DEPARTMENT OF COMPUTER SCIENCE

Passive Energy Expenditure Estimation using Anonymized Video Devices

Jason Hian Hong Chan

A dissertation submitted to the University of Bristol in accordance with the requirements of the degree of Bachelor of Science in the Faculty of Engineering.

Friday 1st November, 2019

Declaration

This dissertation is submitted to the University of Bristol in accordance with the requirements of the degree of BSc in the Faculty of Engineering. It has not been submitted for any other degree or diploma of any examining body. Except where specifically acknowledged, it is all the work of the Author.

Jason Hian Hong Chan, Friday 1st November, 2019

Abstract

Due to the increasing costs of healthcare in our modern society, it is important to ensure that people are taking preventative measures and getting sufficient physical activity. However, most research looking at physical activity measurements do not take into account the individual performing household activities during normal daily living. As such, this paper seeks to measure physical activity levels in the home environment. Energy expenditure (EE) is the best metric of quantifying physical activity levels. Computer vision and deep learning methods have been gaining in popularity of late with regards to EE estimation. However, most of these studies do not produce a direct mapping from video data to EE, but instead take a two-step approach by first performing activity recognition before using activity-specific models to estimate EE for each activity detected. This is typically done using METs tables, but this is largely inaccurate and comes with several problems associated with METs tables.

In this paper, we seek to use silhouettes extracted from RGB-D data along with accelerometer data in order to produce a direct mapping to EE. The use of silhouette data is motivated by privacy and anonymity concerns for individuals being monitored in the home environment. Using a fusion of both temporal silhouettes and accelerometer data in a convolutional network to estimate EE, this technique allows the subjects to move around freely without being inconvenienced by complex instruments. The network is trained and cross-validated on the SPHERE-Calorie dataset [62], achieving comparable results to the work by Masullo *et al.* [45].

In the future, this project can be extended in several different ways, such as by introducing physical activity labels as a target when training the network, or even by altering the network architecture in order to achieve better estimation accuracy.

Acknowledgements

I would like to thank my supervisors Sion Hannuna and Alessandro Masullo for their invaluable support and guidance throughout this project. I would also like to thank my family for their encouragement and support throughout these years. Finally, I would like to thank my friends, without whom I would not have been able to burn the midnight oil with.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Physical activity is important for the health of the individual and of the general populace [27]. There is evidence suggesting that a higher volume of physical activity can result in a reduction in all-cause mortality [41]. A lack of regular physical activity has been linked with a number of chronic diseases, including cardiovascular disease, hypertension, obesity, thromboembolic stroke, anxiety and depression [35]. Conversely, a healthy amount of physical activity has been shown to lead to a reduction in some of these major causes of death, such as for strokes and cardiovascular disease. Older adults also tend to encounter a higher risk of developing problems due to musculoskeletal deficiencies and cardiovascular conditions. However, regular physical activity has been shown to demonstrate a statistically significant drop in the health risks associated with age [18].

In the UK alone, the NHS spends upwards of £6 billion from its annual budget on diseases related to physical inactivity and overweight and obesity [58]. It's clear that improvements need to be made for health care services throughout the world. With the rise of deep learning methods however, these technologies and methods have the potential to pave the way towards a new era of predictive health care systems [46]. There are a wide range of applications for this: disease prediction, treatment recommendation as well as prediction of medication effects, to name but a few of these areas where predictive health care can provide improvements to. One other important use case is in the measurement of physical activity. Being able to measure a person's physical activity is useful not just for

the general health and wellbeing of the population, but also to ensure that people with, or at risk of, certain chronic diseases are able to be monitored for poor health. Typically, a doctor or a medical practitioner has to be present with the patient in order to perform diagostic tests to measure the patient's physical fitness and physical activity levels. This is problematic as these are resources that could be better used for treating a wider range of people. Furthermore, measuring physical activity levels in this way does not take into account the physical activity levels from normal daily living in the home environment. Thus, accurate measurements of physical activity levels in the home environment is vital to monitoring the health of these individuals over a long period of time [57].

When it comes to quantifying physical activity levels, the most commonly used metric is energy expenditure (EE), otherwise known as calorific expenditure. Several different methods for estimating EE have been presented over the years, such as using metabolic equivalents (METs) tables [2], to indirect calorimetry by measuring oxygen uptake and carbon dioxide output [24], to direct calorimetry which measures the total heat dissipated by a person's body [34]. These methods tend to be invasive and costly, while in the case of METs tables, often requires self-reporting by the subject and does not account for individual factors specific to the subject, leading to inaccuracies. With the rise of wearable inertial sensors however, monitoring a person's physical activity levels can be performed much more conveniently and noninvasively. With more and more devices having network capabilities, the Internet of Things (IoT) through the use of frameworks such as Internet of Things based Physical Activity Monitoring (PAMIoT) [53] shows a lot of promise in enabling cost-effective health monitoring. Indeed, embedding sensors into articles of clothing has become much easier in recent years, as can be seen in the work done by Morris *et al.* [47] which integrated sensors into shoes to monitor gait. Furthermore, accessories such as smartphones, smart watches and wireless fitness devices that have become a large part of daily life provide unobtrusive means of obtaining sensor data. Particularly, Fitbit devices have been shown to have the potential to provide accurate and reliable estimations of EE during walking and running. [21].

In comparison, a video based system for estimating EE would be able to provide richer data as direct motion recognition can be extracted without needing any extra de-

vices to be worn. Furthermore, camera systems used in home monitoring do not need to be constantly charged unlike most wearable devices. Vision-based estimation using RGB-D cameras has been shown to give very accurate EE estimations [64]. In addition to this, a system using both vision and accelerometers simultaneously has the potential to improve EE estimation beyond that of just video data [48, 63]. While video data can provide richer analysis of physical activity, this also raises the issues of patient privacy and consent [7]. Patients tend to be quite critical and reluctant towards medical technology within spaces as private as the home environment, fearing unauthorized data transfer and access, and the alteration or loss of data [76]. Privacy has been said to be the greatest obstacle against smart home systems [30]. One of the techniques to address these privacy concerns has been to convert the raw RGB-D data into bounding boxes, silhouettes and skeletons, as has been performed in the study by Hall *et al.* [25] for the Sensor Platform for HEalthcare in a Residential Environment (SPHERE) project [52], in the hopes of increasing patient anonymity and allowing for more scalable data sizes. This method also has implications for IoT platforms, as the switch to silhouette data helps to minimise the leakage of sensitive data [45]. With regards to the effectiveness of this method, this thesis will make use of one of SPHERE's existing datasets, SPHERE-Calorie [62] (first presented in [64]), which contains RGB and depth data along with accelerometer data. The key point about this dataset is the inclusion of gas exchange-based calorimeter data which provide ground truth calorific values, allowing for a very accurate cross-validation of the energy estimation results.

1.2 Contributions of the Thesis

In this paper, we replicate the work done by Masullo *et al.* [45] on online EE estimation using silhouette data in the home environment. The fused convolutional neural network, *CaloriNet*, that was presented in their work will be used as a basis with which to work with the SPHERE-Calorie dataset. As a BSc student, I only have a simplistic understanding of neural networks, having not studied deep learning as a unit. This means that a significant amount of self-studying needs to be put towards understanding neural networks in practice by learning the Keras toolchain along with TensorFlow. In order to train the network,

BlueCrystal is required due to the large computational cost of training. Having to learn how to work with BlueCrystal and submit batch jobs comes with a separate set of problems as well, including the handling of large datasets such as SPHERE-Calorie. This requires some techniques in data organisation and management.

Chapter 2

Background

2.1 Estimation of EE

EE estimation in real-time is far from a simple problem, as there a number of factors besides physical activity that contribute to total EE, including but not limited to a person's fitness, metabolism, pathological conditions and environmental conditions such as temperature [15]. However, there is still a strong correlation between the type and intensity of activtiy performed and EE. In order to compare the coding of physical activities, the Compendium of Physical Activities [2] was developed using a five-digit coding scheme defined as the ratio of work metabolic rate to resting metabolic rate (METs). While this offers a quick method of estimating EE, it does not provide precise measurements of EE for an individual as it does not account for differences in age, gender, body mass etc. Furthermore, different people might characterize the intensity of an activity differently, and similarly for how the person performs the activity. As such, differences in EE for the same activity can be quite large depending on the individual subject, and might not reflect the person's true EE. Another problem with using this system is that new activities might be observed that are not included in the table. While the effect of this has been reduced with an updated version of the Compendium [1] which extended the number of physical activities listed, this solution only remains relevant until more complex activities or combinations thereof are required to be observed.

