No Need for a Lab: Towards Multi-sensory Fusion for Ambient Assisted Living in Real-world Living Homes

Alessandro Masullo[a], Toby Perrett[b], Dima Damen[c], Tilo Burghardt and Majid Mirmehdi[d]

University of Bristol, BS8 1UB, Bristol, U.K.

Keywords: Multi-sensory Fusion, Ambient Assisted Living, Silhouettes, Wearable Devices, Acceleration.

Abstract: The majority of the Ambient Assisted Living (AAL) systems, designed for home or lab settings, monitor one participant at a time – this is to avoid the complexities of pre-fusion correspondence of different sensors since carers, guests, and visitors may be involved in real world scenarios. Previous work from (Masullo et al., 2020) presented a solution to this problem that involves matching video sequences of silhouettes to accelerations from wearable sensors to identify members of a household while respecting their privacy. In this work, we elevate this approach to the next stage by improving its architecture and combining it with a tracking functionality that makes it possible to be deployed in real-world homes. We present experiments on a new dataset recorded in participants' own houses, which includes multiple participants visited by guests, and show an au-ROC score of 90.2%. We also show a novel first example of subject-tailored health monitoring measurement by applying our methodology to a sit-to-stand detector to generate clinically relevant rehabilitation trends.

1 INTRODUCTION

The world is getting old. Continuously improving medical technologies and healthcare systems are contributing to extend life expectancy more than ever before, with the effect that the world's demographic of age 60 or more is expected to double in the next 30 years (Patel and Shah, 2019). This shift in age distribution is accompanied by the demand of an independent lifestyle that can be met by novel technologies through unobtrusive monitoring and the use of artificial intelligence (AI). To this end, Ambient Assisted Living (AAL) is a field of research aimed at developing an ecosystem of sensors that work in cooperation to help monitoring elderly people and their health to live independently (Rashidi and Mihailidis, 2013).

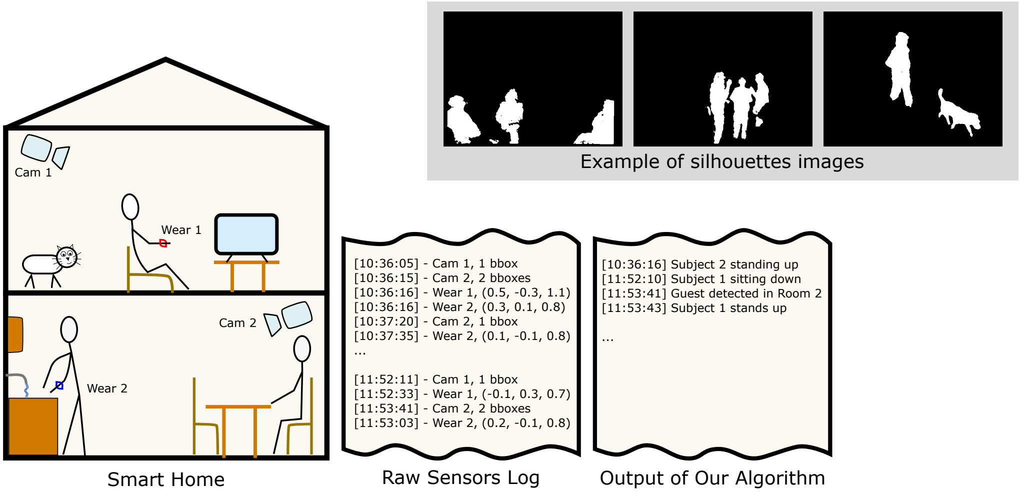
A multitude of sensors may be typically involved in AAL applications, for example wearable accelerometers as the most common of all, Passive Infrared (PIR) sensors (Cook et al., 2013), floor vibration sensors (Dobbler et al., 2014), ambient sensors (like temperature, humidity, power consumption) (Zhu et al., 2015), and a variety of health sensors (heart rate, respiratory rate, VO2 max) (Calvaresi et al., 2017). Videos recorded from camera sensors, especially with the RGB data transformed into relatively anonymous silhouettes, are also being increasingly employed outside of lab conditions in a number of AAL applications, e.g. for fall detection (Akagunduz et al., 2017), the measurement of calorie expenditure (Masullo et al., 2018) and the analysis of transitions from sitting to standing while recovering from surgery (Masullo et al., 2019). Video is indeed a powerful tool in the AAL armoury, e.g. it is a routine part of the multi-platform SPHERE (Sensor Platform for Healthcare in a Residential Environment) system (Zhu et al., 2015).

Each of the sensor types comes with its own pros and cons. For example, cameras are considered the richest source of information, but a significant number may be necessary for full coverage of the home, while also invoking massive data storage needs. Wearable devices can be carried everywhere, but require user interaction to put them on and charge their batteries regularly, and cannot provide data rich enough to deal with the complexities of human behaviour. Other ambient sensors, such as PIR and door opening detectors are inherently passive but often only allow for limited and specific exploration.

In order to enable any AAL system to provide useful trends for health measurements, it is essential

[a] https://orcid.org/0000-0002-6510-835X
[b] https://orcid.org/0000-0002-1676-3729
[c] https://orcid.org/0000-0001-8804-6238
[d] https://orcid.org/0000-0002-6478-1403

Masullo, A., Perrett, T., Damen, D., Burghardt, T. and Mirmehdi, M.
No Need for a Lab: Towards Multi-sensory Fusion for Ambient Assisted Living in Real-world Living Homes.
DOI: 10.5220/0010202903280337
In *Proceedings of the 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2021) - Volume 5: VISAPP*, pages 328-337
ISBN: 978-989-758-488-6



```
[10:36:05] - Cam 1, 1 bbox
[10:36:15] - Cam 2, 2 bboxes
[10:36:16] - Wear 1, (0.5, -0.3, 1.1)
[10:36:16] - Wear 2, (0.3, 0.1, 0.8)
[10:37:20] - Cam 2, 1 bbox
[10:37:35] - Wear 2, (0.1, -0.1, 0.8)
...

[11:52:11] - Cam 1, 1 bbox
[11:52:33] - Wear 1, (-0.1, 0.3, 0.7)
[11:53:41] - Cam 2, 2 bboxes
[11:53:03] - Wear 2, (0.2, -0.1, 0.8)
```

```
[10:36:16] Subject 2 standing up
[11:52:10] Subject 1 sitting down
[11:53:41] Guest detected in Room 2
[11:53:43] Subject 1 stands up

...
```

