sensors

MDPI

Article

Using Computer Vision to Annotate Video-Recoded Direct Observation of Physical Behavior

Sarah K. Keadle [1,*], Skylar Eglowski [2], Katie Ylarregui [1], Scott J. Strath [3], Julian Martinez [3], Alex Dekhtyar [4] and Vadim Kagan [2]

[1] Department of Kinesiology and Public Health, California Polytechnic State University, San Luis Obispo, CA 93407, USA; kylarreg@calpoly.edu
[2] Sentimetrix Inc., Bethesda, MD 20814, USA; skylar@sentimetrix.com (S.E.); kagan@sentimetrix.com (V.K.)
[3] College of Public Health, University of Wisconsin, Milwaukee, WI 53205, USA; sstrath@uwm.edu (S.J.S.); marti994@uwm.edu (J.M.)
[4] Department of Computer Science and Software Engineering, California Polytechnic State University, San Luis Obispo, CA 93407, USA; dekhtyar@calpoly.edu
* Correspondence: skeadle@calpoly.edu

Abstract: Direct observation is a ground-truth measure for physical behavior, but the high cost limits widespread use. The purpose of this study was to develop and test machine learning methods to recognize aspects of physical behavior and location from videos of human movement: Adults (N = 26, aged 18–59 y) were recorded in their natural environment for two, 2- to 3-h sessions. Trained research assistants annotated videos using commercially available software including the following taxonomies: (1) sedentary versus non-sedentary (two classes); (2) activity type (four classes: sedentary, walking, running, and mixed movement); and (3) activity intensity (four classes: sedentary, light, moderate, and vigorous). Four machine learning approaches were trained and evaluated for each taxonomy. Models were trained on 80% of the videos, validated on 10%, and final accuracy is reported on the remaining 10% of the videos not used in training. Overall accuracy was as follows: 87.4% for Taxonomy 1, 63.1% for Taxonomy 2, and 68.6% for Taxonomy 3. This study shows it is possible to use computer vision to annotate aspects of physical behavior, speeding up the time and reducing labor required for direct observation. Future research should test these machine learning models on larger, independent datasets and take advantage of analysis of video fragments, rather than individual still images.

Keywords: physical activity; sedentary behavior; assessment; direct observation; computer vision

Citation: Keadle, S.K.; Eglowski, S.; Ylarregui, K.; Strath, S.J.; Martinez, J.; Dekhtyar, A.; Kagan, V. Using Computer Vision to Annotate Video-Recoded Direct Observation of Physical Behavior. *Sensors* **2024**, *24*, 2359. https://doi.org/10.3390/s24072359

Academic Editors: Veronica Cimolin and Paolo Capodaglio

Received: 14 February 2024
Revised: 19 March 2024
Accepted: 28 March 2024
Published: 8 April 2024

1. Introduction

The numerous health benefits of a physically active lifestyle are well-established, and the implementation of devices to measure physical behavior in prospective cohorts and randomized clinical trials has provided important new insight into the type and amount of physical activity that is needed for optimal health benefits [1–3]. However, there is heterogeneity in data collection procedures and data processing methods (i.e., statistical algorithms applied to device signals to estimate aspects of physical behavior, such as time spent in moderate-intensity physical activity), which causes a great deal of confusion in the field as different studies are reaching different conclusions about the amount of physical activity for optimal health benefits [4–6]. In order for device data to provide consistent, accurate estimates of physical behavior, there is a need for rigorous validation studies that include data collection in a large number of participants, within naturalistic conditions and established ground-truth measures [7].

Video-recorded direct observation, which includes frame-by-frame analysis of a participant, is a ground-truth measure that can be used to validate many physical behavior metrics that are linked to health, including steps, postural transitions, behavior type

(e.g., housework, walking) and location (e.g., work, park) [8–12]. However, collecting and analyzing this ground-truth method is often prohibitively expensive due to the time, costs, and training required to manually annotate images, making its utility in large studies impractical [13]. Recent advances in machine learning technology as applied to computer vision have demonstrated the potential of automating image annotation. Using multi-layered special purpose neural networks (convolutional neural networks, recurrent neural networks), researchers have been able to accurately classify images based on what is depicted in them, recognize the position of objects of interest in an image, recognize humans in an image, and track objects (vehicles, humans) across multiple consecutive frames of a video [14–16].

