MISO: Monitoring Inactivity of Single Older Adults at Home using RGB-D Technology

June 18, 2024

Longfei Chen, Robert B. Fisher
School of Informatics
University of Edinburgh
longfei.chen@ed.ac.uk, rbf@inf.ed.ac.uk

Abstract

A new application for real-time monitoring of the lack of movement in older adults' own homes is proposed, aiming to support people's lives and independence in their later years. A lightweight camera monitoring system, based on an RGB-D camera and a compact computer processor, was developed and piloted in community homes to observe the daily behavior of older adults. Instances of body inactivity were detected in everyday scenarios anonymously and unobtrusively. These events can be explained at a higher level, such as a loss of consciousness or physiological deterioration. The accuracy of the inactivity monitoring system is assessed, and statistics of inactivity events related to the daily behavior of older adults are provided. The results demonstrate that our method achieves high accuracy in inactivity detection across various environments and camera views. It outperforms existing state-of-the-art vision-based models in challenging conditions like dim room lighting and TV flickering. However, the proposed method does require some ambient light to function effectively.

1 Introduction

The proportion of older adults living alone is increasing globally as the aging population grows [8, 22]. According to the Office for National Statistics in 2019, more than 3 million people over the age of 70 lived alone in the UK [11]. Even if they are living alone, more than 90% of older adults express a strong desire to remain independent – they predominantly prefer to continue living in their own homes rather than relocating to nursing homes or other care facilities [13, 16]. However, older adults living alone face more difficulties in daily life and have higher medical needs [8], [4]. Meanwhile, new or worsening symptoms related to chronic health problems or sensory impairments may not be noticed [13].

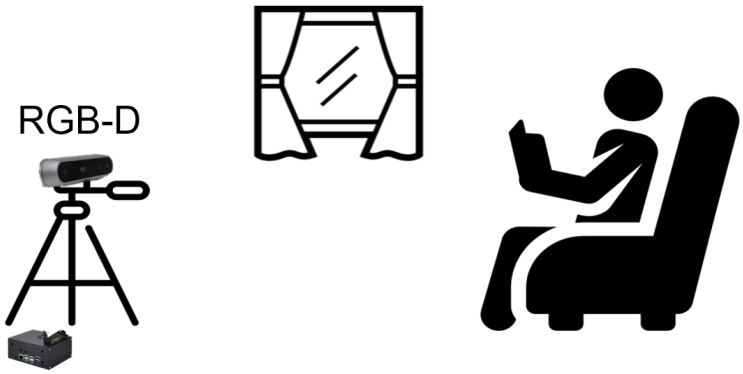


Figure 1: Real-time system for monitoring older adults in common home scenarios (inactivity monitoring of routine seating areas). Anonymous monitoring uses a compact system consisting of an RGB-D camera and a small computer processor. Anonymity is maintained by discarding captured images after processing.

Mobility problems, which encompass a spectrum of limitations including complete immobility (immobility syndrome) and reduced ability to move, can arise when someone experiences a prolonged period of reduced movement. These limitations are particularly prevalent in older adults and can contribute to functional decline, an increased risk of needing long-term care after hospitalization, and various medical complications. Examples include deep vein thrombosis, urinary incontinence, pressure sores, joint contractures, cardiac deconditioning, and muscle weakness [20]. As these complications often develop gradually, our focus lies on monitoring and analyzing behavior in older adults with long-term mobility limitations.

Smart technologies and AI solutions are accelerating the pace of change in healthcare in many areas [24], [28], [18], [31]. Unobtrusive sensors can work 24/7 in the long term, providing faster, more accurate analysis in a user-focused manner. In recent years, many studies have been conducted to monitor older adults in their homes using camera sensors. Stone *et al.* [28] detect falls using the person's vertical state and motion features. Using a depth imagery sensor (Kinect) can largely eliminate the interference of lighting and shadows for visual perception. A patient monitoring system is used to lessen the workload of the nurses: two RGB cameras were placed for monitoring patients, one for bed view and the other for room view, to detect bed occupancy, self-extubation, and falls [18]. The above systems have achieved good accuracy in detecting events, however, one of the major concerns with camera-based systems is *privacy* [7]. This issue must be addressed when convincing people to use such a camera-based system. Other challenges of real-world home camera monitoring include effectiveness

and adaptation. That means accurate detection of anomalous human events in backgrounds for different scenarios, moving objects and pets in the house, and changing lighting conditions, such as sunlight, low light, and TV lighting.

In this study, a new application of real-time monitoring of older adult people in home scenarios is presented. Specifically, the application detects the prolonged inactivity of a person sitting in a standard home location (e.g., on a favorite chair). A compact system consisting of an RGB-D camera and a small computer is deployed, which monitors the person at a specific location. The system is camera-based, but completely anonymous, with no internet connection, no image/video is saved, and only inactivity statistics text logs are kept. The abnormal inactivity events can be related to medical conditions, such as lost consciousness, or long-term decreased physical ability. An example is shown in Fig. 1.

The application was designed for homes where a single aging adult lives, so this is the test scenario that is evaluated. The results show that the system has good accuracy, sensitivity, and robustness in different environmental settings, enabling long-term anonymous monitoring of older adults in their own homes.

This paper introduces a new visual monitoring system that has the following advantages:

(1) A new camera system automatically tracks older adults' inactivity at home, even in difficult situations like dim light, pets around, or TV flickering. It works in real-time and keeps identities anonymous. This system is more accurate than other state-of-the-art vision algorithms for inactivity detection.

(2) The system was piloted in community homes to unobtrusively detect instances of body inactivity in everyday scenarios, revealing the behavior patterns of older adults, such as their daily routines and motion habits.

(3) This affordable, reliable, and zero-interacting system could be used to monitor older adults suffering from frailty and other long-term physical deterioration, as well as detect critical situations, making it a useful tool for everyday health monitoring.

2 Related Work

Many sensors have been investigated for detecting human behavior in indoor scenarios [30]. Individual inactivity is closely associated with hospitalization in older adults [36]. However, few works specifically focus on detecting the inactivity of the human body.

Wearable sensors offer high-accuracy human motion detection. For example, a belt-worn kinematic sensor demonstrated 100% recall in detecting human motionless events, such as several seconds of walking, standing, sitting, or lying, in a lab environment [37]. However, wearable sensors can be intrusive and may require users to wear them continuously, which might not always be acceptable for older adults. Ambient sensors are non-intrusive and capable of detecting various indoor human movements, including whole-body motions, limb movements and chest breathing movements, even in complete darkness. For instance,

Wi-Fi-based human motion sensing [19] can accurately recognize five typical human activities with a 96.6% accuracy; and subtle motions, such as simulated hand tremors, with 95.7% accuracy [5]. Passive Infrared (PIR) sensors [1] have achieved an accuracy of 93% in predicting human relative locations, including stationary individuals. Radar-based sensors [17] have achieved a classification accuracy of over 95% for four basic types of human motion. However, ambient sensors can be triggered by household pets, leading to an increase in data noise and a decrease in the overall predictability of human mobility [32]. Camera-based systems are widely used due to their cost-effectiveness and non-redundant nature in modern buildings. Cameras are less intrusive for long-term monitoring compared to wearable sensors and provide multifunctional and semantically explainable capabilities. They can distinguish between human and non-human movement and identify which body part is moving, and motion scale/speed; they can also efficiently identify and track humans, pets, and various objects. Moreover, the same camera-based system can be used to process different tasks, where ambient motion sensors can only detect some unspecified motion in the environment. For instance, Kinect-based body motion signals have shown moderate to excellent accuracy, with root mean square errors (RMSE) ranging from 20 mm to 89 mm [3]. In video-based pose estimation, mean absolute errors for gait analysis are as low as 0.02 seconds for temporal gait parameters and 0.04 meters for step lengths compared to motion capture technologies [27].

