Given the MoCap–camera calibration $T_{\text{cam}}^{\text{MC}}$, the camera pose in the bone coordinate system is obtained by chaining:

$$T_{\text{cam}}^{\text{bone}} = T_{\text{MC}}^{\text{bone}} \cdot T_{\text{cam}}^{\text{MC}}$$

Extraction of 6-DoF Parameters

The transformation $T_{\text{camera}}^{\text{bone}}$ is expressed as:

$$T_{\text{camera}}^{\text{bone}} = \begin{bmatrix} \mathbf{R} & \mathbf{t} \\ \mathbf{0}^{\top} & 1 \end{bmatrix}$$

The rotation matrix $\mathbf{R}$ is converted into three Euler angles (α, β, γ), using the z–y–x order, representing the camera's orientation.

Since the rendering system requires the position of the bone origin in the camera coordinate system, the translation is re-expressed as:

$$\mathbf{t}_{\text{used}} = -\mathbf{R}^{\top} \cdot \mathbf{t}$$

The final 6-DoF pose consists of the rotation angles (α, β, γ) and the translation vector $\mathbf{t}_{\text{used}}$, ready for input to the `diffdrr` rendering engine.

2.2.5. Initialization Strategy and Parameter Settings for Registration

To align the segmented 3D bone model with the fluoroscopic image, an automatic 2D–3D registration algorithm developed in the same research group was employed. This algorithm consists of a two-stage optimization pipeline that refines the initial pose by maximizing a similarity measure between the fluoroscopic image and a DRR. The optimization was performed using Powell's method, a derivative-free algorithm suitable for low-dimensional nonlinear optimization. The final pose estimate was obtained by first optimizing translations only, followed by full 6-DoF optimization.

The automatic registration builds upon previous work in the group, in which initial poses were generated by adding small perturbations to manually registered ground truth poses. These perturbed poses were then used as the input for automatic optimization, and the results were evaluated by comparing the output with the original manual registration. This approach provided a controlled framework for assessing how well the algorithm could recover accurate poses from known initial deviations.

The manual registrations were performed using an interactive alignment tool in a custom 3D viewer (Figure 2.8). In this procedure, the segmented 3D bone model was overlaid on the fluoroscopic image via DRRs, and the user adjusted all six pose parameters (three translations, three rotations) until the DRR visually matched the anatomical structures. The coordinate system and pose conventions followed the `diffdrr` engine. The final transformation was then saved as the ground truth for evaluating automatic registration accuracy.

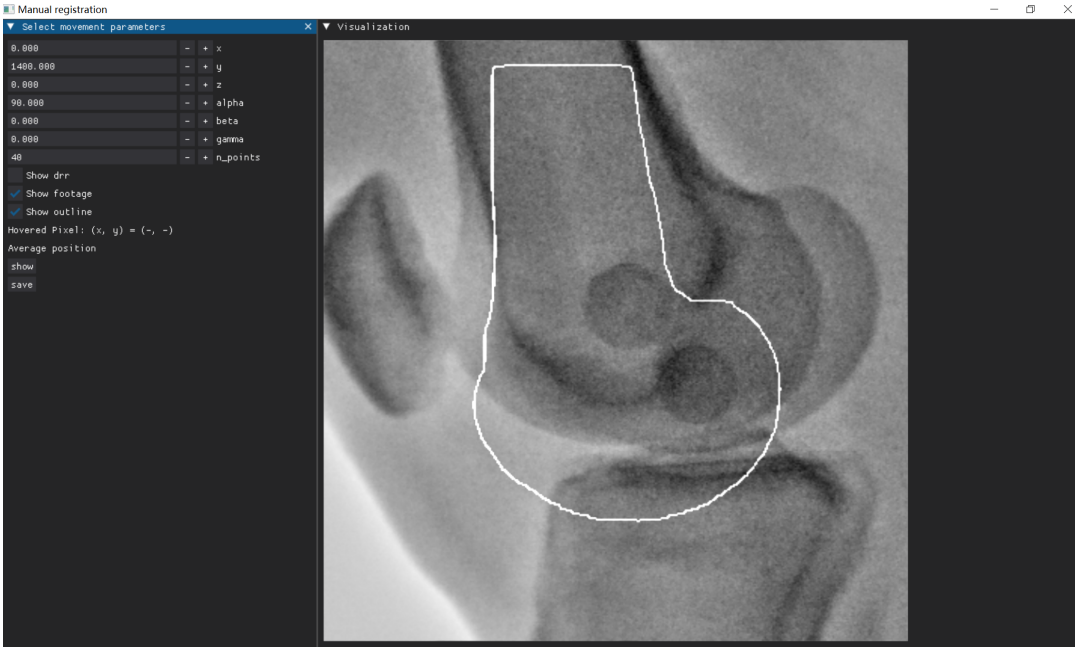


Figure 2.8: Screenshot of the manual registration interface used to generate ground truth poses. The 3D bone model is adjusted in 6-DOF until its DRR matches the fluoroscopy image.

In this study, the registration algorithm was applied to four fluoroscopic trials (Walk01–Walk04) acquired from a single male subject (ID: MOBI003, age: 28). One fluoroscopic frame was selected from each trial based on image clarity, ensuring that the distal femur was clearly visible, with minimal motion blur and no overlap with the contralateral leg. Walk01 to Walk03 correspond to lateral views, while Walk04 was acquired in anterior–posterior (AP) configuration. The source-to-image distance (SID) varied across trials and is summarized in Table 2.1.

Table 2.1: Overview of the four evaluated fluoroscopy trials.

Trial	Subject ID	SID (mm)	Viewpoint	Selected Frame
Walk01	MOBI003	1400	Lateral	12
Walk02	MOBI003	1400	Lateral	49
Walk03	MOBI003	1400	Lateral	29
Walk04	MOBI003	2000	Anterior–Posterior (AP)	39

In this registration framework, the `max_error` parameter defines the optimizer's allowed search range along each degree of freedom relative to the initial pose. In previous studies, `max_error` was tuned based on small synthetic perturbations around the ground truth. However, since MoCap-derived poses exhibit larger discrepancies, those original bounds were no longer sufficient to guarantee successful convergence.

For each selected frame, the automatic registration was evaluated using three different initialization strategies.

In the first and second strategy, the initial pose was generated using motion capture data. This MoCap-derived pose was directly used as input to the automated registration algorithm.

To ensure the optimizer remained effective under these more challenging conditions, two `max_error` configurations were introduced to constrain the search range:

- **Normal**: reflects typical pose discrepancies encountered in MoCap-derived estimates;
- **Large**: accommodates more severe initial misalignments.

Each configuration defines bounds on the allowable deviation from the initial pose across all six degrees of freedom (three translations in mm and three rotations in degrees), as detailed in Table 2.2.

Table 2.2: Bounds defined by the `max_error` parameter. Translations are given along the camera axes in mm: in-plane horizontal (x_{tr}), out-of-plane (y_{tr}), and in-plane vertical (z_{tr}). Rotations are Euler angles about the bone axes in degrees: x_{rt}, y_{rt}, and z_{rt}.

Setting	x_{tr}	y_{tr}	z_{tr}	x_{rt}	y_{rt}	z_{rt}
Normal	10	40	10	20	10	10
Large	15	50	15	30	15	15

As a reference condition, a third initialization strategy referred to as **Reference** was included. This strategy replicates the evaluation protocol used in previous studies within the group, where the performance of the automatic registration algorithm was assessed using initial poses generated by adding small random perturbations to manually registered ground truth poses. For each trial, 10 such poses were created by uniformly sampling each degree of freedom within $\pm 6.2\,\mathrm{mm}$ for translations and $\pm 4.2^{\circ}$ for rotations. This configuration reflects the standard approach previously used to validate the algorithm's accuracy under near-ideal initialization, and serves as a benchmark for comparison with the more challenging MoCap-based initialization introduced in this study.

