IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

Real-time Sit-to-Stand Phase Classification with a Mobile Assistive Robot From Close Proximity Utilizing 3D Visual Skeleton Recognition

Anas Mahdi[1], Zonghao Dong[1], Jonathan Feng-Shun Lin, Yue Hu, Yasuhisa Hirata, Katja Mombaur

***Abstract*—Sit-to-stand (STS) transfer is a fundamental but challenging movement that plays a vital role in older adults' daily activities. The decline in muscular strength and coordination ability can result in difficulties performing STS and, therefore, the need for mobility assistance by humans or assistive devices. Robotics rollators are being developed to provide active mobility assistance to older adults, including STS assistance. In this paper, we consider the robotic walker SkyWalker, which can provide active STS assistance by moving the handles upwards and forward to bring the user to a standing configuration. In this context, it is crucial to monitor if the user is performing the STS and adapt the rollator's control accordingly. To achieve this, we utilized a standard vision-based method for estimating the human pose during the STS movement using Mediapipe pose tracking. Since estimating a user's state from extreme proximity to the camera is challenging, we compared the pose identification results from Mediapipe to ground truth data obtained from Vicon marker-based motion capture to assess accuracy and reliability of the STS motion. The fourteen kinematic features critical for accurate pose estimation were selected based on literature review and the specific requirements of our robot's STS method. By employing these features, we have implemented a phase classification system that enables the SkyWalker to classify the user's STS phase in real-time. The selected kinematics from vision-based human state estimation method and trained classifier can be furthermore generalized to other types of motion support, including adaptive STS path planning and emergency stops for safety insurance during STS.**

Manuscript received: October, 7, 2024; Accepted December, 21, 2024.

This paper was recommended for publication by Editor Angelika Peer upon evaluation of the Associate Editor and Reviewers' comments.

Anas Mahdi, Jonathan Lin, and Katja Mombaur are with the Department of System Design Engineering, CERC Human-Centred Robotics and Machine Intelligence, University of Waterloo, Waterloo, Canada {a7mahdi, jf2lin, katja.mombaur}@uwaterloo.ca

Yue Hu is with the Department of Mechanical Engineering, Active & Interactive Robotics Lab (A.I.R.), University of Waterloo, Canada yue.hu@uwaterloo.ca

Katja Mombaur is with the Karlsruhe Institute of Technology (KIT), Institute of Anthropomatics and Robotics (IAR), Optimization and Biomechanics for Human-Centred Robotics, Karlsruhe, Germany

Zonghao Dong and Yasuhisa Hirata are with the Department of Robotics, Tohoku University, Sendai, Japan {z.dong, hirata}@srd.mech.tohoku.ac.jp

[1] Anas Mahdi and Zonghao Dong are equal first authors of this paper.We gratefully acknowledge funding from the Canada Excellence Research Chair program, the Japan Science and Technology Agency Moonshot R&D Program (JPMJMS2034), the Japan Society for the Promotion of Science KAKENHI (22J10961), and the Tohoku University Graduate Program for Integration of Mechanical Systems. We have obtained ethics approval from the University of Waterloo ethics board to conduct human-involved experiment with SkyWalker.

Digital Object Identifier (DOI): see top of this page.

I. INTRODUCTION

The Sit-To-Stand (STS) action is vital for elderly individuals, facilitating essential daily activities and significantly contributing to their independence and physical fitness [1], [2]. However, executing STS requires complex coordination of lower extremity muscles—a significant hurdle due to age-related muscle weakness and balance issues [3], [4]. These challenges often result in increased reliance on others, adversely affecting mental health and well-being [5], [6].

To mitigate these difficulties, various support methods are employed. Chairs with armrests and mobility aids like walkers and canes provide essential assistance, albeit they may not suffice for those with severe impairments [7], [8]. Innovatively, robotic devices equipped with sensors emerge as a superior solution, offering enhanced mobility and safety by dynamically adjusting support in response to the user's movements, thus preventing falls [9], [10]. While promising, the full potential of these technologies and their integration into daily life warrants further exploration, indicating a significant avenue for future research.

In recent years, advanced artificial intelligence algorithms using instant automatic landmark identification with 2D or 3D images created a new path for real-time markerless human motion recognition, which has become potentially popular with several low-cost customer-grade camera systems [11]. Such algorithms enable the possibility of identifying a full-body skeleton structure describing the posture of a human subject within a given image frame [12]. The research group from Carnegie Mellon University released an image-processing framework called OpenPose, which takes the color image from an RGB camera as input and outputs the recognition of skeletons for one or multiple persons in the same scene [13]. Another famous machine-learning solution is Mediapipe, a high-fidelity body pose recognition framework proposed by Google [14]. Several depth camera hardware like Azure Kinect or Kinect V2 has its own software for providing human body landmarks. These new technologies provide a helpful way for researchers to measure STS features easily. Hsiao et al. used the Kinect system to measure the dynamic balance, and forward motion of the elderly [15]. Aguirre et al. implemented a Kinect depth sensor to determine which STS features a connection with fatigue [16]. They also compared different machine learning models for estimating fatigue in an STS exercise based on the visual measurement of human skeleton movement [17].

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

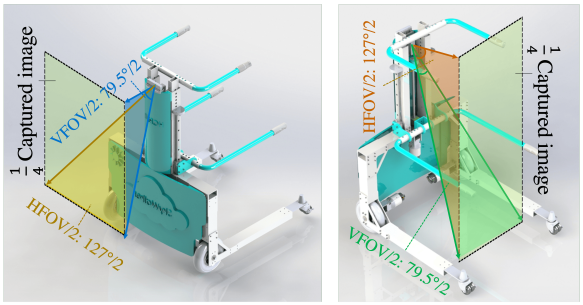


Fig. 1: Field of view (FOV) for navigation camera (left image) and FOV for human phase classification (right image)

Although many promising results have been reported throughout previous research, the human motion recognition using visual-based skeleton tracking on a mobile assistive robot is still facing challenges. Since users always need to keep close to the robot (less than 1 meter), the visual analysis data could be noisy and become not so accurate [18]. Taghvaei et al. reported this challenge while using a Kinect sensor for human state estimation on the assistive robot called RT walker [19]. The human skeleton tracking can also get lost when only part of the human body can be viewed from the camera. Another challenge would be computational cost-effectiveness since the current most fashionable models like OpenPose or Mediapipe still require huge computing resources on host PCs [20], making it difficult to acquire human skeleton information in real time while using low-cost hardware.

