Video-Based Human-Object Interaction Analysis for Patient Behavioral Monitoring

Mostafa Habibi, Zeinab Delaram, and Mehrdad Nourani
Department of Electrical & Computer Engineering
The University of Texas at Dallas
Richardson, Texas 75080, USA

Dennis H. Sullivan
Geriatric Research, Education and Clinical Center
Central Arkansas Veterans Healthcare System
North Little Rock, AR 72114, USA

Outline

- Motivation
  - AI-Powered Solution
- Prior Works
- Proposed Methodology
  - Person & Object Detection and Tracking
  - Posture Detection and Classification
  - Objects Filtering
  - Human-Object Interaction Modeling
- Experimental Results
- Conclusion and Future Work

Motivation

- Hospitalized patients spend significant time in bed, causing
  - Rapid physical deconditioning
  - Higher risk of complications (DVT, UTI, Pneumonia)
  - Increased readmission & nursing home placement
- Early and progressive ambulation (EPA) can shorten the recovery time and reduce the healthcare cost
- Current ambulation monitoring relies on nurse observation; thus subjective, infrequent and inaccurate

- **Reliable, scalable, and non-intrusive monitoring tools are highly in demand.**

AI-Powered Camera-Based HOI Monitoring

- **AI** offers potential to develop accurate, scalable and (if needed) real-time monitoring systems.

- **Computer Vision** is a field of Artificial Intelligence (AI) that enables machines to understand, interpret, and extract meaningful information from visual inputs such as images and videos.

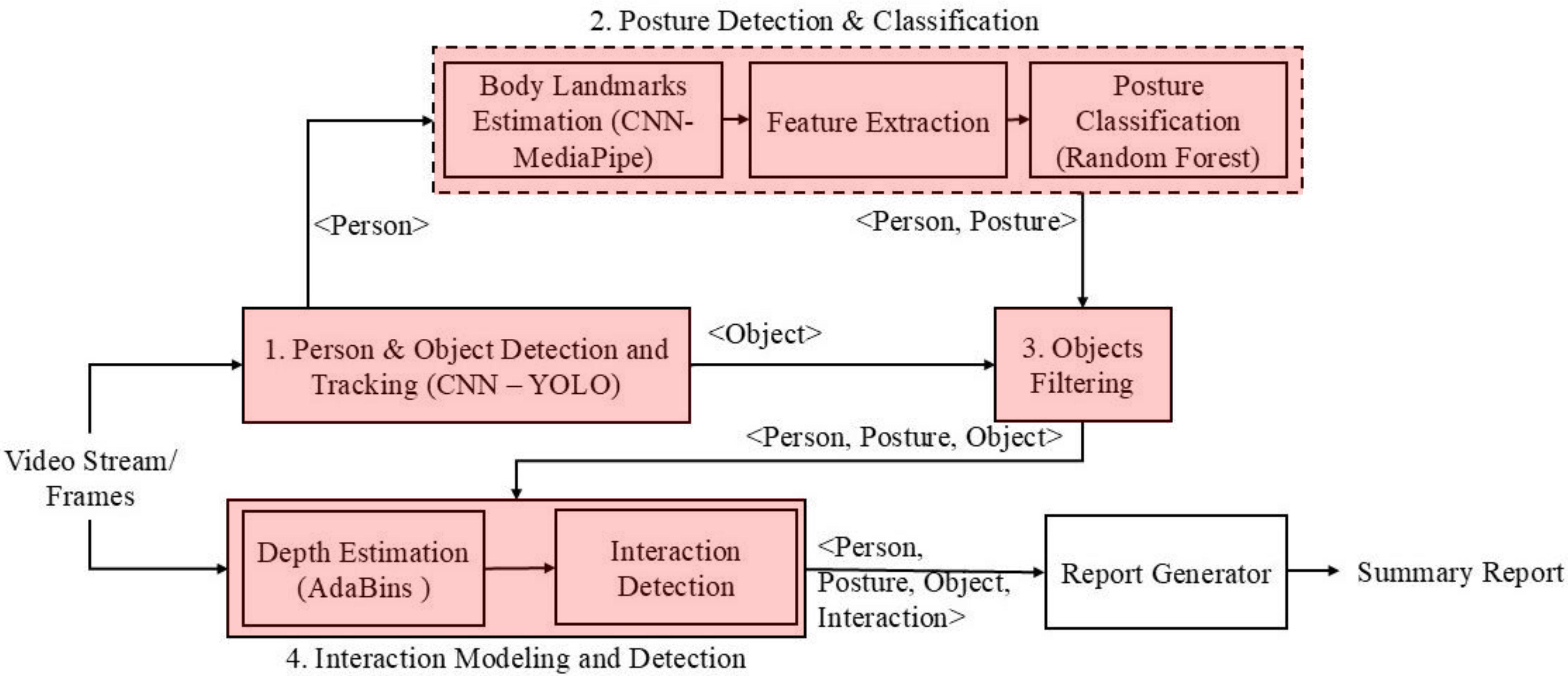
- Key Features
  - Tracks movements, postures, and interactions in real time and continuously
  - Detects sitting, standing, and lying with durations
  - No need for wearables or patient compliance