For human inactivity detection, accurately detecting the person is typically the initial step in monitoring behaviors. Xia *et al.* [34] proposed a model-based approach for indoor person detection using a Kinect sensor to capture a side view of the person. Initially, all 2D circular shapes are localized as head candidates, and then these candidate regions are fitted to a learned 3D human head shape. Most deep learning-based detectors are designed for oblique or front views, but Cho *et al.* [6] trained convolutional neural networks to segment heads using top-view depth data. A side view of the person can be more intrusive, as individuals are aware of the camera's presence. Depth-imaging-based methods are recognized for their superior performance and robustness in detecting humans across various poses, rotations, and lighting conditions. However, they often come with higher computational costs, which can be a limitation in applications requiring real-time processing, such as healthcare applications.

To monitor a person's inactivity while sitting in a preferred location, camera-based motion detection methods can be applied. There is a range of motion detection algorithms designed to address various real-world challenges [15]. One fundamental method is background subtraction [2], which creates a background model and identifies foreground objects by comparing the current frame to this model. Another technique, frame differencing [25], involves comparing the current frame with a reference frame to track the number of differing pixels. However, these methods are sensitive to noise and environmental changes, such as variations in lighting, the presence of shadows, and moving objects. Parametric-based methods, like the Gaussian Mixture model [26], tend to be more robust in the face of noise and artifacts [33]. Non-parametric methods [9], fitting a smooth probability density function to pixel values over a temporal window,

considering both self-similarity and similarity to neighboring pixels. This enhances robustness against camera jitter or minor background movements [15]. When compared to other traditional motion detection methods, non-parametric techniques have shown superior performance in eliminating minor background movements [21]. Common challenges for cameras in achieving accurate human inactivity detection include dealing with complex background noise, coping with fluctuating lighting conditions (including low environmental lighting and abrupt changes, such as TV light flickering), meeting high sensitivity requirements for detecting small body movements (e.g., finger movements), and distinguishing human movements from non-human movements (e.g., pets). The method proposed below can cope with these difficult issues, as demonstrated below.

3 Unobtrusive Monitoring Methodology

The goal is to detect inactivity accurately and sensitively within a home environment where the resident spends a significant amount of time. This environment may include areas for reading, resting, or watching television. An inactivity event is defined as the absence of movement in any body part for a duration exceeding one second[1]. This allows us to record motionless periods relevant for data analysis, particularly in creating long-term mobility profiles. Inactivity events initiate inactivity monitoring, which may reset upon the detection of motion, or may trigger an alert if inactivity persists for too long. Depth-based foreground extraction and color-based motion detection are used to detect if the person has stopped moving.

(i) Foreground detection

When monitoring people sitting in a room, the background can be complex, and the subjects can be in the region actively or inactively for long periods. A robust foreground detection method is applied. A background model is first constructed from a series of depth frames for the first few seconds when no one is present in the view. Each depth frame is smoothed by a median filter. Then, for each new depth frame, foreground pixels are detected by comparing them with the background model using a non-parametric method [10]. The non-parametric method fits a probability density function to the depth values at each pixel over a time window and detects changes. For every pixel at time t, the probability density function that this pixel has a depth value d_t is calculated relative to previous background depth values d_i at the same locations in n recent frames, as:

$$Pr(d_t) = \frac{1}{n} \sum_{i=1}^{n} K_\sigma(d_t - d_i), \tag{1}$$

where K is a 1d Gaussian with parameter σ. The pixel is considered as a foreground pixel if the probability is less than the threshold. Each pixel is compared with several pixels at the same location in several recent background frames to

[1]While ignoring tiny pauses (shorter than one second) between consecutive motions, as these might be inaccurate due to the limited frame rate (3 – 5 fps).

enhance the robustness of foreground detection to small noise or vibrations in the background (e.g., leaf motion [10] or depth estimation errors). To speed up the calculation on a small computer by avoiding the exponential computation for each pixel, the log of (1) is approximated by the quadratic function:

$$f(d_t) = -\log\sqrt{2\pi\sigma^2} - \frac{1}{2n\sigma^2}\sum_{i=1}^{n}(d_t - d_i)^2. \tag{2}$$

The foreground (matrix) is derived as

$$FG_t = D_t^{(f(d_t)<\rho)}. \tag{3}$$

In the implementation, ρ is set to -6.907, and σ is set to $\frac{M}{0.68\sqrt{2}}$ [10], where M is the mean of the absolute value of pixel differences from successive depth frames in the background model. Once the foreground pixels are detected, the traditional post-processing procedures (size filter, tracking, and open processing) are applied to the foreground mask to remove noisy areas. The two most recent foreground regions are also saved (as a binary mask for privacy) for reference in the following misdetection suppression process.

Background frame pixels are selectively updated for every loop, excluding foreground pixels, since some observed body parts may not have any motion for a long time and should not be updated to the background for inactivity detection purposes. The latest background frame is updated and added to the background model sequence, and the oldest frame is removed. For the latest background frame at time t, the foreground area is not updated, it is directly copied from the background frame at time $t-1$, whereas the previous background area is updated with the corresponding pixels from the current depth frame D_t, as:

$$BG_t = BG_{t-1}^{(FG_t)} + BG_{t-1}^{(\neg FG_t)} \cdot (1-\alpha) + D_t{}^{(\neg FG_t)} \cdot \alpha, \tag{4}$$

where α is the background update ratio.

(ii) Motion detection

Motion is detected using a color difference method, by subtracting the color of each pixel in the current frame from the color of the corresponding pixel in the previous frame, taking the absolute value, and comparing it to a threshold. This basic method is sensitive to detect small true motions of the human body (e.g., fingers) and insensitive to slow changes in natural light; however, it is susceptible to small image artifacts and noise. For example, sudden lighting changes when watching TV at home, especially at night when the room light is dim. Sudden changes in TV light reflected on the human body can be mistaken for movement.

Fig. 2 shows an example of color value changes of a pixel in a video that was recorded under the low light condition of a person in the living room watching TV: (a) In the absence of human motion under constant lighting, pixels at the same image location have very small differences in color values between frames; (b) In this sort of environment, when the TV light changes, peaks of the pixel

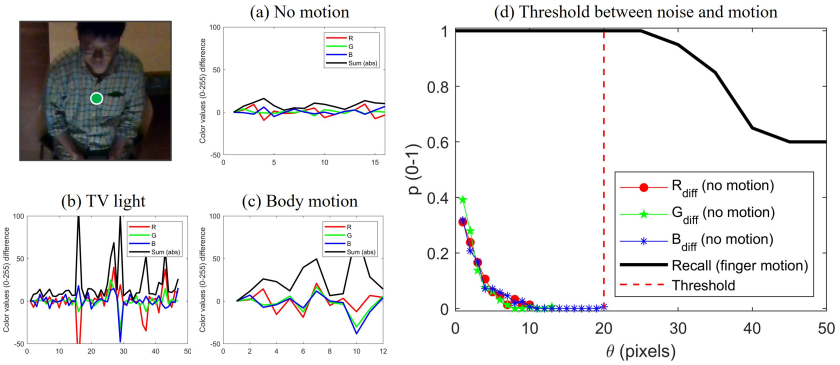


Figure 2: RGB value difference of one color pixel (selected on the human chest) in consecutive frames under the following conditions: (a) no human motion under constant low light; (b) same environment, without human motion but with TV light changes; (c) human movement only. (d) The threshold for distinguishing noise (a) from true body motion (c), see main text for details. (Scene Luma Rec. 601 [12] Y' is 20.5).

color difference in one color channel often appear and disappear in a short period; (c) If it was true human motion, the pixel difference usually showed large peaks in all color channels and lasted longer.

