Journal of Signal Processing Systems (2023) 95:1037–1050
https://doi.org/10.1007/s11265-023-01840-4

Deep Multimodal Habit Tracking System: A User-adaptive Approach for Low-power Embedded Systems

Daniel Deniz[1] · **Gabriel Jimenez-Perera**[1] · **Ricardo Nolasco**[2] · **Javier Corral**[2] · **Francisco Barranco**[1]

Received: 1 April 2021 / Revised: 15 March 2022 / Accepted: 15 January 2023 / Published online: 20 February 2023

Abstract
The pace of population ageing is increasing and is currently becoming one of the challenges our society faces. The introduction of Cyber-Physical Systems (CPS) has fostered the development of e-Health solutions that ease the associated economic and social burden. In this work, a CPS-based solution is presented to partially tackle the problem: a Deep Multimodal Habit Tracking system. The aim is to monitor daily life activities to alert in case of life-threatening situations improving their autonomy and supporting healthy lifestyles while living alone at home. Our approach combines video and heart rate cues to accurately identify indoor actions, running the processing locally in embedded edge nodes. Local processing provides inherent protection of data privacy since no image or vital signs are transmitted to the network, and reduces data bandwidth usage. Our solution achieves an accuracy of more than 80% in average, reaching up to a 95% for specific subjects after adapting the system. Adding heart-rate information improves F1-score by 2.4%. Additionally, the precision and recall for critical actions such as falls reaches up to 93.75%. Critical action detection is crucial due to their dramatic consequences, it helps to reduce false alarms, leading to building trust in the system and reducing economic cost. Also, the model is optimized and integrated in a Nvidia Jetson Nano embedded device, reaching real-time performance below 3.75 Watts. Finally, a dataset specifically designed for indoor action recognition using synchronized video and heart rate pulses has been collected.

Keywords Cyber-Physical System · e-Health · Multimodal Machine Learning · User-adaptive · Edge computing

1 Introduction

E-Health brings together healthcare and Information and Communication Technologies (ICT) to tackle some of the most relevant challenges that our society is currently facing [1, 2]. One of these challenges is our aging population: due to the increasing life expectancy, by 2050 25% of the population in Europe and North America is expected to be over 65 years old [3]. At the same time, there is a growing interest for tools that enable users to take active control of their well-being by monitoring their lifestyle and health [4]. Particularly, Habit Tracking (HT) systems play a crucial role in increasing the efficiency of healthcare systems. For example, by the early detection of risks such as home accidents, contributing to the independent living of the elderly at their own homes [5], or by promoting healthier lifestyles [6].

The evolution of the Internet of Things (IoT) and the introduction of Systems-on-a-Chip (SoC) devices have acted as catalysts for the development of Cyber-Physical Systems (CPS) that provide cost-efficient distributed and scalable e-Health solutions for assisting people with needs [7–9]. CPS integrate distributed computation at their processing nodes, communication and physical processes that respond to their environment, potentially with humans in the loop [6, 10]. Particularly for e-Health, CPS provide distributed solutions for remote care and thus, they are the best-suited candidates for lifestyle monitoring systems [11].

✉ Daniel Deniz
danideniz@ugr.es

Gabriel Jimenez-Perera
gabrieljimenez@ugr.es

Ricardo Nolasco
rruiznolasco@rgb-medical.com

Javier Corral
jcorral@rgb-medical.com

Francisco Barranco
fbarranco@ugr.es

[1] Computer Architecture and Technology CITIC, University of Granada, Granada, Spain

[2] RGB Medical devices, Madrid, Spain

The core processing of lifestyle monitoring systems is action recognition. HAR (Human Activity Recognition) automatically labels human actions from images, videos, or inertial data from wearable devices [12]. With the recent exponential development of Machine Learning, state-of-the-art solutions address action recognition using Deep Learning (DL) models [13, 14]. Multimodal DL architectures provide more robust and accurate HAR by taking advantage of heterogeneous data sources [15]. In particular, previous works have combined video with inertial sensors information [16], or audio and video [17].

In our work, two data sources are combined to develop a *Two-Stream* multimodal architecture for action recognition. The model integrates two streams from: RGB video and heart pulse rate. The video stream analysis is performed using an optimized version of the *RGBI3D* network [18] which is among the state-of-the-art works in recognition performance. The second stream processes the heart pulse rate information collected from a SpO_2 telemedicine module from RGB Medical [19]. A few works have already described solutions that use heart rate data to perform action recognition using Convolutional operations and Recurrent Neural Networks [20, 21]. However, these approaches are limited to ambulatory activities such as *walking* or *running*. To the best of our knowledge, this work represents one of the first approaches that combines these data modalities for action recognition.

Furthermore, edge computing has recently gained importance on distributed CPS [22], specially with the novel edge-cloud paradigms [23]. Local edge processing enables distributed computing, reducing network bandwidth usage and shortening latency, while inherently ensuring the privacy of sensitive information [24]. Our CPS uses high-performance power-efficient embedded devices that provide a good performance vs power consumption trade-off: Jetson Nano SoMs (System on Module) [25, 26]. These cost-optimized embedded devices have limitations in terms of computation capabilities compared to high-performance workstations, generally used for machine learning applications. Therefore, this work proposes optimized DL models that reach real-time performance with the limited available resources while maintaining good accuracy rates for human action recognition tasks.

In this paper, we present an optimized *Multimodal* DL architecture for a CPS that monitors Habit Tracking. Next, the contributions of the work are summarized: 1) a custom dataset for indoor action monitoring is collected, including two modalities namely, video and heart rate data; 2) a novel Two-Stream DL architecture for HAR is developed, combining cues from video and heart rate (see Fig. 1); 3) optimized the HAR models for low-power embedded devices that reach real-time processing are presented; 4) finally, an optimized DL model is described to obtain an enhanced user-adaptive system towards the improvement of accuracy, specifically for the crucial recognition of critical actions.

2 Related Methods

In this section, we describe the state-of-the-art for habit tracking systems and HAR applications, and methods that use telemedicine modules to monitor vital signs.

Generally speaking, Habit Tracking is a component of remote healthcare monitoring systems with a significant socio-economic impact. HT systems aim at improving the users' quality of life and it is also able to alert caregivers in case of emergency. Approaches that include Habit Tracking are currently

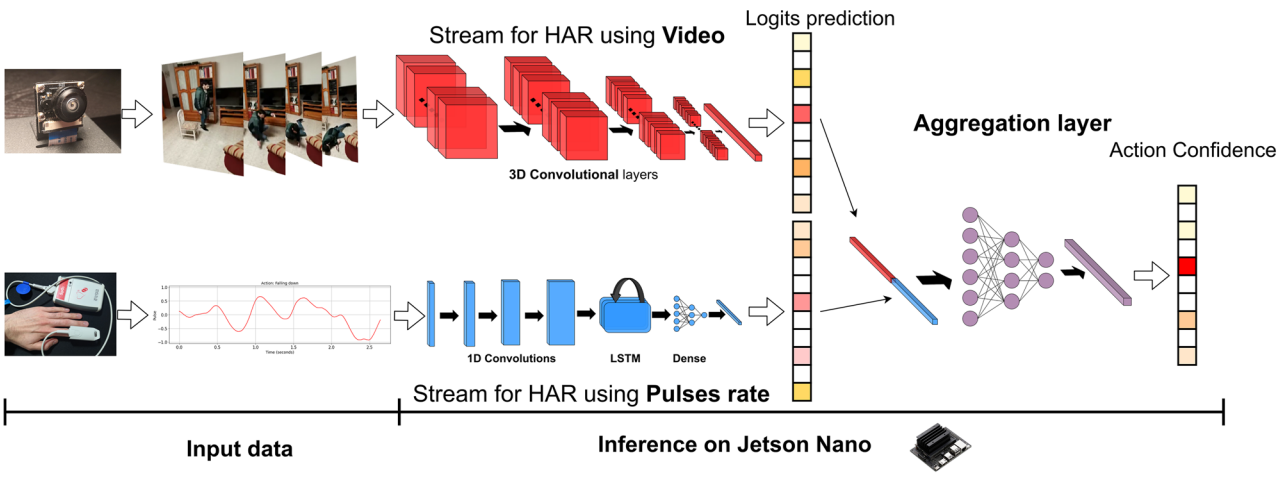

Figure 1 *Multimodal* DL architecture for human action recognition. The proposed Two-Stream model combines (top) RGB videos and (bottom) heart rate data from a SpO_2 monitor (a medical Pulse Oximeter [19] that measures pulse rate and oxygen saturation, although the latter is not used in this work). The action prediction (inference) takes place in power-efficient embedded devices - Nvidia Jetson Nano -, reaching real-time performance.

