Calorific Expenditure Estimation using Deep Convolutional Network Features

Baodong Wang
University of Bristol
bw16221@my.bristol.ac.uk

Lili Tao
University of the West of England
lili.tao@uwe.ac.uk

Tilo Burghardt
University of Bristol
tilo@cs.bris.ac.uk

Majid Mirmehdi
University of Bristol
m.mirmehdi@bristol.ac.uk

Abstract

Accurately estimating a person's energy expenditure is an important tool in tracking physical activity levels for healthcare and sports monitoring tasks, amongst other applications. In this paper, we propose a method for deriving calorific expenditure based on deep convolutional neural network features (within a healthcare scenario). Our evaluation shows that the proposed approach gives high accuracy in activity recognition (82.3%) and low normalised root mean square error in calorific expenditure prediction (0.41). It is compared against the current state-of-the-art calorific expenditure estimation method, based on a classical approach, and exhibits an improvement of 7.8% in the calorific expenditure prediction task. The proposed method is suitable for home monitoring in a controlled environment.

1. Introduction

One aspect of treatment for common conditions and chronic diseases alike, for example diabetes, obesity, dementia, cardiac and respiratory diseases, is regular physical activity. Clinicians require this to be recorded and measured to facilitate a more accurate understanding of its effectiveness and planning for further patient care. Currently, the direct measuring of physical activity intensity levels on a day-to-day basis is subjective and variable when it is reported by patients [25], but the results can be inaccurate and cantankerous to obtain.

Calorific expenditure is one commonly used single metric to quantify physical activity levels over time. A calorimeter may be the most accurate measuring device in existence, which operates based on the respiratory differences of oxygen and CO_2 in the inhaled and exhaled air. Such devices can either be a sealed respiratory chamber [24] or an indirect portable device which requires carrying gas sensors and wearing a breathing mask [1]. However, due to their inherently cumbersome nature and high cost, they are impractical to use in daily life. Recently, wearable devices have become popular for assessing physical activity of individuals [6, 32, 8]. Among these, tri-axial accelerometers are the most broadly used inertial sensors [8]. Whilst appropriate to inferring coarse categorisations of activity intensity levels, based on this data alone it is difficult to produce precise calorific expenditure values [29].

Intelligent visual monitoring has received a great deal of attention in the recent years, and is being increasingly deployed in the development of smart homes, e.g. [33, 12], to assist with the diagnosis and management of health conditions. Deep learning has been successfully applied to various application domains, such as image recognition [16], action classification [14] and pose estimation [30]. There are a few recent studies on healthcare related applications using deep learning framework [9, 21], however to the best of our knowledge, our work is the first one targeted at improving the estimation of energy expenditure using visual data only utilising deep neural networks.

In this work, we use a convolutional neural network (CNN) to extract features for estimating calorific expenditure from video data in an indoor environment as illustrated in Figure 1. The proposed method maps extracted visual features to calorie estimates via activity-specific models. The method explicitly detects activities as an intermediate component to aid the visual estimate of energy expenditure by selecting activity-specific mappings for the calorific estimation.

This is a relatively new application in computer vision for which very few datasets are available. The 'In-home activity recordings' dataset [10] includes three 16-hour days of in-door activities, however, the ground truth is collected by a wearable sensor, which cannot ensure an accurate benchmark. We therefore will evaluate our results on the only other comprehensive dataset available,

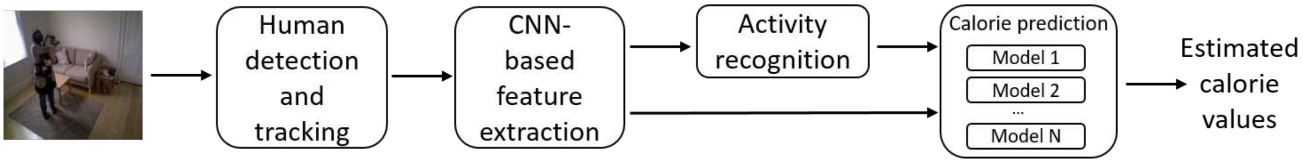


Figure 1. **Overview of the Framework.** The proposed pipeline uses video footage as input and extracts relevant features using a CNN, before exploiting these features for activity recognition, and finally applying an activity-specific regressor towards calorie expenditure.

i.e. the SPHERE-calorie dataset [29, 2], which contains accurate ground truth collected with gas-exchange measurements obtained from a calorimeter and more than ten hours of visual data. We evaluate the performance of our proposed method on this dataset and compare it against the state-of-the-art that uses hand-crafted visual features [27].

The structure of the paper is organized as follows: Section 2 reviews the state-of-the-art techniques on related areas. Section 3 introduces the proposed method for estimating calorific expenditure values from the image data. Section 4 presents the system setup and the experimental results. Conclusion and future work are given in Section 5.

