ODIN: An OmniDirectional INdoor dataset capturing Activities of Daily Living from multiple synchronized modalities

Siddharth Ravi[1] [0000-0002-2301-569X], Pau Climent-Perez[1], Théo Morales[2],
Carlo Huesca-Spairani[1], Kooshan Hashemifard[1], Francisco Florez-Revuelta[1]

[1]University of Alicante [2]Trinity College Dublin

{siddharth.ravi, pau.climent, k.hashemifard, francisco.florez}@ua.es,
moralest@tcd.ie, chs17@alu.ua.es

Abstract

We introduce ODIN (the OmniDirectional INdoor dataset), the first large-scale multi-modal dataset aimed at spurring research using top-view omnidirectional cameras in challenges related to human behaviour understanding. Recorded in real-life indoor environments with varying levels of occlusion, the dataset contains images of participants performing various activities of daily living. Along with omnidirectional images, additional synchronized modalities of data are provided. These include (1) RGB, infrared, and depth images from multiple RGB-D cameras, (2) egocentric videos, (3) physiological signals and accelerometer readings from a smart bracelet, and (4) 3D scans of the recording environments. To the best of our knowledge, ODIN is also the first dataset to provide camera-frame 3D human pose estimates for omnidirectional images, which are obtained using our novel pipeline. The project is open sourced and available at https://odin-dataset.github.io.

1. Introduction

Challenges relating to the analysis of Activities of Daily Living (ADL) have become essential topics of research in computer vision and active and assisted living [3,7,20]. Examples of these challenges include human pose estimation and activity recognition. For the rest of the paper, these will be referred to as human behaviour understanding (HBU) challenges. Most of the research in these fields is done using lateral-view RGB(-D) images as inputs. However, recording these images introduces a practical problem: activities can be easily occluded, such as when the user being monitored faces away from the camera. Additionally, these cameras are obtrusive because they are constantly in the field of view of the user [5]. On the other hand, ceiling-mounted omnidirectional cameras with fisheye lenses provide a worthy solution to these problems. These cameras are generally unobtrusive, have a larger field of view, and can provide largely unoccluded views of the environments being monitored. However, HBU challenges such as pose estimation become all the more challenging due to the viewpoint and due to the heavy distortions introduced by the lens when compared to wide-angle lenses.

The aim of this work is to introduce a new large-scale omnidirectional dataset which contains numerous synchronized modalities. This includes images and videos from cameras of different types recording participants carrying out various activities of daily living, along with their physiological data. ODIN will support research in areas as varied as human pose estimation, activity recognition, person tracking and monitoring, scene understanding, privacy preservation, biometric monitoring, novel view synthesis, generative modelling, 3D scene reconstruction, and image registration. Through our first release, we aim to promote research on 3D human pose estimation using omnidirectional cameras. Research in this area is scarce, arguably due to the difficulty of the problem and the dearth of datasets. For the omnidirectional camera images, the dataset provides associated camera-frame 3D pose estimates. We propose a novel unsupervised pipeline for obtaining these pose estimates in real-life indoor settings while preserving the state of the environment, and also without the use of expensive equipment.

The main contributions of this work are as follows:

- This work introduces a large-scale dataset of omnidirectional images capturing a diverse range of activities of daily living recorded from real-life settings with varying levels of occlusions. Additionally, synchronized data from various viewpoints and of different modalities are also provided. These include:

