applied sciences

MDPI

Article

A Vision-Based System for Monitoring Elderly People at Home

Marco Buzzelli *,† **, Alessio Albé** † **and Gianluigi Ciocca** †

Department of Computer Sciences Systems and Communications, University of Milano—Bicocca, viale Sarca 336, 20126 Milan, Italy; a.albe1@campus.unimib.it (A.A.); gianluigi.ciocca@unimib.it (G.C.)

* Correspondence: marco.buzzelli@unimib.it

† The authors contributed equally to this work.

Received: 22 November 2019; Accepted: 27 December 2019; Published: 3 January 2020

Abstract: Assisted living technologies can be of great importance for taking care of elderly people and helping them to live independently. In this work, we propose a monitoring system designed to be as unobtrusive as possible, by exploiting computer vision techniques and visual sensors such as RGB cameras. We perform a thorough analysis of existing video datasets for action recognition, and show that no single dataset can be considered adequate in terms of classes or cardinality. We subsequently curate a taxonomy of human actions, derived from different sources in the literature, and provide the scientific community with considerations about the mutual exclusivity and commonalities of said actions. This leads us to collecting and publishing an aggregated dataset, called ALMOND (Assisted Living MONitoring Dataset), which we use as the training set for a vision-based monitoring approach.We rigorously evaluate our solution in terms of recognition accuracy using different state-of-the-art architectures, eventually reaching 97% on inference of basic poses, 83% on alerting situations, and 71% on daily life actions. We also provide a general methodology to estimate the maximum allowed distance between camera and monitored subject. Finally, we integrate the defined actions and the trained model into a computer-vision-based application, specifically designed for the objective of monitoring elderly people at their homes.

Keywords: computer vision; action recognition; deep learning; internet of things; assisted living

1. Introduction

Many elderly people require regular assistance for their daily living and healthcare. There is an increased awareness in developing and implementing efficient and cost-effective strategies and systems, to provide affordable healthcare and monitoring services particularly aimed at the aging population. Aging in place is the ability to live in one's own home and community safely, independently, and comfortably, regardless of age, income, or ability level. For elderly people, moving in with the family or entering a nursing home or assisted living facility could be cause of psychological stress, which can lead to health issues and lowering their quality of life.

Allowing elderly people to maintain their quality of life as they get older and as long as possible in their homes is important both for the person as well as for the sustainability of public healthcare systems. According to the "World Population Prospects 2019: Highlights" of the United Nations [1], in 2018, for the first time in human history, people aged 65 years or over outnumbered children under five years of age worldwide. The projections indicate that in 2050 there will be more than twice as many older people as children under five (see Figure 1). Among the aging countries there is Italy. As for 2016, Italy has 22.1% of residents aged more than 65, and 6.7% aged more than 80 years. These percentages are expected to increase up to 33% and 15% respectively within 2070 [2]. With these numbers, Italy has one of the oldest population in Europe (see Figure 2 left) and one of the countries affected by the

highest *old age dependency ratio*, i.e., the ratio of people older than 64 compared with those aged 15–64 (see Figure 2 right).

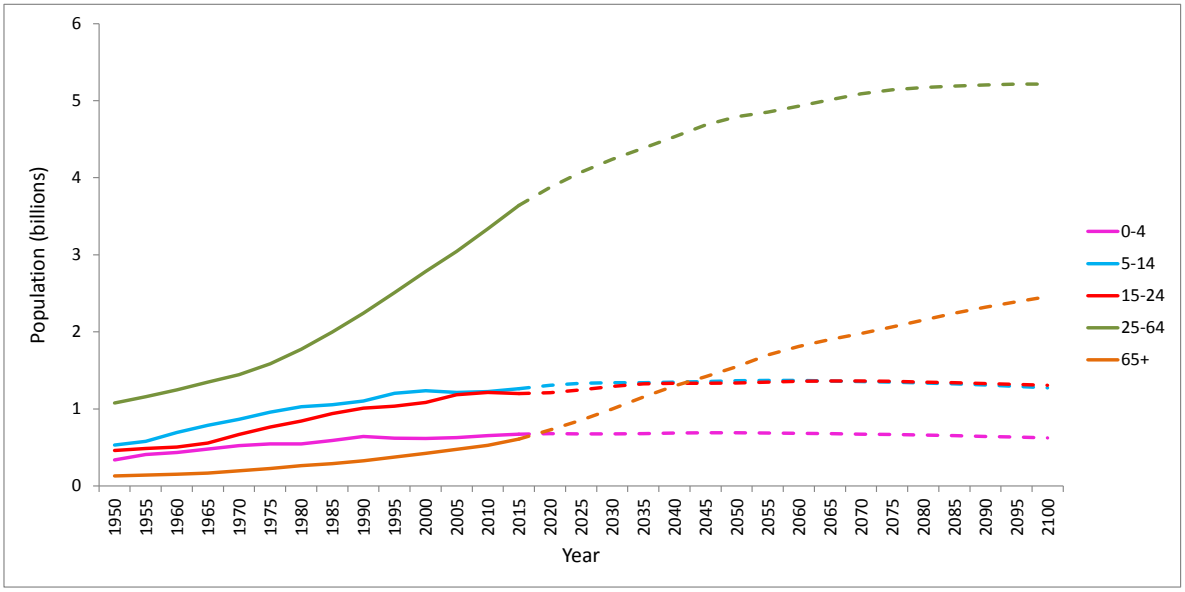


Figure 1. People aged 65+ years old make up the fastest-growing age group worldwide. Data from [3].

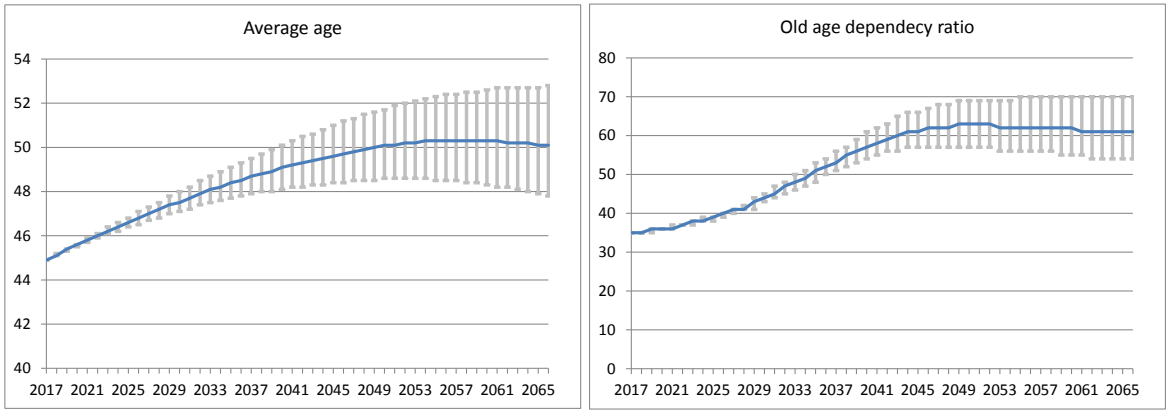


Figure 2. Estimated aging trend in Italy as per the Italian National Institute of Statistics (ISTAT). Estimates are shown within 90% confidence intervals.

If the aging trend is confirmed, there will be fewer people to take care of the elderly in the distant future. It is not surprising that within the "Horizon 2020" Research and Innovation Program of the European Union, many national projects in Italy are oriented toward the older population to promote healthy and active aging, and test new technologies for the sustainability of the healthcare system. The final aim of most of these projects is to develop a novel assistance system with the goal of preserving as long as possible the remaining autonomy of elderly people so that they can live at home instead of being transferred to public or private nursing homes [4]. Always under the "Horizon 2020" umbrella, a noteworthy program called Active and Assisted Living (AAL) [5] stands out. It is a funding program that aims to create a better quality of life for older people and it is placed in the field of healthy aging technology and innovation.

Assisted living technologies can be of great importance to take care of elderly people and help them to live independently. One way to achieve this is to monitor the activities of the elderly in a continuous fashion to detect emergency situations as soon as possible. For example, using ambient or wearable sensors it could be possible to analyze the daily activity of the person and detect if any activity

is outside normal activity patterns. Also, it could be possible to prevent health issue by monitoring the person's behavior with respect to dehydration and lack of food intakes. Finally, emergency events such as falls and pains could be signaled to the healthcare facilities or family members for immediate intervention. These events can be automatically registered by the sensors or signaled by the person requesting help.

Figure 3 shows a general architecture of an Ambient Assisted Living (AAL) system that can be used for monitoring elderly people. Depending on the context in which the system must operate and the needs of the subject to be monitored, one or more sensors can be used and deployed for collecting behavioral and personal data. Several types of sensors can be exploited: audio (i.e., microphones), visual (i.e., 2D and/or 3D cameras), environmental (i.e., pressure, infrared, radar, ...), and physiological (i.e., blood pressure, body temperature, ...). The data acquired by the sensors is stored in a database system for logging and analysis. The data can be continuously processed to detect anomalous activities or danger situations that require intervention. The data can also be used to perform long-term monitoring and analysis of the activities by the family members, caregivers, or experts.

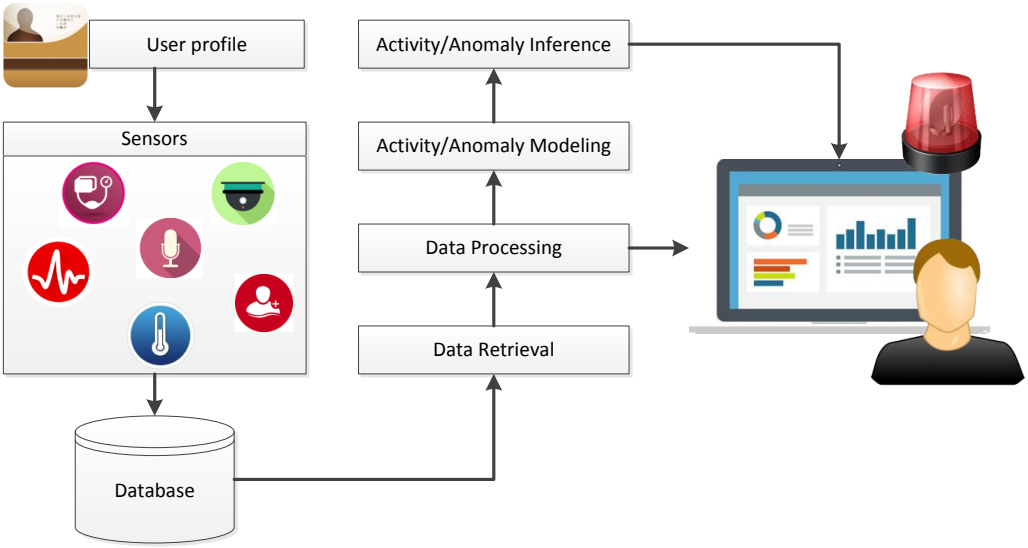


Figure 3. General architecture of an Ambient Assisted Living system.