2. Related Work

Applying computer vision based monitoring techniques to help with the management of health-related conditions has gained considerable attention in the last decade [22, 7]. However, studies on energy expenditure using visual sensors are still strictly limited, possibly due to the absence of labeled data, which is difficult to obtain - not least given the ethical requirements necessary to obtain it. Our work stems from the SPHERE project [33], which has created the circumstances to generate and release a suitable dataset for us and the vision community to explore [27, 29, 2]. The proposed method is built on several relevant subject areas in computer vision.

2.1. Visual Feature Extraction

Visual feature extraction and representation is a core component of human activity analysis framework. Hand-crafted features are typically used to capture low-level information, such as shape, color and appearance. Body configuration and body motion are then also imperative to infer calorific expenditure [27]. The potential features include local interest point configurations [17], holistic approaches like histograms of oriented gradients, and histograms of motion information [26]. In contrast, deep learning models can learn a hierarchy of features by building high-level descriptions from low-level ones. CNNs have been shown to learn powerful and interpretable image features [16]. Encouraged by such positive results in images, we applied CNNs to learning visual features in this work.

Body motion information is an important indicator when estimating calorific expenditure. In fact, calorific expenditure is highly dependent on the motions and actions performed before an estimation point, and thus, the formulation of features for summative interpretations requires the concatenation of per-frame descriptors over time[26]. However, this can result in a high dimensional feature space that comes at a high computational demand.

2.2. Action Recognition

Human actions can be inferred from colour [3], colour and depth [4] and skeleton-based data [23]. Deep neural networks have also significantly improved action recognition compared to traditional techniques [13].The knowledge about the type of action is strongly correlated with calorie expenditure [5]. While great progress has been made on human action recognition [4] there are still many challenges left to address the range and complexity of human motions and actions in practical, real-world applications, such as in a healthcare scenario within the home environment.

In this work, we follow the action recognition stage in [27] as an intermediate component in our work. It should be noted that action recognition is not the focus of this paper, and it can be replaced by any other appropriate action recognition methodology for the calorie estimation strategy proposed here.

2.3. Calorific Expenditure Estimation

Edgecomb and Vahid [10] estimated daily energy expenditure using RGB video, albeit rather coarsely. Their method firstly segmented the foreground subject from the scene background, and then estimated the calorific uptake based on vertical and horizontal velocities and accelerations, and the changes in height and width of the subject's bounding box. A full set of 3D joint movements from the skeleton data (for example from a Microsoft Kinect) can also be used for estimating calorie consumption [31]. However, skeleton data is usually noisy and is only potentially reliably available when the subject is facing the camera, which would make it difficult to provide accurate calorie values in more unconstrained scenarios. A common issue of the above methods is that the ground truth used for training the models is based on wearable accelerometers. The

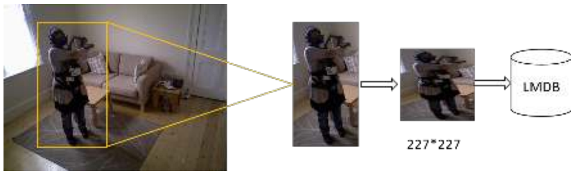


Figure 2. **Image Preprocessing.** Bounding box regions containing humans are extracted using OpenNI, re-scaled to 227 × 227 pixels and normalised before encoding them via an LMDB database.

general layout has been taken further here using a portable calorimeter, to provide more accurate ground truth readings. The recent work in [27] introduced a vision-based framework for estimating calorific expenditure in a home environment. This method has become a baseline method in the area [28, 29]. An extension enables the estimation of physical activity intensity levels in real-time [28]. However, the light-weight features extracted from bounding boxes can only estimate calorific expenditure coarsely. To improve on this setup, a further step was taken to fuse visual sensor data with accelerometer data in order to reduce the estimation error [29]. In this paper, we consider the use of a single visual sensor only and compare a deep architecture against the baseline method.

3. Method

We present a calorific expenditure estimation pipeline using an activity-specific model with CNN-based visual features. {Inspired by [27], the activities are reasoned about first, and then calorie values are estimated based on the extracted CNN features via a set of models, which are each separately trained for certain activities. The block diagram of the proposed method is shown in Figure 1 and consists of four components: (i) human detection and tracking, (ii) CNN-based feature extraction, (iii) activity recognition, and (iv) calorie prediction. The RGB stream is preprocessed by detecting and tracking the subjects and representing the regions of interest by bounding boxes. Per-frame features are then extracted by training a convolutional neural network based on the AlexNet architecture [16] provided with the Caffe Library [15]. Principal Components Analysis is applied to reduce the dimensionality of the resulting features before temporal pooling is applied to represent and encode motion information. These features are then used by a classifier to determine activity types and by a bank of regressors to achieve calorie prediction, one regressor per activity type.