Prior Works

- Current studies primarily focus on patient ambulation & physical activity analysis.
  - IoT-based Systems (Garcia et al.) – for monitoring indoor and outdoor activities.
  - Wearable Technologies & Smartphone-based Solutions (Dhakal et al.) – specific area like dementia.
  - Multi-modal Patient Activity Monitoring System (PAMS) (Steele et al.) – integrating wrist-worn devices and RTLS to classify ambulation and postures.

- Current systems do not explore patient-object interaction, contextual behavior and comprehensive reporting for clinical interpretation.

Proposed Methodology



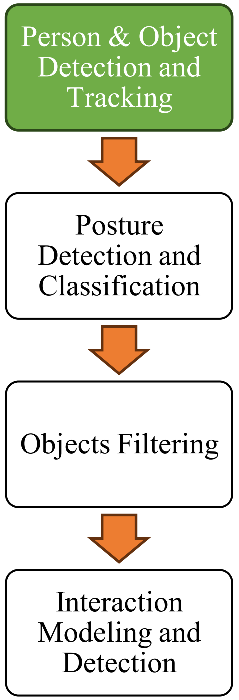
Step 1: Person & Object Detection and Tracking

- Created and annotated a custom dataset to retrain YOLO for real-world hospital objects and environment.

- Dataset of 23,704 annotated images collected from the internet, focused on typical patient-object interactions in hospital environments.

- A set of clinically-relevant objects (e.g. 14 objects of interest) can be customized for scalability and accuracy.

Objects	# of Annotations	Objects	# of Annotations
TV	2393	IV_Pole	1165
Bed	1780	Person	1301
Chair	1812	Sofa	2662
Table	456	Door	801
Cane	2309	Wheelchair	3534
Walker	1298	Crutch	1491
Food Tray	1542	Toilet sign	1050



Step 1: Person & Object Detection and Tracking – Cont.

- After object detection, we perform multi-object tracking to
  - assign a unique ID to each object
  - maintain temporal consistency across video frames

- BoT-SORT is used for tracking
  - Builds on detection results to estimate object state over time
  - Enables real-time tracking of patients and objects in dynamic hospital environments

Person & Object Detection and Tracking

↓

Posture Detection and Classification

↓

Objects Filtering

↓

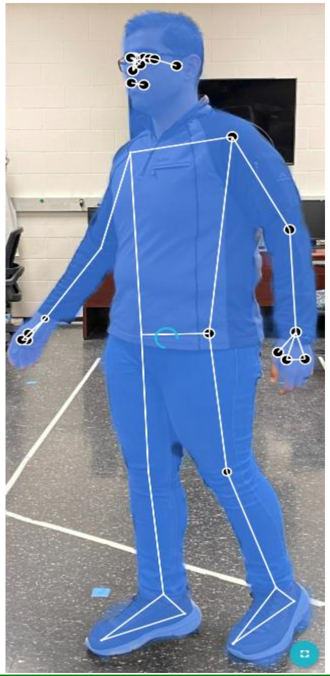
Interaction Modeling and Detection

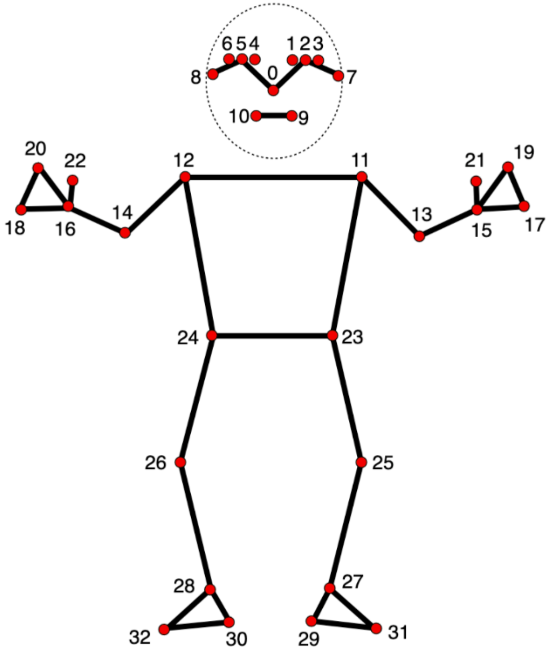
Step 2: Posture Detection and Classification