Figure 1: Application scenario. A Smart Home including cameras recording silhouettes and wearable devices carried by participants. Participants can have guests or pets (see example silhouette images) and live their lives normally in their own homes. Our method can distinguish participants from the silhouettes in the scene and assign each detected activity to a specific monitored subject.

to monitor individual subjects over long periods of time. The system must be able to assign measurements from every sensor to each individual member of the household and discard any data generated by guests, caregivers, pets, and so on. To the best of the authors' knowledge, no current AAL system can deal with such complexity, as the majority of AAL frameworks circumvent the problem by using home-like lab environments or are deployed in single-occupancy houses (Zouba et al., 2009; Skubic et al., 2009; Kurz et al., 2012; Cook et al., 2013; Amato et al., 2016; Holthe et al., 2018; Caroux et al., 2018; Das et al., 2019). This is further discussed in a recent review of AAL systems (Climent-Pérez et al., 2020), which highlights users' concerns in RGB video monitoring systems deployed in peoples homes with respect to mass surveillance and lack of privacy.

In this paper, we present a solution to the long-term monitoring problem for AAL systems which can be deployed in real multi-occupancy houses, as shown in Figure 1. A recent work on video-acceleration fusion (Masullo et al., 2020) provided a ReID capability for actively moving silhouettes in terms of wearable IDs. We now push this work to the next stage by implementing a tracking functionality that propagates the ReID capability to untrimmed videos of any activity content. We show its performance on a new dataset recorded in real homes using the SPHERE video monitoring system (Zhu et al., 2015; Woznowski et al., 2015; Hall et al., 2016). The dataset used in this study is a subset of the houses that volunteered for installing the system in their own habitation; it includes video silhouettes from three different rooms of each house and the accelerations form a wearable device carried by the participants[1]. Further, we also present an example application of subject-tailored analysis of Sit-to-Stand trend plot for a participant who underwent hip/knee replacement surgery. The dual nature of privacy sensitivity through silhouettes and multi-sensory fusion, combined with the maximum level of spontaneity of our participants being recorded in their own homes, make our analysis unique in the field of AAL.

Next in Section 2, we review related works, followed by a review of our dataset in Section 3. In Sections 4 and 5, we present the proposed approach and its evaluation, respectively, and conclude in Section 6.

2 RELATED WORK

The majority of AAL systems do not focus on a single type of sensor but embrace a wide variety of modalities through sensor fusion. Earlier works like the GERHOME project (Zouba et al., 2009) or CASAS (Cook et al., 2013) used a variety of contact sensors on doors, windows and cabinets, together with pressure sensors installed on chairs and power/water consumption, to recognise or discover the activities performed in a home-lab environment. More recently, Holthe et *al.* developed a similar platform of PIR and magnetic sensors that was directly deployed in an elderly volunteer's house (Holthe et al., 2018) and a similar approach was followed in (Caroux

[1] The dataset and code will be provided in a future release on https://github.com/ale152/no-need-for-a-lab

et al., 2018). The very recent project from Toyota Smarthome (Das et al., 2019) pushed the data capture to the next stage by collecting fully unscripted RGB+D data using Kinect sensors installed in a home-lab setting. However, although RGB+D data is the richest form in terms of information, it poses a series of privacy challenges as recently highlighted (Climent-Pérez et al., 2020). These concerns were addressed in the development of the SPHERE project (Hall et al., 2016), which replaced RGB data with binary silhouettes, as a good trade-off between information content of data and respect for privacy.

In spite of the multitude of AAL projects and sensor frameworks developed over the past years, the majority of them have either been deployed in a home-lab setting or involved a single participant living in the home. The DOREMI project (Bacciu et al., 2016) had the particular focus of indoor socialization events and built an automatic system for detecting guests entering and leaving the house. The system was based on a simple classifier of multiple PIR activations and it only provided an approximate indicator of social inclusion. Moreover, due to its simplicity, this approach is unable to provide a real connection between readings of different sensor modalities and participants generating the data. In a recent work, a more advanced approach was presented, where the problem of association between wearable devices and their wearers in video is tackled in the challenging scenario of crowd mingling events (Cabrera-Quiros and Hung, 2019).

The recent work from Masullo et *al.* showed that a deep learning approach to the video-acceleration matching problem (Masullo et al., 2020) produced even better performances than (Cabrera-Quiros and Hung, 2019) on the challenging SPHERE Calorie dataset (Tao et al., 2017). However, while the SPHERE Calorie dataset includes a variety of people in a regular home performing specific actions, it still remains an acted dataset by volunteers. In this paper, we build on the video-acceleration matching for subject ReID (Masullo et al., 2020) by implementing a tracking functionality that allows the system to be deployed in real homes. Together with an improved version of the network architecture, our new algorithm can work on real unscripted data. Our new dataset constitutes a much more difficult challenge as it involves the maximum level of spontaneity as our subjects are living in their own houses and spend a vast part of their time sitting, laying down and generally resting. Moreover, the level of activity, habits and view point for each house are very different from each other, making this dataset even more challenging.

3 THE DATASET

SPHERE is a multi-modal sensor platform designed to provide an intelligent residential space for health monitoring (Woznowski et al., 2015). The project recruited participants who were willing to install the SPHERE platform in their homes, for up to one year. The sensors of the SPHERE platform consist of three different groups: video cameras (used to generate silhouettes) (Hall et al., 2016), a single accelerometer device (Fafoutis et al., 2016) worn by each participant (recording x-y-z acceleration and RSSI signal) and a variety of ambient sensors measuring temperature, humidity, light, power consumption, water consumption and so on (Zhu et al., 2015). This work focuses only on the first two types of modalities, video silhouettes and accelerations, selected from a subset of the homes that were recorded.