3.1. Deep CNN Features

Data pre-processing - We extract the features for bounding boxes returned by the OpenNI SDK [20] person detector and tracker. To normalise the utilised image region due to varying heights of the subjects and their distance to the camera, the bounding box is re-scaled to a square size of 227 × 227 pixels. The image is also normalised by subtracting the mean, which centres the input to 0, and divided by the standard deviation. Considering that the dataset is relatively large, in order to improve the speed of data reading/writing and to reduce the training time, we use the lightning memory-mapped database (LMDB) [18] over the HDF5 file format. LMDB is the database of choice when using Caffe with large datasets. This pre-processing stage is illustrated in Figure 2.

Architecture - A flowchart of the detailed feature extraction procedure is depicted in Figure 3. A deep convolutional neural network, consisting of multiple trainable stages, is employed to extract features hierarchically, as in Figure 3 (top), where we implement the AlexNet network provided with the Caffe Library. We follow the architecture proposed in AlexNet such that the net consists of five convolutional layers. In the first two convolutional layers (fc1 and fc2) convolution operations are followed by Rectified Linear Units (ReLU) as activation functions, followed by max-pooling and normalization operations. The third and fourth convolutional layers (fc3 and fc4) only contain the convolution operations. The fifth convolutional layer (fc5) contains the convolution and max-pooling process, which provides a 4096-dimensional data output to the fully-connected layers. The sixth and seventh layer are fully-connected (fc6 and fc7) using standard 'ReLU' activation functions and dropout. Since there are 11 activities in the dataset, the fully-connected layers lead into a final 11-way softmax layer, which produces a score distribution over class labels.

3.2. Dimensionality Reduction

Instead of deploying the fc7 layer to form feature vectors, as often done, an earlier layer is utilised here in order to capture more general purpose semantics. The proposed method takes the features at the fc6 layer, which is a 4096 dimensional data vector, as a per-frame image descriptor. Although it is possible to encode the motion information as a sequence of per-frame descriptors, its overall dimensionality would be very high (e.g. >40k dimensions in our experiments) exceeding our computational capability. To reduce high dimensionality, we applied a Principal Compoents Analysis (PCA) to vectors for dimension reduction. As it is shown in [19], applying PCA in CNN-based features is helpful not only for dimension reduction, but also for boosting performance. Figure 3 (bottom) shows an example of applying PCA for dimensionality reduction. By keeping the first 1000 dimensions of the PCA-processed data we retain 93.7% of the variance. We use these feature vectors as our final CNN-based per-frame descriptor.

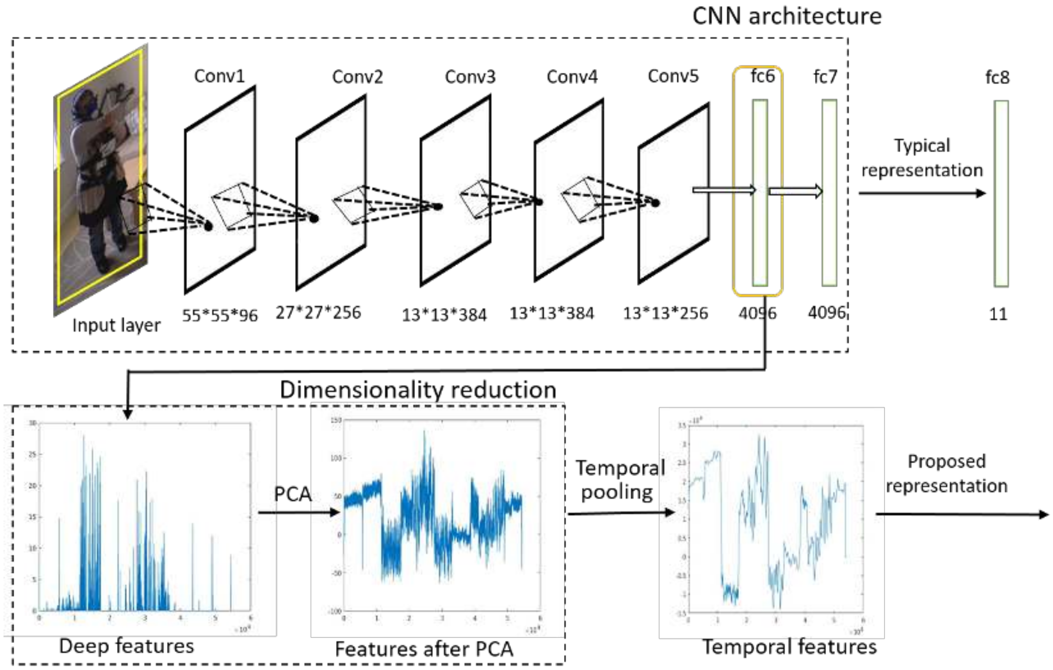


Figure 3. **Overview of Feature Representation.** Fully-connected layer 7 (fc7) of AlexNet pre-trained on ImageNet is commonly used for image classification tasks. However, the proposed approach uses an earlier layer (i.e. fc6) to capture more general purpose semantics (Top); and we show that PCA-compressed and temporally pooled versions of this information are suitable to be used for classification applications related to calorie estimation (Bottom).