of interest for reducing healthcare costs at nursing homes and hospitals [27]. Additionally, Habit Tracking applications enable the monitoring of people with care needs at their own homes, especially relevant for the elderly and people with disabilities that live alone [28]. On the one hand, Habit Tracking systems help to detect potential risks and consequently to trigger alarms [29] offering safety and promoting the user autonomy. On the other hand, these systems also foster healthy lifestyle habits and are able to detect the progressive deterioration of the users [30], a convenient feature for doctors. Regarding some examples of the state of the art on habit tracking: in [31] authors focused on fall detection, proposing a system that notifies caregivers when falls occur using data from inertial sensors; in [32], the work is limited to the monitoring of medical parameters; or in [33], authors describe a system that uses wireless passive sensors that monitor energy consumption, temperature, and motion in order to find out about the behavior of the person at home.

2.1 Human Activity Recognition

Human Activity Recognition (HAR) pursues the analysis and recognition of human actions from different data sources [12]. Based on the kind of source, HAR systems are split into two main categories: 1) systems that recognize human actions from data collected from smartphones, accelerometers in wrist-worn smart devices, or other wearable devices [21, 34]; 2) video-based systems [35]. Both alternatives offer good results in terms of recognition performance using Deep Learning [34, 36]. However, video-based HAR is a less intrusive approach since it merely analyzes video streams from cameras, avoiding the need to wear any device. It also represents an important advantage when considering the reluctance of some users to wearable devices [37].

Regarding the wearable sensors approach, CNNs (Convolutional Neural Networks) and RNNs (Recurrent Neural Networks) are two of the most effective approaches when analyzing signals [38]. Concretely, [39] proposes a model based on RNN layers such as LSTM (Long short-term memory) [40] to extract the temporal patterns of the signals retrieved from body-worn sensors such as accelerometers and gyroscopes to identify daily activities. Moreover, in [41] authors use CNN to extract the salient patterns of signals obtained from body-worn and ambient sensors to identify human actions.

Regarding video-based approaches, the introduction of the 3D Convolutional networks for video processing boosted action recognition performance in terms of accuracy [42]. 3D Convolutional operations simultaneously allow extracting spatial and temporal information from video streams. However, this comes with a high cost in terms of computational requirements compared to conventional 2D Convolutions, more common for single image frame analysis. In particular, action recognition based on 3D convolutions has been addressed through: *3D Residual Network* [43] that proposed a 3D ResNet model [44] improving the state of the art for general action recognition; *Temporal 3D ConvNet* [45] that incorporated 3D filters to a modified DenseNet [46] architecture focusing on the only-temporal cues more than the traditional spatio-temporal approaches; or *RGBI3D* network [18] that inflates the Inception [47] network with 3D operations. Concretely, the *RGBI3D* network [18] is a very interesting approach that achieves high accuracy at a reasonable computational budget for the Kinetics [48] dataset, a widely used dataset for action recognition. Conveniently, the pre-trained weights of the *RGBI3D* model for the Kinetics dataset are publicly available. The availability of the pre-trained model enables boosting the performance of action recognition on smaller datasets through Transfer Learning, also benefiting generalization [49].

2.2 Telemedicine Modules

Telemedicine modules such as electrocardiograms (ECG) or pulse oximeters (SpO_2) contribute to decentralize patient care, moving it outside the hospital, favoring the decrease of hospitalization costs, and increasing equality of care for people that live in isolated environments through e-Health [50, 51].

Our paper describes a solution with a pulse oximeter [19], a module that monitors the heart rate and oxygen saturation of patients through pulse oximetry [52]. Pulse oximetry is a non-invasive technique used to continuously measure the arterial oxygen saturation of the patient [53]. This measurement is performed via photoplethysmography: since arterial blood contains basically two main oxygen absorbents (oxyhemoglobin and reduced hemoglobin), this method uses light emitters with two different wavelengths (red and infrared) to capture the changes of arterial blood volume and to obtain the oxygen saturation and heart rate information [54]. The set of signal data obtained through this method is called a photoplethysmogram (PPG) or pulse rate signal. In summary, PPG signals provide information related to cardiopulmonary system [53].

The information extracted by PPG sensors can be analyzed through Deep Learning for a wide variety of applications such as: in [55, 56], authors analyze pulse rate signal to provide a precise heart rate estimation; biometric identification [57]; emotion recognition [58]; or Activity recognition [20, 21], identifying ambulatory activities such as: *sitting*, *walking*, *jogging*, or *running*.

2.3 SpO_2 Pulse Oximeter

In this work, we have selected the wireless battery-powered SpO_2 (pulse oximeter) module [19] from RGB medical devices for measuring the vital signs of the subjects in a non invasive way. This pulse oximeter generates the Photoplethysmogram (PPG) wave or pulse rate signal that provides

information about the changes of pressure in the blood vessels. Its sampling frequency is 66.67Hz (about 15 ms).

Based on the physiological signal, the module outputs two main measures: 1) Oxygen Saturation (SpO_2) in arterial blood (0-100%); and 2) Heart Rate (30-250 bpm). This telemedicine module provides high resolution and accuracy of the measurements: for Oxygen saturation values, it offers a resolution of 1% step with an accuracy of ±2 *digits* for 70-100% SpO_2; for the heart rate, it offers a resolution of 1 bpm step and an accuracy of ±3%. Additionally, the oximeter alerts the user when the sensor is disconnected or the signal is weak.

For our model, we analyze the PPG signal (pulse rate) to infer indoor human actions. Although other devices can be used to obtain the pulse rate information, this SpO_2 pulse oximeter also monitors the oxygen saturation, enabling the early detection of hypoxemia: the condition of an abnormal below level of oxygen in the blood (< 93%) [52].

3 Our Approach

We present a Two-Stream DL model for recognizing human actions from two modalities: video and heart rate information. First, the two streams are separately developed: a DL model that uses video and another that uses the vital signs (heart pulse rate). Next, a new model that combines both streams is proposed to create a *Multimodal* architecture that obtains the best recognition rates. With the improved performance, we plan to enhance the recognition of critical actions, in order to reduce false alarms. Afterwards, a user-adaptive approach is studied to evaluate the benefits of specializing the DL model to different subjects in terms of recognition performance. This additional adaption is reasonable considering that the system will be of personal use and will also enable the possibility of continuous refinement over time. Finally, to train the DL architectures, a dataset has been collected with synchronized data from both sources of information.

3.1 Deep Learning Models for HAR

The proposed solution for activity recognition is a Two-Stream DL model. The first stream takes care of the video input, and the second analyzes the heart rate. The first stream runs the *RGBI3D* network and uses resource-intensive operations such as 3D convolutions to extract spatio-temporal information from the video. The second stream is a *1DCNN + LSTM* network that extracts temporal patterns from the 1D PPG signal. The latter stream requires significantly less computational requirements than the other one, although the former is more accurate (video information is more discriminant for action recognition). Additionally, the final layers combine features from both data modalities using a custom layer (*WeighPerClass*) that weighs the contribution of the streams to every action for the final prediction. The code is publicly available on GitHub[1].

3.1.1 RGB Video Stream

The architecture for the video-based HAR stream is based on the *RGBI3D* network (see Fig. 2) [18]. This model performs action recognition from a regular video stream recorded at 25 fps with a resolution of 224 × 224. However, it is a resource-demanding network, mainly due to the use of the 3D Convolutional layers. Since our CPS nodes are low-power embedded processors with limited resources, an adaptation of the model is required. The spatial resolution and temporal framerate are downsampled, reducing the operations of the input layers.