The data was recorded in a variety of household sizes and across a wide health spectrum. The completely free living setting involved in the SPHERE project makes the analysis of our data much more challenging than other similar (AAL) projects as all houses are different from each other and our volunteers continued with their normal lives while being monitored, which involved visitors, pets, forgetting to charge their wearable (or to put it on), and even changing furniture arrangements.

3.1 Data Collection and Filtering

When a house is continuously populated by multiple people, it is extremely hard to match the wearable sensor readings to the silhouettes in the scene. Hence, *for generating ground truth*, we selected 4 houses that were populated by single individuals and no pets (H1, H2, H3 and H4), and 1 additional house with multiple occupancy (H5) for testing. This resulted in a total of 38001 pairs of labelled video and acceleration clips for training/validation and 6909 of additional unlabelled pairs for testing. For details, please see Table 1.

Table 1: Some details of the houses H1-H5 in our dataset.

	H1	**H2**	**H3**	**H4**	**H5**
Days Obs.	153	64	104	40	83
Hrs. Rec.	240	28	41	8	58
# clips	28856	3293	4949	903	6909

Since the data was recorded in a completely unscripted fashion, we applied a combination of filters, described below, to produce a ground truth in terms of matching video-acceleration pairs.

Acceleration Magnitude Persistence — We implemented a filter on the acceleration intensity to make sure that the wearable device was indeed being worn.

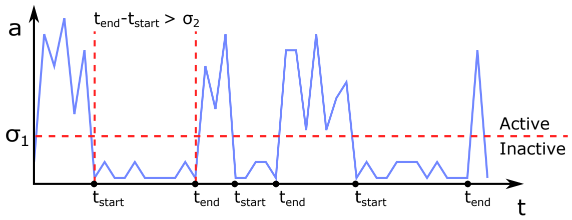


Figure 2: Description of the acceleration filter to discard segments of data where the device is not being worn.

Whenever the acceleration magnitude stayed below a certain threshold for a long time, it is likely that the wearable was not being worn and the corresponding video/acceleration segment pairs should be discarded. Let's consider an acceleration stream,

$$\mathbf{A}' = \{(\mathbf{a}_1,t_1),(\mathbf{a}_2,t_2),...,(\mathbf{a}_n,t_n)\}\,, \tag{1}$$

where $\mathbf{a_i}$ is a vector of the x, y and z components of the acceleration, and t_i is the corresponding timestamp. Inactive segments of the acceleration can at first be identified where the magnitude is lower than a threshold σ_1, i.e.,

$$\mathbf{A}'_{\text{inactive}} = \left\{|\mathbf{a}_i| < \sigma_1, (\mathbf{a}_i,t_i) \in \mathbf{A}'\right\}. \tag{2}$$

If we define the first and last timestamp of each inactive segment as t_{start} and t_{end} respectively, as depicted in Figure 2, we then further examine $\mathbf{A}'_{\text{inactive}}$ to discard those inactive accelerations that persist for longer than a certain time threshold σ_2, i.e.

$$\mathbf{A}'_{\text{discard}} = \left\{t_{\text{end}} - t_{\text{start}} > \sigma_2, (\mathbf{a}_i,t_i) \in \mathbf{A}'_{\text{inactive}}\right\}. \tag{3}$$

All the video/acceleration pairs whose accelerations belong to $\mathbf{A}'_{\text{discarded}}$ are removed from further processing to retain our final set of accelerations $\mathbf{A}$, i.e.

$$\mathbf{A} = \mathbf{A}' - \mathbf{A}'_{\text{discard}}\,. \tag{4}$$

Video-acceleration Sync — Since the accelerations are only recorded when the wearable device is in the range of the house, we filtered to retain the video data by the intersection of timestamps recorded from the video and the wearable devices. This ensures that video data is discarded when the participant is not at home. Let's define a video stream

$$\mathbf{V}' = \{(\mathbf{v}_1,\tau_1),(\mathbf{v}_2,\tau_2),...,(\mathbf{v}_{n_T},\tau_{n_T})\}\,, \tag{5}$$

where $\mathbf{v_i}$ is a silhouette and τ_i is the corresponding timestamp. Considering the acceleration stream defined in Eq. (1), we can filter the video sequences as

$$\mathbf{V}'_{\text{discard}} = \left\{|t_i - \tau_j| > \sigma_3, (\mathbf{v}_j,\tau_j) \in \mathbf{V}', (\mathbf{a}_i,t_i) \in \mathbf{A}\right\}, \tag{6}$$

such that the difference in synchronisation between the two streams is higher than σ_3. The final set of videos clips used for training can be defined as:

$$\mathbf{V} = \mathbf{V}' - \mathbf{V}'_{\text{discard}}. \tag{7}$$

Bounding Box Motion — Our matching algorithm relies on the bounding box detector and tracker implemented in (Hall et al., 2016), which detects all the people that appear in front of the camera. To remove false positive bounding boxes, which are usually stationary objects, we discard all the bounding boxes with a speed lower than σ_4. Hence, given the bounding box centre position $B = (b_x, b_y)$, then it is filtered out if

$$\sqrt{b'^2_x + b'^2_y} < \sigma_4. \tag{8}$$

Once the data was prepared, our network was trained and made ready to be applied to any household of any size without any real limitation.

4 METHODOLOGY

Let us consider a typical AAL house occupied by one or more subjects and frequently visited by guests (e.g. relatives, caretaker, plumber, and so on). If the monitored participants are carrying a wearable, the acceleration measurements can be directly associated with its carrier[2], while sensor measurements from the video cameras and the rest of the environmental sensors cannot. In order to enable a subject-tailored multi-sensory analysis, such measurements need to be assigned to specific individuals and distinguished from guest-triggered readings, and here is where we make our contribution.