Three similarity metrics were integrated into the registration framework, following the laboratory's previous work on MRI-to-fluoroscopy registration. The selected measures capture complementary image features and have been successfully applied in earlier studies [33, 34], and were also explored in previous work conducted within our laboratory.

- **WEMS** (Weighted Edge Matching Score): assigns higher weights to strong edges in the DRR and compares them to edges in the fluoroscopy [34];
- **NED** (Normalized Edge Distance): computes the average distance between detected edges in the DRR and those in the fluoroscopy, previously introduced for fluoroscopy-based registration [33];
- **SM** (Shape Matching): quantifies the contour overlap between the fluoroscopy segmentation mask and the DRR-rendered bone silhouette.

Each metric captures distinct features of image similarity and was used independently to guide the optimization process.

The final manually adjusted poses served as the ground truth for evaluating automatic registration accuracy. The overlay of the DRR outline and the corresponding fluoroscopy frame is visualized in Figure 2.9.

Together, these three initialization strategies, namely MoCap-based poses evaluated under both `Normal` and `Large` `max_error` settings, and perturbed ground truth poses (Reference), offer complementary perspectives on registration robustness. The MoCap-based initialization reflects realistic input conditions with potentially large deviations. The Reference initialization represents near-ideal starting points that were previously used to validate the algorithm. The variation in `max_error` bounds allows for exploration of the optimizer's sensitivity to initialization error.

2.3. Bone Segmentation in Fluoroscopy Images

2.3.1. Training Data

The dataset consisted of six 2D single-plane fluoroscopic images of the right leg acquired from three healthy subjects, each in two distinct walking poses. To avoid occlusion and ensure clear visibility of the femur, only frames where the right leg was clearly visible without overlap from the contralateral leg were selected through visual inspection. The images were originally stored in DICOM format and converted to NIfTI using SimpleITK, an open-source library for medical image analysis, to match the nnU-Net pipeline requirements. Each fluoroscopic image had a native resolution of 1296 × 1328 pixels and was retained without downsampling or cropping to preserve anatomical detail. Figure 2.10 provides an overview of the six training images.

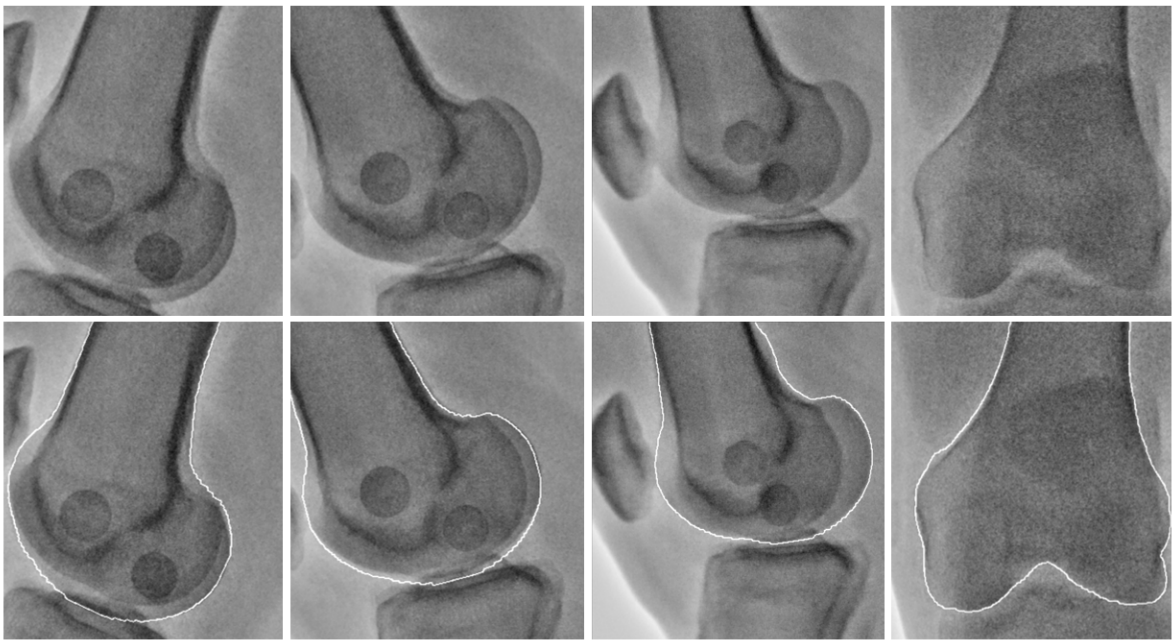


Figure 2.9: Overlay of DRR outline (depicted as the white contour) and fluoroscopic images for all four manually registered trials (Walk01–Walk04), demonstrating alignment accuracy.

Manual segmentation of the femur was performed in 3D Slicer. All masks were initially annotated by a single rater and subsequently reviewed by a second rater with domain expertise to ensure anatomical accuracy. No intensity normalization, cropping, or contrast enhancement was applied; all images preserved their original appearance to assess model generalization under realistic imaging conditions.

2.3.2. Model Training and Inference

The nnU-Net v2 framework was used to train a 2D segmentation model. A five-fold cross-validation strategy was employed, where each fold used five images for training and one for validation. Given the small dataset size, cross-validation was performed at the image level rather than the subject level. The fold assignments were automatically generated and stored in the file `splits_final.json`, which was loaded during training.

All jobs were executed on the DelftBlue high-performance computing cluster managed by TU Delft. Each training job was allocated 16 CPU cores, 64 GB of RAM, and a single NVIDIA Tesla V100S GPU with 32 GB of memory. The software environment was configured with CUDA version 12.8 and driver version 570.124.06. Each fold took approximately 9.5 hours to complete. For a full training script, see Appendix A.

The nnU-Net training was performed using default settings. No manual post-processing or model ensembling was applied during or after training.

To evaluate model generalization to unseen subjects, six fluoroscopic images from a fourth individual were used for inference. These test images were not included in any training or validation folds. Notably, the test subject was female, while the training subjects consisted of thee males and one female. Similar to the training images, the test images were converted to NIfTI format and left unprocessed. Visual inspection was used to assess anatomical plausibility of the predicted masks.