Transfer learning [49] and specifically fine tuning [59] is applied to prevent overfitting and reduce training time, not requiring large amount of training data. Transfer learning and fine tuning are methods to reuse pre-trained DL models for similar tasks, taking advantage of the knowledge extracted by the DL models by retraining them for a few epochs.

Concretely, the model is fed with a stream of 64 frames with a spatial resolution of 112 × 112 (2560 ms of video). Next, the network is trained following these steps:

- A data augmentation procedure is added to prevent overfitting. A spatio-temporal window of each video is randomly cropped and fed to the network every epoch. The video is also randomly flipped and rotated.
- Transfer learning is applied using the *RGBI3D* network [18] to take advantage of the pre-trained weights that extracted the knowledge from the Kinetics dataset. This dataset includes all kinds of actions such as human-object or human-human interactions ranging from e.g. people playing sports or instruments to people hugging.
- Finally, fine tuning is performed to obtain a faster convergence to our own indoor actions, for which we use the dataset presented in [11] for training. The training is done for 40 epochs with a batch size of 8, using the Adam optimizer.

Moreover, after training and in the pursue of creating a more accurate model that adapts to the user behavior, the model is fine tuned again for the user during an initial configuration phase. Specializing the DL model for each user makes the system learn the user model, concerning to the relevant features for action recognition. It is a simple approach for a user-adaptive system [60] that achieves significant improvement in accuracy for action recognition.

[1] https://github.com/DaniDeniz/deep-multimodal-action-recognition

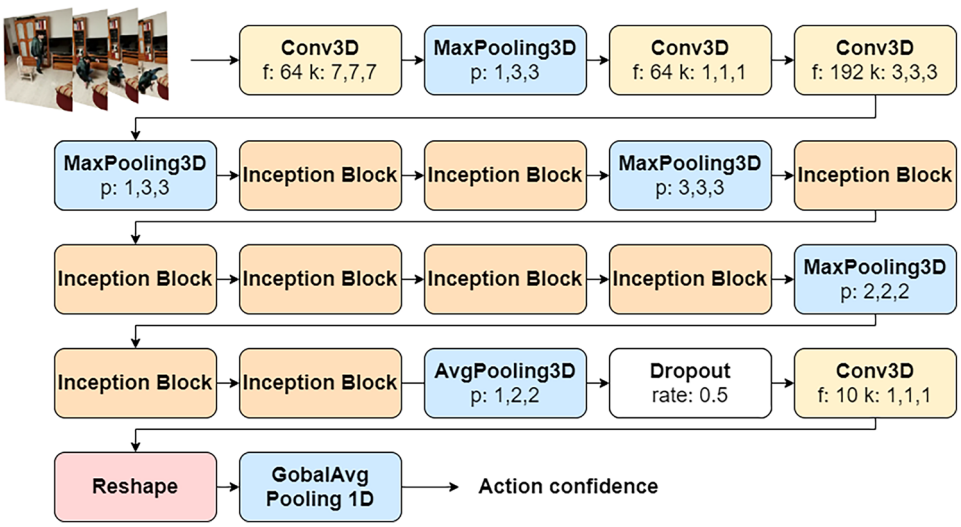

Figure 2 The *RGBI3D*-based stream for action recognition from Video is based on the *Inception* model [47]. It uses 3D convolutional layers to analyze spatio-temporal information from the video, pooling layers to reduce data resolution (and thus, the computation and total parameters of the network), and dropout layers to avoid overfitting. Refer to [18] to inspect the operations carried out on the *Inception* blocks. This is a complex DL model with more than 12 million weight parameters. *f*: Number of filters, *k*: kernel size, *p*: pool size, *rate*: fraction of dropped connections between layers.

3.1.2 Pulse Rate Stream

A medical pulse oximeter measures pulse rate, a 1D signal whose frequency indicates the heart-rate of the patient. The amplitude and frequency of the pulse signal vary according to the activity that the user is performing [55].

We propose a novel model (see Fig. 3) to analyze pulse rate data and infer indoor human actions. This DL model is formed by two main types of layers: 1) 1D Convolutions and 2) LSTM layers. The 1D Convolution layers are used to extract the local variations of the points of the signal. LSTM is a Recurrent Neural Network (RNN) that extracts the temporal information at a global scale via memory to learn sequences of patterns. This DL network also uses dropout layers: it drops some connections between the 1D Convolution layers at training time to prevent overfitting.

Furthermore, the DL model is designed to accept an input of variable size. Thus, the network is fed at training and inference with a variable number of points. In this way, the network is enabled to learn for adaptive time intervals, ensuring a better generalization of the model.

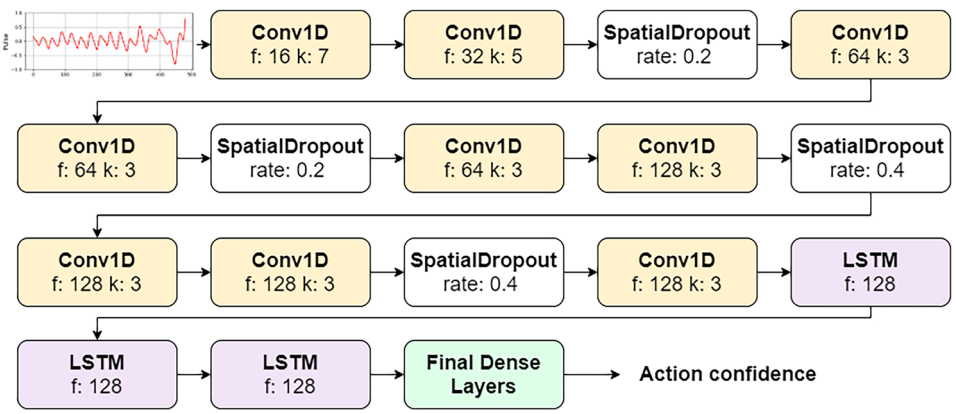

Figure 3 1DCNN + LSTM DL stream for action recognition from pulses. It uses 1D convolutional layers to analyze local variations of the pulses, LSTM layers to learn temporal patterns, and dropout layers to avoid overfitting. This DL model amounts to a total of 650000 weight parameters (approximately 20x less complex than the video-based stream architecture). *f*: Number of filters, *k*: kernel size, *rate*: fraction of the dropped connections between layers.

As for the training procedure, it is done as described below:

- First, data augmentation is performed sampling a temporal window in a different location of the pulses measure at every time. Data augmentation contributes to increasing the accuracy while also preventing overfitting, and it is very helpful when, as in our case, the amount of available data is scarce for DL training. It is ensured that at least, the last 2.5 seconds of the selected signal correspond to the actual action that is being analyzed. The rest of the sample can partially overlap with the previous action (actions in the dataset are recorded in a continuous batch).
- The 1DCNN + LSTM stream is trained from scratch for 100 epochs, using the Adam optimizer with a batch size of 4. Then, the weights that correspond to the epoch with the lowest validation loss are selected. The input is on average 6.6 seconds long (approximately 440 points).

Similarly to the video-based approach, the DL model is specialized in an initial configuration phase, to adapt to the user features, behavior, and scenario. The final stage of specialization makes the system become a user-adaptive solution that achieves better accuracy results.

3.1.3 Two-Stream DL Network for HAR

The proposed multimodal architecture for the Habit Tracking system integrates information from the video and pulse rate streams. Two approaches are followed to combine the streams: 1) the addition of a layer that sums up logits output from both streams; 2) the addition of a new (*WeighPerClass*) layer that learns to weigh the individual contributions for every action. The first simple approach consists in aggregating the unnormalized logit predictions from both streams and pass them to a softmax activation function. The second approach follows the next procedure: the logit predictions from both streams are concatenated (L), and passed into a custom layer named *WeighPerClass*. This layer is built using a weighing matrix W ($NClasses \times NStreams$), initialized to 1 (in our case, *NStreams* is 2 and *NClasses* is 10). Then, a softmax function combines the contribution from each stream and for each class (see Eq. 1)

$$WeighPerClass(W,L)_{i,j} = \frac{e^{W_{i,j}}}{\sum_{k=1}^{nstreams} e^{W_{i,k}}} \cdot L_{i,j} \quad \forall_{i\in\{1,\ldots,nclasses\}} \text{ and } \forall_{j\in\{1,\ldots,nstreams\}} \tag{1}$$

After the weighing phase, another softmax activation function is applied to obtain the final output of the *Multimodal* DL architecture.