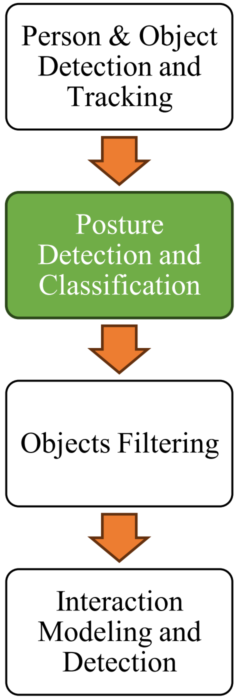
- Created training data for 3 postures: sitting, standing and lying
  - 14 videos (sitting), 13 (standing), 13 (lying)
  - Recorded from 8 volunteers
  - Generated labeled frames for model training & testing

Postures	# of videos	# of Labeled Frames
Sit	14	12657
Stand	13	10597
Lie	13	17235

- Extracted body’s key points using
  - Mediapipe landmark estimation
  - Detection of 33 human body landmarks





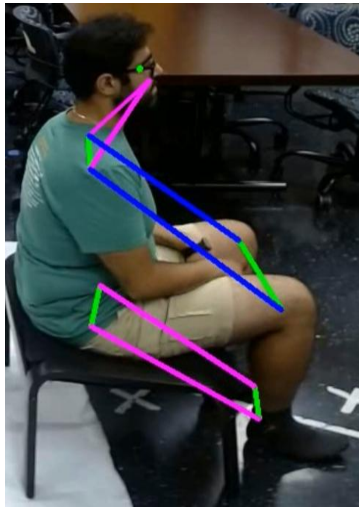


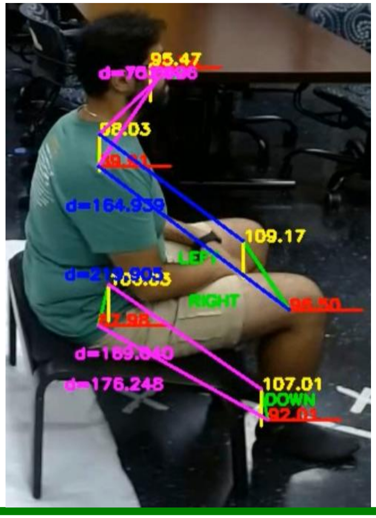
Step 2: Posture Detection and Classification – Cont.

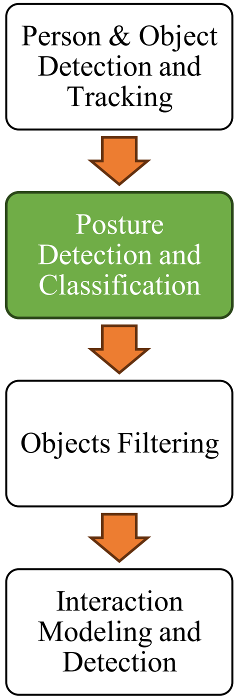
- Utilized 11 landmarks (eyes, shoulders, hips, knees, ankles, nose)
  - 5 main-body vectors, 6 custom vectors
  - 33 features (6 lengths, 27 angles)

- Employed a Random Forest classifier for posture recognition
  - Three Classes: Sitting, Standing, Lying
  - Environment-independent (e.g., sitting on chair or bed is same class)
  - Achieved high accuracy

Main Vector	Custom Vector
Eye Line	Nose → L_Shoulder
Shoulder Line	Nose → R_Shoulder
Hip Line	L_Shoulder → L_Knee
Knee Line	R_Shoulder → R_Knee
Ankle Line	L_Hip → L_Ankle
-	R_Hip → R_Ankle







Step 3: Objects Filtering

- Define meaningful interactions using a relational table

- Relational table links patient postures to meaningful object interactions, using logical real-world constraints
  - Example: A patient can lie in bed, but not in chair

Posture	Objects	Posture	Objects	Posture	Objects
Sit	Bed	Stand	Cane	Lie	Bed
	Chair		Walker		Sofa
	Sofa		Crutch		Table
	Table		IV Pole		
	Wheelchair				

