
Publisher's PDF, also known as Version of record
License (if available): CC BY
Link to published version (if available): 10.1049/iet-cvi.2017.0112

Link to publication record in Explore Bristol Research
PDF-document

This is the final published version of the article (version of record). It first appeared online via IET at http://digital-library.theiet.org/content/journals/10.1049/iet-cvi.2017.0112. Please refer to any applicable terms of use of the publisher.

University of Bristol - Explore Bristol Research

General rights

This document is made available in accordance with publisher policies. Please cite only the published version using the reference above. Full terms of use are available: http://www.bristol.ac.uk/pure/about/ebr-terms
Energy expenditure estimation using visual and inertial sensors

Lili Tao1,2, Tilo Burghardt1, Majid Mirmehdi2✉, Dima Damen1, Ashley Cooper1, Massimo Camplani4, Sion Hannuna1, Adeline Paient1, Ian Craddock1

1 SPHERE, Faculty of Engineering, University of Bristol, Bristol, UK
2 Centre for Machine Vision, Bristol Robotics Laboratory, University of the West of England, Bristol, UK
✉ E-mail: M.Mirmehdi@bristol.ac.uk

Abstract: Deriving a person's energy expenditure accurately forms the foundation for tracking physical activity levels across many health and lifestyle monitoring tasks. In this study, the authors present a method for estimating caloric expenditure from combined visual and accelerometer sensors by way of an RGB-Depth camera and a wearable inertial sensor. The proposed individual-indeterminate visual and inertial method fuses information from both modalities which leads to improved estimates beyond the accuracy of single modality and manual metabolic equivalents of task (MET) lookup table based methods. For evaluation, the authors introduce a new dataset called SPHERE_RGBD + Inertial_calorie, for which visual and inertial data are simultaneously obtained with indirect calorimetry ground truth measurements based on gas exchange. Experiments show that the fusion of visual and inertial data reduces the estimation error by 8 and 18% compared with the use of visual only and inertial sensor only, respectively, and by 33% compared with a MET-based approach. The authors conclude from their results that the proposed approach is suitable for home monitoring in a controlled environment.

1 Introduction

The term ‘energy expenditure’ refers to a human's caloric uptake over time, which is one commonly used single metric to quantify physical activity levels. It is an important determinant in understanding the development of chronic diseases, such as obesity and diabetes. Current evidence-based guidelines [1] indicate that people who are regularly physically active have a 20–40% lower risk of developing conditions such as cardiovascular disease and type 2 diabetes than those who are inactive, and suggest that adults should accumulate at least 150 min of moderate intensity physical activity each week or 75 min of vigorous activity, or a combination of the two. Most research into estimating and understanding caloric expenditure focuses on coarse energy totals over longer time segments or relates to specific activities only, such as walking and running, which generally occur outside the home.

Yet, very little attention has been paid on how activities of normal daily living in an indoor environment can be quantified and understood in terms of energy expenditure. Traditionally, physical activity levels have been measured in Metabolic Equivalents of Task (MET) [2], where a fixed value is assigned to each activity, e.g. 1 MET corresponds to energy expended at rest. However, the method is highly unreliable due to the fact that the activities are monitored using self-report approaches, such as questionnaires and occasional clinical check-ups.

There are various approaches that reliably estimate human energy expenditure using respiratory gas analysis, including both direct and indirect methods. Direct calorimetry measures, such as a sealed respiratory chamber [3], produce accurate outputs, but require a laboratory-based environment. Indirect calorimetry, on the other hand, measures energy expenditure based on inspired and expired respiratory gas flows, volumes and concentrations of oxygen consumption and carbon dioxide production. Some of these measurement devices are portable, less invasive and can produce accurate readings. They form the measurement standard for non-stationary scenarios where the person can move freely. Nevertheless, participants in experiments are required to carry gas sensors and wear a breathing mask [4].

Recently, with an increasing number of wearable devices becoming available, approximating the energy expenditure using inertial sensors has become a popular monitoring choice due to its low cost, low energy consumption, and data simplicity. Acceleration reflects a relation between motion and energy expenditure, thus tri-axial accelerometers are the most broadly used inertial sensors [5]. Recent studies show that more sensors could be involved in the task: breath rate, chest and arm skin temperature also show the correlation with energy expenditure via estimating the oxygen consumption [6]. The data could be obtained by a heart rate monitor and thermometers.

Vision-based systems, as alternative approximative sensors, do not require the wearing of extra devices. In fact, they are already a key part of home entertainment systems today [7], where RGB-Depth sensors allow for a rich and fine-grained analysis of human activity for purposes such as gaming within the field of view. Recent advances in computer vision have now opened up the possibility of integrating these devices seamlessly into home monitoring and assisted living systems [8–10].

Simultaneous utilizations of visual and inertial sensors are not common today, but are receiving growing attention in various areas, including action recognition [5], gesture recognition [11], robotics [12], augmented/virtual reality [13], and assistive technologies applications, such as fall detection [14], food preparation [15] and in a general ambient assisted living system [16]. Although employing multi-modal sensors has the advantage of complementing shortcomings of individual modalities, wearing a multitude of sensors can cause user acceptance issues. With this in mind, in this paper we propose a framework for estimating energy expenditure in living environments based on a non-intrusive RGB-Depth visual sensor and two inertial sensors – worn on the wrist and waist – backed up in experiments by simultaneously taken indirect calorimetry measurements based on the measurement of oxygen consumption and carbon dioxide production for an accurate ground truth provision. This is a new application and to the best of our knowledge no dataset of a similar setup with reliable and accurate ground truth exists. Thus, in order to quantify the performance of the proposed system, we present a new dataset, SPHERE_RGBD + Inertial_calorie, for caloric expenditure estimation collected within a home environment. The dataset contains 11 common household activities performed over
up to 20 sessions, lasting up to 30 min for each session, in each of which the activities are performed continuously. The experimental setup consists of an RGB-Depth Asus Xtion camera mounted at the corner of a living room, two accelerometer sensors, and a COSMED K4b2 [4] indirect calorimeter for ground truth measurement (see Fig. 1). The SPHERE_RGBD + Inertial_calorie dataset is publicly released (The dataset is available online at http://doi.org/cc5k).