The *WeighPerClass* layer of the DL model is trained only for 5 epochs, starting with a learning rate of 0.01 that is reduced by a factor of 2 after every epoch. Note that the learning rate (lr) has a huge impact on the result: large learning rates may lead the network to suboptimal solutions while too small values result in insignificant variations on the original weights, causing the process to last very long or in the worst case, to get stuck.

As a result, the Two-Stream DL model improves action recognition performance, especially for the critical actions such as *falling down* or *lying on the floor* reducing false positives, benefiting from features extracted from both video and heart rate data.

4 Discussion and Results

In this section, we first describe the collection of our custom IAPV dataset for lifestyle monitoring systems and its structure. Next, we present results to prove the benefits of specializing the DL model for each user, supporting our decision to build a user-adaptive system. Also, we present an ablation study of our multimodal DL model using our IAPV dataset, assessing the independent contributions of the video and pulse rate streams. Finally, since our goal is a CPS with nodes that perform HAR, the DL models are optimized and their performance and power consumption (essential qualities for embedded edge nodes) are discussed.

4.1 Multimodal Indoor Action Dataset - IAPV

Since Machine Learning systems learn from examples, one of the most important elements when building a Machine Learning is the availability of datasets for the application field. The quantity and quality of the data is crucial and has a direct impact on the system recognition performance [61]. Currently, there are publicly available datasets for performing activity recognition from video [48, 62] and pulses rate information [56]. However, there are no available datasets that provide video and pulses data synchronously for action recognition.

To overcome this problem, we collected a multimodal Indoor action recognition dataset (IAPV), gathering synchronized videos and pulse rate information. This dataset contains indoor scenes of people performing actions at different scenarios at home such as bedrooms, living rooms and kitchens. The dataset was recorded by 5 actors (3 men and 2 women) at home. Actions were recorded in continuous batches of ten minutes, using the RGB Medical telemedicine (SpO_2) module for the pulse rate data. Actors carried out different actions relevant for lifestyle monitoring as listed in Table 1 such as *cleaning*, *eating*, *sitting down*, *walking*, or *watching tv*. Also, some critical actions were included such as *falling down* or *lying on the floor* that are useful to identify whether a subject has suffered a life-threatening situation that requires assistance. Figure 4 shows three examples that illustrate some of these actions performed by different actors.

Table 1 Number of clips of the IAPV Dataset.

Action name	Train	Val.	Test	Total
blowing nose or sneezing	38	10	15	63
cleaning	55	17	34	106
eating	57	18	32	107
falling down	34	11	16	61
lying on the floor	53	12	36	101
sitting down	66	22	29	117
standing up	95	30	41	166
walking	158	34	90	282
watching tv	60	19	43	122
no action	57	14	32	103
Total	673	187	368	1228

Regarding the data preparation, videos were first manually segmented into clips and then labeled. Next, pulse data were also automatically labeled since they are synchronized with videos during collection. As mentioned, Table 1 shows the nine action labels contained in the dataset. The *no action* class represents scenes of empty rooms without people to help the DL model focus on humans when recognizing the actions instead of focusing on the environment. This class was not taken into account for the pulse rate stream (no humans are involved).

Multimodal instances were manually assigned to each split approximately: 55% training, 15% validation, and 30% testing. Data is carefully split guaranteeing that all data from a batch of actions is assigned to the same set, ensuring the fairness of the experiment.

Table 2 Number of clips recorded by each actor (IAPV).

Actor ID	Gender	Train	Val.	Test	Total
0	M	406	92	123	621
1	M	32	0	85	117
2	M	70	25	29	124
3	F	70	0	63	133
4	F	95	70	68	233

Furthermore, Table 2 shows the number of clips recorded per actor. The actor ID is used to perform individual analysis of the recognition performance depending on the subject, and to enable the user-adaptive system. Bear in mind that, after the system deployment, more data will be collected from subjects to fine-tune the system and improve its performance in real situations in subsequent iterations.

In this section, the performance of the model using the IAPV dataset is presented, focusing mainly in two metrics: *accuracy*, and *macro F1-score*. The *macro F1-score* is the harmonic mean of the precision $\frac{TP}{TP+FP}$ and recall $\frac{TP}{TP+FN}$ values. This metric equally weighs the contribution of every class, thus it takes into account the issues of an unbalanced dataset as it is the case with our IAPV dataset.

Firstly, an evaluation of the recognition performance of the *Two-Stream* multimodal architecture is shown. Next, an evaluation of the customization stage that is added to specialize the system to make it respond better to a specific user is presented. Finally, an ablation study has been included to understand the independent contributions of the streams.

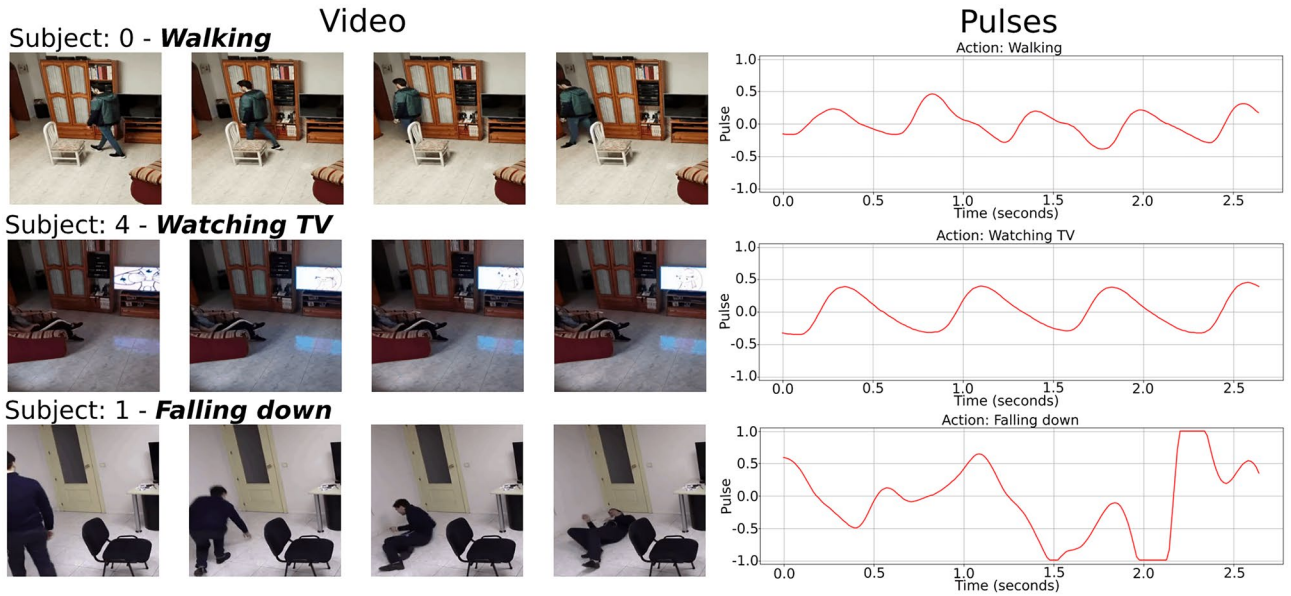

Figure 4 Samples of the IAPV dataset. It shows three of the subjects performing the following actions: *walking, watching tv*, and *falling down*. Note how pulses are synchronized with the video stream and how they greatly vary depending on the specific action.

4.1.1 Two-Stream Multimodal Network Evaluation

The *Two-Stream* network processes input data from two different sources (video and pulses rate). Evaluation is addressed using a sliding window of 64 frames and averaging the inference confidence along the video clip, for every action instance. Regarding the pulse rate signal, inference uses chunks of 6.6 seconds (440 points) synchronized with the video. When the system is deployed, multimodal data is synchronized by timestamping the signals received on the node. Then, the DL model is fed with batches of data representing the same time slice of images and vital sign data.