Therefore, to enhance robustness when classifying the changes as human motion, the change of intensities in all channels of the color pixels should be above a threshold, as:

$$mv = FG\ [\mathbf{R}_{\text{diff}} \geq \theta\ \cap\ \mathbf{B}_{\text{diff}} \geq \theta\ \cap\ \mathbf{G}_{\text{diff}} \geq \theta]\,, \tag{5}$$

where mv is the movement pixels and FG is the foreground region, $\mathbf{R}_{\text{diff}}$ is the absolute value of the difference of pixel values between consecutive frames.

The distribution of the pixel value differences in condition (a) in RGB channels over 300 frames was calculated. Since there is no movement, the difference value is considered as noises (ideally it should be 0). Moreover, in low lighting, the noise is usually larger than in good well-lit environments (i.e., upper bound of such noise). θ was increased to find the lower bound of losing any small movements (recall of 20 times finger movements, as illustrated in Fig. 6), and choose the proper threshold θ that can remove the noise while keeping the true small motions.

In the experiments, a single threshold θ for all RGB channels is set at 20 (pixel value range 0-255). This threshold corresponds to the upper bound of image noise typically observed in low-light conditions. Fig. 2 (d) shows the distribution of pixel value differences of the RGB channels in the no-motion condition over 300 frames from both low-light and well-lit environments. Since there's no movement in these conditions, the differences ideally should be 0, with the values of this distribution indicating the noise level. Note that noise

levels are higher in low-light conditions ($m = 2.23$, $\sigma = 2.29$) compared to well-lit environments ($m = 1.75$, $\sigma = 1.27$). To identify the threshold where motion detection starts to miss small movements, The detection of small finger movements (over 20 trials) was investigated. From this, the detection threshold was increased to $\theta_{max} = 25$, to effectively remove noise while preserving true small motion detections.

A temporal median filter with a window size of 5 frames was then used to mitigate the effects of TV flickering and other sudden illumination changes. The temporal length of TV flickering was calculated across 40 trials. The average temporal length was 1.6 frames ($\sigma = 0.74$). This suggests that sudden illumination changes are usually less than 0.5 seconds. Furthermore, considering the spatial and temporal size of the true motion of the human body (i.e., larger than a square centimeter, usually more than 0.5 seconds), a 2d spatial median filter (5×5 pixels) is applied to the foreground regions to further eliminate tiny isolated noise pixels.

(iii) Misdetection suppression

The motion regions detected using the depth-based method in (i) above are often disconnected when the person adopts some poses, such as reclining or lying on a couch. This is because, in these poses, the depth values of the background (e.g., couch) are very close to the depth values of the body parts, and most detection methods that calculate the difference between foreground and background depths may not be able to distinguish the small differences given inaccurate camera measurements ($< 2\%$ depth error at 2 meters). An example is shown in Fig. 3.

To deal with this incomplete foreground, the detected foreground is grown by referencing its depth and spatial locations on the image. An enlarged bounding box (10% of image width) is set around the foreground area to grow within. The mean (m) and standard deviation (σ) of foreground depth values are calculated as growth reference values. For any pixel adjacent to a foreground region in the enlarged bounding box, if its depth value d is close to the foreground mean m, as:

$$|d - m| < 2.8\sigma, \tag{6}$$

then this pixel is treated as a foreground inlier and will grow into the foreground. If the nearby background regions are not close to the foreground, there will be no region growth. If multiple foreground regions are detected, only the regions that are close to the most recent foregrounds in size and depth are grown.

Human/Pet detector: Large objects in the background, such as chairs or tables that may shift when a person leaves their seat, are sometimes detected during monitoring and remain in the foreground. These objects cannot be eliminated through selective background updates. To address this, an object detector (pre-trained YOLOv5, including humans and pets) is employed. Since object detectors are computationally expensive for real-time processing and may not always perform reliably with complex backgrounds and with various human poses, the detection occurs every 10 seconds. The system aggregates multiple observations and votes for human or non-human. After a minute, if the vote

(a) Before region growing

(b) After region growing

Figure 3: A person sitting on a couch. The grey area shows the foreground detected using the depth map (a) before the region growing and (b) after the region growing. Real human motion (red) is not detected in the incomplete foreground before the region growing.

result ($\geq 80\%$) indicates 'not human', the system updates the foreground region to become part of the background.

Pets, which are often present in the homes of single older adults, may move around humans even when the human is motionless, and this can introduce unwanted motion detections. When both humans and pets are present, the detector runs frame by frame, subtracting the detected pet region (bounding box) from the foreground region. Motion is then detected only within the remaining foreground region, effectively excluding pet motion, as shown in Fig. 4.

3.1 Behavior Statistics and Models

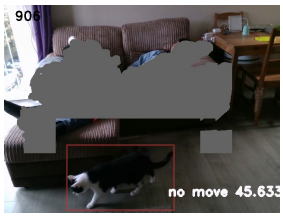
For inactivity detection, if no human motion is detected in the foreground, a timer will start counting the inactivity period in seconds. Once motion is detected in the foreground (and excluding the pet region), or if the foreground is recognized as non-human, the timer will reset. Periods of inactivity (≥ 1 second) and their occurrence times are saved in a log. Median, maximum, and minimum inactivity periods, by time of the day, are extracted from the log data.

Fitting distributions to the data was investigated to enable long-term comparison within the same subjects or among multiple subjects. The number of movement occurrences over a period of time (e.g., 1 minute, 1 hour) can be modeled with a Poisson distribution if one assumes that movements occur independently. Then the inactivity period between any consecutive movement events can be modeled with an exponential distribution as:

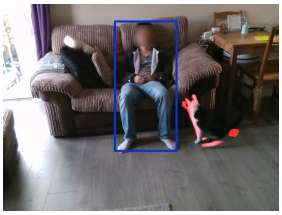
$$f(x;\lambda) = \lambda e^{(-\lambda x)}; x >= 0, \tag{7}$$

where the maximum likelihood estimate of the parameter λ is the inverse of the mean of the data as $\lambda_{mle} = 1/\bar{x}$.

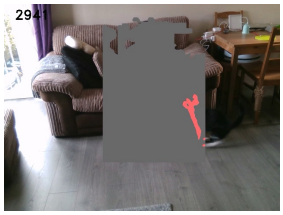
(a) Pet motion only



(b) Removed pet motion



(c) Failed to detect pet



(d) Failed to remove pet motion

Figure 4: Object detector helps remove pet motion. (a) Pet movement occurs while humans are inactive. (b) The pet is successfully detected, and the relative region is removed from the foreground, allowing the human inactivity count to continue. (c) The pet is not detected by the object detector. (d) Pet motion is not removed, resulting in the cessation of the human inactivity count.

4 Results

4.1 System Configuration

The RGB-D camera is an Intel RealSense D415 [23] with an ideal range of 0.5 meters to 3 meters. A USB cable connects the camera to the processor, a Jetson Nano [14], which requires a minimum of 4.75 volts and operates on as little as 5 watts. The overall dimensions of the system are less than 20 cm $\times$ 10 cm $\times$ 15 cm (excluding power cables). The camera captures color and depth imagery at a resolution of 640 $\times$ 480. The images are processed by the processor in real time, extracting the foreground and motion features. For the detection task, both depth and color images are used. 5 fps was observed when no one was present, and 3.5 fps was observed when someone was in view. The process is efficient and has no lag; inactivity events will be reported within 1 second (see details in Section 4.3: Temporal Sensitivity).