3.3. Temporal Pooling

Given the CNN-based descriptor extracted from each frame in a sequence of images, it is important to capture the temporal changes and summarise them to represent the motion in the video. Notice, that the human calorific expenditure depends on both short and long term changes of data. We present a temporal pyramid structure to model information from various temporal window sizes in a single descriptor, as shown in Figure 4.

The figure illustrates the input data and the structure of the image descriptor (left). Implementing a pyramid scheme, time series data is represented over various levels, where per level i there is a set of 2^i time segments as $[\mathbf{S}_i^1, \ldots, \mathbf{S}_i^{(2^i)}]$. The time series data can also be explained in matrix form $\mathbf{S} \in \mathbb{R}^{T \times N}$as T per-frame feature vectors, such that $\mathbf{S} = \{S_1, \ldots, S_N\}$ for a video. N represents the length of the per-frame feature vector (e.g. in our case, $N = 1000$), and T is the number of frames. A time series of a single feature is denoted $S_n = [s_n(1), \ldots, s_n(T)]$ tracing the n^{th} feature across $1, \ldots, T$ frames, where $s_n(t)$ denotes the n^{th} feature at frame t. For each segment $[t_{min}, t_{max}]$, a set of temporal filters with multiple pooling operators is applied, which produces a single feature vector for each segment via concatenation. Frequency domain pooling is used along with two conventional pooling operators, max pooling and sum pooling, respectively defined as:

$$\mathcal{O}_{\max}(S_n) = \max_{t=t_{\min}\cdots t_{\max}} s_n(t), \tag{1}$$

$$\mathcal{O}_{\text{sum}}(S_n) = \sum_{t=t_{\min}}^{t_{max}} s_n(t). \tag{2}$$

Frequency domain pooling is employed to represent the time series S_n in the frequency domain by the discrete cosine transform, where the pooling operator takes the absolute value of the j lowest frequency components of the frequency coefficients,

$$\mathcal{O}_{\text{dct}}(S_n) = |M_{1:j} S_n| , \tag{3}$$

where M is the discrete cosine transformation matrix. The final feature representation is a concatenation of multiple pooling operators applied to each time segment at each level.

3.4. Activity Recognition and Calorie Prediction

The activity is reasoned about first via an SVM using the pooled motion features introduced above. For each activity category, a SVM-based regression model is trained for estimating the calorie expenditure. It is shown in [27] that such an activity specific method has the potential for improved performance compared to predicting the calorific expenditure without knowing the activity.

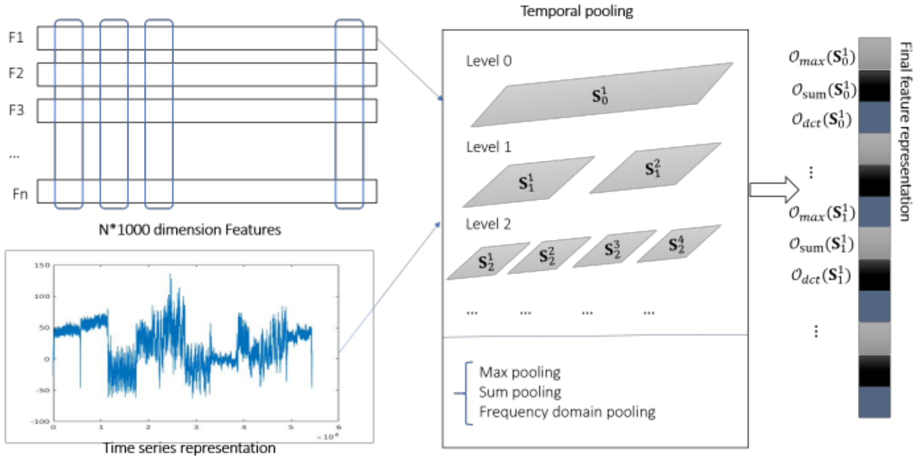


Figure 4. **Temporal Pyramid Pooling.** PCA-compressed feature vectors over time are processed via three different pooling operators and evaluated over a pyramid of varying windows in order to generate the concatenated components of the feature vector used by classifiers for activity type determination and by regressors for calorific uptake estimation.

4. Experimental and Results

4.1. Experimental Settings

Dataset - To evaluate the method, we test the proposed method on the *SPHERE-calorie dataset* [27]. The dataset contains colour images and the corresponding calorie expenditure values. The ground truth was captured by the COSMED K4b2 portable calorimeter, which provides reference calorie readings. The videos were recorded over 20 sessions by 10 subjects in a daily living environment using the Kinect. The total recording time is more than 10 hours. The videos and the calorie values were synchronised for the experiments. The dataset contains up to 11 activity categories per session. All the activities considered are home-based daily activities, i.e.: 1) standing still, 2) sitting still, 3) walking, 4) wiping table, 5) vacuuming, 6) sweeping floor, 7) lying down, 8) exercising, 9) stretching, 10) cleaning floor stain, and 11) reading. The numbers are corresponding to the indices in the experimental results presented below. These actions were recorded under varying angles, distances and lighting conditions. Figure 5 details the dataset corpus further.

