

Jointly Learning Energy Expenditures and Activities using Egocentric Multimodal Signals

Katsuyuki Nakamura^{2*} Serena Yeung¹ Alexandre Alahi¹ Li Fei-Fei¹
¹Stanford University ²Hitachi, Ltd.

katsuyuki.nakamura.xv@hitachi.com, alahi@stanford.edu, {serena, feifeili}@cs.stanford.edu

Abstract

Physiological signals such as heart rate can provide valuable information about an individual’s state and activity. However, existing work on computer vision has not yet explored leveraging these signals to enhance egocentric video understanding. In this work, we propose a model for reasoning on multimodal data to jointly predict activities and energy expenditures. We use heart rate signals as privileged self-supervision to derive energy expenditure in a training stage. A multitask objective is used to jointly optimize the two tasks. Additionally, we introduce a dataset that contains 31 hours of egocentric video augmented with heart rate and acceleration signals. This study can lead to new applications such as a visual calorie counter.

1. Introduction

Wearable sensors such as heart rate monitors and accelerometers are widespread as a way to track physical activity and energy expenditure. To detect activities as accurately as possible, these sensors are used in many combinations and configurations, and they are attached to body parts ranging from the chest to the wrist and foot. More recently, egocentric first-person cameras have gained popularity as a new modality of wearable sensors. Activity recognition in visual data is a widely studied problem in computer vision [22, 63, 35, 2], and a number of works have investigated this task in the domain of egocentric video [48, 8, 5, 30] and in combination with other wearable sensors [59, 44].

However, the degree to which computer vision algorithms alone can be effective for energy expenditure monitoring remains an open question. We would also like to understand how much relevant information still requires other sensor modalities, and how these other signals should be best fused with vision. Pioneering works [59, 44] showed

*This work was done when Katsuyuki Nakamura was a visiting scholar at Stanford University.

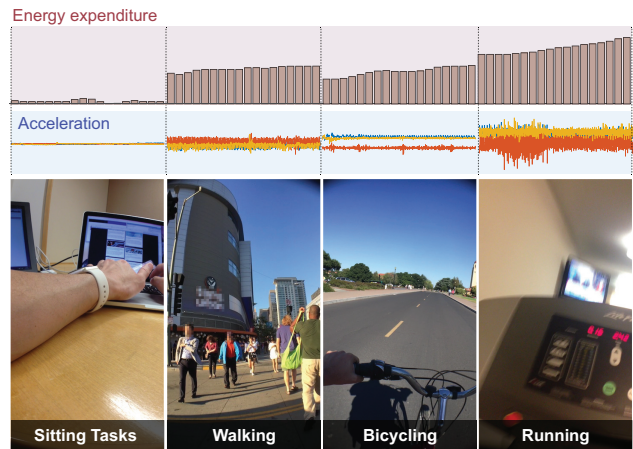


Figure 1. Egocentric videos augmented with heart rate and acceleration signals enable us to further understand activities and energy expenditures.

that systems of inertial measurement units, audio and accelerometer sensors, could improve activity segmentation and recognition over only egocentric cameras in indoor settings. However, these works were based on relatively simple visual features such as Gist and color histograms. Furthermore, they do not address energy expenditure or reason on physiological heart rate signals.

In this work, we present a quantitative study of multimodal activity detection and energy expenditure estimation, using egocentric camera data in conjunction with physiological heart rate data and accelerometer data. Specifically, we model the real-world use case and interaction of two of the most commercially popular wearable sensors: egocentric cameras and heart rate monitors (which often additionally contain an accelerometer). Given a video frame and acceleration signal, we show how to output an activity label and energy expenditure corresponding to each frame. We present a recurrent neural network-based model for multimodal activity detection and energy expenditure estimation and analyze the contributions of the different signals. We use heart rate signals as privileged self-supervision to de-

rive energy expenditure in a training stage. Since we are not aware of any available dataset that extends egocentric video with physiological data, we introduce an egocentric multimodal dataset that contains 31 hours of egocentric video, heart rate, and accelerometer data under free-living settings (Figure 1).

2. Related Work

There is a significant body of work on egocentric vision. Recent surveys can be found in [48, 8, 5, 30], which address various tasks such as activity recognition, object recognition, and video summarization. Here, we focus on activity recognition using egocentric video and wearable sensors.

Egocentric Activity Recognition Work on egocentric activity recognition has typically taken one of three approaches: motion-based, object-based, and hybrid approaches. Motion-based approaches [33, 51, 55, 68, 57] assume that each activity creates coherent motion patterns. Kitani *et al.* [33] introduces unsupervised ego-action discovery using Dirichlet processes. Poleg *et al.* [51] proposed using cumulative motion for classifying ego-actions. Object-based approaches [23, 20, 17, 36, 21, 12] rely on the objects appearing in video sequences. This setting is especially suitable for detecting activities in daily living (ADL). [23, 20] reason about state changes in household objects, and [17, 54] reason about human-object interactions. Adding mid-level cues such as face, gaze, and hands has also been investigated by [37, 19, 18, 49, 6]. Hybrid approaches [43, 38, 56] utilize both object and motion information. However, most of these works do not address datasets in diverse environments but instead focus on single, constrained environments, e.g., kitchens, houses, and laboratories. In contrast, we present a dataset and model that addresses activity detection in untrimmed (hours-long) egocentric video of diverse, natural environments including both indoors and outdoors.

Sensor-based Activity Recognition A number of works have used non-visual wearable sensors for activity recognition, including accelerometers and heart rate sensors [16, 9, 47]. Bao *et al.* [7] use multiple accelerometers on the hip, wrist, arm, ankle and thigh to classify 20 classes of everyday household activities. Abdullah *et al.* [1] combine smartphone accelerometers with crowd-sourcing in order to scale activity detection. Several works have also combined body sensors with egocentric cameras. Alahi *et al.* [4] combine cameras with wireless signals to locate humans in 3D. Spriggs *et al.* [59] use an egocentric camera, inertial measurement units (IMUs), and other sensors to classify 29 classes of kitchen activities. Maekawa *et al.* [44] use a wrist-mounted camera and sensors to detect 15 classes of

ADL recognition. However, many of these works use complex systems of sensors that are impractical for widespread use. Furthermore, they do not explore the combination of visual, heart rate, and acceleration information that is now commonly available with popular commercial devices. In this work, we construct a natural sensing system using a smartphone and wrist-worn heart-rate monitor. We provide a publicly available dataset of egocentric video augmented with heart rate and accelerometer signals, and we investigate the fusion of these signals for activity detection and energy expenditure estimation.