4.2 Evaluation of the DL Architecture

This network fuses the information retrieved by the streams following two approaches: 1) A naive approach that sums up the predictions of both streams; 2) An approach that uses the custom layer *WeighPerClass* to weigh the contribution of every stream on the prediction of every class. Table 3 shows that if the model uses the *WeighPerClass* layer, it reaches better results in terms of accuracy and F1-score, obtaining 1.7% and 2.5% higher values respectively compared to the naive approach. This improvement in performance when using the weighing custom layer is due to the network learning the right contribution of each stream per action.

In particular, using the custom weighing layer, for *standing up* action the video stream has a contribution of 71% for the prediction of this class, meaning that the video is a much more discriminant feature for this action. However, pulse rate data brings relevant information for identifying people *falling down* or *lying on the floor*. For these actions, the video stream contribution is reduced to approximately 60% and the pulse rate network reaches 40% of the contribution for the confidence of the predicted action.

Finally, observe in the confusion matrix (Fig. 5) that the *Two-Stream* model obtains 93.75% of precision and recall values for the class *falling down*, and the F1-score for *lying on the floor* reaches 72%. This model also reaches great recognition performance for the actions *walking* or *cleaning* among others. For example, it offers a high recall for *watching tv* and reasonable accuracy when identifying people *eating*, two relevant actions towards the definition of strategies for promoting healthy lifestyles. However, it has misleading classifications between some actions. Note also that the top 2 accuracy of the DL model is 89.9%. This means that with high probability, the carried action will be identified as one of the top 2 with higher confidence.

The presented performance values show how this model offers good results for indoor action recognition, enabling the lifestyle monitoring application. It also provides very high recognition accuracy for potentially risky situations such as *falling down* or *lying on the floor*. This minimizes the triggering of false alarms and increases the probability of accurate identifications, reducing costs and building social trust in these systems.

Table 3 Two-Stream evaluation.

Model name	Accuracy	F1-score
Two-Stream (naive approach)	79.34	76.35
Two-Stream (weighing custom layer)	**80.70**	**78.27**

Bold emphasis highlights the model with highest recognition performance

4.2.1 Evaluation of the User-adaptive System

The presented *Two-Stream* DL model is trained using the whole IAPV dataset. However, one of the objectives of this work is to study the effect of specializing the DL model for each actor. As mentioned in the introduction, after a first user-adaptation stage, the overall recognition rate of the system is significantly improved leading to a higher user engagement.

The specialization is addressed in a two-step procedure: firstly, each model is trained leaving one actor out; secondly, DL model is fine-tuned using the whole IAPV dataset (including training data of the new subject). Although the second step could achieve better results when training the specialized DL model using only the selected actor's clips, this was rejected due to the lack of training data. Building a larger dataset with more actors and samples would be crucial for the latter.

Table 4 presents the results of the user-adaptation procedure; accuracy and F1-score are evaluated only for the test set of the selected actor before and after the specialization phase with the *Two-Stream* architecture. The accuracy after specialization is increased on average approximately 18%, and the F1-score in 21.85%. This shows how adapting the model to the user remarkably improves recognition rates.

Let us point out, for example, the case of *Subject 0* and *Subject 4*. Both subjects account for the largest number of samples (see Table 2) and consequently, the greatest improvements in terms of F1-score (around 33%) are found for these two subjects. Obviously, the number of samples has an impact on this user-adaptive approach but even a small number of samples (such as the 32 clips from *Subject 1*) may lead to a substantial improvement (20% for the F1-score). In any case, using the samples from other subjects at training benefits generalization, leading also to better recognition performance rates.

4.2.2 Ablation Study

This section discusses the contribution of each stream to the results of the *Two-Stream* multimodal architecture.

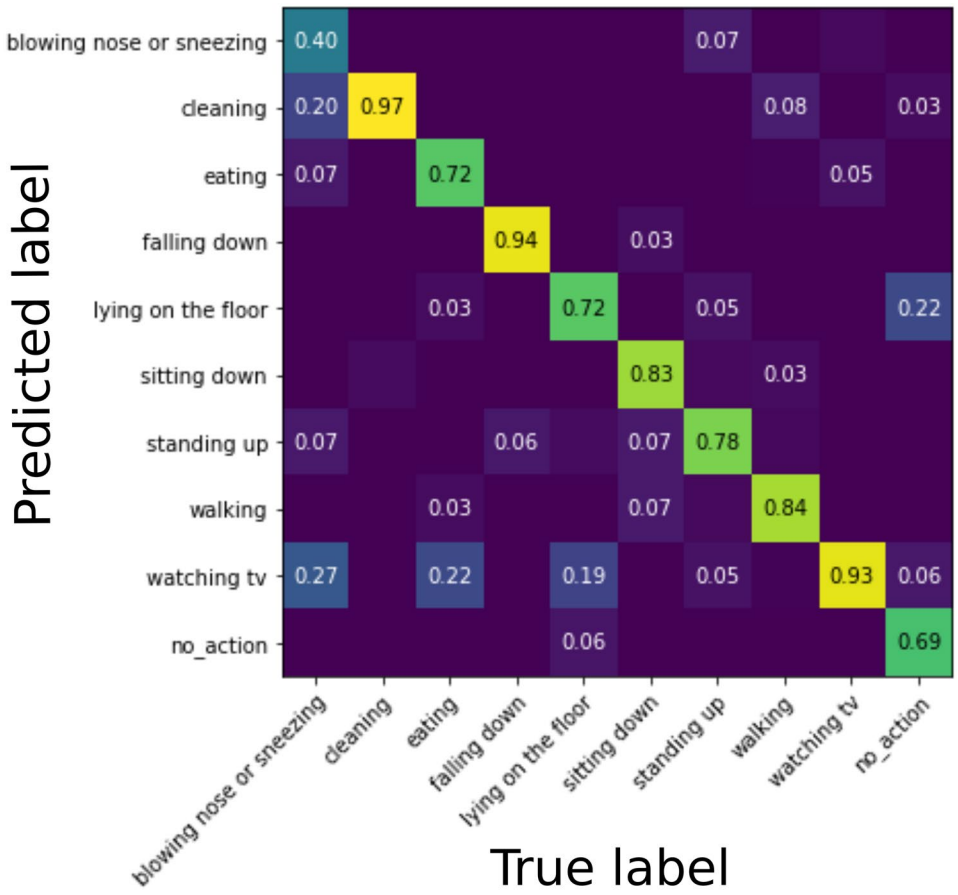

Figure 5 Confusion matrix of the evaluation of the *Two-Stream* model using the weighing custom layer over the IAPV dataset. Note how the DL model obtains an F1-score of 93.75% when identifying the critical action *falling down*. The model also successfully recognizes other actions useful for habit monitoring such as: *walking* or *watching tv*. Recognition is slightly degraded for actions such as *blowing nose or sneezing*, but this usually occurs when the subject is simultaneously carrying out other activities (eg. *cleaning* or *watching tv*).

Figure 6 shows precision-recall curves for the *Two-Stream* architecture and the separated streams. PR-curves show the trade-off between the precision and recall metrics for different thresholds. The area under the curve (AUC) is related to precision-recall values: larger AUC denotes better results. Note how the *1DCNN+LSTM* reaches poor values of AUC when evaluated independently. However, it contributes to improving the AUC of the *Two-Stream* architecture by 2.8% with respect to the *RGBI3D* stream. Since values shown here are the average for all the actions, one could consider this contribution not very relevant. However, the substantial improvement in the recognition for certain actions justifies

Table 4 User-adaptive evaluation.