Person & Object Detection and Tracking

Posture Detection and Classification

Objects Filtering

Interaction Modeling and Detection

Step 4: Interaction Modeling and Detection

- Metrics used for interaction modeling

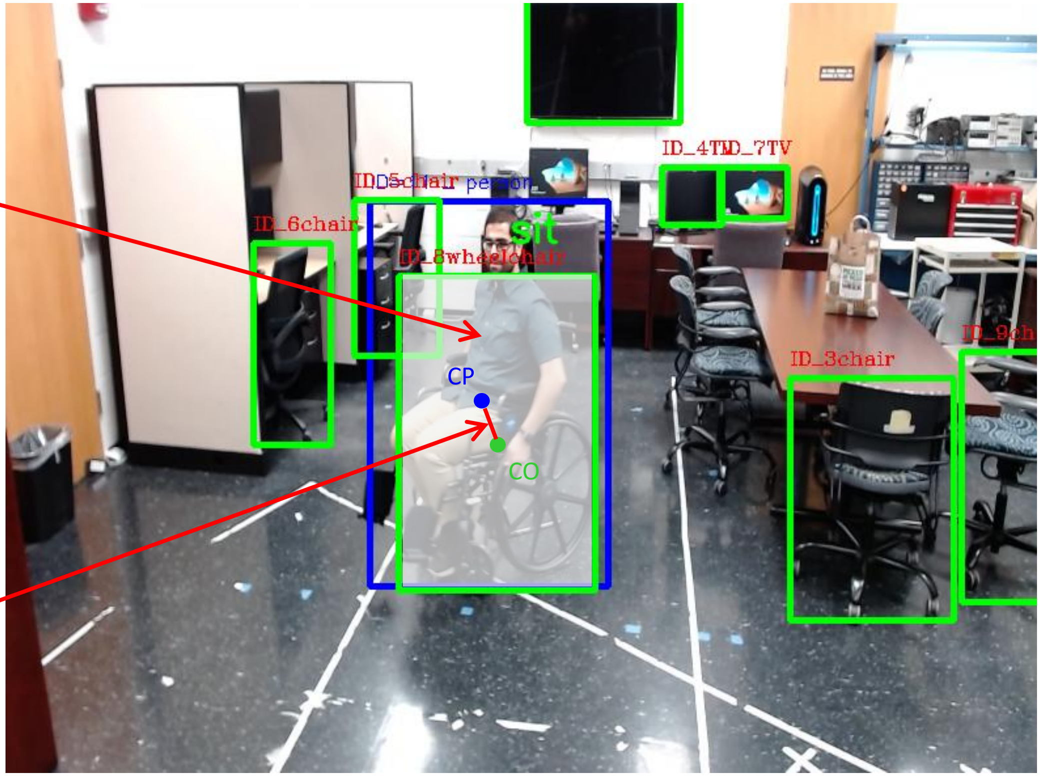
1. **Intersection over Union (IoU):** Measures overlap (of bounding boxes) between objects

$$IoU = \frac{Area_{Intersection}}{Area_{Object} + Area_{Patient} - Area_{Intersection}}$$

- IoU remains unaffected by object size, providing fair comparison

2. **Central Distance**: Measures spatial adjacency between bounding boxes of patient and object
   - Identifies which object is **closest** to the patient