Recurrent Neural Networks for Activity Recognition

Donahue *et al.* [15] introduced recurrent convolutional neural networks for activity recognition in video. Yeung *et al.* uses recurrent neural network-based models for dense activity recognition [66] and action detection from frame glimpses [67]. Haque *et al.* [24] use a Recurrent Attention Model (RAM) to re-identify humans. [65, 32] use recurrent networks for single triaxial accelerometer data, [41] for a heart rate controller, and [50] for multimodal wearable sensors. To the best of our knowledge, our work is the first to investigate the performance of recurrent neural networks for multimodal egocentric video and physiological sensor data.

Energy Expenditure Estimation The most reliable method for energy expenditure estimation is considered to be direct calorimetry, which measures the production of carbon dioxide. However, this method is highly inconvenient. As a proxy, measurements of acceleration and heart rate are widely used [39], e.g., multiple accelerometers [40, 45, 11], heart rate [10], indirect estimation from heart rate and oxygen uptake relationships [34], and heart rate in combination with accelerations [16]. However, estimating energy expenditure from visual data has not yet been explored. The only work reports estimating heart rate from a head-mounted wearable camera and sensors [26, 27]. In our work, we show that reasoning on egocentric video data can be an effective estimate of energy expenditure under a free-living setting.

3. Stanford-ECM Dataset

Physiological data such as heart rate is commonly used to assess physical activity and energy expenditure. However, we are not aware of an existing dataset that supports research in egocentric video augmented with physiological sensors. Table 1 shows a comparison of existing egocentric video datasets. The CMU-MMAC dataset [59] is the only dataset that augments video with wearable sensors, but they do not provide heart rate data.

We therefore introduce a new dataset called the egocentric multimodal dataset (Stanford-ECM), which comprises

Dataset	#Subjects	#Clips	#Hour	#Class	Resolution	Mount	Video	Acc	HR	Scenario	Task
Disney World [18]	8	113	42.0	6	1280×720	Head	✓	-	-	Amusement	Activity
CMU-MMAC [59]	39	175	17.0	29	800×600	Head	✓	✓	-	Kitchen	Activity
UT Ego [36, 42]	4	4	17.0	-	480×320	Head	✓	-	-	Natural	Summary
HUJI EgoSeg [51]	3	44	10.5	-	1280×720	Head	✓	-	-	Natural	Activity
UCI ADL [23]	20	20	10.0	18	1280×720	Chest	✓	-	-	House	Activity
GTEA Gaze+ [19]	5	30	9.0	44	1280×960	Head	✓	-	-	Kitchen	Activity
MILADL [49]	20	122	6.5	23	640×360	Head/Wrist	✓	-	-	House	Activity
Intel-Ego [53]	2	10	2.0	-	1024×768	Head	✓	-	-	Indoor	Object
BEOID [12]	5	58	1.5	-	640×480	Head	✓	-	-	Indoor	Object
GTEA Gaze [19]	14	17	1.0	40	640×480	Head	✓	-	-	Kitchen	Activity
GTEA [21]	4	28	0.6	71	1280×720	Head	✓	-	-	Kitchen	Activity
UEC EgoAction [33]	1	2	0.5	37	840×480	Head	✓	-	-	Indoor	Activity
JPL-Interaction [54]	1	62	0.4	7	320×240	Head	✓	-	-	Indoor	Activity
PAMAP2 [52]	9	-	-	-	-	-	-	✓	✓	-	-
Stanford-ECM (ours)	10	113	31.0	24	720×1280	Chest	✓	✓	✓	Natural	Activity

Table 1. Comparison with available egocentric datasets. ECM is the only dataset that contains video, accelerometer, and heart rate signals.

31 hours of egocentric video augmented with heart rate and acceleration data. We label the data with temporal activity annotations and energy expenditure, specifically metabolic equivalents of task (MET). We believe that our dataset provides a realistic and challenging dataset for multimodal egocentric activity recognition and energy expenditure estimation and that it can also be of interest for additional applications including video summarization. We will make a dataset publicly available after de-identifications.

We describe data collection in Sec. 3.1, the activity and MET annotations in Sec. 3.2, and dataset statistics in Sec. 3.3.

3.1. Data Collection

Data was collected using a sensing system comprising a mobile phone and a wrist-worn heart rate sensor. The mobile phone was placed in the chest pocket of subjects to collect egocentric video and accelerometer data, and the wrist sensor provided corresponding heart rate data. Ten subjects wore the sensing system, recording 113 videos for a total of 31 hours. The lengths of the individual videos covered a diverse range from 3 minutes to about 51 minutes in length. Subjects were only instructed to perform daily life activities without constraint on how, where, or in what environments to spend time. Data collection was therefore performed under natural daily conditions.

The mobile phone collected egocentric video at 720×1280 resolution and 30 fps, as well as triaxial acceleration at 30Hz. The mobile phone was equipped with a wide-angle lens, so that the horizontal field of view was enlarged from 45 degrees to about 64 degrees. The wrist-worn heart rate sensor was used to capture the heart rate every 5 seconds (0.2 Hz). The phone and heart rate monitor was time-synchronized through Bluetooth, and all data was stored in the phone’s storage. Piecewise cubic polynomial interpola-

Activity	MET	Activity	MET
1. BicyclingUphill	14.0	13. Shopping	2.3
2. Running	8.0	14. Strolling	2.0
3. Bicycling	7.5	15. FoodPreparation	2.0
4. PlayingWithChildren	5.8	16. TalkingStanding	1.8
5. ResistanceTraining	5.0	17. TalkingSitting	1.5
6. AscendingStairs	4.0	18. SittingTasks	1.5
7. Calisthenics	3.8	19. Meeting	1.5
8. Walking	3.5	20. Eating	1.5
9. DescendingStairs	3.5	21. StandingInLine	1.3
10. Cooking	3.3	22. Riding	1.3
11. Presenting	3.0	23. Reading	1.3
12. Driving	2.5	24. Background*	N/A

Table 2. Definitions activity classes and their MET values [3]. Each activity corresponds to one of four activity intensities: vigorous (≥ 6 METs), moderate (3.0-5.9 METs), light (1.6-2.9 METs), and sedentary (1.0-1.5 METs). * represents classes added in addition to [3].

tion was used to fill in any gaps in heart rate data. Finally, data was aligned to the millisecond level at 30 Hz.