Based on the work in (Masullo et al., 2020), where a novel deep learning approach was proposed to solve the video-acceleration matching problem, we use additional tracking information to extend its functionality even further. Our algorithm is able to match untrimmed video clips of silhouettes to temporal segments of accelerations and compute the distance between the two to assign each anonymous video clip to a wearable ID (and henceforth to a specific subject).

Next, we briefly summarise the previous work from (Masullo et al., 2020) on video-acceleration matching and then present the details of how we build upon that method.

4.1 Video-acceleration Matching

As postulated in (Masullo et al., 2020), the problem of Video-Acceleration Matching can be handled via a triplet-loss formulation (as typical in face recognition (Schroff et al., 2015)), where the triplet comprises a video clip of silhouettes $\mathbf{V}$ (the anchor) and two acceleration segments $\mathbf{A}_p$ and $\mathbf{A}_n$ (respectively positive

[2] Apart from the unlikely scenario where participants exchange their wearable device.

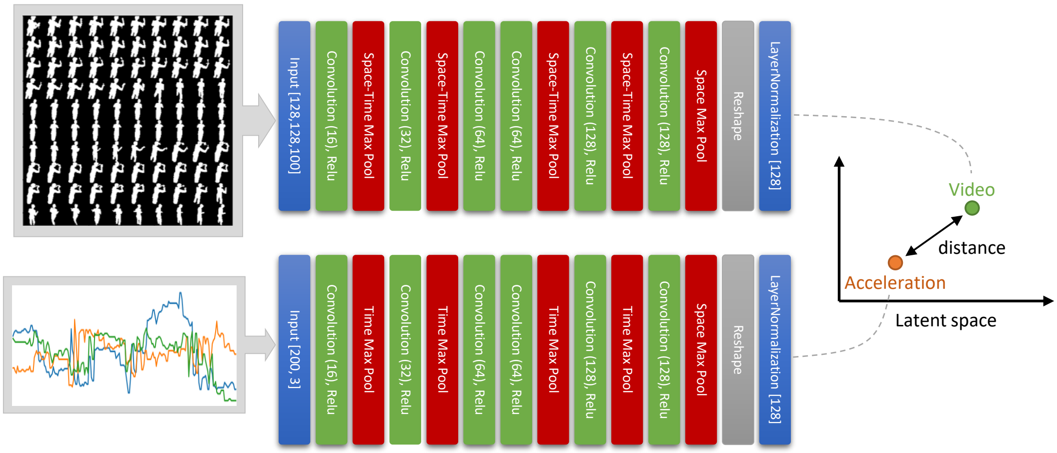


Figure 3: Illustration of the network architecture adopted in our work. Two distinct branches process independently the video silhouettes and the acceleration so that they can be compared on a latent space.

and negative matching samples)

$$(\text{Anchor}, \text{Positive}, \text{Negative}) = (\mathbf{V}, \mathbf{A}_p, \mathbf{A}_n) \ . \qquad (9)$$

During training, the actual matching acceleration is used as a positive sample, whereas a different acceleration from a different wearable is used as a negative sample (more on the choice of negative samples later). Two distinct neural networks $f(\cdot)$ and $g(\cdot)$ encode the videos and the accelerations respectively into an embedding space where the Euclidean distance between the two can be measured to assert the matching

$$d(\mathbf{V},\mathbf{A}) = \sqrt{\sum [f(\mathbf{V}) - g(\mathbf{A})]^2} \ . \qquad (10)$$

Reciprocal Triplet Loss (Masullo et al., 2020) (RTL) is used to train the networks and a threshold is finally employed to discriminate between matching and non-matching pairs as

$$L_{\mathrm{RTL}} = |f(\mathbf{V}) - g(\mathbf{A}_p)|^2 + \frac{1}{|f(\mathbf{V}) - g(\mathbf{A}_n)|^2} \ . \qquad (11)$$

4.2 Network Architecture

The results of previous work (Masullo et al., 2020) showed that a 3D fully convolutional network is the best candidate to encode both video and acceleration data, so we selected the *fully-conv* architecture from (Masullo et al., 2020) and improved it to better generalise to real data. The main changes to the network architecture are in the very last layer, where the *tanh* activation previously used was replaced by a combination of *relu* and Layer Normalization layers, as described in (Ba et al., 2016). The Layer Normalization not only reduces the training time, but it also improves over the validation error providing a more general model that helps fusing the video and the acceleration data streams. In addition, we also increased the size of the frame input from 100x100 pixels to 128x128 pixels, allowing for more details in the video to be exploited for the analysis. This change was also paired with an additional max-pooling layer at the end of the network to compensate for the bigger input size. Figure 3 illustrates for our network architecture.

4.3 Temporal Structure

To improve the matching algorithm to handle in-the-wild data, we increased the temporal window used to match the video and acceleration segments. As already observed in their experiments in (Masullo et al., 2020), longer observation times lead to a drastic increase in the matching performances; however, the 100 frames previously used in (Masullo et al., 2020) for the video clips were not sufficient to enable discrimination between different subjects in a real scenario. Due to the limitations of GPU memory size, increasing the number of frames for the video clips was not viable and we therefore opted for increasing the observation window by a factor of 2 while introducing a frame sub-sampling factor of 2. The observation window of the acceleration stream was doubled without sub-sampling and without significant memory impact.

4.4 Tracking

Our last and most important improvement to the Video-Acceleration Matching algorithm in (Masullo et al., 2020) to deal with in-the-wild data is tracking, as illustrated in Figure 4. The tracklet ID for each user provided by the SPHERE system (Hall et al., 2016) is consistent while the subject is in view, and sometimes even preserved when the user only briefly disappears from the frame - however, it changes to a new ID on a later re-appearance of the subject. Since each tracklet is assigned to a specific individual, we match a wearable with an entire tracklet itself. To do that, we divide each tracklet into N_{ovp} overlapping video and acceleration segments $(\mathbf{V}_i, \mathbf{A}_i)$ and compute the matching distance d as per Eq. (10) across the entire tracklet,

$$\mathbf{D} = \left\{ d(\mathbf{V}_1, \mathbf{A}_1), d(\mathbf{V}_2, \mathbf{A}_2), ..., d(\mathbf{V}_{N_{\text{ovp}}}, \mathbf{A}_{N_{\text{ovp}}}) \right\}. \tag{12}$$

The result is a vector $\mathbf{D}$ of N_{ovp} elements containing the matching distance for each short video/acceleration clip. Based on this vector, we developed three different strategies that perform the matching based on the (1) minimum, $d_{\text{min}} = \min(\mathbf{D})$, (2) median, $d_{\text{median}} = \text{median}(\mathbf{D})$, and (3) average distance, $d_{\text{mean}} = \text{mean}(\mathbf{D})$, across the vector.