    1. Images recorded from multiple calibrated lateral-

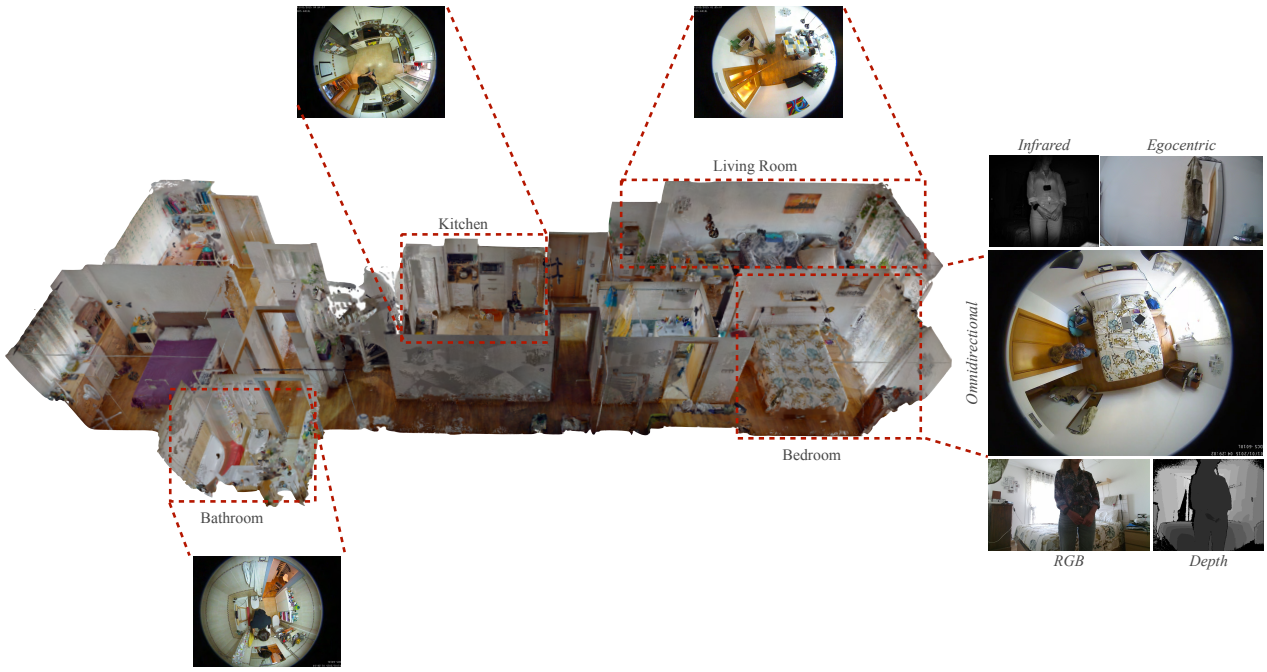

Figure 1. **Overview of ODIN, a large-scale omnidirectional dataset for Human Behaviour Understanding** – Each sequence is composed of the 3D scan of the recording location and omnidirectional RGB images as well as 5 extra modalities: (1) depth, (2) IR, (3) RGB images from side views, (4) RGB egocentric images and (5) biometric signals from a wearable device. Four of the five environments — kitchen, living room, bathroom, and bedroom — are represented in the figure (the activity room can be seen in Fig. 4).

view RGB-D cameras placed in the environment, each providing high-resolution RGB, infrared, and depth images, along with the calibration files.

2. Videos from a chest-mounted egocentric camera worn by the participant.
3. Physiological signal recordings from a smart bracelet.
4. Scans of the recording environments from a 3D scanner.

- Additionally, we provide various processed information derived from the images, such as camera frame 3D body pose estimates from all static camera viewpoints.

To summarize, ODIN is a large multi-modal multi-viewpoint dataset for HBU challenges that makes use of real-world recording environments. To the best of our knowledge, our dataset is the first to provide camera-frame 3D pose estimates on omnidirectional images. Fig. 1 illustrates some different modalities and environments that can be seen in ODIN.

2. Related Work

Tab. 1 compares the modalities present in ODIN to some commonly used omnidirectional datasets along with some relevant datasets in ADL.

Omnidirectional datasets: The works that are closest in scope to ODIN are WEPDTOF [31] and the PIROPO database [10]. WEPDTDOF contains 14 YouTube videos that capture people from top-view omnidirectional cameras and contains annotations for people detection and tracking. PIROPO contains 100k annotated frames of people in various situations, such as walking or standing. The dataset focuses on localizing people with a point-based ground truth located at the centre of their heads. PIROPO is also recorded in an uncluttered and artificial recording setup. While this is beneficial for providing unobstructed recordings, actual living environments contain more clutter, which is what ODIN captures.

Another closely related work is the fisheye dataset proposed by Eichenseer and Kaup [12]. The dataset provides synthetic and real-world video sequences, along with the camera calibration information, to facilitate the creation of image and video processing algorithms that fail on typical omnidirectional cameras. ODIN, on the other hand, is a multi-modal dataset containing top-view omnidirectional images which will help spur research in a multitude of areas, including pose estimation.

Other omnidirectional datasets captured using fisheye lenses do also exist, with these mostly aimed at challenges

such as the improvement of autonomous driving systems [27, 32] and robotics [6]. Synthetic omnidirectional datasets for HBU tasks also exist [22, 26].

ADL datasets: Several datasets, including the ADL dataset [21], HumanEva dataset [29], the Penn Action dataset [33], the Human3.6M dataset [16], the NTU RGB+D dataset [28], the Toyota Smarthome Dataset [8] and the MPII Human Pose dataset [2], have been created for research into HBU challenges. The EPIC-KITCHENS dataset [9] captures multiple modalities (egocentric, RGB, and depth) and contains a wide range of ADLs. The largest effort to date that captures a wide range of modalities is the more recent Ego4D dataset [14]. While primarily being a dataset of egocentric camera videos of ADL, portions of the recordings are also accompanied by stereo, audio, environment meshes from 3D scanners, along with other modalities. These efforts have primarily focused on standard RGB and egocentric cameras.

Omnidirectional cameras in HBU: Although arguably a nascent field, there have been a few attempts to use omnidirectional cameras for HBU challenges, although datasets to facilitate these have never been publicly released to the best of our knowledge. These are mostly in the field of classification of poses using omnidirectional cameras. Akama et al. [1] create a method for the detection of human activities in an indoor environment by using multiple omnidirectional cameras. The method allows the tracking of the participant being monitored and the classification of four typical postures for indoor activities. Georgakopoulos et al. [13] create a method for classification of poses using silhouettes obtained from omnidirectional cameras into 3 distinct poses, namely falling, sitting, and standing.