Implementing a vision-based approach for recognizing human motion regardless of close distance restriction could be promising because it is less expensive compared to other real-time detection methods. It requires less hardware and is more practical to implement on an assistive robot.

In this paper, we introduce an STS assistive robot with real-time human motion recognition functions from close proximity to the camera.

Our objective in implementing real-time phase classifier is to enable the robot to adjust its action based on the user's current state, ensuring safe standing up and preventing falls.

Our contributions in this paper:

- We propose a real-time, visual-based software for classifying the STS phases from close proximity, designed to be used with assistive devices and run on board. The software utilizes a 3D visual of the human skeleton model obtained from MediaPipe using a wide-angle depth camera.
- We compare the 3D skeleton tracking result with a motion capture system to analyze both the kinematics accuracy and similarities for the trend of changes
- We evaluate different machine learning algorithms including support vector machine (SVM), decision tree (DT), random forest (RF), and XGBoost for estimating STS motion phases.
- We validate the proposed approach through a challenging experiment with the mobility assistance robot SkyWalker during which the user is getting very close to the camera and parts of their recognized body skeleton resulting in larger distortion.

This paper is organized as follows: In Section II, we introduce the previous related works. In Section III, we introduce the SkyWalker architecture and how the robot support STS motion. In Section IV, we propose a visual-based phase classification method using Mediapipe and how we acquire the human kinematics data from the perceived skeleton model. We discuss the kinematics features accuracy through a biomechanical evaluation in Section V. In Section VI, We applied those features with different types of classifier and discuss the result of classification. Finally, we conclude our paper in Section VII.

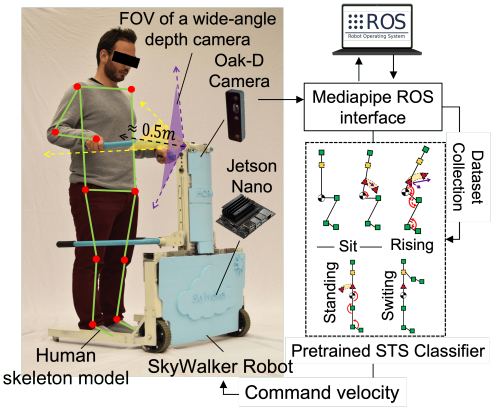


Fig. 2: Pipeline for the system architecture presented in this paper. The subject is around 0.5 m away from the depth camera.

II. The SkyWalker Mobility Assistance Robot

A. Robot hardware system

The SkyWalker is a mobility assistance robotic walker, designed to aid elderly individuals in their walking and standing up/sitting down movements [21]. The SkyWalker robot is designed for people who can perform STS with some degree of independence but require additional body weight support to ensure safety and stability.

Walking motions are supported via the front wheels which are actuated by two brushless motors, while the rear passive caster wheels are smaller to enable easy maneuverability.

The design features two sets of handles at different heights, which are adjustable to the user's size. The upper handles are static and are intended for use during upright standing and walking. The lower handles are designed for active STS (Sit-to-Stand) assistance, aiming to bring the user to about 2/3 of the way to upright standing. The vertical motion of the handles is powered by a brushless motor, while the forward motion during STS is produced by the powered wheels.

B. Robot support for STS

Fig. 3 shows the different phases of the STS support for the robot and the user, distinguishing three phases for the robot and four for the human.

To initiate STS motion from a sitting position, users place their hands vertically down on the lower set of handles. In this first phase, the robot is stationary. The STS movement of the robot starts by lifting the user, simultaneously providing vertical support using the vertical lift and horizontal support from forward motion by moving the chassis for a specific distance. At the same time the user lifts off the chair and starts

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

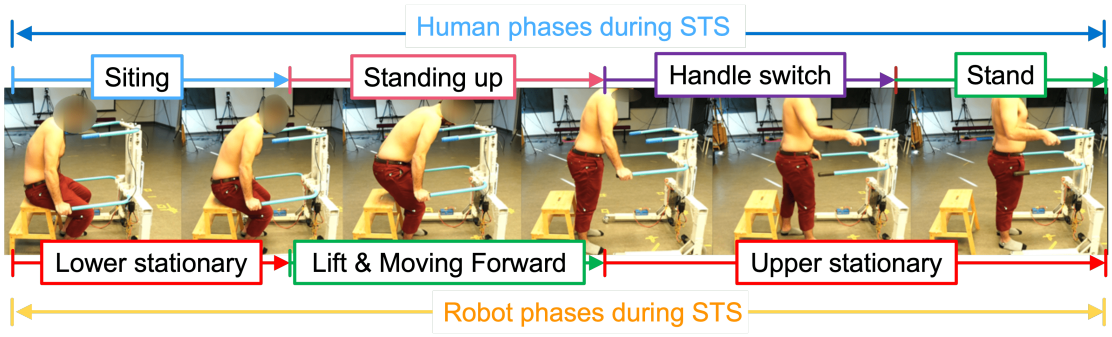


Fig. 3: Image sequence demonstrating STS assistance by the SkyWalker

stretching the legs and moving the thighs from a horizontal to a vertical orientation.

Once the lower handles have reached their highest position both forward and upward motion stop and the robot reaches another stationary phase. The user must now actively switch from the lower to the upper level handles, one hand at a time. Once this is done, the user has reached the last phase, standing upright in contact with both upper handles.

To guarantee user safety and coordinate the robot's motions it is important to check if the user interacts with the robot as planned and is able to perform the motions as expected.