Privacy-preserving: The monitoring is camera-based, but the video frames are only temporally loaded in the RAM for on-the-fly analysis and then discarded. No images or videos are stored or transferred anywhere. Currently, the device has no internet connection, so there is no risk of hacking or people viewing the subject, not even the researchers (see the detailed discussion in Section 5). All that is stored about the recorded events are when someone is seen at

the monitored location, and how often they move. Detected events are stored anonymously in the form of text event logs. It is assumed that these logs will be analyzed on the home device, with the results available to visiting health workers.

4.2 Experimental Methodology

As an anonymous and privacy-preserving monitoring method, obtaining the ground truth for evaluation is challenging. Therefore, controlled experiments were conducted in a variety of environments to evaluate the accuracy of the system. Images were saved, and the ground truth was manually labeled. 116 videos were recorded across 12 indoor scenarios. This accuracy evaluation of the system included the following aspects: (I) Motion detection, (II) Spatial and temporal sensitivity, (III) Robustness in low light conditions and against TV light, and (IV) Robustness against pet motion. Two state-of-the-art motion detection methods are compared. One is a kinetic-based pose estimation method, ViTPose [35]; the other is an optical flow-based method, RAFT [29]. Both methods are pre-trained on transformer-based deep neural networks. The pre-trained coefficients of the networks were used and then tuned both methods to small motion sensitivity levels while keeping the upper bound of the noise threshold to remove as much noise as possible. The results are presented in Section 4.3.

Subsequently, our system was tested in a real-life deployment to capture the inactivity patterns of older adults. The monitoring system was deployed in four older adult households within the community, as an ethically approved study. In this home monitoring, no video or image data was saved, and no ground truth of the detected inactivity events was obtained. The results are presented in Section 4.4.

4.3 Accuracy Evaluation

(I) Motion detection. To evaluate the detection accuracy, video capture sessions with controlled motion and no-motion behaviors were conducted. A person started with continuous movement (about 30 seconds), then went to no movement at all (about 10 seconds), and then started continuous movement again (about 20 seconds). This process was repeated 70 times and 70 short videos were recorded in laboratories, offices, and homes, as illustrated in Fig. 5. The ground truth for these events, i.e., humans appear/disappear, movement starts/ends) is labeled manually as frame numbers in the videos.

Four types of errors are evaluated frame-wise. Human detection false positives (HuFP) and false negatives (HuFN), and motion detection false positives (MoFP) and false negatives (MoFN), are shown in Table 1. [2] To calculate the motion FN, it was assumed that during continuous movement periods, all frames are motion positive, although there are sometimes short pauses ($\ll$1s)

[2] From seventy video recordings with controlled motion/no-motion behaviors.

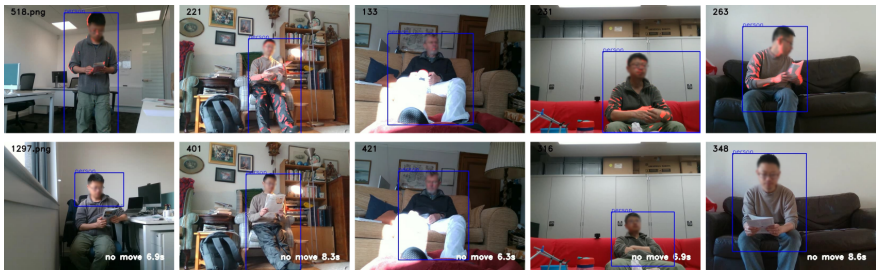


Figure 5: Examples of inactivity detection evaluation in different scenarios (front view in office, lab, and homes). The top row shows people in motion, and the bottom row shows people remaining inactive. Detected humans are marked with blue boxes, and areas with detected motion are highlighted in red pixels.

in transitions of body movement. During periods of complete motionlessness, all frames are labeled as motion-negative.

The results show that MoFP has a low error rate in all scenarios - from 0.48% to 0.00% when the temporal matching tolerance increases from ± 0 to ± 5 frames for the beginning/end moment of the completely no motion periods. This shows that, although a few frames of mismatch are tolerated at the beginning/end of the inactivity period, every no-motion frame is correctly detected in periods of inactivity, where the detection is insensitive to noise, i.e., the detection method can filter out all types of background noise or fake motion caused by lighting changes. MoFN is at 4.31% to 3.66% (± 0 to ± 5), indicating that several frames of true motion during continuous motion periods were missed. HuFP is 1.46% to 0.97% (± 0 to ± 5) in all frames and HuFN is 3.59% to 3.23% (± 0 to ± 5), which shows that the human detector (pre-trained YOLOv5) is more likely to miss real people than detecting background regions as humans in these experimental environments. In this experiment, there are many short (mostly < 1 second) periods of no detected motion during the intervals labeled in the ground-truth as having motion. These are the frame-wise MoFN errors. However, when considering the motion instance level that fuses motion across these brief gaps, all motion and non-motion instances are correctly detected.

(II) Spatial and temporal sensitivity. To evaluate the spatial sensitivity of our method, a subject performed five physical movements repeatedly and was monitored by a side-view camera 2.6 meters away. Fig. 6 illustrates the sensitivity of detecting human movements under different motion patterns. The movement ground truth was: body, wrist, finger, foot, each repeated 20 times (10 to the left, 10 to the right), and 40 head movements (20° each in four directions). The results in Table 2 demonstrate that our method has good sensitivity, with a true positive rate of 1.0 for detecting four motion types, including both large movements (such as sitting) and small movements (such as finger lifting). The true positive rate for detecting small head motion is 0.95, where all head motions were detected, however, few (2 out of 40) happened too quickly (2 frames or less) and were removed by the temporal filter. Whereas

Table 1: Error rates for inactivity detection. The error types are: **HuFP**: detected background region as human, which usually starts the inactivity time counter, resulting in false inactivity events. **HuFN**: missed real human detection, which causes the person to be updated into the background and reset the timer, resulting in missing inactivity events; **MoFP**: detected no motion as true motion, which falsely resets the timer, causing both missed and fewer true long-period inactivity events; **MoFN**: missed true human motion which causes a failure to reset the inactivity time counter and extends the count period, causing false long-period inactivity events. Because of potential mislabeling in the ground truth, a given number of frames of mismatch Tolerance is allowed in the bottom three rows.

Tolerance (frames)	HuFP ($10e{-}2$)	HuFN ($10e{-}2$)	MoFP ($10e{-}3$)	MoFN ($10e{-}2$)
± 0	1.46	3.59	4.8	4.31
± 1	1.31	3.47	1.7	4.10
± 3	1.15	3.31	0	3.82
± 5	0.97	3.23	0	3.66

Table 2: Recall (R) and Precision (P) Under Different Body Motion Patterns (Motion Detection Accuracy)

Lab ($Y'107$)	ViTPose[35]		RAFT[29]		MISO	
	R↑	P↑	R↑	P↑	R↑	P↑
body	**20/20**	**20/20**	**20/20**	**20/20**	**20/20**	**20/20**
head	25/40	25/25	**40/40**	**40/40**	38/40	38/38
wrist	**20/20**	**20/20**	**20/20**	**20/20**	**20/20**	**20/20**
finger	0/20	0/0	19/20	19/19	**20/20**	**20/20**
foot	**20/20**	**20/20**	**20/20**	**20/20**	**20/20**	**20/20**

the ViTPose pose-based model [35] missed all of the finger motions and 25% of the head motions since it only estimated the main body joints, and was not as fine-grained as the RAFT optical-flow-based method [29] nor our proposed method.

