Human Activity Reconstruction Using Accelerometer Based Sensor Readings

An undergraduate project proposal report submitted to the

Department of Electrical and Information Engineering
Faculty of Engineering
University of Ruhuna
Sri Lanka

in partial fulfillment of the requirements for the

Degree of the Bachelor of the Science of Engineering Honours

by

G.G.A.I Dayananda - EG/2021/4458
G.M.I. Weerasiri - EG/2021/4858
B.G.D.D.A. Wickramasinghe - EG/2021/4861
W.D.M Deshappriya - EG/2021/4899

.....
Dr.S.H.Gunawardhana
(Supervisor)

Abstract

Human activity reconstruction is widely used in the health industry, sports, rehabilitation, and virtual reality platforms. Traditional motion capture technology often relies on multiple inertial sensors including accelerometers, gyroscopes, as well as magnetometers, increasing the cost, complexity, and the power consumption. This project presents a simplified human activity reconstruction system using a sparse sensor configuration with only five 3-axis accelerometer sensors placed on the wrists, ankles, and torso which is designed as a simple wearable device. Machine learning algorithms are employed for recognizing patterns in human movement from the 15 accelerometer reading we are obtaining for a certain time period, and inverse kinematics is employed to model overall human movement. Our main aim is to represent the reconstructed model as a customized 3D avatar motion based on the body dimensions. Therefore, the proposed approach demonstrates a low cost and simple wearable alternative to conventional human activity reconstruction systems.

Contents

List of Figures

List of Tables

Acronyms

AI Artificial Intelligence

AMASS Archive of Motion Capture as Surface Shapes

API Application Programming Interface

AR Augmented Reality

biRNN Bidirectional Recurrent Neural Network

CoM Center of Mass

DIP Deep Inertial Poser

DNN Deep Neural Networks

ECE Expected Calibration Error

FID Fréchet Inception Distance

FIP Fast Inertial Poser

FPS Frames Per Second

FSR Foot Sliding Ratio

GPU Graphics Processing Unit

HMR Human Mesh Recovery

IK Inverse Kinematics

IMU Inertial Measurement Unit

LBS Linear Blend Skinning

LPIPS Learned Perceptual Image Patch Similarity

ML Machine Learning

MPJPE Mean Per Joint Position Error

PA-MPJPE Procrustes-Aligned Mean Per Joint Position Error

PCA Principal Component Analysis

PD Proportional Derivative

PIFu Pixel-aligned Implicit Function

PIP Physical Inertial Poser

PSD Pose Space Deformation

PVE Per Vertex Error

RGB Red Green Blue

RNN Recurrent Neural Network

ROMP Regressing Monocular 3D Pose and Shape

SIP Sparse Inertial Poser

SMPL Skinned Multi-Person Linear Model

TIP Transformer Inertial Poser

VR Virtual Reality

Chapter 1

Introduction

1.1 Background

Human activity reconstruction of a user's 3D skeletal configuration has become increasingly important for analyzing and visualizing human movement in areas such as healthcare, sports analysis, rehabilitation, virtual reality, gaming, biomechanical analysis and immersive human computer interaction including virtual and augmented reality since it enables remote monitoring and evaluation. Also,the approach of reconstructing human motion using movement data rather than using video data helps to preserve the privacy of the users of these systems. These systems must function reliably in everyday environments and minimize the intrusiveness and user instrumentation.

Traditional motion capture systems rely on high-end optical based setups with multiple cameras, markers such as Vicon, and high accuracy inertial sensors, which require expensive infrastructure, controlled environments, and complex setups to attach them to the body, making them unsuitable for portable or everyday use.

Sensor based approaches that mount devices directly on the human body can overcome the line-of-sight limitations of the vision based methods. Hence, Inertial Measurement Units (IMUs), which provide acceleration and orientation data are proposed due to their low power consumption and ease of wearability. Dense inertial configurations, typically using more than 17 IMUs placed near major joints, can directly estimate a joint's orientation. Therefore, these systems rely heavily on sensor accuracy. Even though there are commercial solutions based on this system like Xsens, dense IMU setups are intrusive, uncomfortable to wear and require extensive calibration making them undesirable for consumer level usages[1].

In contrast, recent research shows that sparse sensor configurations can effectively capture motion by using the lesser number of sensors which are placed on the key segments of the body. The reduced number of sensors leads to insufficient joint constraints, making the motion reconstruction more complex. The IMUs itself is incapable of measuring the distances directly and the early sparse IMU methods primarily based on optimization approaches which the human motion was reconstructed by matching sensor data to pre recorded motion datasets or by solving constrained convex or nonconvex optimization problems using physical and

kinematic constraints. Therefore, these methods require high computational power and present significant latency making these unsuitable for real time applications.

To address these limitations, learning based methods using Deep Neural Networks (DNNs) have been proposed and Deep Inertial Poser (DIP) was one of the earliest learning based method used to directly map sparse IMU data to full body pose using DNNs and large scale training datasets. DIP significantly improved the real time performance which was a challenge in earlier times but still had some limitations in accuracy and recreation of the complex motions. Then , the Transformer Inertial Poser (TIP) introduced transformer architectures combined with sampling based optimization to capture long term temporal dependencies in motion data improving the reconstruction quality but the computational cost was increased. Physical Inertial Poser (PIP) further incorporated physical constraints and torque modeling through dual proportional derivative (PD) controllers to model the torque and optimize the output of the transpose. But this additional optimization layer increases the computational burden and reduces the suitability for resource constrained platforms.

For real-world applications involving consumers, motion capture systems should be able to operate within mobile platforms such as AR-VR Headsets, Smart Glasses, or wearable technology that impose tight power, size, and processing resource constraints. More importantly, most of the current approaches haven't considered body shape information. This disregard causes bias to body shapes and adds to computational costs involved in pose estimation.

Then the Fast Inertial Poser (FIP) was proposed which is a real time motion capture method which can reduce the computational burden and latency while increasing the accuracy of the human activity reconstruction making it more suitable for embedded platforms with limited computational power and also it consider the body measurements of the user before wearing the device.

Also, the coupling of Machine Learning (ML) methods and biomechanical modeling has emerged as a successful strategy for the estimation of correct 3D skeletal poses using noisy and incomplete sensor data. From the point of view of IMU-based motion capture systems, Machine Learning models can be trained to learn the complex, generally nonlinear relations between the sensor data (Inertial measurements of orientation, Angular Velocity, and Acceleration) and the corresponding joint orientations of the human body. Using the prior knowledge of human motion learned by Machine Learning systems, it is possible to overcome the ambiguities inherent to the application of incomplete sensor setups and noisy sensor readings.

With joint orientation/pose parameters estimated by ML algorithms, a full-body skeletal reconstruction is often obtained via Inverse Kinematics (IK). It imposes constraints such as bone length, joint limits, and body part hierarchical order to ensure a valid skeletal reconstruction that is not only meaningful in a physical sense but anatomically feasible as well. Thus, ML is used for a problem that is undetermined yet important for pose reconstruction and temporal consistency of

the reconstruction sequence, and IK is then used to refine the final reconstruction and prepare it for visualization purposes.

The synergy of ML-based pose estimation and inverse kinematics is even more valuable for real-time computation on embedded systems. ML inference is fast once the model is trained, and IK is a light Post-processing step that maintains human body consistency with relatively inexpensive global optimization. Thus, these two combined can make skeletal data generation from sparse IMU highly efficient for minimally invasive and real-time motion capture systems used for applications like VR/AR, gaming, and human-computer interaction[2].

1.2 Problem Statement

Even though human activity reconstruction is carried out using wearable motion capture systems in areas such as healthcare, sports, virtual reality, and rehabilitation. Most existing IMU-based motion capture approaches rely on multiple sensors incorporating accelerometers and gyroscopes to estimate body orientation and joint movements with high accuracy. Commercial systems such as Xsens MVN and perception neuron typically employ 10-17 IMUs leading to increased hardware cost, complex calibration procedures and higher computational requirements for sensor fusion and drift compensation making such systems less suitable for simple and low hardware cost wearable solutions.

Gyroscope-based motion reconstruction involves sensor fusion, calibration processes, and drift compensation algorithms which increases the computational power, power consumption and the design complexity. Also, over time the drift will accumulate and this leads to inaccuracies in the orientation which affects the precision of the reconstruction over longer spans of the activity. Furthermore, when a larger number of gyroscope and accelerometer equipped sensors are used, it reduces the wearability and simplicity, making them uncomfortable for users and impractical for everyday or long-term use.

Considering all these limitations, a simple, low cost and minimally intrusive system approach is needed for human activity reconstruction. Therefore, the challenge is to utilize a sparse set of accelerometer only sensors and combine evolving engineering technologies in order to reconstruct meaningful human motion while maintaining affordability, simplicity, and user comfort.

1.3 Objectives and Scope

1.3.1 Aim

Develop a low cost and simple wearable accelerometer based human activity reconstruction system that accurately estimates the movements of the joints and reconstructs the full body motion and visualizes it through a personalized 3D

avatar while validating the reconstructed motion against synchronized reference video data.