C. Robot sensory system

The robot has multiple sensors to understand the user's intentions, phase, and navigation capabilities, which include an Alexa smart speaker for voice control and two depth/stereo cameras. One forward-facing depth camera is used for robot navigation and obstacle avoidance, while another rear-facing camera is used for the pose estimation of the user and phase classification as shown in Fig. 1. The cameras used are the Luxonis Oak-D W. The camera processor can perform real-time processing of the data and can run deep learning models without needing to send data to a remote server [22].

The integration of further sensors, such as force torque sensors in the handles will be evaluated in future research.

III. Methodology

This section outlines the methodology used in our study, including real-time vision-based human skeleton tracking in close proximity, a workflow for human key landmark acquisition, and algorithms for kinematic feature calculation.

A. Vision-based STS Skeleton Tracking in Close Proximity

The primary challenge of implementing a vision-based human state estimation system for assistive robotics, especially in the context of the SkyWalker robot, stems from the necessity of operating at extremely close proximities. Traditional vision-based systems typically perform well when there is sufficient distance between the camera and the subject, allowing for unobstructed views and clear data capture [23]. However, in the case of the SkyWalker robot, the user is required to stay within less than one meter of the camera. This proximity leads to several challenges.

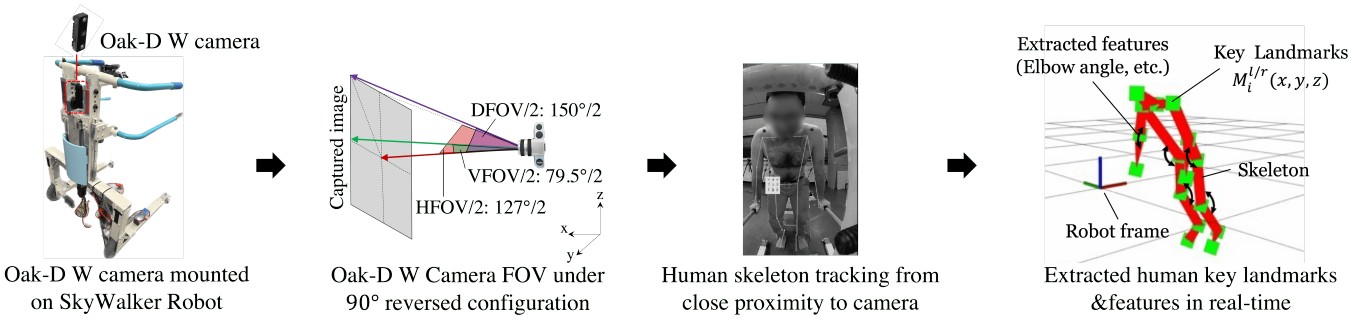
Firstly, at close distances, camera lenses often suffer from distortion, which can result in inaccurate skeleton tracking and pose estimation. Parts of the body might appear disproportionately larger or smaller, complicating the interpretation of the user's posture and movements. Secondly, accurate depth perception is harder to achieve at close distances, leading to potential errors in determining the spatial positioning of the user's body parts. Lastly, real-time processing of visual data at close range with high accuracy requires significant computational resources, which can be a challenge given the need for the system to be used in real-time.

To address these challenges, we have employed a combination of advanced software and strategic hardware placement. Specifically, we utilize Google's Mediapipe framework [14] and strategically mount the camera on the SkyWalker robot to optimize performance in close proximity. The Mediapipe BlazePose GHUM Lite model is particularly suitable for this application scenario due to its lightweight architecture [24], allowing for efficient processing with limited computational resources available on our robotic platform. With the current hardware implementation of the robot, we determined that a minimum frame rate of 15 Hz is sufficient for real-time classification. This is because, when monitoring STS motion in older adults, the robot's actions are not too fast, allowing for accurate detection of user movement at this frame rate. The camera, a Luxonis Oak-D W depth AI camera with a wide field of view (FOV) (150° Depth-FOV with 128° Horizontal-FOV) lens, is mounted on the robot with 90° rotated to adapt its HFOV for the vertical direction. This strategic placement minimizes occlusions and distortions, ensuring a tracking of user movements during the STS phases. The system's architecture is shown in Fig 2.

B. Key Landmark Acquisition Pipeline

Fig.4 illustrates the kinematic features data acquisition process, which is integral to our feature extraction pipeline. The process begins with the Oak-D W camera mounted on the SkyWalker robot, collecting images to be processed to identify key landmarks on the human body. The camera's FOV is configured in a 90° reversed setup, ensuring the camera can fully track the human skeleton, especially when the user is less than one meter away. These landmarks include critical points such as the pelvis, shoulders, elbows, etc. The extracted human key landmarks and features are processed in real time, as shown in the final part of Figure 5. This includes calculating important kinematic features such as the elbow angles and the

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.



Oak-D W camera mounted on SkyWalker Robot | Oak-D W Camera FOV under 90° reversed configuration | Human skeleton tracking from close proximity to camera | Extracted human key landmarks &features in real-time

Fig. 4: Kinematic features data acquisition process

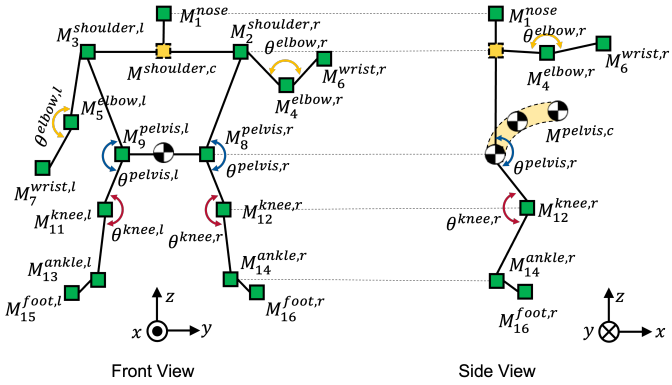


Fig. 5: Simplified mediapipe human skeleton model and landmarks. Pelvis center and each green marker's location can be obtained from mediapipe, while others (yellow point and joint angles) can be obtained through calculation

positions of various joints. The data from these key landmarks is used to create a simplified skeleton model (Fig.5) that can be analyzed to understand the user's movement patterns.

