Should I take a walk? Estimating Energy Expenditure from Video Data

Kunyu Peng*, Alina Roitberg*, Kailun Yang, Jiaming Zhang, and Rainer Stiefelhagen

* denotes equal contribution
Karlsruhe Institute of Technology
firstname.lastname@kit.edu

Abstract

We explore the problem of automatically inferring the amount of kilocalories used by human during physical activity from his/her video observation. To study this under-researched task, we introduce Vid2Burn – an omni-source benchmark for estimating caloric expenditure from video data featuring both, high- and low-intensity activities for which we derive energy expenditure annotations based on models established in medical literature.

In practice, a training set would only cover a certain amount of activity types, and it is important to validate, if the model indeed captures the essence of energy expenditure, (e.g., how many and which muscles are involved and how intense they work) instead of memorizing fixed values of specific activity categories seen during training. Ideally, the models should look beyond such category-specific biases and regress the caloric cost in videos depicting activity categories not explicitly present during training. With this property in mind, Vid2Burn is accompanied with a cross-category benchmark, where the task is to regress caloric expenditure for types of physical activities not present during training. An extensive evaluation of state-of-the-art approaches for video recognition modified for the energy expenditure estimation task demonstrates the difficulty of this problem, especially for new activity types at test-time, marking a new research direction. Dataset and code are available at https://github.com/KPeng9510/Vid2Burn1.

1. Introduction

If you would ask people to honestly answer “Why do you go to the gym?” a frequent reply would be to burn calories. Physical activity is connected with our health and is an important element in prevention of obesity, diabetes or high blood pressure2 – issues which are amplified through the recent Covid-19 lockdowns and the home office regulations [3]. With the rise of health tracking apps, automatic inference of energy expenditure is rapidly gaining attention [2, 5, 36, 39, 59, 70], but almost all prior research has focused on signals obtained from wearable devices, such as smart watches or heart rate monitoring chest straps. While such sensors are not always present at hand or comfortable to wear, most people can easily access a video camera in their phone or laptop. Apart from helping the users interested in tracking their exercise and maintaining active lifestyle, recent studies in gerontology highlight the benefits of automatically tracking the level of physical activity in assistive smart homes in order to support the elderly [4, 22, 45].

As important as it is for our health, understanding physical activity offers new technical challenges in computer vision. Excellent progress has been made in the field of human activity recognition [6, 14, 44, 56, 65] with remarkable accuracies reported on datasets such as HMDB-51 [29] or Kinetics [6]. However, when facing our task of estimating

1Acknowledgments: The research leading to these results was supported by the SmartAge project sponsored by the Carl Zeiss Stiftung (P2019-01-003; 2021-2026).

caloric expenditure from human observations, these methods will face two main obstacles. First, the cornerstone of past research lies in rather rigid categorization into predefined actions. These categories are often relatively coarse, (e.g., “football” vs. “jogging”), so that the scene context provides the network with an excellent shortcut to the decision, leaving the actual moving person behind [10, 62]. Our task however requires fine-grained understanding of human movement, as medical research [7] lists which muscles are active and how hard they work as the main drivers of energy expenditure (although a multitude of further factors influence this complex physiological process). A second key challenge is to encourage the model to capture the essence of energy expenditure instead of memorizing average values of specific activity categories seen during training. Deep neural networks are prone to learning shortcuts [10, 16, 19] and internally casting the calorie regression problem as an “easier” task of activity categorization which might be one of such potential shortcuts. Even if the annotations are continuous caloric values and not rigid categories, in practice, the training set can only cover a finite amount of activity types. Ideally, our model should not be bounded to category-specific biases and indeed learn the nature of activity-induced energy expenditure by, e.g., understanding the type and intensity of bodily movement produced by the skeletal muscles. When developing an energy expenditure benchmark, it is therefore critical to evaluate the results on types of physical activity not present during training.

In this paper, we explore the new research direction of inferring activity-induced caloric cost by observing the human in video as Fig. 1. To tackle the lack of public large-scale datasets, we introduce Vid2Burn - a new open-source benchmark spanning 9789 video examples of people engaged in different activities with corresponding annotations designed based on models established in medical literature [27] from (1) current activity category (2) intensity of the skeleton movement and (3) heart rate measurements obtained for a subset of activities (household activities) in a complementary study. Videos in the dataset are chosen from four diverse activity recognition datasets [25, 29, 50, 53] originally from YouTube, movies or explicitly designed for recognition in household context.