Early studies involving physical activity measurement used self-report methods, the most widely used of which include physical activity questionnaires, physical activity

records and diaries [49]. While these methods provide an easy and affordable way of conducting large cohort studies, they are largely limited by low accuracy and reliability of physical activity measurement [5]. Furthermore, questionnaires also rely on the participant's ability to recall previously performed physical activity which increases the chance of memory bias, leading to more inaccuracy in the data [61].

Direct calorimetry is widely considered to be the gold standard for EE estimation, and is the most accurate method for quantifying metabolic rate [32]. However it has since fallen out of fashion due to the high cost of operation and the technical challenges associated with it. This method directly measures the heat energy dissipated by a subject by placing the subject within a small, insulated chamber [34]. There are four different types of direct calorimeters that measure this heat energy through different methods: isothermal, heat sink, convection and differential [32]. Indirectly calorimetry on the other hand, is much more widely employed for measurements of energy expenditure [34]. This technique estimates energy expenditure by measuring the subject's uptake and output of oxygen and carbon dioxide [42]. Of the different methods of indirect calorimetry, doubly labeled water is often cited as the most reliable method for assessments of physical activity. It is non-invasive and unobtrusive, allowing for a reliable means of monitoring the energy expenditure of free-living subjects over 1-2 week periods [72]. Indirect calorimetry is typically used as a method of validating other methods of estimating energy expenditure, as these tools are usually limited by requiring very costly equipment. In the case of face mask indirect calorimetry, a high burden is placed on the subject as the mask is obtrusive, and while doubly labeled water solves this issue, it is also limited by the fact that obtaining the stable isotopes of water required involves considerable difficulty [72].

2.1.1 Heart Rate Monitoring

It is clear that more individual-oriented approaches to EE estimation are required. One of the more widely used methods nowadays is heart rate monitoring, which is a physiological indicator of physical activity and thus, energy expenditure. It relies on the linear relationship between a person's heart rate and their consumption of oxygen, but this relationship is less accurate during sedentary or low intensity physical activities [20]. Heart rate monitoring is also limited by the fact that other unrelated factors besides physical

activity can affect oxygen consumption, such as body position, stress, food intake, ambient temperature) [43]. It is also important to note that heart rate monitoring is only able to measure a person's response to an activity i.e. their oxygen consumption, but not the identification of the activity itself [43]. The main benefits of heart rate monitoring are that they provide an unobtrusive means of obtaining real-time data from a subject without requiring much effort from the subject [60]. These devices are also much better suited than accelerometers and pedometers for measuring physical activities that involve the upper body, and activities such as swimming and cycling [19].

In a study by Ceesay *et al.* [14] involving minute-by-minute heart rate monitoring using commercially available monitors at the time, the proposed method was found to produce satisfactory predictions of EE for its low cost of resources. However, estimating EE in this way requires some level of individual calibration as everyone has different resting heart rates and fitness. More recently, a study by Altini *et al.* [3] proposed a method that involved estimating cardiorespiratory fitness (CRF) from heart rate, and then using the CRF as a predictor in a hierarchical Bayesian model for EE estimation. This technique was able to reduce the error in EE estimation without requiring individual calibration.

2.1.2 Accelerometry

Accelerometers, sometimes referred to as inertial or wearable sensors, measure acceleration in order to determine body movements in one (uniaxial), two (biaxial) or three (triaxial) orthogonal planes (anteroposterior, mediolateral and vertical) [17]. Typically, triaxial accelerometers are used as they are richer in providing 3-dimensional data. The raw acceleration signals produced by accelerometers are known as counts, which have to be translated or calibrated into some other interesting metric in order to then quantify physical activity. These metrics could be either physical activity patterns (e.g. walking, standing) or a biological variable (e.g. energy expenditure, oxygen consumption) [23]. In the case of energy expenditure, regression equations are used on the raw count data to derive point estimates of EE [23, 29]. Accelerometers have been gaining popularity in recent times as they are objective, lightweight, require minimal effort from subjects, and can be applied over extended periods of time on free-living participants. These devices have also

proven to be reliable in estimating EE, for example, the Tracmore is a model that has shown comparable results when validated with doubly labeled water as a reference [10]. They are also able to provide information about the frequency and intensity of physical activity, by using regression analysis to define accelerometer-derived count cut-points that correspond to different physical activity intensity levels [29]. However, it is important to note that there is a relationship between different cut-points and the estimation of physical activity intensity [44]. This leads to inconsistencies on how to determine suitable cut-points, particularly amongst youth samples, and this lack of standardization requires a consideration of the methodology and sample characteristics [36]. Accelerometers suffer from several other limitations, being unreliable for assessing upper body activities such as lifting or throwing [72], and some devices are unable to differentiate between body positions such as sitting, standing or lying down due to being unable to provide postural information [26].

There are a great number of different combinations of where acccelerometers can be worn, each providing data for different sets of activities that could be performed [17]. More recently however, some studies have attempted to combine multiple accelerometers placed at different body segments with different arrangements in order to improve the accuracy of energy estimation. A study by Zhang *et al.* [74] introduced the Intelligent Device for Energy Expendiutre and Activity (IDEAA) that made use of five miniature accelerometers attached to the chest, midthigh of both legs and both feet. The IDEEA was able to estimate energy expenditure of physical activity with an accuracy >95% when compared to mask calorimetry and respiratory chamber calorimetry, showing that multiple sensors can be applied in estimation of EE. On the other hand, the study performed by Altini *et al.* [4], found that it is sufficiently accurate to use one accelerometer near the person's center of mass provided that an activity-specific EE estimation model is combined with it. It was also found that increasing the number of sensors, provided that the optimal accelerometer positioning is chosen, does not result in any significant error reduction in EE estimation.

Some studies have attempted to combine accelerometry with other physiological measurements. One such study combined accelerometers with heart rate monitoring and

found an improvement in the precision of prediction of oxygen uptake, especially so in the case where accelerometer data was input separately from heart rate data rather than as simultaneous input [10]. The Actiheart is a device that combines both heart rate and unaxial sensors into a single-piece monitor. It was demonstrated to be able to predict EE, using a combined model, to a higher precision than using either parameter alone during standardized technical conditions i.e. walking and running [12]. When combining accelerometers with other physiological measurements, it is worth investigating the impact of the placement of the accelerometers on the accuracy of EE estimation. The research done by Ellis *et al.* [22] combined accelerometer data and heart rate data, and compared the effectiveness of accelerometers worn on the wrist versus on the hip in predicting EE. Using random forest classification and regression trees, this study found that the hip accelerometer was better overall at estimating EE and predicting locomotion, while wrist accelerometers were more suited to predicting physical activites involving substantial arm motions.

There is a growing trend in the use of wearable devices and smart devices equipped with motion sensors for the purposes of tracking physical activity levels for exercise purposes. One of the most common commercially-available wearable devices is the Fitbit, which have the main benefit of having wireless capabilities for interfacing with mobile devices, allowing subjects to efficiently and easily share physical activity information with their physicians. The hip-based Fitbit One and the wrist-based Fitbit Flex have been shown to provide reliable estimations of step count and energy expenditure during walking and running when validated against gas-based indirect calorimetry, with the hip-based model outperforming the wrist-based device [21]. In a study by Noah *et al.* [50], the Fitbit and Fitbit Ultra devices were found to estimate energy expenditure reliably when compared to the Actical, a well-validated accelerometer, and the Cosmed K4b2 indirect calorimetry device. However, it is important to note that both of these studies only tested a small subset of possible physical activites, i.e. walking and running, and these devices have not been validated for measuring non-ambulatory activities, such as weight training and cycling, and free-living household activities. A study by Cvetković *et al.* [75] made use of the accelerometers on both smartphones and wrist-worn devices (smartwatches and sensor-equipped wristbands) to perform accurate activity recognition and estimation of

energy expenditure. Their activity-monitoring algorithm was able to perform with an individual device or a combination of both, and produced comparable estimations of EE to using the BodyMedia Fit Advantage armband, a dedicated sensor device, and the Oxycon mobile indirect calorimeter. More importantly, this study was conducted with both normal daily living activities (e.g. lying down, eating, cleaning) and exercise activities (e.g. walking, running) and shows promise for activity monitoring in the home environment.

2.2 Computer Vision for EE Estimation

Computer vision and machine learning has seen many advancements in the recent decade, allowing researchers to tackle the problem of activity recognition and EE estimation using optical-based methods. While activity recognition using data is a problem that has been studied extensively [33, 39, 67], the effectiveness of computer vision algorithms for EE estimation is a question that remains to be solved in full.