The key landmarks' position $M_i^{l/r}(x,y,z)$, as shown in Fig.5, extracted from the skeleton, are used to derive the necessary kinematic features. These features can then be fed into the phase classification system to determine the user's current phase throughout the STS process.

C. Kinematic Feature Calculation

We utilized the obtained key landmarks, as shown in Fig.5, to calculate the kinematic features representing an STS motion. Initially, all the obtained landmarks are under the camera coordinate, while the camera has a forward movement during STS support. To eliminate the influence of camera movement, we fixed a marker on the top of the camera and utilized a motion capture Vicon to monitor its movement. Considering a coordinate transform from the pelvis frame to the camera frame and finally to the world frame, Eq. 1 can be modified as follows:

$$M_{i,world}^{l/r} = (M^{pelvis,c} + \overline{M_i^{l/r}}) + M^{cam} + \begin{bmatrix}0 & 0 & \Delta l\end{bmatrix} \quad (1)$$

Here the $M^{pelvis,c}$ stands for the pelvis frame while $\overline{M_i^{l/r}}$ are landmarks (other than pelvis center) position corresponding to the pelvis frame. The camera's movement under world coordinate M^{cam} is introduced to transform the point location from the camera to the world coordinate. A shifting matrix $(0,0,\Delta l)$ was utilized to make the camera's marker position correspond to the bottom stereo camera. Δl was set as $-0.097m$ according to the OAK-D W camera datasheet.

The landmark positions derived from Eq.(1) allow us to calculate the user's kinematics properties during STS. Previous studies have demonstrated that several determinants, such as trunk, arm, and knee positioning and movements, are considered significant kinematic features that affect STS performance [21][25]. Based on these biomechanical considerations, we selected 14 unique kinematic features to represent the STS movement while using the Skywalker.

- Pelvis x and z position: $M^{pelvis,c}(x,z)$
- Shoulder x and z position: $M^{shoulder,c}(x,z)$
- Elbows x and z position: $M_5^{elbow,l}(x,z)$, $M_4^{elbow,r}(x,z)$
- Hip sagittal angle: $\theta^{pelvis,l}$, $\theta^{pelvis,r}$
- Knee sagittal angle: $\theta^{knee,l}$, $\theta^{knee,r}$
- Elbow sagittal angle: θ^{elbow}, $\theta^{elbow,r}$

Here we need to note that instead of measuring the coordinates of each key landmark, the Mediapipe package only provides the 3D coordinates of a reference depth point ($M^{pelvis,c}[x,y,z]$). In contrast, other body landmarks are calculated by projecting them onto a 2D image surface, and their coordinates are presented in relation to the center pelvis coordinates. Given this consideration, the actual perception of each landmark is calculated using the following equation:

$$M_i^{l/r} = M^{pelvis,c} + \overline{M_i^{l/r}} \quad (2)$$

where the $M^{pelvis,c}$ is the center pelvis location under camera coordination and is directly acquired from the Mediapipe. While $\overline{M_i^{l/r}}$ is the landmark location with respect to pelvis location. Using Eq. 2, we can transport all landmarks' locations from the pelvis coordinate to the camera coordinate.

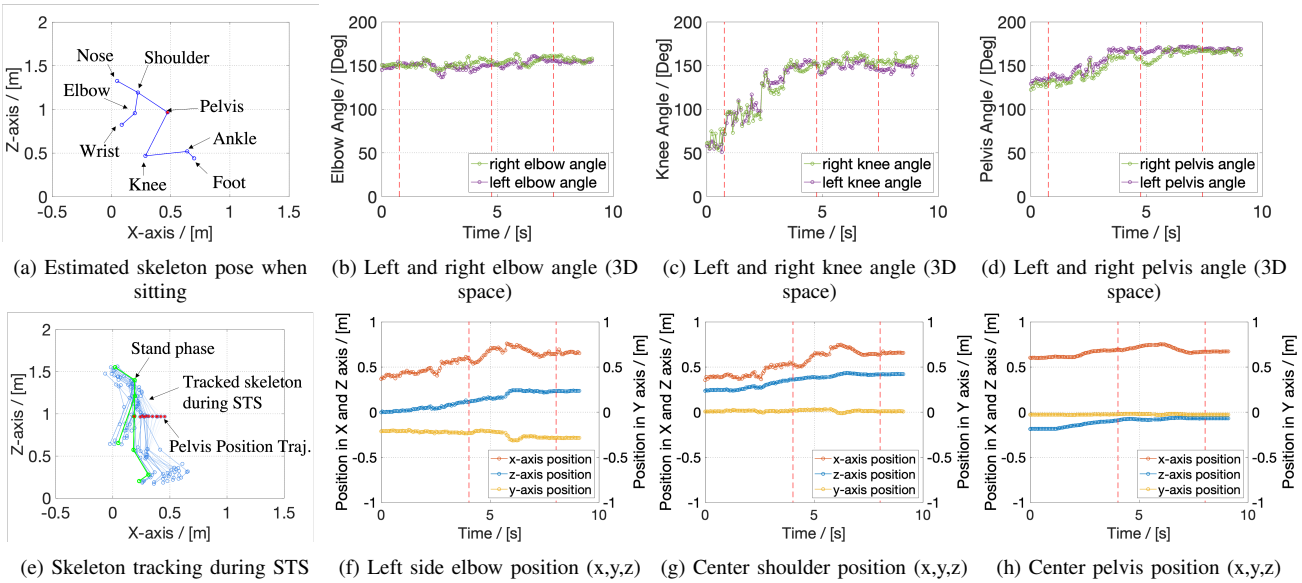
We used the following equations to calculate the positions and derive the angles of the user's center shoulder and pelvis under camera coordination

$$M^{shoulder,c} = \frac{M_9^l + M_8^r}{2} \quad (3)$$

where M_9^l and M_8^r are the measured 3D location under camera coordination, and $M^{shoulder,c}$ is the calculated shoulder's center location. Finally, we can calculate the joint angle position, for example, the left pelvis angle, based on the following equation:

$$\theta^{pelvis,l} = \arccos \frac{\overrightarrow{M_9^l M_3^l} \cdot \overrightarrow{M_9^l M_{11}^l}}{|\overrightarrow{M_9^l M_3^l}||\overrightarrow{M_9^l M_{11}^l}|} \quad (4)$$

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.



