Computer Vision-Based Descriptive Analytics of Seniors' Daily Activities for Long-Term Health Monitoring

Zelun Luo[1]* ALANZLUO@STANFORD.EDU
Jun-Ting Hsieh[1]* JUNTING@STANFORD.EDU
Niranjan Balachandar[1] NIRANJA9@STANFORD.EDU
Serena Yeung[1] SYYEUNG@CS.STANFORD.EDU
Guido Pusiol[1] GUIDO@CS.STANFORD.EDU
Jay Luxenberg[3] JLUXENBERG@ONLOK.ORG
Grace Li[3] GRACELI@ONLOK.ORG
Li-Jia Li[2] LIJIALI@CS.STANFORD.EDU
N. Lance Downing[2] LDOWNING@STANFORD.EDU
Arnold Milstein[2] AMILSTEIN@STANFORD.EDU
Li Fei-Fei[1] FEIFEILI@CS.STANFORD.EDU

[1] *Department of Computer Science, Stanford University, United States*
[2] *Department of Medicine, Stanford University, United States*
[3] *On Lok Inc., United States*

Abstract

Nations around the world face a rising demand for costly long-term care for seniors. Patterns in seniors' activities of daily living, such as sleeping, sitting, standing, walking, etc. can provide caregivers useful clues regarding seniors' health. As the older population continues to grow worldwide, it becomes more and more challenging for caregivers to monitor seniors' daily activities manually and continuously. To improve caregivers' ability to assist seniors, an automated system for monitoring and analyzing patterns in seniors activities of daily living has been long-awaited. A possible approach to implementing such a system involves wearable sensors, but this approach is intrusive in that it requires adherence by seniors, and is specific to a certain activity category. In this paper, using a dataset we collected from an assisted-living facility for seniors, we present a novel computer vision-based approach that leverages non-intrusive, privacy-compliant multimodal sensors and state-of-the-art computer vision techniques to continuously detect seniors' activities and provide the corresponding long-term descriptive analytics. This analytics includes both qualitative and quantitative descriptions of senior daily activity patterns, which can be further interpreted by caregivers. Our work is progress towards a smart senior home that uses computer vision to support caregivers in senior healthcare to help meet the challenges of an aging worldwide population.

1. Introduction

The U.S. is experiencing a shift in the age demographics of its population; the percentage of U.S. residents aged 65 or above is projected to rise from 13% in 2010 to over 20% in

* These authors contributed equally to this work.

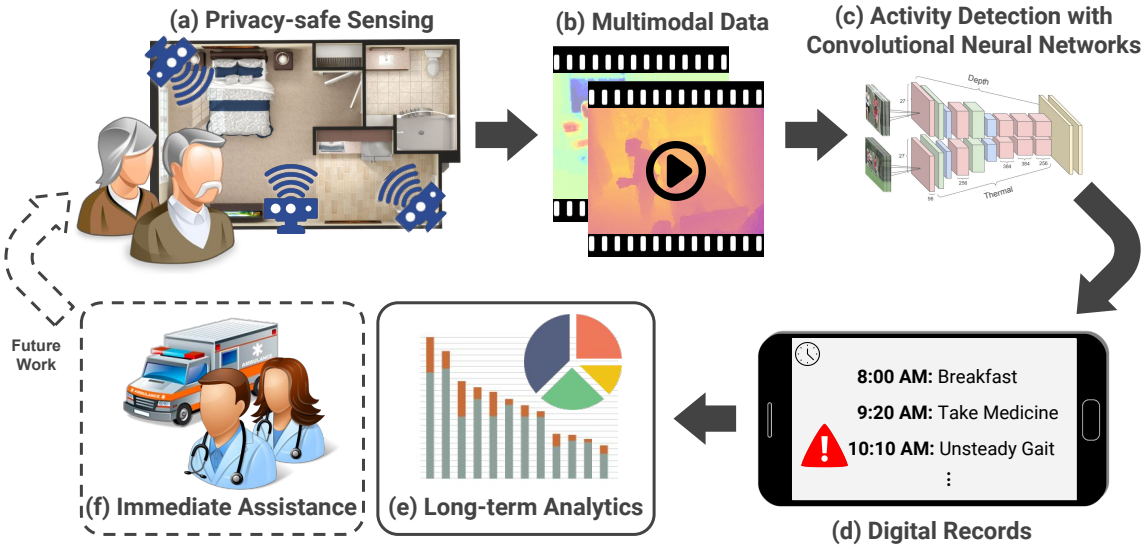


Figure 1: In this work, we propose a system that automatically provides long-term activity detection and analytics of seniors by leveraging the privacy-compliant multimodal sensors and state-of-the-art computer vision algorithms. We will extend our work to provide real-time alerts to caregivers when emergency events such as falls are detected (dashed line).

2030 (Ortman et al., 2014). The aging population has a significant societal impact, with the Congressional Budget office projecting Medicare spending to rise from 3.5% of the U.S. GDP in 2015 to 6.3% in 2040 (Hall, 2015). Additionally, in 2014, 1.4 million seniors in the U.S. resided in senior housing – including assisted living facilities. As this number continues to increase, the average number of potential caregivers per senior is expected to drop from 7 per senior in 2010 to 4 per senior by 2030 (Redfoot et al., 2013; Harris-Kojetin et al., 2016). This declining ratio challenges safe and compassionate access to quality elder care. Beyond the U.S., these demographic shifts and implications are also echoed worldwide (Bloom et al., 2011). Therefore, it is crucial and challenging to have effective long-term care for seniors across individual homes, assisted living facilities, and long-term care facilities.