This paper is built on our recent work in [17–19], with significant extensions and improvements. Tao et al. [17] introduced a fusion framework for recognising human daily activity using visual and inertial sensors. The work did not address the issue of energy expenditure estimation. Tao et al. [19] proposed a framework for calorific expenditure estimation using only a visual sensor, thus, there was no sensor fusion involved. In [18], we presented a system which allows real-time prediction for activity intensity levels, relying on light-weight bounding box features. This makes the method unable to produce precise calorific expenditure values. In this work, we have improved the feature representation for both inertial and visual sensor data by considering spatial and temporal information at the same time, and investigated both early and late fusion approaches of the data from these sensors. The key contributions of this work are as follows. (i) We propose a first-ever framework for the estimation of calorific expenditure from a RGB-Depth sensor and inertial wearable sensors. There is no work published on visual-inertial energy expenditure estimation, there being only very few works that offer purely vision-based estimation [18, 20, 21]. (ii) We improve the feature representation for both inertial and visual features in the previous fusion framework in [17] by extracting rich, multi-level information to give improved estimation accuracy. (iii) We introduce a new dataset, linking more than 10 h of RGB-Depth video data and inertial sensor data to ground truth calorie readings from indirect calorimetry based on gas exchange. (iv) We present a comparative study on the utility of both visual and inertial data when estimating energy expenditure in a living environment. The visual sensor and inertial sensors are evaluated individually first, followed by an evaluation of two fusion approaches. The rest of the paper is organised as follows. Section 2 presents the background and work related to our study. Section 3 describes the proposed framework for estimating energy expenditure from RGB-Depth and inertial sensors alone, as well as in fusion. The experimental setup and the results are presented in Section 4, followed by a discussion and our conclusions in Section 5.

2 Related work

2.1 Inertial sensors

Acceleration, angular velocity, and rotation signals from wearable devices have been used for human action recognition [22], and are popular in healthcare-oriented applications, such as in fall detection systems [23] and medication adherence monitoring systems [24]. Inertial sensors can offer particularly low-cost and ubiquitous monitoring solutions for physical activities. Techniques that can control computational complexity, power consumption, and improve the unobtrusiveness of the wearable computers [25] are applicable to many systems including the one at hand. Here, we first discuss inertial sensor feature extraction methods described in the literature, followed by an outline of existing models of energy expenditure estimation based on them.

Feature representation: Different features extracted from inertial sensor devices have been considered ranging from raw signal samples to high-level descriptors. Raw time series data from accelerometers is most often provided as triples of scalars, where each scalar corresponds to acceleration in one of three orthogonal spatial dimensions. The same fundamental structure applies to angular velocity signals and orientation signals of three directions. There is no computational burden associated with feature extraction when the raw data is used.
However, raw data may not expose enough discriminative structure to achieve high performance on specific classification tasks. Instead, statistical features may be extracted from each of the three axes, where sensor signal sequences are often partitioned into temporal windows over which features are generated. All features extracted from a temporal window are then concatenated to form a single combined descriptor vector.

Commonly used features include the first- and second-order statistics, namely the mean and variance [26]. In [17], apart from these commonly used features, correlation measures between each axis pair are also extracted. Basic statistical measures are computationally efficient and are able to capture structural patterns in inertial data. The feature descriptor can be further quantised into a number of codewords, such as in [27]. Approaches based on deep learning are currently being explored to create more generalised learning methods that generate features directly from the input data and promise to optimise performance further [28].

Energy expenditure estimation: The first automatic methods for inertial energy expenditure estimation [29] were count-based estimation systems applied by fitting a single regression model to all the data regardless of what activity was being performed. However, systems that map from a single wearable to calorie values struggle to accurately estimate the intensity of physical activity across a range of actions. For example, some actions involving only upper or lower body movements are difficult to be recognised via a single wearable device, and therefore a high estimation error [30] occurs. Different activities may require different models to represent them. Activity-specific (AS) methods split the estimation process into two steps, where activity groups are detected and classified first, and only then an AS model is applied to estimate energy expenditure. MET lookup tables are the most common approach to perform the latter, where a static MET value is assigned from a compendium on physical activities [2]. However, METs-based approaches neglect any transitional effects of activities (continued calorie expenditure after rigorous activity has finished), and they overlook the fact that even the activities in same cluster can be performed at varying intensities, for example, walking at different speeds, or body exercise with different intensity.

An attempt to model the transition between activities was proposed in [32], where an accelerometer and a heart rate sensor were used and the transition between sedentary, household activities, and walking were modelled. The work in [33] shows that by using data from multiple inertial sensors one can more accurately predict energy expenditure, although the limitations of wearable devices are considerable particularly with respect to accuracy as emphasised in [17].

Accelerometer feature descriptors are often formed within a temporal window. This brings out another concern that the window sizes are set usually at <10 s in existing works [6, 32]. The length of window would significantly affect the results. It should be short enough to recognise activities as local temporal information are more descriptive, but long enough to predict calorie values since current energy expenditure strongly depends on previous activity intensity level.

2.2 Visual sensors

Visual sensor based techniques have emerged over recent years for which there exists a significant body of literature describing the inference of activities from two-dimensional (2D) colour intensity imagery [34]. Meanwhile, the increasing availability of depth-measuring sensors, especially the introduction of the Microsoft Kinect, has generated an opportunity for utilising depth in conjunction with traditional RGB camera data allowing for richer and more fine-grained analysis of human activity [7]. Applying computer vision techniques to help with the diagnosis and management of health and wellbeing conditions has gained significant momentum over the last years [35]. However, studies on energy expenditure using visual sensors have been relatively limited. Our work explores this field further and builds on several relevant subject areas in computer vision.

Visual feature representation: The visual trace of human activity in video forms a spatio-temporal pattern. To extract relevant properties from this for the task at hand, one aims at compactly capturing this pattern and highlighting important aspects related to the properties of interest. Assuming that both body configuration and body motion [36] are relevant to infer caloric uptake, the pool of potential features is large – ranging from local interest point configurations [37], over holistic approaches like histograms of oriented gradients and histograms of motion information [17], to convolutional neural network features [38].

Motion information in the first place could also be recovered in various ways, e.g. from RGB data using optical flow or from depth data using 4D surface normals [39]. Whilst a composition of these features via concatenation of per-frame descriptors is straightforward, this approach suffers from the curse of dimensionality and unaffordable computational cost. Sliding window methods [40] on the other hand, can limit this by predicting current values only from nearby data within a temporal window. Further compaction may be achieved by converting large feature arrays into a single, smaller vector with a more tractable dimension count via, for instance, bags of visual words [41], Fisher vectors [42], time series pooling [43], or the features extracted from convolutional neural networks [44]. In summary, the challenge of feature representation will require capturing visual aspects relevant to caloric expenditure, whilst limiting the dimensionality of the descriptor.

Activity recognition: There exists a significant body of literature describing the inference of activities from 2D colour intensity imagery [34], RGB-Depth data [7], and skeleton-based data [45]. Knowledge about the type of activity undertaken has been shown to correlate with the caloric expenditure incurred [2]. In alignment with Fig. 2a, we will argue in this work that an explicit activity recognition step in the vision pipeline can aid the visual estimate of energy uptake.

Energy expenditure estimation: 2D video has recently been used by Edgecomb and Vahid [20] coarsely to estimate daily energy expenditure. In their work, subjects are first segmented from the scene background. Changes in height and width of the subject's motion bounding box, together with vertical and horizontal velocities and accelerations, are then used to estimate caloric uptake. Tsou and Wu [21] take this idea further and estimate caloric consumption using full 3D joint movements tracked as skeleton models by a Microsoft Kinect. We note, however, that both of the above methods use wearable accelerometry as the target ground truth, which in fact does not provide an accurate benchmark; skeleton data is commonly noisy and currently only operates reliably when the subject is facing the camera. This limits applicability in more complex, in-the-wild, visual settings as, for instance, contained in the SPhERE RGB + Inertial caloric database [19].