(a) Estimated skeleton pose when sitting

(b) Left and right elbow angle (3D space)

(c) Left and right knee angle (3D space)

(d) Left and right pelvis angle (3D space)

(e) Skeleton tracking during STS

(f) Left side elbow position (x,y,z)

(g) Center shoulder position (x,y,z)

(h) Center pelvis position (x,y,z)

Fig. 6: Calculated input kinematic features from Mediapipe Lite model

Furthermore, We derive the elbow and knee angle using similar equations as shown in Eq. 4. The obtained configuration of a sitting pose and time-lapse image of STS skeleton tracking can be found in Fig.6(a)(e). Here we demonstrate part of the calculated kinematics features in Fig.6. The red dashed lines in Fig.6 divide the raw input into four sections, representing four phases defined in the previous section.

D. Kinematics Features Comparison

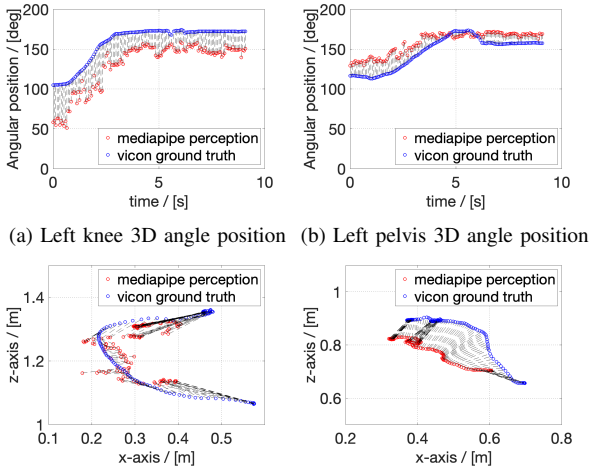
To validate the accuracy of the Mediapipe-based skeleton tracking, we compare its performance against ground truth Vicon. This comparison allowed us to benchmark Mediapipe's performance, identify discrepancies between the two systems within our robotic setup, and evaluate the effectiveness of the 14 selected features for STS motion classification. The Vicon system utilizes multiple cameras and reflective markers placed on the subject's body to capture detailed motion data at frequencies of 100 Hz. Additionally, we transform the kinematic data from the camera coordinate system to a world coordinate system to maintain consistency and accuracy in our measurements, especially given the dynamic movements of the SkyWalker robot during STS assistance. For data synchronization, A Vicon-ROS bridge is used to synchronize data collection between Vicon and Mediapipe, ensuring that both systems record data simultaneously and under the same conditions.

Figure 7(a) shows the left knee 3D angle position over time. While MediaPipe perception generally follows the Vicon ground truth, noticeable deviations occur, particularly in the initial 2-4 seconds, suggesting challenges with rapid initial movements. Similarly, Figure 7(b) presents the left pelvis 3D angle position, where slight deviations are particularly noticeable during dynamic posture changes. Fig.7(c)-(d) plot the center shoulder and pelvis positions in the x-z plane, where x represents depth and z represents vertical position. MediaPipe data points show significant scatter compared to the Vicon ground truth, especially along the depth (x-axis) for both the shoulder and pelvis. This indicates challenges in accurately tracking depth movements, although the vertical position (z-axis) shows better accuracy, albeit with some deviations, particularly for the pelvis. These observations highlight areas where MediaPipe's tracking accuracy could be improved.

Figure 8(a) illustrates the average angular error for the pelvis, shoulder, and elbow angles. Errors are higher for the shoulder and elbow compared to the pelvis, with the right body side consistently showing higher errors. This indicates variability in MediaPipe's performance across different joints. Figure 8(b) presents the accuracy of positional features, showing the average positional errors for the pelvis, shoulder, and elbow in depth (x), horizontal (y), and vertical (z) directions. Errors are generally higher in the x direction for all joints, reflecting difficulties in accurately perceiving x positions. The lowest error is observed in the x direction for the pelvis, indicating better depth perception accuracy. However, the depth accuracy decreases as the features move away from the pelvis, with the elbow showing the highest error. This is because the landmarks in MediaPipe are driven based on the center of the pelvis.

Figure 8(c) assesses the compatibility of angular features between MediaPipe perception and Vicon ground truth through mean correlation coefficients for left and right angular features. Both sides show coefficients above the threshold, indicating reliable tracking of angular features, suggesting MediaPipe's capability in following joint angles critical STS motion. Figure 8(d) shows the compatibility of positional features, with mean correlation coefficients for positional features. The depth and vertical directions exhibit coefficients above the threshold, indicating reasonable compatibility of following positions of the features. However, the horizontal direction shows coefficients below the threshold, reflecting challenges in horizontal position tracking. These analyses highlight that while MediaPipe performs well in depth and vertical tracking, it struggles significantly with horizontal position tracking, particularly for the shoulder positions. In conclusion, these findings demonstrate that MediaPipe perception can effectively

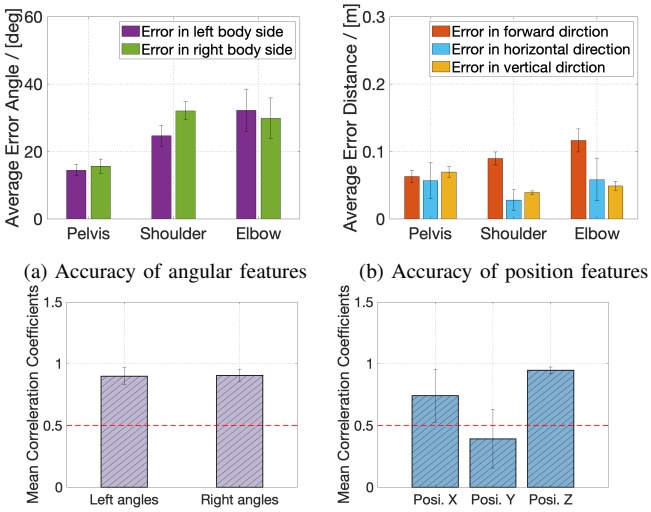
IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.