While traditional care models have depended heavily on in-person care, there is a gaining recognition that types of remote monitoring may afford longer independent living for seniors on the cusp of needing some form of assisted living. Smart monitoring of key activities and health conditions may help prevent acute health declines and also create the necessary layer of safety for those at high risk of losing independence. Furthermore, automating the continuous monitoring of seniors' health-related activities and behaviors have significant potential to support senior home caregivers. One approach to implementing such a system is to use wearable IoT devices to track vital signs, activities, and accidents, such as heart rate, respiration rate, sleeping, and falls. However, this approach requires seniors to wear these devices throughout the day, which can be intrusive and inconvenient, and there has

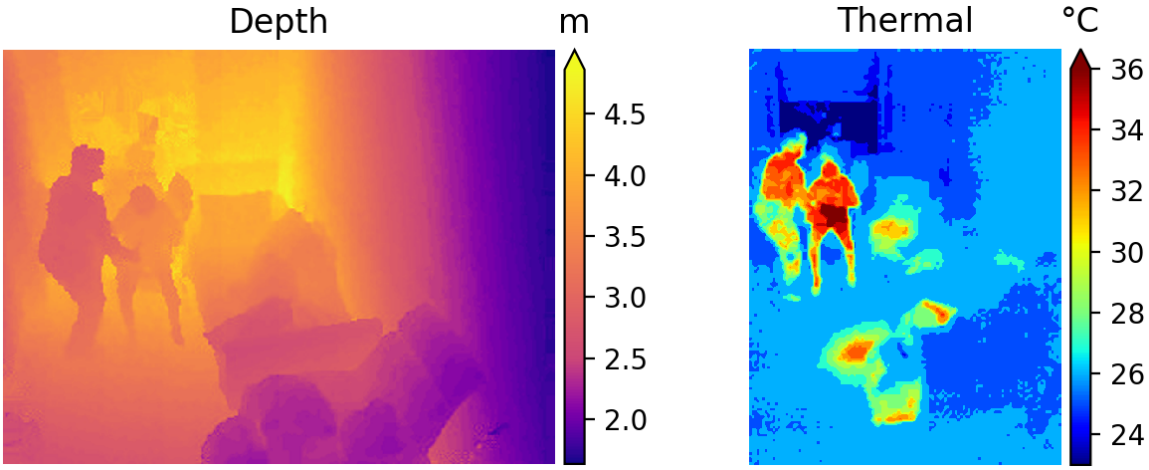


Figure 2: Sample images of depth modality (left) and thermal modality (right). Colormaps are applied for better visualization. Our framework utilizes both depth and thermal signals to detect and monitor senior daily activities. Both modalities are privacy-compliant and thus prevent identification of the person in the images.

been limited success in improving outcomes via these sensor types (González-Valenzuela et al., 2011; Lee and Carlisle, 2011; Park and Jayaraman, 2003). In addition, this approach requires dedicated sensors for different activities, which could be cumbersome and costly if we are interested in a wide range of activities.

Video-recording would provide a non-intrusive alternative to wearable sensors for automated monitoring of seniors' daily activities. Recent advances in computer vision have shown promise in human activity classification from video data (Karpathy et al., 2014; Simonyan and Zisserman, 2014; Tran et al., 2015; Carreira and Zisserman, 2017). If we could apply these techniques to classify clinically relevant activities accurately, an inexpensive, passive, continuous layer of safety and monitoring can be extended to the world's aging population. Thus far, these approaches have focused on detecting the occurrence of acute, short-term events such as falls (Zhang et al., 2015; Chua et al., 2015; Mastorakis and Makris, 2014; Yu et al., 2012; Rougier et al., 2011a,b; Luo et al., 2017). While these are important, long-term monitoring to detect abnormalities from daily senior activities can also provide useful information regarding seniors' health. Prior work for vision-based senior activity detection have various limitations, such as using privacy-violating RGB videos (Cheng et al., 2014; Xiang et al., 2015) or simulated data (Parajuli et al., 2012). Furthermore, none of these prior works provide long-term analytics of continuous senior activities. Thus, to our knowledge, an automated, non-intrusive, privacy-compliant vision-based system trained on real video data for continuous senior activity detection and long-term health monitoring has not been developed prior to our work.

In this work, we introduce an approach that uses computer vision to monitor and produce long-term descriptive analytics of seniors' health. We pilot our implementation in a senior's room in On Lok, San Francisco, an assisted-living facility for seniors. A schematic of our

system, along with future work, is shown in Figure 1. Firstly, we developed multimodal (explained in Figure 2), privacy-preserving, and easy-to-install sensors and deployed them in a senior home facility (Figure 1(a)). Secondly, we recorded long-term continuous video data and annotated the data using the two annotation interfaces we developed (as shown in Figure 1(b)). Next, we demonstrate the ability of our model to classify the clinically-relevant activities in real-time as well as to perform temporal activity detection in the long term using state-of-the-art activity classification and detection models (as shown in Figure 1(c)). We examine the occurrence of 6 types of activities: *sitting, standing, walking, sleeping, using bedside commode, getting assistance.* We choose these activities because they are *fundamental activities of daily living* (ADL) that span the majority of seniors' activity in the environment, and patterns in these activities can provide clinically-relevant information regarding a seniors' health (Phelan et al., 2004; Gangwisch et al., 2008; Buysse et al., 1991; Gómez-Cabello et al., 2012; Pitta et al., 2005; Van Kan et al., 2009; Studenski et al., 2011; Österberg et al., 1996; Charach et al., 2001; Vaizey et al., 1997). Finally, we aggregated these results to build daily activity timelines of seniors and calculated important quantitative and qualitative descriptive analytics that are clinically relevant to the health of seniors (as shown in Figure 1(d) and Figure 1(e)).

Here we demonstrate a real-world implementation and its efficacy of an automated, non-intrusive, privacy-compliant vision-based system for continuous senior activity detection and long-term health monitoring. We first show the ability of our activity classification model to learn and classify a short video clip into the clinically-relevant activities in real-time (as shown in Table 2). We then adapt the model to temporal activity detection task in long-term continuous videos (as shown in Table 3). Finally, we use the results of the temporal activity detection to build daily activity timelines of the senior and generate quantitative and qualitative descriptive analytics (as shown in Table 4, Figure 6, Figure 7, Figure 8, and Figure 9), all of which are clinically-relevant indicators of the seniors health.

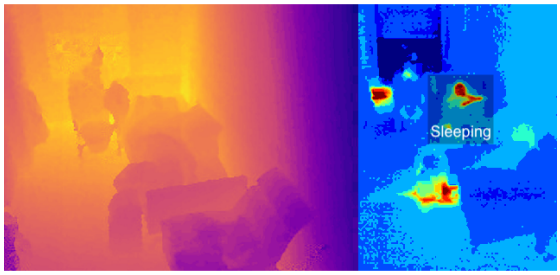
Technical Significance. Firstly, we proposed and implemented an automated, multi-modal, privacy-preserving, and non-intrusive sensing system that fits well in home care settings. Secondly, we collected and annotated a large-scale and realistic dataset in a senior care facility. Lastly, we proposed a Convolutional Neural Networks-based model that effectively detects a wide range of activities by fusing multimodal information and generates insightful analytics.