The heuristic hypothesis behind the use of tracklets to improve the matching is that users spend most of the time at home sitting or sedentary, while the Video-Acceleration Matching algorithm (Masullo et al., 2020) is most effective when subjects are moving. Therefore, if a subject is sitting for a long period of time, using the tracklet we can identify them either while walking to a chair or while standing up from it, and then propagate their ID to the rest of the monitoring sequence.

4.5 Negative Samples

As we described in Section 4.1, training the Video-Acceleration Matching network requires the definition of a triplet constituted by a video clip anchor $\mathbf{V}$, a positive acceleration $\mathbf{A}_p$ and a negative acceleration $\mathbf{A}_n$. While the choice of the $(\mathbf{V}, \mathbf{A}_p)$ pair is trivial (each video clip only has one acceleration that is matching with it), the negative acceleration $\mathbf{A}_n$ can have a large effect on the overall results and different strategies need to be tested. Considering a set of k subjects $S = (S_1, ..., S_k)$ recorded for m days $(\text{day}_1, ..., \text{day}_m)$, a negative acceleration $\mathbf{A}_n$ can be chosen from the Same Subject on the Same Day (SSSD), Same Subject on a Different Day (SSDD), or a Different Subject from a different house (DS). In addition, we can also choose a negative sample that is simply overlapping with the positive sample (OVLP). The possible negative samples are summarised for simplicity in Table 2. In our experiments, we tested all the different strategies to find the most advantageous one.

Table 2: Description of possible negative samples for triplet learning.

	Same Day (SD)	**Diff Day (DD)**
Same Sub. (SS)	*SSSD*	*SSDD*
Diff. Sub. (DS)	*DS*	*DS*
Overlap	*OVLP*	*OVLP*

5 RESULTS

We now present a series of results and ablation studies to show the improvements of the Video-Acceleration Matching algorithm. To measure the performances of each trained model, we use the area under the ROC curve (auROC). If we define True Positive (TP) and False Positive (FP) as:

$$\begin{aligned} TP(\beta) = \{ \, (\mathbf{V}_i, \mathbf{A}_j) \, | f(\mathbf{V}_i)^2 - g(\mathbf{A}_j)^2 < \beta, \\ (\mathbf{V}_i, \mathbf{A}_j) \in P \, \} \, , \end{aligned} \tag{13}$$

$$\begin{aligned} FP(\beta) = \{ \, (\mathbf{V}_i, \mathbf{A}_j) \, || f(\mathbf{V}_i)^2 - g(\mathbf{A}_j)^2 < \beta, \\ (\mathbf{V}_i, \mathbf{A}_j) \in Q \, \} \, , \end{aligned} \tag{14}$$

where P are the true matching pairs of videos and accelerations $(\mathbf{V}, \mathbf{A})$ and Q are the true non-matching ones, we can define True Positive Rate (TPR) and False Positive Rate (FPR) as:

$$TPR = \frac{TP}{P} \text{ and } FPR = \frac{FP}{Q} \, . \tag{15}$$

The area under the TPR v. FPR is what we refer to as auROC.

5.1 Best Negative Strategy

The first step to achieve a functional algorithm for Video-Acceleration Matching is to optimise the video and acceleration encoders described in Section 4.1. The most important aspect of this optimisation process is to find the best training strategy for the negative samples. Therefore in this experiment, we keep the tracking functionality off and focus on the performances of the encoders. As already described in Section 4.5 and in (Masullo et al., 2020), different training strategies lead to very different behaviours of the Video-Acceleration encoder and the best way to find the optimal negative strategy is to test all the possible alternatives. For convenience, we decided to train a model using Houses H1, H2 and H3, and we kept House H4 out for validation. We tested the 4 different strategies for negative samples described in Section 4.5 and we report the results in Table 3.

In addition, to better understand the contribution of our work and compare our results, we also introduced an extra step to the original method (Masullo

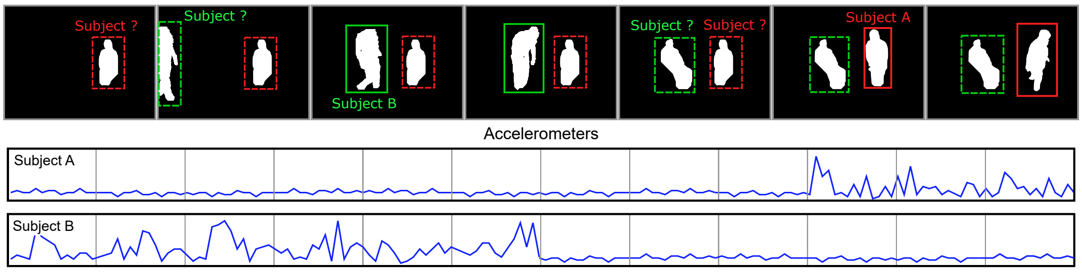


Figure 4: Illustration of the proposed algorithm. Two subjects are tracked over a long sequence. Their identity, initially unknown (dashed bounding box) is revealed when they move (solid bounding box) and propagated back using the tracking information.

et al., 2020), namely (Masullo et al., 2020)+Time as shown in Table 3, which includes our novel temporal sampling described in Section 4.3.

Table 3: Results of auROC for the Video-Acceleration encoder using different negative strategies.