Different monitoring systems have been proposed in the literature and have been surveyed in [6–8]. These systems mostly differ by the sensors used for collecting the raw data, i.e., single vs. multiple sensors and mono or multi-modal sensing devices, and in the final aim, i.e., registering specific activities, daily logging or detecting dangerous situations. All these systems have basic architectures similar to the one shown in Figure 3.

Systems that exploit wearable sensors can be perceived as being intrusive by the users, while systems based on ambient sensors requires the installment of specific hardware in the rooms of the house that can be also problematic on existing buildings. In this work, we tackle the problem of monitoring elderly people at home by exploiting computer vision technologies and visual sensors such as RGB cameras. More specifically, we designed an assisted living monitoring system to record and analyze daily activities of elderly people at home. The system is designed to be as unobtrusive as possible and thus it does not require the person to wear sensors. Instead it relies on a camera to collect video streams that are processed to recognize and store the target's behaviors.

For this work we envision a single-camera scenario, where the acquisition device is located in a strategic position, possibly in the living room above a TV screen. To this extent, we provide experiment-based recommendations about the suggested location, driven by the estimated maximum distance between the camera and the subject. The natural extension of this setup involves installing multiple cameras in the house, possibly including 360° cameras, and leveraging on person

re-identification techniques to provide a consistent analysis, although we reserve this development for future works.

To perform the monitoring, the system integrates advanced action recognition algorithms that are robust for the indoor scenario. The output of the monitoring is then used to provide alert messages in case of anomalous events that can be selected by the user. We designed our system to be accessible and reliable. We also propose an application to support raw data collection, activity monitoring, visual log generation, and support for anomaly inference and alerting, through a user-friendly interface.

The rest of the paper is organized as follow: Section 2 presents an overview of the existing action recognition methods and systems, ending with an overview of commonly used action recognition datasets. In Section 3 we describe the creation of our reference dataset of actions that will be used for the design and evaluation of action recognition algorithms. In Section 4, we introduce our proposed monitoring system based on action recognition, and how we approached the subject localization and the recognition of the action. Experimental results are described and analyzed in Section 5. In Section 6 we present a client-server application based on our monitoring system. Finally, Section 7 concludes the paper.

2. Related Works

In the following section we present a review of literature approaches on the problem of action-based monitoring at home. Section 2.1 covers existing systems for ambient assisted living, with a particular focus on solutions aimed at elderly care. Since our own system is based on action recognition algorithms to perform the monitoring of the human behavior, in Section 2.2 we revise some notable works in human action recognition. The type and source of the data used in the design and validation of action recognition algorithms is very important. To this extent, in Section 2.3 we review the most used action recognition datasets available in the literature.

2.1. Ambient Assisted Living Systems

Several surveys in the literature describe recent trends in smart homes aimed at assisted living systems [6–9]. These monitoring systems can use exclusively ambient sensors (i.e., RGB and/or infrared cameras) to limit user discomfort as much as possible [10–12], can use wearable sensors if health parameters need to be monitored [13], or can exploit different modalities at the same time [6–8].

The following systems make a pervasive use of ambient and wearable sensors. Necesity [14] is an ambient assisted living system, which monitors the states of the elderly (out, active, inactive, resting, sleeping and inactive anomalous), through different ambient sensors (pressure, door and activity) scattered into the environment. Both [15,16] present an elderly healthcare system aimed at monitoring different activities using body sensors. A significant issue for systems based on body sensors is the need to apply them onto the subject, for better accuracy or to detect more actions or activities. This can be considered a critical aspect because wearable sensors can lead to physical discomfort for the user. A different kind of sensor, less invasive and more discreet, is used in the system presented by [17], which can both track and detect the fall of elderly people using smart tiles.

Regarding video-based systems, ref. [18] propose a method for human posture-based and movements-based monitoring, limited however to only 5 postures (standing, bending, sitting, lying and lying toward) and 4 movements (running, jump, inactive, active). IFADS (Image-based FAll Detection System) [19] focuses on falls that might happen while sitting down and standing up from a chair, a situation of potential danger for elderly people.

In this work we design a monitoring system which exploits visual data. This type of information is easily acquired using RGB cameras that can be placed in the environment with minimum effort. As an indication of the pervasiveness and affordability of this kind of sensors, the survey [8] reports more than 60 works on activity monitoring systems exploiting visual data, and about 20 works for wearable sensors. Moreover, different from existing solutions in the literature, our system is carefully designed to support the recognition and monitoring of a wider variety of actions, including status,

different alerting situations as well as daily life activities. These actions have been specifically selected for monitoring elderly people at home.

2.2. Action Recognition Methods

In recent years, deep learning received considerable attention in computer vision applications. Many deep learning-based approaches have been proposed to tackle the problem of human action recognition. In the following we present some works on action recognition mostly based on RGB inputs and exploiting different deep learning strategies. Table 1 summarizes the performance of some of the relevant methods in the state-of-the-art. Results are reported for the most common datasets. The works in the literature are mainly based on one of the following three deep learning strategies: fusing different pieces of information about the video stream (i.e., two-stream networks [20]); including spatio-temporal structure (i.e., 3D convolutional networks [21]); including temporal analysis of video contents (i.e., long short-term memory networks [22]).

Table 1. List of major action recognition methods and their performance on used datasets.

Method	Year	HMDB-51 [23]	UCF-101 [24]	Kinetics [25]	Charades [26]	NTU [27]
Two streams (RGB+OF) [20]	2014	59.4%	88.0%			
C3D+Linear SVM [21]	2015		85.2%			
LSTM30+OF+RGB [28]	2015		88.6%			
S:VGG-16, T:VGG-16 [29]	2016	65.4%	92.5%			
TSN (3 modalities) [30]	2016	69.4%	94.2%			
ST-LSTM+Trust Gate [31]	2016					69.2%
LTC [32]	2017	67.2%	92.7%			
I3D [33]	2017	80.9%	98.0%	74.2%		
T3D(+TSN) [34]	2017	63.5%	93.2%	71.5%		
P3D ResNet [35]	2017		88.6%			
L^2STM [36]	2017	66.2%	93.6%			
STA-LSTM [37]	2018					73.4%
DTMV+RGB-CNN [38]	2018	55.3%	87.5%			
R(2+1)D-Two [39]	2018	78.7%	97.3%	75.4%		
NL I3D [40]	2018			77.7%	39.5%	
VideoLSTM [41]	2018	56.4%	88.9%			
DeepHAR (RGB only) [42]	2018					84.6%
R(2+1)D-152 [43]	2019			81.3%		
PA3D+I3D [44]	2019	82.1%			41.0%	

In two-stream networks, the spatial RGB information, is usually combined with temporal information in the form of motion vectors or optical flows. The two sources of information are used for training two separate networks and the outputs are fused in late layers. In [20], a spatial network is trained on single RGB frames while the temporal network is trained on a stack of optical flow frames. The two networks perform classification and the fusion is applied to the class scores using a Support Vector Machine (SVM). In [38], computed optical flows are substituted with motion vectors that are readily available in video streams, thus improving the efficiency of the two-stream networks making the approach usable in real-time applications. In [29], instead of performing late fusion, the two streams are fused in middle layers using convolution and pooling layers. Long-range temporal structure modeling and warped flows are exploited in [30] in a temporal segment network (TSN) improving the original two-stream network results. Fusing multiple information is also exploited in [42] where a multitask deep architecture is used to perform 2D and 3D pose estimation jointly with action recognition. The model first predicts location of body joints and then, using this information, it predicts the action performed in the video. The joint pose/action learning and recognition is shown to be more robust than using the information separately.

Temporal information can be also incorporated into the network architectures by considering stack of frames and 3D convolutions. The work introducing this rationale is [21]. It is shown that an architecture with small $3 \times 3 \times 3$ convolution kernels in all layers can improve recognition performance. The new architecture (3D ConvNet) is able to produce robust features (C3D) that can be effectively used in a simple linear classifier for action recognition. Carreira et al. [33] extend a two-stream network architecture, based on inception-V1, with 3D convolutions creating a two-stream inflated 3D ConvNet (I3D). The temporal 3D ConvNet [34] extends a DenseNet architecture by introducing a new temporal layer that models variable temporal convolution kernel depths with 3D filters and 3D pooling kernels. An approach to learn video representations using neural networks with long-term temporal convolutions (LTC) is presented in [32]. Different low-level frame representations are considered, and high-quality optical flows are found to be the most relevant for robust action recognition. Standard Convolutional Neural Networks (CNN) analyze information at local neighborhood. Wang et al. [40] introduced non-local operations as a generic family of building blocks for capturing long-range dependencies in action recognition videos. Experiments are performed on the Inflated 3D ConvNet showing improvements to the Kinetics dataset. Tran et al. [39] demonstrated the advantages of 3D CNNs over 2D CNNs within the framework of residual learning. The 3D convolutional filters are factorized into separate spatial and temporal components. The devised R(2+1)D convolutional block (2+1-dimensional ResNet) can achieve comparable or superior results to the state-of-the-art methods.

Training networks for action recognition usually requires a large amount of annotated data. In [43] a study is conducted on how to improve action recognition classification using large-scale weakly supervised pre-training. The reference model used in the experimentation is the R(2+1)D-d [39]. Results shows that notwithstanding data noise, the models significantly improve the state-of-the-art performance. Most 3DCNN models are built upon RGB and optical flow streams and lack information about human pose. Yan et al. [44] proposed a novel model that encodes multiple pose modalities within a unified 3D framework. The model, Pose-Action 3D Machine (PA3D), exploits a novel temporal pose convolution to aggregate spatial poses over frames. Building deep 3DCNN results in expensive computational cost and memory demand. Qiu et al. [35] proposed a new family of Pseudo-3D (P3D) blocks to replace 2D Residual Units in ResNet achieving spatio-temporal encoding for videos.

