ESTIMATING PHYSICAL ACTIVITY INTENSITY AND ENERGY EXPENDITURE USING COMPUTER VISION ON VIDEOS

Philip Saponaro, Haoran Wei, Gregory Dominick, Chandra Kambhamettu

University of Delaware, Newark, DE

ABSTRACT

Estimating physical activity (PA) intensity and energy expenditure (EE) is a problem that typically requires the use of wearable sensors such as a heart rate monitor, or accelerometer. We investigate the accuracy of a computer vision system using videos recorded from a pair of wearable video glasses to estimate PA strength and EE automatically using age, gender, speed, and activity cues. Age and gender are obtained using the Deep EXpectation network, while activity is estimated from joint angles and movement speed. We also present results on a study of 50 participants performing four different activities while measuring corresponding features of interest such as height, weight, age, sex, and ground truth EE and PA strength data collected via accelerometer. We present both the results of each computer vision subsystem and overall accuracy of the PA strength estimation (89.5%) and the average EE difference (1.96 kCal/min).

***Index Terms*—** Convolutional neural networks, energy expenditure, physical activity intensity, action recognition

1. INTRODUCTION

Automatically determining the physical activity (PA) strength and energy expenditure (EE) of people in a scene would dramatically improve PA measurement for public health monitoring as well as applications in park management and design [1], smart gyms [2], or even home gaming Virtual Reality (VR) consoles such as the Playstation VR, which mainly relies on camera sensors.

PA intensity is often derived from known metabolic equivalent (MET) values (ratio of EE to mass) that are associated with a particular workload. The energy requirements at rest are equivalent to 1 MET, while MET ranges between 1.5-2.9, 3.0-5.9, and $>= 6.0$ METs reflect light, moderate, and vigorous intensity, respectively. Accelerometer-derived estimates of PA intensity and EE are based in part, on an estimated MET value. [3].

Another common metric in PA measurement is EE, measured in kCals/min (i.e. calories/min). Similar to METs, EE increases proportionately with increasing intensity. EE can be directly obtained via indirect calorimetry, although accelerometers are more commonly used to estimate EE in free-living conditions. As such, current objective measures of PA rely heavily on wearable sensors which may not be feasible when monitoring population-level PA and other health behaviors.

Fig. 1. Example output from our system. All of the annotations are predictions that have been calculated automatically (offline).

In this paper, we utilize computer vision and machine learning techniques in order to estimate the PA strength category (light, moderate, vigorous) and the average number of calories burned (kCal/min). Computer vision and machine learning have recently progressed to the point where Convolutional Neural Networks (CNNs) have performed extremely well on a variety of tasks. We approach the problem by estimating various features of interest from video sequences using CNNs, and using those features to estimate the PA strength via a random forest. The features of interest we extract from the video sequence are age, sex, activity type, and speed. Age and gender are determined via the Deep EXpectation (DEX) network, while PA type is calculated from joint angles and speed.

The contribution of this work is as follows:

- Investigation of computer vision accuracy for obtaining features of interest on a novel dataset of 50 participants performing four different activities.
- Development and analysis of activity recognition using joint angles and speed features
- Quantitative analysis of PA strength and EE using a random forest to fuse the features extracted by computer vision

The rest of this paper is organized as follows. Section 2 gives an overview of ways to calculate PA strength and EE, and related recent computer vision algorithms. Section 3 details the data collection and the estimation of features via computer vision. Section 4 describes the results of each computer vision subsystem and the overall accuracy for estimating PA strength and EE. Finally, the paper is summarized and concluded in Section 5.

2. BACKGROUND

Estimating EE has been researched for decades. In a comparison from 1989 [4], EE was estimated from an accelerometer, heart rate, and (manual) video analysis for activity rating, and found that the accelerometer data was the most accurate in terms of kCal/min difference. A newer study [5] from 2012 fuses features from various personal body sensors (accelerometer, heart rate monitor, respiration monitor, and skin temperature) via a hidden markov model. Recurrent neural networks for fusing sensor data has also been tried [6]. Accelerometers were found to be usable for distinguishing certain activities (e.g. walking vs running), but lacks the ability to distinguish other activities, such as swimming or cycling [7]. Computer vision lacks this limitation.