$$Distance = \sqrt{(X_{PC} - X_{OC})^2 + (Y_{PC} - Y_{OC})^2}$$



Person & Object Detection and Tracking

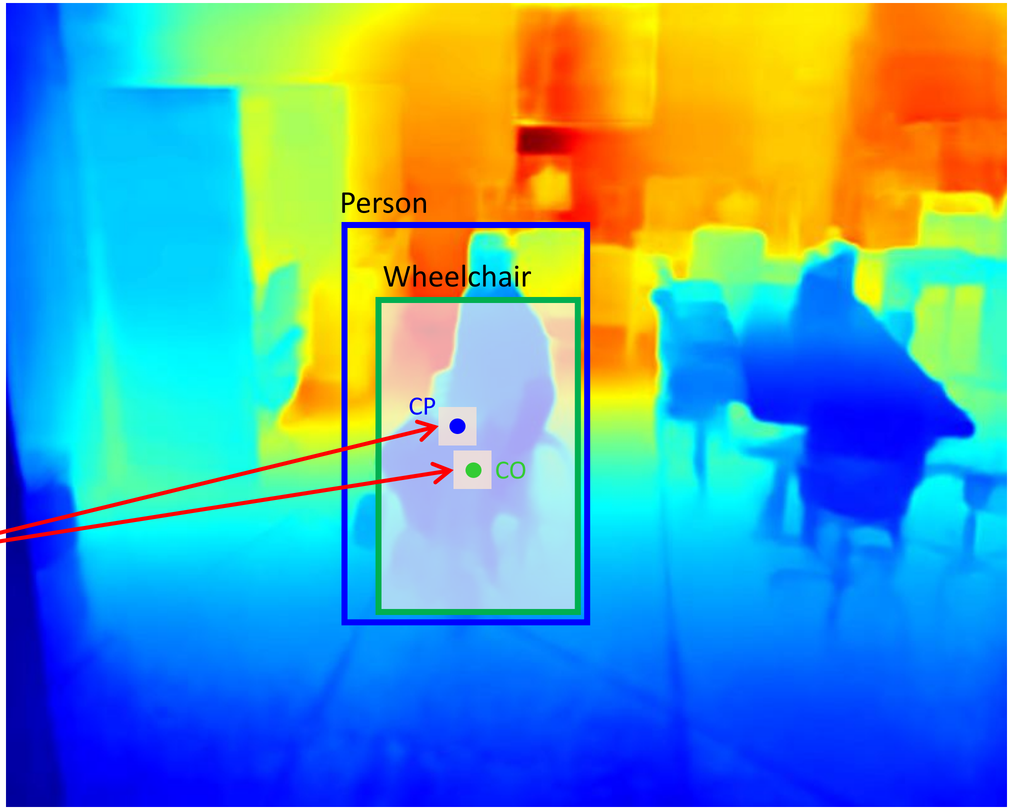
Posture Detection and Classification

Objects Filtering

Interaction Modeling and Detection

Step 4: Interaction Modeling and Detection – Cont.

- Metrics used for interaction modeling
  3. **Depth Estimation:** Handles 2D ambiguity in object size & distance.
     - AdaBins estimates monocular depth from a single image and creates depth heatmaps
     - Addresses issues with scale and object distance using adaptive binning
     - Average depth is calculated in a 10×10 pixel window at the center of the bounding box