Yet, a key challenge when applying energy expenditure models in practice arises from transferring the learned concepts to new activity types. To meet this requirement, Vid2Burn is equipped with a cross-category benchmark, where the caloric cost estimation models are evaluated against activity types not seen during training. In addition to potential mobile health applications, our dataset therefore fills the lack of benchmark studying concise recognition of body movement without relying on category-specific context biases. From the computer vision perspective, the key technical challenges of our benchmark are (1) fine-grained understanding of bodily movement and (2) generalization to previously unseen types of activities. Extensive experiments with multiple state-of-the-art approaches for video- and body pose based action recognition demonstrate the difficulty of our task using modern video classification architectures, highlighting the need for future research.

2. Related work

Activity recognition in videos. Human activity recognition often operates on body poses [9, 31, 33, 51, 67] or learns representations end-to-end directly from the video data using Convolutional Neural Networks (CNNs) [6, 17, 20, 52, 68]. CNN-based approaches often deal with the temporal dimension via 3D convolution [44, 56, 57, 61, 65] or follow the 2D+1D paradigm, chaining spatial 2D convolutions and subsequent 1D modules to aggregate the features temporally [14, 15, 24, 32, 60, 71]. Fueled by multiple publicly released large-scale activity recognition datasets collected from YouTube/Movies [25, 29, 53] or in home environments [50], the research of deep learning based activity recognition became a very active research field also explored in more targeted applications, e.g., in cooking [11, 46], sports [42], robotics [23, 49], and automated driving [35]-related tasks. More specialized activity recognition research also addressed topics such as uncertainty of video classification models [47, 54]. However, all the approaches focus on categorization into previously defined activity classes, while examining their feasibility for capturing complex physiological processes of the body, such as our calorie expenditure task, has been largely overlooked.

Energy expenditure prediction. Visual estimation of caloric values has been mainly investigated in food image analysis (i.e., tracking the amount of caloric intake) [34, 40, 48]. Energy expenditure induced by physical activity is mostly studied from an egocentric perspective featuring data from wearables, such as accelerometers or heart rate monitors [2, 5, 18, 26, 37, 39, 41, 55, 64], with a recent survey provided in [70]. Only very few works address the visual predicted activity-related caloric expenditure [39, 55]. The only dataset collected for energy expenditure prediction by

### Table 1

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Vid2Burn-Diverse</th>
<th>Vid2Burn-ADL</th>
<th>Vid2Burn</th>
</tr>
</thead>
<tbody>
<tr>
<td>Video origin</td>
<td>Youtube/movie datasets</td>
<td>ADL datasets</td>
<td>Youtube/ Movie/ADL</td>
</tr>
<tr>
<td>#Clips</td>
<td>4260</td>
<td>5529</td>
<td>9789</td>
</tr>
<tr>
<td>#Activities</td>
<td>33</td>
<td>39</td>
<td>72</td>
</tr>
<tr>
<td>#Train/test categories</td>
<td>27/6</td>
<td>33/6</td>
<td>60/12</td>
</tr>
<tr>
<td>Unit</td>
<td>kcal/hour</td>
<td>kcal/hour</td>
<td>kcal/hour</td>
</tr>
<tr>
<td>Min</td>
<td>64</td>
<td>153</td>
<td>64</td>
</tr>
<tr>
<td>Max</td>
<td>961</td>
<td>449</td>
<td>961</td>
</tr>
<tr>
<td>Mean</td>
<td>373</td>
<td>276</td>
<td>318</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Video origin</th>
<th>Youtube/movie datasets</th>
<th>ADL datasets</th>
<th>Youtube/ Movie/ADL</th>
</tr>
</thead>
<tbody>
<tr>
<td>#Clips</td>
<td>4260</td>
<td>5529</td>
<td>9789</td>
</tr>
<tr>
<td>#Activities</td>
<td>33</td>
<td>39</td>
<td>72</td>
</tr>
<tr>
<td>#Train/test categories</td>
<td>27/6</td>
<td>33/6</td>
<td>60/12</td>
</tr>
<tr>
<td>Unit</td>
<td>kcal/hour</td>
<td>kcal/hour</td>
<td>kcal/hour</td>
</tr>
<tr>
<td>Min</td>
<td>64</td>
<td>153</td>
<td>64</td>
</tr>
<tr>
<td>Max</td>
<td>961</td>
<td>449</td>
<td>961</td>
</tr>
<tr>
<td>Mean</td>
<td>373</td>
<td>276</td>
<td>318</td>
</tr>
</tbody>
</table>