3DCNN-based approaches incorporate temporal information extending filter and pooling layer to work with group of frames. Approaches based on Long-Short-Term Memory networks (LSTM) process a video as an ordered sequence of frames. Each frame is fed to the network that retains information about previous frames in internal memory states. Ng et al. [28] proposed a recurrent neural network that uses LSTM cells connected to the output of the underlying CNN. Donahue et al. [22] developed a novel recurrent convolutional architecture suitable for large-scale visual learning. The tested RNN models are directly connected to ConvNet models, and are trained to output variable length video descriptions of actions. In [31] a spatio-temporal LSTM for 3D human action recognition is proposed. The standard LSTM learning approach is extended to incorporate both spatial and temporal domains to analyze the hidden sources of action-related information within the input data over both domains concurrently. In standard LSTM approaches it is implicitly assumed that motions in videos are stationary across different spatial locations. To overcome this limitation, Lattice-LSTM [36] extends LSTM by learning independent hidden state transitions of memory cells for individual spatial locations. In [37] a network based on the recurrent neural networks with long short-term memory units is built. The model uses a spatial attention module to assign different levels of importance to different joints in a 3D skeleton. Moreover, a temporal attention module allocates different levels of attention to each frame within a sequence. Attention-LSTMs (ALSTMs) take into account spatial locality in the form of attention. To be applied for video sequence, VideoLSTM [41] enhances an ALSTM architecture by introducing Convolutional ALSTM modules to exploit the spatial correlation in frames, and a Motion-based Attention module to guide the network towards the relevant spatio-temporal locations.

2.3. Action Recognition Datasets

In the literature there exist many datasets that provide a list of different actions depending on the recognition task for which they have been designed. Table 2 shows a list of the most used datasets for human action recognition based on RGB data. For each dataset we report a brief description, the total number of action samples and the number of classes.

Table 2. List of the main RGB datasets used in human action recognition research field.

Dataset	Year	Samples	Classes
IXMAS [45]	2006	396	15
UCF Sport [46]	2008	150	10
Hollywood 2 [47]	2009	1707	12
HMDB-51 [23]	2011	6766	51
MSR Daily Activity 3D [48]	2012	320	16
UCF-101 [24]	2012	13,320	101
UCF-50 [49]	2013	6618	50
N-UCLA [50]	2014	1475	10
Sports 1M [51]	2014	1,133,158	487
UWA3D II [52]	2016	1075	30
Kinetics [25]	2017	306,245	400
DALY [53]	2016	3600	10
Charades [26]	2016	9848	157
NTU [27]	2016	56,880	60

IXMAS [45]: the dataset contains 15 actions captured from different viewpoints. A total of 11 people perform the following actions: nothing, check watch, cross arms, scratch head, sit down, get up, turn around, walk, wave, punch, kick, point, pick up, throw (over head), throw (from bottom up). The acquisition device is a low-resolution 23fps RGB camera. Its position is fixed, just as background, illumination and environment where actions are performed.

UCF Sport [46]: the dataset is a collection of sport videos acquired from a wide range of video sources. It contains 150 video sequences belonging to 10 actions: diving, golf swing, kicking, lifting, riding horse, running, skateboarding, swing-bench, swing-side, and walking.

Hollywood 2 [47]: twelve frequent actions in movies are considered in this dataset: answer phone, drive car, eat, fight person, get out car, hand shake, hug person, kiss, run, sit down, sit up, stand up. These actions have been labelled from 69 movie scripts and the corresponding sequence included in the dataset for a total of 600,000 frames and 7 h of video. The dataset also contains scene labelling.

HMDB-51 [23]: the dataset contains 51 action categories for a total of about 7000 manually annotated clips extracted from a variety of sources ranging from digitized movies to YouTube videos. Each action category contains at least 101 clips. The actions comprise: facial actions (e.g., smile), facial actions involving objects (e.g., smoking), body movements (e.g., clap hands), body movements involving objects (e.g., draw sword), human interaction (e.g., fencing).

MSR Daily Activity 3D [48]: contains 16 different activities acquired with a Kinect device: drink, eat, read book, call cellphone, write on a paper, use laptop, use vacuum cleaner, cheer up, sit still, toss paper, play game, lay down on sofa, walk, play guitar, stand up, sit down. If possible, the subject performs each activity in two different position sitting and standing. The acquisition environment is the same for all the sequences using a fixed camera position.

UCF-50 [49]: the dataset contains a set of 50 actions whose videos are taken from the web. The videos are characterized by random camera motion, poor lighting conditions, clutter, as well as changes in scale, appearance, and viewpoints. The actions in the dataset are very heterogeneous and some examples are tai chi, rowing, play piano, tossing balls, and biking.

UCF-101 [24]: it is an extension of the UCF-50 dataset. 51 new actions are added bringing the total action classes to 101 and the total 13,320 video clips. Each class has an average of 125 clips.

N-UCLA [50]: the dataset includes 10 actions captured from different viewpoints (usually 3) using multiple Kinect devices, 10 actors perform following actions: pick up with one hand, pick up with two hands, drop trash, walk around, sit down, stand up, donning, doffing, throw, carry.

Sports 1M [51]: this is the largest available dataset of human actions. It contains 1 million YouTube videos belonging to 487 classes. On average, 1000–3000 videos comprise each class. The video classes belong to the following macro-category: aquatic sports, team sports, winter sports, ball sports, combat sports, sports with animals.

UWA3D II [52]: contains 30 actions, mainly captured from 4 different viewpoints (front, left, right and top view) with a Kinect device in the same environments. A total of 10 subjects perform the following actions: one hand waving, one hand punching, two hand waving, two hand punching, sitting down, standing up, vibrating, falling down, holding chest, holding head, holding back, walking, irregular walking, lying down, turning around, drinking, phone answering, bending, jumping jack, running, picking up, putting down, kicking, jumping, dancing, mopping floor, sneezing, sitting down (chair), squatting, coughing.

Kinetics [25]: the dataset contains 400 human action classes, with at least 400 video clips for each action. Each clip lasts around 10 s and is taken from a different YouTube video. The videos are annotated using Amazon's Mechanical Turk. The actions cover a wide range of situations including human-object interactions such as playing instruments, as well as human-human interactions such as shaking hands.

DALY [53]: the Daily Action Localization in YouTube contains high-quality temporal and spatial annotations for 3.6k instances of 10 actions in 31 h of videos (3.3 million frames). The actions belong to 10 categories: applying makeup on lips, brushing teeth, cleaning floor, cleaning, drinking, folding textile, ironing, phoning, playing harmonica, taking photos/videos.

Charades [26]: the dataset has been created by hundreds of people recording videos in their own homes, acting out casual everyday activities. The dataset is composed of 9848 annotated videos with an average length of 30 s, showing activities of 267 people from three continents. In total, Charades provides 27,847 video descriptions, 66,500 temporally localized intervals for 157 action classes and 41,104 labels for 46 object classes.

NTU [27]: the dataset includes 60 actions for a total of 56,880 videos acquired with a Kinect v2 device. Video sequences are recorded in different environments and captured from different viewpoints. Actions are gathered under the following categories: daily actions, medical conditions and mutual actions.

As it can be seen, the datasets in the literature are different and designed with different aims in mind. There are datasets specific for a given scenario, and others which are very heterogeneous; datasets containing high-quality videos, and datasets with homemade videos; datasets with a static background and others with a dynamic background; etc. For a robust and reliable action recognition system, selecting the right dataset is crucial if we do not want to introduce any bias towards any specific acquisition condition or set of actions. Moreover, most of the datasets in the literature are not suitable for monitoring the actions of elderly people in an indoor scenario, while others contain a small selection of possible actions of interest. However, to the best of our knowledge, no public dataset contains actions specifically performed by elderly people, depicting instead either adult or young actors. This can represent a bias for elderly monitoring, as older subjects move with a different speed compared to younger people, and could perform actions in a different way. These motivations lead us to the creation of a merged dataset selecting samples from different datasets, with the purpose of better generalizing the variability of actions movements, and partially mitigating the observed representation gap. The resulting dataset is presented in the following section.

3. Dataset Definition

Individual datasets from the literature are not suitable for monitoring the actions of elderly people in an indoor scenario, either due to a limited set of classes or to inadequate environmental

conditions, as shown in Section 2.3. After having extensively analyzed the available datasets, we select only those that include the actions that we consider useful for indoor monitoring. Starting from all the datasets presented in Table 2, only five provide the chosen actions: IXMAS [45], UWA3D II [52], N-UCLA [50], MSR 3D [48] and NTU [27]. Subsequently we analyzed the properties of each action, to find any possible grouping driven by its characteristics. The end result is a composite dataset, which is characterized by a wide variety of environments, illumination conditions, acquisition devices, and relative position of camera and subject.

3.1. Action Grouping

The defined actions and relative properties are shown in Table 3. As reported, not all the actions have the same duration, alert level or movement type. Considering these characteristics, we have implemented a conceptual grouping that resulted in three different action groups. The identified action characteristics are the following: "Long" property means that actions actually can be performed in a long range of time, vice versa "Short" suggests that the actions can be executed quickly in a small amount of time. "Warning" property denotes actions that might represent a potential warning situation for the subject, the opposite "Common" represents common actions that do not show potential danger situations. "Movement" reports actions that need a partial or fully relevant body movements, its opposite "Static" means all actions that required a minimum displacement and body movements. Starting from datasets that provide the requested actions, we created three different groups of actions: Status, Alerting and Daily-life.

Table 3. Defined actions of interest, with the corresponding characteristics.

	Characteristic					
	Duration		**Type**		**Position**	
Actions	**Long**	**Short**	**Warning**	**Common**	**Movement**	**Static**
Drinking	✓			✓		
Eating	✓			✓		
Exercising	✓				✓	
Falling		✓	✓		✓	
Walking	✓			✓	✓	
Lying	✓					✓
On the floor	✓		✓		✓	✓
Reading	✓			✓		
Seated	✓			✓		✓
Coughing/sneezing		✓	✓		✓	
Standing	✓			✓		✓
Touching back		✓	✓			
Touching head		✓	✓			
Touching neck		✓	✓			
Touching torso		✓	✓			
Using phone	✓			✓		
Using laptop	✓			✓		
Vomiting	✓	✓	✓		✓	
Waving hands	✓	✓	✓		✓	
Dressing/undressing	✓			✓		

Status represents all possible poses that a subject can reach, typically are the final state reached after a movement action, in addition we insert into this group *Walking* and *On the floor* that present more common characteristics with the group's actions. This group is composed of the following classes: *Seated*, *Standing*, *Lying*, *On the floor* and *Walking*.

Alerting contains actions that need to be monitored due to potentially representing alerting or helping situations for the subject, included actions are: *Touching head*, *Touching back*, *Touching torso*, *Touching neck*, *Vomiting*, *Coughing/sneezing*, *Waving hands*, *Exercising* and *Falling*. In this group we also

inserted *Exercising* as a movement action class, because it shares common characteristics with other actions of the group.

The last group, Daily-life, contains actions that can be performed in a common relaxing indoor context (typically daily life actions), which for this reason have been grouped together. The group includes: *Drinking*, *Eating*, *Reading*, *Using phone*, *Dressing/undressing*.