Clinical Significance. The system is helpful in objectively recording and analyzing long-term behaviors and capturing seniors' health decline. It could be utilized to better identify patients who need a medical intervention earlier and potentially could lead to better clinical outcomes.

2. Cohort

We conduct this study in an assisted living facility for seniors, after obtaining each participant's consent to deploy sensors in their studio apartments. The typical layout of these rooms is detailed in Figure 9(h). We use a multimodal visual sensing setup to identify important activities in privacy-preserving manner. At the current stage, we focus on the following activities: sitting, standing, walking, sleeping, getting assistance, and using the bedside commode.

(a) An example of a frame in which thermal data are more informative than depth data.



(b) An example of a frame in which depth data are more informative than thermal data.

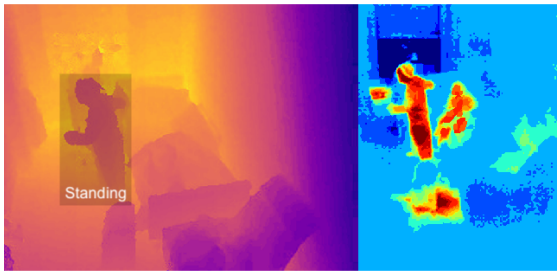


Figure 3: A comparison that shows depth and thermal data provide complementary information. There are various scenarios in which one of the modalities is able to capture information that the other is unable to capture. In (a), the senior is sleeping in the bed, but is covered by blankets so is not well-distinguished from the background by the depth sensor, whereas the thermal sensor is able to more clearly capture the senior in the bed though body temperature. In (b), the senior just got out of the bed and the thermal sensor captures the residual body heat on the bed, which may be problematic for the detector, whereas this noise is absent in the corresponding depth frame.

2.1. Multimodal Visual Sensing

While RGB cameras are standard for visual sensing in computer vision applications, they also capture highly identifiable data and can raise privacy concerns in home care settings. We investigate the use of a multimodal combination of privacy-preserving visual sensors for recognizing daily activities of seniors.

One type of privacy-preserving visual sensor is depth sensors. These sensors capture depth images where pixel values in the image indicate a calibrated distance between corresponding objects and the sensor. Over the past few years, depth based imaging has become a popular method for scene capture, in addition to traditional RGB-based image capture. While they do not capture appearance detail, depth images have several advantages over color images: they are invariant to different lighting conditions, color, and texture, and they contain rich 3D information about structures in a scene. Their lack of appearance detail further enables them to preserve privacy, and this has motivated their use in healthcare applications such as monitoring hand hygiene activity in hospitals (Haque et al., 2017).

Another type of privacy-preserving visual sensor is thermal sensors. In thermal images, pixel values indicate the temperature of corresponding objects. Since the human body is typically significantly warmer than the surrounding environment, this is particularly useful in the context of recognizing human activities. Similar to depth images, thermal images do not suffer in low light conditions and are invariant to color.

We use a multimodal combination of depth and thermal sensors in order to leverage the strengths of both. In particular, we develop a visual sensing setup (Figure 2) where an ASUS Xtion PRO depth sensor and FLIR Lepton 3 thermal sensor are mounted adjacent

to each other near the ceiling of a senior home room, with a 45° downward viewing angle. The depth sensor records images at 240×320 resolution and 30 fps, with distance range of 0.4 to 4.5 meters. The thermal sensor records images at 160×120 resolution and 8.8 fps, with temperature range -10 to 65°C.

Figure 3 illustrates the complementary information provided by depth and thermal sensors. For example, Figure 3(a) shows how a thermal sensor can capture human body information despite coverage by blankets, and Figure 3(b) shows how a depth sensor can indicate accurate volumetric occupancy of a person despite deceptive body heat signatures.

2.2. Dataset Collection and Annotation

We collected a dataset of multimodal depth and thermal video of a senior home room over the course of 1 month, using our visual sensing setup. Collecting an actionable multimodal video dataset for downstream training of recognition models presents several challenges that need to be addressed.

First, when dealing with multiple modalities of data stream, there is increased strain placed on data capture and storage resources. In our setup, depth images are relatively larger in size, and furthermore uninformative during activities such as sleeping, as illustrated in Figure 3. We therefore use real-time, background subtraction-based motion detection in the depth images to determine when to capture and store depth data. In particular, once motion is detected, the depth sensor will start recording for a minimum amount of time. While our depth data is thus filtered, we capture thermal data continuously due to its smaller size and footprint.

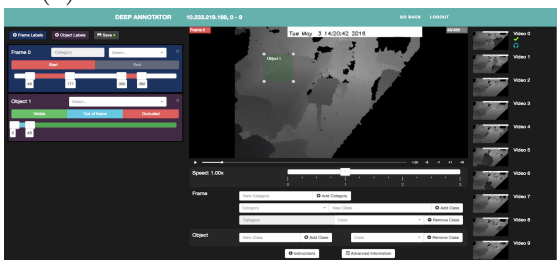
A second issue is alignment of the captured depth and thermal images, both temporally and spatially. Since the depth and thermal sensors have different frame rates, there is no natural alignment of the images. Instead, we use nearest neighbor matching in time to temporally align the two modalities. For spatial alignment, a calibration and post-processing transformation of the depth images can allow more precise alignment with the thermal images. However, we found the rough alignment of the physically nearby sensors to be sufficient for our needs, without requiring more computationally expensive post-processing.

In total, we collected and annotated 7 days of training data and 3 days of testing data. In addition, we collected 14 days of continuous data for the generation of long-term descriptive analytics. We annotated the data with 6 types of daily living activities expected to cover the majority of the day of a senior residing in an assisted living facility, plus a background class when the senior is out of the room. These activities are: sleeping, sitting, standing, walking, using bedside commode, and getting assistance from a caregiver. In order to annotate the data, we developed two types of annotation tools for efficient labeling (Figure 4): a web interface which is more intuitive for clinicians and supports remote access, and a terminal interface which supports faster annotation for annotators familiar with the terminal. The data was annotated and reviewed by a total of 4 students with the supervision of our clinicians. Statistics for the resulting dataset are given in Table 1.

3. Methods

The goal of the activity detection model is to predict an activity timeline over an extended period given a continuous multimodal video. We approach this task by dividing the video

(a) Our web annotation interface.



(b) Our terminal annotation interface.