(a) Left knee 3D angle position (b) Left pelvis 3D angle position

(c) Center shoulder position (x,z) (d) Center pelvis position (x,z)

Fig. 7: Mediapipe perception V.S. Ground truth Vicon



(a) Accuracy of angular features (b) Accuracy of position features

(c) Compatibility of angular features (d) Compatibility of position features

Fig. 8: Skeleton tracking performance comparison (3D space)

track joint angles and positions during STS motion using the robot. MediaPipe provides 3D coordinates of joint centers based on visual input from the camera, relying on a machine learning model to infer positions. This innovative approach can be influenced by occlusions and varying image qualities. In contrast, the Vicon system uses physical markers placed on the skin to capture precise movement, offering high accuracy in positional data, though it does not directly measure joint centers. Given these fundamental differences, our focus was on relative movements and trends rather than direct positional accuracy, ensuring the robustness of the 14 features that will be used for the phase classifier despite inevitable discrepancies between the two systems. Comparisons were primarily conducted in the sagittal plane, which helps mitigate systematic errors and enhances the reliability of our kinematic feature extraction process. This comparison validates the relevance of the 14 selected features for effective STS motion classification extracted from MediaPipe. Validation against ground truth confirms MediaPipe's potential as a valuable tool for real-time STS phase classification in assistive robotic applications from close proximity.

IV. Experiment & Classification Result

A. Experiment protocol

To demonstrate the usability of the proposed phase classification method described in the previous sections, a set of experiments was conducted. 10 able-bodied subjects (weight: 65.75 ± 13.02 kg, height: 167 ± 4.08 cm) performed the STS motion 13 repetitions with the SkyWalker robot while the camera simultaneously recorded their motion. Each participant replicated the STS motion as illustrated in Fig. 3.

We selected able-bodied subjects to prepare our dataset because they can perform STS motions with minimal assistance from the robot and provide consistent and reliable data for training and testing our model.

In the beginning, the participant sat on the chair with both upper limbs vertical to the lower handle. When the user was ready, we sent a start signal to the robot and mentioned it to the participant simultaneously. The robot then followed a predefined trajectory in both forward (0.2m, 0.3m, or 0.4m) and upward (0.3m) directions for 4 and 6 seconds, respectively.

We chose different forward distances to test how the user's kinematics and phases change with varying levels of horizontal support from the robot. The upward distance was fixed to ensure that the user could comfortably and safely reach the upper handles.

Each participant was required to run the experiment for 13 trials, allowing us to collect the raw output data from 3D keypoint landmarks at approximately 15Hz.

We utilized the collected experimental data for both training our classifier and validating the classification accuracy.

Since the SkyWalker moves simultaneously with the user's STS motion, we adopted eq.(4) to eliminate the influence of robot movement and obtained pure kinematics information from Fig. 4 to calculate the selected key features as the input features. To eliminate the feature variability caused by the subject's physical condition, we normalized all the features extracted by dividing each data point by the corresponding maximum value into the range [-1, 1].

We used a depth camera with the Mediapipe model to record the kinematics of the subjects performing STS motions with the SkyWalker robot. As previously mentioned, we selected 14 kinematic features relevant for STS phase classification, such as joint angles and key point positions. We also manually labeled the data according to the four phases defined in Section III.

We chose to use only a subset of the Mediapipe landmarks, rather than the whole skeleton model, for two reasons. First, some landmarks were not visible or reliable when the user was very close to the camera (around 0.5 meters away), such as the nose and foot points. Second, some landmarks were not informative or discriminative for the STS phases, such as the wrist and ankle points. Therefore, we focused on the landmarks that were most related to the STS biomechanics, such as the pelvis, shoulder, elbow, hip, and knee points.

Once we extracted 14 features from each dataset, we prepared our data into training (6053 frames), testing (1513 frames), and validation (1344 frames) sets. For SVM, we utilized the LIBSVM software for model training. The open-source Python library "scikit-learn" was used to prepare and

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

TABLE I: PHASE TRANSITION TIME ERROR IN SECONDS

Validation Set	Phase 1-2					Phase 2-3					Phase 3-4				
	SVM	XGboost	RF	DT	LSTM	SVM	XGboost	RF	DT	LSTM	SVM	XGboost	RF	DT	LSTM
1	-0.67	-0.53	-0.70	-0.35	-0.29	-0.99	-1.01	-1.10	-0.30	-1.07	+1.13	-	+0.74	+1.54	+1.03
2	-0.44	-0.86	-0.00	+2.31	+2.91	-0.24	1.25	-1.25	+0.98	-	+0.29	+0.90	-	-0.12	+1.93
3	-2.34	+0.03	+0.37	+0.37	-1.14	-1.50	+0.49	+0.42	+0.76	+0.42	+0.53	-	-	-0.29	+2.53
4	-0.60	-0.53	-0.77	-1.34	-0.28	-0.89	-0.31	-0.47	+0.12	-1.01	+0.56	-	+0.59	+0.55	+0.56
5	-0.48	-2.00	-0.92	-1.92	-2.00	-1.35	-0.67	-0.75	+1.24	-0.75	-	-	-	+1.90	+0.51
6	-0.66	-1.49	-1.48	-0.23	-1.55	-0.27	-0.27	-0.15	-0.00	-0.40	+1.26	+1.26	+1.00	+1.07	+0.94
7	+0.10	+1.79	-0.48	+1.91	-0.42	+0.16	+0.10	+0.16	+0.16	+0.10	+1.16	+0.97	+0.66	+1.16	+1.22
Average	0.76	1.03	0.67	1.20	1.23	0.77	0.59	0.61	0.51	0.63	0.82	1.04	0.75	0.95	1.25

execute other classifier options (DT, RT, and XGBoost). Finally, we recorded the video from the camera and determined each participant's ground truth phase visually, then compared the prepared label with the predicted phases.