3.2. Activity and MET Annotations

It is challenging to define activity classes in untrimmed, unconstrained egocentric video due to the diversity and continuity of activities that may exist, compared with most existing activity recognition datasets (e.g., ActivityNet [25], THUMOS [29], UCF-101 [58], and Sports-1M [31]). However, an important goal of egocentric activity detection, and the focus of our work, is physical activity and health tracking.

We therefore define the activity classes as subsets of a compendium of physical activities [3]. Each of these classes is associated with a metabolic equivalent of task (MET)

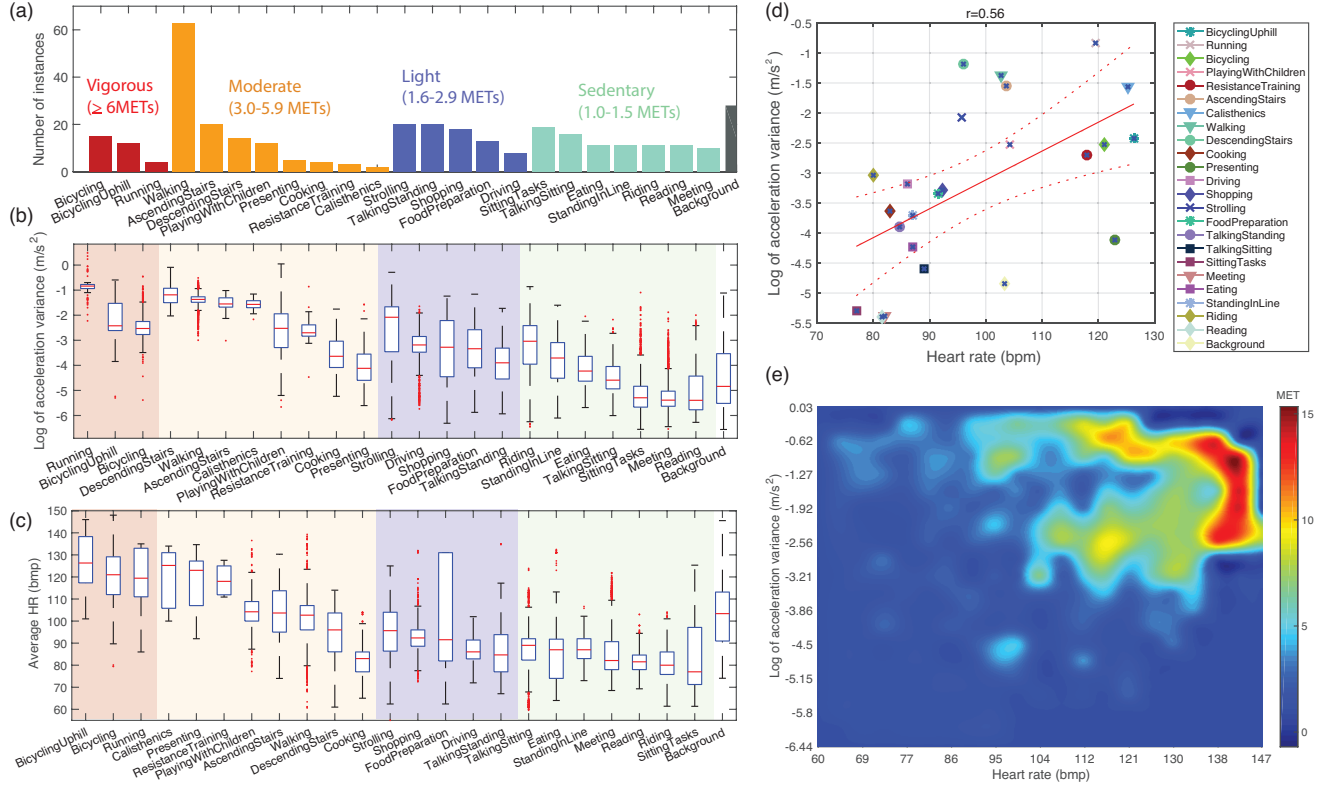


Figure 2. **Statistics of dataset.** (a) Number of instances, (b) acceleration variance for each activity, (c) average heart rate for each activity, (d) scatter plot of heart rate and acceleration variance, and (e) MET distribution for heart rate and acceleration variance.

measure [3]. MET is a physiological measure defined relative to the resting metabolic rate and expresses the energy cost of physical activities. For instance, quiet sitting is considered to be 1.0 MET, walking to be 3.0 MET, and bicycling to be 7.5 MET ($\text{kcal} \cdot \text{kg}^{-1} \cdot \text{h}^{-1}$). MET is a simple and practical measure and is widely used to quantify energy expenditure.

For the first release of a dataset, we temporally annotated a total of 24 MET-associated activities. We selected 23 activity classes from a compendium [3] and added 1 additional class, *background*, that corresponds to miscellaneous activities such as *taking pictures*, *riding an elevator*, and *parking a bicycle*. A single activity can co-occur at the same time. Table 2 shows the definition and MET value of each class.

3.3. Dataset Statistics

Figure 2(a) shows the number of video instances depicting each activity. *Walking* appears in the greatest number of videos; more than 63 of the 113 videos contain some walking. *Running* and *calisthenics* appears in the fewest number of videos. The average number of distinct activity types per video is 4.1, and there are often multiple occurrences of the same activity in a video. The durations span a range from several seconds to a half hour. The longest duration is *meeting*, which occurred for 2,498 seconds in a single video. The

shortest duration is an instance of *ascending stairs* which occurred for 1.4 seconds. This variety presents challenging scenarios for activity detection.

Figure 2(b) shows the distribution of acceleration variance measurements by activity class in log scale, and Figure 2(c) shows the distribution of heart rate measurements by activity class. Both classes with higher heart rate and higher acceleration variance tend to have higher MET values, indicating their usefulness for predicting physical activity and MET. Several classes such as *walking*, *talking-sitting*, *riding*, and *meeting* display higher variance due to frequent co-occurrence. This suggests that the sensor data alone is insufficient for accurate prediction.

Figure 2(d) shows a scatter plot of median heart rate and acceleration variance per class. This plot quantifies the correlation between heart rate and acceleration variance, which has a correlation coefficient of $r = 0.56$. Finally, Figure 2(e) visualizes the MET distribution for heart rate and acceleration variance, and shows that both heart rate and acceleration are strong indicators of MET.