Figure 4: We developed two annotation interfaces for different users and needs. The web interface (as shown in (a)) is more user-friendly for people without a programming background and can access the data remotely. The terminal interface (as shown in (b)) functions locally and is ideal for rapid annotation for people with a programming background.

Table 1: Statistics of the classification dataset. Each frame in a video will be classified into one of these 7 activity classes. The dataset is highly imbalanced in terms of the frequencies and durations of different activities, which makes our problem very challenging.

Activity	Clips	Frames	Frames per clip
Sitting	589	41,042	69.7
Standing	365	22,946	62.9
Walking	252	7,099	28.2
Sleeping	58	10,585	182.5
Getting Assistance	231	43,282	187.4
Using Bedside Commode	124	6,100	85.9
Background	124	19,887	160.4
Total	1,690	150,941	89.3

into different segments, each corresponding to one activity. Equivalently, we need to assign an activity label to each frame in the video. We first train a model that predicts an activity label given a short video clip (classification), then apply the classification model on the long continuous video in a sliding-window fashion for activity detection.

This section describes our pipeline for activity classification of short video clips, temporal activity detection of a long continuous video, and improving the performance by combining the information of depth and thermal modalities.

3.1. Activity Classification

We first formally define the activity classification problem: consider a K-way classification problem with a training set $\mathcal{D}_{train} = \{(x_i, y_i)\}_{i=1}^{|\mathcal{D}_{train}|}$, where $x_i \in \mathbb{R}^{T\times H\times W\times C}$ is a video of an activity, $y_i \in [1, K]$ is an integer denoting the class label, and $|\cdot|$ denotes the set cardinality. Note that in our scenario, $K = 7$ classes and $C = 1$ for both depth and thermal modalities.

Our objective is to learn a model f that minimizes the empirical risk on unseen test data $\mathcal{D}_{test} = \{(x_i, y_i)\}_{i=1}^{|\mathcal{D}_{test}|}$:

$$f^* = \arg\min_{f\in\mathcal{F}} \frac{1}{|\mathcal{D}_{test}|} \sum_{(x_i,y_i)\in\mathcal{D}_{test}} \mathbb{E}[\ell(f(x_i), y_i)], \tag{1}$$

where $\mathcal{F}$ is a class of functions and ℓ represents the loss function. Here $f(x_i) \in \mathbb{R}^K$ is the softmax scores, i.e. the probability of each category.

In practice, the empirical risk is often optimized on the seen training data under a specific loss. For video classification, a popular choice for ℓ is the softmax cross entropy loss:

$$\ell_c(f(x_i), y_i) = -\sum_{j=1}^{K} \mathbb{1}(y_i = j) \log f_j(x_i), \tag{2}$$

where K is the number of categories; $\mathbb{1}$ is the indicator function which equals 1 when the condition is true and 0 otherwise; $f_j(x_i)$ is the probability score of class j.

At test time, the output prediction of an input video x_i is

$$\hat{y}_i = \arg\max_j f_j(x_i).$$

3.2. Temporal Activity Detection

Once we have a trained model f, we can apply it to long continuous videos for temporal activity detection. Given a long continuous video X, we first segment it into L overlapping clips of T frames: $X = \{x_i\}_{i=1}^{L}$, where $x_i \in \mathbb{R}^{T\times H\times W\times C}$. Each video clip is independently fed into the model to obtain the clip-level softmax scores $f(x_i)$.

Following the procedure in Montes et al. (2016), we apply a smoothing window W to smooth the scores through time, i.e.,

$$\tilde{f}(x_i) = \frac{1}{W} \sum_{j=i-W/2}^{i+W/2} f(x_j).$$

Next, we calculate the frame-level scores by averaging the softmax scores of the clips that overlap the corresponding frame. The frame-level scores give us the final prediction output of each frame, and they are used to calculate the frame-level mean Average Precision (Yeung et al., 2017) for evaluation. We focus on frame-level evaluation for detection in long continuous videos since obtaining accurate frame-level labeling and duration statistics is desired for long-term senior health monitoring.

3.3. Multimodal Recognition

With both depth and thermal modalities available, we can utilize a network fusion scheme to combine the two information streams to yield better performance. Given a depth and thermal input pair $\{x_D, x_T\}$, our goal is to obtain a fusion model f_F such that the prediction outputs $\hat{y} = f_F(x_D, x_T)$ are more accurate than models trained only on a single modality, f_D and f_T. We identify two general fusion strategies on multimodal data, namely: *early fusion* and *late fusion*, as they have been shown to work well in various video analysis tasks (Simonyan and Zisserman, 2014; Karpathy et al., 2014; Feichtenhofer et al., 2016).

Early fusion. This refers to the fusion scheme that integrates unimodal features before learning concepts. In our experiment, we combine the two modalities directly on the pixel level by concatenating the input data channel-wise. The model takes in the concatenated input and is trained the same ways as a single network.

Late fusion. This refers to the fusion scheme that first reduces unimodal features to separately learned concept scores, which are then integrated to learn concepts. In our experiments, we first train two separate models for depth and thermal, f_D and f_T. At test time, for a given video clip with two modalities (x_D, x_T), we average the softmax scores $f(x_D, x_T) = \frac{1}{2}\big(f_D(x_D) + f_T(x_T)\big)$ to make the final prediction. This is similar to the two-stream activity recognition strategy used in (Simonyan and Zisserman, 2014).

This score fusion can be applied to both classification and detection tasks. As observed in many previous work (Simonyan and Zisserman, 2014; Tran et al., 2015; Carreira and Zisserman, 2017), a simple fusion of different modalities usually yields better performance. In Section 4.3 and 4.4, we will show that this is true for us as well. The question arises whether early fusion or late fusion is the preferred method for senior home activity understanding. In this paper, we discuss both multimodal fusion approaches and perform a comparative evaluation, as shown in Table 2.

4. Results

The following sections describe the details of our experiments and results for activity classification, temporal activity detection, and interpretable qualitative and quantitative descriptive analytics of the senior's daily activities.