ODIN is the first dataset to provide full-body 3D pose estimates for top-view omnidirectional images, which allows for the training of omnidirectional 3D pose estimation models. The available data is also expected to expand in scope in the future to directly enable other HBU challenges.

3. The Dataset

Consisting of 55 distinct sequences from 15 participants, the dataset comprises recordings from 4 locations, in 5 different types of environments (kitchen, bathroom, bedroom, living room, and activity room). The sequences in the recording are of varying times, with most lasting under 10 minutes. An overview of the statistics of the dataset can be seen in Tab. 2, and in Fig. 2.

In addition to synchronized images from RGB-D devices, we provide synchronized physiological data and wrist motion accelerometer readings from a smart bracelet and 3D meshes of the recording areas from a 3D scanner. The omnidirectional camera images are accompanied by a camera frame 3D pose estimate, obtained through a novel pipeline explained in Sec. 4. Additionally, we provide the intrinsic and extrinsic calibration information, i.e., camera matrices that allow users to perform perspective projection between the lateral-view and top-view cameras.

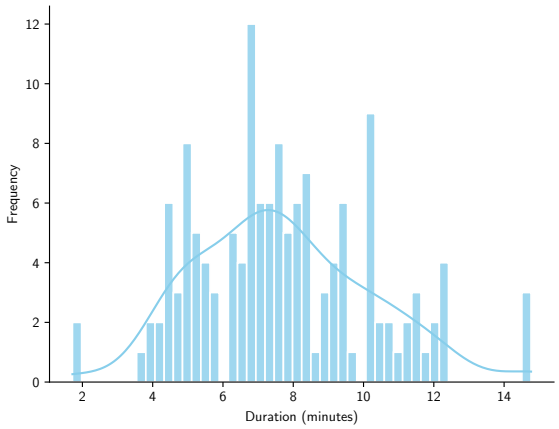

Figure 2. **Distribution of the duration of sequences in ODIN** – Most sequences last between 6 and 10 minutes in duration, with the longest ones being around 15 minutes.

3.1. Collection Methodology

Equipment Used: The recordings are conducted using a static ceiling-mounted D-Link DCS-6010L omnidirectional camera with a fisheye lens. It has a 180-degree field of view when mounted on ceilings and a 2-megapixel sensor that can capture 1080p videos. Up to 3 Microsoft Kinect v2 devices monitor the same recording environment synchronously as static lateral-view devices. The Kinect v2 features an array of sensors that enable it to detect and track users' movements in three dimensions, including a high-definition RGB camera, a depth sensor, an infrared emitter, and a microphone array. The field of view of the device is 70 degrees horizontal and 60 degrees vertical, with a range of 0.5 to 4.5 meters. The embedded depth sensor uses time-of-flight technology to calculate the distance between the sensor and objects in the environment, allowing it to create a detailed 3D map of the environment. For ODIN, all static cameras record at a frame rate of 15 fps.

A Xiaomi Mi action camera 4K was used in a chest-mounted modality as the egocentric camera, recording at a frame rate of 25fps. This is a compact and lightweight action camera that can capture high-quality video and photos. It features a Sony IMX317 sensor, a 145-degree wide-angle lens, and a built-in 2.4-inch touchscreen display. A Matterport Pro2 3D scanner was used to scan the recording environments to obtain 3D meshes. The Matterport Pro2 is a professional-grade 3D scanner designed for creating high-quality 3D models of real-world spaces. The scanner has a range of up to 3 to 4.5 meters and can capture up to 134 megapixels of visual data per scan, resulting in highly de-

Dataset	Omni	Ego	RGB	3D scans	Stereo	IMU	Synced-cam	Phys. signals	Pose	Activity labels	Audio
ODIN	✓	✓	✓	✓	✓	(Partial)	✓	✓	✓	(×)	×
PIROPO Database [10]	✓	×	×	×	×	×	✓	×	×	✓	×
WEPDTOF [31]	✓	×	×	×	×	×	×	×	×	×	×
Fisheye dataset [12]	✓	×	×	×	×	×	×	×	×	×	×
MPII Human Pose [2]	×	×	✓	×	×	×	×	×	✓	×	×
Human3.6M [16]	×	×	✓	×	×	×	✓	×	✓	×	×
Toyota Smarthome [8]	×	×	✓	×	✓	×	✓	×	✓	✓	×
NTU RGB+D Dataset [28]	×	×	✓	×	✓	×	✓	×	✓	✓	×
ADL Dataset [21]	×	×	✓	×	×	×	×	×	×	✓	×
EPIC KITCHENS [9]	×	✓	×	×	×	×	×	×	×	✓	×
Ego4D [14]	×	✓	✓	✓	✓	✓	✓	×	×	✓	✓

Table 1. **Comparison of modalities in state-of-the-art datasets related to ODIN** – RGB points to static RGB camera images, synced-cam refers to synchronized multi-camera setups, stereo refers to depth/infrared images, and IMU refers to inertial measurement unit readings. ODIN only provides wrist-motion accelerometer readings, and hence is marked as partial. Activity labels are planned for a future release

Modality/characteristic	Amount
Omnidirectional RGB images	332K
Lateral-view RGB images	1.464M
Lateral-view infrared images	1.464M
Lateral-view depth images	1.453M
Environment meshes	3
Egocentric videos	52
Physiological readings	39
Accelerometer measurements	39
Participants	15
Locations	4
Types of environments	5

Table 2. **Dataset statistics** – ODIN consists of more than 300K omnidirectional images, making it the first large-scale dataset for HBU.

tailed and realistic models.