B. Classification result

In this study, we focus on classifying the different phases of the Sit-to-Stand (STS) motion using a mobile assistive robot, SkyWalker. Accurate phase classification is crucial for the robot to provide timely and appropriate assistance, ensuring user safety and optimizing support during the STS movement. To achieve this, we employed various machine learning classifiers, including Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), XGBoost, and Long Short-Term Memory (LSTM), to determine the user's current phase in real-time based on 14 selected kinematic features extracted from Mediapipe's 3D visual skeleton tracking.

The classification results are evaluated by comparing the predicted phase transitions to the ground truth transitions, with the transition time errors (in seconds) recorded for each classifier across three critical phase transitions: Phase 1-2, Phase 2-3, and Phase 3-4. These transitions represent the user's movement from the initial sitting position to the intermediate phases and finally to the standing position.

Table 1 presents the phase transition time errors for different classifiers (SVM, XGBoost, RF, DT, and LSTM) across three phase transitions. For Phase 1-2, SVM has an average error of 0.76 seconds, with errors ranging from -2.34 to +0.10 seconds, indicating it generally performs well but struggles with rapid movements. XGBoost shows significant variation with an average error of 1.03 seconds, especially in validation sets 5 and 7. RF demonstrates consistent performance with an average error of 0.67 seconds, while DT exhibits considerable variation with an average error of 1.20 seconds. LSTM shows the highest variation, with an average error of 1.23 seconds, indicating difficulties with initial transitions.

For Phase 2-3, SVM has an average error of 0.77 seconds, showing moderate performance. XGBoost performs well with an average error of 0.59 seconds, and RF is consistent with an average error of 0.61 seconds. DT indicates good performance with fewer large errors, averaging 0.51 seconds. LSTM, with an average error of 0.63 seconds, shows variability but generally moderate performance.

Phase 3-4 shows higher errors. SVM has an average error of 0.82 seconds, XGBoost shows significant variation with an average error of 1.04 seconds, RF demonstrates consistent performance with an average error of 0.75 seconds, DT shows considerable variation with an average error of 0.95 seconds, and LSTM has the highest variability with an average error of 1.25 seconds, indicating challenges in final phase transitions.

Overall, the RF and DT classifiers generally exhibit the most consistent performance across all phase transitions, with smaller average errors and narrower error ranges. The SVM and XGBoost classifiers also perform well but show slightly higher variation. The LSTM classifier demonstrates the highest variability, suggesting it might be less reliable for real-time applications where consistency is crucial. The errors in the Phase 1-2 transition are generally higher across all classifiers, indicating the challenge of accurately identifying the initial phase transition. Errors decrease in the Phase 2-3 transition, suggesting improved classifier performance once the initial transition is accurately detected. However, the Phase 3-4 transition exhibits higher errors, indicating challenges in accurately identifying the final phase transition, likely due to not having enough features for the classifier to observe since only the elbow kinematics are changing.

To enhance classifier performance, fine-tuning parameters, incorporating more diverse training data, exploring advanced models or hybrid approaches. These improvements aim to achieve higher accuracy and reliability, particularly for the initial and final phase transitions, thereby enhancing the system's overall robustness and responsiveness.

Furthermore, it is important to note that we trained the classifier using the data collected from healthy subjects, which may limit the generalizability of our approach to other populations with different mobility impairments or STS patterns. Therefore, one of our future work is to collect data from individuals with mobility impairments and evaluate how our model performs on them.

C. Discussion

The estimated phase from the classifiers provides a clear understanding of the user's current STS (Sit-to-Stand) phase among the four defined phases, with a relatively small delay of 0.88s and a high accuracy of 89%. This understanding enables the robot to adjust its actions accordingly. For instance, the robot will begin moving only after the user's phase has transitioned from phase 1 to 2. This classification result also helps us identify whether the user has fully stood up by monitoring the transitions from phases 2 to 4. If the user's phase does not change as expected, the robot can stop its motion, thus preventing potential falls. Finally, the robot can start moving only after detecting that the user has fully stood up, which corresponds to phase 4.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

D. Real-time applications

Key real-time applications include adaptive STS assistance, where the SkyWalker continuously monitors both able and disabled subjects' posture and phase to provide optimal support during the movement, thus reducing the risk of falls and improving stability. The system can also detect deviations from the expected STS movement pattern with a rapid response time of less than one second, enabling the robot to initiate emergency stops or other safety measures, which is crucial for preventing accidents. Additionally, the phase classification system enhances user interaction by allowing the robot to respond intuitively to the user's real-time movements, creating a seamless and supportive experience.

V. Conclusion and Future Work

This paper presents a vision-based algorithm for estimating human phase from close proximity for use with a robotic rollator providing STS assistance. Our study offers a two-fold contribution: first, by comparing the pose identification accuracy of the Lite version of Mediapipe with Vicon, and second, by developing a robust phase classification framework for STS movements. The comparison showed that certain joint angles and key points from Mediapipe can reliably estimate the human state.

In our phase estimation experiment, predefined trajectories were executed while recording the subjects' kinematics using a depth camera with the Mediapipe model. The kinematic data from the 14 selected features were used to train human phase classifiers, with Random Forest (RF) and Support Vector Machine (SVM) achieving the best performance, reaching an accuracy of around 89 percent. The small average time error when switching between phases indicates that the model is suitable for phase classification.

Several limitations were encountered during the study. The Lite version of Mediapipe was used instead of the Heavy or Full model, potentially contributing to the errors observed. Future work will involve using a more powerful processing unit to evaluate the performance of the Heavy model, which is documented to have higher accuracy. The camera's position also impacted the results, as being too close to the subjects affected the skeleton tracking performance. This can be addressed by training users before they can use the walker properly. Future studies will explore varying camera positions to optimize tracking performance. We also plan to train a customized Mediapipe model, in order to improve the performance of pose tracking accuracy furthermore. Lastly, the human state estimation and phase classification for both able and disabled subjects will be integrated into SkyWalker to adjust its STS trajectory in real time, providing optimal support for safe movement and fall prevention.