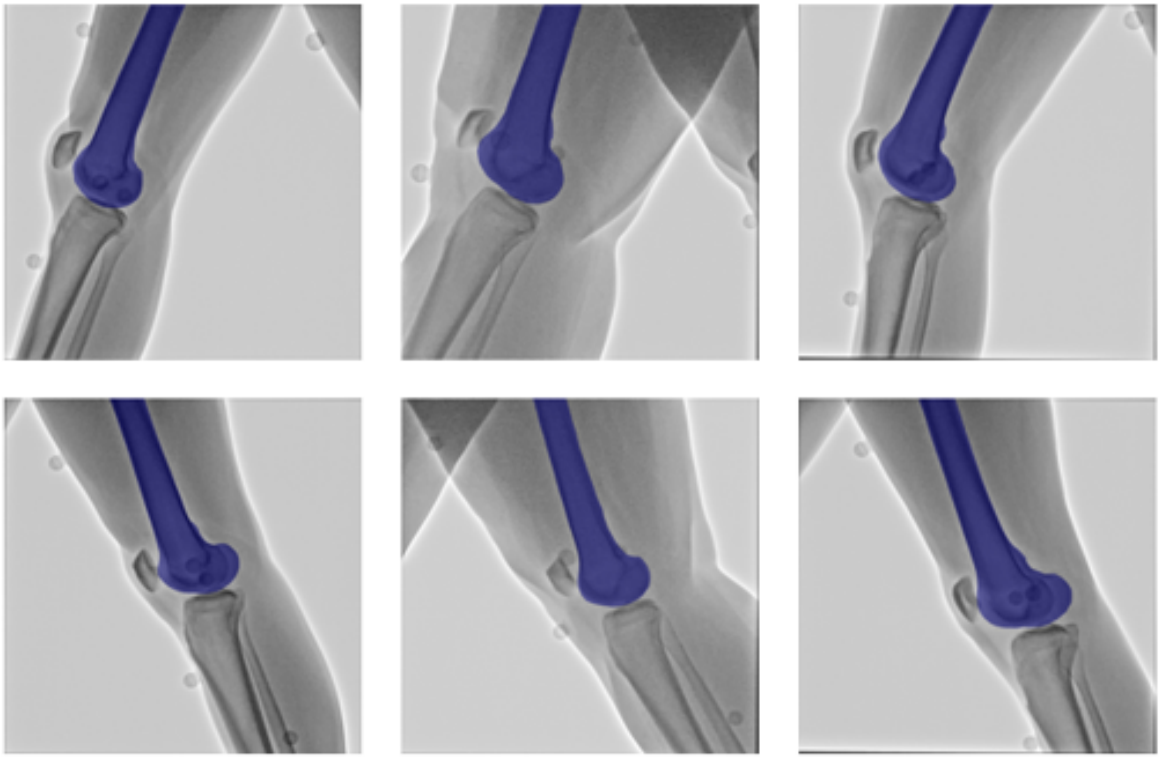


Figure 2.10: Overview of training images used in this study. Each image corresponds to a different leg posture with the femur clearly visible

3

Results

3.1. System Synchronization and Calibration

3.1.1. Assessment of Temporal Synchronization

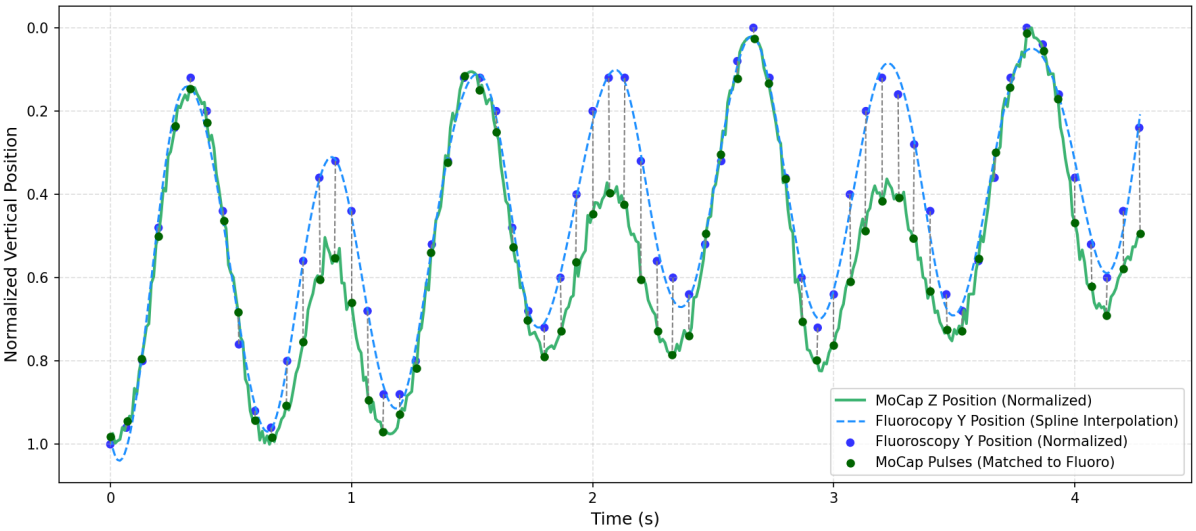


Figure 3.1: Normalized vertical position time series from MoCap (solid green) and fluoroscopy (dashed blue). Fluoroscopy frames are shown as blue dots, and a spline interpolation through these points is shown as the dashed curve. MoCap pulse timestamps are shown as green dots.

Electronic pulses registered in the motion-capture (MoCap) system showed a one-to-one correspondence with fluoroscopic frames. No missing or duplicated pulses were observed across the analyzed sequences, indicating a stable hardware link between the X-ray tube and the MoCap input.

After spline interpolation of the fluoroscopy time series to the MoCap timeline and normalization of both signals to $[0, 1]$, the vertical position curves closely overlapped over multiple oscillation cycles (Figure 3.1). Pulse time stamps aligned with the expected fluoroscopy frame times along the MoCap curve, further supporting correct frame matching. The mean temporal difference between fluoroscopy frame timestamps and the corresponding MoCap pulse timestamps was 2.2 ms (SD 1.6 ms), with a maximum difference of 3.3 ms. Cross-correlation between the two normalized time series exhibited a clear maximum at zero temporal lag with a peak correlation of approximately $r \approx 0.93$ (Figure 3.2). The correlation decreased for positive and negative delays within the tested ± 50 ms window.

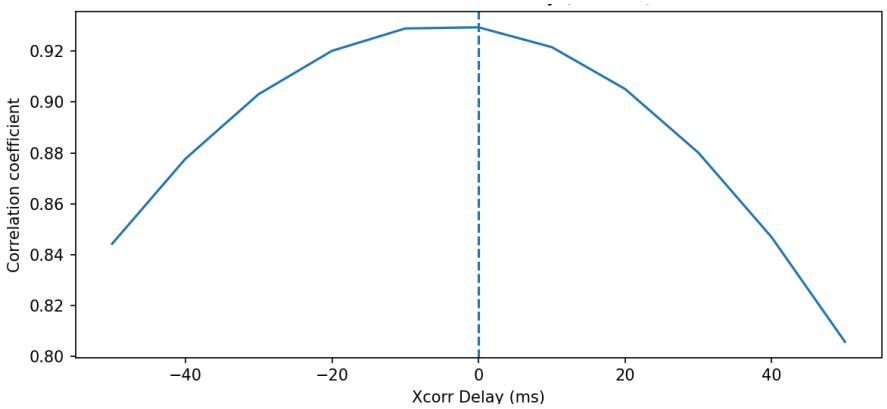


Figure 3.2: Cross-correlation coefficient as a function of temporal lag (ms) between the normalized MoCap and fluoroscopy time series. The maximum occurs at zero lag (vertical dashed line).

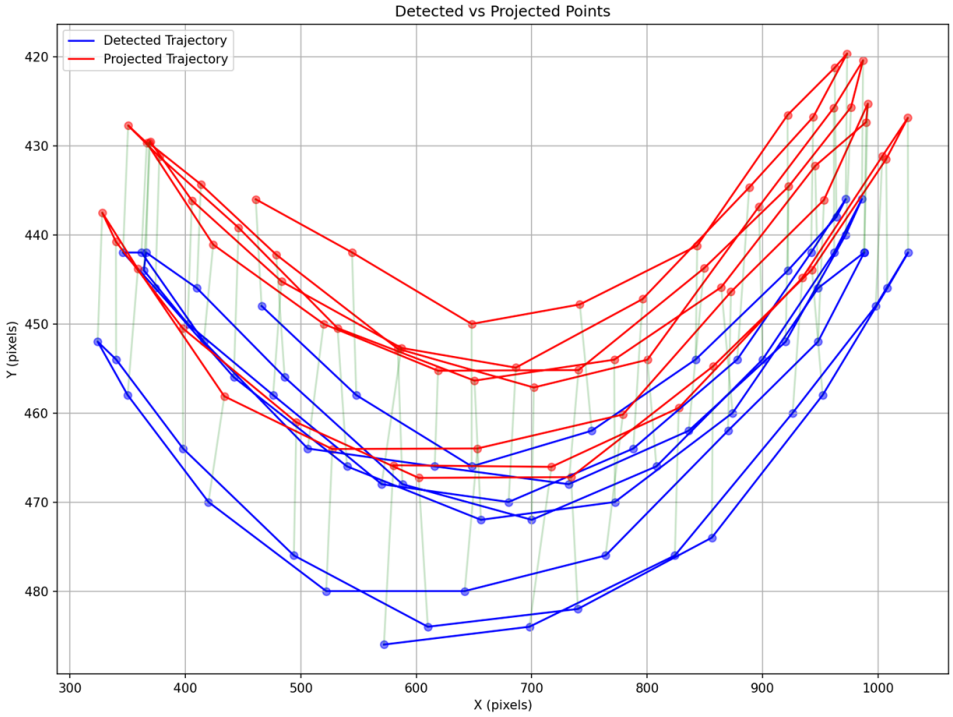


Figure 3.3: Comparison between projected (red) and detected (blue) marker trajectories. Green lines indicate 2D projection error at each time point.