1.3.2 Objectives

The specific objectives of the project are as follows,

1. Accurately acquire raw time synchronized motion data from the user during movement
   Expected outcome – Accurate accelerometer-based sensor readings with synchronized timestamps representing the user's physical movement to be used for further processing.

2. Design and develop a method to estimate joint postures and orientations using the collected raw accelerometer data
   Expected outcome – Accurate joint posture parameters, such as joint angles and orientations, that represent the user's physical movement for use in skeletal mapping.

3. Design a skeletal model using joint posture parameters and create a 3D avatar
   Expected outcome – A realistic, 3D human avatar that reconstructs the full-body motion of the user's physical movement.

4. Design and implement a subsystem to validate the reconstructed human activity and the avatar output through a reliable evaluation method
   Expected outcome – Validation results representing the accuracy, consistency, plausibility, and reliability of the human activity reconstruction system.

1.3.3 Scope

1. Development of computational methods to estimate the posture or the movement of the human body using only accelerometer data, without reliance on gyroscopes or magnetometers. To compensate for the limited input which utilizes only 5 accelerometer based sensors, advanced predictive and calculative methods will be needed. Therefore, due to computational latency and processing constraints, real time motion reconstruction will not be feasible. Consequently, the system operates in an offline capacity; data is recorded during the session and manually processed post-hoc for activity reconstruction.

2. Integration of estimated posture or the movement into a skeletal kinematic model for a high fidelity representation of the user's movements.

3. Creation of a personalized 3D human avatar that animates the reconstructed skeletal motion, incorporating skin generation based on individual user parameters such as limb lengths, body proportions and skin tone providing a realistic visual representation of the user's physical movements while enhancing the user experience. However, a primary limitation of the current model is the absence of fine-grained joint articulation-specifically finger movements, subtle rotational variances, and head orientation as the reconstruction is constrained by the sparse five-sensor input.

4. Design and implementation of a dedicated validation subsystem that utilizes a synchronized video feed as the ground-truth reference. The validation process involves a side-by-side comparison of the raw video footage against the reconstructed avatar motion. This subsystem focuses on qualitative assessment and approximate accuracy rather than exhaustive quantitative precision, as the primary objective is to validate the model's architectural integrity from a visualization perspective. The validation process may be limited by the availability, synchronization accuracy and quality of reference video data or ground-truth motion capture systems.

Chapter 2

Literature Review

2.1 Previous Work

2.1.1 Developing the Models using Sparse IMUs to Estimate the Joint Postures and the Orientations

IMU-based human activity reconstruction aims to model full-body pose while avoiding the limitations of camera-based systems. Over time, research on this topic has evolved from optimization-based approaches to deep learning models while reducing the number of sensors, computational power, cost, and improving durability. SIP, DIP, TIP, PIP, and FIP represent key milestones along this research journey.

(a) Sparse Inertial Poser (SIP)

SIP is an optimization-based method that formulates pose estimation as a global or temporal optimization problem using motion priors and kinematic constraints. Although this method achieves high accuracy with sparse IMUs, it is computationally expensive and exhibits higher latency, which limits its applicability in real-time and mobile systems [1].

(b) Deep Inertial Poser (DIP)

DIP employs bidirectional RNNs to estimate joint rotations from IMU data, incorporating learned kinematic constraints to achieve faster inference compared to SIP while maintaining good accuracy.However, this system is not suitable for real time applications because of the non causal inference and the latency of the system where it is depending on past and future frames. [3]. Figure 2.1 illustrates the overview of the DIP pipeline.

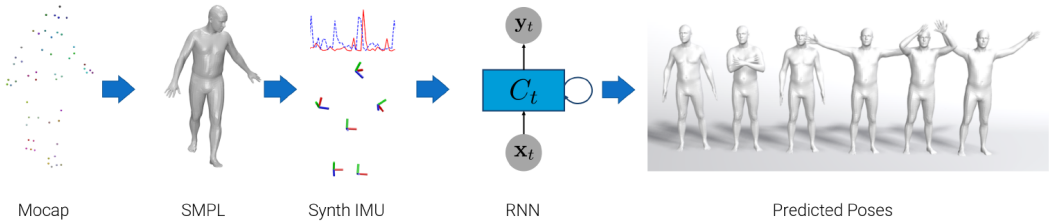

Figure 2.1: Overview of the DIP pipeline

(c) Transformer Inertial Poser (TIP)

TIP is a transformer based system which can handle complex motions accurately since it captures long range temporal dependencies without using inverse kinematics solver. But this system requires high computational power and memory resources limiting the usage in embedded and wearable applications. [1].Figure 2.2 presents the overview of the TIP pipeline.

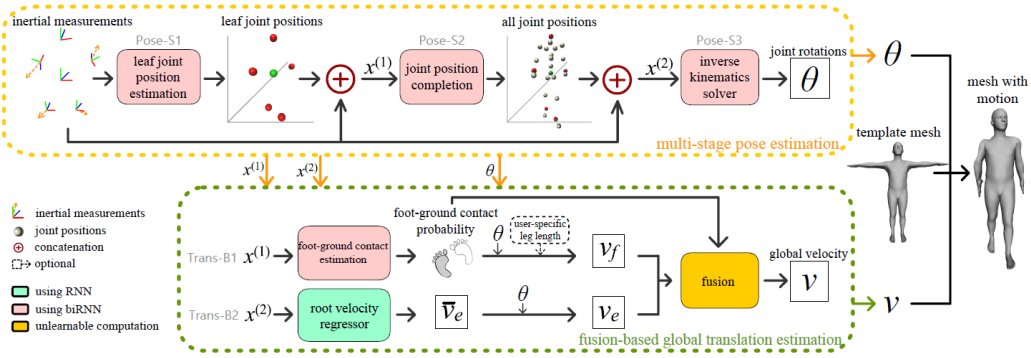

Figure 2.2: Overview of the TIP pipeline

(d) Physics-based Inertial Poser (PIP)

PIP is a physics-based model with optimization and inverse kinematics to enhance the balance constraints and ground contact consistency. This system has improved realistic human motion reconstruction and contact handling than the learning based models. But high computational power and complex system design limits the real time applicability[2].

(e) Fast Inertial Poser (FIP)

FIP is a real time human motion reconstruction model which estimates full body motion using only six sparse IMUs while considering the human body shape information. This model uses a bidirectional RNN that and enables low-latency inference. Body parameters such as height and limb lengths help resolve pose ambiguities caused by sparse sensing. Experimental validation on the AMASS and DIP-IMU datasets demonstrates state-of-the-art performance, achieving over 60 FPS with less than 15 ms latency [2]. Figure 2.3 shows the overview of the FIP pipeline.

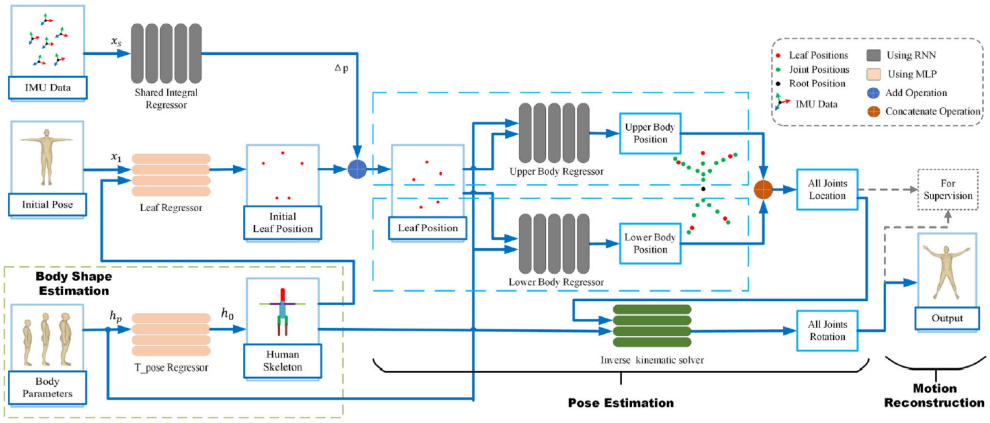

Figure 2.3: Overview of the FIP pipeline

2.1.2 Developing the Models using Accelerometers to Estimate the Joint Postures and the Orientations

1. Human activity reconstruction for full-body pose has been performed using only four sparse 3-axis accelerometers placed on the wrists and ankles. Instead of using accelerometer readings directly, which can introduce significant drift, this method used a data-driven approach leveraging a large motion capture database. A simulate-and-optimize approach was utilized to generate synthesized accelerometer readings to align with sensor data through nearest-neighbor search and graph optimization. Inverse kinematics and reduced-dimensional pose representations were used to recover plausible full-body poses in real-time [4]. Following Figure 2.4 shows the overview of the system proposed in tautges 2011.

