A Deep Learning Approach to 3D Human Walking Pose Estimation From SensFloor Capacitive Signals

Anonymous Author(s)

Abstract

While camera-based systems are the dominant approach for human pose estimation, they face challenges in terms of privacy concerns and occlusion problems. These issues are of particular relevance in domains such as elderly care, where pose estimates can be used to monitor residents health or analyze incidents retrospectively. To assess an alternative for camera-based pose estimation, this paper aims to predict 3D walking poses using SensFloor: a capacitive-based floor which registers movement activity. We analyze the potential to utilize the floor's low resolution signals to estimate poses and to what extent certain joint positions can be predicted accurately.

To this end, we collected synchronized SensFloor signals and video data, from which we extracted 3D human poses using MediaPipe to serve as ground-truth labels for training. These signals and their corresponding labels were then used for supervised training of an LSTM neural network. To estimate the person's position on the floor, a Kalman filter was applied to smooth the noisy SensFloor measurements.

Our results demonstrate that it is possible to predict human walking poses using the proposed methods, establishing a proof-of-concept for an alternative way of activity monitoring.

Keywords

SensFloor, Human pose estimation, Deep Learning, LSTM, Kalman Filter, Capacitive Floor, MediaPipe

Code: https://github.com/sensfloor

Create README for public organization view

1 Introduction

2 Methods

The final goal of this project is to design a system that reads SensFloor signals, processes them, and visualizes the person's gait within a virtual simulation. To achieve this goal, we split up the problem into two components: pose estimation and localization. The process of both components is illustrated in figure 1. By separating these tasks, the system is able to use a small localized active area of the floor for pose estimation. Afterward this pose is mapped across the global floor coordinates. This enables the system to be ported to various floors of different dimensions.

For pose estimation component, we implemented a supervised fine-tuning pipeline. For that we collected reference pose estimates from a MediaPipe to serve as labels. Using these labels, we trained a Long Short-Term Memory (LSTM) to predict human poses based on SensFloor activations within an extracted Region of Interest (ROI). To localize the estimated pose, a Kalman Filter processes the noisy sensor signals to determine the person's global coordinates. Finally, the combined data of pose and position is streamed to a frontend application for real-time visualization.

[1]

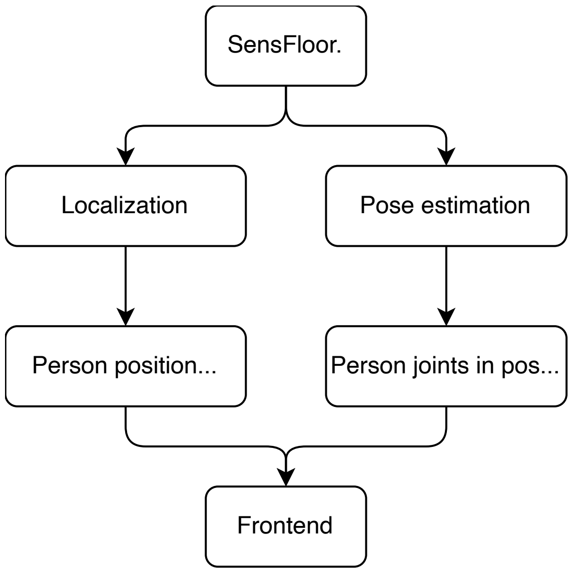

Figure 1: TEST TEST

2.1 Data collection

The model developed in this paper takes SensFloor signals as input to predict the 3D joint coordinates of a person walking on the floor. To train and evaluate such a model, we recorded in two sessions a total of eight hours of synchronized SensFloor signals and video from three participants. To synchronize the data, we mapped the SensFloor signals to the corresponding video frame number at their time of occurrence. In post-processing, we used MediaPipe to extract 3D pose estimates from the video to serve as training targets. It is important to mention here that the origin of MediaPipe's pose estimates coordinate system is located at the center of the hips. This allows us to decompose the paper's problem into the previously mentioned two components of pose estimation relative to the hip and localization on the whole floor.

Look up actual duration

cite

2.2 Training Setup

We used a CNN-LSTM architecture for training our model. It is based on the model from . Instead of rotating the step, we use a CNN to account for the Translation invariance. Figure STH shows the full architecture. Instead of giving the model all signals of the floor, we construct a Region of Interest (ROI) which is a part of the floor containing the most signals in a defined sequence. A sequence is a collection of frames. For each frame we have a state of the floor with certain signals. Our model receives the signals in the ROI frame by frame and produces a vector with continuous values as logits. These logits are given into our Loss function which can be mathematically described in the following We substract

cite correctly

replace with correct equation

the predictions from the reference labels, square and multiply the weights of the landmarks. The link loss is identical to the one in We assume an upright position, this is why the upper body and the head will only rotate and the focus is more on the knee and feet Mediapipe extracts 33 Joints, we only used 13, excluding the fingers, face expression, heels and foot index. For each epoch we calculate the following metrics: Mean joint position error (MJPE), percentage correct keypoits (PCK) with a 5 and 10 threshold

replace with correct cite

We assume only one person is walking on the floor

3 Results

4 Discussion and Conclusion

References

[1] Mic Bowman, Saumya K. Debray, and Larry L. Peterson. 1993. Reasoning About Naming Systems. *ACM Trans. Program. Lang. Syst.* 15, 5 (November 1993), 795–825.