In the current work, we will argue that an explicit visual based framework for estimating caloric expenditure in a home environment, and we then extended it to be able to estimate physical activity intensity levels in real time [18]. Although the method is practically applicable to more complex settings, the light-weight features extracted from bounding boxes (velocity vector and the ratio of height and width of the bounding box) can only help make a gross estimate of caloric expenditure. In this paper, instead of using only simple bounding box features, we simultaneously collect RGB and depth imagery and then encode appearance and motion features via spatial pyramids. The temporal information is encoded using a pyramidal temporal pooling with multiple pooling operators. This has the aim of extracting rich, multi-level information to give improved estimation accuracy, whilst maintaining applicability to detect complex human activities.

2.3 Sensor fusion

It is reasonable to expect that the use of multiple sensor types would improve the overall performance compared with single sensor settings, since sensors may complement the limitations of each other. Given an accurate temporal synchronisation between the different modality sensors, learning from multi-modal data is applicable. In general, sensor-level fusion (early fusion)
decision-level fusion (late fusion) are the two approaches most often employed to fuse multiple modalities. Both early and late fusion strategies are explained in further detail in [46].

**Feature-level fusion:** This methodology involves carrying out fusion of features right after features are extracted from raw data. This scheme only requires one learning stage and allows to take advantage of mutual information from data. For instance, in [47], the depth and inertial sensor data were concatenated, then an Hidden Markov Model (HMM) classifier was employed for recognising basic gestures on the fused data. The results reveal significant improvements when the fusion scheme is applied compared to using each sensor individually. The work in [17] investigates the practical home-use of body-worn mobile phone inertial sensors together with an RGB-Depth camera to achieve monitoring of daily living scenarios. The results indicate that the vision-based approach significantly outperforms the wearable-based method, while fusion of both sensors slightly improves the performance further. Clearly, feature-level fusion can be applied effectively in practical settings; however, it may suffer from the ‘curse of dimensionality’.

**Decision-level fusion:** This approach fuses the decisions made by individual classifiers, each of which corresponds to one sensor. Since decision information is of low complexity, the curse of dimensionality can effectively be targeted. In [48], for instance, a Bayesian co-boosting training framework combines multiple hidden Markov model classifiers of two modalities – a Kinect sensor and an inertial measurement unit. The result is the construction of a strong classifier for gesture recognition, which achieved the best performance in the multi-modal gesture recognition challenge. A real-time action recognition system in [49] uses Dempster–Shafer theory to combine the classification outcomes from a depth camera and several inertial sensors. A Bayesian model for sensor fusion is introduced in [16], which aims at addressing the challenges of fusion of heterogeneous sensor modalities in ambient assisted living.

**Comparisons:** In this work, we consider both fusion approaches and provide a direct comparison. In the feature-level fusion approach, features generated from the two modality sensors are merged before classification, and the decision-level fusion is performed by forming a linear combination of different classifiers using stacking regression [50] to improve overall accuracy. As outlined in the following section, our work attempts to use skeleton-independent, RGB-Depth-based vision, together with two wearable accelerometer devices to estimate calorific expenditure against a standardised calorimetry sensor COSMED-K4b2 based on gas exchange.

### 3 Method

To describe our framework for estimating calorific expenditure, we initially introduce the methods for visual and wearable sensors separately, and then describe two approaches for their fusion.

Fig. 2a shows a flowchart of the visual method – mapping visual flow and depth features to calorie estimates using AS models. The method implements a cascaded and recurrent approach, which explicitly detects activities as an intermediate to select type-specific mapping functions for the final calorific estimation. Importantly, our setup as a video-based system is designed to reason about activities first, before estimating calorie expenditure via a set of models which are each separately trained for particular activities. In contrast to this, our direct mapping (DM) method designed for wearable sensor data directly maps inertial features to calorie estimates via a monolithic classifier. A flowchart of the wearable approach is shown in Fig. 2b. In our fusion system, we consider both feature-level and decision-level fusion of these two approaches. Finally, we compare these methods against a ground truth of gas-exchange measurements and off-the-shelf alternatives, that is manual mapping from activity classes to calorie estimates via METs lookup tables [2] as it is often applied in clinical practice today.

#### 3.1 Visual features

We obtain RGB and depth imagery using an Asus Xtion. For each frame $t$, appearance and motion features are extracted, with the latter being computed with respect to the previous frame (level 0). A set of temporal filters is then applied to form higher level motion features (level 1). We extract features from within the bounding box returned by the OpenNI SDK [51] person detector and tracker, which allows to follow up to six persons in the camera view at the same time. To normalise the utilised image region due to varying heights of the subjects and their distance to the camera, the bounding box is scaled by fixing its longer side to $M = 60$ pixels, a size recognised as optimal for human action recognition [52], while maintaining aspect ratio. The scaled bounding box is then centred in a $M \times M$ square box and horizontally padded.

**Motion feature encoding:** Inspired by Tran and Sorokin [52], optical flow measurements are taken over the bounding box area and split into horizontal and vertical components. These are re-

---

**Fig. 2** Overview of our visual-based and wearable-based frameworks

(a) Visual-based framework. RGB-Depth videos are represented by a combination of flow and depth features. The proposed recurrent method then selects AS models which map to energy expenditure estimates. (b) Wearable-based framework. Inertial features are formed from two accelerometer sensor data, then features are mapped directly to calorie estimates via a monolithic classifier.

---

**Table 3** Visual features

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Motion feature encoding</td>
<td>Optical flow measurements are taken over the bounding box area and split into horizontal and vertical components.</td>
</tr>
</tbody>
</table>
sampled to fit the normalised box and a median filter with kernel size $5 \times 5$ is applied to smooth the data. A spatial pyramid structure is used to form hierarchical features from this. Such partitioning of the image into an iteratively growing number of sub-regions increases discriminative power. The normalised bounding box is divided into a $n_g \times n_g$ non-overlapping grid, where $n_g$ depends on the pyramid level, and the orientations of each grid cell are quantised into $n_b$ bins. The parameters for our experiments are empirically determined as $n_b = 9$ and $n_g = 1$ and 2 for levels 0 and 1, respectively. Fig. 3a exemplifies optical flow patterns and their encoding in two different example activities.

**Appearance feature encoding:** We extract depth features by applying the histogram of oriented gradients feature on raw depth images [53] within the normalised bounding box. We then apply principal component analysis and keep the first 150 dimensions of this high-dimensional descriptor, which retains 95% of the total variance.