4. Model

In this section, we describe our model for joint activity detection and energy expenditure regression. We formulate

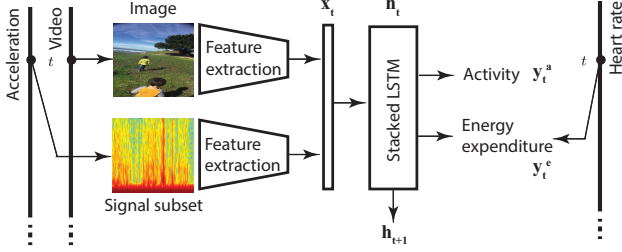


Figure 3. **EnergyRCN**: a model for joint activity detection and energy expenditure regression from egocentric multimodal signals. The model takes as input multimodal video and acceleration features from long, untrimmed sequences. A multi-task loss is used to jointly optimize activity detection y_t^a and energy expenditure regression y_t^e at each timestep. The heart rate signal is used as a self-supervision to derive energy expenditure.

our model as a recurrent long short-term memory (LSTM) neural network that takes as input a multimodal feature representation of video and acceleration signals and outputs an activity label and energy expenditure corresponding to each frame (overview in Figure 3). In the following, we describe how we adapt this model for our multimodal input and how we use heart rate as a self-supervised signal for energy expenditure regression.

4.1. Feature Representation

The input to our model is a sequence of video frames $\mathcal{V} = \{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_T\}$ and a sequence of triaxial acceleration signals $\mathcal{A} = \{\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_T\}$. In the following, we describe our corresponding feature representations \mathbf{x}_t^v and \mathbf{x}_t^a for each of the signals.

Visual features Motivated by the recent success of recurrent convolutional neural networks (CNN) for video analysis [15], we visually represent each video frame using frame-level CNN features $\mathbf{x}_t^v = \text{CNN}_{\theta_c}(\mathbf{v}_t)$. Here, CNN_{θ_c} uses a convolutional neural network to transform video frame \mathbf{v}_t at time t into a M -dimensional feature vector. In our experiments, we use an Inception network [60] pretrained on ImageNet [14] and extract pool 5 features, such that $\mathbf{x}_t^v \in \mathbb{R}^{1024}$.

Acceleration features Acceleration features are directly extracted from the time-series of the raw signals. We followed the standard time-domain and frequency-domain features [39] using the sliding window of the most recent observations. For time-domain features, we compute mean, standard deviation, skewness, kurtosis, percentiles (10th, 25th, 50th, 75th, 90th), and acceleration counts for each axis. Also, we add the correlation coefficients between each axis $\mathbf{a}_t = \{a_{x_t}, a_{y_t}, a_{z_t}\}$. For frequency-domain features,

we use the spectral entropy J , where

$$J = - \sum_{i=0}^{N/2} \bar{P}_i \cdot \log_2 \bar{P}_i. \quad (1)$$

Here, \bar{P}_i is the normalized power spectral density computed from a short-time Fourier transform (STFT). We use a temporal pooling [55] of four types of windows (1, 3, 5, and 10 seconds). All features are concatenated so that the acceleration feature vector has dimension $\mathbf{x}_t^a \in \mathbb{R}^{144}$.

Multimodal fusion We adopt an early fusion scheme. The above mentioned features are concatenated into a single feature vector $\mathbf{x}_t = (\mathbf{x}_t^v, \mathbf{x}_t^a)^\top$ to produce the input to the LSTM network at each time step. Therefore, the final feature vector is a dimension of $\mathbf{x}_t \in \mathbb{R}^{1168}$.

4.2. Multitask Training with Self-supervision

Our intuition is that the tasks of activity detection and energy expenditure estimation are highly related and that providing a network with additional supervision in the form of energy expenditure will help the network achieve stronger performance for activity detection as well. We therefore introduce a multi-task training objective to jointly optimize the tasks of activity detection and energy expenditure regression.

Multitask loss Given training data in the form of $(\mathbf{x}_t, y_t^a, y_t^e)$ for each frame, where $\mathbf{x}_t \in \mathbb{R}^d$ is the input feature vector, $y_t^a \in \mathbb{R}^{24}$ is the ground truth activity label, and $y_t^e \in \mathbb{R}$ is the derived energy expenditure, we optimize the following multi-task loss.

$$L = L_{\text{act}} + \lambda L_{\text{EE}} \quad (2)$$

Here the first term L_{act} is a cross-entropy loss for activity detection. The second term L_{EE} is a Huber loss for energy expenditure regression

$$L_{\text{EE}}(r) = \begin{cases} \frac{1}{2}r^2 & \text{if } |r| \leq \delta \\ \delta \left(|r| - \frac{1}{2}\delta \right) & \text{otherwise} \end{cases}, \quad (3)$$

where $r = y_t^e - \hat{y}_t^e$. Since derived energy expenditures have outliers, we handle them by using a robust Huber loss function. λ controls the trade-off between the two tasks.

Energy expenditure Utilizing linear relationships between heart rate (HR) and oxygen uptake (VO_2) [28], we estimate energy expenditure by using HR signals:

$$y_t^e = \alpha \text{HR}_t + \beta \text{weight} + \gamma \text{HR}_t \text{weight}, \quad (4)$$

where y_t^e is the energy expenditure ($\text{kcal} \cdot \text{min}^{-1}$) at frame t . We set as $\alpha = 4.56$, $\beta = 0.0265$, $\gamma = 0.1506$, and body weight as 60 kg. We also apply a moving average of 10 seconds centered at each frame t to smooth HR signals.

5. Experiments

We evaluated our model on the dataset for two tasks: activity detection and energy expenditure regression. We describe the results for each of these in this section.

5.1. Setup

All videos were down-sampled to 1 fps and resized to 227×227 . Horizontal reflection and random cropping were performed for data augmentation during the training phase. The training set consisted of 91,043 images and the test set of 30,930 images. The test set contained two unknown (zero-shot) subjects that were not contained in the training set. The EnergyRCN model was trained using truncated BPTT (Back-Propagation Through Time) [46]. The RMSPProp [13] per-parameter adaptive update rule was used with a base learning rate of 2×10^{-5} and decay of 0.95. Hyperparameters were set via cross-validation; the LSTM consisted of 2 hidden layers with 2048 hidden units. λ was set to 1.0. The sequence length was 10 frames (10 seconds), and the batch size was 60 sequences. Chainer [61] was used for implementation.

5.2. Activity Detection

In this task, the input was untrimmed, egocentric sequences of video and acceleration, and the output was the activity labels of each frame. Performance was evaluated using mean average precision (mAP). We compare our model with different baseline models explained below¹.