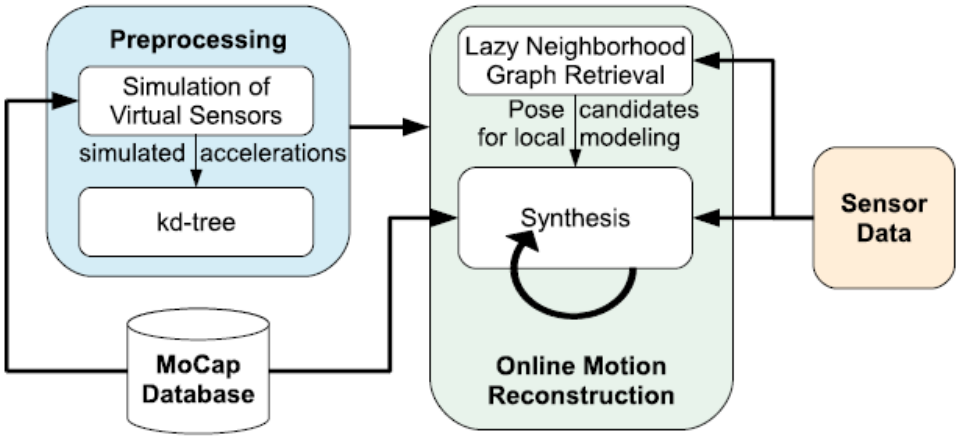

Fig. 1: *Overview of the animation system.*

Figure 2.4: Overview of the system proposed in tautges 2011

2. Human motion can also be reconstructed using wearable accelerometers in complete isolation from the requirement for optical motion capture systems. A portable, low-cost framework where several body-mounted tri-axial accelerometers were used to estimate full-body motion, especially for sports and outdoors for which optical systems are impractical. Instead of directly computing the joint angles from sensor data, their approach relies on a data-driven motion database and a motion graph constructed from pre-recorded motion capture data. On this database, virtual accelerometers were simulated, and a dynamic programming–based matching algorithm was used for finding pose sequences whose simulated accelerations best matched the measured sensor readings. While the method indeed produced naturalistic motions using only accelerometer data, it was prone to positional drift over time, showing one major limitation of accelerometer-based motion reconstruction approaches.[5] Following Figure 2.5 shows the overview of the system proposed in kelly 2010.

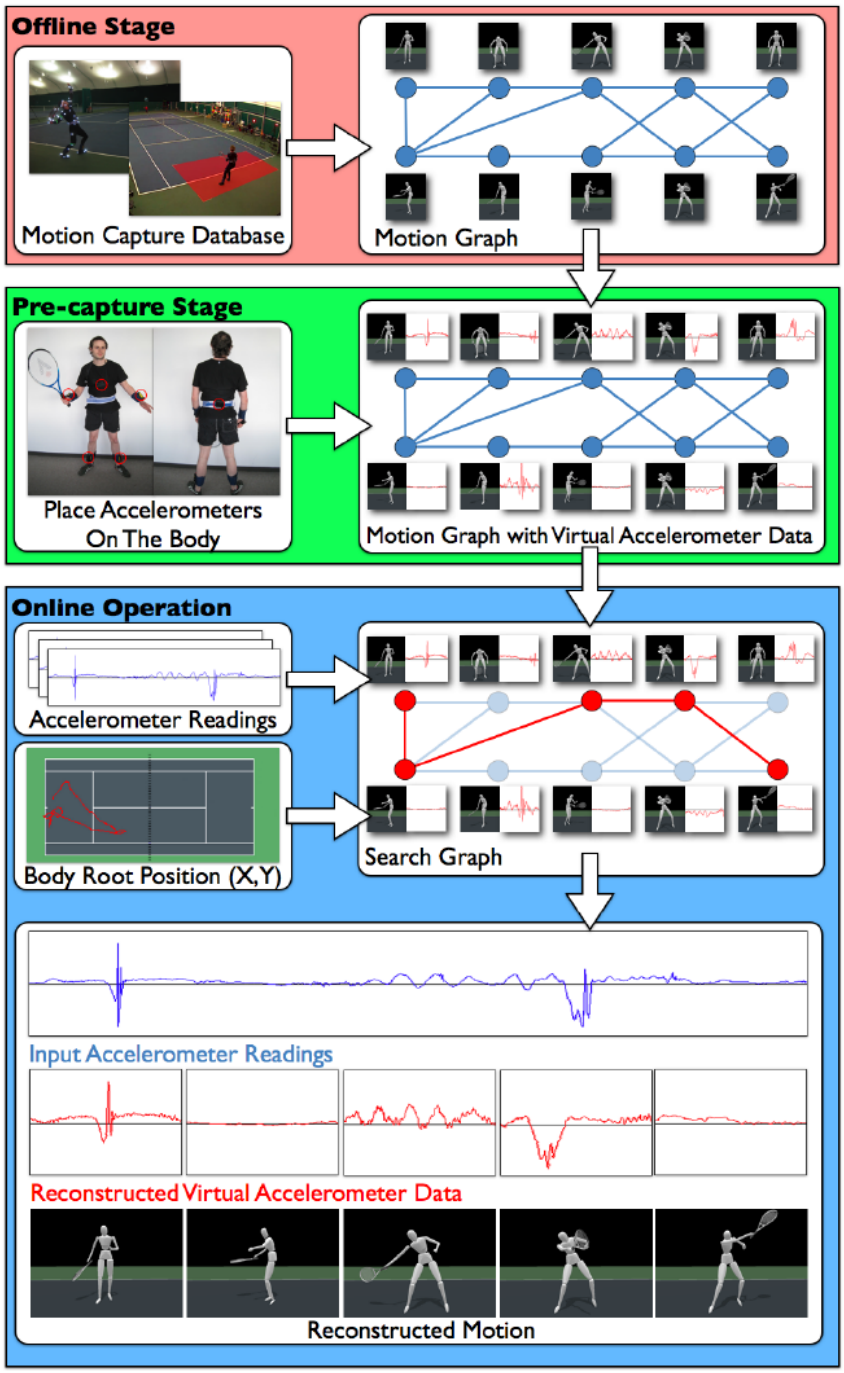

Figure 2.5: Overview of the system proposed in Kelly (2010)

2.1.3 Human Activity Detection Using WWSN and Artificial Intelligence

Beyond skeletal pose estimation, localized research has demonstrated the efficacy of a Wearable Wireless Sensor Network (WWSN) for direct activity classification using sparse accelerometer data. This study, from the University of Ruhuna, accurately differentiated six specific physical activities: cycling, pushups, running, squats, walking, and table tennis. The study also aimed to discover the optimal configurations of the sensors to lessen the burden placed on the user.

The research developed two distinct paradigms for activity classification based on the dependencies they capture:

1. Detection by Capturing Spatial Dependencies: When an activity is seen as a single snapshot of sensor readings, conventional machine learning approaches are used. The Random Forest Classifier (RFC) outperformed KNN, SVM, and Naïve Bayes, with an accuracy of over 96%.

2. Detection by Capturing Spatial and Temporal Dependencies: For activities defined as a series of instances, Convolutional Neural Networks (CNNs) were used to capture complex physical movements over time. This strategy achieved an accuracy and F1-score of more than 97%.

   A key conclusion of this study was the identification of wrist sensors (Sensor IDs 1 and 4) as the most significant pair for classifying target activities. [6]. The results show that a reduced arrangement of two sensors can give reliable activity recognition with only slight reduction in performance, provided the models are trained on a balanced and sufficiently large dataset. Following Figure 2.6 shows the sensor positioning of WWSN and the three axes of accelerometer.

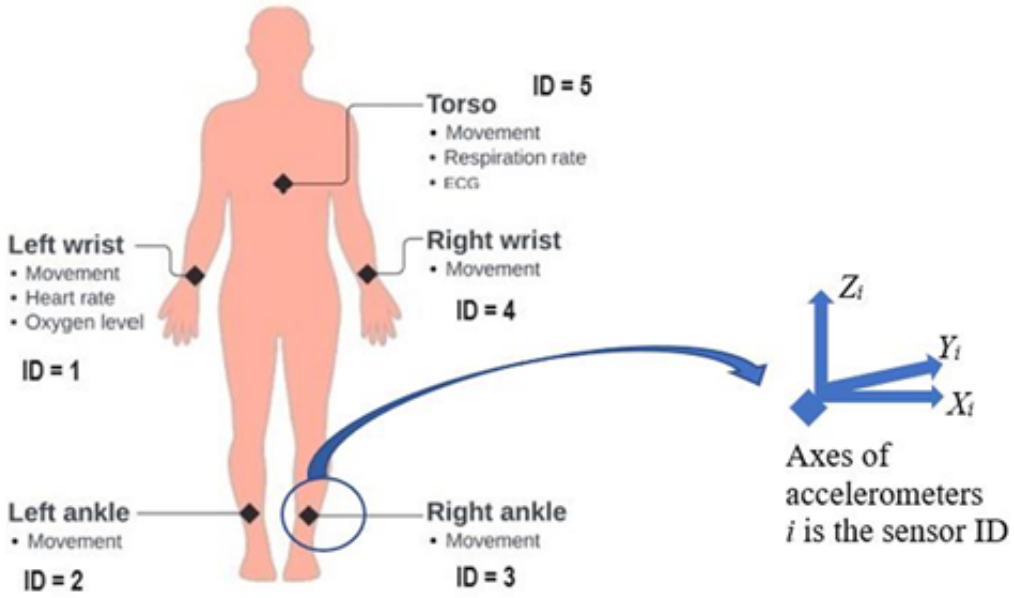

Figure 2.6: Sensor positioning of WWSN and the three axes of accelerometers.