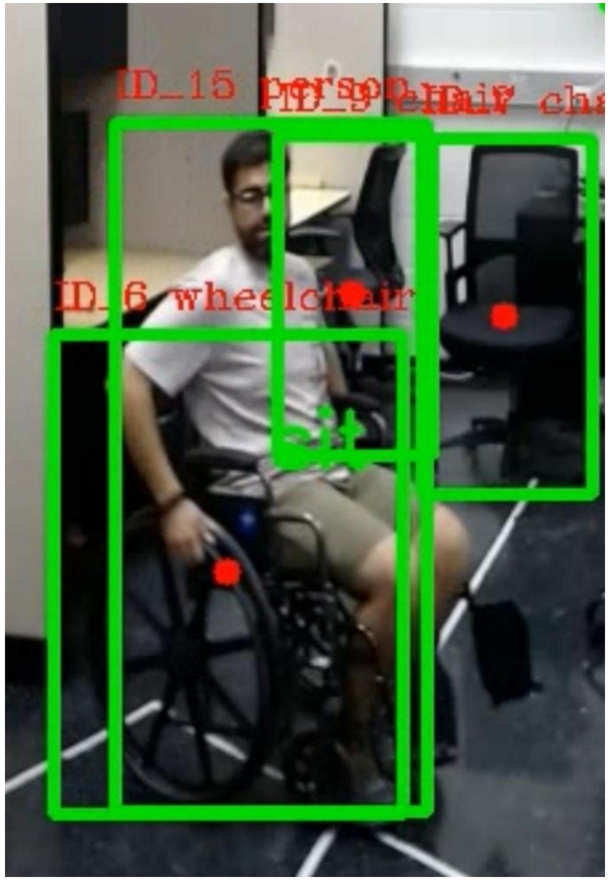


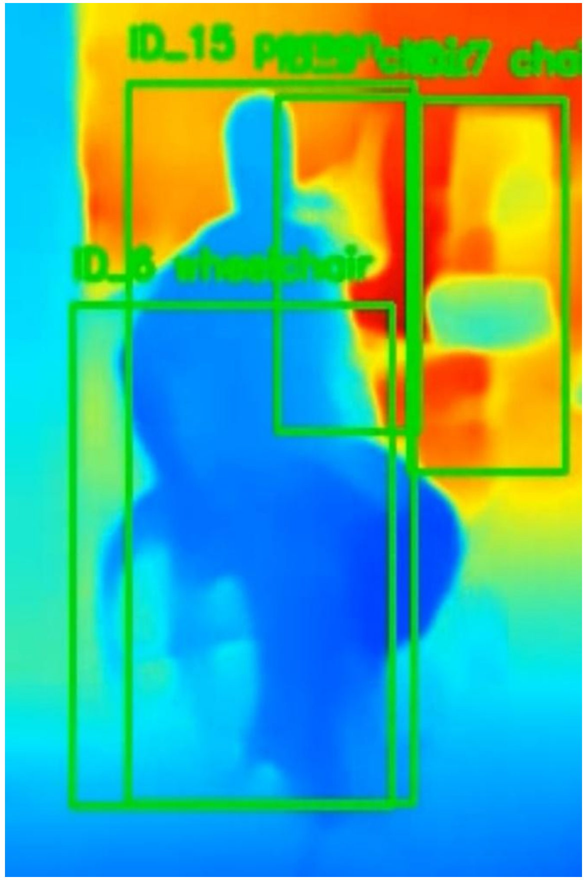
Person & Object Detection and Tracking → Posture Detection and Classification → Objects Filtering → Interaction Modeling and Detection

Step 4: Interaction Modeling and Detection – Cont.

The NOTES is good here. Those red dots are also good that I talk about.

- Object Interaction Detection
  - Sort objects based on maximum IoU values
  - Prioritize objects with minimum central distance to the patient
  - Refine the selection using depth value
  - Select the object whose depth value is closest to the patient's depth





Person & Object Detection and Tracking

Posture Detection and Classification

Objects Filtering

Interaction Modeling and Detection

Experimental Results – Setup & Data Collection

- Computational Platform
  - Laptop with 32GB of RAM, an Intel i9-12900HK CPU, and an NVIDIA GeForce RTX 3050 GPU
  - Customized MediaPipe using Python, and TensorFlow Lite for efficient model execution
  - A regular webcam Logitech C920S camera

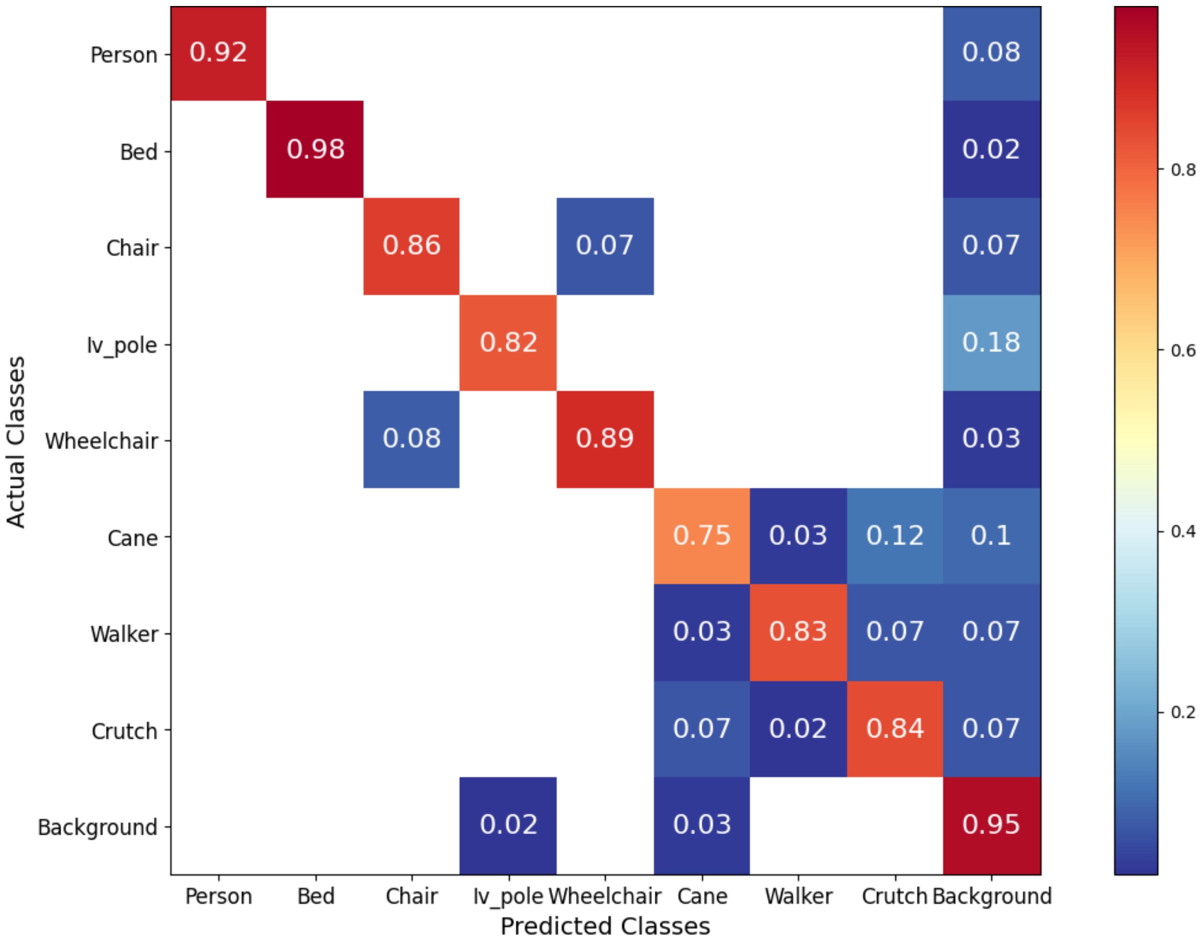
- Eight Subjects
  - Institutional Review Board IRB-24-776 was approved by University of Texas at Dallas.
  - Each volunteer performing three actions (Sitting, Standing, Lying) in controlled setting.