- Liu *et al.* [40]: an acceleration-based baseline, which combines an RBF kernel SVM with acceleration features \mathbf{x}_t^a .
- Inception [60]: a frame-level baseline, which uses an RBF kernel SVM trained on top of InceptionNet.
- LRCN [15]: a temporal baseline that combines CNN and naive LSTM. We do not back-propagate into the CNN layers.
- C3D [62]: a spatial-temporal CNN baseline, which is combined with a linear SVM. A publicly available model that was pre-trained on Sports-1M was used.
- iDT+FV [64]: a non-deep learning based baseline. The hand-crafted features (HOG+HOF+MBH+Traj) are first dimensions reduced by a factor of 6 with PCA and then encoded with a Gaussian mixture model (GMM) and Fisher vectors (FV).

Table 3 summarizes detection results at activity intensities. Each column corresponds to one of four activity intensities: vigorous, moderate, light, and sedentary. Fusing modalities together achieved comparable or the highest performance for all activity intensities, but the difference was

Method	Vig	Mod	Lig	Sed	All
Liu <i>et al.</i> [40]	68.0	76.9	52.7	39.7	53.9
Inception [60]	80.9	86.1	72.4	66.6	73.2
LRCN [15]	84.9	92.3	76.6	61.2	72.9
C3D [62]	82.0	80.6	65.7	58.0	66.3
iDT+FV [64]	80.8	72.8	59.8	49.7	59.8
EnergyRCN (A)*	58.1	71.9	49.4	42.4	52.3
EnergyRCN (V)*	82.3	89.9	70.5	71.0	76.2
EnergyRCN (V+A)*	82.2	92.1	78.3	70.8	78.1

Table 3. Activity detection results (mAP). Each column corresponds to activity intensity. Vig: vigorous (≥ 6 METs), Mod: moderate (3.0-5.9 METs), Lig: light (1.6-2.9 METs), and Sed: sedentary (1.0-1.5 METs). Only our method jointly predicts activity detection and energy expenditure estimation (denoted as *) (A: acceleration feature, V: visual feature).

not huge, illustrating the strength of vision models for activity detection.

Activities belonging to vigorous and moderate can be distinguished by motion patterns (see Fig. 2(b)), which lead to higher success rate of motion-based method (e.g., iDT, C3D, and Liu *et al.*). On the other hand, large-variance movements under light activities lead to a drop in mAP. Visual features show effectiveness for improving performance in that case.

Figure 4(a) shows the confusion matrix of an acceleration-based baseline. Significant ambiguities can be seen, especially for light activities (right hand side in the matrix), whereas Figure 4(c) shows that adding visual features make it possible to resolve confusion between many classes such as *meeting*, *sitting tasks*, and *standing in line*, which are all fairly sedentary.

5.3. Energy Expenditure Regression

We also show the results for our model on the task of energy expenditure regression. We compare our model with an additional baseline of MET-based regression. The MET values can be used to compute energy expenditure using the formula

$$EE = \alpha \sum_{k=1}^K d_k m_k, \quad (5)$$

where m_k is the MET value of detected activity k , d_k is the activity's duration in hours, and α is a constant defined as body weight $\times 1.05$. The constant 1.05 represent a resting energy expenditure ($\text{kcal} \cdot \text{METs}^{-1} \cdot \text{kg}^{-1} \cdot \text{h}^{-1}$) defined by American College of Sports Medicine (ACSM). EE is the energy expenditure ($\text{kcal} \cdot \text{h}^{-1}$).

Table 4 shows the ability of our model to estimate energy expenditures. Since the baselines did not estimate both activity and energy expenditures simultaneously, we trained

¹We used the original video frame rate of 30 fps for [62][64]

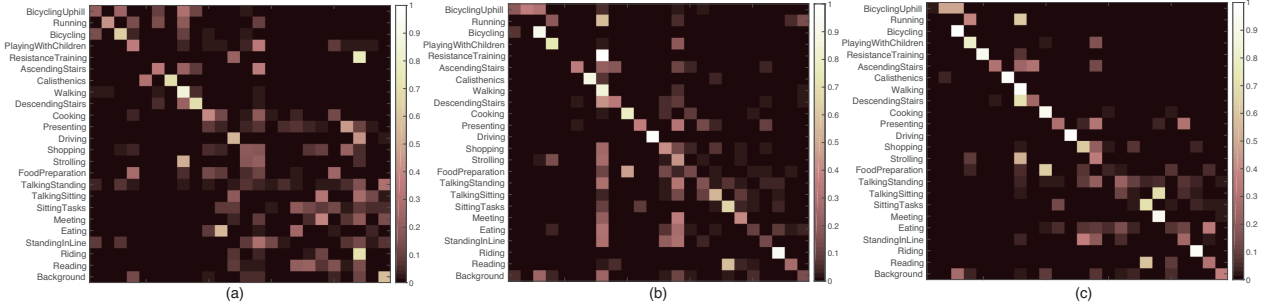


Figure 4. Confusion matrices of activity detection: (a) acceleration-based baseline, (b) Inception network, and (c) EnergyRCN (V+A). Significant ambiguities can be seen in (a), which used only acceleration signals.

Method	$\mu_{err} \pm \sigma_{err}$	ρ
MET _{GT} (upper bound)	6.84 ± 3.92	N/A
Liu <i>et al.</i> [40]	0.74 ± 0.29	0.346
Inception [60]	0.72 ± 0.24	0.341
LRCN [15]	0.73 ± 0.28	0.365
EnergyRCN (A)*	0.77 ± 0.36	0.386
EnergyRCN (V)*	0.75 ± 0.29	0.382
EnergyRCN (V+A)*	0.70 ± 0.27	0.407

Table 4. Results of energy expenditure regression. μ_{err} and σ_{err} are absolute errors and standard deviations, respectively ($\text{kcal} \cdot \text{m}^{-1}$). ρ are cross correlation coefficients between true and estimated energy expenditures. MET_{GT} utilizes ground-truth activity labels to determine energy expenditure, while our method jointly predicts the activity and energy expenditure (denoted as *).

individual regression models to output energy expenditures. Namely, SVR and LSTM were trained for [40][60][15].

EnergyRCN (V) could predict energy expenditure with similar error rate while outperforming activity detection baselines. Visual features were again a good individual feature, but the gap between visual features and acceleration was smaller for energy expenditure regression than for activity detection. This also demonstrates that acceleration signals alone bring benefits to estimating energy expenditures. This was backed up with dataset statistics, in which the acceleration correlated to MET value.