Table 1. An overview of the main properties of Vid2Burn and its two versions (including statistics of the caloric annotations).
Figure 2. An overview of the dataset statistics. The statistics of the caloric cost values are summarized as pie charts in (a) and (b) for Vid2BurnDiverse and Vid2BurnADL, respectively. The caloric cost annotation statistics grouped by the individual activity categories are provided in (c), where blue bars represent the category-wise annotations, orange bars stand for the average of the sample-wise values and the bar index on the vertical axis indicates the action ID (which correspond to the order of activities listed below the blue histogram). The sample frequencies for different action categories in Vid2BurnDiverse are visualized in the blue bar chart (d) with multiple visual examples.


Given the growing demand for eHealth apps\(^3\), it is surprising that there is not a larger body of work on estimating physical intensity of activities in videos. This might be due to the general focus of video classification research evolving mostly around activity categorization [6, 8, 13, 15, 17], while virtually all exercise intensity assessment datasets focus on wearable sensors [2, 5, 39] delivering, e.g., heart rate or accelerometer signals. To promote the task of visually estimating the hourly amount of kilocalories burned by the human during the current activity, we introduce the novel Vid2Burn dataset, featuring >9K videos of 72 different activity types with both caloric expenditure annotations on category- and sample-level.

3.1. Dataset Collection

Vid2Burn is an omni-source dataset developed with a diverse range of movements and settings in mind. Our data collection procedure comprised the following steps. We started by surveying the well-known available datasets for categorical activity classification, (e.g., [25, 29, 50, 53]). Then, we identified categories which are not only accessible from these public datasets but also have great technically feasibility to infer caloric cost annotations. The main sources of our dataset are UCF-101 [53], HMDB51 [29], test set of Kinetics [25] and NTU-RGBD [50]. We manually identified 72 activity types for which the hourly caloric cost can be estimated based on the established physiological models, (e.g., [1, 27, 58]). Then, we estimated the labels for the energy expenditure based on these models on the

\(^3\)Grand View Research. mHealth Apps Market Size, Share & Trends Analysis Report By Type (Fitness, Medical), By Region (North America, APAC, Europe, MEA, Latin America), And Segment Forecasts, 2021 - 2028. 2021. Available from: www.grandviewresearch.com/industry-analysis/mhealth-app-market.
category- and sample-level described in Section 3.3.

3.2. Dataset Structure

The benefits of understanding caloric cost from videos extend to many applications, such as tracking of active exercise routines [21] or monitoring the daily physical activity level for elderly care [4, 22, 45]. From the technical perspective, it is also useful to distinguish settings with higher and lower differences between the samples. Lastly, while it is feasible to derive proper ground-truth for coarse behaviours or situations with well-studied energy expenditure, (e.g., types of sports and exercises), many daily living activities do not fall into this category and should be addressed with different techniques. Motivated by this, we group the content of Vid2Burn in two subsets: Vid2Burn_Diverse and Vid2Burn_ADL. Table 1 gives an overview of Vid2Burn and both its variants.

**Vid2Burn-Diverse** is collected from Youtube- and movie-based sources [25, 29, 53] and therefore features a highly uncontrolled environment (camera movement, diverse inside/outside backgrounds). Since we focused on activities with well-studied energy expenditure models, a large portion of behaviours are related to sports, (e.g., PushUps). However, the database also covers certain everyday activities, such as walking, standing or eating. The distribution of different activity types is summarized in Figure 2. On average, the dataset features 129 video clips per category using category labels inherited from the original sources, with walking being unsurprisingly the most common behaviour for 548 videos while stretching and shopping are the least frequent ones for 47 videos.