An Empatica E4 smart bracelet was employed for recording the physiological measurements. The Empatica E4 is a Class IIa medical device that measures accelerometer data, in additional to the physiological readings relating to blood volume pulse, heart rate variability, inter-beat interval, skin temperature, and electrodermal activity. The device's accelerometer sensor measures the continuous gravitational force along the three spatial directions (x, y, and z axes) with a sampling frequency of 32 Hz and a range of $\pm 2g$. The conversion factor between raw acceleration samples and true values is $g/64$ (where $g = 9.81\text{m/s}^2$). Additionally, electrodermal activity (EDA), i.e. changes in conductivity, are sampled at 4 Hz (range is $[0.01, 100]$ μSiemens); blood volume pulse (BVP) is sampled at 64 Hz (range $[-500, 500]$), the inter-beat interval (IBI) and heart rate (HR) are derived from it (HR at 1 Hz). Finally, the skin temperature (not representative of core body temperature), is sampled at 4 Hz. The device calibrates automatically during the initial 15 seconds of each session. The device was used in recording mode, in which it stores data directly on the device's internal memory.

The settings for the equipment used during recordings can be found in Tab. 3.

Participants: 15 adult participants have taken part in the creation of the dataset, and they have been requested to participate through word-of-mouth. Recordings have been captured in 4 real-life indoor environments in the south-east of Spain. The participants are told to carry through, in no particular order, a set of single-person activities that can naturally be performed in the corresponding indoor environment. As a large amount of emphasis is placed on the realism of the recordings, the participants are requested to act naturally as they would have if the recording equipment were not present. They are also told to stop the session when they feel they have exhausted the set of activities they are comfortable performing during the session. A non-exhaustive list of activities carried out by the participants per type of environment can be seen in Tab. 4.

Environment Layout: Before recording each participant, the environment is analysed to understand where to mount the cameras for recording, while also preserving the original layout. An illustration of the placement of the recording devices in a typical environment can be seen in Fig. 3.

3.2. Recordings

A variety of indoor environments have been chosen for the recordings, including living rooms, bathrooms, kitchens, and activity rooms. A set of activities relevant to

Device Type	Specification	Brand
Omnidirectional camera	15fps, 1200×900px	D-Link DCS-6010L
RGB-D cameras	15fps, 512×424px (depth/IR) 1280×720px (RGB)	Microsoft Kinect v2
Egocentric camera	25fps, 3840×2160px	Xiaomi Mi Action camera 4K
Smart bracelet	Accel.: 32 Hz, 0.015 g resolution EDA: 4 Hz, $[0.01, 100]$ μS Temperature: 4 Hz BVP: 64 Hz, $[-500, 500]$ HR & IBI: 1 Hz	Empatica E4
Environment scanner	134.2 MP, Export images-8092×4552px	Matterport Pro2

Table 3. **Devices used and their specifications while recording ODIN** – 5 different types of devices were used to record the dataset, resulting in a plethora of modalities being collected.

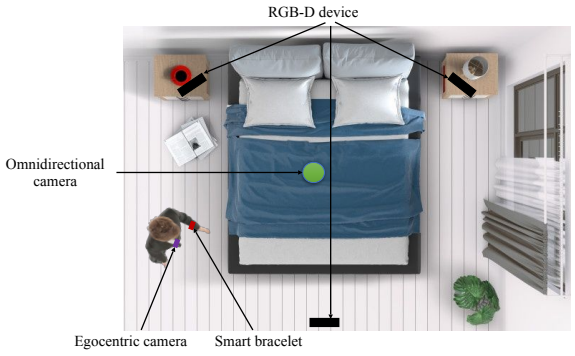

Figure 3. **Illustration of a typical environment layout** – The RGB-D devices are positioned so that their viewpoints capture nearly all possible angles of the room, while the ceiling mounted omnidirectional camera is placed towards the centre of the room.

each environment is determined and communicated to the participants. All recordings are done non-scripted, and the participant is free to choose which activities to perform and to interact freely with the environment during the recording.

Calibration and Synchronization: The recordings are collected using calibrated cameras. This allows us to do perspective projections of the pose estimates obtained from using the RGB-D camera images as input. Immediately before recording at each location, a checkerboard is placed at a location in the environment visible through one RGB-D device and the omnidirectional camera simultaneously. This is then used to calculate the extrinsic matrix required to project from the viewpoint of the RGB-D device to the top-view omnidirectional camera. This procedure is repeated for each RGB-D device present.