2.1.4 3D Skeletal Reconstruction using Parametric Pose Angles

Traditional motion capture often relies on deriving absolute 3D coordinates (x, y, z) for individual joints. However, for applications requiring high anatomical fidelity, simple coordinate-based methods are insufficient as they lack surface information and rigid biomechanical constraints. Parametric skeletal reconstruction addresses these limitations by estimating pose angles of a digital skin. This approach ensures reconstructed motion adheres to physiological limits and serves as the foundation for "Digital Twin" technologies.

SMPLify (Skinned Multi-Person Linear Model)

The SMPL model is a vertex-based statistical model defining the human body surface. It is parametric, controlled by two low-dimensional vectors: Shape (β), representing coefficients derived from PCA, and Pose (θ), representing axis-angle rotations of joints in a kinematic tree. The skin consists of 6,890 vertices bound using Linear Blend Skinning (LBS) weights [7]. The effectiveness of this reconstruction approach is illustrated in Figure 2.7, which demonstrates the SMPLify model's performance across varying accuracy percentiles.

1. **Mathematical Formulation**
   The body mesh is a function of shape (β) and pose (θ):

   a) **Shape (β)**
   A vector of 10 coefficients derived from Principal Component Analysis (PCA) of body scans. Changing these values alters the subject's height, weight, and muscularity without affecting their pose.

   b) **Pose (θ)**
   A vector of 24×3 parameters representing the axis-angle rotation of the 23 joints relative to their parents in the kinematic tree, plus the global orientation of the root joint (Pelvis).

2. **Kinematic Tree & Skinning**
   The model utilizes a hierarchical kinematic tree (e.g., the shoulder moves the elbow, which moves the wrist). The "Skin" consists of **6,890 vertices** that are bound to this skeleton using **Linear Blend Skinning (LBS)** weights. This ensures that when the skeleton moves (driven by the reconstructed angles), the skin stretches and deforms realistically, providing the necessary surface data for regeneration analysis.

Figure 2.7: Evaluation of SMPLify reconstruction performance across varying accuracy percentiles

HMR (Human Mesh Recovery)

HMR utilizes a neural network to predict pose and shape coefficients from a single image in one step. This confirmed real-time skeletal reconstruction while preserving human anatomy [8]. The qualitative capabilities of this end-to-end regression are presented in Figure 2.8, demonstrating the recovery of 3D meshes from unconstrained RGB images.

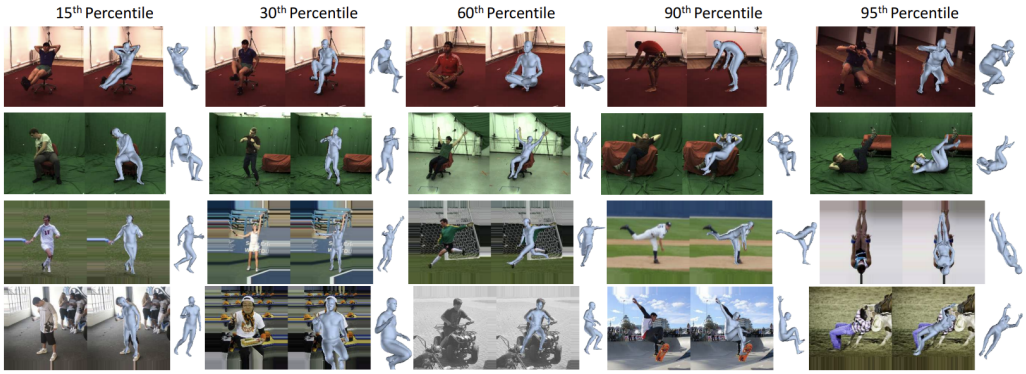

Figure 2.8: Qualitative results of Human Mesh Recovery (HMR) on unconstrained monocular RGB images

ROMP (Regressing Monocular 3D Pose and Shape)

While SMPL provides the container for the data, ROMP is the engine used to extract that data from monocular images. It is a "one-stage" regression network designed to operate efficiently on local GPU hardware[9].

1. **Single-Stage Architecture**
   Unlike older "two-stage" methods (like SMPLify) that first detect 2D keypoints

and then separately optimize a 3D model to fit them—a slow and iterative process—ROMP utilizes a ResNet-50 backbone to perform end-to-end regression. It takes an input image and simultaneously predicts the camera parameters, body center, and SMPL parameters in a single forward pass.

2. **Pixel-Level Center Map**
   ROMP distinguishes itself by treating the 3D body center as a pixel-level representation. It generates a "Center Map" (heatmap) to locate subjects even in crowded scenes.

3. **Parameter Map Construction:** For every detected body center, ROMP samples a feature vector that is decoded into:

   a) **3D Pose (θ):** The rotation matrices for all 24 joints.

   b) **Body Shape (β):** The coefficients describing the user's unique anatomy.

The complete one-stage pipeline, from the backbone feature extraction to the final parameter regression, is visualized in Figure 2.9, highlighting the pixel-level localization strategy.

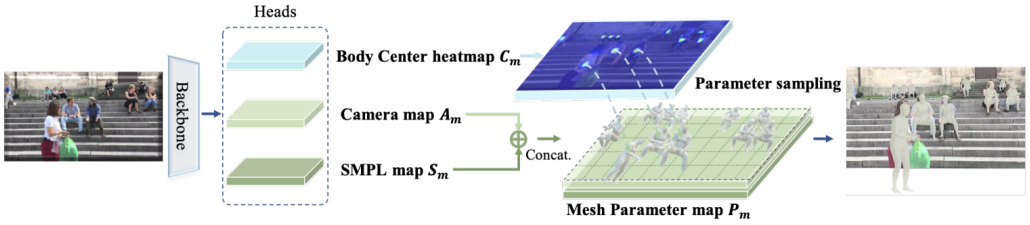

Figure 2.9: The ROMP One-Stage Regression Architecture for 3D Mesh Recovery

2.1.5 3D Skin Generation and Skeletal Binding

Human digitization has evolved from primitive skeletal models to complex, mesh-based representations containing both identity and motion. Current research in 3D skin generation and skeletal binding is dominated by three major technological pillars:

1. **Parametric Body Modeling**
   The most influential progression in skin creation is the **Skinned Multi-Person Linear (SMPL)** model [7]. Other than the stick-figure skeleton model, SMPL is a vertex-based representation of simplified human bodies that provides a low-dimensional parameter space to describe human body shapes. It deforms a generic template with shape parameters β to correspond to distinctive physical proportion and pose parameters θ that define joint rotations. The visual separation of these parameters is depicted in Figure 2.10, illustrating how the model mathematically decouples body shape (identity) from skeletal pose (motion).

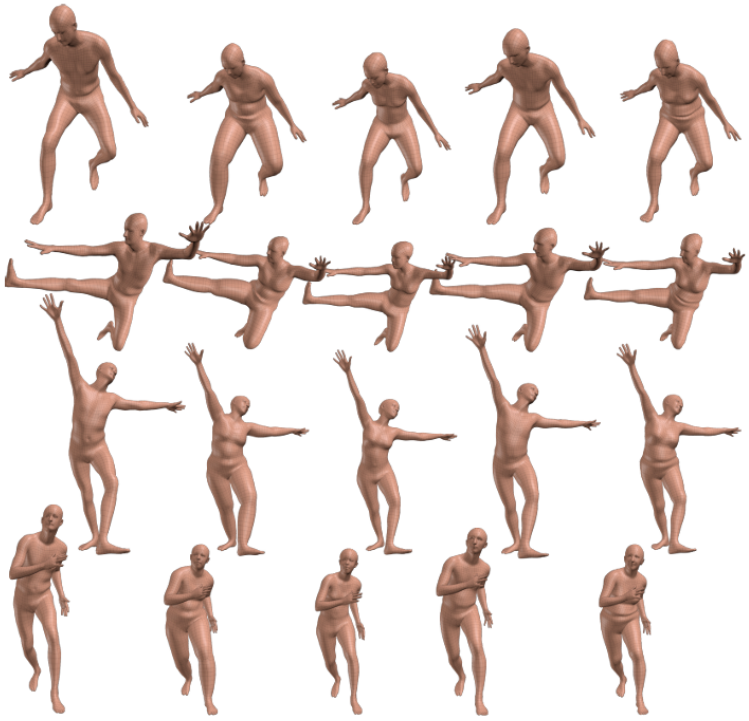

Figure 2.10: Decomposition of the SMPL model into shape (β) and pose (θ) parameters

2. **Automated Mesh Recovery**
   The alternative to manual character creation is solved with use of the so called **Human Mesh Recovery (HMR)** methods. Research by Kanazawa et al. [8] showed that it is possible to directly regress SMPL parameters from a single RGB image using end-to-end deep learning.This regression capability is visualized in 2.11, showing the model's output on diverse input images. **Pixel-aligned Implicit Functions (PIFu)** [10] go further in this direction, enabling digitization of "clothed" humans and obtaining high-resolution texture and clothing folds that parametric models generally fail to represent. The ability of PIFu to capture these fine-grained surface details and clothing deformations is demonstrated in 2.12.