- Videos Collected
  - Eight recorded video segments (each approximately 20 minutes) covering 24 scenarios.

Experimental Results – Object Detection Confusion Matrix

- Relevant hospital-related objects
  - Higher detection accuracy
  - Improved training efficiency

- Figure shows the confusion matrix for 8 most common objects of interest



Experimental Results – Posture Detection

- We first created our own dataset and then used MediaPipe to extract 33 features from 11 key body landmarks

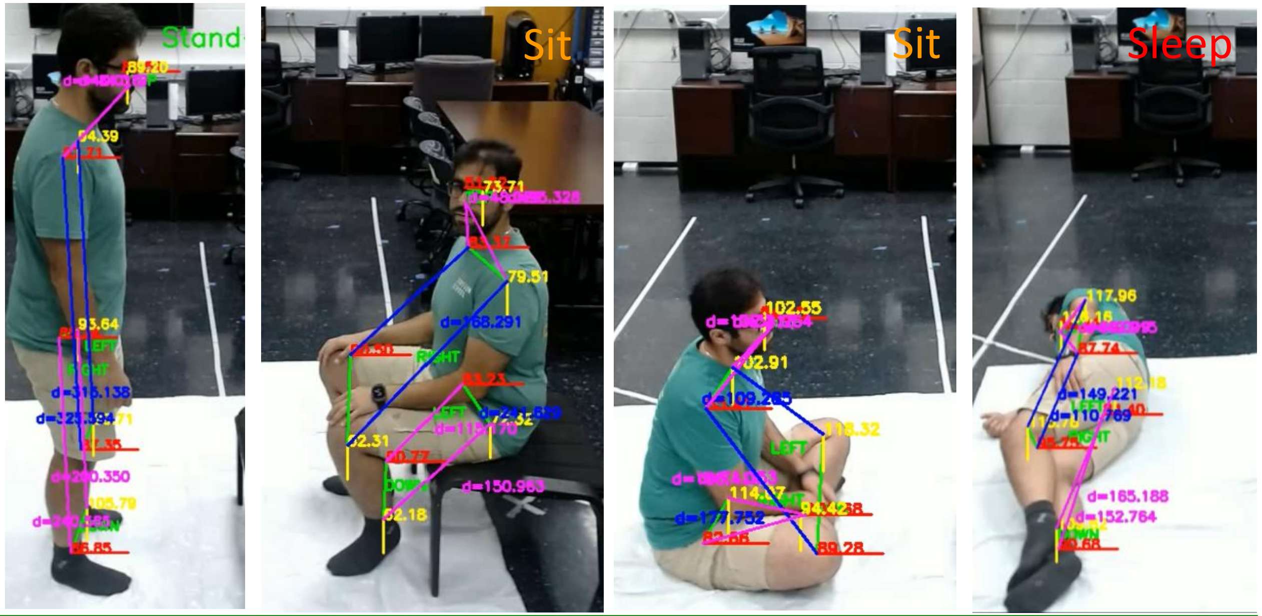
Postures	# of videos	# of Labeled Frames
Sit	14	12657
Stand	13	10597
Lie	13	17235