2.2.1 Physical Activity Recognition

There are two main complementary features that affect activity recognition in video-based systems: appearances and temporal dynamics [71]. With advancements in deep learning research, there has been a rise in the use of deep learning techniques in the field of computer vision, particularly for image classification [37]. In particular, deep convolutional neural networks (CNNs) have been shown to be well-suited to the task of action recognition in videos [59, 65] as they are capable of extracting accurate assumptions about representations from raw visual inputs. However, in the region of activity recognition in videos, features learned from end-to-end deep CNNs have yet to show significant advancements over the traditional method of using hand-crafted features [71]. Hand-crafted features refer to low-level descriptors that capture the appearance and motion information of the video, such as histograms of oriented gradients (HOG) and histograms of optical flow (HOF) [38], motion boundary histograms (MBH) [70], Local Trinary Patterns (LTP) [73], and so on.

Due to how activity recognition is affected by the temporal durations of the video

input, it is important that the temporal structure of the video data is modeled. Pirsiavash *et al.* [51] proposed a method of capturing the temporal structure of complex actions in a hierarchical manner through the use of a segmental regular grammar of actions. Some recent works have made attempts to model long-range temporal structures through the use of CNNs and recurrent neural networks (RNNs). The work by Ng *et al.* [31] compared the use of a CNN with feature pooling and a RNN using a Long Short Term Memory (LSTM) architecture, against the state of the art at the time, finding a significant increase in accuracy for videos up to two minutes in duration. A study by Wang *et al.* [71] focused on creating a general framework for video-level learning, coined temporal segment network (TSN). By working with a sequence of short snippets sampled from the entire video, this segment based sampling method was able to outperform previous works that operated on a single frame or a short frame stack.

2.2.2 EE Estimation Algorithms

When it comes to using computer vision to estimate EE, there are two main ways of approaching this problem: activity-specific models that split the process into activity recognition followed by EE estimation specific to the activity detected, or models that produce a direct mapping from visual data to energy expenditure. For the former, one commonly used approach is using METs tables to estimate the EE of each activity detected [69]. However, this technique would still be plagued by the inaccuracies and drawbacks of using METs tables as described previously. The Microsoft Kinect is a popular device due to its ability to capture depth data and to track body joints simultaneously in 3D, and was used in a study by Tsou and Wu [66]. Using the movement tracking of body joints as accelerometers, this study was able to accurately estimate calorie expenditure. That being said, this method suffers from some drawbacks as the study was validated on calorific values obtained from heart rate monitoring devices, which is not the gold standard for calorific ground truth, and the use of skeleton data from a Kinect requires the user to be facing the camera in order for joint tracking to be accurate.

A study conducted by Nakamura *et al.* [48] introduced an egocentric approach to using computer vision in EE estimation, combining combining egocentric video data with accelerometer and heart rate data to produce both activity recognition and estimation

of EE. This technique was shown to produce comparable results to the current state-of-the-art. A more relevant approach is in the work by Tao *et al.* [63], which fused both RGB-D data and accelerometer data in order to perform EE estimation. This method does not perform direct mapping from visual data to calorie estimation, but instead performs activity recognition first before estimating EE. Using both feature-level and decision-level fusion, they were able to accurately estimate EE when validated against face mask indirect calorimetry data. While this is very promising, it also requires the use of full RGB-D data which is unsuitable in use cases where privacy is a requirement [45].

Chapter 3

Technical Background

3.1 Silhoutte Extraction

In order to restrict the amount of visual information of the subjects in private environments, it is important to perform silhouette extraction on the images being captured. This is important for the privacy and anonymity of the individuals, and also to ensure that if data is leaked from the system, it does not contain easily identifiable information that could be traced back to the subject [16]. One such method of converting RGB-D data to foreground silhouttes, proposed in [45], is to perform RGB depth-based segmentation on the image. Images are first processed using OpenPose to extract skeleton data of the subjects along with their associated bounding boxes. By then performing k-means clustering on the RGB-D values within each bounding box, it is possible to produce the silhouette of the subject. Once this is done, the RGB-D images are simply discarded so that only the silhoutte data is left.

OpenPose [13] is an open-source realtime system for 2D pose estimation (body, foot, hand and facial keypoints) capable of handling images with multiple persons. This technique uses a multi-stage convolutional neural network (CNN) where an image is used as input for the network to predict both the confidence map of body part locations and the degree of association between body parts, known as the part affinity field (PAF). A greedy inference is then used to parse both of these parameters in order to associate body part candidates, producing the output of all the keypoints for each person in the image. While this seems very computationally expensive, OpenPose is able to perform in realtime

through the use of a CUDA-enabled Nvidia GPU. While OpenPose can be run on CPU-only devices, it is significantly slower on such devices. Another benefit of using OpenPose is that it provides an integrated pipeline for processing images, including a frame reader, a means of visualising the results and also generates the output in easily readable JSON files.

An alternative to using k-means clustering is to use more sophisticated techniques of segmenting the foreground and background. GrabCut [55] is a foreground extraction technique using iterated graph cuts that could be applied to segmenting the silhouette foreground from the background. It first models the colour data in the image using a Gaussian mixture model (GMM) for both the foreground and the background. It then iteratively performs the minimisation of an energy function using a minimum cut algorithm, described in [11], until convergence is achieved. The benefit of this approach is that it is robust, being able to achieve accurate segmentation in images where there is no clear distinction between foreground and background colour distributions. However, while its runtime is reasonable for normal applications, it may not be suitable for processing large amounts of video data, as compared to the quick runtime of k-means clustering in practice.

Temporal Silhouettes

As mentioned before, the temporal durations of the video input have a strong effect on both activity recognition and EE estimation. Modelling the temporal structure of video data is one method of handling this effect, but there are some considerations when attempting this. A dense approach to temporal sampling by loading a large buffer of images with a pre-defined sampling interval into the network would require both an excessive amount of computational power and a large amount of memory [71]. Due to the impracticality of dense temporal sampling for untrimmed video data, there is value in looking at different ways of compressing video data without the loss of important information. Some studies have looked into different techniques of transforming video data into more compact representations, such as by converting an entire video sequence into a binary cumulative motion image known as motion-energy images (MEI) [9], or by summarising a video sequence into a single dynamic image through the use of of a ranking classifier [6].

While these techniques may not be as suitable when dealing with silhouette data from SPHERE-Calorie, they help to show that a compact representation of video data can be useful in the estimation of EE. The work by Massulo *et al.* [45] introduced the idea of using average silhouettes over variable timescales in order to reduce the dependency of the network on any specific temporal scale chosen.

3.2 Neural Networks and Network Architectures

3.2.1 Theory of Neural Networks

The topic of neural networks has been mentioned several times throughout this paper without a proper explanation. The theory behind neural networks is not a new one, but has seen a surge in popularity in the recent decade with advancements in computational power. Much of this section has been adapted from Bishop [8]. The principal idea behind neural networks is the use of a fixed number of basis functions in which the parameter values are adapted during training. These parametric nonlinear basis functions usually follow the form

$$y(\mathbf{x}, \mathbf{w}) = f\left(\sum_{j=1}^{M} w_j \phi_j(\mathbf{x})\right) \tag{3.1}$$

where $f(\cdot)$ is a nonlinear activation function. A typical hidden layer in a neural network is constructed using M linear combinations of the input variables $x_1, \ldots, x_D$ such that

$$a_j = \sum_{i=1}^{D} w_{ji}^{(1)} x_i + w_{j0}^{(1)} \tag{3.2}$$

The superscript (1) indicates that the parameters here correspond to the first layer of the network. Here, $w_{ji}^{(1)}$ and $w_{j0}^{(1)}$ are the weight and bias parameters respectively. The activation a_j is then transformed using a nonlinear activation function $h(\cdot)$, typically chosen to be sigmoidal functions.

$$z_j = h(a_j) \tag{3.3}$$

The resulting values z_j are called hidden units, which are then linearly combined to construct the output layer of the network

$$a_k = \sum_{j=1}^{M} w_{kj}^{(2)} z_j + w_{k0}^{(2)} \tag{3.4}$$

for K number of outputs, where $k = 1, \ldots, K$. Similar to the hidden layer, the output units are transformed using an appropriate activation function, except that the choice of activation function now depends on the nature of the data, producing the final set of network outputs y_k

$$y_k = f(a_k) \tag{3.5}$$

Combining all these stages together, we can see that the neural network model is just a nonlinear function that takes a set of input variables $\{x_i\}$ and returns a set of output variables $\{y_k\}$, controlled by adjustable parameter weights in a vector $\mathbf{w}$, taking the final form

$$y_k(\mathbf{x}, \mathbf{w}) = \sigma\left(\sum_{j=1}^{M} w_{kj}^{(2)} h\left(\sum_{i=1}^{D} w_{ji}^{(1)} x_i + w_{j0}^{(1)} \right) + w_{k0}^{(2)} \right) \tag{3.6}$$