**Pyramidal temporal pooling:** Given the motion and appearance features extracted from each frame in a sequence of images, it is important to capture both short- and long-term temporal changes, and summarise them to represent the motion in the video. Pooled motion features were first presented in [43], even though designed for egocentric video analysis. We modify their pooling operator to empirically determine as increases discriminative power. The normalised bounding box is applied to each time segment at each level. The time series data can be represented as a $T$-dimensional descriptor, which retains 95% of the total variance.

The frequency components of the frequency coefficients $X$ are defined as:

$$X_m(S_n(n)) = \left\{ \begin{array}{l}
\frac{1}{\sqrt{2}}(s_{n,1} + s_{n,2}) & \text{if } m = 0 \\
\frac{1}{\sqrt{2}}(s_{n,1} - s_{n,2}) & \text{if } m = 1 \\
\frac{1}{\sqrt{2}}s_{n,m} & \text{if } m \geq 2
\end{array} \right.$$
changes. Thus, we apply three pooling operators (max pooling, sum pooling, and frequency domain pooling) to the inertial data.

3.3 Learning and recurrency

Energy expenditure estimation can be formulated as a sequential and supervised regression problem. We train a support vector regressor to predict calorie values from given features over a training set. The sliding window method is used to map each input window of width $w$ to an individual output value $y_t$. The window contains the current and the previous $w-1$ observations. The window feature is represented by temporal pooling from the time series $S = S_t-w+1, \ldots, S_t$.

We note that energy values for a particular time are highly dependent on the energy expenditure history. In our system, these are most directly expressed by previous calorific predictions during operation. Thus, employing recurrent sliding windows offers an option to not only use the features within a window, but also take the most recent $d$ predictions $\hat{y}_{t-d}, \ldots, \hat{y}_{t-1}$ into consideration to help predict $y_t$. During learning, as suggested in [54], the ground truth labels in the training set are used in place of recurrent values.

3.4 Fusion approach

Both feature-level and decision-level fusion are considered in our work.

Feature-level fusion: This is an early fusion approach, for which all features from all modalities are concatenated together, and employed as a single unified feature stream to the learning components. Given visual features in $d_1$-dimensional feature space $S_v \in \mathbb{R}^{d_1}$ and accelerometer features in $d_2$-dimensional feature space $S_a \in \mathbb{R}^{d_2}$, the fused feature set can be represented as $S \in \mathbb{R}^{d_1+d_2}$, where the feature set is constructed as $S = (S_v, S_a)$. The fused feature vector is then used as input to the classifiers of the system. Fig. 4a shows a flowchart of this feature-level fusion approach.

Decision-level fusion: In this approach, a collection of models are learned, and the predictions are combined together only at the last stage to form the final decision. We apply the decision-level fusion via a stacking regression method, which forms linear combinations of different classifiers to improve overall estimation accuracy.

Consider that there are $K$ predicted values $\hat{y}_1, \ldots, \hat{y}_K$ estimated from each regressor individually. Then, the final predictor value $\hat{Y}(S)$ can be represented as a linear combination of a set of predicted values with different weighting coefficients, constructed as

$$\hat{Y}(S) = \sum_{k=1}^{K} \alpha_k \hat{y}_k(S)$$

(3)

Given a set of training data $\{(S^i, y^i), \ldots, (S^T, y^T)\}$ with $T$ training samples, where each $S^i$ is an input vector, the goal is to minimise the distance of the ground truth $y^i$ and the predicted values $\hat{Y}(S)$ from the combined regressor. This optimised distance can be obtained by

$$\arg \min \left\{ \sum_{i=1}^{T} \left[ y^i - \sum_{k=1}^{K} \alpha_k \hat{y}_k(S^i) \right]^2 \right\}$$

(4)

with the constraints $0 \leq \alpha_k \leq 1, \ k = 1, \ldots, K$. The resulting combined predicted value $\sum_{k=1}^{K} \alpha_k \hat{y}_k(S)$ is then used as prediction. Fig. 4b shows a flowchart of this decision-level fusion approach.
4 Experimental results

4.1 Dataset and ground truth

We introduce the SPHERE_RGBD+Inertial_calorie dataset for human calorific expenditure estimation, comprising RGB-Depth and inertial sensor data captured in a real living environment. The ground truth was captured by the COSMED K4b2 portable metabolic measurement system. The dataset was generated over 20 sessions by 10 subjects with varying anthropometric measurements. Participants were seven males and three females, with mean age of 27.2 ± 3.8 years, mean weight of 72.3 ± 15.0 kg, mean height of 173.6 ± 9.8, mean body mass index of 23.7 ± 2.8. Ethics approval was obtained and each participant signed a consent form agreeing to share their data for research purposes. The dataset contains up to 11 activity categories per session, and totalling around 10 h recording time. The activities were captured in daily-living scenarios containing a variety of body positions, viewpoints, and distances naturally associated with the various actions performed. Fig. 5 shows frames from the vacuuming activity depicting this variety. It is also shown that the sequences are captured in different time of the day which contains various lighting conditions. All the activities, the associated intensity categories, and MET values are shown in Table 1. In addition, Table 2 lists the number of frames for each action and sequence [Some actions in certain sequences are missing due to various reasons (hence they have 0 frames), for example exercise is missing in sequences 7 and 17 as the participants had difficulty in performing the action.]

Colour and depth images were acquired at a rate of 30 Hz. The accelerometer data was captured at about 100 Hz and sampled down to 30 Hz, a frequency recognised as optimal for human action recognition [55]. The calorimeter gives readings per breath, which occurs approximately every 3 s. To model transitions better between activity levels, we consider the nine different combinations of the three activity intensities (light, light+, moderate) in the design of each session. Fig. 1 shows a detailed example of calorimeter readings and associated sample RGB images from the dataset (top) and the accelerometer data reading (bottom). The raw breath data is noisy (in red). We apply an average filter with a span of ~20 breaths (in blue). The participants were asked to perform the activities based on their own living habits without any extra instructions.

4.2 Parameter settings

In our experiments, we use non-linear Support Vector Machines (SVMs) with radial basis function kernels for activity classification and a linear support vector regressor for energy expenditure prediction. The libsvm [56] implementation was used. We perform a grid search algorithm to estimate the hyper-parameters of the SVM. To test our individual-independent approach, we implement leave-one-subject-out cross-validation on the dataset in which each subject's data are tested in turn using models trained with all other subject data combined. This process iterates through all subjects, and the average testing error and standard deviation of all iterations are reported. We use the normalised root-mean-squared error (normalised RMSE) as a standard evaluation metric to facilitate the comparison between data with different scales for the deviation of estimated calorie values from the ground truth.

4.3 Evaluation of individual modalities

We start with tests on each sensor type (visual and inertial), and compare their performance in situations when used independently. Temporal window size: The accuracy of predicted calorie values is linked to the window of previous information utilised for making the prediction. In a first experiment, we look at the relation...
illustrate the activity recognition rates and the average normalised calorie prediction errors on the other. All sequences are tested between window length on the one hand, and activity recognition.