Along with the omnidirectional camera images and the corresponding top-view pose estimates, we provide, as an additional modality, synchronized images from multiple RGB-D devices recording the environment. The recording files from each static camera are named according to the UTC timestamp at the time of the recording. As these are bound to have errors due to the different computers not being in sync, the network transfer speeds, and the speed of writing to disk, we also utilized a visual cue to synchronize the feeds. At the beginning of each recording, the participant is requested to toggle a light in the environment. The moments when the light is toggled is used to fine-tune and label the images that correspond to each other from the different static cameras, and also to mark the timestamp on the egocentric video.

Environment	Activities
Kitchen	Wash dishes, drink water, prepare a snack, eat a snack, take a pill, prepare a hot beverage, drink beverage, wash hands, dry hands, use a microwave, Use the stove top, use a kettle
Bathroom	Wash hands, comb hair, wash face, brush teeth, use the toilet, dry hands, dry face, use cell phone
Bedroom	Make phone calls, use cell phone, lie on bed, exercise, drink a beverage, take a pill, take off/put on shoes, take off/put on jacket, use laptop
Living room	Watch TV, lie on the couch, play with tablet, use cell phone, make phone calls, drink water, eat a meal, lie on the floor, take off/put on glasses, take off/put on shoes, cough
Activity room	Dance, do a handstand, exercise, read a book, use cell phone, drink a beverage, use laptop, walk around, take off/put on glasses, take off/put on shoes, sit on the floor, lie on the floor

Table 4. **Non-exhaustive list of activities performed in ODIN** – 5 types of environments were chosen, with a variety of relevant activities done per environment

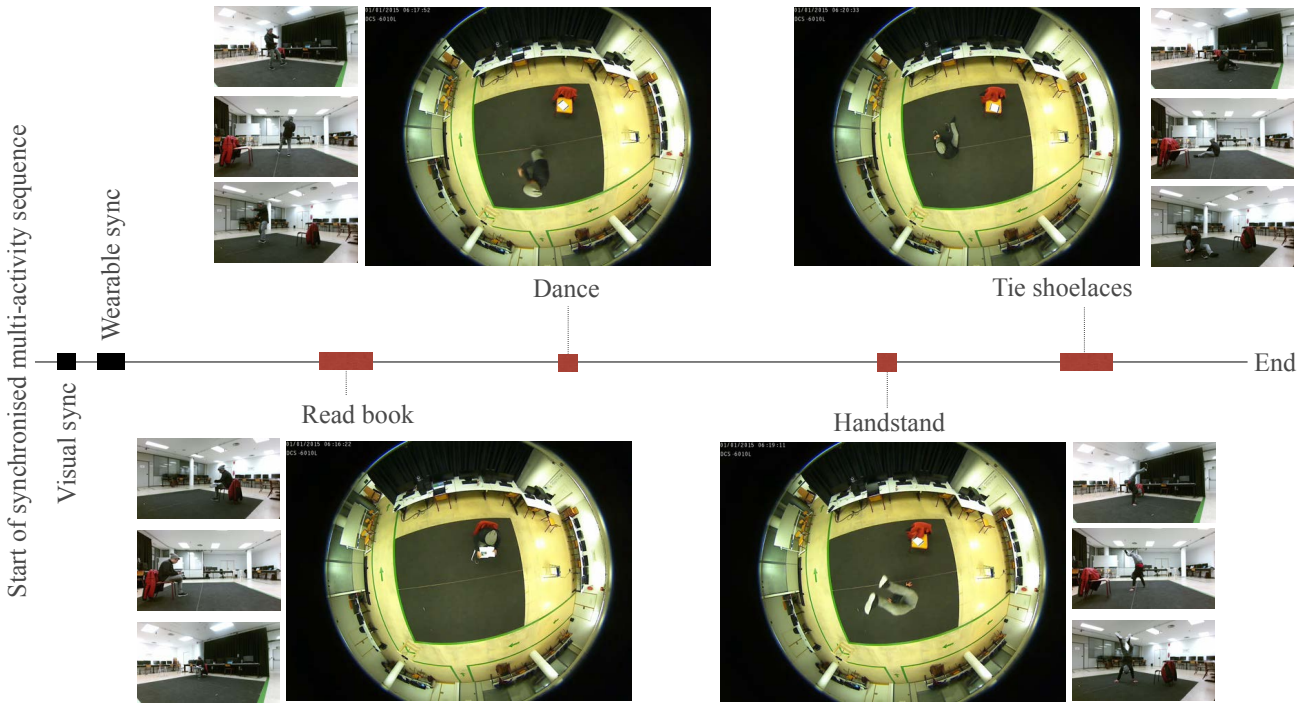

Figure 4. **Timeline of a multi-activity sequence recorded from the activity room environment** – Synchronized RGB visuals from every static camera are shown in the image. The participant starts with the visual device synchronization procedure using a light toggle, followed by the wearable device synchronization. They then proceed to do a set of non-scripted activities.

To synchronize the smart bracelet, the device is shown during its calibration phase to one of the RGB-D cameras. This enables the images to take note of the moment when the device starts recording. The device represents this moment through a visual cue, an LED indicator changing colour.