5.4. Ablation study

To understand how well multi-task training helps to improve the performance, we compared with the single task model as well as feature combinations (Table 5). Multi-task training consistently improved the mAP for activity detection. Also, the performances of energy expenditure regression are comparable with a single task, showing that our model learned the shared representations for both tasks.

Finally, qualitative results of energy expenditure regression are shown in Figure 5. MET-based regression utilizes ground-truth activity labels; however, expenditure estima-

	Detection			Regression		
	A	V	V+A	A	V	V+A
Single-task	51.9	72.9	76.7	0.771	0.732	0.699
Multi-task	52.3	76.2	78.1	0.768	0.751	0.696

Table 5. Results of ablation study. Multi-task training improved activity detection performance (mAP) and also achieved comparable errors of energy expenditure regression ($\text{kcal} \cdot \text{m}^{-1}$).

tion performance is worse than our model since the discrete outputs are poor in accuracy.

5.5. Discussion

The power of vision Through comprehensive study on multimodal activity detection and energy expenditure regression, we achieve greater understanding of the power of visual information. Intuitively, multimodal data can provide additional rich information not contained in vision, for both activity detection and energy expenditure regression. Our experimental results quantify this gain, but also show that visual information alone can also be a powerful signal for these tasks. This insight leads us to new applications such as a visual calorie counter.

Activity quantification We introduce models for estimating energy expenditure with vision alone, and also utilizing METs associated with vision. To the best of our knowledge, these methods are novel and enable us to reason on activities in a way that quantifies physical workload. Activity recognition itself does not provide such internal/invisible states, so our work brings the benefits of further understanding activities.

Future work Physiological signals can encode information not only about physical activities but also mental activities. Distinguishing and characterizing these will provide rich information that can be used for tasks such as video segmentation and video summarization. Further exploration of this will enable a deeper understanding of human activity.

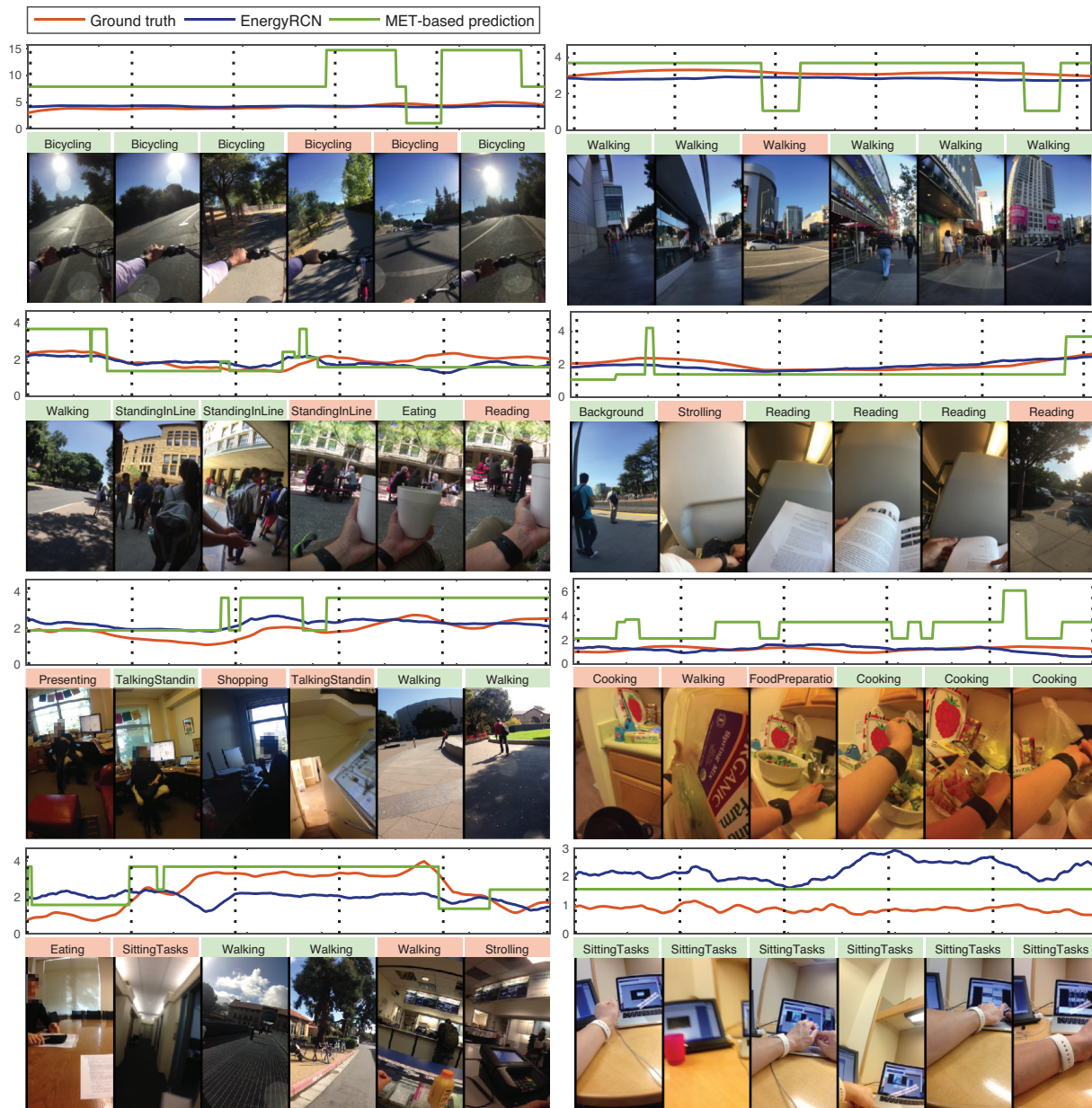


Figure 5. Examples of EnergyRCN predictions. EnergyRCN estimates activities (bottom figures) and energy expenditures (above graph by $\text{kcal} \cdot \text{min}^{-1}$) simultaneously. In the above graph, the red line is the ground truth, the blue line is EnergyRCN estimates, and the green line is the MET_{GT} estimates. While MET_{GT} only produces discrete values with large errors, our model can produce continuous energy expenditures accurately. The bottom examples represent hard cases for energy expenditure regression that have relatively large differences.