Method	auROC score
(Masullo et al., 2020)	81.3%
(Masullo et al., 2020) + Time	85.1%
Ours: SSSD	80.6%
Ours: SSDD	80.2%
Ours: DS	71.9%
Ours: OVLP	**86.1%**

Results show that the best training strategy is obtained when negative samples are overlapping with the anchor (OVLP), with an auROC of 86.1%. The optimal strategy found in this work differs from the optimal strategy discovered in (Masullo et al., 2020) on the Calorie dataset and it is to be expected. In fact, the Calorie dataset is constituted by a set of acted actions, for which a negative strategy that leads the model to learn the different activities is the most efficient solution. When dealing with unscripted data from SPHERE, the network cannot exploit any other information than the actual correlation between the videos and the acceleration to disclose the matching, which explains why an overlapping strategy produces such a high auROC score.

A comparison with previous work also reveals that our encoder, even without the tracking functionality, produces better results in terms of auROC score. We can see that even implementing just the temporal sampling already contributes to an improvement of 3.8 percentage points. In Section 5.4 we will see that this improvement is even more remarkable once we introduce the tracking functionality.

5.2 Cross-validation of Houses

In order to further assess the results from Section 5.1, we performed an experiment of leave-one-out cross-validation using the 4 labelled houses that we have available. We selected the best negative strategy of OVLP and trained 3 different models using only 3 of the houses available and leaving a different one out for validation for each model. Results are presented in Figure 5 in terms of ROC curves. The auROC scores, highlighted in the legend, support our findings from Section 5.1, with an average auROC score across all the houses of 81.5%.

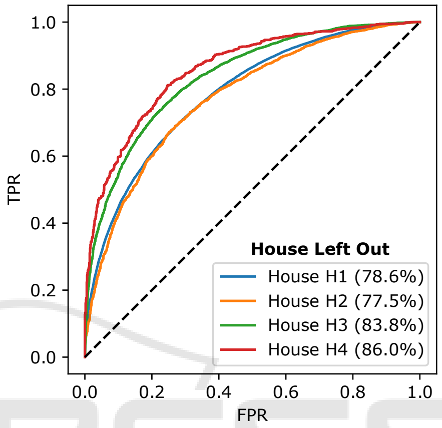


Figure 5: ROC curves for cross-validation using the OVLP negative strategy. Each curve represents the results for a specific validation House left out from the training process. The average auROC across all tested models is 81.5%.

5.3 Test for Number of Houses

Next, we investigated how the number of houses used during training affected the performances of our Video-Acceleration encoder. We kept one house at a time out for validation, as in Section 5.2, and used an increasing number of houses for training among the remaining three. The results, plotted for each left-out house in Figure 6 illustrate that, as expected, increasing the number of houses used for training leads to better performance. It is also interesting to note that the performance improvement is particularly steep when House H1 is left out for validation. This can be explained by noting that House H1 (see Table 1) is the

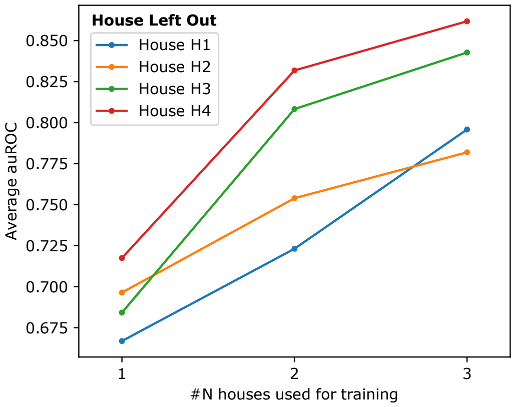


Figure 6: Test for variable number of houses used in training. Each house is kept out for validation and the remaining houses are used for training in increasing size.

largest house in terms of data collected and therefore leaving it out of the training set leads to a model with the poorest performances. Results for House H4 confirm this observation presenting the highest auROC score and constituting the smallest house that is left out of the training data.

5.4 Results for Tracking Sequences

Finally, we test the performance of our full proposed method, including the tracking functionality described in Section 4.4. For this test, we trained on houses H1, H2, and H3 using the best negative strategy OVLP and tested using entire tracklets on house H4. As described in Section 4.4, the output of our algorithm is a vector of N_{ovp} distances between video clips and accelerations evaluated for each tracklet. The final decision on the matching can be taken by aggregating the tracklet vector in a single number by considering its mean, median or minimum value. In order to study the performance of the algorithm, we simulate a scenario where two wearables are detected in the same house but only one person is in front of the camera and we want to know who that person is. We achieve this by comparing each video sequence both to the matching acceleration and to the acceleration of the same person at a different time, as already done in previous studies (Shigeta et al., 2008; Masullo et al., 2020). We then calculate the auROC score for all *mean*, *median* and *minimum* strategies as seen in Table 4. For convenience, we also report here the comparison results from (Masullo et al., 2020) of Table 3.

We notice how the *minimum* strategy is the least effective, since it maximises the chances that a single clip in the tracklet is erroneously matched between video and acceleration. The mean and median strategies are comparable in performance, although the former has a stronger discrimination power with an auROC score of 90.2%. When comparing our results with previous works, we can see that even without the tracking functionality our algorithm is already able to achieve 86.1% auROC agasint the 81.3% of (Masullo et al., 2020). When we introduce tracking, the auROC jumps up to 90.2%, outperforming previous works by 8.9 percentage points.

Table 4: Results of auROC for different tracking methods.

Method	auROC score
(Masullo et al., 2020)	81.3%
(Masullo et al., 2020) + time	85.1%
Ours: Tracking minimum	77.3%
Ours: No tracking	86.1%
Ours: Tracking median	89.1%
Ours: Tracking mean	**90.2%**

5.5 Tailored Sit-to-Stand Measurements

Finally, in order to provide with a real application of our novel video-acceleration matching algorithm, we tested our methodology on House H5, which is completely new to the model and was not filtered or labelled in any way. The house was occupied by two participants, only one of whom was being monitored using the accelerometer. The monitored subject underwent a hip/knee replacement during the experiment and we analysed their transition from sitting to standing to study their recovery progress over time.