The result of actions grouping and the contribution of each selected datasets are shown in Table 4.

Table 4. Number of samples provided by each dataset that compose the merged dataset.

Action	IXMAS [45]	UWA3D II [52]	N-UCLA [50]	MSR 3D [48]	NTU [27]	Total
Status						
Seated		36	54	20	948	1058
Standing	36	36	53	20	948	1093
Lying				20		20
On the floor	36	70				106
Walking	36	35	70	20		161
Alerting situations						
Touching head		36			948	984
Touching back		36			948	984
Touching torso		36			948	984
Touching neck					948	948
Vomiting					948	948
Coughing/sneezing		71			948	1019
Waving hands		36				36
Exercising		36				36
Falling		36			948	984
(Reject)	36	35	170	120	6636	6997
Daily life actions						
Drinking				20	948	968
Eating				20	948	968
Reading				20	948	968
Using phone				20	948	968
Dressing/undressing			100		1896	1996
Using laptop				20	948	968
(Reject)		287			6636	6923

3.2. Our Merged Action Dataset: ALMOND

By gathering samples from the five existing datasets that provide a meaningful contribution to our defined set of actions, we can produce an aggregated dataset that is inherently heterogeneous.

The adopted subdivision into three groups of actions implies that no action belonging to a specific group excludes those from a different group. Conversely, actions inside each group are mutually exclusive to each other, i.e., they cannot be performed at the same time. Furthermore, it is fundamental to consider the inclusion of a reject class, in order to contemplate the inference on actions that were never seen during the training of a specific group. The Status group is an exhaustive set, and as such it does not need a reject class (i.e., the subject must necessarily perform one of the involved actions). Conversely, for Alerting and Daily-life we explicitly included a sample of instances from all the other action groups. The underlying idea is to train a model that is robust with respect to collateral movements of a subject performing actions outside the action group. This is necessary for a real-life application, as most of the datasets are performed by a still person, only moving the strictly necessary body parts, and as such depicting an artificial execution of the action.

Table 4 shows the number of samples provided by each dataset for each action. As reported, the Status group is composed of the largest variety of datasets, as all datasets give samples for almost all its actions. The remaining groups are markedly bimodal, meaning that for each action the samples

are provided by two datasets only. Touching neck and vomiting are the only actions whose samples came from a single dataset (i.e., NTU [27]).

From Table 4 we observe that the number of samples from the NTU dataset greatly outnumbers those from the other datasets. This could be a potential issue, as the resulting merged dataset will be strongly unbalanced, especially in some classes, toward samples from NTU. We tried to minimize the impact of this unbalancing on the merged dataset by adopting specific criteria for re-balancing the samples: the training set was subject to per-class balancing through sample duplication, with the objective of preventing a bias on given classes. The test set was balanced instead on the different datasets, with the objective of producing a heterogeneous and thus significant benchmark. More specifically, for the test set, we selected from each dataset 10 samples for each action. Notice that multiple actions given by one or more datasets can be used to compose a single action of our merged dataset. For the training set the applied rule was the previous one, except that we took the available samples (after excluding those used in test set) up to a maximum of 25. With this operation the training set still results unbalanced, so we balanced each action by randomly duplicating the inner samples to reach the same number of samples of the most populated action in the group. With this method we obtained a balanced and heterogeneous training set, and an unbalanced and heterogeneous test set. The test set unbalancing is not an issue since the performance will not be computed globally but over each class. The dataset was the result of conducted experiments that will be detailed in Section 5.

The final dataset, called ALMOND (Assisted Living MONitoring Dataset) is made available for public download [54], and will be used for all the following experiments for action recognition. Classes population details are shown in Table 5, with number of samples of test and training set respectively of 790 and 6775 for a total of 7565. The acquisition resolutions of ALMOND vary from 320 × 240 pixels of images coming from the UWA3DII dataset, to 1920 × 1080 pixels from NTU. N-UCLA and MSR3D have the camera positioned at eye-level, while UWA3D II and NTU have acquisitions at both ceiling-level and eye-level, and the IXMAS is only at ceiling-level. All datasets depict the subject from different sides.

Table 5. ALMOND dataset training and test set division, with classes cardinality.

Action	Train	Test
Seated	110	40
Standing	110	50
Walking	110	40
Lying	110	10
On the floor (seated/lying)	110	30
Touching head	360	20
Touching back	360	20
Touching torso	360	20
Touching neck	360	10
Vomiting	360	10
Coughing/sneezing	360	30
Waving hands	360	10
Exercising	360	10
Falling	360	20
Reject	360	180
Drinking	375	20
Eating	375	20
Reading	375	20
Using phone	375	20
Dressing/undressing	375	40
Using laptop	375	20
Reject	375	150
TOTAL	6775	790

This variety in the acquisition setup creates the conditions for high generalization capabilities in recognizing the performed action in different environments.

4. Proposed Monitoring Approach

In this section, we will define our approach for assisted living monitoring, leveraging on the sets of actions defined in Section 3.1. The proposed solution, shown in Figure 4, is composed of two processing steps:

1. Localizing the monitored subject inside the scene
2. Recognizing the action performed by the subject.

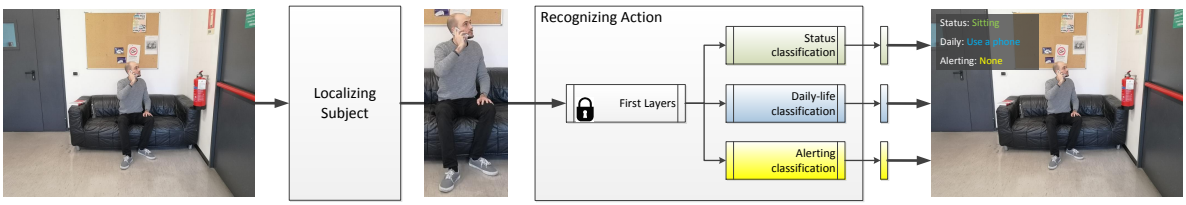


Figure 4. Schema of the proposed monitoring approach. Each processing step is implemented with algorithms based on Deep Learning. Subject localization is based on the Faster R-CNN network. For action recognition two networks are evaluated: I3D and DeepHar. We modified these networks to take advantage of a multi-branch approach where groups of actions and states are recognized separately.

Each individual step, as well as their joint application, will be rigorously evaluated in Section 5 and integrated in a final system as presented in Section 6.

4.1. Localization of the Subject

People detection has been successfully addressed in the past, with state-of-the-art models reaching excellent performance in terms of Average Precision (AP) [55]. Faster R-CNN (Regions with CNN features) is considered a state-of-the-art object detector, successfully applied to the detection of human subjects [56]. It falls in the category of two-stage neural detection models. During the first stage, a list of object proposals is generated in the form of bounding boxes coordinates. In the second stage, the neural features corresponding to each proposed region are brought to common size and classified into a defined set of classes. For our solution, we trained the detection model focusing on only the "person" class, which makes the classification stage reject false positive detections that were generated at the proposal stage. At inference time, we resort to selecting only the largest detected subject, which is supposed to be the closest to the camera, for the subsequent step of action recognition. In future developments, we will consider the integration of person re-identification techniques [57] to ensure the execution of a proper analysis.

As we will show with proper experiments (Section 5.2), however, the same models quickly degrade in performance when tasked with detecting subjects in horizontal position or lying down. We consider this a critical aspect for our application, where correctly detecting the presence of a lying subject is a potential trigger for an alerting situation. Based on the hypothesis that this behavior is not an inherent limit of people detection models, but lies in the training data used for learning, we used digitally rotated images for training and testing the people detector. We chose this strategy to exploit the inherent richness and high cardinality of existing datasets for people detection, which however contain little to no samples of people lying down.

4.2. Recognition of the Action

In Section 2.2 we have presented a synthetic overview of existing methods for action recognition. In the following, we focus our attention on two of such methods for possible integration in our

system for assisted living: I3D [33] and DeepHAR [42] (Deep Human Action Recognition). The I3D model has produced groundbreaking results in challenging domains such as the Kinetics datasets [25]. Given the large cardinality (up to 700 classes in the latest release), such a well-performing model is expected to successfully discriminate between subtle differences in very similar classes, a characteristic that we particularly cherish for the Daily-life action group. DeepHAR has been recently proposed as a model that infers the depicted action through an explicit representation of the subject inferred skeleton. We therefore expect this method to perform especially well for full-body actions such as those belonging to the Status and Alerting groups.

Each action group presented in Section 3.1 has been designed to be independent from the others. In practice, this translates to the creation of three separate action recognition models, continuously processing a common stream of data when the system is running live. Although our proposed system does not have the constraint of real-time processing, a short-term response is indispensable, as the detection of potentially alerting situations must allow for timely intervention. To reduce the overall computational burden at inference time, we adopted the following approach during training:

1. We pre-trained the action recognition model on a wide dataset, characterized by high cardinality both in terms of classes and examples. For DeepHAR we pre-trained on the NTU dataset [27] as suggested by the authors, while I3D was pre-trained on the Kinetics-400 dataset.
2. We fine-tuned three models (one per action group) on our aggregated dataset, freezing the gradient backpropagation in the first layers. Specifically, we blocked all pose estimation layers in DeepHAR, and all layers before and including the fourteenth layer of I3D, which is a $2 \times 2 \times 2$ 3D max-pooling layer.
3. We combine the three models into a unique multi-objective neural network, which performs an initial common processing and then eventually branches out into three independent paths.

This approach is inherited from the concept of transfer learning itself [58], where low-level features extracted in the first layers are hypothesized to represent pieces of information that can be exploitable throughout the entire domain.