6. Conclusion

In this paper, we introduced the task of energy expenditure estimation from an egocentric video perspective. To address this task, we first constructed the Stanford-ECM dataset of egocentric video augmented with heart rate and acceleration signals. We explored the statistics of the dataset and presented a recurrent neural network architecture for reasoning on egocentric multimodal data. Our EnergyRCN model jointly predicts energy expenditures and

activities, achieving state-of-the-art results. We believe this work will lead to new directions of research on various aspects of egocentric video understanding, ranging from activity detection to video summarization.

Acknowledgements Serena Yeung, Alexandre Alahi and Li Fei-Fei are partially supported by a grant from Intel. We would like to thank Emily Tang, Juan Carlos Niebles, Cewu Lu, Marius Cătălin Jordan, Guido Pusiol, and all the Vision lab members for helpful comments and data collection.

References

- [1] S. Abdullah, N. D. Lane, and T. Choudhury. Towards population scale activity recognition: a framework for handling data diversity. In *AAAI*, 2012. 2
- [2] J. K. Aggarwal and M. S. Ryoo. Human activity analysis: A review. *ACM Computing Surveys*, 43(3):1–43, 2011. 1
- [3] B. E. Ainsworth, W. L. Haskell, S. D. Herrmann, N. Meckes, D. R. Bassett, C. Tudor-Locke, J. L. Greer, J. Vezina, M. C. Whitt-Glover, and A. S. Leon. 2011 compendium of physical activities: a second update of codes and MET values. *Medicine and Science in Sports and Exercise*, 43(8):1575–1581, 2011. 3, 4
- [4] A. Alahi, A. Haque, and L. Fei-Fei. RGB-W: When vision meets wireless. In *ICCV*, 2015. 2
- [5] S. Bambach. A survey on recent advances of computer vision algorithms for egocentric video. In *arXiv:1501.02825*, 2015. 1, 2
- [6] S. Bambach, D. J. Crandall, and C. Yu. Lending a hand : detecting hands and recognizing activities in complex egocentric interactions. In *ICCV*, 2015. 2
- [7] L. Bao and S. S. Intille. Activity recognition from user-annotated acceleration data. In *Pervasive Computing*, 2004. 2
- [8] A. Betancourt, P. Morerio, C. S. Regazzoni, and M. Rauterberg. The evolution of first person vision methods: a survey. *IEEE Trans. Circuits and Systems for Video Technology*, 25(5):744–760, 2015. 1, 2
- [9] A. G. Bonomi, A. H. C. Goris, B. Yin, and K. R. Westerterp. Detection of type, duration, and intensity of physical activity using an accelerometer. *Medicine and Science in Sports and Exercise*, 41(9):1770–1777, 2009. 2
- [10] S. M. Ceesay, A. M. Prentice, K. C. Day, P. R. Murgatroyd, G. R. Goldberg, W. 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(02):175–186, 1988. 2
- [11] S. E. Crouter, K. G. Clowers, D. R. Bassett, E. Scott, and A. Jr. A novel method for using accelerometer data to predict energy expenditure. *Journal of Applied Physiology*, 100(4):1324–1331, 2006. 2
- [12] D. Damen and A. Calway. You-do , I-learn : discovering task relevant objects and their modes of interaction from multi-user egocentric video. In *BMVC*, 2014. 2, 3
- [13] Y. N. Dauphin, J. Chung, and Y. Bengio. RMSProp and equilibrated adaptive learning rates for non-convex optimization. In *arXiv:1502.04390*, 2015. 6
- [14] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009. 5
- [15] J. Donahue, L. A. Hendricks, S. Saenko, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, 2015. 2, 5, 6, 7
- [16] K. Ellis, J. Kerr, S. Godbole, G. Lanckriet, D. Wing, and S. 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, 2014. 2
- [17] A. Fathi, A. Farhadi, and J. M. Rehg. Understanding egocentric activities. In *ICCV*, 2011. 2
- [18] A. Fathi, J. K. Hodgins, and J. M. Rehg. Social interactions: a first-person perspective. In *CVPR*, 2012. 2, 3
- [19] A. Fathi, Y. Li, and J. M. Rehg. Learning to recognize daily actions using gaze. In *ECCV*, 2012. 2, 3
- [20] A. Fathi and J. M. Rehg. Modeling actions through state changes. In *CVPR*, 2013. 2
- [21] A. Fathi, X. Ren, and J. M. Rehg. Learning to recognize objects in egocentric activities. In *CVPR*, 2011. 2, 3
- [22] D. M. Gavrilu. The visual analysis of human movement: a survey. *CVIU*, 73(1):82–98, 1999. 1
- [23] P. Hamed and D. Ramanan. Detecting activities of daily living in first-person camera views. In *CVPR*, 2012. 2, 3
- [24] A. Haque, A. Alahi, and L. Fei-Fei. Recurrent attention models for depth-based person identification. In *CVPR*, 2016. 2
- [25] F. C. Heilbron, V. Escorcia, B. Ghanem, J. C. Niebles, and U. Norte. ActivityNet : a large-scale video benchmark for human activity understanding. In *CVPR*, 2015. 3
- [26] J. Hernandez, Y. Li, J. M. Rehg, and R. W. Picard. BioGlass: physiological parameter estimation using a head-mounted wearable device. In *Intl. Conf. Wireless Mobile Communication and Healthcare (MobiHealth)*, 2014. 2
- [27] J. Hernandez, Y. Li, J. M. Rehg, and R. W. Picard. Cardiac and respiratory parameter estimation using head-mounted motion-sensitive sensors. *EAI Endorsed Trans. Pervasive Health and Technology*, 1(1), 2015. 2
- [28] H. K. Hiilloskorpi, M. E. Pasanen, M. G. Fogelholm, R. M. Laukkanen, and A. T. Mänttari. Use of heart rate to predict energy expenditure from low to high activity levels. *International Journal of Sports Medicine*, 24(05):332–336, 2003. 5
- [29] H. Idrees, A. R. Zamir, Y.-G. Jiang, A. Gorban, I. Laptev, R. Sukthankar, and M. Shah. The THUMOS challenge on action recognition for videos "in the wild". In *arXiv:1604.06182*, 2016. 3
- [30] T. Kanade and M. Hebert. First-person vision. *Proc. IEEE*, 100(8):2442–2453, 2012. 1, 2
- [31] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014. 3
- [32] A. M. Khan, Y. K. Lee, S. Y. Lee, and T. S. Kim. A triaxial accelerometer-based physical-activity recognition via augmented-signal features and a hierarchical recognizer. *IEEE Trans. Information Technology in Biomedicine*, 14(5):1166–1172, 2010. 2
- [33] K. M. Kitani, T. Okabe, Y. Sato, and A. Sugimoto. Fast unsupervised ego-action learning for first-person sports videos. In *CVPR*, 2011. 2, 3
- [34] A. Kurpad, R. Raj, K. Maruthy, and M. Vaz. A simple method of measuring total daily energy expenditure and physical activity level from the heart rate in adult men. *European Journal of Clinical Nutrition*, 60(1):32–40, 2006. 2
- [35] G. Lavee, E. Rivlin, and M. Rudzsky. Understanding video events: a survey of methods for automatic interpretation