**Vid2Burn-ADL** on the other hand targets Activities of Daily Living (ADL) and might be used for physical workload tracking in smart homes. The activity types and video examples are derived from the public NTU-RGBD [50] dataset for ADL classification and, compared to Vid2Burn_Diverse, this dataset contains activities of rather lower physical intensities (e.g., pickup, take off jacket, read, drink water). The environment of Vid2Burn_ADL is much more controlled and the differences between the individual samples are at smaller scale. In other words, Vid2Burn_ADL can be regarded as a much more fine-grained benchmark for caloric cost regression. In contrast to Vid2Burn_Diverse, the categories of Vid2Burn_ADL are rather well-balanced and the number of examples per activity type is 142 on average (detailed frequency statistics provided in the supplementary).

3.3. Caloric Expenditure Annotations

To adequately represent activity-induced caloric expenditure, we conducted a literature review on this physiological process [1, 37, 43, 58, 63]. Tracking the heat expended by nutrients oxidation (i.e., monitoring oxygen intake and carbon dioxide production) is considered the most accurate way of estimating energy expenditure [43]. While this method is invasive and not practical for large-scale use, a multitude of topical studies conduct and publish such measurements for specific groups of activities, which is often summarized by meta-reviews in the form of compendiums [1]. Such catalogues provide energy expenditure values for specific activities that are often available online as look-up tables. A common way of estimating caloric cost, which is also leveraged by us, is deriving it from the heart-rate with validated physiological models [27, 37]. Our annotation scheme leverages three methods to estimate hourly energy expenditure: established medical compendiums [1], heart-rate based measurements [27] as well as adjustments based on the captured body movement [58].

Next, we describe the three different ways for obtaining caloric cost ground truth (Section 3.3.1) and explain how we leveraged them to annotate the Vid2Burn_Diverse (Section 3.3.2) and the Vid2Burn_ADL (Section 3.3.3) datasets.

### 3.3.1 Annotation Methods

Our derived annotation scheme leverages three types of sources: (1) current activity category, (2) intensity of the skeleton movement, as well as, (3) heart rate measurements obtained for a subset of activities (household activities) in a complementary study.

**Caloric cost values from published compendiums.** First, we leverage activity-specific metabolic rate values from published compendiums [1] often summarized as look-up tables available on the web.\(^1\) For simplicity, we assume the body weight as 150 lb, since this is also the average
\(^1\)https://captaincalculator.com/health/calorie/
body weight of subjects captured in our heart rate measurement study. Examples of the category-wise caloric expenditure annotations are marked as stars in Figure 3 for the Vid2BurnDiverse dataset.

Heart-rate based annotations. Vid2BurnADL focuses on daily living activities, which naturally exhibit lower intensity of movement. The differences are at a much smaller scale compared to Vid2BurnDiverse and the average expected energy expenditure has not been well-studied for many such concise types of physical activity. However, due to the more restricted nature of the environment, the NTU-RGBD setting is easy to reproduce. In such cases, we recreate the environment of 39 activities of Vid2BurnADL and estimate their average caloric cost based on heart rate measurements captured in a study with four volunteer participants. Four people, one female and three males, participated in the data collection (1 female, 3 male, 27.75 years old on average, average weight 150 lb). The participation in our study was voluntary, and the subjects were instructed about the scope and purpose of the data collection and have given their written consent according to the requirements of our institution.

The heart rate of all participants was recorded using a wrist band activity tracker. The subjects were asked to execute 39 activities of the Vid2BurnADL with a resting period in between to ensure the heart rate recovery. More information about the study setup is provided in the supplementary.

Given the measured heart rate, we compute the caloric cost of the activity in accordance to [27] as,

\[
\text{Cal}_M = 60 \times T \times ((-55.0969) + (0.6309 \times HR) + (0.1988 \times W) + (0.2017 \times A))/4.184,
\]

\[
\text{Cal}_F = 60 \times T \times ((-20.4022) + (0.4472 \times HR) - (0.1263 \times W) + (0.074 \times A))/4.184,
\]

where \(\text{Cal}_M/\text{Cal}_F\) indicates the hourly caloric expenditure for male/female, \(HR\) is heart rate, \(W\) is the body weight, \(A\) indicates participant’s age and \(T\) is the time length in hour.