To evaluate the temporal sensitivity, 30 video clips of inactivity events ($\geq 5s$) were collected in a well-lit laboratory, manually noted the beginning and end moment of each event as ground truth, and then compared these with the detected log. The temporal accuracy of the inactivity detection result is ± 1.4 frames (at 3 to 5 fps), which indicates that the inactivity events will be reported promptly within 1 second.

(III) Robustness in low light conditions and against TV light flicker. Motion detection assessments were then performed under varying lighting conditions. The cameras recorded the subjects' activities with an oblique view at 1.5 meters, as illustrated in Fig. 7. The detection performance is then evaluated by the number of motions missed given 20 ground-truth movements under

Figure 6: Examples of spatial resolution of human motion detection under good lighting conditions in the laboratory (side view, Luma Y' 107). Five body movements were performed, 20 repetitions each (the head 40 repetitions for 20° rotation). Recall rates: Head (1.0), Body (1.0), Wrist (1.0), Fingers (1.0), Feet (1.0).

each lighting condition. The lighting condition has 5 levels from high to low, as daylight (Rec. 601 Luma Y' 97), night light (Y' 75), low light (Y' 36), dim light (Y' 24), and dark (Y' 4). The results (Table 3) show that under the first four lighting conditions (Y' 24 to 97), 1 or 2 out of the 20 movements were not detected (error rate 5% to 10%); whereas in the darkest environment ($Y'4$), the misdetection rate of true motions was 40% (8 out of 20 events were not detected). This indicates the limitation of our detection method under extremely low-lighting environments.

As mentioned previously, watching TV is a common activity for older adults, especially in dimly lit rooms, where the light from the TV can easily scatter colors onto people to affect the chromatic-based detection methods. To test the robustness of the motion detection against changes in TV lighting, an experiment was conducted under different room lights and at different distances from the TV. Examples are shown in Fig. 8. Subjects sat at 0.7 meters and 1.5 meters from the TV under 4 different room lighting conditions and then changed the TV light by changing the TV channel, 20 times at each environment setting

Table 3: Recall (R) and Precision (P) Under Different Lighting Conditions (Motion Detection Accuracy)

Room Lighting	ViTPose[35]		RAFT[29]		MISO	
	R↑	P↑	R↑	P↑	R↑	P↑
day (*Y*'97)	16/20	16/17	16/20	16/17	**19/20**	**19/19**
night (*Y*'75)	**19/20**	**19/21**	17/20	17/17	18/20	18/18
low (*Y*'36)	20/20	20/200*	19/20	19/39	**19/20**	**19/19**
dim (*Y*'24)	nan^		15/20	15/210*	**18/20**	**18/18**
dark (*Y*'4)	nan^		nan^		**12/20**	**12/12**

^ : Excessive noise led to many false positives.
* : Approximated by the average ratio of the number of FP instances to each TP instance.

Table 4: Human Motion False Positive(FP) Under TV Light Flickering Conditions

Lighting	TV Dis.	ViTPose[35] FP↓	RAFT[29] FP↓	MISO FP↓
Y'90	0.7m	**0/20**	3/20	**0/20**
	1.5m	**0/20**	5/20	**0/20**
Y'41	0.7m	3/20	6/20	**0/20**
	1.5m	**0/20**	8/20	**0/20**
Y'14	0.7m	14/20	nan^	**0/20**
	1.5m	19/20	nan^	**0/20**
Y'4	0.7m	15/20	nan^	**0/20**
	1.5m	nan^	nan^	nan*

^ : Excessive noise led to many false positives.
* : The human detector failed in the dark environment.

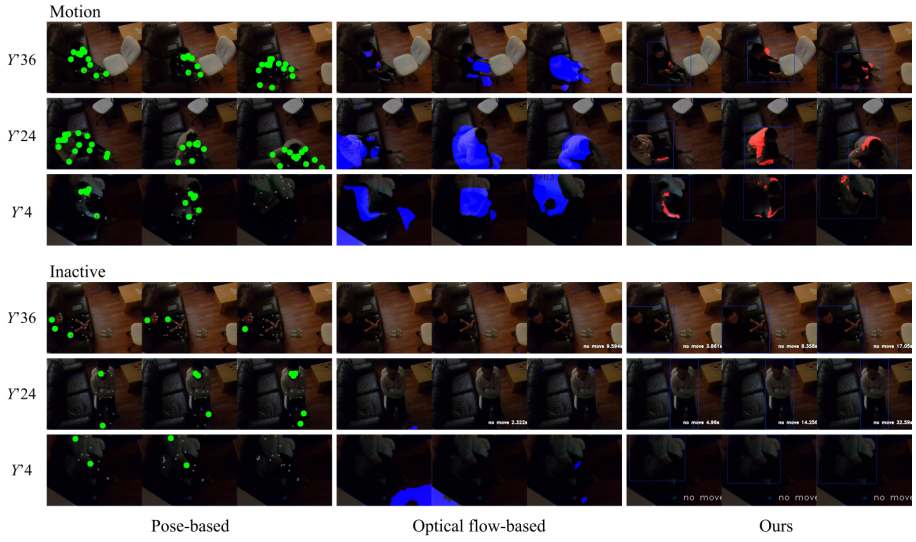


Figure 7: Performance comparison in low lighting conditions (top view). From left to right: VitPose[35], RAFT[29], and our method. Detected motion is shown in green, blue, and red, respectively, for the methods.

as ground. The results in Table 4 show that the false positive rates [3] are 0% in all settings, except in the darkest and furthest condition (1.5 m, Luma $Y'4$), where the person detector often failed to identify a person sitting in the scene. Although a few changes in TV global illumination were detected (see Fig. 8), they were not classified as true motion by our method, as these changes only occurred briefly (less than 0.5 seconds and thus were removed by the temporal filter). On the other hand, the pre-trained deep neural network models both performed worse in low-light and TV flickering conditions compared to our method (see Table 3 and Table 4). The optical flow-based method exhibited significant false positives in non-human regions under low-light conditions, whereas the pose-based method could detect the human joints but also detected substantial jitters, even when the person remained inactive, as illustrated in Fig. 7.

(IV) Robustness against pet motion. To test our detection method against pet motion, two subjects participated in three inactivity trials in a home scenario with a cat present (see Fig. 4). Each inactivity trial lasted approximately 10 minutes, resulting in a total inactivity duration of about 30 minutes (10,898 frames). The object detector runs frame by frame to detect the cat's region and remove it from the foreground.

The recall rate[4] for the true human inactivity duration, is presented in Table

[3]Detection accuracy is represented by the false positive rate. False positives occur when the TV light reflected onto the body is classified as motion.

[4]The rate is calculated as the ratio of recalled inactivity seconds to the total inactivity seconds.

5. Without removing the pet region, an average of 63.7% of the true human inactivity duration was recalled. When the pet moved around, it introduced random motions that significantly affected the accuracy of human inactivity detection. With the removal of the pet region during inactivity detection, the recall rate for the true human inactivity duration improved to 82.4% for our method. For comparison, the flow-based method was easily affected by both pet and background movement, whereas pose-based methods detected only human joints and achieved comparable performance with our method. The primary reason for this limitation was the pre-trained object detector's failure to detect the pet under certain conditions, such as specific poses or when the pet was moving quickly (resulting in motion blur), as illustrated in Fig. 4 (c) and (d). Enhancing the pet detector could help address this problem, but this falls outside the scope of the current study.