4.1. Implementation Details

For our model, we use the ResNet-34 architecture (He et al., 2016). The inputs to our model are depth and thermal video clips. The original input sizes of depth and thermal images are 240×320 and 160×120 respectively. We scale all inputs to size 224×224, which is the size required by the ResNet-34 network. We normalize all inputs by the mean and standard deviation in the training set, which is $(77.0, 20.1)$ for depth and $(37.0, 2.1)$ for thermal. For data augmentation, we only use random horizontal flipping. We initialize our network with ImageNet pre-trained weights. Since the pre-trained checkpoint is trained on RGB images, it has a different number of input channels. Thus, using the method from Wang et al. (2016), we average the weights across the RGB channels and replicate this average by the channel number. We train using SGD for 40 epochs, with batch size 8, momentum 0.9 and

Table 2: Activity classification results in instance-level mAP. "Early fusion" and "Late Fusion" refer to the two methods we deployed to combine multimodal information, mentioned in Section 3.3. The thermal-only model outperforms the depth-only model on 5 activity categories, while the depth-only model outperforms the thermal-only model on 2 activity categories. The joint model using late fusion achieves the best overall results.

	Thermal	Depth	Early Fusion	Late Fusion
Sitting	**0.989**	0.967	0.967	**0.985**
Standing	**0.966**	0.963	0.949	**0.983**
Walking	0.965	**0.972**	0.955	**0.989**
Sleeping	**1.000**	0.960	0.960	**1.000**
Getting Assistance	**0.850**	0.817	0.759	**0.857**
Using Bedside Commode	0.953	**1.000**	0.974	**1.000**
Background	**0.855**	0.846	0.861	**0.867**
Mean AP	**0.940**	0.932	0.918	**0.958**

weight decay 5×10^{-4}. Our initial learning rate is 0.001, and we decay the learning rate by a factor of 10 at epoch 20 and 30, similar to the training schedule as He et al. (2016).

For classification, we set the length of the video clips T to be 10, and at test time we sample $N = 5$ clips from each video. For detection, video clip length T is also 10, the smoothing window W is set to 5.

4.2. Evaluation Metrics

Activity Classification. We evaluate our model performance using classification accuracy and instance-level mean average precision (instance-level mAP). Classification accuracy is a standard metric for classification. However, since our dataset is highly imbalanced (see Table 1), accuracy can be misleading. For example, a trivial model that simply outputs the dominant class for every instance can still achieve a high accuracy. In this situation, instance-level mAP is a more suitable metric. Note that mean Average Precision (mAP) is calculated by averaging the average precision (AP) of each category, and Average Precision (AP) is calculated as the area under the precision-recall curve.

Temporal Activity Detection. Our task is predicting dense labels of long videos, i.e., assigning a label for each frame. Frame-level mean Average Precision (frame-level mAP) is a common evaluation metric and has been used in several papers (Yeung et al., 2017; Gorban et al., 2015).

4.3. Activity Classification Results

We train our classification model on 7 days of data and test on 3 other days of data. The results are shown in Table 2. Both the depth and thermal models achieve over 0.9 mAP. For all classes except Walking and Using Bedside Commode, thermal outperforms depth

Table 3: Temporal activity detection results in frame-level mAP. We achieve 0.903 frame-level mean AP across all 7 activity classes (6 activities + background).

	Late Fusion
Sitting	0.940
Standing	0.910
Walking	0.717
Sleeping	0.947
Getting Assistance	0.994
Using Bedside Commode	0.995
Background	0.991
Mean AP	0.903

modality. This is expected since it is easier to locate the senior's location in a thermal video by body temperature. The confusion matrix (Figure 3(a)) gives some interesting insights. The most common mistake is confusion between Background with Getting Assistance, since there are often caregivers in the room for these actions. Another reasonable mistake is confusing the class *standing* with the class *walking*, as the definitions of these two activities are sometimes ambiguous (e.g. when the senior is standing but slightly shifting the feet).

The late fusion of depth and thermal modalities achieves the best results, which is consistent with previous work (Simonyan and Zisserman, 2014; Tran et al., 2015; Carreira and Zisserman, 2017). This shows that different modalities provide complementary information, which improves performance when combined together. Interestingly, early fusion does not improve model performance.

4.4. Temporal Activity Detection Results

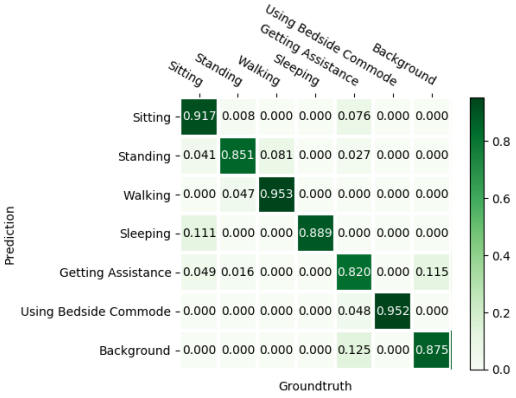
We evaluate our model on 3 days of long continuous videos. Table 3 shows the results of temporal activity detection, and Figure 3(b) shows the corresponding confusion matrix. We show the frame-level mean Average Precision (AP), as well as the AP values for individual action classes. Our model achieves over 0.9 mAP for all classes except for the class Walking. This allows us to build an accurate activity timeline of the senior, and calculate significant healthcare analytics, which will be discussed in the next section.

4.5. Descriptive Analytics

We then provide various qualitative and quantitative descriptive analytics of the senior's activities from the temporal activity detection results on 14 consecutive days of continuous unlabeled data. These qualitative and quantitative analytics provide useful clinical information that can be interpreted by a caregiver.

Duration and Number of Instances of Activities. The activity timeline allows us to easily calculate duration and number of instances of activity, which can be used to further extract useful information. In Table 4, we show the mean and standard deviation of the

(a) The confusion matrix of activity classification results in Table 2.



(b) The confusion matrix of temporal activity detection results in Table 3.

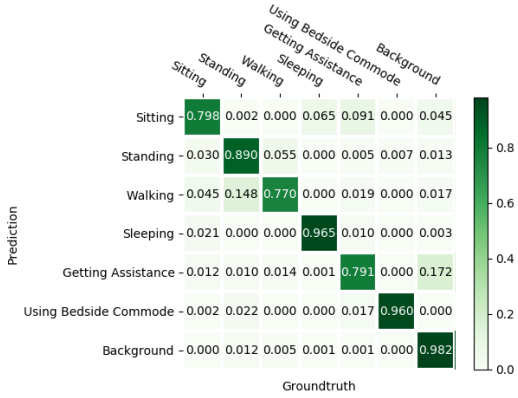


Figure 5: Confusion matrices of activity classification and temporal activity detection across different classes.