Body movement-based annotations. Next, we approximate the caloric cost induced by the movement by leveraging the model of Tsou et al. [58]. People can engage in the same type of activities in different ways and, since the amount of calories burned is directly linked to the amount/types of active muscles and the intensity, more active bodily movements lead to higher caloric cost. Tsou et al. [58] formalizes and validates a model corresponding based on the movement of eight body regions \(r\). We estimate the skeleton movement using AlphaPose [12, 30, 66] for Vid2BurnDiverse, while for Vid2BurnADL we use the skeleton data provided by the authors of the original datasets [38, 50]. Following [58], we group skeleton joints into eight regions-of-interest to approximate the energy consumption as:

\[
E_{\text{body}} = \sum_{t=1}^{F} \sum_{r=1}^{8} \omega_r \lambda_r \left( \frac{1}{2} M_r \lambda \Delta x_{rt}^2 + \frac{1}{2} M_r \lambda \Delta y_{rt}^2 + \frac{1}{2} M_r \lambda \Delta z_{rt}^2 \right)
\]

where \(\Delta x_{rt}, \Delta y_{rt}, \Delta z_{rt}\) indicate the position difference of the body region between frame \(t\) and \(t-1\) (the average position of all body joints inside same region), \(F\) indicates the frame number and \(\lambda\) is the frame frequency, \(M_r\) indicates the mass of \(r\)th body region. The final calorie consumption is obtained via multiplication with \(0.239(Cal/J) \times F \times \lambda \times 3600(h/s)\) in per hour-wise expression. The weighting factors \(\omega_r\) of the different body regions are obtained from [58].

The main purpose of the body-pose based caloric cost estimation is to enable more concise annotations at sample-level, since the same activity can be executed with different intensities.

3.3.2 Vid2Burn-Diverse annotations

One strategy behind the design of Vid2BurnDiverse was to select behaviour types for which the average caloric costs have been well-studied and easily accessible [1] (for that reason, sports-related videos constitute a significant portion of Vid2BurnDiverse). The main source for the category-level values is therefore derived from the published average category-specific which have been well-studied and easily accessible in this regard (on category level). We then correct the estimations of the individual videos based on the previously explained body-movement model [58], resulting in more concise sample-level annotations.

3.3.3 Vid2Burn-ADL annotations

Energy expenditure is not well-studied for many of the more concise daily living situations in Vid2BurnADL. We therefore take a detour by conducting a study with participants’ heart rate recorded during the 39 target activities (as described in the heart rate paragraph of Section 3.3.1). For each activity type, we estimate (1) the average heart rate-based caloric cost value obtained from our study, (2) the average values based on the skeleton movement captured in the Vid2BurnADL videos and (3) if available, values established from published medical compendiums [1]. Note, that while for Vid2BurnDiverse such estimations from published compendiums represent the main ground truth source, which are only available for 4 out of 39 Vid2BurnADL activities. The final category-level annotations are then computed as the average of the estimations delivered by the two/three aforementioned methods. Similar to Vid2BurnDiverse, we then derive more accurate sample-level annotations through previously described skeleton-based corrections.
### 3.4. Same- and intra-activity splits

Since we specifically aim to rate generalization of the caloric estimation models for new activity types, we construct two testing scenarios covering (1) **Known activity types** evaluation, with videos covering the same behaviours as the training set and (2) **unknown activity types** evaluation where the train and test samples are drawn from different activity types. We randomly select 27 (Vid2Burn\textsubscript{Diverse}) and 33 (Vid2Burn\textsubscript{ADL}) activity types for the training set, while for both benchmark versions, the 6 remaining categories are used for evaluation. Next, the data of the 27/33 training activity types is further split into training/testing (with ratio 7:3) for the same-category evaluation. Note that the category annotations used for constructing the splits are inherited from the source datasets. Overall, our dataset comprises 2782/3243 videos for training, 1192/1390 samples for the validation on the same activity type and 286/886 samples for new-activity-type evaluation for the Vid2Burn\textsubscript{Diverse} and Vid2Burn\textsubscript{ADL} databases, respectively.