Automatically extracting the EE from video requires three steps: detection, tracking, and energy estimation. Each of these are separate computer vision tasks with a dense history of work, of which we will mention a few state-of-the-art methods. The state-of-the-art for detection and tracking leverage CNNs, such as FasterRCNN [8] or YOLOv3 [9] for detection, and Simple Online and Realtime Tracking (SORT) [10] for tracking. For activity recognition, Long-Short-Term-Memory networks have been used [11]. Skeletal pose information was used [12] to match similar skeletons to each other for action recognition.

Studies that combine computer vision with EE studies are described below. The CAM system [13] uses an overhead camera with semi-automatic algorithms with user input and Kalman filters for tracking [14], and movement speed is used to calculate the PA intensity. The most recent and similar study to ours uses CNNs to extract features for each activity type, and then use a LSTM for estimating METs [15]. In our study, we automatically detect, track, and extract individual features such as age, gender, activity type, and speed, and use those features to estimate the final EE from a ground-based viewpoint.

3. METHODS

In this section, we describe the data collection, the detection, tracking, and feature extraction computer vision subsystems, and the random forest classifier and regressor for PA strength and EE estimation.

3.1. Data

50 adults participated in this study. After assessing age, gender, height and weight, subjects were fitted with an ActiGraph GT3X accelerometer prior to performing four different timed activities on an outdoor field: stand (1 min), walk (3 min), jog (3 min), and a single 50 meter sprint. A pair of GoGloo video glasses was used to record each activity at 1920x1080 resolution and 30 frames per second from the side of the track with markers placed every 5 meters.

Ground truth EE (average kCals/min) and strength classifications were estimated from the ActiGraph GT3X accelerometer data (sampling rate = 100Hz) using the Freedson Vector Magnitude 3 [16] algorithm. In our data, the EE ranges from 3.1 kCal/min to 18.27kCal/min.

3.2. Computer Vision Subsystems

3.2.1. Detection and Tracking

Detection was performed using the FasterRCNN [17] detector with a VGG16-base classifier [18] that is pre-trained with the Caltech pedestrian benchmark [19] and trained on 501 manually annotated samples from an in-house dataset collected by us around various parks in Delaware. Tracking is somewhat implicit, with most of our data including only the subject of interest in the frame. However, in some rare instances there are people who walk by, that would disrupt the implicit tracking assumption. To ensure robustness, we apply the DeepSORT [20, 21] algorithm to track individuals across frames.

3.2.2. Age and Gender Estimation

To perform age and gender estimation, first the face of the subject needs to be extracted. We use the Voila-Jones face detector [22] to detect faces in the "standing" action videos. The Deep EXpectation (DEX) [23] network, which is trained on the IMDB-WIKI dataset [23], is used to estimate age and gender.

3.2.3. Activity Recognition

We chose to implement activity recognition by using joint angles (specifically elbow and knee angles) and speed as features. The general idea is to create a reference (average) distribution of joint angles and speed for each activity class (sprinting, jogging, walking, standing), and then use a nearest neighbor approach to map features from new data to the closest model activity class.

To calculate the joint angles, the 2D skeletal pose of the subject is estimated in every frame using the Realtime Multi-Person Pose Estimation [24, 25] algorithm. From the ordering of joints from Realtime Multi-Person Pose Estimation [24, 25], joints 12-14 were used to calculate the knee angle, and joints 6-8 were used to calculate the elbow angle. The angle is calculated by

$$\mathbf{a} = \begin{bmatrix} pt1_x - pt2_x \\ pt1_y - pt2_y \end{bmatrix} \mathbf{b} = \begin{bmatrix} pt3_x - pt2_x \\ pt3_y - pt2_y \end{bmatrix} \theta = cos^{-1} \frac{\mathbf{a} \cdot \mathbf{b}}{|\mathbf{a}| \cdot |\mathbf{b}|}. \quad (1)$$

The collection of all joint angles from each frame of an activity video form a distribution. To compute distribution sim-

ilarity between the reference distribution and a new video's distribution, we use the Kolmogorov–Smirnov distance [26].

To calculate the speed, we use the distance each subject traveled (20m) in the activity videos and divided by the number of seconds the video contained. This distance is standardized across all videos and subjects and can be assumed. For the "standing" activity, the subject never moves and the distance traveled is zero. This distance over time gives an average speed for the subject for a given activity video. The average speed is compared against the reference distribution of speeds using a Mahalanobis distance [27].