Table 4: Mean and standard deviation of total daily duration and number of instances of each activity. As expected, the class *sleep* has the longest average duration and the class *using bedside commode* has the shorted average duration.

Activity	Duration (min)	# instances
Sitting	81.75 ± 30.06	82.36 ± 67.28
Standing	27.23 ± 7.84	70.71 ± 16.52
Walking	4.76 ± 1.01	54.50 ± 11.10
Sleeping	626.11 ± 25.90	28.86 ± 18.13
Getting Assistance	114.01 ± 34.98	68.93 ± 31.05
Using Bedside Commode	1.69 ± 1.99	2.86 ± 4.76

total daily duration and number of daily instances of each activity. In Figure 6, we show the duration of sleep across 14 days. These results allows caregivers to easily observe changes over time or anomalous behavior, e.g. less sleep than normal or using bedside commode more frequently than normal.

Transitions between Activities. We also calculate the average number of transitions between activities in Figure 7. The number of transitions is a good indicator of how active the senior is throughout the day. Some interesting examples include how fragmented the senior's sleep is (*Sleeping* $\rightarrow$ *Sitting*), and how many times the senior stands up (*Sitting* $\rightarrow$ *Standing*).

Temporal Heatmap. As shown in Figure 8, we provide a temporal heatmap for each activity, which displays the average temporal density of the activity across a 24 hour period.

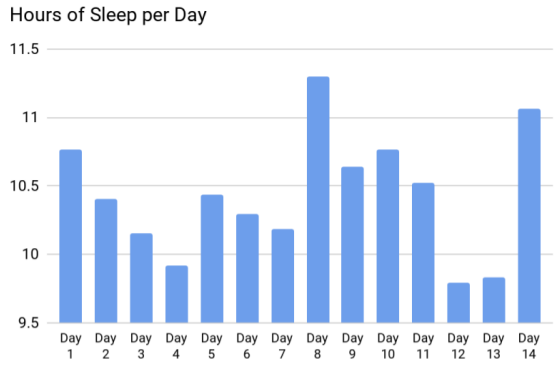


Figure 6: Duration of sleep across 14 days.

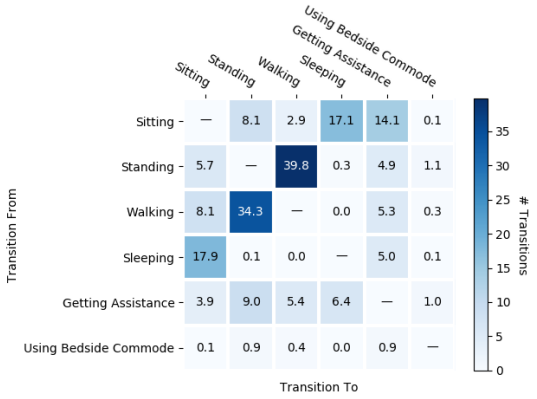


Figure 7: Average number of transitions per day between activities.

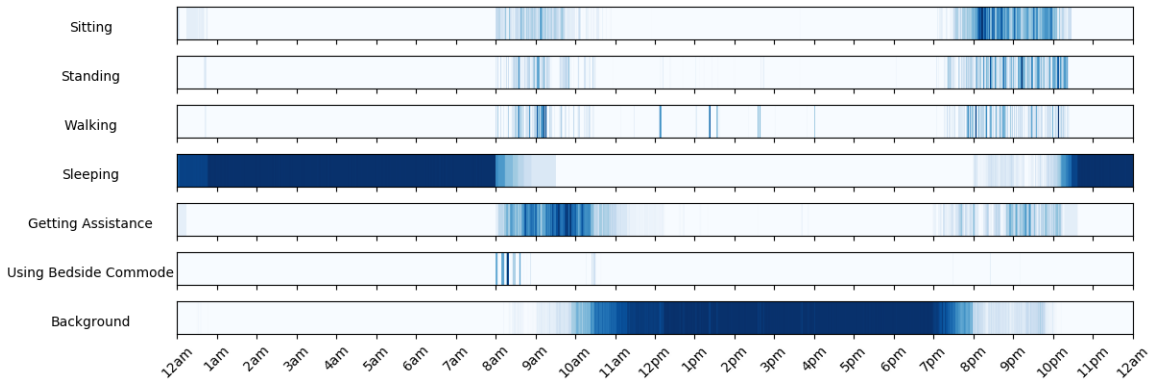


Figure 8: The temporal heatmap of each of the 7 activity classes (6 activities + background). At a certain time, the darker the color is, the more frequently the corresponding activity was performed, averaged across 14 days.

Temporal Heatmap provides information about the habits of the senior, and it also allows caregivers to monitor changes in the senior's daily living patterns.

Spatial Heatmap. In Figure 9, for each activity we provide a spatial heatmap indicating average physical location of the senior for that activity. We generate the spatial heatmaps by performing background subtraction with the thermal data to segment out the senior from the background. A segmentation of a background frame is given in Figure 9(h) to provide a clear visualization of the room layout. This provides a visualization of the locations in which the senior performs each activity. Note that for the background class, even though the senior is not present, caregivers sometimes come in the room and thus detected by our algorithm.