Table 5: Average Recall (%) of Human Inactivity when Pets are Around

Trial	MISO (without pet removal)	MISO (with pet removal)	ViTPose[35]	RAFT[29]
T1	59.8%	74.6%	**94.30%**	52.60%
T2	69.4%	85.5%	**97.70%**	47.90%
T3	60.3%	**85.8%**	60.5%	68.10%

4.4 Field Study

The monitoring system was deployed in four households in the older adult community. Residents ranged in age from 65 to 80 years old, with two male older adults and two female older adults. For each participant's house, the system captured data over a period of approximately three days. For the inactivity detection, all cameras were in the living room on a tripod, two observed a couch, one observed a chair, and the last one viewed both a chair and a couch. In total, 23,393 log events of inactivity ($\geq$1s) were recorded.

Table 6 summarizes the inactivity statistics of four older adult participants on a chair or couch in the living room. A median of 2.0 to 2.9 seconds for periods of non-movement was detected across all participants, with the minimum 25% intervals of 1.0 to 1.5 seconds. When the maximum 25% periods of inactivity periods are considered, the duration varied between participants, ranging from 7.5 seconds to 15.6 seconds. This demonstrated that a person tends to remain completely still for only very short periods of time when they are awake. Fig. 9 shows the statistics.

Table 7 presents the overall percentages of inactivity periods in different time ranges among all participants, showing that more than 90% of the inactivity instances were under 10 seconds, and approximately 99% of the instances were under 30 seconds. On the other hand, several periods of inactivity over 100 seconds and one over 800 seconds were observed from Participant 1, who was napping on the couch at the time. Fig. 10 shows an example of monitoring

data for Participant 4, where the participant's presence was between 10 am and 22 pm, with a peak inactivity duration at 19 pm. Meanwhile, the exponential distribution fits all the inactivity data well (see also Fig. 11 for Participants 1-3).

The authors acknowledge that one cannot make strong claims on the basis of only 4 participants, but it is hoped that the similarity of the inactivity distributions in Figures 9, 10, and 11 across the four volunteers suggests that it is possible to discriminate between normal short periods of inactivity and more serious longer periods.

Table 6: Average Duration (seconds) of Inactivity for Four Older Adult Participants

	Median	Max25%	Min25%
P1	2.07	15.61	1.37
P2	2.90	13.76	1.26
P3	2.30	7.58	1.43
P4	2.40	9.31	1.12
Mean	2.42	11.57	1.30
SD	0.30	3.25	0.12

Table 7: Percentage of Inactivity Duration for Different Time Ranges

Time range (s)	[1,2)	[2,5)	[5,10)	[10,30)	[30,60)	[60,200)	[200,500)	500+
P1	44.74%	33.22%	11.60%	8.08%	1.75%	0.55%	0.03%	0.03%
P2	35.46%	35.28%	15.64%	12.33%	1.15%	0.13%	0.02%	0.00%
P3	44.66%	37.86%	11.65%	5.83%	0.00%	0.00%	0.00%	0.00%
P4	39.33%	38.80%	14.02%	7.45%	0.35%	0.04%	0.00%	0.00%
Mean	41.05%	36.29%	13.23%	8.42%	0.81%	0.18%	0.01%	0.01%
SD	3.49%	1.96%	1.52%	2.15%	0.61%	0.20%	0.01%	0.01%
Cumulative	41.05%	77.34%	90.57%	98.99%	99.80%	99.98%	99.99%	100%

5 Discussion

In this study, the MISO was proposed, a camera-based system for monitoring inactivity among single older adults in home environments during daily activities. The zero-interaction system offers advantages over the wearable, which requires constant wearing or charging. Additionally, compared to ambient sensors, our system is multi-functional. The same device can perform various tasks using different algorithms. Furthermore, it provides high-level features that are semantically meaningful. For instance, it can understand fine-grained motion (such as which part of the body is moving) and interpret environmental context and interactions.

For inactivity detection, non-parametric-based methods in depth maps and motion detection methods in RGB are good for environmental adaptivity. There is no need to retrain to adapt to new environments. The proposed method is fast enough to be used in a compact processor in real-time as well. The system was tested in different scenarios, and TV conditions at different distances, and the results remain constant in indoor scenarios for motion detection. The results show that compared to SOTA pre-trained models, our method excelled in accurately detecting small body motion, demonstrated robustness in low-light conditions, as well as resistance to environmental factors such as TV light flickering and the presence of pets.

The recorded anonymized data can reveal the activity characteristics of older adults, such as daily habits of body movement patterns while staying at their favorite home places. It offers real-time tracking, enabling timely detection of excessive inactivity events. Additionally, because the approach saves longer-term inactivity statistics, it supports the analysis of chronic mobility issues.

Currently, the real-time text data is captured and stored locally for privacy-preserving, without the need for an internet connection. This facilitates long-term mobility records, such as weekly or monthly inactivity distributions. The system can also be configured to provide local user reminders, for example, a blinking light on the device to encourage older adults to stay active. However, for emergency situations involving critical inactivity events like falls or loss of consciousness, future internet connectivity would be necessary to enable identification and localization of the user. This integration would necessitate careful consideration of infrastructure safety and privacy concerns.

The final decisions regarding how the monitoring data is used will depend on factors outside the scope of this paper, such as the target user's behavior context, medical conditions, stakeholders, and relevant healthcare policies.

6 Conclusion

A system for inactivity detection in older adult residents' homes using an RGB-D camera and a small computer processor was presented. Collecting several days of data from each local household characterized the device's performance under real-home conditions. The method was tested in different living environments and various lighting conditions. Data processing for analysis is carried out in real-time. The system runs the inactivity detection task at 3–5 frames per second. The devices are small, data is anonymous, unobtrusive, and low-cost, which can be distributed well at homes for long-term use. A lack of motion is unlikely to be missed (a 0% false positive rate with ± 3 frames temporal tolerance), and a 3% false negative rate of missing true body motion on controlled short-term experiments. True lack of motion is likely to be long-term and this will be noticed.

The main limitation of the described method is its performance in extremely dark environments, where the object detector often fails to detect a person or a pet. Additionally, the study faces constraints related to the small dataset and

the lack of ground truth for real-life data.

Future work should focus on enhancing detection methods, capturing long-term personalized behavioral profiles of individual older adults to identify slowly deteriorating conditions, and designing decision rules based on historical patterns to provide warnings about potentially dangerous medical situations.

7 Acknowledgments

This research was funded by the Legal & General Group (research grant to establish the independent Advanced Care Research Centre at the University of Edinburgh). The funder had no role in the conduct of the study, interpretation or the decision to submit for publication. The views expressed are those of the authors and not necessarily those of Legal & General. Approval for the experiments was granted by the School of Informatics Ethics Committee.

References

[1] Jack Andrews, Meghana Kowsika, Asad Vakil, and Jia Li. 2020. A motion induced passive infrared (PIR) sensor for stationary human occupancy detection. In *2020 IEEE/ION position, location and navigation symposium (PLANS)*. IEEE, 1295–1304.

[2] Yannick Benezeth, Pierre-Marc Jodoin, Bruno Emile, Hélène Laurent, and Christophe Rosenberger. 2010. Comparative study of background subtraction algorithms. *Journal of Electronic Imaging* 19, 3 (2010), 033003.