Datasets for recognizing humans and for classifying on-screen human actions and activities abound, as does research on human action detection. Herath et al. [17] note in their survey of work on automated recognition of human activities that researchers interpret the notion of "action" to be detected in drastically different ways: from capturing atomic limb movements [18,19] to recognizing "simple motion patterns . . . lasting for a very short duration (order of tens of seconds)" [20] to attempting to build a hierarchical representation of the notion of "action" [21]. Yet, there is a difference between these varied views on what constitutes an action and the notion of physical behavior used in public health research, which are based on constructs and taxonomies that are defined by the American Time Use Survey, Compendium of Physical Activities and the Sedentary Behavior Research Network Consensus Terminology Project [22–24]. To our knowledge, none of the widely used datasets in computer vision research contain ground-truth annotations compatible with the standardized terminology in the fields of physical behavior and health, nor do they contain video footage taken during human activity studies in naturalistic conditions; nor have there been any open-source models trained on such taxonomies.

To date, the primary application of this computer vision to physical activity and health studies has focused on quantifying aspects of the built environment that may affect physical activity levels [25–28]. Carlson et al. demonstrated the utility of computer vision for evaluating aspects of the physical environment, including the number of people in a park and identifying how many were engaging in physical activity [26]. This work primarily has used two types of video or still image data—a first-person (or "ego-centric") perspective where the images reflect what the participant can see externally but do not contain their body within the image or a fixed camera where the camera stays in the same place, but people come in and out of view [26,29,30]. Limited work has been done using third-person images, where the full body of a single participant is shown continuously. The third-person point of view is particularly beneficial when video-recorded direct observation provides ground truth for device algorithm development because the participant's whole body can be seen, enabling detailed annotation of body posture transitions and activity types and locations within their natural environment [13].

The primary aim of this study was to develop and test machine learning methods that estimate physical behavior metrics from video-recorded direct observation that includes a single participant for a 2-h period. Specifically, we examined the following taxonomies that are aligned with consensus labels in physical activity and health research: (1) sedentary versus not; (2) general activity type (sedentary, mixed movement, walking and running); and (3) activity intensity (sedentary, light, moderate, and vigorous) [22,24].

2. Materials and Methods

Participants were recruited from San Luis Obispo, CA, and the surrounding communities through word of mouth and fliers. All participants read and signed an Informed Consent Document approved by the Cal Poly Institutional Review Board. Participants were aged between 18 and 59 years old and were generally healthy with no major orthopedic injuries that inhibited their ability to perform exercise and/or walk. Information including age (yrs), sex, race/ethnicity, height and weight, and physical activity status (response options range from 0—avoided walking or exertion to 7- ran more than 10 miles per week

or sport over 3 h per week in comparable activity) were assessed at the first study visit. In total, 26 participants consented and were enrolled with an average mean (SD) of 30.5 (11.5) years old and an average BMI of 24.6 (3.8) mg/m^2; 63% of participants were female.

Participants were scheduled for two, 2-h video-recorded direct observation (DO) sessions. Participants were recorded using a GoPro (GoPro, Inc., San Mateo, CA, USA) Hero 5 camera. Two observers were present during each session. The primary observer held the Hero 5 and the secondary observer recorded details of the DO session. Observers were trained to avoid interacting with or influencing the participant's behavior. For each session, participants were instructed to complete their normal daily activities. Each DO session was assigned one of five time-use-based categories to ensure a wide range of activities and settings were observed. The categories and general constraints were as follows: (1) work sessions took place in the participant's typical work setting. If their occupation was sedentary office work, they were instructed to take at least two breaks from their desk during the two-hour period; (2) household sessions took place in a participant's house, and they were instructed that at least 45 min of the session should be household or personal care activities such as meal preparation, gardening, house cleaning, caring for others, laundry, and/or dishwashing; (3) sedentary leisure consisted of the participant being seated and/or lying on furniture while performing a leisure activity such as watching television, playing a video game, and/or reading a book. They were instructed to take at least one break from sitting; (4) active leisure sessions consisted of participants spending at least 45 min in a leisure-time physical activity (e.g., walking, weightlifting, running, playing frisbee, sports, and hiking); and (5) community sessions consisted of observing a participant taking some form of transportation (e.g., bike, bus, car) from their initial environment to a new environment located in a community setting such as a grocery store, a sports stadium, a restaurant, or a local venue. The participant must have a reason for going to the new environment, such as buying groceries, meeting up with friends, or attending an event.