Finally, to predict the activity, a linear combination of the joint angle and speed distance scores is used

$$D_{total} = \alpha_1 * D_{knee} + \alpha_2 * D_{elbow} + \alpha_3 * D_{speed}. \quad (2)$$

The α's were determined empirically, by trying all combinations of α's at a step size of 0.05 and reporting the best results. The best α's are $0, 0.75, 0.25$, respectively. This is surprising because the system found that both the elbow and knee angles are not needed, setting $alpha_1 = 0$. Thus, only the elbow angles and speed were used in the final version of the model.

The final activity is the one in which D_{total} is minimized.

3.3. Predicting PA Intensity and Energy Expenditure

A random forest regressor was trained to estimate the EE, and a random forest classifier was used to estimate the PA strength (light, moderate, vigorous) based on the above features extracted via computer vision. 20% of the data was used for testing, with 80% for training. The implementation of the models has the following parameters:

Classification model: NumTrees: 300, MinLeafSize : 1
Regression model: NumTress: 800, MinLeafSize: 5

4. ANALYSIS AND RESULTS

In this section, we measure the accuracy of each computer vision subsystem, measure the accuracy of the random forest regressor and classifier with ground truth features, and then, finally, measure the accuracy of the overall system. For each test, 80% was used for training, with 20% for testing, with the average 5-fold cross validation accuracy reported. Accuracy is reported as the number correct divided by the total.

The quantitative results for age, gender, and action recognition are given below. For detection, tracking, and face detection, we do not have the ground truth annotations for all frames in our dataset (there would be tens of thousands of frames to annotate). However, we did verify manually that these systems were correctly working and will submit supplementary videos showing this is the case.

Gender Acc	Age ≤ 5 Err	Age Mean Err	Age Std Err
77%	81%	3.2	3.9

Table 1. Age and gender prediction accuracy using Deep EXpectation (DEX).

4.1. Computer Vision Subsystems

4.1.1. Age and Gender Estimation

81% of the age estimates are within 5 years of the true age, with 57% of the estimates within 3 years of the true age. The maximum difference was 12 years, but that subject was wearing sunglasses. For gender estimation, the overall average accuracy was 77%. A few of the incorrectly labeled genders were scrunching up their face due to the wind. These numbers are summarized in Table 4.1.1

4.1.2. Action Recognition

We present the results of action recognition for using speed only, joint only, and then the final model described in Section 3. Confusion matrices are shown in Figure 2, with the overall accuracies 91.7%, 84.9%, and 96,8% respectively. For the joints-only model, the mistakes are made between jogging and sprinting. For the speed-only model, walking/jogging and jogging/sprinting are confused. The final model only makes one of each type of mistake. These results show that joint angles and speed are both useful features.

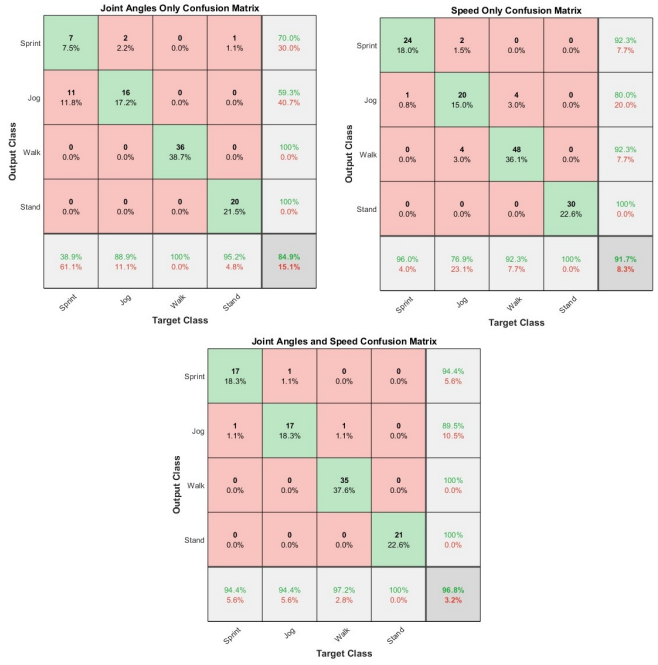

Fig. 2. Confusion matrices for action recognition using joint angles only, speed only, and the final combined model.

4.2. PA Strength and EE without Computer Vision using Ground Truth Features

This section tests the accuracy of the random forest classifier and regressor from ground truth age, gender, activity, and speed.