References

[1] G. Baer and A. Ashburn, "Trunk movements in older subjects during sit-to-stand," *Arch Phys Med Rehab*, vol. 76, pp. 844–849, 1995.

[2] E. Phillips, J. Schneider, and G. Mercer, "Motivating elders to initiate and maintain exercise," *Arch Phys Med Rehabil*, vol. 85, pp. 52–57, 2004.

[3] Z. Matjačić, M. Zadravec, and J. Oblak, "Sit-to-stand trainer: an apparatus for training "normal-like" sit to stand movement," *IEEE Trans Neural Syst Rehab Eng*, vol. 24, pp. 639–649, 2015.

[4] P.-T. Cheng, C.-L. Chen, C.-M. Wang, and W.-H. Hong, "Leg muscle activation patterns of sit-to-stand movement in stroke patients," *Am J Phys Med Rehab*, vol. 83, pp. 10–16, 2004.

[5] Z. Liao, J. V. S. Luces, and Y. Hirata, "Human navigation using phantom tactile sensation based vibrotactile feedback," *IEEE Robotics and Automation Letters*, vol. 5, no. 4, pp. 5732–5739, 2020.

[6] Z. Liao, J. Salazar, and Y. Hirata, "Robotic guidance system for visually impaired users running outdoors using haptic feedback," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021, pp. 8325–8331.

[7] N. Reider and C. Gaul, "Fall risk screening in the elderly: a comparison of the minimal chair height standing ability test and 5-repetition sit-to-stand test," *Arch Gerontol Geriatr*, vol. 65, pp. 133–139, 2016.

[8] H. Mollenkopf and J. L. Fozard, "Technology and the good life: Challenges for current and future generations of aging people," *nnu Rev Gerontol Geriatr*, vol. 23, no. 1, pp. 250–279, 2003.

[9] C. Werner, M. Geravand, P. Z. Korondi, A. Peer, J. M. Bauer, and K. Hauer, "Evaluating the sit-to-stand transfer assistance from a smart walker in older adults with motor impairments," *Geriatr Gerontol Int*, vol. 20, pp. 312–316, 2020.

[10] Z. Dong, J. V. S. Luces, A. A. Ravankar, S. A. Tafrishi, and Y. Hirata, "A performance evaluation of overground gait training with a mobile body weight support system using wearable sensors," *IEEE Sensors Journal*, vol. 23, no. 11, pp. 12 209–12 223, 2023.

[11] R. Clark, B. Mentiplay, E. Hough, and Y. H. Pua, "Three-dimensional cameras and skeleton pose tracking for physical function assessment: A review of uses, validity, current developments and kinect alternatives," *Gait Posture*, vol. 68, pp. 193–200, 2019.

[12] S. Colyer, M. Evans, D. Cosker, and A. Salo, "A review of the evolution of vision-based motion analysis and the integration of advanced computer vision methods towards developing a markerless system," *Sports Med Open*, vol. 4, no. 1, pp. 1–15, 2018.

[13] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *IEEE Comput Soc Conf Comput Vis Pattern Recognit*, 2017, pp. 7291–7299.

[14] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, and J. Lee, "Mediapipe: A framework for building perception pipelines," *arXiv preprint arXiv:1906.08172*, 2019.

[15] M.-Y. Hsiao, C.-M. Li, I.-S. Lu, Y.-H. Lin, T.-G. Wang, and D.-S. Han, "An investigation of the use of the kinect system as a measure of dynamic balance and forward reach in the elderly," *Clin Rehabil*, vol. 32, no. 4, pp. 473–482, 2018.

[16] A. Aguirre, J. Casas, N. Céspedes, M. Múnera, M. Rincon-Roncancio, A. Cuesta-Vargas, and C. A. Cifuentes, "Feasibility study: Towards estimation of fatigue level in robot-assisted exercise for cardiac rehabilitation," in *IEEE Int Conf Rehabil Robot*, 2019, pp. 911–916.

[17] A. Aguirre, M. J. Pinto, C. A. Cifuentes, O. Perdomo, C. A. Díaz, and M. Múnera, "Machine learning approach for fatigue estimation in sit-to-stand exercise," *Sensors*, vol. 21, no. 15, p. 5006, 2021.

[18] T.-H. Ou-Yang, M.-L. Tsai, C.-T. Yen, and T.-T. Lin, "An infrared range camera-based approach for three-dimensional locomotion tracking and pose reconstruction in a rodent," *J Neurosci*, vol. 201, no. 1, pp. 116–123, 2011.

[19] S. Taghvaei, Y. Hirata, and K. Kosuge, "Vision-based human state estimation to control an intelligent passive walker," in *IEEE/SICE Int Symp Syst Int*. IEEE, 2010, pp. 146–151.

[20] C. Chen, P. Zhang, H. Zhang, J. Dai, Y. Yi, H. Zhang, and Y. Zhang, "Deep learning on computational-resource-limited platforms: a survey," *Mob Info Syst*, vol. 2020, pp. 1–19, 2020.

[21] A. Mahdi, J. F.-S. Lin, and K. Mombaur, "Maintaining mobility in older age - design and initial evaluation of the robot skywalker for walking and sit-to-stand assistance," in *IEEE RAS/EMBS Int Conf Biomed Robot Biomech*, 2022, pp. 01–08.

[22] "Oak-d w." [Online]. Available: https://shop.luxonis.com/products/oak-d-w

[23] M. Zago, M. Luzzago, T. Marangoni, M. De Cecco, M. Tarabini, and M. Galli, "3d tracking of human motion using visual skeletonization and stereoscopic vision," *Frontiers in bioengineering and biotechnology*, vol. 8, p. 181, 2020.

[24] "Pose | mediapipe." [Online]. Available: google.github.io/mediapipe/solutions/pose

[25] W. G. Janssen, H. B. Bussmann, and H. J. Stam, "Determinants of the sit-to-stand movement: a review," *Physical therapy*, vol. 82, no. 9, pp. 866–879, 2002.