2.1. Direct Observation Annotation

When downloaded, the GoPro files were spliced into 20-min clips, which were edited into a single video using Adobe Premiere (Adobe, Inc., San Jose, CA, USA). The videos were annotated using the Noldus Observer XT 14 program (Noldus Information Technology, Wageningen, The Netherlands). Research assistants were trained using established protocols adapted for the use of video recording [11,13,31]. All assistants read a written manual and then completed 12 h of practice coding. Observers then coded two practice videos that included a range of behaviors, postures, and intensities using the full coding scheme and were required to obtain an intraclass correlation (ICC) >0.9. To score the actual study data, the coders used a multi-pass method. In the first pass, they coded for behavioral domains consistent with the American Time Use Survey [23]. In the second pass, observers coded for posture/movement (lying down, sitting/reclining, kneeling/squatting, stretching, standing, stand and move, walk, walk with load, running, biking, ascending stairs, descending stairs, muscle strengthening activities, and sport movement) and intensity (sedentary, light, moderate, and vigorous). The posture coding scheme was designed to align with the consensus taxonomy on physical behaviors, and intensity was defined using the Compendium of Physical Activity and SBRN Consensus Terminology Project as a reference [22,24]. A random sample of 20% of the videos were dual-coded by all coding staff, with high inter-rated reliability (ICC = 0.95).

The annotated data were exported from Noldus into Python (version 3.5.4) in event-based format and converted to second-by-second estimates. If multiple behaviors or postures occurred within the same second, these were labeled as transitions, and the behavior/posture that occurred for >50% of the second was assigned as the primary behavior/posture. For this initial proof of concept, the full annotation schemes were consolidated into the following taxonomies, and we trained and evaluated separate models for each of the taxonomies (See Figure 1).

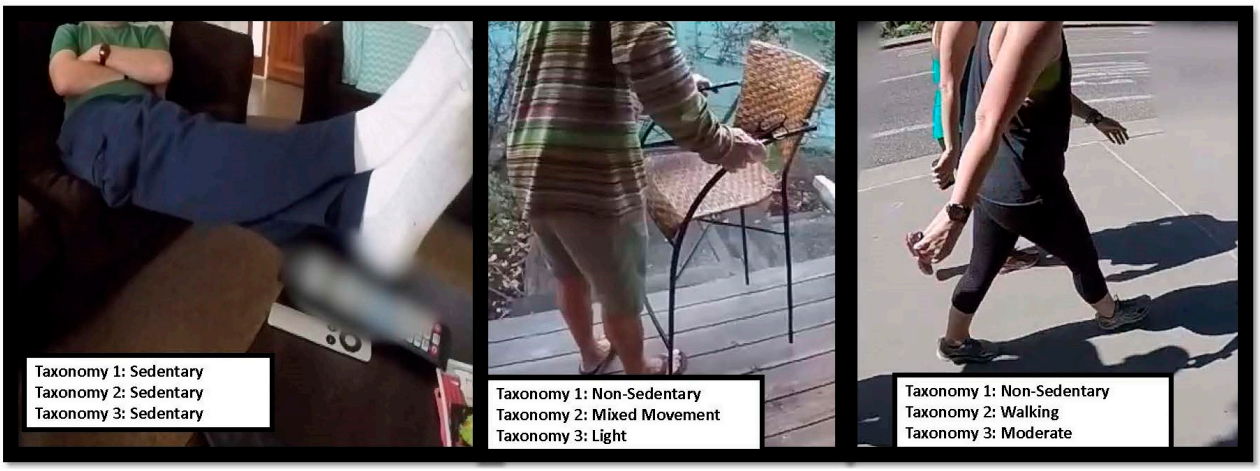

Figure 1. Sample images from videos with corresponding labels by taxonomy. Note: Taxonomy 1 is sedentary or not, Taxonomy 2 is activity type (four classes), and Taxonomy 3 is activity intensity (four classes).