4. Pose Annotation

To tackle many of the complex HBU challenges in omnidirectional computer vision, full-body pose estimates obtained in camera or image frame and the activity labels per time frame are required. These provide rich information on the temporal scene and allow training deep-learning-based methods to solve high-dimensional problems from large data sets. Automatically annotating images from omnidirectional cameras is a hard problem due to the extreme image distortion in such images and low visibility along the edges. Motion capture (MoCap) systems can be used to obtain accurate pose estimates by tracking the positions of markers on a participant's body to estimate the location of their joints. MoCap systems typically use a rig of cameras along with other sensors to track the movement of these markers in 3D space, and specialized software is then used to accurately compute the position and orientation of the joints based on the marker data. This, however, requires the use of expensive equipment and software. Moreover, MoCap systems require a clear line of sight to obtain accurate estimates, which is not always possible while recording in indoor environments [11].

For ODIN, the problem of obtaining camera-frame 3D pose estimate annotations for omnidirectional cameras is approached by obtaining pose estimates using a state-of-the-art 3D temporal pose estimation model (HuMoR [24]) on the RGB-D camera images. These are then combined, and then projected to the view of the omnidirectional camera using perspective projection. All this is obtained through a 4 stage pipeline, starting with the curation, calibration, and synchronization of the images.

Calibration, and Synchronization: After the recorded images are curated, the images are synchronized between all static cameras using the visual cues mentioned in Sec. 3.2. The extrinsic matrices required to do a perspective projection from the camera frame of the RGB-D devices to the camera frame of the omnidirectional camera is then estimated using Zhang's algorithm [34] implemented in OpenCV [4]. For the intrinsics of the omnidirectional camera images, a modified version of Zhang's algorithm is utilized.

A state-of-the-art temporal pose estimation model is

(a) Lateral-view (b) Top-view

Figure 5. **Example of body pose annotation of the same sequence from two views** – Fig. 5a shows the pose estimated when using an image from a synchronized RGB-D device as input. Fig. 5b showcases the same pose estimate reprojected to the omnidirectional camera.

used for the pose estimation. Computing 3D pose estimates using the model on a large number of images is a computationally intensive task. So, we first filter out the frames that have no person in them to decrease the amount of computation necessary. This also prevents error accumulation in the temporal pose estimation model. The filtering is done using a pose estimation model, Densepose [15].

Person Detection: Densepose is a pose estimation model that provides accurate correspondences between RGB images and a surface representation of the human body. The model generates IUV maps, which encode body part indices (I), and UV coordinates of each visible body part. Using the generated IUV maps and a segmentation map derived from it, we create a metric to determine whether a person is in an image or not. We extract the pixels of a body part in the IUV map, and then compute the intersection of those pixels with the pixels corresponding to the person's segmentation mask. We then compute the weighted visible surface area of the body part a_p according to the function $a_p = \frac{\max(50, \sum_{i \in I_p} I_i \cap S)}{\sum_{j \in S} S_j}$, where $I_i = \mathbb{1}_i(I_p)$ and $S_j = \mathbb{1}_j(S)$. Here I_p is the set of indices of pixels corresponding to part p in the IUV map, and I_i is an indicator variable that is 1 if the pixel i is part of I_p and 0 otherwise. S is the set of indices of pixels corresponding to the person's body in the segmentation mask, and S_j is an indicator variable that is 1 if pixel j is part of S and 0 otherwise. This is then compared to the computed surface areas with an ideal distribution, and the KL divergence between the two distributions is computed. The ideal distribution D is computed as $D = \{d_p\}_{p \in P}$ with P being the set of 8 body parts comprising the whole human body. We compute the distribution of surface areas for each part p that is visible in the frame as $a_p \in A$, with A being the distribution of visible parts. We then compute the Kullback-Leibler (KL) divergence [17] between D and the visible distribution A as $\text{div} = \text{KL}(D \parallel A) = \sum_{p \in P} d_p \log \frac{d_p}{a_p}$.

We then apply a sigmoid function to the KL divergence to obtain a *diversity* metric that ranges from 0 to 1, with higher values indicating more even body coverage across different body parts. It is predicted as diversity $= \frac{1}{1+e^{-\text{div}}}$. If the diversity of body parts is determined to be more than a threshold in an image, a person is ascertained to be in the image.

Pose Estimation: For the images where participants are in a frame according to the diversity metric, we use HuMoR [24] to obtain the 3D pose estimates on the lateral-view images. HuMoR is a state-of-the-art pose estimation model that predicts full-body temporal poses in camera frame. It uses a conditional variational autoencoder (CVAE) [30] to learn the distribution of the transition in pose at each step of a motion sequence. The model also relies on test-time optimization during inference to optimize the model's performance.