- of semantic occurrences in video. *IEEE Trans. Systems, Man and Cybernetics Part C: Applications and Reviews*, 39(5):489–504, 2009. 1
- [36] Y. J. Lee, J. Ghosh, and K. Grauman. Discovering important people and objects for egocentric video summarization. In *CVPR*, 2012. 2, 3
- [37] Y. Li, A. Fathi, and J. M. Rehg. Learning to predict gaze in egocentric video. In *CVPR*, 2013. 2
- [38] Y. Li, Z. Ye, and J. M. Rehg. Delving into egocentric actions. In *CVPR*, 2015. 2
- [39] S. Liu, R. Gao, and P. Freedson. Computational methods for estimating energy expenditure in human physical activities. *Medicine and Science in Sports and Exercise*, 44(11):2138–2146, 2013. 2, 5
- [40] S. Liu, R. X. Gao, D. John, J. W. Staudenmayer, and P. S. Freedson. Multisensor data fusion for physical activity assessment. *IEEE Trans. Biomedical Engineering*, 59(3):687–696, 2012. 2, 6, 7
- [41] C.-H. Lu, W.-C. Wang, C.-C. Tai, and T.-C. Chen. Design of a heart rate controller for treadmill exercise using a recurrent fuzzy neural network. *Computer Methods and Programs in Biomedicine*, 128:27–39, 2016. 2
- [42] Z. Lu and K. Grauman. Story-driven summarization for egocentric video. *CVPR*, 2013. 3
- [43] M. Ma, H. Fan, and K. M. Kitani. Going deeper into first-person activity recognition. In *CVPR*, 2016. 2
- [44] T. Maekawa, Y. Yanagisawa, Y. Kishino, K. Ishiguro, K. Kamei, Y. Sakurai, and T. Okadome. Object-based activity recognition with heterogeneous sensors on wrist. In *Pervasive*, 2010. 1, 2
- [45] B. Mortazavi, N. Alsharufa, S. I. Lee, M. Lan, M. Sarrafzadeh, M. Chronley, and C. K. Roberts. MET calculations from on-body accelerometers for exergaming movements. In *IEEE Conf. Body Sensor Networks (BSN)*, 2013. 2
- [46] M. C. Mozer. A focused backpropagation algorithm for temporal pattern recognition. *Complex Systems*, 3:349–381, 1989. 6
- [47] M. Nakanishi, S. Izumi, S. Nagayoshi, H. Sato, H. Kawaguchi, M. Yoshimoto, T. Ando, S. Nakae, C. Usui, T. Aoyama, and S. Tanaka. Physical activity group classification algorithm using triaxial acceleration and heart rate. In *EMBC*, 2015. 2
- [48] T. H. C. Nguyen, J. C. Nebel, and F. Florez-Revuelta. Recognition of activities of daily living with egocentric vision: A review. *Sensors*, 16(1), 2016. 1, 2
- [49] K. Ohnishi, A. Kanehira, A. Kanazaki, and T. Harada. Recognizing activities of daily living with a wrist-mounted camera. In *CVPR*, 2016. 2, 3
- [50] F. Ordóñez and D. Roggen. Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition. *Sensors*, 16(1), 2016. 2
- [51] Y. Poleg, C. Arora, and S. Peleg. Temporal segmentation of egocentric videos. In *CVPR*, 2014. 2, 3
- [52] A. Reiss and D. Stricker. Introducing a new benchmarked dataset for activity monitoring. In *ISWC*, 2012. 3
- [53] X. Ren and C. Gu. Figure-ground segmentation improves handled object recognition in egocentric video. In *CVPR*, 2010. 3
- [54] M. S. Ryoo and L. Matthies. First-person activity recognition: what are they doing to me? In *CVPR*, 2013. 2, 3
- [55] M. S. Ryoo, B. Rothrock, and L. Matthies. Pooled motion features for first-person videos. In *CVPR*, 2015. 2, 5
- [56] S. Singh, C. Arora, and C. V. Jawahar. First Person Action Recognition Using Deep Learned Descriptors. In *CVPR*, 2016. 2
- [57] S. Singh, C. Arora, and C. V. Jawahar. Trajectory aligned features for first person action recognition. *Pattern Recognition*, 62:45–55, 2017. 2
- [58] K. Soomro, A. Zamir, and M. Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. In *arXiv:1212.0402*, 2012. 3
- [59] E. H. Spriggs, F. De La Torre, and M. Hebert. Temporal segmentation and activity classification from first-person sensing. In *CVPR Workshop on Egocentric Vision (CVPRW)*, 2009. 1, 2, 3
- [60] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, 2015. 5, 6, 7
- [61] S. Tokui, K. Oono, S. Hido, and J. Clayton. Chainer: a next-generation open source framework for deep learning. In *NIPS Workshop on Machine Learning Systems*, 2015. 6
- [62] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3D convolutional networks. In *ICCV*, 2015. 6
- [63] P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea. Machine recognition of human activities: A survey. *IEEE Trans. Circuits and Systems for Video Technology*, 18(11):1473–1488, 2008. 1
- [64] H. Wang and C. Schmid. Action recognition with improved trajectories. In *ICCV*, 2013. 6
- [65] J. Y. Yang, J. S. Wang, and Y. P. Chen. Using acceleration measurements for activity recognition: An effective learning algorithm for constructing neural classifiers. *Pattern Recognition Letters*, 29(16):2213–2220, 2008. 2
- [66] S. Yeung, O. Russakovsky, N. Jin, M. Andriluka, G. Mori, and L. Fei-Fei. Every moment counts: Dense detailed labeling of actions in complex videos. *arXiv preprint arXiv:1507.05738*, 2015. 2
- [67] S. Yeung, O. Russakovsky, G. Mori, and L. Fei-Fei. End-to-end learning of action detection from frame glimpses in videos. In *CVPR*, 2016. 2
- [68] R. Yonetani and K. M. Kitani. Recognizing micro-actions and reactions from paired egocentric videos. In *CVPR*, 2016. 2