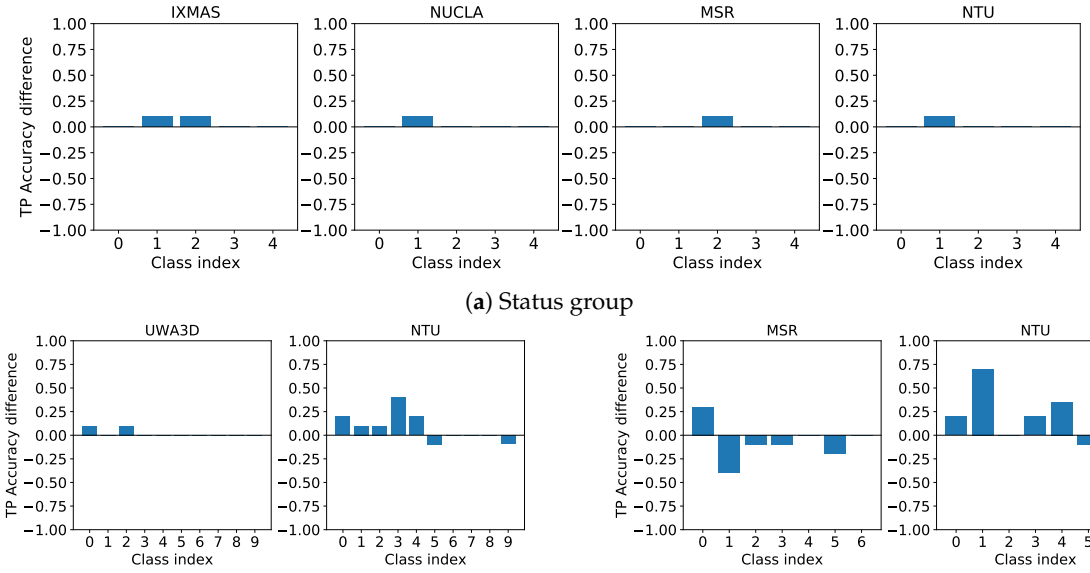
5. Evaluation of the Proposed Monitoring Approach

In this section, we present the results of the experiments we have conducted on the processing steps described in Section 4: from tackling the problem of detecting the subject in various poses using the COCO dataset [55] Common Objects in COntext, up to seeking the best architecture for action classification using our ALMOND dataset. We also conduct preliminary experiments to understand if the inclusion of more samples from NTU dataset can be useful with the scope of creating an ALMOND with more variance for models training.

5.1. Effects of Unbalanced Datasets

As presented in Section 3 it is possible to observe that ALMOND uses only a small portion of samples provided by NTU [27], whose cardinality dominates the other datasets, as visible in Table 4. We explore deeper this issue with some preliminary experiments, our intent was investigating if the inclusion of a larger number of samples could create a better dataset. Performed experiments were structured comparing ALMOND, presented in Section 3, and its version with two edits, the former including 200 samples more from NTU, the latter including 400 NTU samples. Extra samples have been added only for classes where NTU provided it. In all the cases the dataset has been balanced by actions, except for the two reject classes, duplicating samples of each action to reach the samples number of group's most populated action, as indicated in Section 3.1. As experiments results, we observed that an increment of samples from NTU did not lead to better performance or a richer dataset in terms of provided information. Figure 5 shows the true positive accuracy difference for each datasets that form ALMOND, missing datasets mean there were no changes in terms of accuracy, the positive values mean the percentage of increment in action recognition, otherwise negative values mean worse

performance. We can notice that an increment of NTU samples corresponds to worse generalization capacity for the model, giving only a relevant increase of accuracy on samples provided by NTU. With the inclusion of 400 NTU samples the behavior is the same as with 200, but even more pronounced. Generally, increasing the number of samples from a single dataset does not lead to a merged dataset with more useful information, rather shows a loss of heterogeneity. Facing the results of preliminary experiments, we use ALMOND as described in Section 3 without the inclusion of additional samples from NTU.



(**a**) Status group

(**b**) Alerting group

(**c**) Daily-life group

Figure 5. True prediction accuracy difference for each group, between standard merged dataset and the same dataset with the addition of 200 samples from NTU for each action where provided. Charts omitted if there are no differences for the specific dataset. Positive values represent the percentage increment of accuracy, negative values the deterioration of performance.

5.2. Results on Subject Localization

To localize the subject inside the scene we used a Faster R-CNN network [56] pre-trained on the COCO dataset [55]. We noticed an increment of missing detections when subjects are lying down, or assume a horizontal position. Exploring more in detail the COCO dataset, we noticed a lack of samples of lying down subjects compared to people in other common positions. This led us to go deeper into the issue, by investigating if a fine-tuning, on the same dataset extended with digitally rotated by $\pm 90^{\circ}$ images, can lead to better subject detection. Table 6 reports the comparison between pre-trained Faster R-CNN and the fine-tuned one. Results show that the detection performance are comparable for unrotated images, where a few lying down people is present, while the fine-tuned model provides better results with rotated images. Taking into account the results, the fine-tuned model correctly detect more lying people and provides more stable detection in those cases, in other cases performance are comparable to pre-trained version.

Table 6. People detectors performance, comparison between pre-trained detector and fine-tuned.

Model	AP@0.5	AP
	2cUnrotated	
Faster R-CNN orig.	0.819	0.516
Faster R-CNN fineT.	0.801	0.497
	2cRotated +90	
Faster R-CNN orig.	0.246	0.106
Faster R-CNN fineT.	0.651	0.345
	2cRotated -90	
Faster R-CNN orig.	0.244	0.105
Faster R-CNN fineT.	0.648	0.341

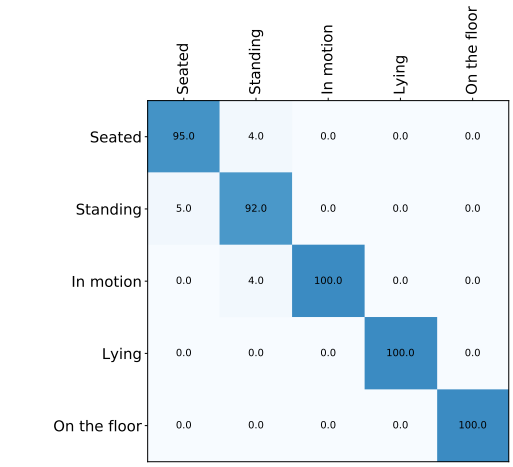
5.3. Results on Action Recognition

First, we trained and tested the model with a fixed crop position, cropping a center square image portion for each frame of the sequences. Subsequently, we used a people detector to center the crop on the subject. We explored this matter to understand if the use of a person detector can help the models to reach better classification performance. Results in Table 7 shows that the use of a person detector allows models to achieve a better action classification. Most relevant performance increment is in Alerting group, follow by Daily-life, as expected the increment for Status is very small, this could be due to already high performance without the use of a people detector. As reported, a bigger increment of performance is obtained in Alerting and Daily-life, where the number of the classes is higher and the difference between some actions is concentrated in small details or minor movements of subjects, faced the results the use of a people detector is desirable.

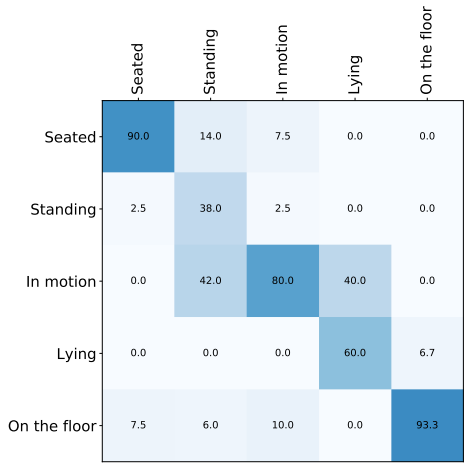
Table 7. Results of average per-class accuracy for compared architectures. We conduct experiments with fixed central crop and with the use of people detector for both selected models.

Subject Localization	Action Group	I3D [33]	DeepHAR [42]
Fixed central crop	Status	0.947	0.664
	Alerting	0.770	0.518
	Daily-life	0.630	0.446
People detector	Status	0.974	0.723
	Alerting	0.829	0.613
	Daily-life	0.715	0.498

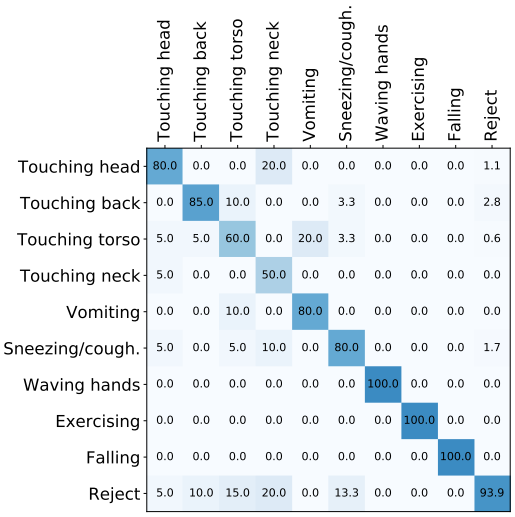
In parallel, we performed experiments comparing the two different architectures, presented before in Section 4.2. The results highlight that I3D with its architecture based on 3D convolutions made it possible to learn properly spatio-temporal features, reaching good performance in action classification. This means that the expansion of 2D image classification models to 3D convolutional networks allows learning efficiently spatio-temporal relations. I3D performs better in all conditions, reaching the best performance on ALMOND when combined with the subject detector. As speculative observation 3D convolution architecture could perform better than DeepHAR because the latter cannot reach the same expressive power and it is potentially limited by the pose estimation. It is worth noticing that also for the I3D architecture it results difficult to discriminate actions with similar movements and small differences, like those appearing in *Drinking* and *Eating*. This behavior can be mainly noticed into Daily-life group, and reported in confusion matrix in Figure 6. In conclusion, I3D architecture obtains the best results for action classification; however it is not free from misclassification problems also with the use of a people detector. As shown in Figure 6e there is some confusion between the actions *Drinking*, *Eating* and *Using phone*, which differ only for the object that the subject held on hand. The same situation is visible for *touching* actions in Alerting group.