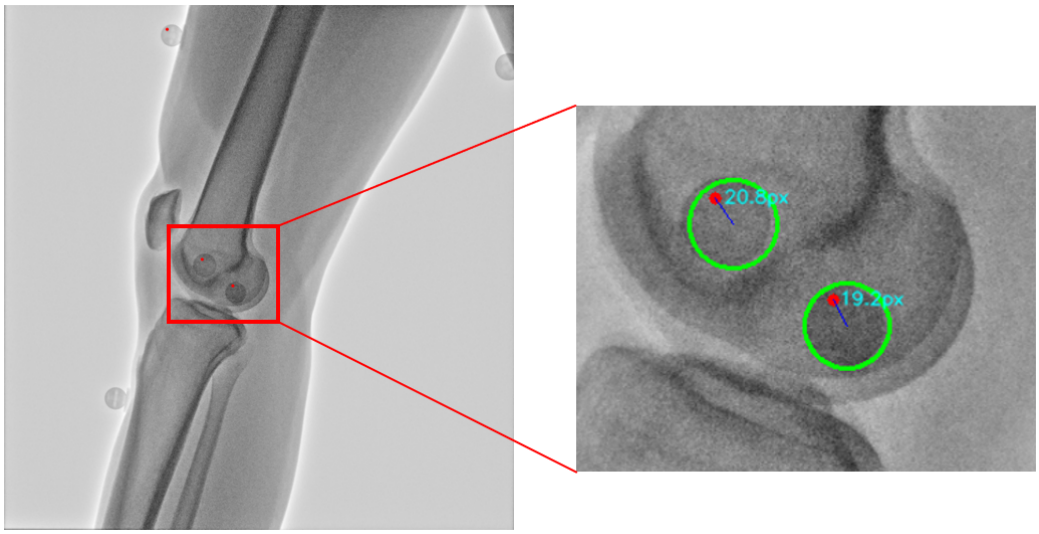


Figure 3.4: Walking trial example: projected 3D MoCap marker positions (red dots) overlaid on a fluoroscopic frame. The detected marker centers are shown as green circles. Blue lines connect each red dot to the corresponding marker center, representing the projection error.

3.1.2. Assessment of Camera Pose Adjustment

Using the updated intrinsic and extrinsic parameters, the 3D positions of the pendulum marker captured by the MoCap system were projected onto the fluoroscopic image plane. These projected trajectories were compared against the marker centers detected directly from the fluoroscopic frames.

Figure 3.3 shows the comparison over 64 consecutive frames, where the projected (red) and detected (blue) marker trajectories are overlaid. The green lines indicate the 2D projection error at each time point. The mean projection error was 16.48 px ($\approx$ 5.3 mm), with a minimum of 11.10 px ($\approx$ 3.6 mm) and a maximum of 26.20 px ($\approx$ 8.4 mm). Errors were smaller near the swing endpoints, where marker velocity was minimal, and larger near the oscillation midpoint, where velocity peaked.

Projection accuracy was also illustrated on real walking trials. For a representative fluoroscopic frame containing visible markers, the corresponding 3D MoCap positions were projected using trial-specific camera parameters. As shown in Figure 3.4, the projection error was approximately 20 px ($\approx$ 6 mm) in the image plane, given the pixel spacing of 0.32 mm.

3.2. Initial Pose Estimation Using MoCap Data

3.2.1. Comparison Between Two Femoral Shaft Definition Methods

To examine the differences between the two MoCap-based femoral shaft definition methods, AD_AP and THI, the resulting 6-DoF poses at the selected frame were compared across four walking trials. The translation and rotation values obtained from each method are summarized in Table 3.1.

Table 3.1: Initial pose translation (mm) and rotation (°) values for each walking trial. Translations are given along the camera axes in mm: in-plane horizontal (x), out-of-plane (y), and in-plane vertical (z). Rotations are Euler angles about the bone axes (RAS) in degrees: x_{R} (right), y_{A} (anterior), and z_{S} (superior).

Trial	Translation (mm)						Rotation (°)					
	In-plane x		Out-of-plane y		In-plane z		Rot. around x (bone)		Rot. around y(bone)		Rot. around z(bone)	
	AD_AP	THI	AD_AP	THI	AD_AP	THI	AD_AP	THI	AD_AP	THI	AD_AP	THI
Walk01	-33.28	-33.57	1109.45	1109.45	-2.70	-2.41	96.29	96.01	-1.07	-1.51	14.73	12.18
Walk02	-59.33	-59.39	1121.88	1121.85	-1.89	-1.43	93.03	92.61	-3.32	-3.66	-22.96	-25.92
Walk03	3.82	3.73	1142.46	1142.43	-6.92	-6.48	91.30	90.89	-3.19	-3.44	-18.18	-21.09
Walk04	79.40	79.40	1748.52	1748.67	-19.92	-19.76	-174.77	-174.93	13.24	11.83	1.52	1.75

For each trial of this subject, the Euclidean distance between the translation vectors obtained from AD_AP and THI was calculated. The average translation difference across the four trials was $0.39 \pm$

0.10 mm. Rotational discrepancies, expressed as absolute differences in Euler angles, were similarly small ($0.32° \pm 0.11°$ around x, $0.61° \pm 0.46°$ around y, and $2.16° \pm 1.13°$ around z). These results indicate that, for this subject, the two femoral shaft definition methods yielded highly consistent 6-DoF initial poses. Since the choice between them makes little practical difference, the AD_AP method was selected for subsequent experiments.

3.2.2. Comparison Between MoCap-Based and Manually Annotated Initial Poses

To evaluate the accuracy of the initial pose estimation derived from MoCap data, the 6-DoF parameters obtained using the AD_AP cube construction method were directly compared against manually annotated ground truth poses for each trial. Signed translation and rotation errors are reported in Table 3.2.

Table 3.2: 6-DoF errors of the initial pose estimate (AD_AP) compared to manual registration. Translations are given along the camera axes in mm: in-plane horizontal (x_{tr}), out-of-plane (y_{tr}), and in-plane vertical (z_{tr}). Rotations are Euler angles about the bone axes (RAS) in degrees: x_{rt}, y_{rt}, and z_{rt}.

Trial	x_{tr}	y_{tr}	z_{tr}	x_{rt}	y_{rt}	z_{rt}
Walk01	-8.28	-40.55	7.30	6.28	2.94	-8.27
Walk02	0.31	-38.84	7.05	14.91	-3.79	-8.36
Walk03	0.86	-41.86	8.58	15.96	-2.86	-9.87
Walk04	-0.60	-24.48	7.08	11.23	-11.36	3.02

Across all four trials, the mean translation error was approximately -1.93 mm (in-plane horizontal), -36.4 mm (out-of-plane), and 7.5 mm (in-plane vertical), with corresponding rotational errors of $12.1°$, $-3.8°$, and $-5.9°$ around the x-, y-, and z-axes respectively.

To illustrate this level of error, Figure 3.5 shows the estimated initial pose from Walk03 within the manual registration interface. As seen in the central panel, the femur obtained from the MoCap-based initial pose largely overlaps with the fluoroscopic bone in terms of overall position, and the general orientation (e.g., anterior tilt) is consistent. Nevertheless, the model does not perfectly coincide with the fluoroscopic bone image.