CNN implementation - In our work, we discriminatively trained a deep convolutional neural network to classify action type. We initialize the parameters from Conv1 to fc7 using a standard pre-trained model, namely "bvlc reference caffenet [15]". During the training procedure, the parameters are learned using stochastic gradient descent with momentum. We set momentum to be 0.9 and weight decay to be 5×10^{-4}. The training uses a batch size of 256 for all training sets.

Evaluation settings - In our experiments, we use linear SVMs for activity classification and a linear support vector regressor for energy expenditure prediction. The SVM is implemented using a Liblinear library [11]. Liblinear can cope well with large-scale datasets, and also significantly improves the performance and efficiency of program execution. A grid search algorithm is performed to estimate the hyper-parameters of the SVM. For testing, we cross-validate the method using "one left out subject". The process iterates through all subjects, and the average testing error of all iterations are reported.

We use the normalised root mean squared error (Normalised RMSE) as a standard evaluation metric for the deviation of predicted calorie values from the ground truth calorie expenditure.

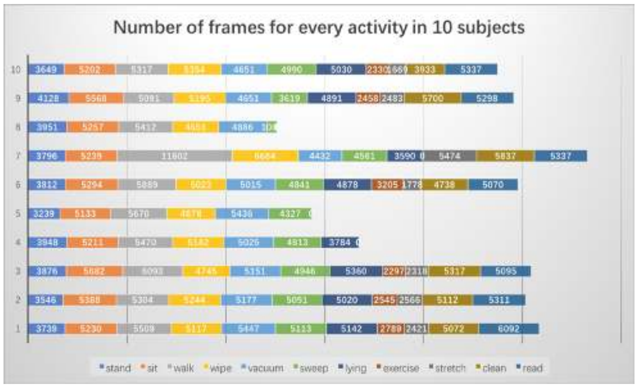
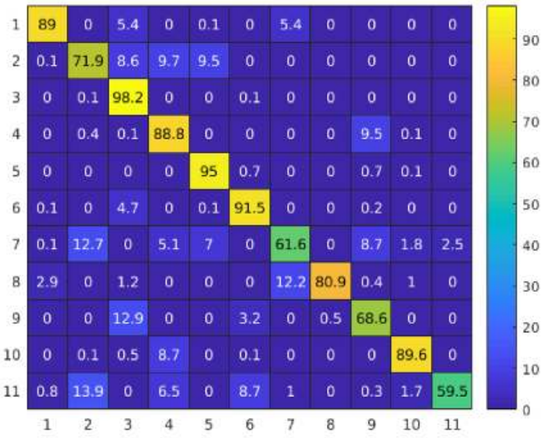


Figure 5. **Dataset Overview.** The graphic illustrates the number of frames per sequence of each activity in the dataset.



	1	2	3	4	5	6	7	8	9	10	11
1	89	0	5.4	0	0.1	0	5.4	0	0	0	0
2	0.1	71.9	8.6	9.7	9.5	0	0	0	0	0	0
3	0	0.1	98.2	0	0	0.1	0	0	0	0	0
4	0	0.4	0.1	88.8	0	0	0	0	9.5	0.1	0
5	0	0	0	0	95	0.7	0	0	0.7	0.1	0
6	0.1	0	4.7	0	0.1	91.5	0	0	0.2	0	0
7	0.1	12.7	0	5.1	7	0	61.6	0	8.7	1.8	2.5
8	2.9	0	1.2	0	0	0	12.2	80.9	0.4	1	0
9	0	0	12.9	0	0	3.2	0	0.5	68.6	0	0
10	0	0.1	0.5	8.7	0	0.1	0	0	0	89.6	0
11	0.8	13.9	0	6.5	0	8.7	1	0	0.3	1.7	59.5

Figure 6. **Activity Recognition Performance.** The confusion matrix shows results using the proposed method with window size $w = 15s$.

4.2. Action Recognition

As the proposed method is built upon the baseline method in [27], where hand-crafted features are extracted from images instead of using CNN-based features, we approach our experiments with a view to compare against it. In the first experiment, we look at the performance of the two approaches w.r.t. correct activity type classification rate considering two different temporal window lengths for encoding the temporal information - 15 seconds and 60 seconds, i.e. $w = \{15, 60\}$.

As illustrated in Table 1, in general, the better performance is achieved when a shorter window is applied, e.g. $w = 15$. This is particularly evident for the high physical intensity level activities, such as exercising and stretching. In such cases, activities are likely to be better explained within a relatively small temporal segment, for which local temporal information are more meaningful.