In order to then train the network and determine the network parameters, one approach when dealing with a feed-forward network is to minimise a sum-of-squares error function given K input vectors $\{\mathbf{x}_k\}$ and K target vectors $\{\mathbf{t}_k\}$,

$$E(\mathbf{w}) = \frac{1}{2} \sum_{k=1}^{K} \Big(y(\mathbf{x}_k, \mathbf{w}) - \mathbf{t}_k \Big)^2 \tag{3.7}$$

allowing us to arrive at the error on the output layer. A simple technique, originally presented by Rumelhart *et al.* [56], of performing this minimisation is to backpropagate this error and iteratively update the weights on each layer of the network through the use of gradient descent. Assuming that the dataset is i.i.d, we can say that the error function is comprised of a sum of terms, one term for each data point in the training set such that

$$E(\mathbf{w}) = \sum_{n=1}^{N} E_n(\mathbf{w}) \tag{3.8}$$

Keeping in mind that each unit simply performs a computation over all its inputs, producing a weight sum such that,

$$a_j = \sum_i w_{ji} z_i \tag{3.9}$$

where w_{ji} is the weight associated with the connection between a unit i to a unit j, and z_i is the activation of unit i. As we did before, we use a nonlinear activation function $h(\cdot)$ to transform this sum, producing the activation z_j of unit j

$$z_j = h(a_j) \tag{3.10}$$

Since E_n only depends on the weight w_{ji} through the summed input a_j into unit j, we evaluate the derivative of E_n with respect to the weight w_{ji},

$$\begin{aligned}\frac{\partial E_n}{\partial w_{ji}} &= \frac{\partial E_n}{\partial a_j}\frac{\partial a_j}{\partial w_{ji}} \\ &= \frac{\partial E_n}{\partial a_j}\frac{\partial}{\partial w_j i}\sum_i w_{ji} z_i \\ &= \delta_j z_i\end{aligned} \tag{3.11}$$

where we have used the notation $\delta_j \equiv \frac{\partial E_n}{\partial a_j}$. This equation shows that calculating the derivative is as simple as a multiplication of δ_j and z_i, where δ_j represents the error for the unit at the output end of the weight and z_i represents the activation at the input end. Thus, in order to evaluate the δ value for the hidden layer, we can perform a similar operation on the next layer

$$\begin{aligned}\delta_j \equiv \frac{\partial E_n}{\partial a_j} &= \sum_k \frac{\partial E_n}{\partial a_k}\frac{\partial a_k}{\partial a_j} \\ &= h'(a_j)\sum_k w_{kj}\delta_k\end{aligned} \tag{3.12}$$

for all units k to which unit j sends connections to. This final equation is the backpropagation formula, which tells us that the value of δ for a certain unit, usually referred to as an error, can be obtained by propagating the errors backward from units higher up in the network.

3.2.2 Network Architectures

Convolutional neural networks are an approach to building invariance properties into the structure of a network in order to deal with transformations of the inputs, first introduced by LeCun *et al.* [40]. CNNs are very well suited for the task of processing images and thus, very useful in the field of computer vision. This is due to how pixels in images have a stronger correlation with other nearby pixels as compared to pixels further away. A typical CNN makes use of two operations: convolutions and pooling. In image processing, a convolution can be thought of as a weighted sum within a spatial region when a small kernel image K is passed over that region, and then performed over the whole image I.

$$S_{ij} = \sum_m \sum_n I(i-m, j-n)K(m,n) \tag{3.13}$$

Using different kernels produces different types of convolved images, some examples include a kernel for blurring that simply averages the values in a region, or for edge detection by using the values -1 and 1 on two adjacent pixels which produces a non-zero value for adjacent pixels with very different values, i.e. edges. For neural networks, the convolution operation can be thought of as using the weights as the kernel in order to extract local features from the data. Pooling, on the other hand, is an operation that allows the network to be invariant to transformations in the data such as translations and scaling. One popular pooling method is called max-pooling, which takes the maximum of features over small blocks of a previous layer,

$$S_{ij} = max\{x_{ij}\}_{i=1,j=1}^{i=N,j=M} \tag{3.14}$$

This operation allows us to find out if a feature is present in a region of the previous layer, but not its exact location. This means that it is more important for something to be present, rather than where it is present, allowing the network to be invariant to transformations in the data.

Recurrent neural networks are distinct in that they are designed to model sequential data. The recurrency comes from performing the same task for every element in a sequence, where the output is dependent on the previous computation, such that

$$s^{(t)} = f(s^{(t-1)}, x^{(t)}) \tag{3.15}$$

In this sense, $s^{(t)}$ can be thought of as the "memory" of the network, capturing information about previous calculations. Since the same task is performed at every layer, a RNN shares the same parameters at each layer unlike other neural networks. RNNs seem to be very well suited to the task of natural language processing but has also seen some applications in activity recognition using computer vision [31].

Chapter 4

Methodology

4.1 Dataset

The method used in this paper is evaluated on the publicly available SPHERE-Calorie [62] dataset. This challenging dataset was captured in a home environment and includes RGB-D and accelerometer data along with ground truth values captured using a COSMED K4b2 indirect calorimeter. A total of 10 subjects with varying human body measurements were used to generate this dataset over 20 sessions. The participants had an average age of 27.2 ± 3.8 years, with a mean height and weight of 173.6 ± 9.8cm and 72.3 ± 15.0kg respectively. The 7 male and 3 female participants had a mean body mass index of 23.7 ± 2.8. Each participant was recorded with an RGB-D Asus Xtion camera placed in the corner of a living room, and fitted with two accelerometer sensors on the waist and wrist respectively, and a calorimeter. Colour and depth images were recorded at a rate of 30Hz,

Figure 1: Sample images of each activity performed by different subjects in the dataset.

Activity	MET Value
sit still	1.3
stand still	1.3
lying down	1.3
reading	1.5
walking	2.0
wiping table	2.3
cleaning floor stain	3.0
vacuuming	3.3
sweeping floor	3.3
upper body exercise	4.0
stretching	5.0

Table 1: Activities and their associated MET values

while accelerometer data was sampled down to 30Hz from abaout 100Hz.

11 daily-living activity categories were recorded in a predetermined sequence per session: stand still, sit still, walking, wiping the table, vacuuming, sweeping floor, lying down, upper body exercise, stretching, cleaning stain and reading. The MET values for each activity type is shown in Table 1, while sample images of each activity can be seen in Figure 1. The activities performed contained a range of body positions, viewpoints and distances associated with each activity. There are some gaps within the dataset for some recorded sequences, due to various reasons such as some participants having difficulty performing the *exercise* activity. Since silhouettes could not be generated in these cases, the gaps were filled by randomly sampling input from the sequences of the same subject with the same activity label.

4.2 Silhouettes

In order to keep the results consistent, a similar method to the method proposed by Masullo *et al.* [45] was used to extract silhouettes from the video data. As such, average silhouettes are also used to model the temporal structure in the video data. For N time intervals Δt_n of decreasing length, a multi-scale temporal template is used to produce

average silhouettes over various timescales

$$\Delta t_1 > \Delta t_2 > \cdots > \Delta t_n \tag{4.1}$$

Using silhouettes within the interval $i = [t - \Delta t_k, t]$, an average silhouette was produced for each time interval Δt_k

$$\bar{S}_k = \frac{1}{\Delta t_k} \sum_{i=t-\Delta t_k}^{t} S(i) \tag{4.2}$$

to produce one temporal silhouette $\bar{S}_k$ for each Δt, giving a total of N multi-scale temporal silhouettes. In order to obtain the average silhouette at time t, all the multi-scale temporal silhouettes are stacked into a 3D tensor S^*

$$S_t^* \equiv \bar{S}_1, \bar{S}_2, \ldots, \bar{S}_N \tag{4.3}$$

where the third dimension is the stacked multi-scale temporal silhouette.

4.3 Data Augmentation

Data augmentation is a means of enlarging a dataset by performing minor alterations to the existing dataset and introduce more variations into the data. This is particularly important when working with a smaller dataset like SPHERE-Calorie [62], as this removes biases that could be associated with the recording location or any specific body poses and motions. As temporal silhouettes are being used in this paper, it is not sufficient to perform the traditional method of dealing with moving subjects by cropping the active area and then resizing this area to a fixed size [28]. This is due to how the size of the average silhouette is dependent on the subject's motion for that time interval.