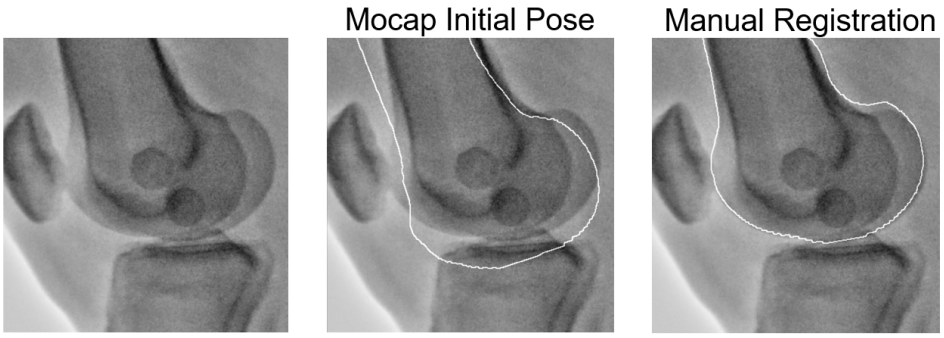


Figure 3.5: Example of MoCap-based initial pose (Walk03) visualized in the manual registration interface. The pose roughly captures the correct bone position and orientation.

The maximum absolute translation error observed was approximately 41.9 mm, and the maximum rotation error reached $15.96°$. These empirical values were used to define the `max_error` bounds in the subsequent registration experiments described in the Method section.

3.2.3. Registration Performance under Different Error Boundaries

Figure 3.6 displays the registration errors across four trials for each degree of freedom under different initialization strategies and similarity metrics. The results show large variations across trials, particularly for the MoCap-based initializations (Normal and Large). Except for in-plane translations, the remaining four degrees of freedom, including out-of-plane translation and the three rotational directions, exhibited trial-specific variations. For the same similarity metric, degree of freedom, initialization range, and trial, the resulting error patterns were not consistent, indicating limited generalizability of registration performance across different trials.

Figure 3.6: Registration errors across four trials for each of the six degrees of freedom. Each subplot shows the effect of three initialization strategies (Reference, MoCap-Normal, MoCap-Large) under three similarity measures (NED, SM, WEMS).

Figure 3.7 provides a summary of these results by aggregating data across all trials. For in-plane translations along the x and z axes, both MoCap-Normal and Reference strategies produced similar mean errors and standard deviations. In contrast, the out-of-plane translation along the y axis and the three rotational directions resulted in higher errors and larger variability for the MoCap-based conditions. The Reference strategy maintained relatively low standard deviations across all degrees of freedom.

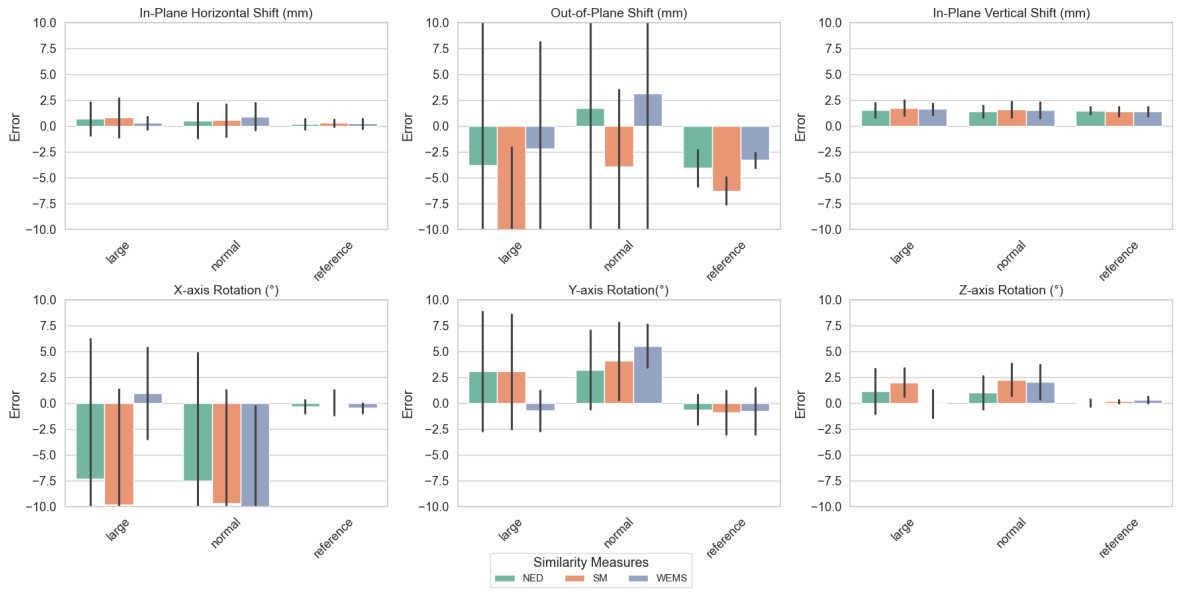


Figure 3.7: Summary of registration errors across all trials, grouped by degree of freedom. Bars represent the mean and standard deviation of the errors for each condition.

3.3. Bone Segmentation in Fluoroscopy Images

The trained nnU-Net model was evaluated on six fluoroscopic images from an independent subject. Figure 3.8 shows the predicted femur masks overlaid on the corresponding images.

Quantitatively, the cross-validation on the training dataset yielded a mean Dice coefficient of 0.988, and the Dice scores for the six images were all above 0.98. For the independent test subject, only qualitative evaluation was performed. Visual inspection confirmed that the predicted contours closely followed the outlines of the femoral shaft and condyles. Segmentation accuracy appeared stable under different pose conditions, and no systematic differences were observed between gait phases.

Minor inaccuracies were identified in several frames. In some cases, projected markers located near the distal femoral condyles were partly included in the predicted masks, despite being outside the true bone boundaries (see lower right image in Figure 3.8). Small deviations of the contours were also noted at the distal edges of the femur (upper right image in Figure 3.8).

In addition, when applied to continuous fluoroscopic sequences where both legs were sometimes simultaneously visible, the model failed to consistently restrict the segmentation to the target femur, resulting in outputs that were often inconsistent and lacked a representative pattern.

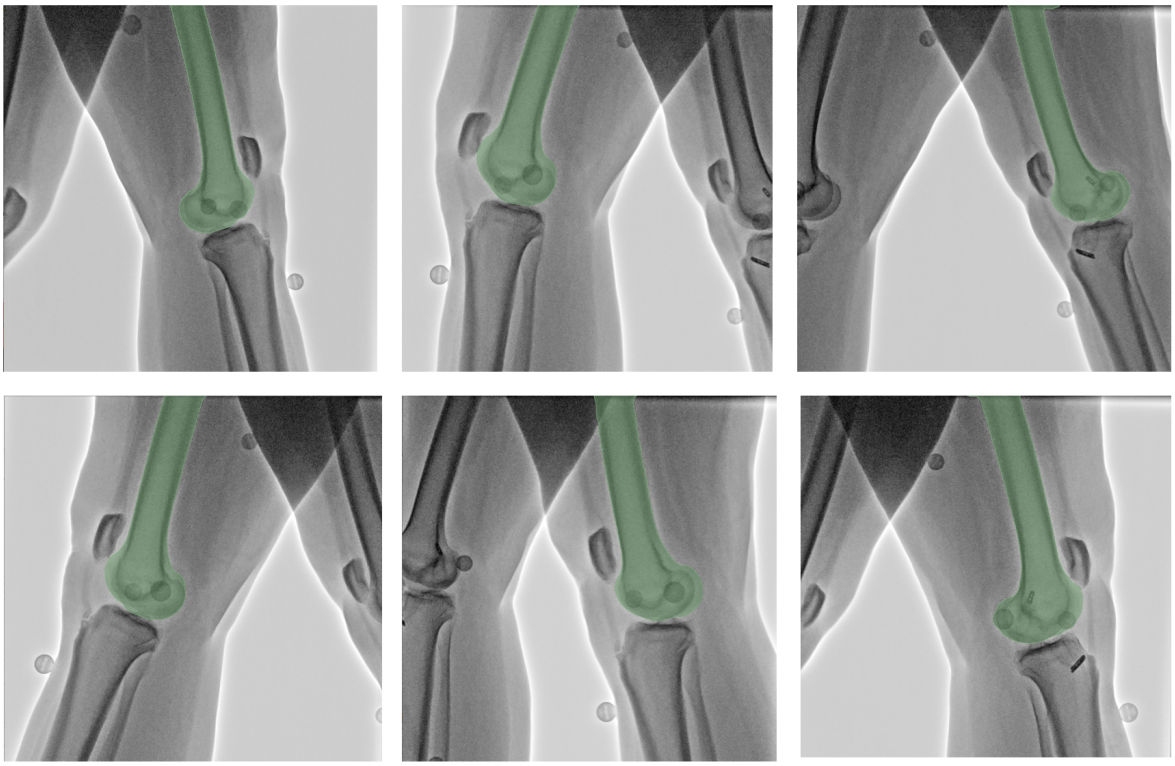


Figure 3.8: Predicted femur masks (green) overlaid on six test fluoroscopic images from an unseen subject.