1. Sedentary versus non-sedentary (two classes): Sedentary was considered sitting or lying, and all other postures/whole-body movements were non-sedentary.
2. Activity type (four classes): Sedentary was defined the same as Taxonomy 1, with the non-sedentary broken up into three new categories: walking, running, and mixed movement. Mixed movement covers all other postures/whole-body movements.
3. Activity intensity (four classes): Sedentary is low energy expenditure while seated, light is standing and <3 metabolic equivalents (METs), Moderate is 3–5.99 METs, and Vigorous is greater than or equal to 6 METS) [24,32].

2.2. Image Data Processing

Our dataset was composed of approximately 5000 min (81.5 h) of annotated activity filmed in 43 different videos. For each second of annotated video, we extracted a single representative frame for a total of 293,700 images. Videos are recorded at high resolution and fps (1080p with 30 fps) and then downscaled to 2 fps to align with ground truth. To fit the specific demanded resolution from our pre-trained models, either 224 × 224 or 384 × 384, we downscaled the longest edge to the desired resolution, keeping the aspect ratio intact. We then padded in the excess space with black pixels. This enabled the handling of a diverse number of orientations and input resolutions.

We then placed each input video into one of three categories: training, testing, or evaluation. These groups are referred to as "folds". The training fold constituted 80% of the data and was used to train the machine learning models. The testing fold constitutes 10% of the data. This set served multiple purposes related to the regular assessment of intermediate or prototype versions of each model. For example, the testing fold was used to help address "overfitting", which occurs when the model begins to form overly specific assumptions about incidental correlations in the training data that do not generalize well to unseen data. By periodically evaluating a deep learning model during its training, we can pinpoint an inflection point where the reported accuracy of the training fold stops increasing and begins to decrease. Once this trend is consistent, further training stops and reverts the model to the state it was at the inflection point. The third category is evaluation, which also made up 10% of the data. This data was held in reserve until, after sufficient experimentation, the results of the training fold were considered acceptable. The trained model was run against the evaluation fold, and we reported these results as an estimation of how the model would perform on an independent dataset, which was not seen in model training.

The standard but naive approach would be to shuffle all the frames together into these datasets, divided so that each class belonged in weighted proportion to each fold.

This works for many machine learning problems, but not this one. Suppose in video A, frame #3 belongs to the training fold, and frame #4 belongs to the evaluation fold. As those two frames are substantially similar in their visual content and likely have identical labels, this would create a false impression of the true performance of our model: the model might simply memorize frame #3 and report its label rather than make any difficult determinations. An additional challenge is that the dataset's videos are of varying lengths and do not feature a similar distribution of activities or intensities, so we cannot simply select random videos. Instead, we calculated the class distribution of all 79,507 possible video-fold assignments and compared the actual distribution of classes in the folds to the overall class distribution in the dataset. We then selected the distribution that has the smallest $\chi 2$ Distance [33], subject to a few additional constraints. First, we ensured that our splits had one example of each class in every fold possible, and second, we targeted the 80%/10%/10% fold splits. $\chi 2$ distance is a statistical measure of goodness of fit and is an effective measurement to compare the difference between a randomly sampled group and the overall population. The video-fold assignment selected is where the weighted distance of all three folds is minimized subject to the constraints used. This resulted in the class distributions shown in Table 1 for each of the taxonomies. Note that even though the distribution of videos in each taxonomy's respective folds happened to be very similar across all three taxonomies, this was just a coincidence: there is no requirement that a video needs to be in the same or different folds taxonomy-to-taxonomy. Note that because we only had two videos that depicted running, it was impossible to create a fold that satisfied the three-fold class constraint for running: for all metrics involving Taxonomy 2's running, we report the testing score rather than the evaluation score.