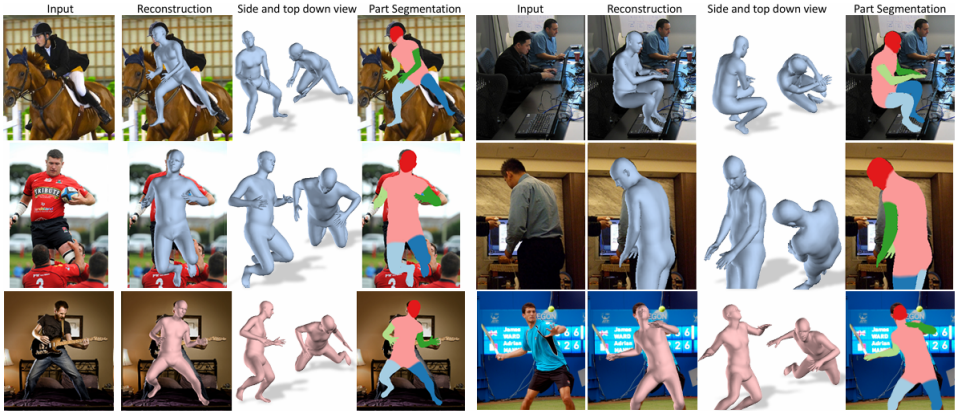

Figure 2.11: Real-time 3D Human Mesh Recovery (HMR) from a single RGB image

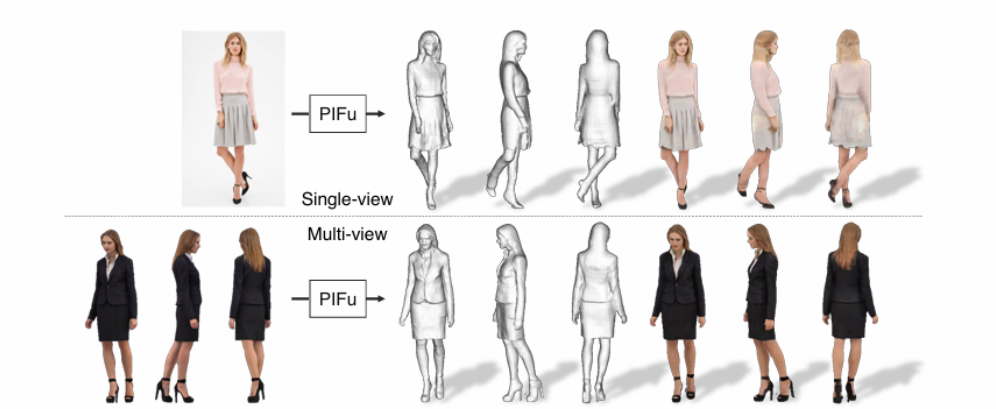

Figure 2.12: High-resolution surface reconstruction using Pixel-aligned Implicit Functions

3. **Vertex-to-Bone Mathematical Binding**
   To ensure that the generated skin can deform with the skeleton without visual artifacts or slippage, Linear Blend Skinning (LBS) is used. LBS was developed as skeletal subspace deformation. While basic LBS can suffer from volume loss, the foundational work in Pose Space Deformation (PSD) [11] created the mathematical means to connect the weighted influence of mesh vertices to the motions of the bones. This balances the surface deformation to the active bone movements while reducing joint distortion artifacts. The improvement in volume preservation offered by this technique compared to standard skinning is illustrated in Figure 2.13.

Figure 2.13: Comparison of joint volume loss in standard Linear Blend Skinning (LBS) vs. the corrected Pose Space Deformation

2.1.6 Validation for Accelerometer-Based Activity Reconstruction

A primary objective of this research is to validate the proposed activity reconstruction model through comprehensive assessment of its accuracy, consistency, plausibility, and reliability. To achieve rigorous validation, this study adopts a hierarchical framework comprising four distinct levels, each addressing a different dimension of reconstruction quality. This multi-tiered approach is particularly critical for accelerometer-based systems, where inherent challenges such as sensor drift and the limited observational scope of inertial measurement units compared to optical motion capture systems necessitate evaluation beyond conventional geometric accuracy metrics alone.

i. **Geometric Accuracy Assessment:**
Geometric validation forms the foundation of motion reconstruction evaluation, primarily relying on Euclidean distance metrics to quantify positional deviations from ground truth data obtained through optical motion capture systems[12].

   a) **MPJPE (Mean Per Joint Position Error):** The average distance (in millimeters) between predicted and true joint coordinates.

   b) **PA-MPJPE (Procrustes-Aligned MPJPE):** This rigid alignment removes errors related to global rotation and scale, focusing purely on the "pose" structure.

   c) **PVE (Per Vertex Error):** For full avatars (meshes), this measures the error of every vertex on the body surface, capturing body shape (e.g., fat/muscle) accuracy better than just joints.

ii. **Physical Plausibility Verification:**
Given that accelerometers fundamentally measure force-related quantities governed by Newtonian mechanics, validation must extend beyond geometric correspondence to encompass physical realism[13].

   a) **Foot Sliding (FSR):** A critical metric for generative models. It measures the distance feet travel while they should be planted on the ground. High foot sliding is a primary indicator of poor reconstruction quality.

   b) **PhysCap & NeuralPhysCap:** These methods validate motion by running it through a physics simulator (like MuJoCo). If the simulator cannot replicate the motion without applying impossible external forces (e.g., invisible strings pulling the puppet), the reconstruction is flagged as "implausible".

   c) **Center of Mass (CoM) Stability:** This verifies if the avatar is balanced or if it would physically fall over given the reconstructed pose.

iii. **Perceptual Quality Evaluation:**
Recent scholarship acknowledges that mathematical accuracy does not guarantee perceptually natural motion, addressing the potential for technically correct but visually uncanny results[12].

   a) **LPIPS (Learned Perceptual Image Patch Similarity):** A deep-learning metric used to compare rendered frames of the avatar against real video. It correlates better with human visual perception than pixel-level errors.

   b) **FID (Fréchet Inception Distance):** Measures the distance between the distribution of generated motions and real human motions. A lower FID means the motion "looks" more natural and human-like.

   c) **PP-Motion:** A newly proposed metric (2025) that combines physical constraints with human perceptual scores to give a single "Fidelity" rating.

iv. **Reliability and Uncertainty Quantification:**
The most sophisticated validation tier addresses model confidence calibration and epistemic uncertainty. This reliability assessment is particularly pertinent for accelerometer-based systems operating in unconstrained environments where activity diversity may exceed training data coverage[14].

a) **Expected Calibration Error (ECE)** measures the alignment between predicted confidence levels and actual accuracy rates, with elevated ECE values signaling unreliable confidence estimates.

b) **Epistemic uncertainty quantification** identifies knowledge gaps in the model's training distribution, proving essential for detecting when the system encounters novel activities beyond its learned repertoire.

2.2 Gaps in Literature

2.2.1 Developing the Models using Accelerometers to Estimate the Joint Postures and the Orientations

There are several gaps remaining in accelerometer-only based motion human activity reconstruction existing research. First, dependence on large fixed motion databases like AMASS limits the generalizationabilty and the quality of reconstruction of out-of-distribution motions as they fail to reconstruct complex and novel movements that were not present in the training data which leads to stiff motions. Second, optimization and retrieval methods being used in earlier approaches for motion reconstruction require high computational costs limiting the applicability in lightweight systems. Additionally, existing accelerometer based methods face ambiguities in rotational pose estimation and require very strong assumptions in the case of missing information when it comes to yaw rotations. The fine grained motion information, global translation estimation and adaption capability for different human morphologies and complex motions are also less represented in the existing research. These facts point towards more sophisticated methods being designed for more effective and efficient research in human activity reconstruction using accelerometers[4].

2.2.2 Parametric Reconstruction and Anatomical Binding

While notable progress has been achieved separately in the areas of 3D surface modeling and inertial pose estimation, effectively combining these two technologies remains a major challenge—particularly when operating with sparse sensor setups.The following sections outline the specific deficiencies in current research that this study aims to address.

Metric Scale and Manual Measurement Ambiguity Current regression-based approaches such as ROMP[9] and HMR[8] are capable of reliably estimating 3D body shape and proportions; however, their outputs are typically expressed in a unit-less coordinate space. Most existing works focus on visual pose appearance rather than absolute metric correctness. This limitation is a major unresolved

challenge for clinical applications, such as skin regeneration, where real-world measurements are essential. Moreover, existing sparse inertial systems such as the Fast Inertial Poser[2] still rely heavily on manual measurements. Users are required to input pre-measured limb lengths and biological parameters to properly calibrate the skeletal model. Currently there is no fully autonomous pipeline that bridges automated data derived from single-perspective visual inputs with the real-time requirements of sparse inertial sensing

Biometric Mismatch and Kinematic Drift A recurring problem in motion reconstruction is the Body Size Mismatch caused by the use of "standard" digital templates. Research shows that even slight differences between a digital skeleton's segment lengths (e.g., the humerus) and a user's actual anatomy can cause the Inverse Kinematics (IK) solver to compute inaccurate joint angles[15]. This often leads to kinematic drift and visually unrealistic artifacts, like feet sinking through the floor. Current studies do not offer a reliable method for automatically resizing a digital skeleton to match a user's specific limb lengths using just a single-photo calibration before inertial tracking.