Subject	**Before** specialization		**After** specialization	
	Accuracy	F1-score	Accuracy	F1-score
0	67.47	57.34	82.92	**76.71**
1	78.82	72.57	91.76	**87.34**
2	62.06	41.88	72.41	**51.74**
3	61.90	**58.06**	68.25	57.53
4	76.47	71.80	95.58	**95.15**

Bold emphasis points out at which stage of the user-adaptation procedure the F1-score is highest

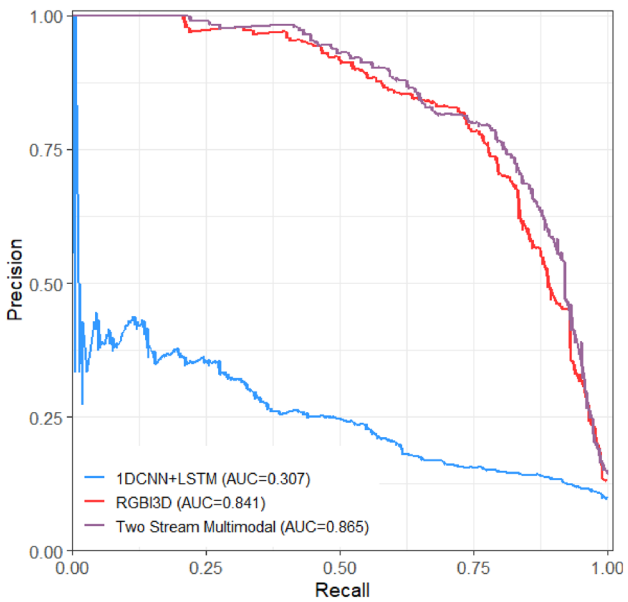

Figure 6 Macro-average Precision-Recall curves for the *Two-Stream* architecture and the two separated streams. Although the pulse rate stream (*1DCNN+LSTM*) obtains poor precision-recall values, this computationally efficient stream brings gains in terms of AUC (2.8%) for the *Two-Stream* network compared to the *RGBI3D* model, and helps increasing recognition for critical actions.

Table 5 Evaluation of individual streams per class.

Action	1DCNN+LSTM		RGBI3D	
	Precision	Recall	Precision	Recall
blowing nose or sneezing	0.00	0.00	58.33	46.67
cleaning	41.67	29.41	64.70	97.05
eating	13.63	9.37	85.18	71.87
falling down	20.00	56.25	93.75	93.75
lying on the floor	33.33	41.67	66.67	55.55
sitting down	16.67	10.34	92.00	79.31
standing up	38.00	46.34	82.05	78.04
walking	50.68	41.11	94.73	80.00
watching tv	47.45	65.11	64.51	93.02
no action	-	-	76.67	71.87

the combination (see Table 5 and clarification in the next three paragraphs).

The *1DCNN+LSTM* stream analyzes the pulse rate signal using 1D convolutions and LSTM layers. The low complexity of this architecture coupled with the one-dimensional signal feeding the network makes it computationally efficient and thus, it offers very fast inference. The analysis of Table 5 shows that using only this stream is not enough for recognition. This is mainly due to the lack of distinctive features in the pulse rate for some actions such as *eating* or *sitting down*. However, the model is much more accurate to discriminate between actions such as *walking*, *falling down* or *lying on the floor*.

Regarding the *RGBI3D* stream, it provides great recognition results at the expense of high computation requirements, mainly due to the use of 3D convolutional operations. Note that it has 20x more weight parameters compared to the *1DCNN+LSTM* model. Nevertheless, it achieves high recognition performance for lifestyle monitoring. In contrast to the pulse rate stream, it accurately recognizes actions such as *eating*, *sitting down* or *watching TV*. Remarkably, the video model achieves an F1-score of 93.75% for the critical action *falling down*. On average, it reaches an accuracy of 78.26% for the IAPV dataset. Despite of this, it presents significant confusion between *lying on the floor* and *no action*. Obviously, in both cases the network does not detect any motion in the scene.

As shown, both streams are complementary and their combination boosts the model performance, including the critical actions with most interest for our case. For example, the F1-score of *lying on the floor* improves by 18%. Additionally, F1-score of daily activities such as *walking* or *cleaning* are also improved by 3% and 8.9% respectively (see also Fig. 5). Therefore, despite its low performance when evaluated independently, the *1DCNN+LSTM* network delivers more robust predictions for the *Two-Stream* architecture at a very low cost.

4.3 Edge Processing Nodes

The proposed *Two-Stream* multimodal DL architecture for action recognition has been designed and optimized to be integrated on a CPS for Habit Tracking system that runs on low-powered execution boards. Concretely, we use the Nvidia Jetson Nano [26] embedded systems.

As mentioned in the introduction, local processing in a distributed CPS presents advantages such as inherent data privacy, or reduction of data bandwidth usage and latency. However, it presents limitations in terms of energy and resource availability that lead to design solutions that provide good performance vs accuracy trade-offs. As described in Section 4.1.1, the proposed *Two-Stream* architecture offers great results for the recognition of critical actions and thus, it would minimize false alarms.

The presented DL models were optimized through TensorRT 7 [63] to embed them on the Jetson Nano devices. TensorRT allows the DL model to be optimized quantizing the parameter bitwidths (eg. INT8 - 8-bit integers, or FP16 - 16-bit floating point, instead of the standard 32-bit floating point values) of weights and activation functions. It also enables additional resource savings by fusing layers, or reusing memory to reduce communication latencies. The overall optimization reduces inference time, on average 10% faster, and minimizes GPU memory usage up to 25%. Table 6 compares inference times, energy consumption, and F1-scores for the two streams independently and the final multimodal solution. The multimodal architecture reaches real-time performance with inference times under 1.7 s using input batches of 64 frames or 2560 ms. The efficient *1DCNN + LSTM* stream is able to analyze a 6.6-second long signal in just 16 ms at a very low energy budget. As a result, fusing the video and pulses rate streams leads the solution to an enhanced recognition performance in terms of F1-score of 2.4% with a low cost in terms of time performance and energy.

Finally, in order to illustrate the power consumption, the *Multimodal* network would continuously run on a $5V$ $25.000mAh$ battery-powered Jetson Nano for more than 30 hours.

Table 6 Performance on the Nvidia Jetson Nano.

DL model	Inference time (ms)	GPU (W)	Device (W)	F1-score
RGBI3D	1604.15 ± 0.702	1.24 ± 0.87	3.68 ± 0.97	76.41
1DCNN + LSTM	16.60 ± 0.480	0.09 ± 0.02	2.42 ± 0.05	30.48
Multimodal network	1626.96 ± 2.812	1.25 ± 0.87	3.75 ± 0.97	78.27

5 Conclusions

In this work we have presented a Deep multimodal (Two-Stream) neural network for an e-Health CPS for monitoring the elderly at their home. This solution fuses video and heart rate information to recognize human daily activities with high accuracy. In particular, the IAPV dataset was introduced to train and evaluate the performance of the habit tracking system on a real-world scenario where the subject is monitored using a video camera and a device that monitors the pulse rate (eg. the SpO_2 telemedicine module from RGB Medical devices). The proposed multimodal *Two-Stream* architecture reaches a very high recognition performance improving the only-video action recognition, specially for critical actions that are of utmost interest for our case. Additionally, the performance improvements come at a very low cost in terms of energy and inference time.

The correct recognition of critical actions is crucial to trigger alarms in case of accident, for example. Particularly, our *Two-Stream* model is able to recognize people *falling down* and *lying on the floor* with a F1-score of 93.75% and 72%. Additionally, we have proposed a one-time user-adaptive stage that improves further the recognition of these critical actions. This stage shows how by recording only a few samples and refining the DL model for a few epochs, the recognition performance (F1-score) for a new subject is boosted in average by 21%.

Finally, in our CPS, the multimodal *Two-Stream* model runs on power-efficient Nvidia Jetson Nano devices. These embedded devices enable local processing and allow our CPS to reach real-time inference, consuming less than 3.75 W. This means that in a battery-powered device it would continuously run for more than a day. Additionally, this CPS is easily scalable in terms of computational resources due to its low energy consumption.

Our e-Health cost-efficient solution has multiple applications helping with adapted habit monitoring of patients with a specific disease pathology, or assisting health-care professionals to enhance tailored activity programs for healthy lifestyle to their patients, or in general helping people with needs to safely live autonomously. In future works, we plan to include new telemedicine modules for concrete medical conditions such as heart arrhythmias. We are also interested in working on monitoring systems that evaluate the deterioration of patients or recovery after serious injuries, to help medicine professionals create improved targeted therapies based on the evolution of the collected data.