4

Discussion

4.1. System Synchronization and Calibration

The pendulum experiment confirmed that the MoCap and fluoroscopy systems were temporally aligned at the frame level. Residual offsets were within a few milliseconds, and the cross-correlation analysis showed maximum agreement at zero lag. These results suggest that the two data streams can be treated as effectively synchronous.

In the pendulum experiment, projection errors between MoCap-derived 3D marker positions and fluoroscopy-detected marker centers were in the millimeter range (approximately 3–8 mm). Projection errors tended to increase with marker velocity, with larger deviations observed at higher speeds. This pattern can be explained by motion blur during X-ray exposure, which affects centroid localization when the marker traverses multiple pixels within a single frame. However, even when the marker reached the turning points of the swing, where its velocity was minimal, , errors still exceeded 5 mm in some frames. This indicates that residual inaccuracies in the intrinsic and extrinsic calibration, as well as in the tube-marker-based pose adjustment, are the dominant contributors.

When extended to anatomical data, such discrepancies become more critical. Although a few millimeters represent only a small fraction of the detector field, they are not negligible relative to the projected dimensions of a single bone, where millimeter-level precision is required for reliable initialization.

4.2. Initial Pose Estimation Using Mocap data

Visual inspection of the MoCap-derived initial poses revealed that the projected anatomy and DRR contours were largely consistent in overall position and orientation (Figure 3.5), yet noticeable discrepancies remained across all six degrees of freedom. The discrepancies are likely to arise from a combination of several factors, which can be grouped into four categories:

First, soft tissue artifact. Because the markers were attached to the skin, their positions were affected by relative motion between the skin and the underlying bone. Given the placement of the five markers used in this study, soft tissue motion would theoretically manifest predominantly in the in-plane directions for the lateral view, whereas for the AP view it would be expected to affect the out-of-plane direction. However, the results did not show smaller in-plane errors for the AP view compared to the lateral view, suggesting that soft tissue artifact was not the primary contributor to the observed discrepancies.

Second, approximations in constructing the cube within the bone coordinate system. In the bone model, the femoral shaft axis was approximated by the superior direction of the RAS coordinate system. Although this approach is convenient, it may not accurately reflect the true anatomical axis. Because the extracted femur was relatively short, methods such as principal component analysis or shape-based alignment could not be applied, as they typically require a longer bone segment to robustly define the major axis. A potential improvement would be to register the segmented femur to a standardized full-length femur model and transfer the anatomical axis of the reference model to the partial bone.

This strategy could provide a more consistent and anatomically meaningful shaft orientation, thereby reducing angular errors in the initial pose and improving the robustness of orientation estimation.

Third, errors from constructing the cube using markers in the MoCap coordinate system. For the subject in this study, the AD_AP and THI configurations produced comparable results. The AD_AP setup, with markers placed on the anterior thigh, can provide reliable performance in individuals with moderate body size. However, in larger subjects, anterior–posterior markers may be more strongly affected by soft tissue motion, and therefore less representative of the femoral shaft axis. In contrast, the THI configuration places markers along the lateral thigh, closer to the femoral axis, which in theory provides a more direct and anatomically consistent estimate of shaft orientation. At the same time, this approach is more dependent on operator placement and more prone to soft tissue artifact, since lateral skin surfaces tend to move more during gait. Future work should evaluate both configurations in a larger cohort with diverse body types. Such data are needed to determine the relative robustness of the two setups.

Fourth, inaccuracies in constructing the transformation between fluoroscopy and MoCap. As reported in the previous section, the projection error reached the millimeter scale, indicating that inaccuracies in calibration may also contribute to errors in the initial pose estimation. In this study, camera parameters were derived from a single calibration trial using a rigid box with reflective markers. However, the spatial layout of the markers provided only limited depth variation (about 5 cm), which may have reduced the amount of depth information available for extrinsic parameter estimation and limited its robustness. In addition, intrinsic parameters were computed from theoretical geometry, based on nominal source-to-image distance and pixel spacing. Real-world deviations from the nominal geometric values could therefore contribute to projection errors. Future improvements could involve designing calibration phantoms with greater depth extent to provide richer spatial constraints for extrinsic parameter estimation. Furthermore, adopting multi-view calibration procedures or improved calibration objects could help recover the true imaging geometry more robustly and reduce projection errors in dynamic trials.

4.3. Registration Performance

The automatic registration algorithm was effective when initialized close to the ground truth, as seen with the Reference strategy. In contrast, MoCap-based initializations exhibited substantial deviations under both the MoCap-Normal and MoCap-Large conditions. Although in-plane errors were often corrected successfully, out-of-plane translation and rotational components remained difficult to recover (Figures 3.7).This highlights the sensitivity of the optimization framework to the quality of the initial pose. Large initial misalignments, particularly in depth or rotation, can result in convergence to local minima. Furthermore, good registration performance in in-plane directions does not necessarily indicate optimizer robustness. Rather, based on the results, a plausible explanation is that the initial errors in the in-plane directions were already smaller than those in the out-of-plane and rotational components. Therefore, interpretation of registration outcomes must account for the distribution and magnitude of initial pose errors across all degrees of freedom.

It should be noted that the ground truth used for evaluation was based on manual registration. Because manual alignment relies on 2D overlay, its accuracy in the out-of-plane direction is inherently limited. Therefore, the comparison between MoCap-based and manual initialization does not necessarily indicate which method is more anatomically accurate, but rather highlights the limitations of both approaches. In this context, MoCap-derived poses may still hold potential advantages for out-of-plane initialization, even though such benefits were not clearly demonstrated in our results. Building on this rationale, a hybrid initialization strategy could be considered in future work: MoCap could provide constraints for the out-of-plane components, while in-plane directions are refined via image-based cues or manual adjustment. Such a strategy would, however, depend on improving the accuracy of MoCap-derived estimates through refinements such as better bone axis definitions, shape-aware reference cube registration, or subject-specific femoral morphology.

4.4. Bone Segmentation in Fluoroscopy Images

The 2D nnU-Net model trained on only six fluoroscopic images demonstrated promising generalization to unseen test data. Predicted femur masks generally aligned well with anatomical boundaries, despite

variations in pose and subject anatomy (Figure 3.8). This suggests potential for use in fully automated workflows.

Nonetheless, some limitations were observed. In several images, projected skin markers near the distal condyles were mistakenly included in the predicted bone mask. These artifacts likely resulted from the absence of such structures in the training data. Additionally, minor errors tended to occur near bone edges, particularly close to the image boundaries where intensity gradients are weaker and contextual information is limited. These findings highlight the need for a more diverse and comprehensive training set that includes examples with markers and bones extending to the image edges. Future improvements to the segmentation model could include incorporating images with markers and training on larger datasets. These steps may help improve robustness under varying conditions.