As such, average silhouette images were randomly flipped horizontally, tilted using a rotation range of $\theta = \pm 5°$, and translated both horizontally and vertically using a random shift of $t_x = t_y = \pm 20\%$ during training. The data augmentation parameters used are similar to that of Masullo *et al.* [45] in order to keep the data consistent. It is worth noting that some of the data augmentation procedures produced cropped silhouette images, although this does not pose a problem as there are situations where the subject could be in partial view of the camera.

The data augmentation procedure for the accelerometer data involved multiplying the acceleration magnitude with a scalar randomly sampled from a Gaussian distribution with a mean of 1 and a standard deviation of 0.1, taking inspiration from Um *et al.* [68]. Along with the random changes to the magnitude, the axes of each accelerometer were randomly swapped according to random permutations.

4.4 Network Architecture and Implementation

Since CNNs are commonly used in computer vision, a CNN architecture was chosen using two separate modalities as input, which consists of the 3D average silhouette S_t^* from equation 4.3, and a buffer of accelerometer data using the same time interval $[t - max_k(\Delta t_k), t]$. This network was implemented and trained using Keras, as depicted in Figure 2, and the backend for this was chosen to be Tensorflow. A feature-level fusion approach was used when combining the two silhouette and accelerometer branches in the network. A shallow architecture was implemented using two stacks of layers for both

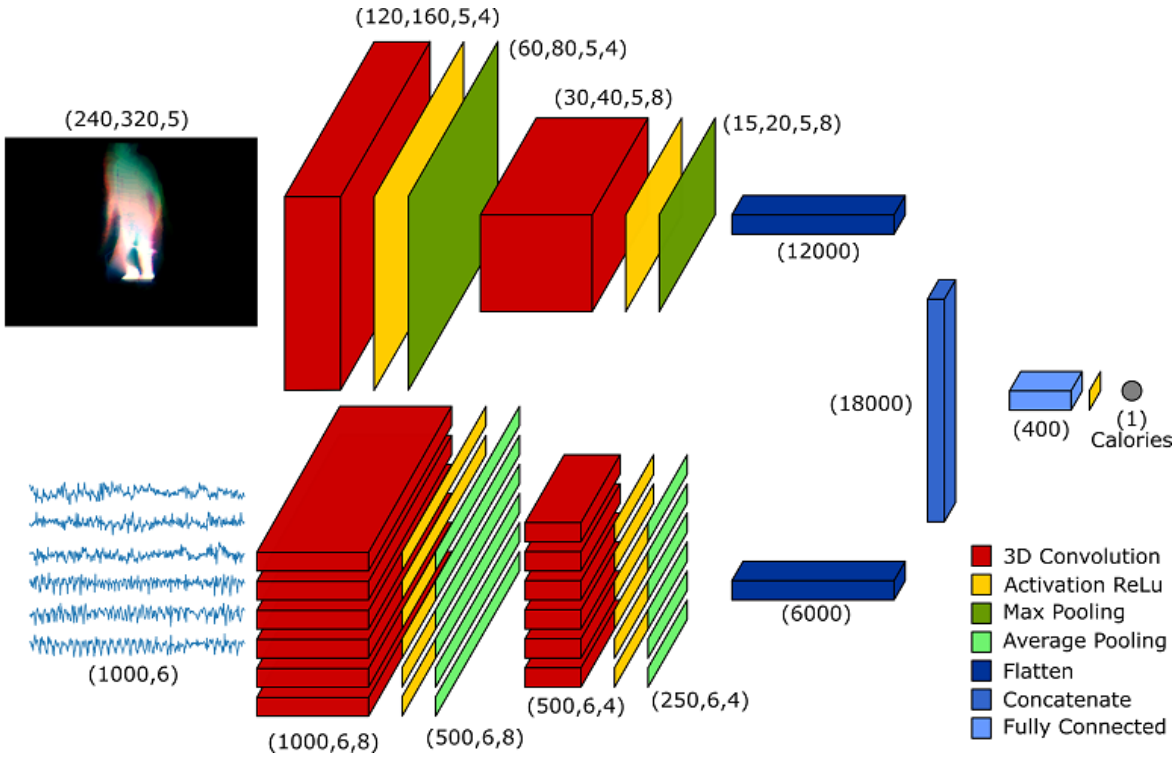

Figure 2: A visualisation of the network architecture, where silhouette data (upper branch) and accelerometer data (lower branch) inputs are fused to produce EE estimation as output.

branches. The silhouette input consisted of a $240 \times 320 \times 5$ tensor, where 5 time intervals Δt are used such that

$$\Delta t_k = \frac{T}{3^k}, k = [0, \ldots, N] \tag{4.4}$$

where T is the maximum buffer size of 1000 in the multi-scale temporal silhouette image, and $N = 4$. For the accelerometer input, a 1000×6 tensor was obtained by using both accelerometers as input and combining them into a 6-channel input. A high pass gravity filter was then applied to the accelerometer data by using a Wiener filter [54] with a window size of 1 second to estimate the gravity vector. The direction from this gravity vector is then subtracted from the accelerometer data to remove the minor effect of gravity on the accelerometer data.

In the case of the silhouettes branch, each stack of layers consisted of a 3D convolution layer followed by a rectified linear unit (ReLu) activation function layer and a max pooling layer. This branch used a pooling size of 2 and a stride length of 2 for each layer. The accelerometer branch similarly made use of two stacks of sequential convolution-activation-pooling layers. In this branch, average pooling was used, along with a kernel size of 5 and a stride length of 2 for each layer as well. The first stack made use of 8 filters, while the second stack used 4.

The silhouette and accelerometer features that were extracted were then concatenated and entered into one final fully connected layer for the estimation of calorific expenditure using regression. The network was trained over 1000 epochs using the mean squared error loss function which compared the error between the ground truth calorie measurements C_{GT} and the estimated calorie expenditure C_P over all times t

$$Loss = \sum_t (C_P^t - C_{GT}^t)^2 \tag{4.5}$$

By selecting the model with the lowest validation loss after the network has been trained for a minimum of 30 epochs, the optimal parameters for the network are obtained. Thus, the network is trainable from end-to-end.

Chapter 5

Results

In order to test my results, leave-one-subject-out cross-validation was used on my network. When training the network, a loss versus epochs graph was produced for each subject being left out. Figure 3 shows this plot for subject 9 being left out. It is observed that both training loss and validation loss does indeed decrease over time as expected, however there is a very long plateau without an increase in loss at the end. This shows that the model achieves convergence but gains no improvements for the majority of epochs, indicating that the training could possibly have been stopped earlier without losing accuracy. Another observation is that the validation loss across all subjects is fairly noisy, which is most likely due to the small size of the validation set even with data augmentation applied, which could be improved with more samples to train the network on. The high variability in training loss can also be interpreted to be due to the learning rate being too high for the dataset.

The root mean square error (RMSE) between the EE estimation (per minute) and the ground truth for each activity was calculated in order to provide an understanding of the accuracy of the model. This was obtained by first calculating the mean errors for each activity type, before averaging across the errors for each subject. In order to evaluate the network's performance, the results obtained from my network were compared against a baseline of the results obtained from METs tables, along with previous state-of-the-art work on the same dataset from Tao *et al.* [63], which used hand-crafted features and non-linear Support Vector Machines (SVMs) for activity classification and a linear support vector regressor for EE estimation, and Masullo *et al.* [45], which made use of a CNN