Table 3 Number of frames per sequence and action in SPHERE_RGBD + Inertial_calorie dataset

<table>
<thead>
<tr>
<th>seq ID</th>
<th>stand</th>
<th>sit</th>
<th>walk</th>
<th>wipe</th>
<th>vacuum</th>
<th>sweep</th>
<th>lying</th>
<th>exercise</th>
<th>stretch</th>
<th>clean</th>
<th>read</th>
<th>overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3739</td>
<td>5230</td>
<td>5509</td>
<td>5117</td>
<td>5447</td>
<td>5113</td>
<td>5142</td>
<td>2789</td>
<td>2421</td>
<td>5072</td>
<td>6092</td>
<td>51,671</td>
</tr>
<tr>
<td>2</td>
<td>3546</td>
<td>5388</td>
<td>5304</td>
<td>5244</td>
<td>5177</td>
<td>5091</td>
<td>5020</td>
<td>2545</td>
<td>2566</td>
<td>5112</td>
<td>5311</td>
<td>50,304</td>
</tr>
<tr>
<td>3</td>
<td>3876</td>
<td>5682</td>
<td>6093</td>
<td>4745</td>
<td>5151</td>
<td>4946</td>
<td>5360</td>
<td>2297</td>
<td>2318</td>
<td>5317</td>
<td>5095</td>
<td>50,880</td>
</tr>
<tr>
<td>4</td>
<td>3948</td>
<td>5211</td>
<td>5470</td>
<td>5182</td>
<td>5025</td>
<td>4813</td>
<td>3784</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>33,433</td>
</tr>
<tr>
<td>5</td>
<td>3229</td>
<td>5133</td>
<td>5670</td>
<td>4878</td>
<td>5436</td>
<td>4327</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>28,683</td>
</tr>
<tr>
<td>6</td>
<td>3812</td>
<td>5294</td>
<td>5889</td>
<td>5023</td>
<td>5015</td>
<td>4841</td>
<td>4878</td>
<td>3205</td>
<td>1778</td>
<td>4738</td>
<td>5070</td>
<td>49,543</td>
</tr>
<tr>
<td>7</td>
<td>3796</td>
<td>5239</td>
<td>11,602</td>
<td>6684</td>
<td>4432</td>
<td>4561</td>
<td>3590</td>
<td>0</td>
<td>0</td>
<td>5474</td>
<td>5837</td>
<td>5337</td>
</tr>
<tr>
<td>8</td>
<td>3951</td>
<td>5257</td>
<td>5412</td>
<td>5244</td>
<td>4886</td>
<td>1005</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>25,755</td>
</tr>
<tr>
<td>9</td>
<td>4128</td>
<td>5568</td>
<td>5091</td>
<td>5195</td>
<td>4651</td>
<td>3619</td>
<td>4891</td>
<td>2458</td>
<td>2483</td>
<td>5700</td>
<td>5298</td>
<td>49,082</td>
</tr>
<tr>
<td>10</td>
<td>3649</td>
<td>5202</td>
<td>5317</td>
<td>5354</td>
<td>4651</td>
<td>4990</td>
<td>5030</td>
<td>2330</td>
<td>1669</td>
<td>3933</td>
<td>5337</td>
<td>47,462</td>
</tr>
<tr>
<td>11</td>
<td>4367</td>
<td>4901</td>
<td>5103</td>
<td>5133</td>
<td>5006</td>
<td>4761</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>29,671</td>
</tr>
<tr>
<td>12</td>
<td>3697</td>
<td>5270</td>
<td>5618</td>
<td>5010</td>
<td>5107</td>
<td>4988</td>
<td>4991</td>
<td>2891</td>
<td>2335</td>
<td>5299</td>
<td>5412</td>
<td>50,618</td>
</tr>
<tr>
<td>13</td>
<td>4250</td>
<td>5936</td>
<td>4644</td>
<td>5162</td>
<td>5259</td>
<td>4517</td>
<td>4944</td>
<td>899</td>
<td>2839</td>
<td>5495</td>
<td>6206</td>
<td>50,151</td>
</tr>
<tr>
<td>14</td>
<td>4263</td>
<td>4732</td>
<td>5370</td>
<td>5150</td>
<td>4769</td>
<td>4847</td>
<td>4861</td>
<td>3008</td>
<td>3224</td>
<td>4366</td>
<td>5765</td>
<td>50,355</td>
</tr>
<tr>
<td>15</td>
<td>3457</td>
<td>5784</td>
<td>4789</td>
<td>4745</td>
<td>5159</td>
<td>4911</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>28,845</td>
</tr>
<tr>
<td>16</td>
<td>3919</td>
<td>5466</td>
<td>5062</td>
<td>5308</td>
<td>2716</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>22,471</td>
</tr>
<tr>
<td>17</td>
<td>3613</td>
<td>5343</td>
<td>5432</td>
<td>4914</td>
<td>5032</td>
<td>4461</td>
<td>4979</td>
<td>0</td>
<td>5063</td>
<td>5902</td>
<td>4873</td>
<td>49,612</td>
</tr>
<tr>
<td>18</td>
<td>3715</td>
<td>5340</td>
<td>5422</td>
<td>5013</td>
<td>5839</td>
<td>4743</td>
<td>4517</td>
<td>1977</td>
<td>2793</td>
<td>5948</td>
<td>5012</td>
<td>50,373</td>
</tr>
<tr>
<td>19</td>
<td>4521</td>
<td>5434</td>
<td>5787</td>
<td>4740</td>
<td>5015</td>
<td>4459</td>
<td>5480</td>
<td>3174</td>
<td>2707</td>
<td>4803</td>
<td>6342</td>
<td>52,462</td>
</tr>
<tr>
<td>20</td>
<td>4040</td>
<td>5255</td>
<td>5597</td>
<td>5472</td>
<td>5309</td>
<td>4551</td>
<td>4926</td>
<td>1797</td>
<td>1691</td>
<td>6356</td>
<td>6443</td>
<td>51,437</td>
</tr>
</tbody>
</table>

Table 4 Vision-based prediction results. Activity recognition rate (%) and calorific expenditure prediction error (normalised RMSE) with different window lengths $w$, stated in seconds. The best results for each activity are in bold.