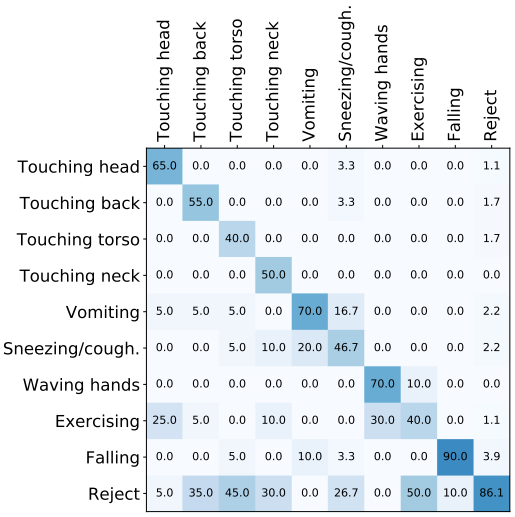
(**a**) I3D Status group



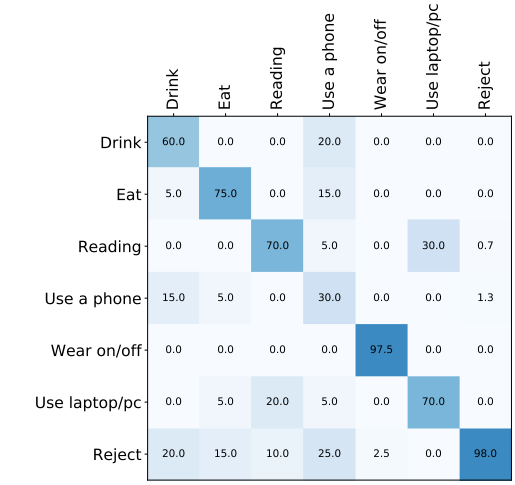
(**b**) DeepHAR Status group



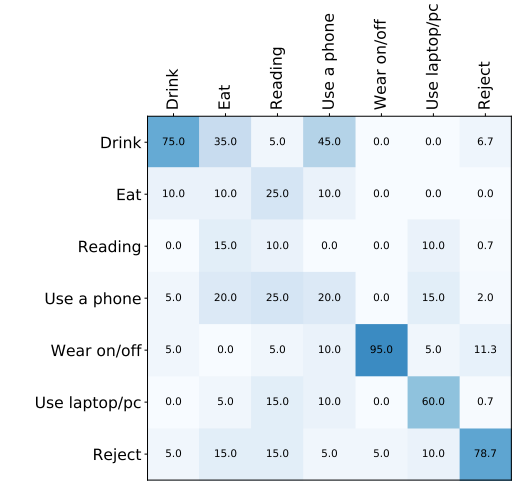
(**c**) I3D Alerting group



(**d**) DeepHAR Alerting group



(**e**) I3D Daily-life group



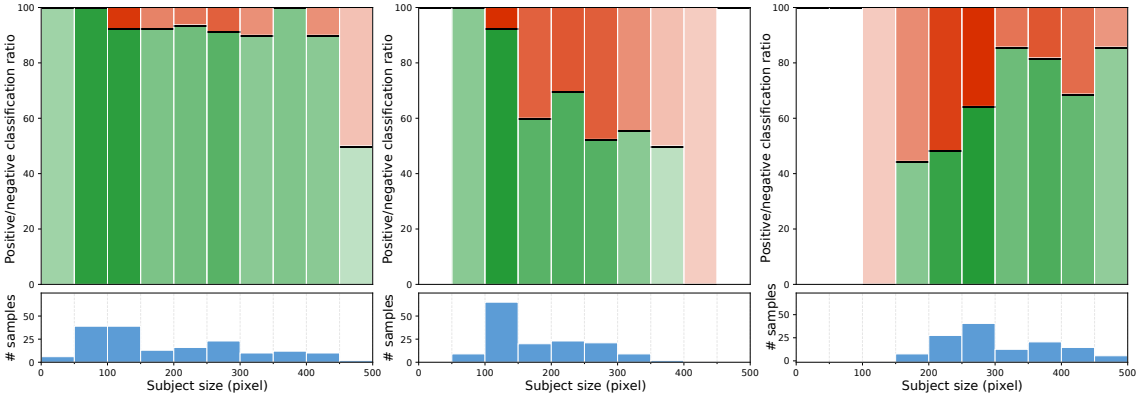
(**f**) DeepHAR Daily-life group

Figure 6. Action classification confusion matrix for each defined actions group. Left column refers to I3D, right to DeepHAR. Results refer to experiments with the people detector.

5.4. Analysis of Environment Setup Constraints

In the following, we assess the dependence of classification accuracy on the apparent size of the subject in the image frame. We hypothesize a drop in performance of action recognition when the subject's image is not large enough. The objective is to quantify this intuition, to eventually lay out some guidelines on the allowed distance between camera and monitored subject.

The test set instances are partitioned according to the size (in pixels) of the detected subject, and each block is evaluated in terms of average accuracy. Figure 7 presents the results of this analysis: the accuracy is reported as a stacked bar plot showing the ratio between correctly identified examples and misclassifications. The analysis could be biased by the effective number of samples in each block, as fewer data lead to less reliable estimates. To compensate for this, we also report the samples distribution, encoded both in the opacity of the stacked bar plot, as well as a separate bar plot.



(**a**) Status group (**b**) Alerting group (**c**) Daily-life group

Figure 7. Dependence of classification accuracy on the apparent size of the subject in the image frame.

The Status group shows excellent results at all scales, therefore not putting any concrete constraint on subject size. The Alerting group presents relatively stable results for all size blocks for which there is a significant amount of data, i.e., from 100 pixels on. Finally, the Daily-life group establishes a possible constraint on the subject size to be at least 300 pixels, to guarantee a high enough recognition level.

By knowing or estimating the intrinsic parameters of an acquisition device, it is possible to exploit the following correspondence between the subject's apparent size in pixels, and their distance from the camera [59]:

$$distance = \frac{real\_size}{apparent\_size}\left(\frac{image\_size}{sensor\_size}F_{metric}\right) = \frac{real\_size}{apparent\_size}F \quad (1)$$

As an example, with a consumer device such as the Microsoft LifeCAM HD 3000 working at 1280×800 pixels resolution, the estimated focal length F is 1165 pixels. If we assume an average person to be 1.65m tall (considering the average between mean height of adult males and females in year 2014, from a global study published by NCD Risk Factor Collaboration in 2016 [60]), putting a constraint on the subject being at least 300 pixels large means imposing the camera to be no farther than 6m from the monitored subject.

6. Design of the Monitoring System

In this section, we design and present a system for monitoring elderly subjects, based on the proposed monitoring approach, set of actions, collected datasets, and trained action recognition model, described in the previous Sections. We will start by defining the high-level goals of such system, in order to gradually refine them in terms of features of a client-server application.

The typical end-user, called a "guardian" in the following, is a person in charge of monitoring a subject who can take care for him/her-self, but who is at high risk of domestic accident when living alone. The main goal of our application is to give the guardian an effective exploration of the events regarding the monitored subject. Specifically, the application should satisfy the following set of requirements:

1. A quick and intuitive way to reach the desired information.
   If the guardian has a precise idea of what he/she wants (a specific timestamp for example).
2. A synthetic yet exhaustive abstraction of the detected events.
   If the guardian is broadly exploring a given time range.
3. Timely notifications for situations of interest.
   The application should actively reach the guardian when specific conditions are met.

These requirements are fulfilled by the designed client-server application, whose architecture is shown in Figure 8. In the envisioned scenario, the monitored subject is recorded through a camera installed in the chosen environment. The recorded video is sent to the server, where it is processed for action recognition. The guardian accesses a client web application, and uses it to request to the server the desired analysis. The web application renders the video stream and the corresponding analysis, both received by the server (requirements 1 and 2). If the server-side processing triggers a condition of interest, a notification is sent to the guardian (requirement 3).

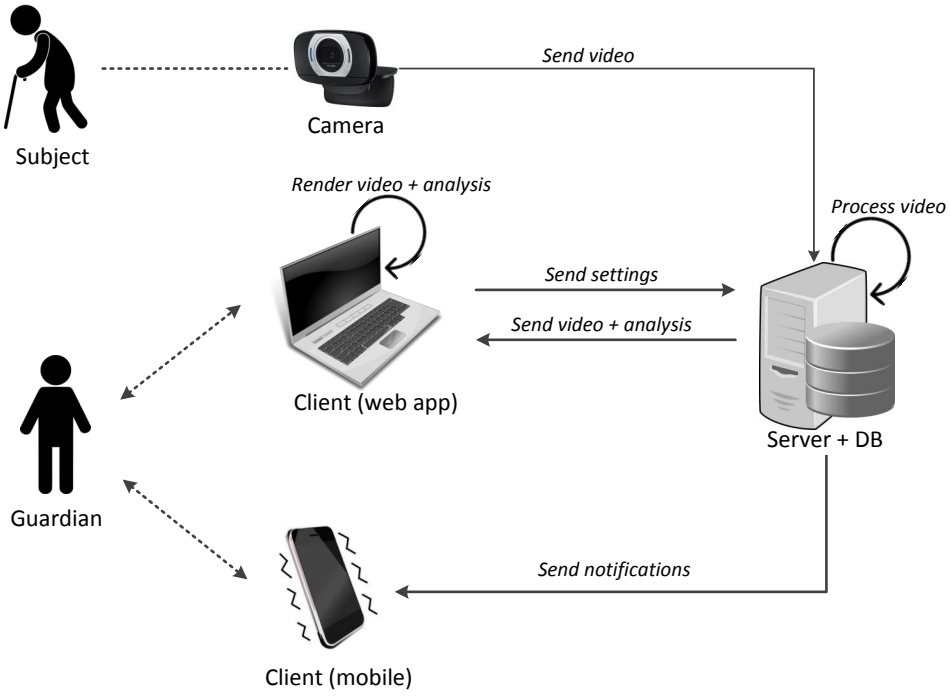


Figure 8. Architecture of our client-server application for assisted living monitoring.

The high-level requirements are concretely implemented with the following set of tools:

- Coarse-to-fine temporal navigation (time range and timeline view)
- Instant-level description of the subject behavior (snapshot)
- Automatic events partitioning and description (storyboard view)
- Global statistics for the selected time range (global view)
- User-customizable filters and notifications

These tools are shown in Figure 9 in context of the designed web application, and are further described in the following.