It is also important to note that the current segmentation model was trained and evaluated exclusively on images where only the target leg was visible. However, additional scenarios may arise in practice. For example, some fluoroscopic images may contain both legs within the field of view, while only the target leg is relevant for registration and analysis. In other cases, the two legs may overlap, resulting in partial occlusion of the target anatomy. These more complex conditions were not addressed in the present model and may require specialized handling. To address this, a preliminary strategy was tested in which the non-target leg was manually masked prior to segmentation, allowing the model to correctly extract the femur of interest. While effective, this approach still requires substantial manual effort and does not scale to fully automated workflows. Future work could therefore explore dedicated strategies for such cases, for example by adopting multi-label approaches to distinguish between left and right limbs, or by developing models specifically trained to identify and isolate the target leg under conditions of overlap or occlusion.

5

Conclusion

This thesis presented a multimodal framework that integrates motion capture with single-plane fluoroscopy to enable dynamic 2D–3D bone registration during gait. The system addressed three challenges: temporal and spatial synchronization, MoCap-based initial pose estimation, and automated femur segmentation. Experiments demonstrated sub-frame temporal alignment and millimeter-level projection accuracy after dynamic camera calibration. MoCap-derived initialization provided anatomically plausible in-plane alignment but showed a systematic depth offset of 30–40 mm, highlighting the sensitivity of registration to starting pose. A two-stage optimization algorithm refined most in-plane errors, though large depth deviations remained difficult to correct. For bone segmentation, a 2D nnU-Net trained on six annotated images reached a mean Dice of 0.988 in cross-validation, while qualitative evaluation on an independent subject showed anatomically plausible contours.

Overall, the framework demonstrates the feasibility of integrating MoCap and fluoroscopy for quantitative knee kinematic analysis, while also highlighting limitations in depth accuracy and the robustness of initialization that point toward future improvements.s

References

[1] Han-Zheng Li et al. “Global, regional, and national burdens of osteoarthritis from 1990 to 2021: findings from the 2021 global burden of disease study”. In: *Frontiers in Medicine* 11 (2024), p. 1476853.

[2] Dragan Primorac et al. “Knee osteoarthritis: a review of pathogenesis and state-of-the-art non-operative therapeutic considerations”. In: *Genes* 11.8 (2020), p. 854.

[3] Thomas P Andriacchi and Annegret Mündermann. “The role of ambulatory mechanics in the initiation and progression of knee osteoarthritis”. In: *Current opinion in rheumatology* 18.5 (2006), pp. 514–518.

[4] Timothy M Griffin and Farshid Guilak. “The role of mechanical loading in the onset and progression of osteoarthritis”. In: *Exercise and sport sciences reviews* 33.4 (2005), pp. 195–200.

[5] JGM Oonk et al. “Quantification of the methodological error in kinematic evaluation of the DRUJ using dynamic CT”. In: *Scientific Reports* 13.1 (2023), p. 3159.

[6] Luca Buzzatti et al. “Dynamic CT scanning of the knee: Combining weight bearing with real-time motion acquisition”. In: *The Knee* 44 (2023), pp. 130–141.

[7] Richard D Komistek et al. “Knee mechanics: a review of past and present techniques to determine in vivo loads”. In: *Journal of biomechanics* 38.2 (2005), pp. 215–228.

[8] Andrea Ancillao, Erwin Aertbeliën, and Joris De Schutter. “Effect of the soft tissue artifact on marker measurements and on the calculation of the helical axis of the knee during a gait cycle: A study on the CAMS-Knee data set”. In: *Human Movement Science* 80 (2021), p. 102866.

[9] Wenjin Wang et al. “Effects of soft tissue artifacts on the calculated kinematics of the knee during walking and running”. In: *Journal of biomechanics* 150 (2023), p. 111474.

[10] Brigitte M Potvin et al. “A practical solution to reduce soft tissue artifact error at the knee using adaptive kinematic constraints”. In: *Journal of Biomechanics* 62 (2017), pp. 124–131.

[11] Ann-Kathrin Einfeldt et al. “A new method called MiKneeSoTA to minimize knee soft-tissue artifacts in kinematic analysis”. In: *Scientific Reports* 14.1 (2024), p. 20666.

[12] Albert Planta et al. “A dual-plane fluoroscope to track joint kinematics during dynamic daily activities”. In: *PloS one* 20.7 (2025), e0328351.

[13] Javad Fotouhi et al. “Pose-aware C-arm for automatic re-initialization of interventional 2D/3D image registration”. In: *International journal of computer assisted radiology and surgery* 12.7 (2017), pp. 1221–1230.

[14] Stacey Acker et al. “Accuracy of single-plane fluoroscopy in determining relative position and orientation of total knee replacement components”. In: *Journal of biomechanics* 44.4 (2011), pp. 784–787.

[15] Guoan Li, Samuel K Van de Velde, and Jeffrey T Bingham. “Validation of a non-invasive fluoroscopic imaging technique for the measurement of dynamic knee joint motion”. In: *Journal of biomechanics* 41.7 (2008), pp. 1616–1622.

[16] Primoz Markelj et al. “A review of 3D/2D registration methods for image-guided interventions”. In: *Medical image analysis* 16.3 (2012), pp. 642–661.

[17] Andreas Varnavas, Tom Carrell, and Graeme Penney. “Fully automated 2D–3D registration and verification”. In: *Medical image analysis* 26.1 (2015), pp. 108–119.

[18] Wentao Ye et al. “A Robust Method for Real Time Intraoperative 2D and Preoperative 3D X-Ray Image Registration Based on an Enhanced Swin Transformer Framework”. In: *Bioengineering* 12.2 (2025), p. 114.

[19] Zhenzhou Shao et al. “Robust and fast initialization for intensity-based 2D/3D registration”. In: *Advances in Mechanical Engineering* 6 (2014), p. 989254.

[20] Minheng Chen et al. “Embedded feature similarity optimization with specific parameter initialization for 2d/3d medical image registration”. In: *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2024, pp. 1521–1525.

[21] Yehyun Suh, J Ryan Martin, and Daniel Moyer. “Better Pose Initialization for Fast and Robust 2D/3D Pelvis Registration”. In: *arXiv preprint arXiv:2503.07767* (2025).

[22] Yuxin Cui et al. “Robust and Accurate Multi-view 2D/3D Image Registration with Differentiable X-ray Rendering and Dual Cross-view Constraints”. In: *arXiv preprint arXiv:2506.22191* (2025).

[23] Roman Flepp et al. “Automatic multi-view X-ray/CT registration using bone substructure contours”. In: *International Journal of Computer Assisted Radiology and Surgery* (2025), pp. 1–8.

[24] Wang Jiangtao, Nur Intan Raihana Ruhaiyem, and Fu Panpan. “A Comprehensive Review of U-Net and Its Variants: Advances and Applications in Medical Image Segmentation”. In: *IET Image Processing* 19.1 (2025), e70019.

[25] Fabian Isensee et al. “nnu-net revisited: A call for rigorous validation in 3d medical image segmentation”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2024, pp. 488–498.

[26] Fabian Isensee et al. “nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation”. In: *Nature methods* 18.2 (2021), pp. 203–211.

[27] Ying Peng et al. “The nnU-Net based method for automatic segmenting fetal brain tissues”. In: *Health information science and systems* 11.1 (2023), p. 17.

[28] Yunxiang Li et al. “Plug-and-play segment anything model improves nnUNet performance”. In: *Medical physics* 52.2 (2025), pp. 899–912.

[29] Jieneng Chen et al. “TransUNet: Rethinking the U-Net architecture design for medical image segmentation through the lens of transformers”. In: *Medical Image Analysis* 97 (2024), p. 103280.

[30] Eleonora Croci et al. “Fully automatic algorithm for detecting and tracking anatomical shoulder landmarks on fluoroscopy images with artificial intelligence”. In: *European Radiology* 34.1 (2024), pp. 270–278.