Table 1. Frame Class Distribution by Taxonomy and Fold.

Taxonomy 1: Sedentary					
Fold	**# Vid**	**Sedentary**	**Active**		
Training	32	116,759	111,184		
Testing	6	19,329	16,816		
Evaluation	5	14,618	14,994		
Taxonomy 2: Activity					
Fold	**# Videos**	**Sedentary**	**Mixed movement**	**Walking**	**Running**
Training	32	69,858	13,206	34,884	4090
Testing	6	1848	731	687	5106
Evaluation	5	2627	618	2099	0
Taxonomy 3: Intensity					
Fold	**# Videos**	**Sedentary**	**Light**	**Moderate**	**Vigorous**
Training	32	125,363	70,081	24,707	7551
Testing	6	15,325	13,427	1752	5228
Evaluation	5	10,009	18,519	1453	267

2.3. Model Selection

We developed four specialized machine learning models and applied them to each of the three taxonomies. The models are based on analysis of a single still image from a video. In our search for a baseline, we considered several prior studies that presented successful results when using machine learning techniques for physical activity detection [18,34,35]; however, none of these studies used the taxonomies we selected, which are aligned with standard definitions in physical activity and health research. Additionally, to evaluate the performance obtained from training the four models discussed in this paper, we considered a fifth model trained on one of the taxonomies (Taxonomy 1: sedentary vs. non-sedentary) as a benchmark to which the performance of other models for that taxonomy can be compared.

The four machine learning models that were evaluated for each taxonomy, three deep and one linear, are ResNet with Split Attention (ResNetSt) [36], Vision Transformer (ViT) [37], Convolutional Vision Transformer (CvT) [38], and XGBoost [39]. The first three are pre-trained deep learning models, which were fine-tuned on our dataset, while the XGBoost was the sole linear model we studied. The fine-tuning process, traditionally referred to as transfer learning, leveraged pre-trained open-source deep learning models that are provided by large tech companies that are trained on general object detection datasets. For each transfer learning exercise, we removed the final output layer of the pre-trained model, which does not correspond to our new target taxonomy labels. We replaced that output layer with an uninitialized output layer that was optimized on our own dataset and contained our target taxonomy labels (e.g., sedentary or non-sedentary). We then performed classic machine learning training, only adjusting the parameters in the output layer, leaving all the rest of the pre-trained architecture unmodified. Each model had slightly different expected input, but they all worked on raw video frames that had been cropped, resized, and normalized to a specific value range. In addition, we augmented our training data by randomly modifying it each epoch: randomly modifying lightness values (+/−10%), flipping the image horizontally, or randomly cropping within the frame itself. The purpose of this was to prevent the model from overfitting on many similar frames—particularly on sedentary examples where the subject and camera frame were relatively unchanged for long periods of time—and to allow the model to be more tolerant of different quality video or video perspectives. The utility of this technique has been thoroughly documented [40].

Our fifth benchmark Machine Learning Model is a simple ResNet50 convolutional neural network architecture [41] trained fully and completely from scratch on the images from the dataset used in this study and only on those. Since it was introduced in 2016 [41], ResNet50 has become one of the most popular CNN architectures for the analysis of still images. Trained fully and solely on the data from our study, this model serves as a good benchmark because it is sufficiently sophistcated, as it uses a well-known and successful convolutional neural network architecture. We use the benchmark model to illustrate the difficulty of learning the taxonomies studied in this paper. The benchmark model was trained for Taxonomy 1 (sedentary/non-sedentary). After reviewing its performance, we decided to skip training this benchmark model on the more complex Taxonomies 2 and 3 in order to save computing time—Section 3 discusses this decision in more detail.