Funding Funding for open access publishing: Universidad de Granada/CBUA. This project has received funding from the ECSEL Joint Undertaking (JU) under grant agreement No 783162. The JU receives support from the European Union's Horizon 2020 research and innovation programme and Netherlands, Czech Republic, Finland, Italy, and Spain (via the Spanish National grant funded by MINECO through APCIN PCI2018-093184). The work was also partially supported by the National Grant PID2019-109434RA-I00/ SRA (State Research Agency /10.13039/501100011033). Gabriel Jimenez-Perera's research has been supported by "Consejería de Transformación Económica, Industria, Conocimiento y Universidades de la Junta de Andalucía" (Grant Ref. PREDOC_00280).

References

1. Li, J., Han, X., Zhang, X., & Wang, S. (2019). Spatiotemporal evolution of global population ageing from 1960 to 2017. *BMC public health, 19*(1), 1–15.
2. Taramasco, C., Rodenas, T., Martinez, F., Fuentes, P., Munoz, R., Olivares, R., De Albuquerque, V. H. C., & Demongeot, J. (2018). A novel monitoring system for fall detection in older people. *IEEE Access, 6*, 43563–43574.
3. Nations, United and Social Affairs, Department of Economic and Population Dynamics. (2019). 2019 revision of world population prospects. In *World Population Prospects 2019*.
4. Rockmann, R., & Gewald, H. (2015). Elderly people in Ehealth: Who are they? *Procedia Computer Science, 63*, 505–510.
5. Pace, P., Aloi, G., Caliciuri, G., Gravina, R., Savaglio, C., Fortino, G., Ibáñez-Sánchez, G., Fides-Valero, A., Bayo-Monton, J., Uberti, M., et al. (2019). Inter-health: An interoperable IoT solution for active and assisted living healthcare services. In *2019 IEEE 5th World Forum on Internet of Things (WF-IoT)* (pp. 81–86). IEEE.
6. Al-Ars, Z., Basten, T., de Beer, A., Geilen, M., Goswami, D., Jääskeläinen, P., Kadlec, J., de Alejandro, M.M., Palumbo, F., Peeren, G., et al. (2019). The FitOptiVis ECSEL project: Highly efficient distributed embedded image/video processing in cyber-physical systems. In *Proceedings of the 16th ACM International Conference on Computing Frontiers* (pp. 333–338).
7. Calderita, L. V., Vega, A., Barroso-Ramírez, S., Bustos, P., & Núñez, P. (2020). Designing a cyber-physical system for ambient assisted living: A use-case analysis for social robot navigation in caregiving centers. *Sensors, 20*(14), 4005.
8. Farahani, B., Firouzi, F., & Chakrabarty, K. (2020). Healthcare IoT. In *Intelligent Internet of Things* (pp. 515–545). Springer.
9. Shah, T., Yavari, A., Mitra, K., Saguna, S., Jayaraman, P. P., Rabhi, F., & Ranjan, R. (2016). Remote health care cyber-physical system: Quality of service (QOS) challenges and opportunities. *IET Cyber-Physical Systems: Theory & Applications, 1*(1), 40–48.
10. Marwedel, P. (2021). *Embedded system design* (Vol. 1). Springer.
11. Deniz, D., Barranco, F., Isern, J., & Ros, E. (2020). Reconfigurable cyber-physical system for lifestyle video-monitoring via deep learning. In *2020 25th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA)* (Vol. 1, pp. 1705–1712). IEEE.
12. Lara, O. D., & Labrador, M. A. (2012). A survey on human activity recognition using wearable sensors. *IEEE Communications Surveys & Tutorials, 15*(3), 1192–1209.

13. Huang, J., Lin, S., Wang, N., Dai, G., Xie, Y., & Zhou, J. (2019) TSE-CNN: A two-stage end-to-end CNN for human activity recognition. *IEEE Journal of Biomedical and Health Informatics*.
14. Nweke, H. F., Teh, Y. W., Al-Garadi, M. A., & Alo, U. R. (2018). Deep learning algorithms for human activity recognition using mobile and wearable sensor networks: State of the art and research challenges. *Expert Systems with Applications, 105*, 233–261.
15. Sun, Z., Liu, J., Ke, Q., & Rahmani, H. (2020). Human action recognition from various data modalities: A review. Preprint retrieved from http://arxiv.org/abs/2012.11866
16. Wei, H., Jafari, R., & Kehtarnavaz, N. (2019). Fusion of video and inertial sensing for deep learning-based human action recognition. *Sensors, 19*(17), 3680.
17. Owens, A., & Efros, A. A. (2018). Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 631–648).
18. Carreira, J., & Zisserman, A. (2017). Quo vadis, action recognition? A new model and the kinetics dataset. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 6299–6308).
19. RGB Medical. (2021). *Telemedicine modules*. Retrieved February 26, 2021, from https://www.rgb-medical.com/en-gb/telemedicina
20. Boukhechba, M., Cai, L., Wu, C., & Barnes, L. E. (2019). ACTIPPG: Using deep neural networks for activity recognition from wrist-worn photoplethysmography (PPG) sensors. *Smart Health, 14*, 100082.
21. Brophy, E., Muehlhausen, W., Smeaton, A. F., & Ward, T. E. (2020). CNNs for heart rate estimation and human activity recognition in wrist worn sensing applications. In *2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)* (pp. 1–6). https://doi.org/10.1109/PerComWorkshops48775.2020.9156120
22. Liu, B., Zhang, Y., Zhang, G., & Zheng, P. (2019). Edge-cloud orchestration driven industrial smart product-service systems solution design based on CPS and IIoT. *Advanced Engineering Informatics, 42*, 100984.
23. Sharma, S. K., & Wang, X. (2017). Live data analytics with collaborative edge and cloud processing in wireless IoT networks. *IEEE Access, 5*, 4621–4635.
24. Wang, X., Xue, H., Liu, X., & Pei, Q. (2019). A privacy-preserving edge computation-based face verification system for user authentication. *IEEE Access, 7*, 14186–14197.
25. Isern, J., Barranco, F., Deniz, D., Lesonen, J., Hannuksela, J., & Carrillo, R. R. (2020). Reconfigurable cyber-physical system for critical infrastructure protection in smart cities via smart video-surveillance. *Pattern Recognition Letters, 140*, 303–309.
26. NVIDIA. (2020). *Jetson Nano developer kit*. Retrieved February 15, 2021, from https://developer.nvidia.com/embedded/jetson-nano-developer-kit
27. Azimi, I., Rahmani, A. M., Liljeberg, P., & Tenhunen, H. (2017). Internet of things for remote elderly monitoring: a study from user-centered perspective. *Journal of Ambient Intelligence and Humanized Computing, 8*(2), 273–289.
28. Meng, L., Miao, C., & Leung, C. (2017). Towards online and personalized daily activity recognition, habit modeling, and anomaly detection for the solitary elderly through unobtrusive sensing. *Multimedia Tools and Applications, 76*(8), 10779–10799.
29. Pomante, L., Palumbo, F., Rinaldi, C., Valente, G., Sau, C., Fanni, T., Van Der Linden, F., Basten, T., Geilen, M., Peeren, G., et al. (2020). Design and management of image processing pipelines within CPS: 2 years of experience from the FitOptiVis ECSEL project. In *2020 23rd Euromicro Conference on Digital System Design (DSD)* (pp. 378–385). IEEE.
30. Zdravevski, E., Lameski, P., Trajkovik, V., Kulakov, A., Chorbev, I., Goleva, R., Pombo, N., & Garcia, N. (2017). Improving activity recognition accuracy in ambient-assisted living systems by automated feature engineering. *IEEE Access, 5*, 5262–5280.
31. Torti, E., Fontanella, A., Musci, M., Blago, N., Pau, D., Leporati, F., & Piastra, M. (2019). Embedding recurrent neural networks in wearable systems for real-time fall detection. *Microprocessors and Microsystems, 71*, 102895.
32. Rivero-Espinosa, J., Iglesias-Pérez, A., Gutiérrez-Duenas, J. A., & Rafael-Palou, X. (2013). SAAPHO: An AAL architecture to provide accessible and usable active aging services for the elderly. *ACM SIGACCESS Accessibility and Computing, 107*, 17–24.
33. Karvonen, H., Matilainen, A., & Niemelä, V. (2019). Remote activity monitoring using indirect sensing approach in assisted living scenario. In *2019 13th International Symposium on Medical Information and Communication Technology (ISMICT)* (pp. 1–6). IEEE.
34. Hassan, M. M., Uddin, M. Z., Mohamed, A., & Almogren, A. (2018). A robust human activity recognition system using smartphone sensors and deep learning. *Future Generation Computer Systems, 81*, 307–313.
35. Poppe, R. (2010). A survey on vision-based human action recognition. *Image and vision computing, 28*(6), 976–990.
36. Zhang, H. B., Zhang, Y. X., Zhong, B., Lei, Q., Yang, L., Du, J. X., & Chen, D. S. (2019). A comprehensive survey of vision-based human action recognition methods. *Sensors, 19*(5), 1005.
37. Nascimento, B., Oliveira, T., & Tam, C. (2018). Wearable technology: What explains continuance intention in smartwatches? *Journal of Retailing and Consumer Services, 43*, 157–169.
38. Wang, J., Chen, Y., Hao, S., Peng, X., & Hu, L. (2019). Deep learning for sensor-based activity recognition: A survey. *Pattern Recognition Letters, 119*, 3–11.
39. Edel, M., & Köppe, E. (2016). Binarized-BLSTM-RNN based human activity recognition. In *2016 International Conference on Indoor Positioning and Indoor Navigation (IPIN)* (pp. 1–7). IEEE.
40. Gers, F. A., Schmidhuber, J., & Cummins, F. (1999). Learning to forget: Continual prediction with LSTM.
41. Yang, J., Nguyen, M. N., San, P. P., Li, X., & Krishnaswamy, S. (2015). Deep convolutional neural networks on multichannel time series for human activity recognition. In *IJCAI* (Vol. 15, pp. 3995–4001). Buenos Aires, Argentina.
42. Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning spatiotemporal features with 3D convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 4489–4497).
43. Hara, K., Kataoka, H., & Satoh, Y. (2017). Learning spatio-temporal features with 3D residual networks for action recognition. In *Proceedings of the IEEE International Conference on Computer Vision Workshops* (pp. 3154–3160).
44. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770–778).
45. Diba, A., Fayyaz, M., Sharma, V., Karami, A. H., Arzani, M. M., Yousefzadeh, R., & Van Gool, L. (2017). Temporal 3D ConvNets: New architecture and transfer learning for video classification. Preprint retrieved from http://arxiv.org/abs/1711.08200
46. Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4700–4708).
47. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1–9).
48. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al. (2017). The kinetics human action video dataset. Preprint retrieved from https://arxiv.org/abs/1705.06950
49. Thrun, S., & Pratt, L. (1998). Learning to learn: Introduction and overview. In *Learning to Learn* (pp. 3–17). Springer.