Volume Preservation and Physiological Binding Gaps The final challenge lies in maintaining the visual realism of a 3D surface during extreme motions. Conventional techniques for skeletal binding, like Linear Blend Skinning (LBS), often struggle with "volume loss"[11]. This results in joints—particularly elbows and knees—appearing to collapse or form a pinched, "candy-wrapper" effect when bent deeply. While generic corrective models exist, they fail to consider the unique muscle and soft-tissue volumes of an individual. Currently, there is no integrated approach that leverages a user's initial photograph to guide the 3D model in realistically stretching and folding the skin according to that person's specific anatomy.

Table 2.1: Summary of Identified Gaps in Literature

Gap Category	**Current Limitation**	**Impact on Reconstruction**
Metric Ambiguity	Models are unit-less and require manual height/limb entry.	Lack of clinical accuracy in skin regeneration.
Biometric Mismatch	Reliance on generic "average" skeletons.	Calculated joint angles are mathematically incorrect.
Binding Artifacts	Standard LBS causes joint collapse and volume loss.	Unrealistic visual representation during movement.

Chapter 3

Methodology

3.1 Research Design

This research follows a structured approach to human activity reconstruction. First, joint postures and orientations are estimated using sensor data and computational models. In parallel, a 3D skeletal model is reconstructed, and a virtual avatar is generated to represent human motion. These two outputs are integrated to form a complete motion representation. Finally, the proposed system is validated using experimental data, and its performance is evaluated based on accuracy, consistency, and reliability metrics. Following Figure 3.1 illustrates the proposed methodology.

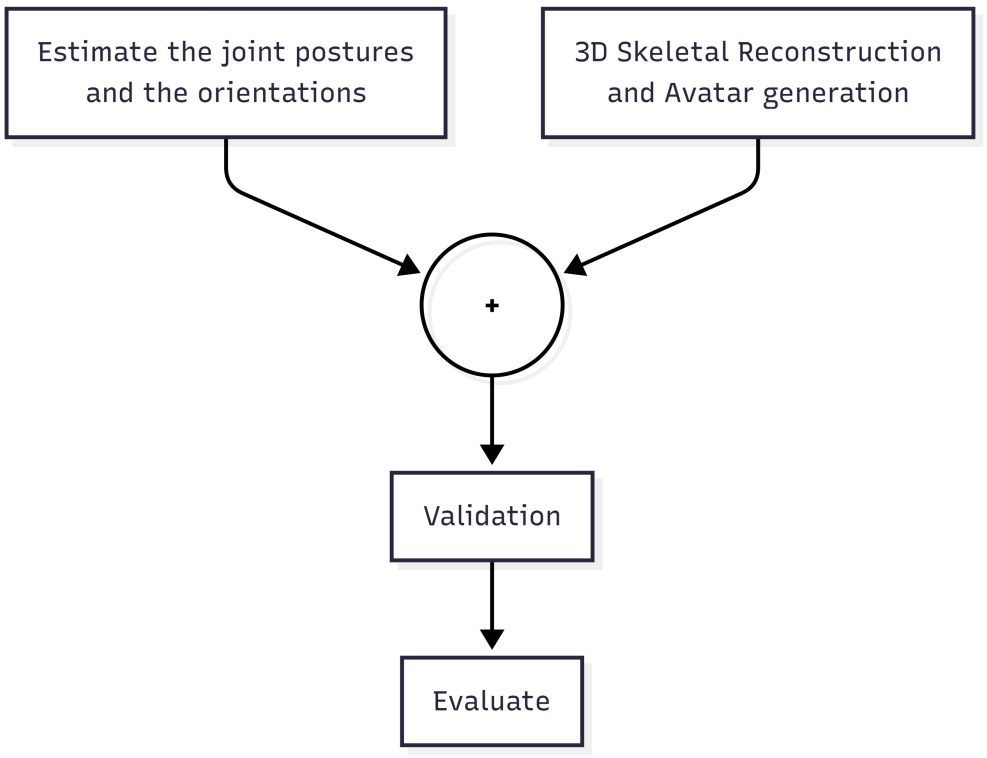

Figure 3.1: Block diagram of the proposed methodology

Estimate the joint postures and the orientations

Since joint postures and the orientations cannot be directly obtained by only using accelerometer data, we are planning to work on different approaches at the same time as follows. Following Figure 3.2 shows the flowchart of the estimate the joint postures and the orientations.

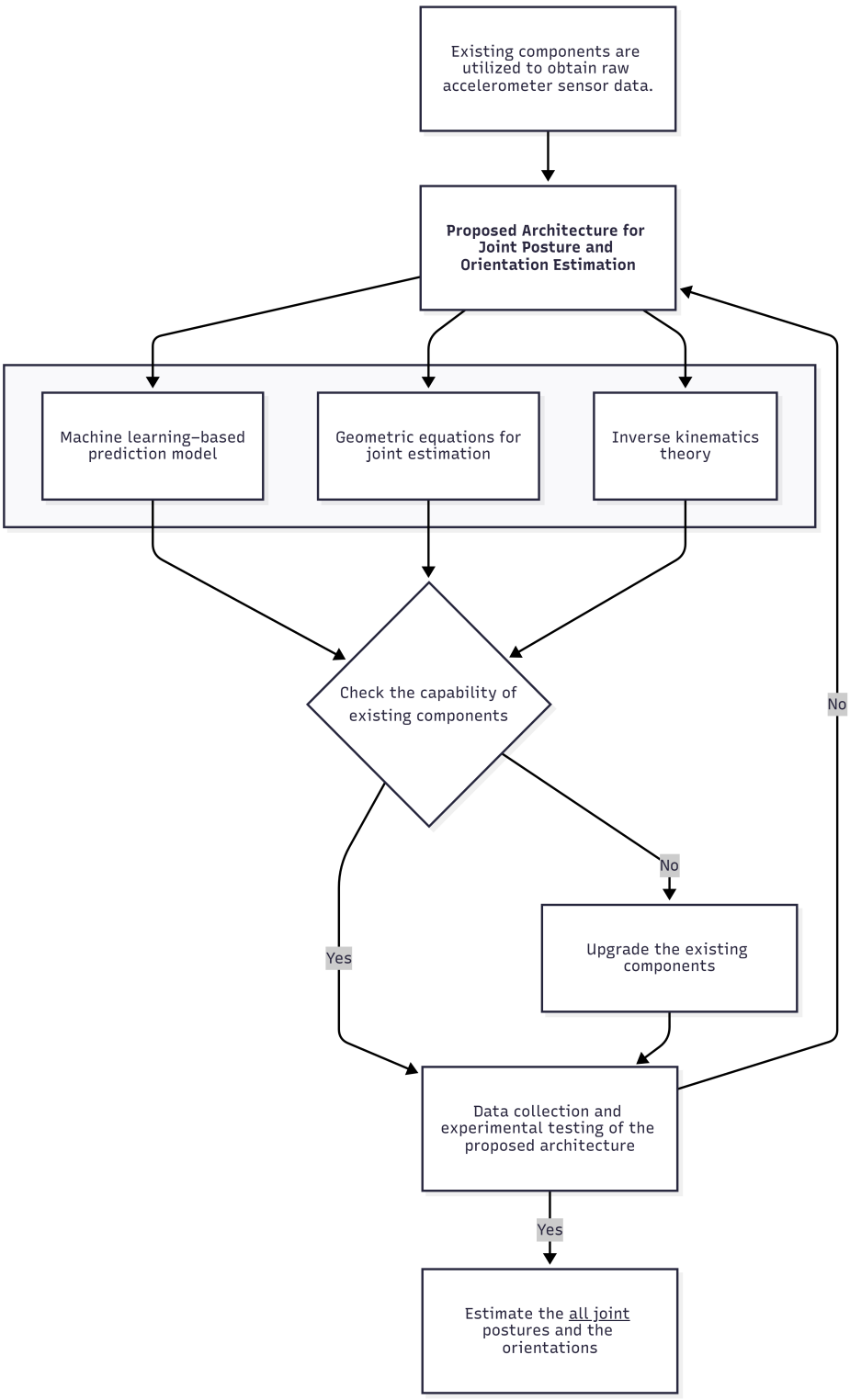

Figure 3.2: Flowchart of the Estimate the joint postures and the orientations

The process commences with directly employing the available hardware components for retrieving raw data from wearable devices pertaining to the accelerometer sensor. The hardware components are solely responsible for providing initial inertial measurements to facilitate the analysis of joint posture and orientation. Wearable sensor-retrieved raw data constitutes the input for the proposed system addressing the estimation of joint posture and orientation by employing the strategy of integrating various estimation techniques.