Each model was trained three times using recommended parameters referenced from their respective source papers and evaluated on our testing fold after each epoch of the training fold. We selected the best epoch which had the best testing fold score on all three runs for each model, running as many epochs as required until we had 10 epochs without showing any improvement. For ViT, unexpectedly, the very first epoch performed the best, as it started to rapidly overfit the dataset. The other two models saw a more typical number of epochs before starting to overfit (5–10 epochs). Below is a very brief description of each model:

ResNetSt [36] was trained on Common Objects in Context (CoCo) [42], which contains 330k images with over 1.5 million tagged images and 250k tagged people. Based on the classic ResNet50 model [41], ResNet50 uses a chain of many convolutional layers, which allow the model to first identify simple shapes such as lines vs. circles and then to combine those into more complex shapes such as a nose or a cup and then finally learn the context around those objects, such as a face or a table. ResNetSt adds an Attention model, which effectively allows those earlier simple details to be forwarded to future layers rather than dropped entirely. The "Split" Attention has to do with only some of those said features being forwarded.

ViT [37]: This model was trained on ImageNet-21k [43], which has 14 million images with 21k different types of objects. The Transformer Architecture has become the current state of the art for Natural Language Processing (NLP) Taxonomies and has only recently been applied to vision taxonomies. ViT uses the NLP Transformer model by treating patches of

16 × 16 pixel data as something analogous to a "word". Rather than sentences or paragraphs, we have a picture that is simply a composition of many of these visual "words".

CvT [38]: Similarly to ViT, this model was trained on ImageNet-21k [43]. The ViT Architecture has a strict ordering of its words, which makes sense as grammar and semantics matter tremendously for conveying information in NLP taxonomies. However, when applied to pictures, the exact order of color patches does not really matter. A slight tilt of a camera, taking the picture in shadow, or capturing some leaves twisting in the background of the frame rarely factually changes the content of the image from a taxonomical perspective. To make the input less sensitive to these small perturbations, CvT re-introduces convolutional patches, which breaks the strict order-dependence of color patches as well as makes the model smaller, giving us the opportunity to train a higher number of epochs in the same amount of time as well as reducing the run time of this model when run by our users.

XGBoost [39] is a linear model built around gradient-boosted decision trees: many very small and simple machine learning models that all vote. The final determination is a weighted average of their collective output. The weights vary from class to class and tree to tree. For example, one of the trees may be a "one issue voter" that is very good at detecting a specific type of label in a specific context but has very little predictive power otherwise. The booster will assign a high value to that tree *only* when it detects a label it is good at detecting and more-or-less ignores it otherwise. Other trees may be more generalist voters, and so taking the average of their individual decisions to form a collective decision gives better accuracy than any one individual tree. XGBoost works from tabular data rather than pixel data; therefore, we first converted the frames to high-quality tabular features. We used AlphaPose [44] as a skeletal-extractor model, which detected people in the frames and output the location and orientation of their limbs. Since XGBoost is a linear model, it trains extremely quickly, giving us the ability to do a high number of epochs and thoroughly experiment with different parameter combinations.

ResNet50 [41] (our benchmark model) is a deep convolutional neural network model in which every three convolutional layers in a row are organized in the form of a residual network, i.e., a neural network (block) designed to predict the residual between the learned function $F(x)$ and the input x. He et al. showed that a residual network architecture learns faster and more accurately than a regular CNN [41]. ResNet50 works well as our benchmark model. It is a well-known neural network architecture that has been proven to be accurate in many settings and is compatible with image recognition. At the same time, to our knowledge, there are no publicly available pre-trained versions of ResNet50 on any of the taxonomies used in our study, giving us the opportunity to train a ResNet50 model from scratch using only the still frames from our dataset. Our benchmark model uses a freshly initialized 50-layer ResNet architecture with a learning rate of 0.001, trained for ten epochs, which are industry standard starting points for a ResNet architecture. We evaluated the model after each epoch against the test fold, saving a copy of its parameter weights. We then selected the copy that performed best on the test fold and evaluated it to our evaluation fold.