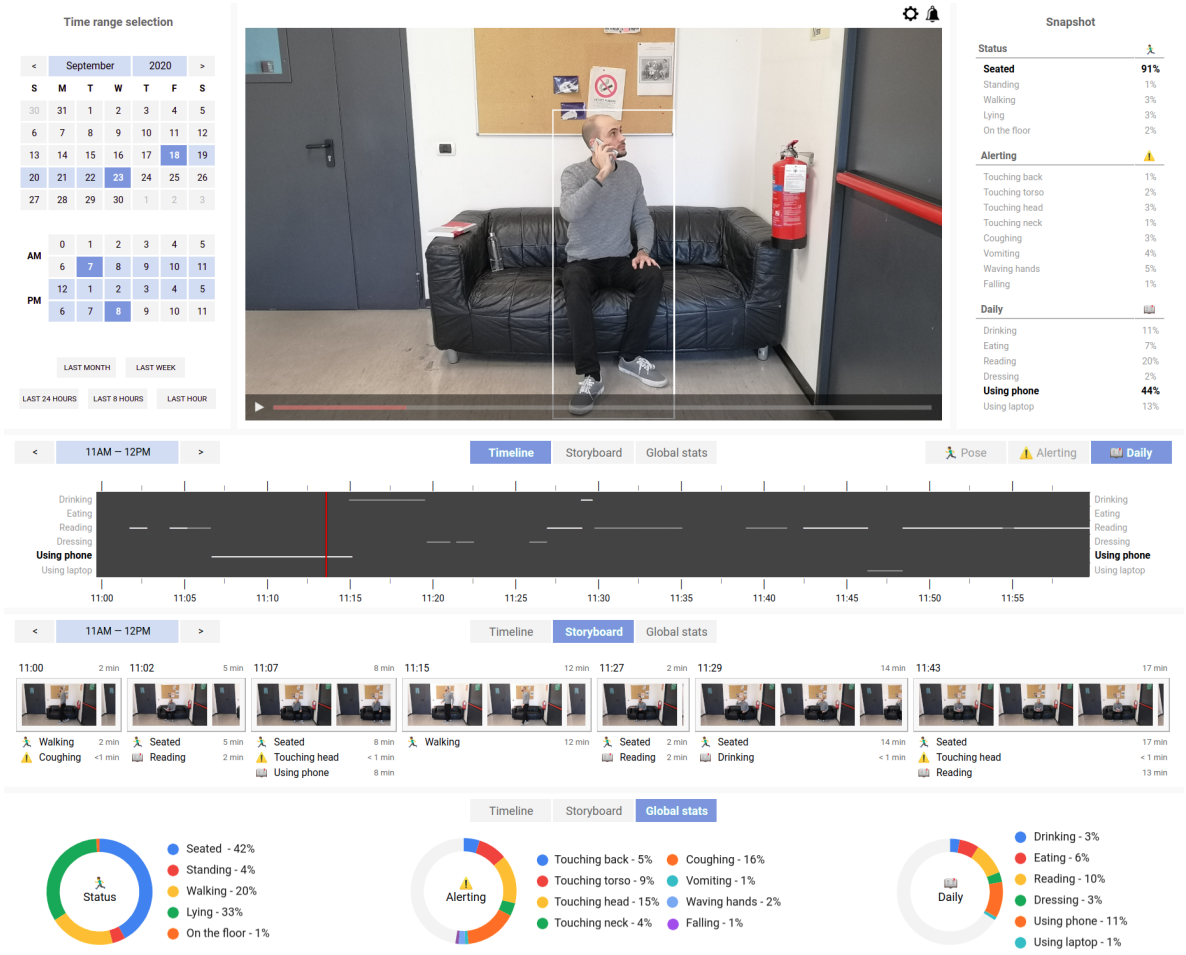


Figure 9. Interface of the client-side web application. After the time period of interest is selected, the registered actions can be browsed and visualized using different modalities: in a timeline, as a visual storyboard, or as global statistics. The user can also setup alarms to be automatically notified.

Since the guardian can watch the subject's videos, an important matter that must be kept in consideration for this kind of applications is the privacy of the subject. The proposed system was designed to be used in a private setting, keeping in mind that there are different laws for different countries which impose different rules. Assuming the guardian can see the monitored subject, the problem of guest privacy can be instead addressed with the use of people identification approaches, anonymizing all the individuals in the scene that are not the monitored subject. An investigation into the most suitable and performant methods will be carried out in future works.

6.1. Time Range and Timeline View

The guardian can browse the video playback (and corresponding analysis) through a hierarchical navigation system. At the highest level, a time-range selector is used to specify the days and hours of interest. A typical use case would have the guardian access the application at late afternoon for a recap of the events of the day. As this broad time range is selected, the other elements in the application render information accordingly: the timeline view, the storyboard view, and the global view. The timeline displays a linear representation of the recognized actions: for any action group selected by the guardian, all detected actions are reported as a horizontal line, whose intensity is related to the recognition confidence. Finally, the timeline also serves as a browsing tool for video playback.

6.2. Snapshot

The snapshot presents an instant-level description of the video frame currently displayed. All action groups are visible at the same time, and each action is accompanied by the associated confidence derived from the recognition model.

6.3. Storyboard View

The storyboard is, at high level, a partitioning of the time range selected by the guardian. For the purpose of our application, each part should be describable in a concise way (i.e., it should not encompass an excessive number of events), but, at the same time, the total number of parts should be bounded. Striking a good balance between such constraints would successfully prevent an overload of information to the guardian. In practice, this can be obtained by generating the largest parts such that each action group has at most one occurrence of any sub-action in every part. Further reduction of information burden can be obtained by excluding uninteresting actions based on user preferences, and by considering low-confidence events only when they appear close to high-confidence events of the same class. This approach is formally described in Algorithm 1.

Algorithm 1 Storyboard partitioning

INPUT: $A = \{a_g(t)\}$ ▷ Recognized actions, for each action group g at timestamp t
INPUT: $C = \{c_g(t)\}$ ▷ Recognition confidence values, for each action group g at timestamp t
OUTPUT: $S = \{s_p\}$ ▷ Starting timestamps, for each part p
OUTPUT: $D = \{d_p\}$ ▷ Descriptions, for each part p

for all timestamps t **do** ▷ Selectively propagate high-confidence recognitions
  for all action groups g **do**
    if $c_g(t) = LOW$ **then** ▷ If recognition has low confidence
      if $c_g(t-1) = HIGH$ & $a_g(t-1) = a_g(t) \mid c_g(t+1) = HIGH$ & $a_g(t+1) = a_g(t)$ **then** ▷ If neighbor recognition is high confidence of the same class
        $c_g(t) \leftarrow HIGH$ ▷ Propagate high confidence to current recognition
      end if
    end if
  end for
end for

$S = \emptyset$ ▷ Initialize partitioning
$D = \emptyset$
s', d', $L \leftarrow$ INIT_PARTITION(0) ▷ Create first part

for all timestamps t **do** ▷ Perform partitioning
  for all action groups g **do**
    if $c_g(t) = HIGH$ **then** ▷ If a high-confidence action is found
      if $l_g = -1$ **then** ▷ If it is the first class found
        $d' \leftarrow d' \cup a_g(t)$ ▷ Add class to part description
        $l_g \leftarrow a_g(t)$ ▷ Update last-seen action
      else
        if $a_g(t) \neq l_g$ **then** ▷ If it is a new, different class
          $S \leftarrow S \cup \{s'\}$ ▷ Add current part to output sets
          $D \leftarrow D \cup \{d'\}$
          s', d', $L \leftarrow$ INIT_PARTITION(t) ▷ Create new part
        end if
      end if
    end if
  end for
end for

function INIT_PARTITION(t)
  $s' \leftarrow t$ ▷ Set starting timestamp
  $d' \leftarrow \emptyset$ ▷ Initialize empty description
  $L \leftarrow \emptyset$ ▷ Initialize dummy last-seen action, for each action group g
  for all action groups g **do**
    $l_g \leftarrow -1$
    $L \leftarrow L \cup \{l_g\}$
  end for
  return s', d', L
end function

6.4. Global View

The global view serves the purpose of providing a synthetic summary of the events, just like the storyboard view, but operates at a higher abstraction level. It consists of a pie chart for each action group, which renders information relative to the entire time range selected by the guardian.

6.5. User-Customizable Filters and Notifications

As an additional tool for browsing and reducing the information burden, we provide the guardian with the possibility of filtering the set of actions recognized by our system. For example, one specific user might decide that the set of actions relative to touching different body parts should not influence the partitioning process involved in the storyboard view.

In a similar fashion, we allow the guardian to create his/her own set of rules for triggering alarms. This is addressed with three types of trigger: a one-shot event (e.g., the act of falling), a time-range event (e.g., being on the floor for a certain amount of time), and a time-range absence (e.g., not drinking for a long time). More elaborate conditions could be implemented by resorting to dedicated high-level programming languages [61], although we consider expanding this topic in future works.

7. Conclusions

We have developed a monitoring system for taking care of elderly people, and for fulfilling their right to aging in place. By approaching the task as a machine-learning problem, we have carefully analyzed existing datasets for action recognition, and concluded that no single dataset matched the required criteria in terms of classes and cardinality. This led us to the definition of a new hierarchy of actions, and to the creation of a corresponding composite dataset, called ALMOND. We then developed and trained a monitoring approach that consists of localizing the subject, and recognizing the performed actions among the defined set. We aimed at reaching high accuracy in a wide range of subject poses, while keeping the computational effort under control. Finally, we presented the end-user application to be exploited by an assigned guardian, defining its functional requirements and designing its main components. Particular attention was given to offering the guardian an effective exploration of the events regarding the monitored subject, without overloading them with information. The evaluation of usability of the developed system will be addressed in the near future.

As further direction for future work, we plan on introducing techniques for person re-identification [57], in order to allow our system to track the subject across multiple acquisition devices, and to be robust to the presence of healthcare assistants. Consequently, we will also expand the set of recognized classes to interactions between two or more subjects. Other developments would include alternative forms of storyboarding, leaning for example on video summarization techniques [62,63], as well as more advanced forms of user-customizable trigger conditions for alerting situations.

Author Contributions: Conceptualization, M.B. and G.C.; methodology, M.B., A.A. and G.C.; software, M.B. and A.A.; validation, A.A.; formal analysis, M.B. and A.A.; investigation, A.A.; resources, M.B. and G.C.; data curation, A.A.; writing–original draft preparation, M.B., A.A. and G.C.; writing–review and editing, M.B., A.A. and G.C.; visualization, M.B., A.A. and G.C.; supervision, M.B. and G.C.; project administration, M.B. and G.C. All authors have read and agreed to the published version of the manuscript.

Funding: The research leading to these results has received funding from The Home of Internet of Things (Home IoT), CUP (Codice Unico Progetto - Unique Project Code): E47H16001380009 - Call "Linea R&S per Aggregazioni" cofunded by POR (Programma Operativo Regionale - Regional Operational Programme) FESR (Fondo Europeo di Sviluppo Regionale - European Regional Development Fund) 2014–2020.