The table also shows that the our method significantly outperforms the baseline method by 9% irrespective of the window sizes. The proposed method performs especially well for individually highly variable activity types, such as exercising, where the accuracy has increased by 41.7%.

Figure 6 shows the recognition confusion matrices from the activity recognition results of our method with $w = 15s$.

4.3. Calorie Expenditure Prediction

Similarly to the results on action recognition, the accuracy of calorie expenditure prediction is also related to the choice of the temporal window length. Figure 7 shows the results for the second experiment where we compare our method for calorie expenditure prediction with the baseline method with window sizes $w = \{15, 60\}$. The proposed method produces fewer errors for the majority of activities. In contrast to activity recognition, calorie expenditure estimation tends to perform better when larger window sizes are applied. This can be explained as human body adaptation causes an exponential change to a plateau in oxygen consumption until it reaches a steady state, thus retaining a relatively long history will help with calorie expenditure prediction.

One may argue that the calorie value is influenced by the performance of action recognition. To test this, we consider the effect of varying action recognition quality. We first test a system in which the ground truth labels are given to select the activity specific model for calorie prediction (see top row in Table 2), and then compare the action recognition at window sizes $w = \{15, 60\}$ seconds. The window length for calorific expenditure estimation is fixed at $w = 60$ seconds. The results show that for most activities the calorie prediction error is the smallest when there is a small activity recognition error only. This indicates that higher action recognition accuracy may indeed help calorie prediction.

To analyse the performance further, we select the model with the best window configurations, that is using window size 15 seconds for activity recognition and 60 seconds for calorie expenditure prediction. Note that this is the best configuration for both our method and the baseline method in [27]. Figure 8 shows the average calorie prediction errors for both methods when ground truth labels are used to select the activity-specific model. In general, the proposed method clearly produces smaller errors than [27]. This is particularly prominent in certain activities, such as sweeping, lying down, wiping table, vacuuming, standing still and sitting still. The proposed method reduces the overall error from 0.44 to 0.41 when 15s window size is employed and from 0.43 to 0.39 when ground truth labels are used.

5. Conclusion and Future work

In this work, we studied the problem of calorie expenditure estimation and introduced a new feature representation designed for the problem. The proposed feature representation captures the data dynamics over a time interval by capturing both global and local changes in high dimensional feature descriptors. The per-frame descriptor is formed based on deep CNN features. The method is evaluated on a public calorie expenditure estimation video dataset, and results show that the proposed method outperforms the previous baseline. Potential future directions include an extension of the presented approach towards full end-to-end calorie value estimation within a single deep learning approach with novel network architectures.

6. Acknowledgements

This work was performed in the SPHERE IRC project funded by the UK Engineering and Physical Sciences Re-

w	**method**	*stand*	*sit*	*walk*	*wipe*	*vacuum*	*sweep*	*lying*	*exercise*	*stretch*	*clean*	*read*	**overall**
15	Baseline	86.5	**77.6**	88.3	69.4	79.0	76.5	**62.3**	39.2	61.1	**91.4**	38.9	73.7
	Ours	**89.0**	71.9	**98.2**	**88.8**	**95.0**	**91.5**	61.6	**80.9**	**68.6**	89.6	**59.6**	**82.3**
60	Baseline	81.1	**79.7**	85.1	66.0	77.2	72.9	33.0	29.3	52.7	90.0	35.9	68.2
	Ours	**81.9**	70.1	**96.1**	**80.9**	**82.1**	**85.0**	**57.6**	**86.0**	**56.8**	**89.3**	**46.2**	**77.8**

Table 1. Activity recognition rate (%) with different method and with different window length. The best results in each activity are in bold.

calorie w	action w	*stand*	*sit*	*walk*	*wipe*	*vacuum*	*sweep*	*lying*	*exercise*	*stretch*	*clean*	*read*	**overall**
60	n/a	**0.33**	**0.28**	0.28	**0.22**	**0.13**	**0.31**	0.45	**0.37**	0.84	**0.21**	**0.40**	**0.39**
	15	0.27	0.38	**0.22**	0.29	0.20	0.33	**0.25**	0.46	**0.55**	0.42	0.49	0.41
	60	0.39	0.35	0.36	0.39	0.33	0.30	0.38	0.46	0.77	0.34	0.46	0.42

Table 2. Calorific expenditure prediction error (normalised RMSE) when ground truth labels are used to select the activity-specific model (top row), and when action recognition is employed at different window lengths.

search Council (EPSRC), Grant EP/K031910/1.

References

[1] Cosmed K4b2. `http://www.cosmed.com/`.

[2] SPHERE calorie dataset. `http://doi.org/cc5k`.

[3] J. Aggarwal and M. Ryoo. Human activity analysis: A review. *ACM Computing Surveys*, 43(3):16, 2011.