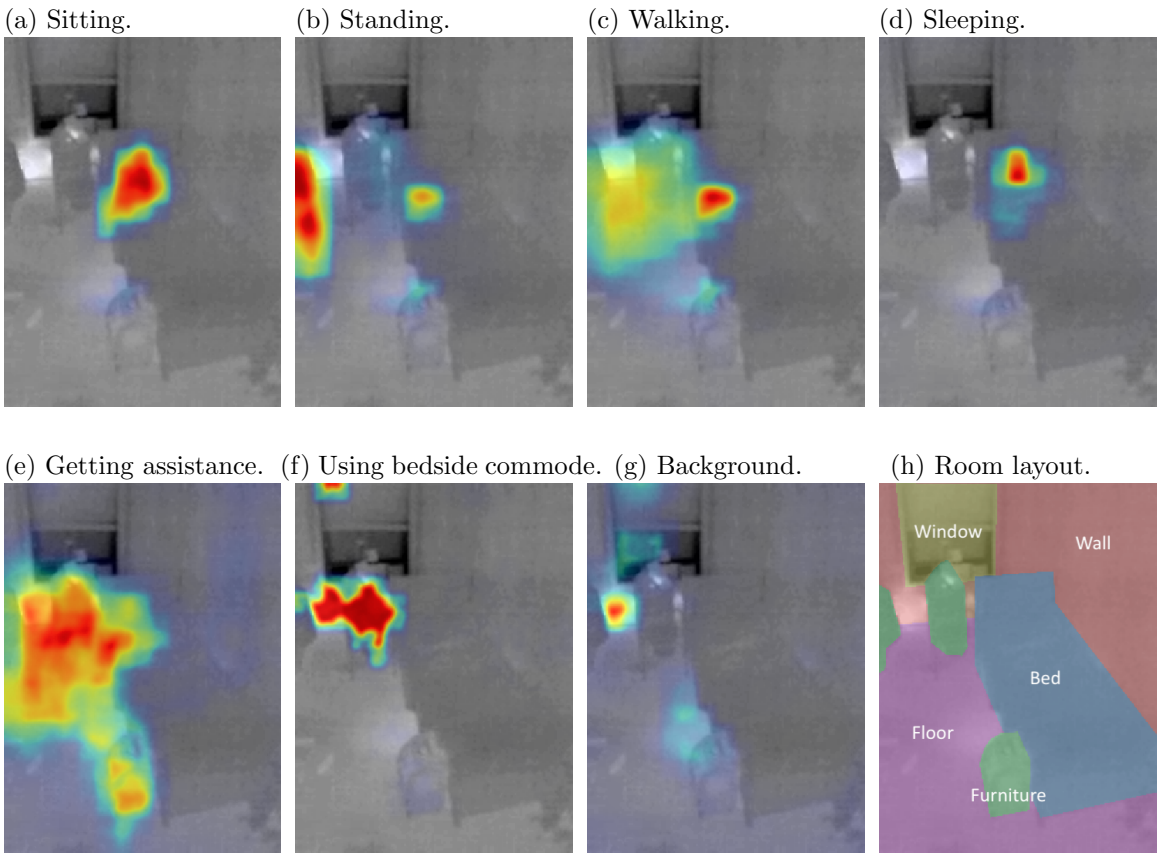


Figure 9: The spatial heatmap of each of the 7 activity classes (6 activities + background). The area highlighted in red corresponds to the area where the activity happened most frequently. (h) shows the background room layout for reference.

5. Discussion and Related Work

Most of the existing approaches to activity monitoring require wearable devices (González-Valenzuela et al., 2011; Lee and Carlisle, 2011; Park and Jayaraman, 2003), but due to their required adherence and other sensor limitations, there has been little evidence that these approaches improve health outcomes or quality of life. Non-intrusive vision-based approaches based upon stationary cameras or sensors may allow passive detection of important activities. A number of such vision-based approaches have also been developed for automated senior health monitoring, many of which are specifically for fall detection (Zhang et al., 2015). While these approaches are important for identifying critical acute conditions, they have not focused on detecting deviations from normal behaviors – another important category of activities. Parajuli et al. (2012) used Kinect depth sensors to detect sitting, standing, and normal and abnormal walking. However, they did not include other clinically-relevant activities such as sleeping and using the bedside commode. Furthermore, their data was simulated by researchers and may not be representative of real-world senior data from senior homes. Cheng et al. (2014) performed temporal activity detection of various

clinically-relevant activities on RGB videos, which is usually impractical because of privacy considerations. Xiang et al. (2015) performed person tracking and posture recognition by using multiple RGB cameras, which could be costly and intrusive to the seniors. To our knowledge, this study is the first automated, privacy-compliant, vision-based system for continuous senior activity detection and long-term health analysis.

Our study introduces a system that accurately monitors clinically-relevant daily activities of seniors. We have introduced a newly collected activity classification and detection dataset with depth and thermal modalities, and our model achieves excellent results on both activity classification and continuous temporal activity detection tasks. Furthermore, we have provided interpretable qualitative and quantitative descriptive analytics of the senior's daily activities over a long period. This analytics can allow caregivers and medical professionals to more efficiently identify patterns of abnormal activity that might be representative of health concerns.

In our future work, we aim to diversify the dataset by incorporating data from more seniors and room layouts. This allows us to improve the robustness of our model and provide more accurate descriptive analytics across different sites and settings. We also plan to include gait, posture analysis as well as fall detection in our system.

Acknowledgments

This work was supported in part by On Lok, Inc., Clinical Excellence Research Center, and Stanford Computer Science Department. We specially thank Julia A. Lee, Wanda Chin, Alexandre Alahi, Federico Polacov, Jose Tiradocrespo, Manuel Ramirez for their managerial and engineering supports.

References

David E Bloom, Axel Boersch-Supan, Patrick McGee, and Atsushi Seike. Population aging: facts, challenges, and responses. *Benefits and compensation International*, 41(1):22, 2011.

Daniel J Buysse, Charles F Reynolds III, Timothy H Monk, Carolyn C Hoch, Amy L Yeager, and David J Kupfer. Quantification of subjective sleep quality in healthy elderly men and women using the pittsburgh sleep quality index (psqi). *Sleep*, 14(4):331–338, 1991.

Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017.

Gideon Charach, Alexander Greenstein, Pavel Rabinovich, Itamar Groskopf, and Moshe Weintraub. Alleviating constipation in the elderly improves lower urinary tract symptoms. *Gerontology*, 47(2):72–76, 2001.

Hong Cheng, Zicheng Liu, Yang Zhao, Guo Ye, and Xinghai Sun. Real world activity summary for senior home monitoring. *Multimedia Tools and Applications*, 70(1):177–197, 2014.

Jia-Luen Chua, Yoong Choon Chang, and Wee Keong Lim. A simple vision-based fall detection technique for indoor video surveillance. *Signal, Image and Video Processing*, 9 (3):623–633, 2015.

Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *CVPR*, 2016.

James E Gangwisch, Steven B Heymsfield, Bernadette Boden-Albala, Ruud M Buijs, Felix Kreier, Mark G Opler, Thomas G Pickering, Andrew G Rundle, Gary K Zammit, and Dolores Malaspina. Sleep duration associated with mortality in elderly, but not middle-aged, adults in a large us sample. *Sleep*, 31(8):1087–1096, 2008.

Alba Gómez-Cabello, Raquel Pedrero-Chamizo, Pedro R Olivares, Rayco Hernández-Perera, José A Rodríguez-Marroyo, Esmeralda Mata, Susana Aznar, José G Villa, Luis Espino-Torón, Narcis Gusi, et al. Sitting time increases the overweight and obesity risk independently of walking time in elderly people from spain. *Maturitas*, 73(4):337–343, 2012.