### 4. Activity Recognition Models in the Context of Caloric Cost Estimation

#### 4.1. Learning continuous caloric values

Given a video input, our goal is to infer hourly energy cost of the activity in which the depicted human is involved. Note, that we target the *intensity* of the bodily activity and not its duration, *i.e.* our goal is to infer kilocalories burnt per hour. Since our targets are *continuous* caloric values our task naturally suits regression-based losses, such as the Euclidean L2 loss. However, we observed that regression optimisation converges to a constant value in our case (a similar effect has been reported before in multimodal problems, *e.g.*, in [69]). We therefore address this problem as multinomial classification with additional label softening. Similar to [54], we binarize each caloric value annotation $l$ with resolution of 1 kcal inside a range $n \in [0,N]$, where $N$ is set to 1000 kcal. To keep certain regression properties, such as penalizing predictions which fall closer to the ground-truth bin less, we soften the labels through Gaussian distribution with a given standard deviation (STD) denoted as $\sigma$. Then, for each ground truth annotation $l$, we obtain the softened label as a distribution over $N$ bins:

$$l_s[n] = \frac{1}{\sigma \sqrt{2\pi}} \exp\left(-\frac{(l-n)^2}{2\sigma^2}\right),$$

where $l_s$ indicates the soft label used for the supervision. We then use the the Kullback-Leibler (KL) divergence between the ground truth and predicted distributions:

$$KL_{\text{class}}(y, l_s) = \frac{1}{N} \sum_{n=0}^{N-1} (l_s[n]\log(l_s[n]) - y[n]),$$

where $y$ indicates the predicted distribution with $n \in [0,N]$.

### Table 3. Prediction quality for individual activity types on Vid2Burn\textsubscript{Diverse} for (two known and two new categories).

<table>
<thead>
<tr>
<th>Method</th>
<th>Known activity types</th>
<th>New activity types</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MAE ↑</td>
<td>SPC ↓</td>
</tr>
<tr>
<td>ST-GCN</td>
<td>13.51</td>
<td>310.9</td>
</tr>
<tr>
<td>SF-AVR</td>
<td>11.76</td>
<td>43.8</td>
</tr>
<tr>
<td>R3D-AVR</td>
<td>6.14</td>
<td>302.3</td>
</tr>
<tr>
<td>SF-LSTM</td>
<td>7.69</td>
<td>62.6</td>
</tr>
<tr>
<td>R(2+1)D-AVR</td>
<td>12.88</td>
<td>288.0</td>
</tr>
<tr>
<td>R(2+1)D-LSTM</td>
<td>12.59</td>
<td>96.6</td>
</tr>
<tr>
<td>SF-AVR</td>
<td>6.54</td>
<td>262.6</td>
</tr>
<tr>
<td>ST-GCN</td>
<td>6.09</td>
<td>147.7</td>
</tr>
<tr>
<td>R(2+1)D-LSTM</td>
<td>11.22</td>
<td>37.1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>New activity types</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MAE ↑</td>
</tr>
<tr>
<td>ST-GCN</td>
<td>19.65</td>
</tr>
</tbody>
</table>
4.3. Temporal fusion

Temporal windows captured by the above backbones vary between 16 frames for I3D [6], R(2+1) [57] and R3D [56] and 32 frames for SlowFast [15] are considerably smaller than the durations of the video clips captured in Vid2Burn. Given an input video of length $T$ and a model $f_θ(·)$ which takes as input $F$ frames, we sequentially pass $K$ video snippets $\{t_1, t_2, \ldots, t_K\}$ using sliding window with overlapping resulting in $K$ predictions $\{f_θ(t_1), f_θ(t_2), \ldots, f_θ(t_K)\}$. We now consider two different strategies for fusing these results: (1) averaging of the output of the last fully connected layer and (2) learning to fuse the output with an additional LSTM network. The first method employs simple average pooling of the final representation: $\text{pred}(t) = \frac{1}{K} \sum_{i=1}^{K} f_θ(t_i)$. Our second fusion strategy passes the last fully connected layer of our video representation backbone to an LSTM network with two layers with the number of neurons corresponding to the input size, trained together with the backbone model in an end-to-end fashion. Since an LSTM also produces sequential output, we also average the resultant sequence to obtain the final prediction.

5. Experiments

5.1. Evaluation Protocols

We adopt Mean Average Error (MAE) as our main evaluation metric and additionally report the Spearman Rank Correlation (SPC), and the Negative Logarithm Likelihood (NLL). MAE is an intuitive metric reporting the mean disarray between prediction and ground truth in our target units, (i.e., kilocalories). Note, that while SPC illustrates the association strength between the ground truth and the predictions (it is $+1$ if one is a perfect monotone function of the other and $-1$ if they are fully opposed), it should be interpreted with care, since it ignores scaling and shifting of the data. In other words, SPC would not reflect if the number of kilocalories is constantly over/underestimated by similar amounts. It therefore should only be viewed as a complementary metric. Note, that we report SPC in % for better readability (i.e., we multiply the result by 100). Our experiments are carried out in two annotation settings: category- and sample-wise annotations (see Section 3.3 for details). We view the sample-wise version as a more concise choice, but using static category labels is the protocol used in the past energy expenditure work from egocentric data [39] and is adopted for consistency. As explained in Section 3.4, we conduct the evaluations on both: behaviours present during training and new activity types.