[4] J. Aggarwal and L. Xia. Human activity recognition from 3D data: A review. *Pattern Recognition Letters*, 48:70–80, 2014.

[5] B. Ainsworth et al. Compendium of physical activities: an update of activity codes and met intensities. *Medicine and science in sports and exercise*, 32(9):498–504, 2000.

[6] M. Altini et al. Estimating oxygen uptake during nonsteady-state activities and transitions using wearable sensors. *IEEE journal of biomedical and health informatics*, 20(2):469–475, 2016.

[7] A. A. Chaaraoui et al. A review on vision techniques applied to human behaviour analysis for ambient-assisted living. *Expert Systems with Applications*, 39(12):10873–10888, 2012.

[8] C. Chen, R. Jafari, and N. Kehtarnavaz. Improving human action recognition using fusion of depth camera and inertial sensors. *IEEE Transactions on Human-Machine Systems*, 45(1):51–61, 2015.

[9] B. Crabbe et al. Skeleton-free body pose estimation from depth images for movement analysis. In *Proceedings of the IEEE ICCV Workshops*, pages 70–78, 2015.

[10] A. Edgcomb and F. Vahid. Estimating daily energy expenditure from video for assistive monitoring. In *International Conference on Healthcare Informatics*, pages 184–191. IEEE, 2013.

[11] R.-E. Fan et al. Liblinear: A library for large linear classification. *Journal of machine learning research*, 9(Aug):1871–1874, 2008.

[12] J. Hall et al. Designing a video monitoring system for aal applications: The sphere case study. 2016.

[13] S. Herath et al. Going deeper into action recognition: A survey. *Image and Vision Computing*, 60:4–21, 2017.

[14] S. Ji et al. 3d convolutional neural networks for human action recognition. *IEEE transactions on PAMI*, 35(1):221–231, 2013.

[15] Y. Jia et al. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the ACM international conference on Multimedia*, pages 675–678. ACM, 2014.

[16] A. Krizhevsky et al. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[17] I. Laptev. On space-time interest points. *IJCV*, 64(2-3):107–123, 2005.

[18] B. Lin et al. Database-oriented storage based on lmdb and linear octree for massive block model. *Transactions of Nonferrous Metals Society of China*, 26(9):2462–2468, 2016.

[19] S. Matsuo and K. Yanai. Cnn-based style vector for style image retrieval. In *Proceedings of the ACM on International Conference on Multimedia Retrieval*, pages 309–312. ACM, 2016.

[20] OpenNI organization. *OpenNI User Guide*, November 2010.

[21] F. J. Ordóñez and D. Roggen. Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. *Sensors*, 16(1):115, 2016.

[22] R. Planinc et al. Computer vision for active and assisted living. *Active and Assisted Living: Technologies and Applications*, page 57, 2016.

[23] L. L. Presti and M. La Cascia. 3d skeleton-based human action classification: A survey. *Pattern Recognition*, 2016.

[24] E. Ravussin et al. Determinants of 24-hour energy expenditure in man. methods and results using a respiratory chamber. *Journal of Clinical Investigation*, 78(6):1568, 1986.

[25] J. J. Reilly et al. Objective measurement of physical activity and sedentary behaviour: review with new data. *Archives of disease in childhood*, 93(7):614–619, 2008.

[26] L. Tao et al. A comparative home activity monitoring study using visual and inertial sensors. In *International Conference on E-health Networking, Application & Services*, pages 644–647. IEEE, 2015.

[27] L. Tao et al. Calorie counter: Rgb-depth visual estimation of energy expenditure at home. In *ACCV*, pages 239–251. Springer, 2016.

[28] L. Tao et al. Real-time estimation of physical activity intensity for daily living. 2016.

[29] L. Tao et al. Energy expenditure estimation using visual and inertial sensors. *IET Computer Vision*, 2017.

[30] A. Toshev and C. Szegedy. Deeppose: Human pose estimation via deep neural networks. In *CVPR*, pages 1653–1660, 2014.