<table>
<thead>
<tr>
<th>$w$</th>
<th>stand</th>
<th>sit</th>
<th>walk</th>
<th>wipe</th>
<th>vacuum</th>
<th>sweep</th>
<th>lying</th>
<th>exercise</th>
<th>stretch</th>
<th>clean</th>
<th>read</th>
<th>overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.5</td>
<td>63.5</td>
<td>75.2</td>
<td>90.5</td>
<td>70.2</td>
<td>79.7</td>
<td>73.4</td>
<td>58.6</td>
<td>36.8</td>
<td>54.1</td>
<td>93.8</td>
<td>39.9</td>
<td>72.3</td>
</tr>
<tr>
<td>activity</td>
<td>0.94</td>
<td>0.70</td>
<td>0.29</td>
<td>0.52</td>
<td>0.31</td>
<td>0.41</td>
<td>0.74</td>
<td>0.63</td>
<td>0.52</td>
<td>0.41</td>
<td>0.67</td>
<td>0.58</td>
</tr>
<tr>
<td>calorie</td>
<td>86.5</td>
<td>77.6</td>
<td>88.3</td>
<td>69.4</td>
<td>79.0</td>
<td>76.5</td>
<td>62.3</td>
<td>39.2</td>
<td>61.1</td>
<td>91.4</td>
<td>36.9</td>
<td>73.7</td>
</tr>
<tr>
<td>15</td>
<td>83.6</td>
<td>66.0</td>
<td>0.46</td>
<td>0.34</td>
<td>0.43</td>
<td>0.45</td>
<td>0.74</td>
<td>0.66</td>
<td>0.54</td>
<td>0.40</td>
<td>0.64</td>
<td>0.53</td>
</tr>
<tr>
<td>activity</td>
<td>85.0</td>
<td>79.1</td>
<td>89.4</td>
<td>71.9</td>
<td>81.1</td>
<td>75.2</td>
<td>54.3</td>
<td>40.3</td>
<td>57.8</td>
<td>90.4</td>
<td>36.8</td>
<td>71.1</td>
</tr>
<tr>
<td>calorie</td>
<td>0.73</td>
<td>0.52</td>
<td>0.30</td>
<td>0.41</td>
<td>0.36</td>
<td>0.41</td>
<td>0.66</td>
<td>0.55</td>
<td>0.54</td>
<td>0.37</td>
<td>0.54</td>
<td>0.49</td>
</tr>
<tr>
<td>60</td>
<td>81.1</td>
<td>79.7</td>
<td>85.1</td>
<td>66.0</td>
<td>77.2</td>
<td>72.9</td>
<td>33.0</td>
<td>29.3</td>
<td>52.7</td>
<td>90.0</td>
<td>35.9</td>
<td>68.2</td>
</tr>
<tr>
<td>activity</td>
<td>0.54</td>
<td>0.45</td>
<td>0.32</td>
<td>0.44</td>
<td>0.34</td>
<td>0.39</td>
<td>0.58</td>
<td>0.42</td>
<td>0.52</td>
<td>0.38</td>
<td>0.50</td>
<td>0.45</td>
</tr>
<tr>
<td>calorie</td>
<td>0.54</td>
<td>0.45</td>
<td>0.32</td>
<td>0.44</td>
<td>0.34</td>
<td>0.39</td>
<td>0.58</td>
<td>0.42</td>
<td>0.52</td>
<td>0.38</td>
<td>0.50</td>
<td>0.45</td>
</tr>
</tbody>
</table>

Fig. 6 Recognition confusion matrices from the best activity recognition results corresponding to the use of (a) Visual sensor, (b) Inertial sensors.

between window length on the one hand, and activity recognition and calorie prediction errors on the other. All sequences are tested with temporal windows of $w = \{7.5, 15, 30, 60\}$ s. Tables 3 and 4 illustrate the activity recognition rates and the average normalised RMSEs for calorie prediction at different window length $w$ using visual and accelerometer data, respectively. It can be seen that, in both modalities, the best performance across the set for recognising activities is achieved when a relatively small size of window is applied. The confusion matrices corresponding to the use of visual and inertial sensors are depicted in Fig. 6.

In a second experiment, we test how the estimated calorie value is influenced by the performance of action recognition. The results
Table 4  Inertial-based prediction results. Activity recognition rate (%) and calorific expenditure prediction error (normalised RMSE) with different window lengths, in seconds. The best results for each activity are in bold.

<table>
<thead>
<tr>
<th>w</th>
<th>activity</th>
<th>stand</th>
<th>sit</th>
<th>walk</th>
<th>wipe</th>
<th>vacuum</th>
<th>sweep</th>
<th>lying</th>
<th>exercise</th>
<th>stretch</th>
<th>clean</th>
<th>read</th>
<th>overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.5</td>
<td>activity</td>
<td>91.1</td>
<td>58.7</td>
<td>78.1</td>
<td>43.6</td>
<td>50.1</td>
<td>48.2</td>
<td>66.0</td>
<td>51.8</td>
<td>65.2</td>
<td>59.8</td>
<td>48.0</td>
<td>60.1</td>
</tr>
<tr>
<td></td>
<td>calorie</td>
<td>0.54</td>
<td>0.50</td>
<td>0.49</td>
<td>0.52</td>
<td>0.51</td>
<td>0.41</td>
<td>0.64</td>
<td>0.83</td>
<td>0.72</td>
<td>0.51</td>
<td>0.77</td>
<td>0.63</td>
</tr>
<tr>
<td>15</td>
<td>activity</td>
<td>89.7</td>
<td>86.7</td>
<td>43.9</td>
<td>50.5</td>
<td>37.9</td>
<td>63.4</td>
<td>66.7</td>
<td>68.1</td>
<td>66.6</td>
<td>61.8</td>
<td>60.3</td>
<td>60.3</td>
</tr>
<tr>
<td></td>
<td>calorie</td>
<td>0.50</td>
<td>0.51</td>
<td>0.41</td>
<td>0.45</td>
<td>0.46</td>
<td>0.38</td>
<td>0.53</td>
<td>0.80</td>
<td>0.68</td>
<td>0.44</td>
<td>0.73</td>
<td>0.59</td>
</tr>
<tr>
<td>30</td>
<td>activity</td>
<td>88.7</td>
<td>56.9</td>
<td>83.9</td>
<td>46.6</td>
<td>48.7</td>
<td>36.4</td>
<td>69.4</td>
<td>51.8</td>
<td>64.1</td>
<td>61.0</td>
<td>52.1</td>
<td>58.0</td>
</tr>
<tr>
<td></td>
<td>calorie</td>
<td>0.46</td>
<td>0.49</td>
<td>0.48</td>
<td>0.46</td>
<td>0.52</td>
<td>0.35</td>
<td>0.54</td>
<td>0.77</td>
<td>0.60</td>
<td>0.43</td>
<td>0.71</td>
<td>0.56</td>
</tr>
<tr>
<td>60</td>
<td>activity</td>
<td>90.8</td>
<td>55.3</td>
<td>74.8</td>
<td>52.3</td>
<td>49.6</td>
<td>45.7</td>
<td>73.1</td>
<td>44.9</td>
<td>59.0</td>
<td>64.2</td>
<td>56.0</td>
<td>58.3</td>
</tr>
<tr>
<td></td>
<td>calorie</td>
<td>0.41</td>
<td>0.41</td>
<td>0.50</td>
<td>0.50</td>
<td>0.47</td>
<td>0.40</td>
<td>0.57</td>
<td>0.64</td>
<td>0.55</td>
<td>0.41</td>
<td>0.71</td>
<td>0.54</td>
</tr>
</tbody>
</table>

Table 5  Calorific expenditure prediction error (normalised RMSE) using the visual sensor when ground truth labels are used to select the AS model (top row) and when action recognition is employed at different window lengths. The best results for each activity are in bold.