Acknowledgments: We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Jetson TX1 Kit and the Titan Xp GPU used for this research.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. United Nations - Department of Economic and Social Affairs. World Population Prospects 2019 - Highlights. 2019. Available online: https://population.un.org/wpp/Publications/Files/WPP2019_Highlights.pdf (accessed on 31 December 2019).
2. European Commission - Economic and Financial Affairs. The 2018 Ageing Report. 2018. Available online: https://www.age-platform.eu/publications/ageing-report-2018 (accessed on 31 December 2019).
3. United Nations - Department of Economic and Social Affairs. World Population Prospects 2019 - Download Center. 2019. Available online: https://population.un.org/wpp/Download/Standard/Population/ (accessed on 31 December 2019).
4. Mazzola, P.; Rimoldi, S.M.L.; Rossi, P.; Noale, M.; Rea, F.; Facchini, C.; Maggi, S.; Corrao, G.; Annoni, G. Aging in Italy: The Need for New Welfare Strategies in an Old Country. *Gerontologist* **2015**, *56*, 383–390.
5. AAL Association. AAL Home 2020 - AAL Programme. 2019. Available online: https://www.aal-europe.eu/ (accessed on 31 December 2019).
6. Al-Shaqi, R.; Mourshed, M.; Rezgui, Y. Progress in ambient assisted systems for independent living by the elderly. *SpringerPlus* **2016**, *5*, 624.
7. Majumder, S.; Aghayi, E.; Noferesti, M.; Memarzadeh-Tehran, H.; Mondal, T.; Pang, Z.; Deen, M.J. Smart homes for elderly healthcare—Recent advances and research challenges. *Sensors* **2017**, *17*, 2496.
8. Uddin, M.; Khaksar, W.; Torresen, J. Ambient sensors for elderly care and independent living: A survey. *Sensors* **2018**, *18*, 2027.
9. Mshali, H.; Lemlouma, T.; Moloney, M.; Magoni, D. A survey on health monitoring systems for health smart homes. *Int. J. Ind. Ergon.* **2018**, *66*, 26–56.
10. Jalal, A.; Kamal, S.; Kim, D. A depth video sensor-based life-logging human activity recognition system for elderly care in smart indoor environments. *Sensors* **2014**, *14*, 11735–11759.
11. Susnea, I.; Dumitriu, L.; Talmaciu, M.; Pecheanu, E.; Munteanu, D. Unobtrusive Monitoring the Daily Activity Routine of Elderly People Living Alone, with Low-Cost Binary Sensors. *Sensors* **2019**, *19*, 2264.
12. Motiian, S.; Siyahjani, F.; Almohsen, R.; Doretto, G. Online human interaction detection and recognition with multiple cameras. *IEEE Trans. Circuits Syst. Video Technol.* **2016**, *27*, 649–663.
13. Malasinghe, L.P.; Ramzan, N.; Dahal, K. Remote patient monitoring: a comprehensive study. *J. Ambient Intell. Humaniz. Comput.* **2019**, *10*, 57–76.
14. Botia, J.A.; Villa, A.; Palma, J. Ambient Assisted Living system for in-home monitoring of healthy independent elders. *Expert Syst. Appl.* **2012**, *39*, 8136–8148.
15. Bourouis, A.; Feham, M.; Bouchachia, A. Ubiquitous mobile health monitoring system for elderly (UMHMSE). *arXiv* **2011**, arXiv:1107.3695.
16. Huo, H.; Xu, Y.; Yan, H.; Mubeen, S.; Zhang, H. An elderly health care system using wireless sensor networks at home. In Proceedings of the 2009 Third International Conference on Sensor Technologies and Applications, Athens, Glyfada, Greece, 18–23 June 2009; pp. 158–163.
17. Daher, M.; Diab, A.; El Najjar, M.E.B.; Khalil, M.A.; Charpillet, F. Elder tracking and fall detection system using smart tiles. *IEEE Sens. J.* **2016**, *17*, 469–479.
18. Nasution, A.H.; Zhang, P.; Emmanuel, S. Video surveillance for elderly monitoring and safety. In Proceedings of the TENCON 2009-2009 IEEE Region 10 Conference, Singapore, 23–26 January 2009; pp. 1–6.
19. Lu, K.L.; Chu, E. An Image-Based Fall Detection System for the Elderly. *Appl. Sci.* **2018**, *8*, 1995.
20. Simonyan, K.; Zisserman, A. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: New York, NY, USA, 2014; pp. 568–576.
21. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning spatiotemporal features with 3D convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Las Condes, Chile, 11–18 December 2015; pp. 4489–4497.
22. Donahue, J.; Anne Hendricks, L.; Guadarrama, S.; Rohrbach, M.; Venugopalan, S.; Saenko, K.; Darrell, T. Long-term recurrent convolutional networks for visual recognition and description. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 2625–2634.
23. Kuehne, H.; Jhuang, H.; Garrote, E.; Poggio, T.; Serre, T. HMDB: A large video database for human motion recognition. In Proceedings of the 2011 International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 2556–2563.

24. Soomro, K.; Zamir, A.R.; Shah, M. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv* **2012**, arXiv:1212.0402.
25. Kay, W.; Carreira, J.; Simonyan, K.; Zhang, B.; Hillier, C.; Vijayanarasimhan, S.; Viola, F.; Green, T.; Back, T.; Natsev, P.; et al. The kinetics human action video dataset. *arXiv* **2017**, arXiv:1705.06950.
26. Sigurdsson, G.A.; Varol, G.; Wang, X.; Farhadi, A.; Laptev, I.; Gupta, A. Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; pp. 20–36.
27. Shahroudy, A.; Liu, J.; Ng, T.T.; Wang, G. NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016.
28. Yue-Hei Ng, J.; Hausknecht, M.; Vijayanarasimhan, S.; Vinyals, O.; Monga, R.; Toderici, G. Beyond short snippets: Deep networks for video classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 4694–4702.
29. Feichtenhofer, C.; Pinz, A.; Zisserman, A. Convolutional two-stream network fusion for video action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1933–1941.
30. Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y.; Lin, D.; Tang, X.; Van Gool, L. Temporal segment networks: Towards good practices for deep action recognition. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; pp. 20–36.
31. Liu, J.; Shahroudy, A.; Xu, D.; Wang, G. Spatio-temporal LSTM with trust gates for 3D human action recognition. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; pp. 816–833.
32. Varol, G.; Laptev, I.; Schmid, C. Long-term temporal convolutions for action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 1510–1517.
33. Carreira, J.; Zisserman, A. Quo vadis, action recognition? A new model and the kinetics dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,Honolulu, HI, USA, 21–26 July 2017; pp. 6299–6308.
34. Diba, A.; Fayyaz, M.; Sharma, V.; Karami, A.H.; Arzani, M.M.; Yousefzadeh, R.; Van Gool, L. Temporal 3D convnets: New architecture and transfer learning for video classification. *arXiv* **2017**, arXiv:1711.08200.
35. Qiu, Z.; Yao, T.; Mei, T. Learning spatio-temporal representation with pseudo-3D residual networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 5533–5541.
36. Sun, L.; Jia, K.; Chen, K.; Yeung, D.Y.; Shi, B.E.; Savarese, S. Lattice Long Short-Term Memory for Human Action Recognition. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.
37. Song, S.; Lan, C.; Xing, J.; Zeng, W.; Liu, J. Spatio-temporal attention-based LSTM networks for 3D action recognition and detection. *IEEE Trans. Image Process.* **2018**, *27*, 3459–3471.
38. Zhang, B.; Wang, L.; Wang, Z.; Qiao, Y.; Wang, H. Real-time action recognition with deeply transferred motion vector CNNs. *IEEE Trans. Image Process.* **2018**, *27*, 2326–2339.
39. Tran, D.; Wang, H.; Torresani, L.; Ray, J.; LeCun, Y.; Paluri, M. A closer look at spatiotemporal convolutions for action recognition. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6450–6459.
40. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7794–7803.
41. Li, Z.; Gavrilyuk, K.; Gavves, E.; Jain, M.; Snoek, C.G. VideoLSTM convolves, attends and flows for action recognition. *Comput. Vision Image Underst.* **2018**, *166*, 41–50.
42. Luvizon, D.C.; Picard, D.; Tabia, H. 2D/3D pose estimation and action recognition using multitask deep learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5137–5146.
43. Ghadiyaram, D.; Tran, D.; Mahajan, D. Large-scale weakly-supervised pre-training for video action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 12046–12055.

44. Yan, A.; Wang, Y.; Li, Z.; Qiao, Y. PA3D: Pose-Action 3D Machine for Video Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 7922–7931.
45. Weinland, D.; Ronfard, R.; Boyer, E. Free viewpoint action recognition using motion history volumes. *Comput. Vision Image Underst.* **2006**, *104*, 249–257.
46. Rodriguez, M.D.; Ahmed, J.; Shah, M. Action MACH a spatio-temporal Maximum Average Correlation Height filter for action recognition. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 23–28 June 2008; Volume 1, p. 6.
47. Marszałek, M.; Laptev, I.; Schmid, C. Actions in Context. In Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition, Miami Beach, FL, USA, 20–25 June 2009.
48. Wang, J.; Liu, Z.; Wu, Y.; Yuan, J. Mining actionlet ensemble for action recognition with depth cameras. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 1290–1297.
49. Reddy, K.K.; Shah, M. Recognizing 50 human action categories of web videos. *Mach. Vis. Appl.* **2013**, *24*, 971–981.
50. Wang, J.; Nie, X.; Xia, Y.; Wu, Y.; Zhu, S.C. Cross-view action modeling, learning and recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 24–27 June 2014; pp. 2649–2656.
51. Karpathy, A.; Toderici, G.; Shetty, S.; Leung, T.; Sukthankar, R.; Fei-Fei, L. Large-scale Video Classification with Convolutional Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 24–27 June 2014.
52. Rahmani, H.; Mahmood, A.; Huynh, D.; Mian, A. Histogram of oriented principal components for cross-view action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 2430–2443.
53. Weinzaepfel, P.; Martin, X.; Schmid, C. Human Action Localization with Sparse Spatial Supervision. *arXiv* **2016**, arXiv:1605.05197.
54. Imaging and Vision Laboratory. Monitoring Elderly People. 2019. Available online: http://www.ivl.disco.unimib.it/activities/monitoring-elderly-people/ (accessed on 31 December 2019).
55. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common objects in context. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 740–755.
56. Zhang, L.; Lin, L.; Liang, X.; He, K. Is faster R-CNN doing well for pedestrian detection? In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; pp. 443–457.
57. Li, W.; Zhao, R.; Xiao, T.; Wang, X. Deepreid: Deep filter pairing neural network for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 24–27 June 2014; pp. 152–159.
58. Bengio, Y. Deep learning of representations for unsupervised and transfer learning. In Proceedings of the 2011 International Conference on Unsupervised and Transfer Learning Workshop - Volume 27, Washington, DC, USA, 2 July 2011; pp. 17–37.
59. Bianco, S.; Buzzelli, M.; Schettini, R. A unifying representation for pixel-precise distance estimation. *Multimed. Tools Appl.* **2019**, *78*, 13767–13786.
60. Roser, M.; Appel, C.; Ritchie, H. Human Height - Our World in Data. 2019. Available online: https://ourworldindata.org/human-height (accessed on 31 December, 2019).
61. García-Herranz, M.; Haya, P.A.; Alamán, X. Towards a Ubiquitous End-User Programming System for Smart Spaces. *J. UCS* **2010**, *16*, 1633–1649.
62. Ciocca, G.; Schettini, R. Dynamic key-frame extraction for video summarization. In *Internet Imaging VI*; Santini, S., Schettini, R., Gevers, T., Eds.; International Society for Optics and Photonics, SPIE: Washington, DC, USA, 2005; Volume 5670, pp. 137–142.
63. Ciocca, G.; Schettini, R. An innovative algorithm for key frame extraction in video summarization. *J. Real-Time Image Process.* **2006**, *1*, 69–88.