All the silhouettes recorded in the house were analysed using the sit-to-stand detection algorithm from (Masullo et al., 2019), resulting in the trend plot in Figure 7, which depicts the stand-up speed as a proxy of musculoskeletal functionality. This plot is an ensemble of all the sit-to-stand transitions detected in the house, including the monitored subject, their partner and any potential guest, which results in particularly large error-bars of 0.0747 m/s. In spite of the error, the plot still reveals a clear decay of the stand-up speed in the 2 weeks soon after the surgery, followed by a slow but steady increase in the following months. This behaviour is to be expected since the surgery patient is impaired post-operation and tends to be more careful standing up, resulting in lower stand up speeds. Once the wounds start healing, the patient regains confidence and starts standing up more quickly, reaching a final stand up speed that is higher than before the surgery.

In order to obtain a more accurate trend for the recovery of the monitored patient, we can use our novel video-acceleration matching algorithm to analyse only the stand-up transitions that were generated by them. Results are presented in terms of tai-

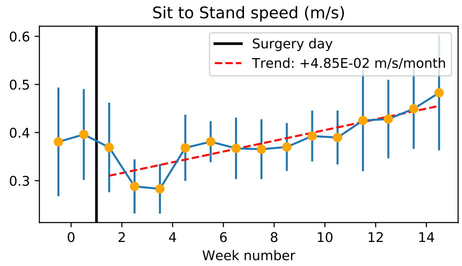


Figure 7: Speed of the transition from sitting to standing calculated on all the subjects appearing in front of the camera and averaged per week. The red light represents the trend from the surgery day until the end of the experiment.

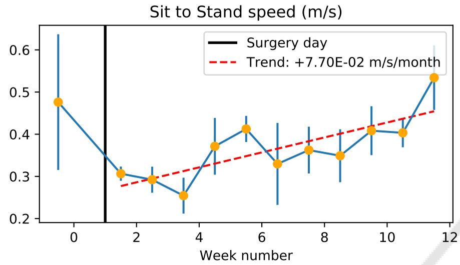


Figure 8: Speed of the transition from sitting to standing calculated only for the subject undergoing hip/knee surgery replacement, isolated using our matching algorithm and averaged per week. The red light represents the trend from the surgery day until the end of the experiment.

lored stand-up transitions in Figure 8, where a much sharper decrease of the stand-up speed soon before and after the surgery are observed, from $\approx 0.5\ m/s$ to $\approx 0.3\ m/s$. A sharper slope can also be observed for the recovery trend, which increases from 0.0482 to 0.0770 $m/s/month$. This steeper slope is an additional element confirming that our matching algorithm is working properly and it is tailoring the video measurements to our target subject. A comparison of the average error-bar size between the two plots provides us with a further confirmation of our results, with an average error of 0.0747 m/s of the old results compared with 0.0611 m/s of our tailored trend.

6 CONCLUSIONS

The majority of AAL systems are designed and implemented under very strictly-controlled conditions, often tailored to home-lab environments with limited occupancy and often including acted scenarios. The reality is different, where people live their own lives and regularly have guests, caretakers, pets, and they leave their house, and even move their furniture around. Moreover, many studies have shown an increasing concern towards the privacy aspects of AAL systems, especially in terms of RGB cameras and mass surveillance.

In this paper, we proposed a solution to the above concerns by developing a framework for AAL that allows monitoring and ReID through multi-sensory fusion of silhouettes and accelerations from wearable devices. Our method is an improved version of the work in (Masullo et al., 2020) to develop a tracking functionality which allows the system to be deployed in real-world multi-occupancy environments. We tested the algorithm in different conditions and dissected its performances through a series of ablation studies, showing an average auROC score of 81.5%. Additionally, we also presented a clinically relevant real-world application of our AAL framework by monitoring a hip/knee replacement surgery patient during their recovery period. Results suggest that the application of our methodology allows for a more detailed trend plot that can better help clinicians to follow their patient's quality of rehabilitation.

ACKNOWLEDGEMENTS

This work was performed under the SPHERE Next Steps Project funded by the UK Engineering and Physical Sciences Research Council (EPSRC), Grant EP/R005273/1.

REFERENCES

Akagunduz, E., Aslan, M., Sengu, A., Wang, H., and Ince, M. C. (2017). Silhouette Orientation Volumes for Efficient Fall Detection in Depth Videos. *IEEE Journal of Biomedical and Health Informatics*, 21(3):756–763.

Amato, G., Bacciu, D., Chessa, S., Dragone, M., Gallicchio, C., Gennaro, C., Lozano, H., Micheli, A., O'Hare, G. M. P., Renteria, A., and Vairo, C. (2016). A Benchmark Dataset for Human Activity Recognition and Ambient Assisted Living. In Lindgren, H., De Paz, J. F., Novais, P., Fernández-Caballero, A., Yoe, H., Jiménez Ramírez, A., and Villarrubia, G., editors, *Advances in Intelligent Systems and Computing*, volume 476 of *Advances in Intelligent Systems and Computing*, pages 1–9. Springer International Publishing, Cham.

Ba, J. L., Kiros, J. R., and Hinton, G. E. (2016). Layer Normalization. *arXiv*.

Bacciu, D., Chessa, S., Ferro, E., Fortunati, L., Gallicchio, C., La Rosa, D., Llorente, M., Micheli, A., Palumbo, F., Parodi, O., Valenti, A., and Vozzi, F. (2016). Detecting socialization events in ageing people: The experience of the DOREMI project. *Proceedings - 12th International Conference on Intelligent Environments, IE 2016*, pages 132–135.

Cabrera-Quiros, L. and Hung, H. (2019). A Hierarchical Approach for Associating Body-Worn Sensors to Video Regions in Crowded Mingling Scenarios. *IEEE Transactions on Multimedia*, 21(7):1867–1879.