For strength classification, there are three categories: light, moderate, and vigorous. The dataset is made up of 27.94% light, 22.55% moderate, and 49.51% vigorous samples, which is mainly due to jogging and sprinting both mostly being vigorous intensity exercises. The overall average accuracy was 98.5%.

For EE regression, the mean, standard deviation, maximum, and median kCal/min difference are 1.23, 1.02, 4.55, and 1.03 respectively. These results are summarized in Figure 3.

Finally, we used the neighborhood component analysis algorithm [28] to analyze how important each feature is, with feature weights shown below in Figure 4. For regression of EE, the activity type and the age of the subject were the most important features. For classification of PA strength, the speed and activity were the most important features.

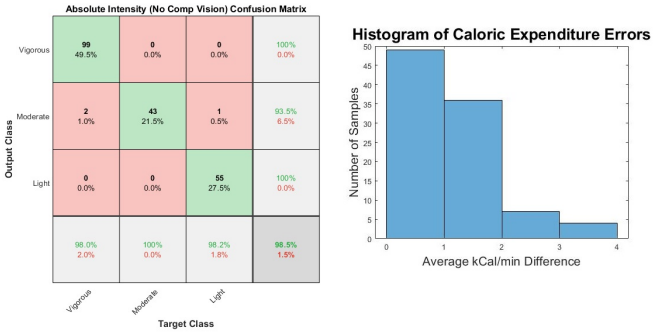

Fig. 3. PA strength classification and EE estimation from ground truth features using a random forest.

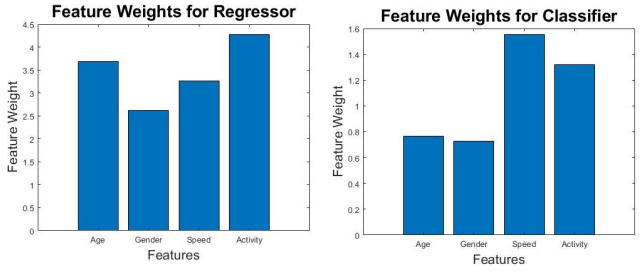

Fig. 4. Feature weights as calculated by the neighborhood component analysis for regression of EE and classification of PA strength

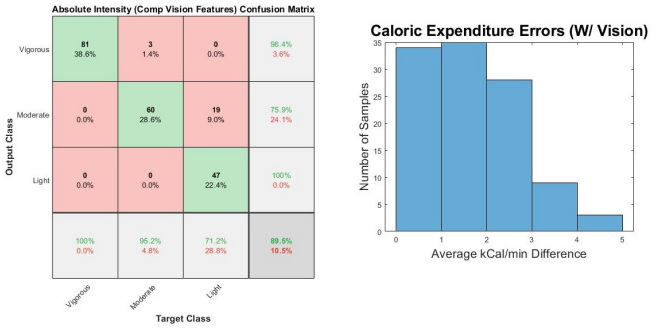

Fig. 5. PA strength classification and EE estimation from computer vision estimated features using a random forest.

4.3. PA Strength and EE using Computer Vision Features

This section tests the accuracy of the random forest on computer vision estimated age, gender, activity, and speed.

The same strength categories (light, moderate, vigorous) are used. The overall accuracy was 89.5%. The most errors came from mistakenly classifying a light strength workout as moderate. This makes sense as jogging and sprinting are vigorous exercises, while walking quickly can be on the boundary between light and moderate.

For EE regression, the mean, standard deviation, maximum, and median kCal/min differences are 1.96, 1.50, 5.83, 1.53 kCal/min, respectively. This means our system, on average on our dataset, can estimate the average number of calories burned by a subject within two calories.

5. CONCLUSION

On our dataset of 50 subjects performing four different activities, we are able to use computer vision to fully automate estimating the PA strength with 89.5% accuracy and EE within an average of 2 kCal/min. We use computer vision to estimate age, gender, activity type, and speed using Deep Expectation and Realtime Multi-Person Pose Estimation, and show through Neighborhood Component Analysis that these features provide value to the estimation of PA strength and EE. We present the accuracy of these individual components as well as the overall accuracy. The main limitation of this work is the lack of common dataset to compare against other methods. In the future, we plan to collect and annotate a much larger dataset to create a common ground for comparison and analysis. We also intend to add additional features to the computer vision system, such as weight and height.

6. ACKNOWLEDGEMENTS

This work was supported by the Center of Innovative Health Research at the University of Delaware.