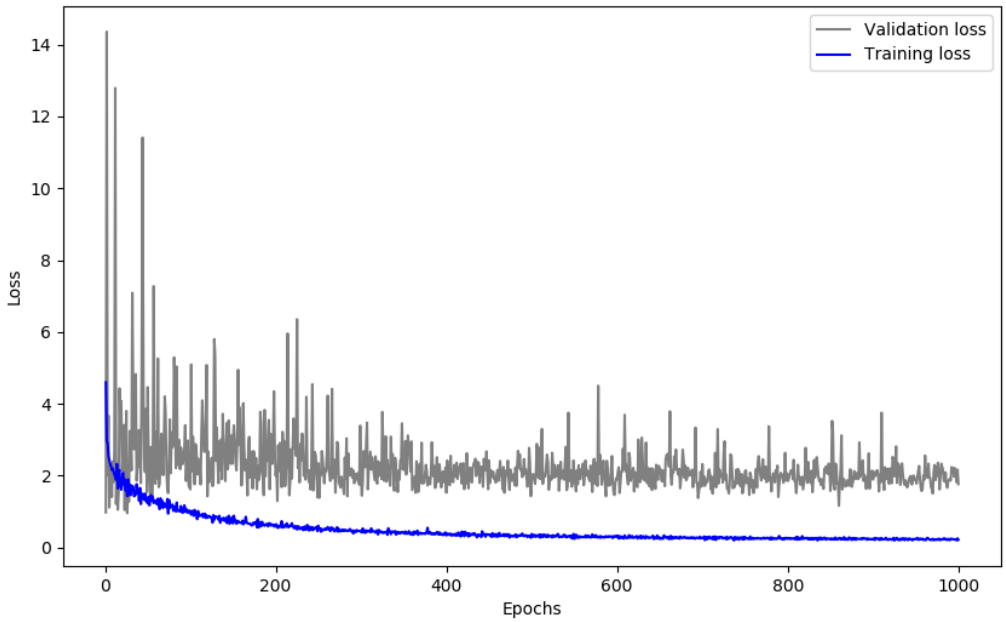

Figure 3: The training loss and validation loss for a single subject being left out (Subject 9) plotted against epochs.

that was trainable end-to-end. Both of these works performed a fusion of both modalities of video data and accelerometer data. A comparison of the results is shown in Figure 4, where the overall error was calculated by calculating the RMSE across all subjects without taking into account the activity performed. Both the methods of METs lookup and Tao *et al.* [63] do not produce EE estimations for activities with no label and hence, have no RMSE associated.

Figure 4 shows that the technique of using METs tables produces the highest overall error of 1.50 cal/min. Since METs tables do not take the individual into account but instead use a statistical approach as mentioned previously, this result is to be expected in a dataset with a small number of subjects. The work from Tao *et al.* [63] produced an overall RMSE of 1.30 cal/min, showing an improvement over that of METs tables for most cases. *CaloriNet*, as proposed by Masullo *et al.* [45], was found to still produce the best results at 0.88 cal/min for overall RMSE, although my network produced very comparable results to Masullo *et al.* [45], producing an overall RMSE of 0.91 cal/min, resulting in a 3.4% difference in overall error. This could be due to differences in how my

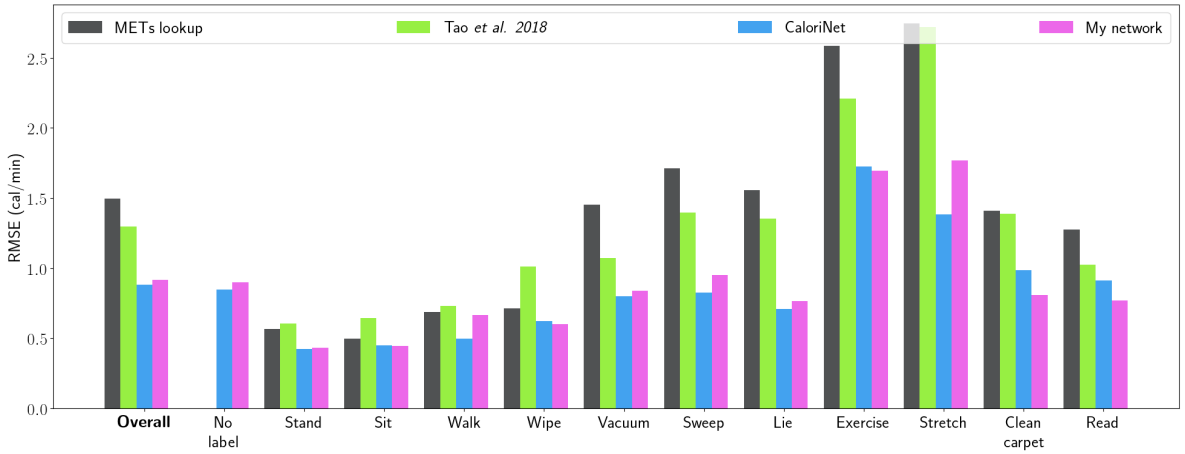

Figure 4: Comparison of the average per-activity RMSE between calorie expenditure estimation and ground truth values.

network was trained, but shows that I've managed to replicate the results of Masullo *et al.* [45] within a reasonable margin of error.

It is also noted that, similar to the findings of Masullo *et al.* [45], all the techniques tested were found to produce a much higher error when estimating EE for the *Exercise* and *Stretch* activity classes. This is highly likely to be due to the inter-class and intra-class variance of these activities which is much higher as compared to the variance of other classes of activities, estimated to be at least 20 times higher [45]. In order to solve the problem of a small training dataset, a wider range of activity classes could be sampled, with more subjects from different backgrounds to introduce more variability in subject metabolism and physical fitness. A richer dataset such as this could reduce this error and lead to better accuracy in EE estimation.

Chapter 6

Conclusion

In this thesis, we have recreated the work by Masullo *et al.* on *CaloriNet* to a reasonable degree of accuracy. The fused CNN created is able to succesfully estimate EE from a combination of silhouette data and accelerometer data with the accuracy of the current state-of-the-art. This method is thus suitable for the use of online EE estimation of household activities during normal daily living in the home environment as it preserves the privacy and anonymity of the individual, and allows the individual to move around freely without being inconvenienced by wearing calorimetry instruments.

In Chapter 2 we provided a comprehensive overview of the different technologies and methodologies for EE estimation. We then provided a thorough discussion of the current approaches to activity recognition and EE estimation using computer vision.

In Chapter 3, we analysed the different technologies required to perform EE estimation in the home environment while maintaining privacy. We looked at two techniques of extracting silhouette data from RGB-D data, before providing a discussion of temporal modelling through the use of average silhouettes. We then provided a broad explanation of the theory underlying neural networks and the properties of some neural network architectures.

In Chapter 4, we outline the details of the developed method for EE estimation using silhouettes and accelerometer data. First, we described the SPHERE-Calorie dataset in section 4.1. In section 4.2, we explain the process of generating multi-scale temporal sil-

houettes from raw silhouettes. Next, in section 4.3, we discussed how data augmentation was performed on the SPHERE-Calorie dataset in order to enlarge the existing dataset. Finally, section 4.4 provides a detailed explanation of my implementation of the network architecture proposed by Masullo *et al.* [45]

In Chapter 5, we presented the process of evaluating the trained network. First, we analysed the loss versus epochs graph when training the network. We then evaluated my model by comparing the performance, in terms of RMSE, of my model against a standard METs table approach and against previous work on the same dataset: the work of Tao *et al.* [63] and Masullo *et al.* [45]. This was followed by a discussion of the results and the challenges associated with the developed method.

6.1 Future Work

Going forward, there are still improvements that could be made to the estimation accuracy of *CaloriNet* while still maintaining the data-fusion approach. We propose three different directions in which to extend this project for future work.

First, the trained model is only able to produce calorie expenditure estimations as its output. It might be useful for it to be trained in such a way that the physical activity labels are also produced as a target. This could potentially increase the accuracy of the network when trained end-to-end to produce activity labels on its own. Moreover, this might also increase the robustness of the network to be able to handle even more activity classes.

Another extension would be to explore and compare the effectiveness of using a different network architecture entirely. Due to how little the area of EE estimation using computer vision has been explored, not as much research has been done on the benefit of using different network architectures such as RNNs, and using different fusion approaches for the two modalities of silhouettes and accelerometer data. As such, there is great value in investigating how different network architectures and fusion approaches might affect the accuracy and efficiency of EE estimation.

Lastly, the model could be extended to be trained on a larger and more comprehensive dataset. Since there are very few studies looking at EE estimation using both RGB-D data and accelerometer data, an ambitious extension for this project would be to extend the dataset further to include more participants and a wider range of activity classes, allowing for better accuracy of EE estimation.

Bibliography

[1] Barbara E. Ainsworth, William L. Haskell, Stephen D. Herrmann, Nathanael Meckes, David R. Bassett, Jr., David R. Jacobs, Jr., Catrine Tudor-Locke, Jennifer L. Greer, Jesse Vezina, Melicia C. Whitt-Glover, and Arthur S. Leon. 2011 Compendium of Physical Activities: A Second Update of Codes and MET Values. *Medicine & Science in Sports & Exercise*, 43(8):1575–1581, 2011.

[2] Barbara E. Ainsworth, William L. Haskell, Arthur S. Leon, David R. Jacobs, Jr., Henry J. Montoye, James F. Sallis, and Ralph S. Paffenbarger, Jr. Compendium of Physical Activities: classication of energy costs of human physical activities. *Medicine & Science in Sports & Exercise*, 25(1):71–80, 1993.