In HuMoR the temporal state of a moving person x is estimated as a sequence of pose parameters partly represented with the SMPL+H body model [18, 25]. The probability of a time sequence of states pr_θ is modelled, with each state assumed to be dependent only on the previous one, and θ being learned parameters. The probability is modelled through the use of a conditional variational autoencoder, with a learned conditional prior. The probability model for the pose transition used by HuMoR can be described as $Pr_\theta(x_t|x_{t-1}) = \int_{z_t} Pr_\theta(z_t|x_{t-1}) Pr_\theta(x_t|z_t, x_{t-1})$, with $z_t \in \mathbb{R}^{48}$ being the latent variable described by the conditional prior. $Pr_\theta(x_t|x_{t-1})$ aims at capturing the plausibility of a state transition. Thus, HuMoR models the most plausible temporal pose per given sequence.

Reprojection: The obtained poses from each of the synchronized cameras are averaged to obtain one singular pose

Figure 6. **Example of incorrect pose annotation** – the pose annotation is shifted, and the mesh is incorrectly predicted.

estimate. Using the camera extrinsic parameters calculated through the camera calibration, the pose estimates are then reprojected to the corresponding omnidirectional camera using a perspective projection transformation. Thus, full-body pose estimates are finally obtained for the recorded omnidirectional camera images. An example of the pose obtained can be seen in Fig. 5.

5. Possible Sources of Bias / Errors

While the dataset aims to spur research in novel directions using omnidirectional cameras, we also acknowledge a few sources of bias and errors in the data. We create pose estimates using an automated pipeline relying on state-of-the-art deep learning models. While being accurate, this does sporadically lead to errors, an example of which can be seen in Fig. 6. Efforts are ongoing to address this issue. Furthermore, while the activities and locations are diverse, the corpus of 15 participants is also some way away from capturing the whole range of poses and body shapes that the human body can take.

6. Future Work

ODIN is created to advance research in multiple fields using omnidirectional cameras. The first version released is designed for omnidirectional pose estimation research, and we plan to release a pose estimation benchmark in the near future. Subsequent versions are aimed at helping researchers to create models related to other challenges in HBU, including omnidirectional action recognition. For this purpose, the dataset is to be annotated for understanding the activities happening on screen. The multiple viewpoints and synchronized modalities existing in ODIN also allow research in fields including generative modelling, 3D scene reconstruction, and image registration. Another distinct research direction concerns scene understanding. As the recordings are conducted in real living and working spaces, the environments contain a multitude of real-life objects. When annotated, this allows for research in fields as varied as object pose estimation, object detection, recognition, and tracking. Overlaying the pose estimates on the images do provide body privacy, but research can also be done in the direction of providing environmental privacy [23], as objects in the environment could be causes for privacy breaches.

7. Conclusion

In this paper, we present ODIN, a large-scale omnidirectional dataset with numerous synchronized modalities for promoting research on top-view pose estimation using images from highly distorted fisheye lenses. The proposed dataset presented in this paper is unique because it contains recordings from numerous synchronized cameras of different types, recording individuals performing various activities of daily living in actual living and working scenarios. We hope that ODIN will spur further research in the field of ADL understanding using omnidirectional cameras.

Ethics and Privacy: In creating ODIN, ethics and privacy were paramount to the collection effort. Data was collected in compliance with applicable laws and regulations, with all participants providing informed consent before participation. The study protocol was approved by the Ethics Committee of the University of Alicante (UA-2022-10-16_2). Homeowners also granted consent for inclusion of their 3D scanned properties in the dataset.

We acknowledge the ethical concerns surrounding the analysis of human behaviour using surveillance cameras. To address these concerns and promote responsible research practices, we have implemented the following measures. Firstly, we employed an anonymization pipeline using Mediapipe [19] to detect faces in the images, which were then blurred to protect individuals' identities. Secondly, we restrict access to the dataset to researchers and practitioners who agree to use the data for academic and research purposes, adhering to the guidelines outlined in our data usage agreement. We also encourage transparency, open discussion of ethical issues, and best practices to maximize benefits and minimize harm from computer vision technologies.

Acknowledgements: The authors declare no potential conflicts of interest regarding commercial or financial relationships in the research, and thank the participants in the dataset. This work is part of the visuAAL project on Privacy-Aware and Acceptable Video-Based Technologies and Services for Active and Assisted Living. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 861091.

References

[1] Shunsuke Akama, Akihiro Matsufuji, Eri Sato-Shimokawara, Shoji Yamamoto, and Toru Yamaguchi. Successive human tracking and posture estimation with multiple omnidirectional cameras. In *2018 Conference on Technologies and Applications of Artificial Intelligence (TAAI)*, pages 46–49, 2018. 3

[2] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014. 3, 4

[3] Djamila Romaissa Beddiar, Brahim Nini, Mohammad Sabokrou, and Abdenour Hadid. Vision-based human activity recognition: a survey. *Multimedia Tools and Applications*, 79:30509–30555, 2020. 1

[4] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000. 6

[5] Timothy Callemein, Kristof Van Beeck, and Toon Goedemé . How low can you go? privacy-preserving people detection with an omni-directional camera. In *Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*. SCITEPRESS - Science and Technology Publications, 2019. 1

[6] D. Caruso, J. Engel, and D. Cremers. Large-scale direct SLAM for omnidirectional cameras. In *International Conference on Intelligent Robots and Systems (IROS)*, sept 2015. 3