Three parallel strategies are incorporated within the proposed architecture for estimating joint posture parameters:

1. **Machine Learning Prediction:** A prediction model using machine learning algorithms has been employed for patterns within the accelerometer data and estimating joint posture and orientation information.

2. **Geometrical Equations:** Geometrical equations for joint estimation, considering relative orientations among the sensors and gravity constraints, have been employed for joint parameter derivation.

3. **Inverse Kinematics:** The theory of inverse kinematics for the refinement of joint orientation estimates, based on anatomical and biomechanical constraints, has been introduced.

The parallel strategies illustrated above permit a comparative assessment and enhance the robustness of the posture estimation process. The results are assessed based on sensor accuracy, computational ability, and estimation robustness. If inadequate, the system upgrades existing elements, such as optimizing sensor locations or sampling. Once standards are satisfied, data collection and experimentation take place to accomplish a comprehensive joint posture estimation system.

Avatar Generation and Animation

In human activity reconstruction, the visualization layer must go beyond basic skeletal overlays to deliver a biologically realistic and metrically precise representation of the individual. Below Figure 3.3 shows the flowchart of the avatar generation and animation which we are planning to work on. A significant limitation identified in current research is the reliance on generic skeletal models that fail to account for individual variations in limb length and body mass.The primary objective of this subsystem is to develop a unified framework that generates a personalized "Digital Twin" and animates it using a high-fidelity skin deformation pipeline. Instead of manual rigging, this research automates the transition from a 2D image to a moving 3D surface through three stages:

1. **Skeletal Reconstruction:** The system investigates Deep Regression methods (e.g., ROMP or HMR) to automatically predict skeletal parameters (θ) and shape coefficients (β) from a single monocular image in a single pass.

2. **Surface Modelling and Binding:** The avatar utilizes the Skinned Multi-Person Linear (SMPL) model, defining the surface with 6,890 vertices attached to a skeletal structure using Linear Blend Skinning (LBS).

3. **Animation Pipeline:** A real-time processing loop receives a standardized Pose Vector (θ) composed of axis-angle rotation parameters. This vector drives the LBS engine to apply rotations to the scaled skeleton with a target latency of less than 15ms .

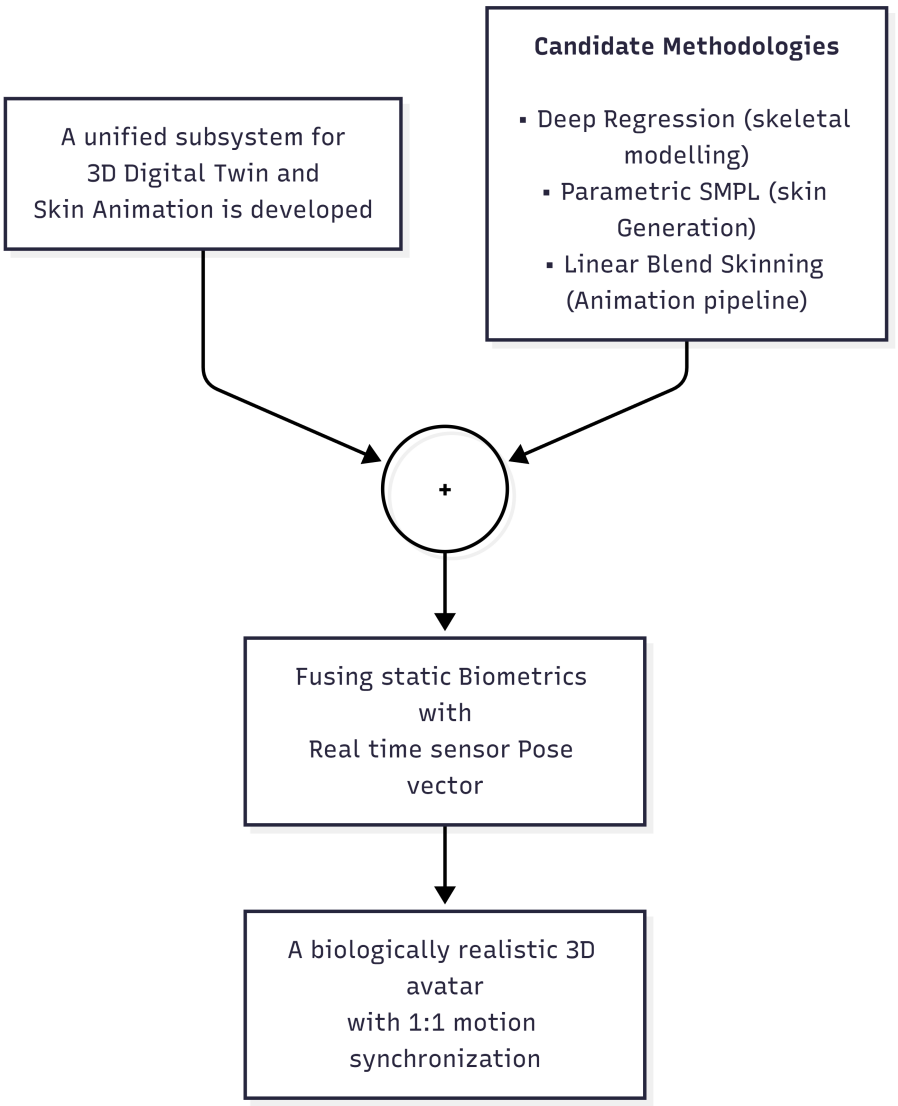

Figure 3.3: Flowchart of the Avatar Generation and Animation

Validation

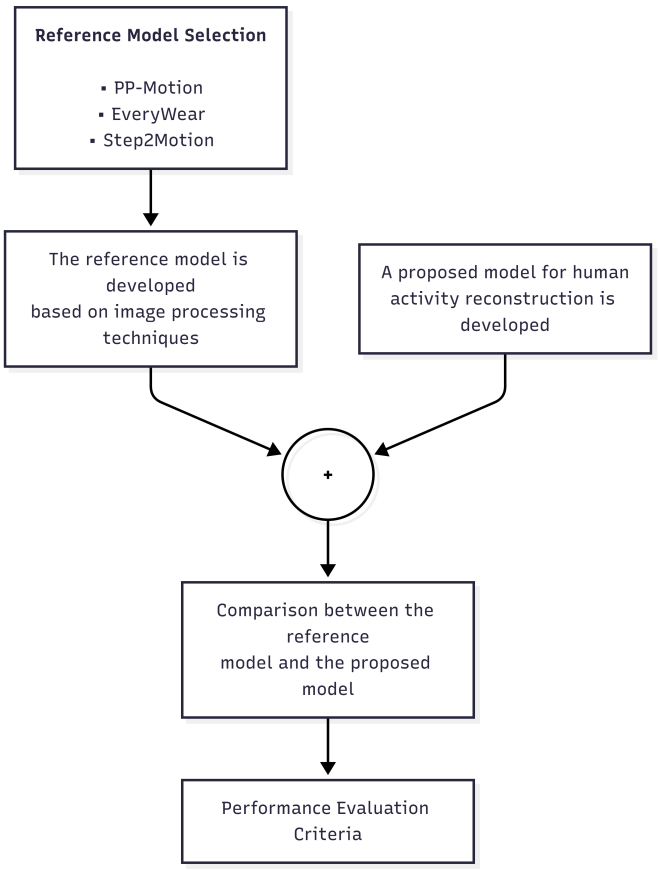

Figure 3.4: Flowchart of the Validation

In human action reconstruction, an ideal system is expected to follow a hierarchical framework comprising four distinct levels as shown in the above Figure 3.4. However, to date, no existing model fully satisfies all four levels of this framework. The primary objective of our methodology is not to validate or improve an existing model, but rather to develop a new model that addresses the identified limitations in current approaches.

Therefore, instead of validating our system against an ideal but unavailable benchmark, we validate our proposed model through comparative analysis with well-established and validated models that are currently used in real-world applications. Based on our literature review, we identified several relevant models for this purpose, including

1. PP-Motion[12]
2. EveryWear[13]
3. Step2Motion[14]

Among these, we select a representative model that employs computer vision and image processing techniques and compare its performance with our proposed

model. The comparison focuses on key evaluation criteria such as accuracy, consistency, plausibility, and reliability. This comparative validation allows us to assess the effectiveness of our approach while positioning our model within the context of existing, validated systems.

3.2 Data Collection

Accelerometer Based Sensor Readings Collection

1. Five 3-axis accelerometer sensors are placed on key body segments: both wrists, both ankles, and the torso.
2. Each sensor measures linear acceleration along three orthogonal axes (x, y, z), including both gravitational and motion-induced acceleration.
3. Prior to acquisition, sensor calibration corrects offset and scale errors. Sensors are securely attached to minimize relative motion.
4. Sensor axes are aligned with anatomical reference frames, and signals are sampled at a fixed frequency as continuous time-series data.
5. Predefined activity sets ensure consistency across data samples.