<table>
<thead>
<tr>
<th>calorie</th>
<th>activity</th>
<th>w</th>
<th>stand</th>
<th>sit</th>
<th>walk</th>
<th>wipe</th>
<th>vacuum</th>
<th>sweep</th>
<th>lying</th>
<th>exercise</th>
<th>stretch</th>
<th>clean</th>
<th>read</th>
<th>overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>60</td>
<td>n/a</td>
<td>0.40</td>
<td>0.45</td>
<td>0.28</td>
<td>0.35</td>
<td>0.32</td>
<td>0.38</td>
<td>0.55</td>
<td>0.36</td>
<td>0.55</td>
<td>0.55</td>
<td>0.50</td>
<td>0.43</td>
<td></td>
</tr>
<tr>
<td>7.5</td>
<td>activity</td>
<td>0.51</td>
<td>0.43</td>
<td>0.28</td>
<td>0.38</td>
<td>0.33</td>
<td>0.38</td>
<td>0.58</td>
<td>0.43</td>
<td>0.35</td>
<td>0.43</td>
<td>0.45</td>
<td>0.45</td>
<td></td>
</tr>
<tr>
<td></td>
<td>calorie</td>
<td>0.51</td>
<td>0.43</td>
<td>0.41</td>
<td>0.41</td>
<td>0.32</td>
<td>0.39</td>
<td>0.57</td>
<td>0.45</td>
<td>0.54</td>
<td>0.36</td>
<td>0.44</td>
<td>0.44</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>activity</td>
<td>0.47</td>
<td>0.44</td>
<td>0.30</td>
<td>0.42</td>
<td>0.31</td>
<td>0.37</td>
<td>0.56</td>
<td>0.44</td>
<td>0.53</td>
<td>0.46</td>
<td>0.44</td>
<td>0.44</td>
<td></td>
</tr>
<tr>
<td></td>
<td>calorie</td>
<td>0.54</td>
<td>0.45</td>
<td>0.32</td>
<td>0.44</td>
<td>0.34</td>
<td>0.39</td>
<td>0.58</td>
<td>0.42</td>
<td>0.52</td>
<td>0.38</td>
<td>0.50</td>
<td>0.45</td>
<td></td>
</tr>
</tbody>
</table>

Table 6  Calorific expenditure prediction error (normalised RMSE) using the inertial sensors when ground truth labels are used to select the AS model (top row), and when action recognition is employed at different window lengths. The best results for each activity are in bold.

<table>
<thead>
<tr>
<th>calorie</th>
<th>activity</th>
<th>w</th>
<th>stand</th>
<th>sit</th>
<th>walk</th>
<th>wipe</th>
<th>vacuum</th>
<th>sweep</th>
<th>lying</th>
<th>exercise</th>
<th>stretch</th>
<th>clean</th>
<th>read</th>
<th>overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>60</td>
<td>n/a</td>
<td>0.47</td>
<td>0.41</td>
<td>0.33</td>
<td>0.26</td>
<td>0.37</td>
<td>0.45</td>
<td>0.46</td>
<td>0.57</td>
<td>0.47</td>
<td>0.40</td>
<td>0.50</td>
<td>0.45</td>
<td></td>
</tr>
<tr>
<td>7.5</td>
<td>activity</td>
<td>0.51</td>
<td>0.47</td>
<td>0.41</td>
<td>0.45</td>
<td>0.39</td>
<td>0.40</td>
<td>0.62</td>
<td>0.72</td>
<td>0.52</td>
<td>0.42</td>
<td>0.63</td>
<td>0.52</td>
<td></td>
</tr>
<tr>
<td></td>
<td>calorie</td>
<td>0.53</td>
<td>0.57</td>
<td>0.40</td>
<td>0.48</td>
<td>0.44</td>
<td>0.43</td>
<td>0.57</td>
<td>0.72</td>
<td>0.46</td>
<td>0.39</td>
<td>0.57</td>
<td>0.52</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>activity</td>
<td>0.48</td>
<td>0.46</td>
<td>0.46</td>
<td>0.50</td>
<td>0.43</td>
<td>0.43</td>
<td>0.57</td>
<td>0.68</td>
<td>0.55</td>
<td>0.41</td>
<td>0.58</td>
<td>0.53</td>
<td></td>
</tr>
<tr>
<td></td>
<td>calorie</td>
<td>0.41</td>
<td>0.41</td>
<td>0.51</td>
<td>0.50</td>
<td>0.47</td>
<td>0.40</td>
<td>0.57</td>
<td>0.64</td>
<td>0.55</td>
<td>0.41</td>
<td>0.71</td>
<td>0.54</td>
<td></td>
</tr>
</tbody>
</table>

For each activity using visual and inertial sensors are listed in Tables 5 and 6, respectively. We first test a system in which the ground truth labels are used to select the AS model for calorific prediction (top rows in Tables 5 and 6). We then compare actual action recognition at varying window lengths $w = \{7.5, 15, 30, 60\}$ s. In all cases, we use a fixed window length $w = 60$ s for calorific expenditure estimation to focus on the effect of varying action recognition quality. As expected, it can be observed that for most activities, the calorific estimation error is smallest when there is no activity recognition error (top rows in Tables 5 and 6). For a more detailed visualisation, we also show the results in Fig. 7; the 11 actions are grouped into three clusters based on their intensity level (see Table 1). The figure summarises the calorific prediction error for different intensities and action recognition rates using the visual system and the inertial system, respectively.

Model comparison: As just observed, activity recognition accuracy affects the calorific prediction results. To determine if the AS model provides a predictive advantage, we compare the estimation performance of the AS approach against the DM approach for each sensor modality. For both sensor systems, we select a fixed window length of $w = 60$ s to analyse performance for calorific value prediction in both DM and AS, and $w = 15$ s for

---

Fig. 7 Prediction accuracy of calorific expenditure. Average calorie prediction errors (normalised RMSE) when ground truth labels are used to select the AS model (in yellow), and when action recognition is employed at different window length using (a) Visual sensor, (b) Two accelerometer sensors.

For Table 4, the best results for each activity are in bold. The window length $w$ ranges from 7.5 to 60 seconds. For Table 5, the window length $w$ is 60 seconds. For Table 6, the window length $w$ is 60 seconds. For all tables, the best results for each activity are in bold.
activity recognition in AS. The results are shown in Fig. 8 for each activity. It can be seen that for the visual-based system, the AS provides best prediction results overall and significantly outperforms DM in most activities. For the inertial-based system, the error associated with AS is significantly higher compared with DM. This is in part due to poor activity recognition results in an inertial measurement setup, which effectively leads to using wrong models to estimate calorie values.