7. REFERENCES

[1] Lindsay K. Campbell et al., "A social assessment of urban parkland: Analyzing park use and meaning to in-

form management and resilience planning," *Environmental Science & Policy*, vol. 62, pp. 34 – 44, 2016.

[2] A. Jain, "A smart gym framework: Theoretical approach," in *2015 IEEE International Symposium on Nanoelectronic and Information Systems*, Dec 2015, pp. 191–196.

[3] Shuhei Yamamoto et al., "The simple method for predicting metabolic equivalents using heart rate in patients with cardiovascular disease," *IJC Heart & Vasculature*, vol. 19, pp. 88–89, 06 2018.

[4] Douglas L. Ballor et al., "Comparison of three methods of estimating energy expenditure: Caltrac, heart rate, and video analysis," *Research quarterly for exercise and sport*, vol. 60, pp. 362–8, 01 1990.

[5] Shaopeng Liu et al, "Computational methods for estimating energy expenditure in human physical activities.," *Medicine and science in sports and exercise*, vol. 44 11, pp. 2138–46, 2012.

[6] Deepika Singh et al., "Human activity recognition using recurrent neural networks," in *Machine Learning and Knowledge Extraction*, Cham, 2017, pp. 267–274, Springer International Publishing.

[7] Andrew Peter Hills et al., "Assessment of physical activity and energy expenditure: An overview of objective measures," in *Front. Nutr.*, 2014.

[8] Shaoqing Ren et al., "Faster R-CNN: towards real-time object detection with region proposal networks," *CoRR*, vol. abs/1506.01497, 2015.

[9] Joseph Redmon and Ali Farhadi, "Yolov3: An incremental improvement," *arXiv*, 2018.

[10] Alex Bewley et al., "Simple online and realtime tracking," *2016 IEEE International Conference on Image Processing (ICIP)*, pp. 3464–3468, 2016.

[11] Yu Zhao et al, "Deep residual bidir-lstm for human activity recognition using wearable sensors," *CoRR*, vol. abs/1708.08989, 2017.

[12] Alessia Saggese et al., "Action recognition by learning pose representations," *CoRR*, 2017.

[13] P Silva et al., "Assessing physical activity intensity by video analysis," *Physiological Measurement*, vol. 36, no. 5, pp. 1037, 2015.

[14] Greg Welch and Gary Bishop, "An introduction to the kalman filter," Tech. Rep., Chapel Hill, NC, USA, 1995.

[15] Jordan et al. A. Carlson, "Automated ecological assessment of physical activity: Advancing direct observation," *International Journal of Environmental Research and Public Health*, vol. 14, pp. 1487, 12 2017.

[16] Kate Lyden et al., "A comprehensive evaluation of commonly used accelerometer energy expenditure and met prediction equations," *European Journal of Applied Physiology*, vol. 111, pp. 187–201, 2010.

[17] Shaoqing Ren et al., "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2015.

[18] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[19] "Caltech pedestrian detection benchmark," http://www.vision.caltech.edu/Image_Datasets/CaltechPedestrians/.

[20] Nicolai Wojke et al, "Simple online and realtime tracking with a deep association metric," in *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2017, pp. 3645–3649.

[21] Nicolai Wojke and Alex Bewley, "Deep cosine metric learning for person re-identification," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018, pp. 748–756.

[22] Paul Viola and Michael J. Jones, "Robust real-time face detection," *Int. J. Comput. Vision*, vol. 57, no. 2, pp. 137–154, May 2004.

[23] Rasmus Rothe et all, "Deep expectation of real and apparent age from a single image without facial landmarks," *International Journal of Computer Vision*, vol. 126, no. 2, pp. 144–157, Apr 2018.

[24] Zhe Cao et al., "Realtime multi-person 2d pose estimation using part affinity fields," in *CVPR*, 2017.

[25] Shih-En Wei et al, "Convolutional pose machines," in *CVPR*, 2016.

[26] M. A. Stephens, "Edf statistics for goodness of fit and some comparisons," *Journal of the American Statistical Association*, vol. 69, no. 347, pp. 730–737, 1974.

[27] P. C. Mahalanobis, "On the generalised distance in statistics," in *Proceedings National Institute of Science, India*, Apr. 1936, vol. 2, pp. 49–55.

[28] Jacob Goldberger et al., "Neighbourhood components analysis," in *Advances in Neural Information Processing Systems 17*, L. K. Saul, Y. Weiss, and L. Bottou, Eds., pp. 513–520. MIT Press, 2005.