5.2. Experiment Results

All implemented models greatly outperform the random and average baselines in all measures (Table 2), showing that our problem is feasible. However, our experiments also underline the difficulty of estimating the energy cost if the model is evaluated in new previously unseen situations, motivating further research of models with deeper and more fine-grained understanding of physical activities.

Tables 2 and 4 illustrate the performance of the employed algorithms as well as the random and average baselines for Vid2BurnDiverse in the category- and sample-labelled settings together with different temporal fusion schemes.
Table 6. Recognition results on the Vid2Burn<sub>ADL</sub> benchmark in the sample-wise caloric annotations setting.

<table>
<thead>
<tr>
<th>Method</th>
<th>Known activity types MAE ↓</th>
<th>SPC ↓</th>
<th>NLL↓</th>
<th>New activity types MAE ↓</th>
<th>SPC ↓</th>
<th>NLL↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>3D-Reg</td>
<td>251.3</td>
<td>-</td>
<td>-</td>
<td>281.7</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Average</td>
<td>76.2</td>
<td>-</td>
<td>-</td>
<td>59.4</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Random</td>
<td>311.7</td>
<td>-</td>
<td>-</td>
<td>309.8</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SF-AVR</td>
<td>20.1</td>
<td>68.6</td>
<td>5.57</td>
<td>36.4</td>
<td>72.30</td>
<td>6.69</td>
</tr>
<tr>
<td>R3D-AVR</td>
<td>35.1</td>
<td>76.91</td>
<td>6.01</td>
<td>44.3</td>
<td>79.78</td>
<td>6.21</td>
</tr>
<tr>
<td>I3D-AVR</td>
<td>22.9</td>
<td>72.97</td>
<td>5.66</td>
<td>39.6</td>
<td>70.45</td>
<td>6.38</td>
</tr>
<tr>
<td>R3+1D-Reg</td>
<td>26.2</td>
<td>71.31</td>
<td>5.99</td>
<td>29.5</td>
<td>75.30</td>
<td>6.28</td>
</tr>
<tr>
<td>SF-LSTM</td>
<td>36.9</td>
<td>77.19</td>
<td>6.05</td>
<td>47.1</td>
<td>80.43</td>
<td>6.24</td>
</tr>
<tr>
<td>R3D-LSTM</td>
<td>39.6</td>
<td>70.45</td>
<td>6.38</td>
<td>51.1</td>
<td>80.72</td>
<td>6.66</td>
</tr>
<tr>
<td>SF-LSTM</td>
<td>36.4</td>
<td>72.30</td>
<td>6.69</td>
<td>48.1</td>
<td>82.30</td>
<td>6.72</td>
</tr>
<tr>
<td>R3+1D-LSTM</td>
<td>88.8</td>
<td>44.79</td>
<td>6.59</td>
<td>72.9</td>
<td>54.97</td>
<td>6.52</td>
</tr>
</tbody>
</table>

Marked as AVR (for the average pooling fusion) and LSTM, with AVR consistently leading notably better results. Slow-Fast (SF) with AVG consistently achieves the best recognition quality with only 34.5 kcal/20.1 kcal MAE. As expected, the results are lower in the more fine-grained sample-wise setting, since the backbones were initially developed for coarser categorization. The task of caloric cost regression in previously unseen situations is much more difficult and the performance drops significantly. The MAE is > 4 and > 2 times higher in the category- and sample-wise settings separately for SF-AVG. Interestingly, the best model in the case of known activities is usually not the top-performing approach for the new activity types, which is SF-AVG in case of the sample-wise evaluation of Vid2Burn<sub>Diverse</sub> with the MAE of 130.8 kcal, but the gap to SF-AVG (MAE of 134 kcal) is very small (Table 4).