Evaluation of a recurrent system layout: To evaluate the use of recurrency, we set the AS method using the sliding window technique as our baseline method for the vision-based comparison, and the DM method for inertial-based comparison. We now introduce two methods, which are based on recurrent sliding window approaches. The first method (Recurrent1) uses the most recent predictions of the baseline method as input together with visual/inertial features to predict current calorie value. Thus, it implements indirect recurrency utilising the predicted values from the baseline as recent predictions. The second method (Recurrent2) implements full recurrency, i.e. it uses its own output as recurrent input together with visual/inertial features.

Table 7 shows the effect of using recurrent information, with the best results for each activity highlighted. In general, the full recurrency model, Recurrent2, suffers from drift and produces the worst results for half of the activities and also overall. When the visual sensor is used, indirect recurrency, Recurrent1, outperforms the other approaches at an average normalised RMSE of 0.42, while in inertial-based systems indirect recurrency increases the estimation error by 7% comparing to its baseline.

Table 7  Average calorific expenditure prediction errors (normalised RMSE) for each activity with different learning approaches. The best results for each activity are in bold

<table>
<thead>
<tr>
<th></th>
<th>w</th>
<th>stand</th>
<th>sit</th>
<th>walk</th>
<th>wipe</th>
<th>vacuum</th>
<th>sweep</th>
<th>lying</th>
<th>exercise</th>
<th>stretch</th>
<th>clean</th>
<th>read</th>
<th>overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>visual</td>
<td>baseline</td>
<td>0.41</td>
<td>0.43</td>
<td>0.30</td>
<td>0.41</td>
<td>0.32</td>
<td>0.39</td>
<td>0.57</td>
<td>0.45</td>
<td>0.54</td>
<td>0.36</td>
<td>0.44</td>
<td>0.44</td>
</tr>
<tr>
<td></td>
<td>recurrent1</td>
<td>0.36</td>
<td>0.39</td>
<td>0.28</td>
<td>0.38</td>
<td>0.30</td>
<td>0.38</td>
<td>0.54</td>
<td>0.41</td>
<td>0.50</td>
<td>0.35</td>
<td>0.41</td>
<td>0.42</td>
</tr>
<tr>
<td></td>
<td>recurrent2</td>
<td>0.35</td>
<td>0.49</td>
<td>0.31</td>
<td>0.52</td>
<td>0.59</td>
<td>0.53</td>
<td>0.56</td>
<td>0.38</td>
<td>0.46</td>
<td>0.42</td>
<td>0.45</td>
<td>0.52</td>
</tr>
<tr>
<td>inertial</td>
<td>baseline</td>
<td>0.41</td>
<td>0.45</td>
<td>0.32</td>
<td>0.30</td>
<td>0.33</td>
<td>0.40</td>
<td>0.35</td>
<td>0.75</td>
<td>0.49</td>
<td>0.35</td>
<td>0.36</td>
<td>0.46</td>
</tr>
<tr>
<td></td>
<td>recurrent1</td>
<td>0.41</td>
<td>0.54</td>
<td>0.37</td>
<td>0.32</td>
<td>0.34</td>
<td>0.41</td>
<td>0.45</td>
<td>0.78</td>
<td>0.49</td>
<td>0.36</td>
<td>0.44</td>
<td>0.49</td>
</tr>
<tr>
<td></td>
<td>recurrent2</td>
<td>0.50</td>
<td>1.06</td>
<td>0.63</td>
<td>0.49</td>
<td>0.45</td>
<td>0.49</td>
<td>0.92</td>
<td>0.95</td>
<td>0.52</td>
<td>0.44</td>
<td>0.85</td>
<td>0.68</td>
</tr>
</tbody>
</table>
4.4 Comparing sensor fusion approaches

Having tested the two modalities individually, we now study modality fusion approaches against the use of individual sensor systems, and also compare against the MET lookup table method for completeness.

In feature level fusion, we apply our AS approach with $w = 15\, \text{s}$ for activity recognition and $w = 60\, \text{s}$ for calorific expenditure estimation. In decision level fusion, we again use the most suitable model for each sensor data to fuse, which is the AS approach for visual sensor data, and DM approach for inertial sensor data. The estimation performance of the two fusion approaches is compared with the performance of each sensor modality individually, as shown in Fig. 8b. It can be seen that both fusion approaches on average outperform unimodal prediction. In particular, by combining the features from the visual and the inertial data, the overall prediction error decreases from 0.46 (inertial sensors alone) and 0.42 (visual sensor alone) to 0.39. The calorie prediction accuracy for most activities is improved when using fusion approaches. We also observed that the two fusion frameworks perform similarly.

Finally, we present the results produced by MET, which is commonly used by clinicians and physiotherapists, and compare our proposed methods against it. It assumes $N$ clusters of activity $A = \{A_1, A_2, \ldots, A_N\}$ are known. A MET value is assigned to each cluster, together with anthropometric characteristics of individuals. The amount of AS energy expended can then be estimated as energy = 0.0175(kcal/kg/min) × weight (kg) × MET values. Here, we use the ground truth labels to select activities to keep this procedure identical to the commonly used manual estimate. Table 8 presents the detailed results for each sequence. The accuracy is calculated over the total calorie expended in each recording session. We also measure the correlation between the ground truth and the observed values [Note that the total calorie values for sequence 4, 5, 8, 11, 15, and 16 are relatively low due to shorter sequences.]. We can see that the fusion of visual and inertial sensors achieves higher accuracy and correlation in more sequences than the MET model or unimodal approaches, and obtains better rates on average, which points towards an advantage of using visual-inertial set-ups for the task of calorific expenditure prediction.

5 Conclusion and future directions

We have presented a system for calorific expenditure estimation using data from two different modality sensors, a RGB-Depth camera sensor, and wearable inertial sensors (accelerometers). We have demonstrated the effectiveness of the fusion approach through a comprehensive comparative study with single modality setups and widely used METs prediction. The proposed fusion system used pooled spatial and temporal pyramids of visual and accelerometer features, which subsequently are fed in both early and late fusion approaches. To test the methodology, we introduced the challenging SPHERE_RGBD + Inertial calorie dataset, which covers a wide variety of home-based human activities. The proposed fusion method demonstrates its ability to outperform the METs estimation approach and the use of single modality sensors. The focus of the paper has been on presenting a system for estimating calorific expenditure from combined visual and accelerometer sensors, where the purpose of the study has been to show that the fusion of both modalities improves the estimates beyond the accuracy of single modality, and the proposed system outperforms manual metabolic lookup table based methods – the main measure used in clinical practice today. We acknowledge that applying more advanced fusion approaches and different feature representations may improve the performance further. Possible future directions include introducing deep learning models and investigating advanced data fusion methodologies for different modality sensors. We hope this work, and the new dataset, will establish a baseline for future research in the area.

6 Acknowledgment

This work was performed under the SPHERE IRC project funded by the UK Engineering and Physical Sciences Research Council, Grant EP/K031910/1. The data from this study is available by request from the University of Bristol research data repository via http://www.sphere-sphere.ac.uk/work-package-2/calorie or http://doi.org/ce5k.

7 References


This is an open access article published by the IET under the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0/)


This is an open access article published by the IET under the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0/)