[31] Vivek Gopalakrishnan and Polina Golland. “Fast auto-differentiable digitally reconstructed radiographs for solving inverse problems in intraoperative imaging”. In: *Workshop on Clinical Image-Based Procedures*. Springer. 2022, pp. 1–11.

[32] Stéphane Armand, Morgan Sangeux, and Richard Baker. “Optimal markers’ placement on the thorax for clinical gait analysis”. In: *Gait & posture* 39.1 (2014), pp. 147–153.

[33] Benjamin J Fregly, Haseeb A Rahman, and Scott A Banks. “Theoretical accuracy of model-based shape matching for measuring natural knee kinematics with single-plane fluoroscopy”. In: (2005).

[34] Tsung-Yuan Tsai et al. “A volumetric model-based 2D to 3D registration method for measuring kinematics of natural knees with single-plane fluoroscopy”. In: *Medical physics* 37.3 (2010), pp. 1273–1284.

A

Slurm Example

```
#!/bin/bash
#SBATCH --job-name=nnunet_527_fold_1
#SBATCH --time=12:00:00
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=16
#SBATCH --mem-per-cpu=4GB
#SBATCH --partition=gpu
#SBATCH --gpus-per-task=1
#SBATCH --account=Research-ME-BME
#SBATCH --output=nnunet_train_%j.log
#SBATCH --error=nnunet_train_%j.err


previous=$(/usr/bin/nvidia-smi --query-accounted-apps='gpu_utilization,mem_utilization,
    max_memory_usage,time' --format='csv' | /usr/bin/tail -n '+2')

export PYTHONUNBUFFERED=1

# Activate environment
echo "Activating␣nnU-Net␣environment..."
source ~/nn_Unet_env/bin/activate

# Set nnU-Net paths
export nnUNet_raw=/scratch/ywu19/nnUNet/nnUNet_raw
export nnUNet_preprocessed=/scratch/ywu19/nnUNet/nnUNet_preprocessed
export nnUNet_results=/scratch/ywu19/nnUNet/nnUNet_results

echo "===␣Initial␣GPU␣Status␣==="
nvidia-smi --query-gpu=name,memory.used,memory.total --format=csv

echo "===␣Start␣training␣$(date)␣==="
nnUNetv2_train 527 2d 1 --npz


/usr/bin/nvidia-smi --query-accounted-apps='gpu_utilization,mem_utilization,max_memory_usage,
    time' --format='csv' | /usr/bin/grep -v -F "$previous"

echo "===␣Done␣training␣$(date)␣==="
nvidia-smi
```

B

Cross-validation output

The full nnU-Net summary output for cross-validation is provided below:


```
{
   "foreground_mean": {
      "Dice": 0.9878267058524681,
      "FN": 1538.5,
      "FP": 1650.6666666666667,
      "IoU": 0.9759679676684043,
      "TN": 1587384.6666666667,
      "TP": 130514.16666666667,
      "n_pred": 132164.83333333334,
      "n_ref": 132052.66666666666
   },
   "mean": {
      "1": {
         "Dice": 0.9878267058524681,
         "FN": 1538.5,
         "FP": 1650.6666666666667,
         "IoU": 0.9759679676684043,
         "TN": 1587384.6666666667,
         "TP": 130514.16666666667,
         "n_pred": 132164.83333333334,
         "n_ref": 132052.66666666666
      }
   },
   "metric_per_case": [
      {
         "metrics": {
            "1": {
               "Dice": 0.9866119071193833,
               "FN": 1190,
               "FP": 2285,
               "IoU": 0.9735775603153964,
               "TN": 1589571,
               "TP": 128042,
               "n_pred": 130327,
               "n_ref": 129232
            }
         },
         "prediction_file": "/scratch/ywu19/nnUNet/
nnUNet_results/Dataset527_Femur/
     nnUNetTrainer__nnUNetPlans__2d/
     crossval_results_folds_0_1_2_3_4/bone_001.nii.gz",
         "reference_file": "/scratch/ywu19/nnUNet/
              nnUNet_raw/Dataset527_Femur/labelsTr/
              bone_001.nii.gz"
      },
      {
         "metrics": {
            "1": {
               "Dice": 0.9904493862143129,
               "FN": 2136,
               "FP": 857,
               "IoU": 0.9810794750549978,
               "TN": 1562900,
               "TP": 155195,
               "n_pred": 156052,
               "n_ref": 157331
            }
         },
         "prediction_file": "/scratch/ywu19/nnUNet/
              nnUNet_results/Dataset527_Femur/
              nnUNetTrainer__nnUNetPlans__2d/
              crossval_results_folds_0_1_2_3_4/bone_002.
              nii.gz",
         "reference_file": "/scratch/ywu19/nnUNet/
              nnUNet_raw/Dataset527_Femur/labelsTr/
              bone_002.nii.gz"
      },
      {
         "metrics": {
            "1": {
               "Dice": 0.9878749473400997,
               "FN": 1370,
               "FP": 2055,
               "IoU": 0.9760404060189298,
               "TN": 1578139,
               "TP": 139524,
               "n_pred": 141579,
               "n_ref": 140894
            }
         },
         "prediction_file": "/scratch/ywu19/nnUNet/
              nnUNet_results/Dataset527_Femur/
              nnUNetTrainer__nnUNetPlans__2d/
              crossval_results_folds_0_1_2_3_4/bone_003.
              nii.gz",
         "reference_file": "/scratch/ywu19/nnUNet/
              nnUNet_raw/Dataset527_Femur/labelsTr/
              bone_003.nii.gz"
      },
      {
         "metrics": {
            "1": {
               "Dice": 0.9815149911165274,
               "FN": 615,
               "FP": 3890,
               "IoU": 0.9637009701227963,
               "TN": 1596980,
               "TP": 119603,
               "n_pred": 123493,
               "n_ref": 120218
            }
         },
         "prediction_file": "/scratch/ywu19/nnUNet/
              nnUNet_results/Dataset527_Femur/
              nnUNetTrainer__nnUNetPlans__2d/
              crossval_results_folds_0_1_2_3_4/bone_004.
              nii.gz",
         "reference_file": "/scratch/ywu19/nnUNet/
              nnUNet_raw/Dataset527_Femur/labelsTr/
              bone_004.nii.gz"
```

```
        },
        {
            "metrics": {
                "1": {
                    "Dice": 0.988279480442464,
                    "FN": 2632,
                    "FP": 214,
                    "IoU": 0.9768305192373448,
                    "TN": 1598254,
                    "TP": 119988,
                    "n_pred": 120202,
                    "n_ref": 122620
                }
            },
            "prediction_file": "/scratch/ywu19/nnUNet/
                nnUNet_results/Dataset527_Femur/
                nnUNetTrainer__nnUNetPlans__2d/
                crossval_results_folds_0_1_2_3_4/bone_005.
                nii.gz",
            "reference_file": "/scratch/ywu19/nnUNet/
                nnUNet_raw/Dataset527_Femur/labelsTr/
                bone_005.nii.gz"
        },
        {
            "metrics": {
                "1": {
                    "Dice": 0.992229522882021,
                    "FN": 1288,
                    "FP": 603,
                    "IoU": 0.9845788752609603,
                    "TN": 1598464,
                    "TP": 120733,
                    "n_pred": 121336,
                    "n_ref": 122021
                }
            },
            "prediction_file": "/scratch/ywu19/nnUNet/
                nnUNet_results/Dataset527_Femur/
                nnUNetTrainer__nnUNetPlans__2d/
                crossval_results_folds_0_1_2_3_4/bone_006.
                nii.gz",
            "reference_file": "/scratch/ywu19/nnUNet/
                nnUNet_raw/Dataset527_Femur/labelsTr/
                bone_006.nii.gz"
        }
    ]
}

                ...
            }
        },
        "prediction_file": ".../bone_001.nii.gz",
        "reference_file": ".../bone_001.nii.gz"
    },
    ...
    ]
}
```