[3] Johannes Bertram, Theresa Krüger, Hanna Marie Röhling, Ante Jelusic, Sebastian Mansow-Model, Roman Schniepp, Max Wuehr, and Karen Otte. 2023. Accuracy and repeatability of the Microsoft Azure Kinect for clinical measurement of motor function. *Plos one* 18, 1 (2023), e0279697.

[4] John T Cacioppo and Stephanie Cacioppo. 2014. Older adults reporting social isolation or loneliness show poorer cognitive function 4 years later. *Evidence-based nursing* 17, 2 (2014), 59–60.

[5] Hui-Hsin Chen, Chi-Lun Lin, and Chun-Hsiang Chang. 2023. WiFi-Based Detection of Human Subtle Motion for Health Applications. *Bioengineering* 10, 2 (2023), 228.

[6] Sung In Cho and Suk-Ju Kang. 2018. Real-time people counting system for customer movement analysis. *IEEE Access* 6 (2018), 55264–55272.

[7] George Demiris, Debra Parker Oliver, Jarod Giger, Marjorie Skubic, and Marilyn Rantz. 2009. Older adults' privacy considerations for vision based recognition methods of eldercare applications. *Technology and Health Care* 17, 1 (2009), 41–48.

[8] Kathryn Dreyer, Adam Steventon, Rebecca Fisher, and Sarah R Deeny. 2018. The association between living alone and health care utilisation in older adults: a retrospective cohort study of electronic health records from a London general practice. *BMC geriatrics* 18, 1 (2018), 1–7.

[9] Ahmed Elgammal, Ramani Duraiswami, David Harwood, and Larry S Davis. 2002. Background and foreground modeling using nonparametric kernel density estimation for visual surveillance. *Proc. IEEE* 90, 7 (2002), 1151–1163.

[10] Ahmed Elgammal, David Harwood, and Larry Davis. 2000. Non-parametric model for background subtraction. In *European conference on computer vision*. Springer, 751–767.

[11] Office for National Statistics. 2020. *People living alone aged 65 years old and over, by specific age group and sex, UK, 1996 to 2019*. Retrieved March, 2022 from `https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/families/adhocs/11446peoplelivingaloneaged65yearsoldandoverbyspecificagegroupandsexuk1996to2019`

[12] ITU-R. 2008. *BT.601*. Retrieved Nov, 2023 from `https://www.itu.int/rec/R-REC-BT.601/`

[13] Ausubel Jacob. 2020. *Older people are more likely to live alone in the US than elsewhere in the world*. Retrieved March, 2022 from `https://www.pewresearch.org/fact-tank/2020/03/10/older-people-are-more-likely-to-live-alone-in-the-u-s-than-elsewhere-in-the-world/`

[14] Jetson. 2022. *Jetson Nano*. Retrieved Dec, 2022 from `https://developer.nvidia.com/embedded/buy/jetson-nano`

[15] Pierre-Marc Jodoin, Sébastien Pierard, Yi Wang, and Marc Van Droogenbroeck. 2014. Overview and benchmarking of motion detection methods. (2014).

[16] S Juntapim. 2020. Mental Health Problem in Older Adults Living Alone: Challenges Issues. *Am J Nurs Stud* 1, 1 (2020), 1007.

[17] Sung-wook Kang, Min-ho Jang, and Seongwook Lee. 2021. Identification of human motion using radar sensor in an indoor environment. *Sensors* 21, 7 (2021), 2305.

[18] Panachit Kittipanya-Ngam, Ong Soh Guat, and Eng How Lung. 2012. Computer vision applications for patients monitoring system. In *2012 15th International Conference on Information Fusion*. IEEE, 2201–2208.

[19] Heju Li, Xin He, Xukai Chen, Yinyin Fang, and Qun Fang. 2019. Wi-motion: A robust human activity recognition using WiFi signals. *IEEE Access* 7 (2019), 153287–153299.

[20] Carlos Guido Musso, José Ricardo Jauregui, Juan Florencio Macías-Núñez, and Adrian Covic. 2020. *Frailty and Kidney Disease: A Practical Guide to Clinical Management*. Springer Nature.

[21] Manoranjan Paul, Shah ME Haque, and Subrata Chakraborty. 2013. Human detection in surveillance videos and its applications-a review. *EURASIP Journal on Advances in Signal Processing* 2013, 1 (2013), 1–16.

[22] Sabine Pleschberger, Elisabeth Reitinger, Birgit Trukeschitz, and Paulina Wosko. 2019. Older people living alone (OPLA)–non-kin-carers' support towards the end of life: qualitative longitudinal study protocol. *BMC geriatrics* 19, 1 (2019), 1–8.

[23] Intel RealSense. 2022. *Intel RealSense D415*. Retrieved Dec, 2022 from `https://store.intelrealsense.com/buy-intel-realsense-depth-camera-d415.html`

[24] Sandeep Reddy, John Fox, and Maulik P Purohit. 2019. Artificial intelligence-enabled healthcare delivery. *Journal of the Royal Society of Medicine* 112, 1 (2019), 22–28.

[25] Nishu Singla. 2014. Motion detection based on frame difference method. *International Journal of Information & Computation Technology* 4, 15 (2014), 1559–1565.

[26] Chris Stauffer and W. Eric L. Grimson. 2000. Learning patterns of activity using real-time tracking. *IEEE Transactions on pattern analysis and machine intelligence* 22, 8 (2000), 747–757.

[27] Jan Stenum, Cristina Rossi, and Ryan T Roemmich. 2021. Two-dimensional video-based analysis of human gait using pose estimation. *PLoS computational biology* 17, 4 (2021), e1008935.

[28] Erik E Stone and Marjorie Skubic. 2014. Fall detection in homes of older adults using the Microsoft Kinect. *IEEE journal of biomedical and health informatics* 19, 1 (2014), 290–301.

[29] Zachary Teed and Jia Deng. 2020. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*. Springer, 402–419.

[30] Dipti Trivedi and Venkataramana Badarla. 2020. Occupancy detection systems for indoor environments: A survey of approaches and methods. *Indoor and Built Environment* 29, 8 (2020), 1053–1069.

[31] Toshifumi Tsukiyama. 2015. In-home health monitoring system for solitary elderly. *Procedia Computer Science* 63 (2015), 229–235.

[32] Tinghui Wang, Diane J Cook, and Thomas R Fischer. 2022. The Indoor Predictability of Human Mobility: Estimating Mobility With Smart Home Sensors. *IEEE Transactions on Emerging Topics in Computing* 11, 1 (2022), 182–193.

[33] C Wren. 1995. Real-time tracking of the human body. *Photonics East, SPIE* 2615 (1995).

[34] Lu Xia, Chia-Chih Chen, and Jake K Aggarwal. 2011. Human detection using depth information by kinect. In *CVPR 2011 workshops*. IEEE, 15–22.

[35] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. 2022. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in Neural Information Processing Systems* 35 (2022), 38571–38584.

[36] Marek Zak, Magdalena Wasik, Tomasz Sikorski, Krzysztof Aleksandrowicz, Renata Miszczuk, Daniel Courteix, Frederic Dutheil, Aneta Januszko-Szakiel, and Waldemar Brola. 2023. Rehabilitation in Older Adults Affected by Immobility Syndrome, Aided by Virtual Reality Technology: A Narrative Review. *Journal of Clinical Medicine* 12, 17 (2023), 5675.

[37] Shumei Zhang, Paul McCullagh, Chris Nugent, and Huiru Zheng. 2009. A theoretic algorithm for fall and motionless detection. In *2009 3rd International Conference on Pervasive Computing Technologies for Healthcare*. IEEE, 1–6.