50. Raj, C., Jain, C., & Arif, W. (2017). Heman: Health monitoring and nous: An IoT based e-health care system for remote telemedicine. In *2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)* (pp. 2115–2119). IEEE.
51. Angelucci, A., Kuller, D., & Aliverti, A. (2020). A home telemedicine system for continuous respiratory monitoring. *IEEE Journal of Biomedical and Health Informatics*.
52. Adochiei, F., Rotariu, C., Ciobotariu, R., & Costin, H. (2011) A wireless low-power pulse oximetry system for patient telemonitoring. In *2011 7th International Symposium on Advanced Topics in Electrical Engineering (ATEE)* (pp. 1–4). IEEE.
53. Joseph, G., Joseph, A., Titus, G., Thomas, R. M., & Jose, D. (2014). Photoplethysmogram (PPG) signal analysis and wavelet de-noising. In *2014 Annual International Conference on Emerging Research Areas: Magnetics, Machines and Drives* (AICERA/iCMMD) (pp. 1–5). IEEE.
54. Jung, S. J., Lee, Y. D., Seo, Y. S., & Chung, W. Y. (2008). Design of a low-power consumption wearable reflectance pulse oximetry for ubiquitous healthcare system. In *2008 International Conference on Control, Automation and Systems* (pp. 526–529). IEEE.
55. Biswas, D., Everson, L., Liu, M., Panwar, M., Verhoef, B. E., Patki, S., Kim, C. H., Acharyya, A., Van Hoof, C., Konijnenburg, M., et al. (2019). Cornet: Deep learning framework for PPG-based heart rate estimation and biometric identification in ambulant environment. *IEEE Transactions on Biomedical Circuits and Systems, 13*(2), 282–291.
56. Reiss, A., Indlekofer, I., Schmidt, P., & Van Laerhoven, K. (2019). Deep PPG: Large-scale heart rate estimation with convolutional neural networks. *Sensors, 19*(14), 3079.
57. Jindal, V., Birjandtalab, J., Pouyan, M. B., & Nourani, M. (2016). An adaptive deep learning approach for PPG-based identification. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (pp. 6401–6404). IEEE.
58. Lee, M. S., Lee, Y. K., Pae, D. S., Lim, M. T., Kim, D. W., & Kang, T. K. (2019). Fast emotion recognition based on single pulse PPG signal with convolutional neural network. *Applied Sciences, 9*(16), 3355.
59. Tajbakhsh, N., Shin, J. Y., Gurudu, S. R., Hurst, R. T., Kendall, C. B., Gotway, M. B., & Liang, J. (2016). Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE Transactions on Medical Imaging, 35*(5), 1299–1312.
60. Frias-Martinez, E., Magoulas, G., Chen, S., & Macredie, R. (2005). Modeling human behavior in user-adaptive systems: Recent advances using soft computing techniques. *Expert Systems with Applications, 29*(2), 320–329.
61. Anaby-Tavor, A., Carmeli, B., Goldbraich, E., Kantor, A., Kour, G., Shlomov, S., Tepper, N., & Zwerdling, N. (2020). Do not have enough data? Deep learning to the rescue! In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 34, pp. 7383–7390).
62. Chaquet, J. M., Carmona, E. J., & Fernández-Caballero, A. (2013). A survey of video datasets for human action and activity recognition. *Computer Vision and Image Understanding, 117*(6), 633–659.
63. NVIDIA. (2020). *TensorRT*. Retrieved February 27, 2021, from https://developer.nvidia.com/tensorrt

Daniel Deniz is with the Department of Computer Architecture and Technology at the University of Granada, Spain. He received the M.Sc. degree in Data Science and Computer Engineering in 2019 and is currently a PhD Student at the University of Granada. He works on Computer Vision, focusing on machine learning solutions for distributed edge-cloud processing.

Gabriel Jimenez-Perera is with the Department of Computer Engineering, Automatics and Robotics at the University of Granada, Spain. He received the M.Sc. degree in Data Science and Computer Engineering in 2018 and is currently a PhD Candidate. He works on federated learning, focusing on computer vision and embedded systems.

Ricardo Nolasco is the head of the "New Products" department and project manager at RGB Medical Devices. He has a degree in Computer Science Engineering from the ETSIInf of the UPM and has work experience in different departments of RGB Medical Devices since 2015.

Javier Corral is the head of the Medical Instrumentation Research Department of the company RGB Medical Devices, and is specialized in Hardware and Software development of medical equipment. He has a degree in Industrial Engineering from the ETSIIM of the UPM and has work experience in different departments of RGB Medical Devices since 1990.

Francisco Barranco received his M.Sc. degree in Computer and Network Engineering, and his Ph.D. degree in Computer Engineering from the University of Granada, Spain, in 2008 and 2012, respectively. Currently, he is an Associate Professor with the Department of Computer Architecture and Technology, University of Granada. His research interests include robotics, embedded real-time machine vision, bio-inspired processing, and cognitive vision.