- These features were used to classify standing, sitting, and lying postures

- Employed 5-fold cross-validation to train and evaluate our Random Forest classifier

Classifier	Accuracy	Mean Error	F1 Score
Random Forest	99%	0.8%	98%

- Output images of feature extraction and detection for 3 postures



Experimental Results – HOI Detection Accuracy

- Accuracy for different postures when interacting with objects

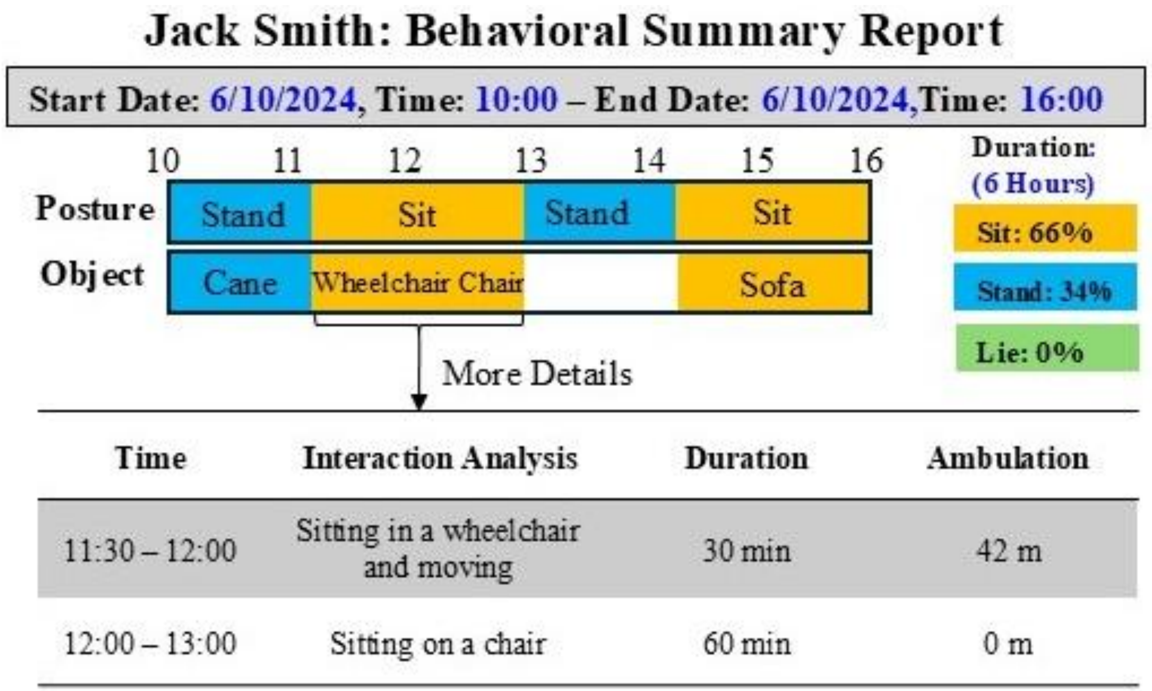
Postures	# of Test Frames	# of Correct Detections	Accuracy(%)
Siting	2174	1783	85
Laying	1947	1938	89
Standing	1870	1851	94

- Findings
  - Standing detection (highest) – clear posture distinction
  - Lying detection – slightly lower due to possible occlusions
  - Sitting detection (lowest) – more challenging due to environmental variations and occlusions

Experimental Results – Patient Activity Reporting

- By combining object detection, posture classification, and patient-object interactions, a comprehensive summary report is automatically generated.

- Summary report as an assistive tool
  - Generated on demand
  - Visually informative (includes time-stamped activities, behavioral routines and anomalies)
  - for alerting caregivers and clinical decision-making



Conclusions & Future Work

- A practical patient monitoring in clinical environments needs to be reliable, scalable, and non-intrusive.
- We combined image processing, computer vision and machine learning techniques to integrate
  - Person and object detection & tracking
  - Posture classification
  - Human-object interactions
- Future Works
  - Handling object, person, and environment variability
  - Enhancing summary reports using generative AI models

Thanks for Your Attention!

Mostafa Habibi, Zeinab Delaram, Mehrdad Nourani
Department of ECE
The University of Texas at Dallas
{mostafa.habibidehsheikhi, zeinab.delaram, nourani}@utdallas.edu

Dennis H. Sullivan
Geriatric Research, Education and Clinical Center
Central Arkansas Veterans Healthcare System
dennis.sullivan3@va.gov