[3] Marco Altini, Pierluigi Casale, Julien Penders, and Oliver Amft. Personalized cardiorespiratory fitness and energy expenditure estimation using hierarchical bayesian models. *Journal of Biomedical Informatics*, 56:195–204, 2015.

[4] Marco Altini, Julien Penders, Ruud Vullers, and Oliver Amft. Estimating energy expenditure using body-worn accelerometers: A comparison of methods, sensors number and positioning. *IEEE Journal of Biomedical and Health Informatics*, 19(1):219–226, Jan 2015.

[5] Ignacio Ara, Raquel Aparicio-Ugarriza, David Morales-Barco, Wysllenny Souza, Esmeralda Mata, and Marcela González-Gross. Physical activity assessment in the general population; validated self-report methods. *Nutricion Hospitalaria*, 31(Suppl 3):211–218, 02 2015.

[6] Hakan Bilen, Basura Fernando, Efstratios Gavves, Andrea Vedaldi, and Stephen Gould. Dynamic Image Networks for Action Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3034–3042, June 2016.

[7] Giles Birchley, Richard Huxtable, Madeleine Murtagh, Ruud ter Meulen, Peter Flach, and Rachael Gooberman-Hill. Smart homes, private homes? an empirical study of technology researchers' perceptions of ethical issues in developing smart-home health technologies. *BMC Medical Ethics*, 18(1):23, Apr 2017.

[8] Christopher M. Bishop. *Pattern Recognition and Machine Learning*, chapter 5, pages 225–269. Springer-Verlag New York, 2006.

[9] A. F. Bobick and J. W. Davis. The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(3):257–267, March 2001.

[10] C. V. Bouten, W. P. Verboeket-van de Venne, K. R. Westerterp, M. Verduin, and J. D. Janssen. Daily physical activity assessment: comparison between movement registration and doubly labeled water. *Journal of Applied Physiology*, 81(2):1019–1026, 1996. PMID: 8872675.

[11] Y. Y. Boykov and M. P. Jolly. Interactive graph cuts for optimal boundary & region segmentation of objects in N-D images. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, volume 1, pages 105–112, July 2001.

[12] S. Brage, N. Brage, P. W. Franks, U. Ekelund, and N. J. Wareham. Reliability and validity of the combined heart rate and movement sensor Actiheart. *Eur J Clin Nutr*, 59(4):561–570, Apr 2005.

[13] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *CoRR*, abs/1812.08008, 2018.

[14] Sana M. Ceesay, Andrew M. Prentice, Kenneth C. Day, Peter R. Murgatroyd, Gail R. Goldberg, Wendy Scott, and G. B. Spurr. The use of heart rate monitoring in the estimation of energy expenditure: a validation study using indirect whole-body calorimetry. *British Journal of Nutrition*, 61(2):175186, 1989.

[15] Francesco S Celi, Robert J Brychta, Joyce D Linderman, Peter W Butler, Anna Teresa Alberobello, Sheila Smith, Amber B Courville, Edwin W Lai, Rene Costello, Monica C Skarulis, Gyorgy Csako, Alan Remaley, Karel Pacak, and Kong Y

Chen. Minimal changes in environmental temperature result in a significant increase in energy expenditure and changes in the hormonal homeostasis in healthy adults. *European Journal of Endocrinology*, 163(6):863–872, 2010.

[16] A. A. Chaaraoui, J. R. Padilla-Lopez, F. J. Ferrandez-Pastor, M. Nieto-Hidalgo, and F. Florez-Revuelta. A vision-based system for intelligent monitoring: human behaviour analysis and privacy by context. *Sensors (Basel)*, 14(5):8895–8925, May 2014.

[17] Kong Y Chen and David R Bassett. The Technology of Accelerometry-Based Activity Monitors: Current and Future. *Medicine and science in sports and exercise*, 37(11 Suppl):S490–S500, November 2005.

[18] Wojtek J. Chodzko-Zajko, David N. Proctor, Maria A. Fiatarone Singh, Christopher T. Minson, Claudio R. Nigg, George J. Salem, and James S. Skinner. Exercise and physical activity for older adults. *Medicine & Science in Sports & Exercise*, 41(7):1510–1530, 2009.

[19] Scott Crouter, Carolyn Albright, and David Bassett. Accuracy of polar S410 heart rate monitor to estimate energy cost of exercise. *Medicine & Science in Sports & Exercise*, 36(8):1433–1439, Aug 2004.

[20] M. J. Dauncey and W. P. T. James. Assessment of the heart-rate method for determining energy expenditure in man, using a whole-body calorimeter. *British Journal of Nutrition*, 42(1):113, 1979.

[21] Keith M. Diaz, David J. Krupka, Melinda J. Chang, James Peacock, Yao Ma, Jeff Goldsmith, Joseph E. Schwartz, and Karina W. Davidson. Fitbit®: An accurate and reliable device for wireless physical activity tracking. *International Journal of Cardiology*, 185:138–140, 2015.

[22] Katherine Ellis, Jacqueline Kerr, Suneeta Godbole, Gert Lanckriet, David Wing, and Simon Marshall. A random forest classifier for the prediction of energy expenditure and type of physical activity from wrist and hip accelerometers. *Physiological Measurement*, 35(11):2191–2203, Oct 2014.

[23] Patty Freedson, David Pober, and Kathleen F. Janz. Calibration of accelerometer output for children. *Medicine & Science in Sports & Exercise*, 37(11 Suppl):S523–530, Nov 2005.

[24] Riddhi Das Gupta, Roshna Ramachandran, Padmanaban Venkatesan, Shajith Anoop, Mini Joseph, and Nihal Thomas. Indirect calorimetry: From bench to bedside. *Indian Journal of Endocrinology and Metabolism*, 21(4):594–599, 2017.

[25] J. Hall, S. Hannuna, M. Camplani, M. Mirmehdi, D. Damen, T. Burghardt, L. Tao, A. Paiement, and I. Craddock. 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)*, pages 1–6, Oct 2016.

[26] L. L. Hardy, A. P. Hills, A. Timperio, D. Cliff, D. Lubans, P. J. Morgan, B. J. Taylor, and H. Brown. A hitchhiker's guide to assessing sedentary behaviour among young people: deciding what method to use. *J Sci Med Sport*, 16(1):28–35, Jan 2013.

[27] William L. Haskell, I-Min Lee, Russell R. Pate, Kenneth E. Powell, Steven N. Blair, Barry A. Franklin, Caroline A. Macera, Gregory W. Heath, Paul D. Thompson, and Adrian Bauman. Physical activity and public health: Updated recommendation for adults from the american college of sports medicine and the american heart association. *Circulation*, 116(9):1081–1093, 2007.

[28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 346–361. Springer International Publishing, 2014.

[29] D. Hendelman, K. Miller, C. Baggett, E. Debold, and P. Freedson. Validity of accelerometry for the assessment of moderate intensity physical activity in the field. *Medicine & Science in Sports & Exercise*, 32(9 Suppl):S442–449, Sep 2000.

[30] Jason I. Hong and James A. Landay. An architecture for privacy-sensitive ubiquitous computing. In *Proceedings of the 2nd International Conference on Mobile Systems, Applications, and Services*, MobiSys '04, pages 177–189, 2004.

[31] Joe Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici. Beyond short snippets: Deep networks for video classification. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4694–4702, June 2015.

[32] Karl J. Kaiyala and Douglas S. Ramsay. Direct animal calorimetry, the underused gold standard for quantifying the fire of life. *Comparative Biochemistry and Physiology Part A: Molecular & Integrative Physiology*, 158(3):252–264, 2011. The challenge of measuring energy expenditure: current field and laboratory methods.

[33] Shian-Ru Ke, Hoang Le Uyen Thuc, Yong-Jin Lee, Jenq-Neng Hwang, Jang-Hee Yoo, and Kyoung-Ho Choi. A Review on Video-Based Human Activity Recognition. *Computers*, 2(2):88–131, 2013.

[34] Glen P. Kenny, Sean R. Notley, and Daniel Gagnon. Direct calorimetry: a brief historical review of its use in the study of human metabolism and thermoregulation. *European Journal of Applied Physiology*, 117(9):1765–1785, Sep 2017.

[35] Antero Y. Kesaniemi, Elliot Danforth Jr., Michael D. Jensen, Peter G. Kopelman, Pierre Lefébvre, and Bruce A. Reeder. Dose-response issues concerning physical activity and health: an evidence-based symposium. *Medicine & Science in Sports & Exercise*, 33(6):S351–S358, 2001.