Sergio González-Valenzuela, Min Chen, and Victor CM Leung. Mobility support for health monitoring at home using wearable sensors. *IEEE Transactions on Information Technology in Biomedicine*, 15(4):539–549, 2011.

A Gorban, H Idrees, YG Jiang, A Roshan Zamir, I Laptev, M Shah, and R Sukthankar. Thumos challenge: Action recognition with a large number of classes. In *CVPR workshop*, 2015.

Keith Hall. The 2015 long-term budget outlook. CONGRESSIONAL BUDGET OFFICE (US CONGRESS) WASHINGTON DC, 2015.

Albert Haque, Michelle Guo, Alexandre Alahi, Serena Yeung, Zelun Luo, Alisha Rege, Jeffrey Jopling, Lance Downing, William Beninati, Amit Singh, et al. Towards vision-based smart hospitals: A system for tracking and monitoring hand hygiene compliance. In *MLHC*, 2017.

Lauren Harris-Kojetin, Manisha Sengupta, Eunice Park-Lee, Roberto Valverde, Christine Caffrey, Vincent Rome, and Jessica Lendon. Long-term care providers and services users in the united states: data from the national study of long-term care providers, 2013-2014. *Vital & health statistics. Series 3, Analytical and epidemiological studies*, (38):x–xii, 2016.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014.

Raymond YW Lee and Alison J Carlisle. Detection of falls using accelerometers and mobile phone technology. *Age and ageing*, 40(6):690–696, 2011.

Zelun Luo, Alisha Rege, Guido Pusiol, Arnold Milstein, Fei-Fei Li, and N. Lance Downing. Computer vision-based approach to maintain independent living for seniors. In *AMIA Annual Symposium*, 2017.

Georgios Mastorakis and Dimitrios Makris. Fall detection system using kinects infrared sensor. *Journal of Real-Time Image Processing*, 9(4):635–646, 2014.

Alberto Montes, Amaia Salvador, and Xavier Giro-i Nieto. Temporal activity detection in untrimmed videos with recurrent neural networks. *arXiv preprint arXiv:1608.08128*, 2016.

Jennifer M Ortman, Victoria A Velkoff, and Howard Hogan. *An aging nation: the older population in the United States*. United States Census Bureau, Economics and Statistics Administration, US Department of Commerce, 2014.

A Österberg, W Graf, U Karlbom, and L Påhlman. Evaluation of a questionnaire in the assessment of patients with faecal incontinence and constipation. *Scandinavian journal of gastroenterology*, 31(6):575–580, 1996.

Monish Parajuli, Dat Tran, Wanli Ma, and Dharmendra Sharma. Senior health monitoring using kinect. In *Communications and Electronics (ICCE), 2012 Fourth International Conference on*, pages 309–312. IEEE, 2012.

Sungmee Park and Sundaresan Jayaraman. Enhancing the quality of life through wearable technology. *IEEE Engineering in medicine and biology magazine*, 22(3):41–48, 2003.

Elizabeth A Phelan, Barbara Williams, Brenda WJH Penninx, James P LoGerfo, and Suzanne G Leveille. Activities of daily living function and disability in older adults in a randomized trial of the health enhancement program. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, 59(8):M838–M843, 2004.

Fabio Pitta, Thierry Troosters, Martijn A Spruit, Vanessa S Probst, Marc Decramer, and Rik Gosselink. Characteristics of physical activities in daily life in chronic obstructive pulmonary disease. *American journal of respiratory and critical care medicine*, 171(9): 972–977, 2005.

Donald Redfoot, Lynn Feinberg, and Ari N Houser. *The aging of the baby boom and the growing care gap: A look at future declines in the availability of family caregivers*. AARP Public Policy Institute Washington, DC, 2013.

Caroline Rougier, Edouard Auvinet, Jacqueline Rousseau, Max Mignotte, and Jean Meunier. Fall detection from depth map video sequences. In *International Conference on Smart Homes and Health Telematics*, pages 121–128. Springer, 2011a.

Caroline Rougier, Jean Meunier, Alain St-Arnaud, and Jacqueline Rousseau. Robust video surveillance for fall detection based on human shape deformation. *IEEE Transactions on circuits and systems for video Technology*, 21(5):611–622, 2011b.

Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, 2014.

Stephanie Studenski, Subashan Perera, Kushang Patel, Caterina Rosano, Kimberly Faulkner, Marco Inzitari, Jennifer Brach, Julie Chandler, Peggy Cawthon, Elizabeth Barrett Connor, et al. Gait speed and survival in older adults. *JAMA*, 305(1):50–58, 2011.

Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, 2015.

CJ Vaizey, AJ Roy, and MA Kamm. Prospective evaluation of the treatment of solitary rectal ulcer syndrome with biofeedback. *Gut*, 41(6):817–820, 1997.

Gabor Abellan Van Kan, Y Rolland, S Andrieu, J Bauer, O Beauchet, M Bonnefoy, M Cesari, LM Donini, S Gillette-Guyonnet, M Inzitari, et al. Gait speed at usual pace as a predictor of adverse outcomes in community-dwelling older people an international academy on nutrition and aging (iana) task force. *The journal of nutrition, health & aging*, 13(10):881–889, 2009.

Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks towards good practices for deep action recognition. In *ECCV*, 2016.

Yun Xiang, Yi-ping Tang, Bao-qing Ma, Hang-chen Yan, Jun Jiang, and Xu-yuan Tian. Remote safety monitoring for elderly persons based on omni-vision analysis. *PloS one*, 10(5):e0124068, 2015.

Serena Yeung, Olga Russakovsky, Ning Jin, Mykhaylo Andriluka, Greg Mori, and Li Fei-Fei. Every moment counts: Dense detailed labeling of actions in complex videos. *International Journal of Computer Vision*, 126(2-4):375–389, 2017.

Miao Yu, Adel Rhuma, Syed Mohsen Naqvi, Liang Wang, and Jonathon Chambers. A posture recognition-based fall detection system for monitoring an elderly person in a smart home environment. *IEEE transactions on information technology in biomedicine*, 16(6):1274–1286, 2012.

Zhong Zhang, Christopher Conly, and Vassilis Athitsos. A survey on vision-based fall detection. In *Proceedings of the 8th ACM international conference on PErvasive technologies related to assistive environments*, page 46. ACM, 2015.