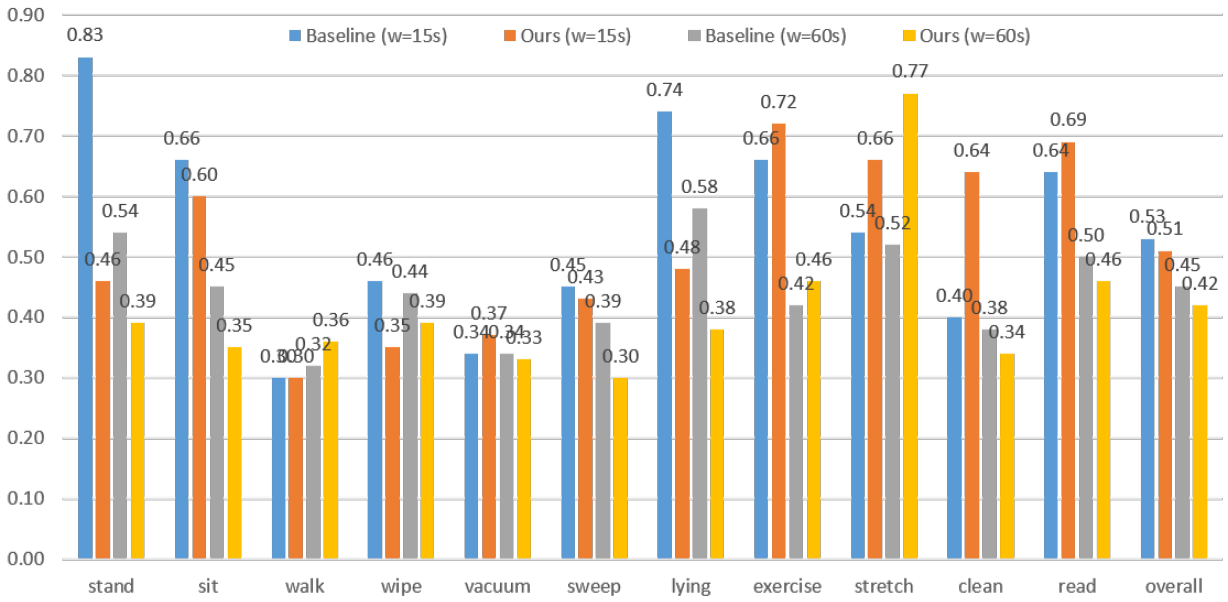


Figure 7. **Average Calorie Prediction Errors.** Errors are shown as normalised RMSE of the proposed method and the baseline method using window sizes 15 and 60 seconds, respectively.

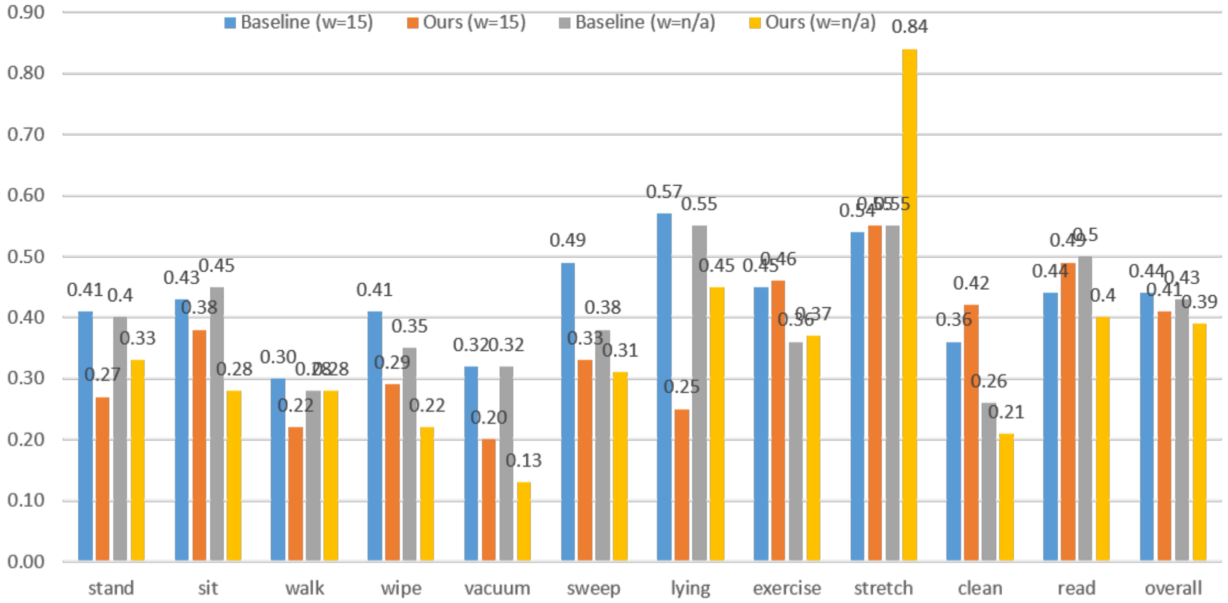


Figure 8. **Average Calorie Prediction Errors.** Errors are shown as normalised RMSE of the proposed method and the baseline method when ground truth labels are used to select the activity-specific model (w=n/a), and when action recognition is employed at window size 15s (w=15).

[31] P.-F. Tsou and C.-C. Wu. Estimation of calories consumption for aerobics using kinect based skeleton tracking. In *Systems, Man, and Cybernetics,International Conference on*, pages 1221–1226. IEEE, 2015.

[32] J. Zhu et al. Using deep learning for energy expenditure estimation with wearable sensors. In *International Conference on E-health Networking, Application & Services*, pages 501–506. IEEE, 2015.

[33] N. Zhu et al. Bridging ehealth and the internet of things: The SPHERE project. *IEEE Intelligent Systems*, 2015.