Biometric and Kinematic Data Integration

This module defines the inputs required to build the Digital Twin:

1. **Visual Calibration Data:** A single high-resolution monocular RGB image of the subject in a neutral A-pose, resized to 512×512 pixels.
2. **Manual Biometric Metadata:** The subject's true height (H_{real}) is used for Height-Anchor Scaling to transform unit-less bone lengths into centimeter measurements.
3. **Kinematic Interface:** The system receives 24×3 axis-angle rotation parameters from the Motion Reconstruction Module.
4. **Synchronization:** The avatar initializes in a "Zero Pose" at $t = 0$ to align the virtual skeleton and mesh with calibrated sensor orientations.

Chapter 4

Timeline and Resource Required

4.1 Timeline

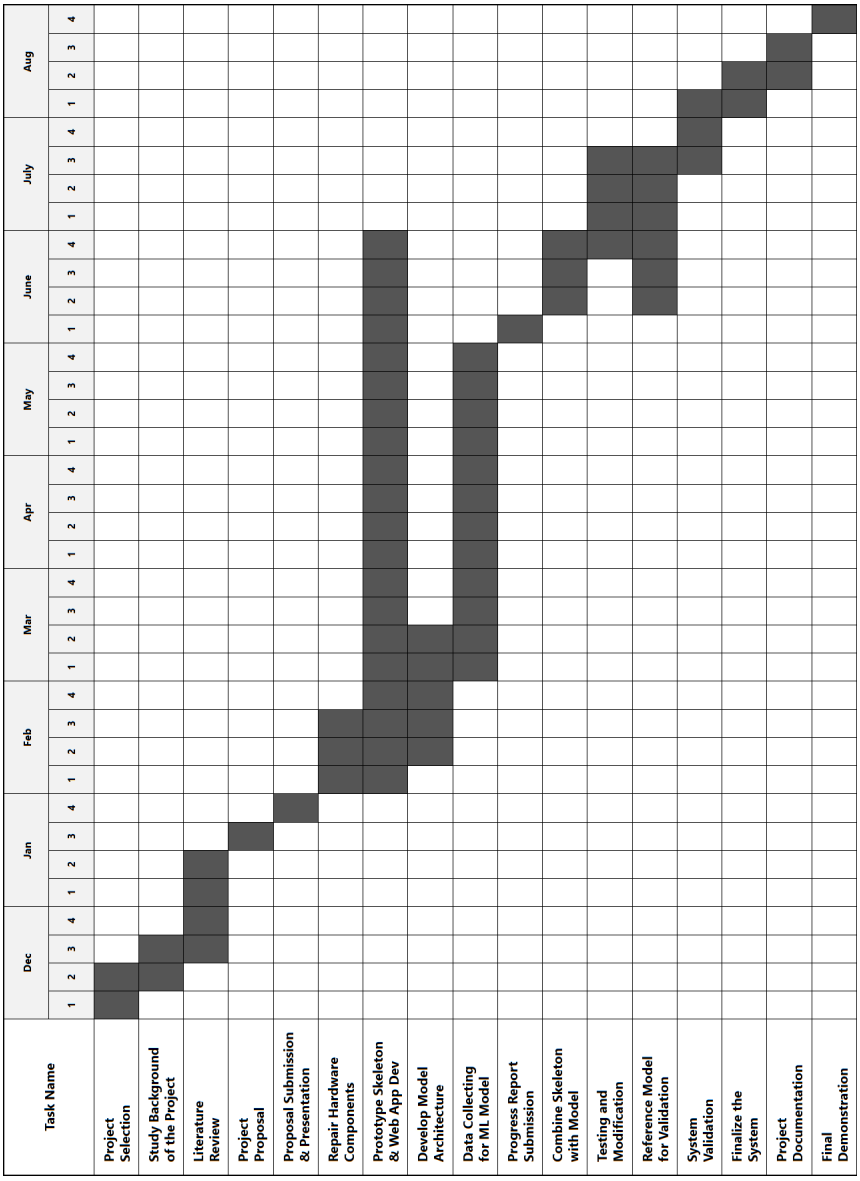

Figure 4.1: TimeLine

4.2 Resource Required

This is a general estimation of budget for this project. These prices can be vary since the exact electronics/ services are not decided yet.

Table 4.1: Estimated Budget for Project Components and Services

No	**Item / Service**	**price**
1	Repair for existing components	5 000
2	Web hosting	15 000
3	ML training platform	8 000
4	Cloud computing	20 000
5	API s	10 000
	Total	**58 000**

Chapter 5

Conclusion

This project proposal outlines a comprehensive framework for accomplishing high fidelity human activity reconstruction with a sparse configuration of only five 3-axis accelerometers. By strategically positioning these sensors on the wrists, ankles, and torso, the suggested system addresses the critical need for a low-cost, minimally intrusive, and portable alternative to standard motion capture technologies which often rely on dense, expensive, and complex sensor arrays.

The methodology uniquely bridges the gap between raw inertial data and realistic visualization by combining three parallel estimation strategies: machine learning based pattern recognition, geometric orientation equations, and inverse kinematics theory. This combined approach serves to maintain robustness of the joint posture and orientation estimation despite ambiguities in the accelerometer-only data.

A significant contribution of this research is the development of a personalized "Digital Twin" through an automated 3D avatar generation pipeline. The system eliminates the need for generic digital templates and manual measurements by utilizing deep regression models like ROMP or HMR, as well as the Skinned Multi Person Linear (SMPL) model. This provides metric accuracy and biologically realistic skin deformation, which successfully reduces common visual defects such as joint volume loss and kinematic drift.

Finally, the project implements a robust four-tier validation framework that includes geometric, physical, perceptual, and reliability assessments - to ensure the system's performance fulfils real-world expectations. This project's effective integration of sparse inertial sensing with advanced computer vision and biomechanical modeling, results in a scalable and efficient solution for applications in healthcare, sports analysis, and immersive virtual environments.

References

[1] X. Yi, Y. Zhou, and F. Xu, "Transpose: Real-time 3d human translation and pose estimation with six inertial sensors," *ACM Transactions On Graphics (TOG)*, vol. 40, no. 4, pp. 1–13, 2021.

[2] X. Xiao, J. Wang, P. Feng, A. Gong, X. Zhang, and J. Zhang, "Fast human motion reconstruction from sparse inertial measurement units considering the human shape," *Nature Communications*, vol. 15, no. 1, p. 2423, 2024.

[3] Y. Huang, M. Kaufmann, E. Aksan, M. J. Black, O. Hilliges, and G. Pons-Moll, "Deep inertial poser: Learning to reconstruct human pose from sparse inertial measurements in real time," *ACM Transactions on Graphics (TOG)*, vol. 37, no. 6, pp. 1–15, 2018.

[4] J. Tautges, A. Zinke, B. Krüger, J. Baumann, A. Weber, T. Helten, M. Müller, H.-P. Seidel, and B. Eberhardt, "Motion reconstruction using sparse accelerometer data," *ACM Transactions on Graphics (ToG)*, vol. 30, no. 3, pp. 1–12, 2011.

[5] P. Kelly, C. Ó Conaire, and N. E. O'Connor, "Human motion reconstruction using wearable accelerometers," 2010.

[6] S. Gunawardena, P. Amarasingha, S. Gunawardhana, and S. Sabaragamuwa, "Wireless sensor network aided human activity detection using artificial intelligence," June 2025. Preprint on ResearchGate.

[7] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "Smpl: A skinned multi-person linear model," *ACM Transactions on Graphics (TOG)*, vol. 34, no. 6, pp. 1–16, 2015.

[8] A. Kanazawa, M. J. Black, D. W. Jacobs, and E. Enqvist, "End-to-end recovery of human shape and pose," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7122–7131, 2018.

[9] Y. Sun, Q. Bao, W. Liu, W. Gao, Y. Fu, M. Chen, and T. Mei, "Monocular, one-stage, regression of multiple 3d people," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11179–11188, 2021.

[10] S. Saito, Z. Huang, R. Natsume, S. Morishima, A. Kanazawa, and H. Li, "Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2304–2314, 2019.

[11] J. P. Lewis, M. Cordner, and N. Fong, "Pose space deformation: a unified approach to shape interpolation and skeleton-driven deformation," in *Proceedings of the 27th annual conference on Computer graphics and interactive techniques (SIGGRAPH)*, pp. 165–172, 2000.

[12] S. Zhao, Z. Wang, T. Luan, J. Jia, W. Zhu, J. Luo, J. Yuan, and N. Xi, "Pp-motion: Physical-perceptual fidelity evaluation for human motion generation," *arXiv preprint arXiv:2508.08179*, 2025.

[13] S. Zhu, Y. Li, J. Li, Q. Wu, Z. Wang, H. Ma, and W. Liang, "Human motion estimation with everyday wearables," *arXiv preprint arXiv:2512.21209*, 2025.

[14] J. L. Ponton, E. Alvarado, L. G. Foo, N. Pelechano, C. Andujar, and M. Habermann, "Step2motion: Locomotion reconstruction from pressure sensing insoles," *arXiv preprint arXiv:2510.22712*, 2025.

[15] P. Molloy, Y. Zhou, M. Habermann, and C. Theobalt, "The impact of anthropometric scaling on inertial motion capture accuracy," *Journal of Biomechanics*, vol. 152, p. 111571, 2023.