[7] Yucheng Chen, Yingli Tian, and Mingyi He. Monocular human pose estimation: A survey of deep learning-based methods. *Computer Vision and Image Understanding*, 192:102897, 2020. 1

[8] Rui Dai, Srijan Das, Saurav Sharma, Luca Minciullo, Lorenzo Garattoni, Francois Bremond, and Gianpiero Francesca. Toyota smarthome untrimmed: Real-world untrimmed videos for activity detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2022. 3, 4

[9] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The EPIC-KITCHENS dataset. In *European Conference on Computer Vision (ECCV)*, 2018. 3, 4

[10] Carlos R. del Blanco, Pablo Carballeira, Fernando Jaureguizar, and Narciso García. Robust people indoor localization with omnidirectional cameras using a grid of spatial-aware classifiers. *Signal Processing: Image Communication*, 93:116135, 2021. 2, 4

[11] Yann Desmarais, Denis Mottet, Pierre Slangen, and Philippe Montesinos. A review of 3d human pose estimation algorithms for markerless motion capture. *Computer Vision and Image Understanding*, 212:103275, 2021. 6

[12] Andrea Eichenseer and André Kaup. A data set providing synthetic and real-world fisheye video sequences. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1541–1545, 2016. 2, 4

[13] S.V. Georgakopoulos, K. Kottari, K. Delibasis, V.P. Plagianakos, and I. Maglogiannis. Pose recognition using convolutional neural networks on omni-directional images. *Neurocomputing*, 280:23–31, 2018. Applications of Neural Modeling in the new era for data and IT. 3

[14] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina González, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jáchym Kolář, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbeláez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18995–19012, June 2022. 3, 4

[15] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 7

[16] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, 2014. 3, 4

[17] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951. 7

[18] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL. *ACM Transactions on Graphics*, 34(6):1–16, nov 2015. 7

[19] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. Mediapipe: A framework for building perception pipelines, 2019. 8

[20] Tewodros Legesse Munea, Yalew Zelalem Jembre, Halefom Tekle Weldegebriel, Longbiao Chen, Chenxi Huang, and Chenhui Yang. The progress of human pose estimation:

A survey and taxonomy of models applied in 2d human pose estimation. *IEEE Access*, 8:133330–133348, 2020. 1

[21] H. Pirsiavash and D. Ramanan. Detecting activities of daily living in first-person camera views. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, jun 2012. 3, 4

[22] Xavier Puig, Kevin Ra, Marko Boben, Jiaman Li, Tingwu Wang, Sanja Fidler, and Antonio Torralba. Virtualhome: Simulating household activities via programs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 3

[23] Siddharth Ravi, Pau Climent-Pérez, and Francisco Florez-Revuelta. A review on visual privacy preservation techniques for active and assisted living, 2021. 8

[24] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J. Guibas. Humor: 3d human motion model for robust pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11488–11499, October 2021. 6, 7

[25] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands. *ACM Transactions on Graphics*, 36(6):1–17, nov 2017. 7

[26] Tobias Scheck, Roman Seidel, and Gangolf Hirtz. Learning from theodore: A synthetic omnidirectional top-view indoor dataset for deep transfer learning. *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 932–941, 2020. 3

[27] Ahmed Rida Sekkat, Yohan Dupuis, Varun Ravi Kumar, Hazem Rashed, Senthil Yogamani, Pascal Vasseur, and Paul Honeine. SynWoodScape: Synthetic surround-view fisheye camera dataset for autonomous driving. *IEEE Robotics and Automation Letters*, 7(3):8502–8509, jul 2022. 3

[28] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. NTU RGB+D: A large scale dataset for 3d human activity analysis. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2016. 3, 4

[29] Leonid Sigal, Alexandru O Balan, and Michael J Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International journal of computer vision*, 87(1-2):4, 2010. 3

[30] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. 7

[31] M. Ozan Tezcan, Zhihao Duan, Mertcan Cokbas, Prakash Ishwar, and Janusz Konrad. Wepdtof: A dataset and benchmark algorithms for in-the-wild people detection and tracking from overhead fisheye cameras. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1381–1390, 2022. 2, 4

[32] Senthil Yogamani, Ciaran Hughes, Jonathan Horgan, Ganesh Sistu, Sumanth Chennupati, Michal Uricar, Stefan Milz, Martin Simon, Karl Amende, Christian Witt, Hazem Rashed, Sanjaya Nayak, Saquib Mansoor, Padraig Varley, Xavier Perrotton, Derek Odea, and Patrick Pérez. Woodscape: A multi-task, multi-camera fisheye dataset for autonomous driving. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9307–9317, 2019. 3

[33] Weiyu Zhang, Menglong Zhu, and Konstantinos G. Derpanis. From actemes to action: A strongly-supervised representation for detailed action understanding. In *2013 IEEE International Conference on Computer Vision*, pages 2248–2255, 2013. 3

[34] Z. Zhang. A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(11):1330–1334, 2000. 6