Calvaresi, D., Cesarini, D., Sernani, P., Marinoni, M., Dragoni, A. F., and Sturm, A. (2017). Exploring the ambient assisted living domain: a systematic review. *Journal of Ambient Intelligence and Humanized Computing*, 8(2):239–257.

Caroux, L., Consel, C., Dupuy, L., and Sauzéon, H. (2018). Towards context-aware assistive applications for aging in place via real-life-proof activity detection. *Journal of Ambient Intelligence and Smart Environments*, 10(6):445–459.

Climent-Pérez, P., Spinsante, S., Mihailidis, A., and Florez-Revuelta, F. (2020). A review on video-based active and assisted living technologies for automated lifelogging. *Expert Systems with Applications*, 139:112847.

Cook, D. J., Crandall, A. S., Thomas, B. L., and Krishnan, N. C. (2013). CASAS: A Smart Home in a Box. *Computer*, 46(7):62–69.

Das, S., Dai, R., Koperski, M., Minciullo, L., Garattoni, L., Bremond, F., and Francesca, G. (2019). Toyota smarthome: Real-world activities of daily living. *Proceedings of the IEEE International Conference on Computer Vision*, 2019-Octob:833–842.

Dobbler, K., Fišer, M., Fellner, M., and Rettenbacher, B. (2014). Vibroacoustic Monitoring: Techniques for Human Gait Analysis in Smart Homes. In Wichert, R. and Klausing, H., editors, *Ambient Assisted Living*, pages 47–58, Berlin, Heidelberg. Springer Berlin Heidelberg.

Fafoutis, X., Janko, B., Mellios, E., Hilton, G., Sherratt, R. S., Piechocki, R., and Craddock, I. (2016). SPW-1: A low-maintenance wearable activity tracker for residential monitoring and healthcare applications. In *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, volume 1, pages 294–305. Springer International Publishing.

Hall, J., Hannuna, S., Camplani, M., Mirmehdi, M., Damen, D., Burghardt, T., Tao, L., Paiement, A., and Craddock, I. (2016). Designing a Video Monitoring System for AAL applications: the SPHERE Case Study. In *2nd IET International Conference on Technologies for Active and Assisted Living (TechAAL 2016)*, volume 2016, pages 12 (6 .)–12 (6 .). Institution of Engineering and Technology.

Holthe, T., Casagrande, F. D., Halvorsrud, L., and Lund, A. (2018). Disability and Rehabilitation : Assistive Technology The assisted living project : a process evaluation of implementation of sensor technology in community assisted living . A feasibility study. *Disability and Rehabilitation: Assistive Technology*, 0(0):1–8.

Kurz, M., Holzl, G., Ferscha, A., Calatroni, A., Roggen, D., Troster, G., Sagha, H., Chavarriaga, R., del R. Millan, J., Bannach, D., Kunze, K., and Lukowicz, P. (2012). The OPPORTUNITY Framework and Data Processing Ecosystem for Opportunistic Activity and Context Recognition. *International Journal of Sensors Wireless Communications and Control*, 1(2):102–125.

Masullo, A., Burghardt, T., Damen, D., Hannuna, S., Ponce-Lopez, V., and Mirmehdi, M. (2018). CaloriNet : From silhouettes to calorie estimation in private environments. *Proceedings of BMVC*, pages 1–14.

Masullo, A., Burghardt, T., Damen, D., Perrett, T., and Mirmehdi, M. (2020). Person Re-ID by Fusion of Video Silhouettes and Wearable Signals for Home Monitoring Applications. *Sensors*, 20(9):2576.

Masullo, A., Burghardt, T., Perrett, T., Damen, D., and Mirmehdi, M. (2019). Sit-to-Stand Analysis in the Wild Using Silhouettes for Longitudinal Health Monitoring. In *Image Analysis and Recognition*, pages 1–26. Springer Nature Switzerland.

Patel, A. and Shah, J. (2019). Sensor-based activity recognition in the context of ambient assisted living systems: A review. *Journal of Ambient Intelligence and Smart Environments*, 11(4):301–322.

Rashidi, P. and Mihailidis, A. (2013). A survey on ambient-assisted living tools for older adults. *IEEE Journal of Biomedical and Health Informatics*, 17(3):579–590.

Schroff, F., Kalenichenko, D., and Philbin, J. (2015). FaceNet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 07-12-June, pages 815–823. IEEE.

Shigeta, O., Kagami, S., and Hashimoto, K. (2008). Identifying a moving object with an accelerometer in a camera view. In *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3872–3877. IEEE.

Skubic, M., Alexander, G., Popescu, M., Rantz, M., and Keller, J. (2009). A smart home application to eldercare: Current status and lessons learned. *Technology and Health Care*, 17(3):183–201.

Tao, L., Burghardt, T., Mirmehdi, M., Damen, D., Cooper, A., Hannuna, S., Camplani, M., Paiement, A., and Craddock, I. (2017). Calorie Counter: RGB-Depth Visual Estimation of Energy Expenditure at Home. In *Lecture Notes in Computer Science*, volume 10116 LNCS, pages 239–251.

Woznowski, P., Fafoutis, X., Song, T., Hannuna, S., Camplani, M., Tao, L., Paiement, A., Mellios, E., Haghighi, M., Zhu, N., Hilton, G., Damen, D., Burghardt, T., Mirmehdi, M., Piechocki, R., Kaleshi, D., and Craddock, I. (2015). A multi-modal sensor infrastructure for healthcare in a residential environment. In *IEEE International Conference on Communication Workshop*, pages 271–277. IEEE.

Zhu, N., Diethe, T., Camplani, M., Tao, L., Burrows, A., Twomey, N., Kaleshi, D., Mirmehdi, M., Flach, P., and Craddock, I. (2015). Bridging e-Health and the Internet of Things: The SPHERE Project. *IEEE Intelligent Systems*, 30(4):39–46.

Zouba, N., Bremond, F., Thonnat, M., Anfosso, A., Pascual, É., Malléa, P., Mailland, V., and Guerin, O. (2009). A computer system to monitor older adults at home: Preliminary results. *Gerontechnology*, 8(3):0–13.