[36] Youngwon Kim, Michael W. Beets, and Gregory J. Welk. Everything you wanted to know about selecting the right Actigraph accelerometer cut-points for youth, but: A systematic review. *Journal of Science and Medicine in Sport*, 15(4):311–321, 2012.

[37] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, pages 1097–1105, USA, 2012. Curran Associates Inc.

[38] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, June 2008.

[39] G. Lavee, E. Rivlin, and M. Rudzsky. Understanding video events: A survey of methods for automatic interpretation of semantic occurrences in video. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 39(5):489–504, Sep. 2009.

[40] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation*, 1(4):541–551, Dec 1989.

[41] I-Min Lee and Patrick Skerrett. Physical activity and all-cause mortality: what is the dose-response relation? *Medicine and science in sports and exercise*, 33(6 Suppl):S459S471; discussion S4934, June 2001.

[42] William R. Leonard. Laboratory and field methods for measuring human energy expenditure. *American Journal of Human Biology*, 24(3):372–384, March 2012.

[43] M. B. E. Livingstone. Heart-rate monitoring: the answer for assessing energy expenditure and physical activity in population studies? *British Journal of Nutrition*, 78(6):869871, 1997.

[44] Paul D. Loprinzi, Hyo Lee, Bradley J. Cardinal, Carlos J. Crespo, Ross E. Andersen, and Ellen Smit. The relationship of actigraph accelerometer cut-points for estimating physical activity with selected health outcomes. *Research Quarterly for Exercise and Sport*, 83(3):422–430, 2012. PMID: 22978192.

[45] Alessandro Masullo, Tilo Burghardt, Dima Damen, Sion L. Hannuna, Víctor Ponce-López, and Majid Mirmehdi. CaloriNet: From silhouettes to calorie estimation in private environments. *CoRR*, abs/1806.08152, 2018.

[46] Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T Dudley. Deep learning for healthcare: review, opportunities and challenges. *Briefings in Bioinformatics*, 19(6):1236–1246, 05 2017.

[47] S. J. Morris and J. A. Paradiso. Shoe-integrated sensor system for wireless gait analysis and real-time feedback. In *Proceedings of the Second Joint 24th Annual Conference and the Annual Fall Meeting of the Biomedical Engineering Society] [Engineering in Medicine and Biology*, volume 3, pages 2468–2469, Oct 2002.

[48] Katsuyuki Nakamura, Serena Yeung, Alexandre Alahi, and Li Fei-Fei. Jointly Learning Energy Expenditures and Activities Using Egocentric Multimodal Signals. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6817–6826, July 2017.

[49] Didace Ndahimana and Eun-Kyung Kim. Measurement Methods for Physical Activity and Energy Expenditure: a Review. *Clinical Nutrition Research*, 6(2):68–80, April 2017.

[50] J. Adam Noah, David K. Spierer, Jialu Gu, and Shaw Bronner. Comparison of steps and energy expenditure assessment in adults of fitbit tracker and ultra to the actical and indirect calorimetry. *Journal of Medical Engineering & Technology*, 37(7):456–462, 2013.

[51] Hamed Pirsiavash and Deva Ramanan. Parsing Videos of Actions with Segmental Grammars. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '14, pages 612–619, Washington, DC, USA, 2014. IEEE Computer Society.

[52] SPHERE Project. https://www.irc-sphere.ac.uk/research, 2018.

[53] Partha P. Ray. Internet of things based physical activity monitoring (pamiot): An architectural framework to monitor human physical activity. In *Proceedings of IEEE Calcutta Conference*, pages 32–34, 2014.

[54] Peter Rizun. Optimal Wiener Filter for a Body Mounted Inertial Attitude Sensor. *Journal of Navigation*, 61(3):455472, 2008.

[55] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. "GrabCut": Interactive Foreground Extraction using Iterated Graph Cuts. In *ACM SIGGRAPH 2004 Papers*, SIGGRAPH '04, pages 309–314, New York, NY, USA, 2004. ACM.

[56] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, October 1986.

[57] Guenther Samitz, Matthias Egger, and Marcel Zwahlen. Domains of physical activity and all-cause mortality: systematic review and doseresponse meta-analysis of cohort studies. *International Journal of Epidemiology*, 40(5):1382–1400, 09 2011.

[58] Peter Scarborough, Prachi Bhatnagar, Kremlin K. Wickramasinghe, Steve Allender, Charlie Foster, and Mike Rayner. The economic burden of ill health due to diet, physical inactivity, smoking, alcohol and obesity in the UK: an update to 2006-07 NHS costs. *Journal of Public Health*, 33(4):527–535, December 2011.

[59] Karen Simonyan and Andrew Zisserman. Two-Stream Convolutional Networks for Action Recognition in Videos. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 568–576. Curran Associates, Inc., 2014.

[60] John R. Sirard and Russell R. Pate. Physical Activity Assessment in Children and Adolescents. *Sports Medicine*, 31(6):439–454, May 2001.

[61] Louisa G. Sylvia, Emily E. Bernstein, Jane L. Hubbard, Leigh Keating, and Ellen J. Anderson. A Practical Guide to Measuring Physical Activity. *Journal of the Academy of Nutrition and Dietetics*, 114(2):199–208, Feb 2013.

[62] Lili Tao. SPHERE-Calorie, 2017.

[63] Lili Tao, Tilo Burghardt, Majid Mirmehdi, Dima Damen, Ashley Cooper, Massimo Camplani, Sion Hannuna, Adeline Paiement, and Ian Craddock. Energy expenditure estimation using visual and inertial sensors. *IET Computer Vision*, 12(1):36–47, February 2018.

[64] Lili Tao, Tilo Burghardt, Majid Mirmehdi, Dima Damen, Ashley Cooper, Sion Hannuna, Massimo Camplani, Adeline Paiement, and Ian Craddock. Calorie counter: Rgb-depth visual estimation of energy expenditure at home. In Chu-Song Chen, Jiwen Lu, and Kai-Kuang Ma, editors, *Computer Vision – ACCV 2016 Workshops*, pages 239–251, Cham, 2017. Springer International Publishing.

[65] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning Spatiotemporal Features with 3D Convolutional Networks. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, pages 4489–4497, Washington, DC, USA, 2015. IEEE Computer Society.

[66] P. Tsou and C. Wu. Estimation of Calories Consumption for Aerobics Using Kinect Based Skeleton Tracking. In *2015 IEEE International Conference on Systems, Man, and Cybernetics*, pages 1221–1226, Oct 2015.

[67] P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea. Machine Recognition of Human Activities: A Survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(11):1473–1488, Nov 2008.

[68] Terry T. Um, Franz M. J. Pfister, Daniel Pichler, Satoshi Endo, Muriel Lang, Sandra Hirche, Urban Fietzek, and Dana Kulić. Data Augmentation of Wearable Sensor Data for Parkinson's Disease Monitoring Using Convolutional Neural Networks. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, ICMI '17, pages 216–220, New York, NY, USA, 2017. ACM.

[69] Vincent T. van Hees, Rob C. van Lummel, and Klaas R. Westerterp. Estimating Activity-related Energy Expenditure Under Sedentary Conditions Using a Tri-axial Seismic Accelerometer. *Obesity*, 17(6):1287–1292, 2009.

[70] H. Wang, A. Klser, C. Schmid, and C. Liu. Action recognition by dense trajectories. In *CVPR 2011*, pages 3169–3176, June 2011.

[71] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool. Temporal segment networks for action recognition in videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2018.

[72] Gregory J. Welk, Charles B. Corbin, and Darren Dale. Measurement Issues in the Assessment of Physical Activity in Children. *Research Quarterly for Exercise and Sport*, 71(sup2):59–73, 2000. PMID: 25680015.

[73] L. Yeffet and L. Wolf. Local Trinary Patterns for human action recognition. In *2009 IEEE 12th International Conference on Computer Vision*, pages 492–497, Sep. 2009.

[74] K. Zhang, F. X. Pi-Sunyer, and C. N. Boozer. Improving energy expenditure estimation for physical activity. *Med Sci Sports Exerc*, 36(5):883–889, May 2004.

[75] Božidara Cvetković, Robert Szeklicki, Vito Janko, Przemyslaw Lutomski, and Mitja Luštrek. Real-time activity monitoring with a wristband and a smartphone. *Information Fusion*, 43:77–93, 2018.

[76] Martina Ziefle, Carsten Rocker, and Andreas Holzinger. Medical technology in smart homes: Exploring the user's perspective on privacy, intimacy and trust. In *2011 IEEE 35th Annual Computer Software and Applications Conference Workshops*, pages 410–415, July 2011.