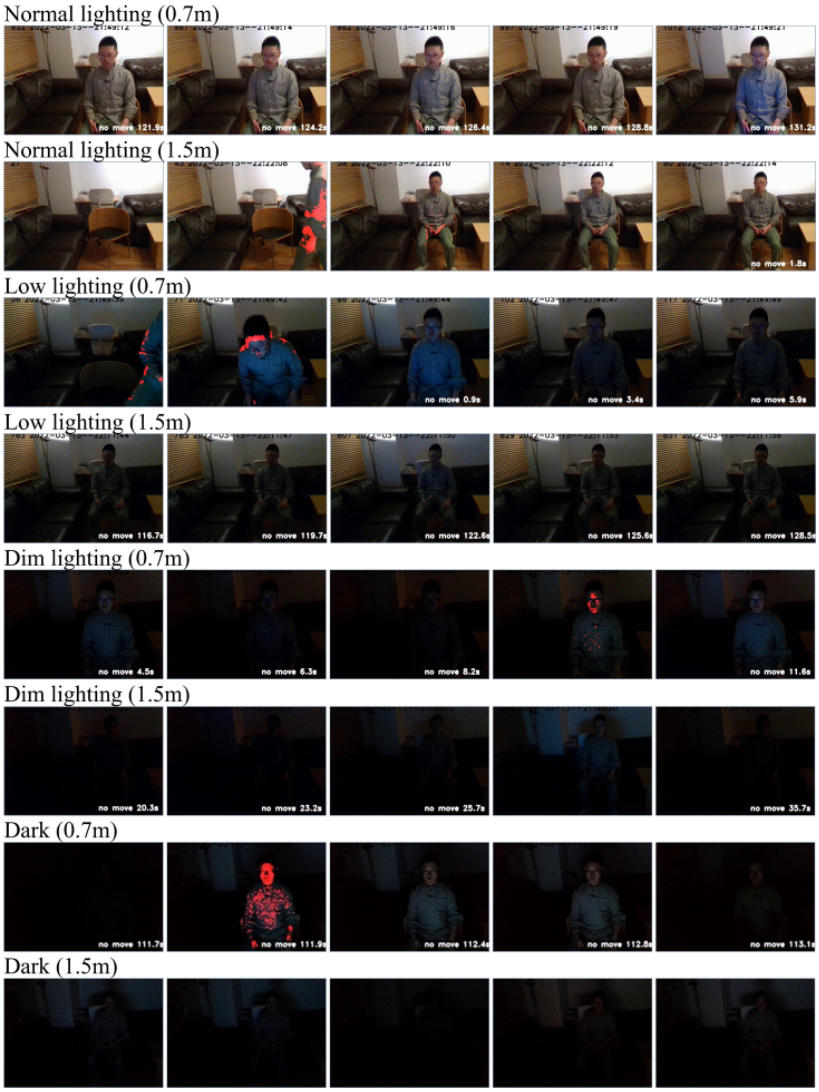
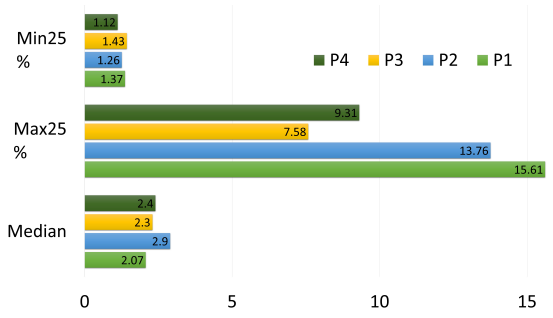
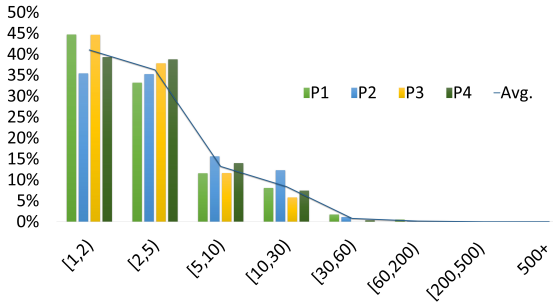


Figure 8: Examples showing motion detection is robust to TV light changes at different person-to-TV distances (0.7m or 1.5m) and four different lighting conditions (top-to-bottom, Luma Y' are 90, 90, 41, 41, 14, 14, 4, 4). False positive rate (top-to-bottom): 0/20, 0/20, 0/20, 0/20, 0/20, 0/20, 0/20, nan.

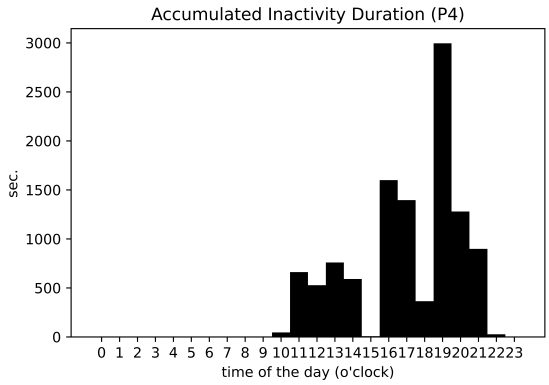


(a)

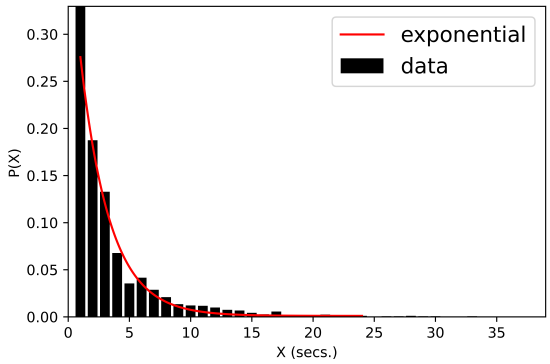


(b)

Figure 9: Inactivity detection statistics for four participants. (a) Average inactivity time of each participant (Median, Max 25%, Min 25%). (b) Percentage of inactivity duration for each different time range (seconds).

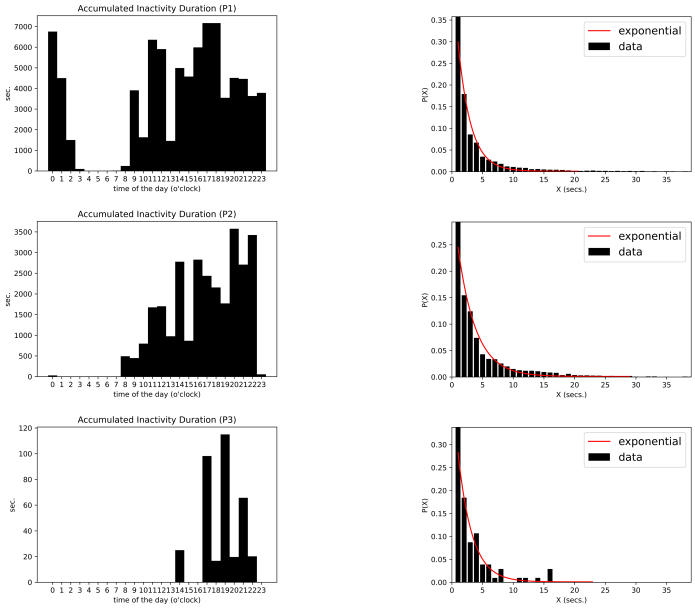


(a) Inactivity time of day



(b) Inactivity duration distribution

Figure 10: Example of inactivity statistics (Older Adult Participant 4).



(a) Inactivity time-of-day (b) Inactivity distribution

Figure 11: Inactivity statistics (Older Adult Participants 1-3).