Presumably due to a more restricted environment and smaller average caloric cost values, the models are more accurate on Vid2Burn<sub>ADL</sub> (Table 6). Consistently with the Vid2Burn<sub>Diverse</sub> results, SF-AVG yields the best recognition quality on known activities (MAE of 20.1 kcal) and is the second best performing model on the new ones. Table 6 also lists our results achieved with a regression loss (L2), marked as R3D-Reg. As explained in Section 4.1, we observe convergence to a constant value resulting in a very high MAE. Note that SPC cannot be computed in this case, since the output is a number while not a distribution. We also consider the trade-off between the performance and the computational cost (Table 5). SF-AVG offers a good balance between speed and accuracy, while the I3D backbone is a more lightweight model, but the MAE is higher.

We further look at the recognition for the individual activity types: two known (running, climbing) and two unknown (yoga, shopping) behaviours as presented in Table 3 (more results are provided in the supplementary). The recognition quality varies greatly depending on whether the activity is familiar. For example, SF-AVG is only off by 43.8 kcal for running, but the MAE is 174.4 kcal for shopping. Finally, in Figure 4 we showcase multiple qualitative results by visualizing the activation region of a CNN (I3D backbone) with multiple examples of representative qualitative results for the Vid2Burn<sub>Diverse</sub> (top) and Vid2Burn<sub>ADL</sub> (bottom) datasets. Additionally to the predicted caloric value and the ground truth, we visualize the activation regions of an intermediate CNN layer (in this case we choose the second convolutional layer). It is evident, that the largest focus is put on the body region activated during movement which we view as a positive property, since the energy expenditure is a direct result of the muscle movement [7]. However, in several cases the objects (e.g., the television) are clearly highlighted despite no direct interaction, indicating category-specific biases which are presumably the leading cause of mistakes in cross-category settings.

**Implementation Details.** Our models are trained with ADAM [28] using a weight decay of $1e^{-5}$, a batch size of 4, a learning rate of $1e^{-4}$ for 40 epochs, the model weights from Kinetics [25] and a Quadro-RTX 6000 graphic card (parameter numbers and inference times reported in Table 5). For binarization of the continuous label space, the maximum calorie prediction limit is set to 1000 kcal for Vid2Burn<sub>Diverse</sub> and 500 kcal for Vid2Burn<sub>ADL</sub> with resolution of 1 kcal. A more detailed description of the parameter settings is provided in the supplementary.

### 6. Conclusion

We introduced the novel Vid2Burn benchmark for estimating the amount of calories burned during physical activities by visually observing the human. Through our experiments, we found that the generalization ability of modern video classification CNNs is limited in this challenging task and we will keep tackling this issue in the future work. Vid2Burn will be publicly released, opening new perspectives on specific challenges in human activity analysis, such as fine-grained understanding of bodily movement and generalization to new physical activity types, since our benchmark specifically evaluates the quality of energy expenditure estimation in new situations. We hope to foster research of human understanding models which are able to capture cues of the underlying physiological processes, (e.g., active muscles and their intensity) instead of learning rigid category-specific biases seen during training.

**Broader Impact and Limitations.** This work targets energy expenditure estimation from videos. The benefits of such methods extend to multiple applications, such as supporting healthy lifestyle, e.g., by tracking exercise routines [21] or monitoring the daily physical activity level for elderly care [45]. However, both annotations in our dataset and the results inferred by our models are approximations and not exact measurements, which should be carefully used in medical care applications, as they are simplified by assuming the body weight of 150 lb, while gender, height and age are not taken into account. Furthermore, our data-driven algorithms may learn shortcuts and biases present in the data potentially resulting in a false sense of security.
References


[21] Ara Jo, Bryan D. Coronel, Courtney E. Coakes, and Arch G. Mainous III. Is there a benefit to patients using wearable devices such as fitbit or health apps on mobiles? A systematic review. *The American Journal of Medicine*, 2019. 4, 8


[34] Ara Jo, Bryan D. Coronel, Courtney E. Coakes, and Arch G. Mainous III. Is there a benefit to patients using wearable devices such as fitbit or health apps on mobiles? A systematic review. *The American Journal of Medicine*, 2019. 4, 8


[37] Li Meina, Jang Myoung Kim, and Youn Tae Kim. A combined heart rate and movement index sensor for estimating the energy expenditure. In *Sensors*, 2010.


[65] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In ECCV, 2018. 1, 2