2.4. Statistical Evaluation

The Area Under the Receiver-Operator-Characteristic Curve (AUC) was used as the primary metric for selecting the final model. An AUC of 1.0 represents perfect discrimination, and 0.5 represents chance discrimination. The AUC visualizes and quantifies the trade-off between too many false positives and too many false negatives. The AUC works well with imbalanced class distributions. We defined the "optimal" decision point as simply the one with the fewest cumulative sum of false positives and false negatives for each classification taxonomy. By default, the AUC only works on the binary decision taxonomy. To convert it to our taxonomies with more than two classes, we used a technique called "One Versus Rest". We broke the problem into multiple subproblems considering only one class at a time, which were called the "positive class". We created the Receiver-Operator-

Characteristic Curve for each positive class and summed up all the other class predictions as just one generic negative class. We then took the average of all these sub-AUCs to get our averaged AUC. This technique made the evaluation of each class in a taxonomy an independent affair and allowed us to substitute the results of detecting the "running" class in Taxonomy 2 in the evaluation dataset with the results from the test dataset without affecting all other results.

For the evaluation fold, we present confusion matrices with our frame-by-frame hand-annotated ground truth compared to the predicted class. A perfectly accurate model would only have numbers along the diagonal, and non-diagonal values indicate errors. To help summarize each table, we also calculated the precision and recall. Precision ((true positives/(true positive + false positive)) is a measure of how often the predicted value is correct for a *specific* class. Recall (true positive/(true positive + false negative) is the percentage of examples in a specific class that were correctly detected. High precision but low recall indicates that the model has a high false negative rate, but when it does report a class, it is almost always correct. Low precision but high recall indicates that the model is oversensitive and has a high false positive rate. This can also be common in imbalanced problems; suppose you have 10,000 examples of Class A but only 100 examples of Class B. You correctly identify every Class B, but you incorrectly identify 1% of Class A as Class B. Class A was much more common, and 1% of its examples were equal to the total amount of Class B instances. Therefore, in this example, it would have 100% Precision and 99% Recall for Class A, 50% Precision and 100% Recall for Class B, and a total accuracy of 99%. We also calculated the F1 statistic, which is a weighted harmonic mean of precision and recall ($2 \times$ [precision $\times$ recall]/[precision + recall]). The F1 provides an overall estimate, particularly for situations with an uneven class distribution. The F1 metric takes values between 0 and 1, with values closer to 1 being better and a value of exactly one corresponding to perfect precision and recall. For multiple class problems, we report the weighted precisions, accuracies, and F1 scores. As previously mentioned, since we only had two videos that depicted running, it was impossible to create a fold that satisfied the three-fold class constraint for running: for all metrics involving Taxonomy 2′s running, we report the testing score rather than the evaluation score. All analyses were done in Python, using JupyterLab as our primary visualization tool.

3. Results

3.1. Deep Learning Model Comparison and Selection

The results of the evaluation of each model are summarized in Table 2. Overall, ViT performance was consistent across taxonomies; however, it is the largest model requiring 13 GB of vRAM. XGBoost was by far the fastest model to train once AlphaPose's feature extraction has been completed and does acceptable on Taxonomy 1 and Taxonomy 3 (where it performs the best), but on Taxonomy 2 it underperforms both Transformer models. However, AlphaPose extraction is approximately 3:1, making this method overall slower (0.3 frames per sec [fps]) (video length to wall time). ResNetSt did not perform well on Taxonomies 2 and 3, was slightly behind the other three models on Taxonomy 1, and was approximately eight times slower than CVT (8 fps). CVT performed best on the simplest taxonomy and was within 2% of the best-performing models or the other two taxonomies, and it is comparatively fast during both training and inference (64 fps), making it a compelling option as a single model to be deployed on off-the-shelf computers. The final model selection was based on the AUCs, as this showcases approximately how reliable signals each classifier is loading from compared to random chance and can be adapted to work well for multiclass problems. For Taxonomy 1 (sedentary vs. non-sedentary) the final model selected was CvT (AUC = 92.1%, which is a 41.1% lift over the benchmark). For Taxonomy 2 (Activity type), the final model selected was ViT (AUC = 72.3%). For Taxonomy 3 (Activity Intensity) the final model selected was XGB (AUC = 73.4%).

Table 2. Comparison of Area Under the Curve for Four Machine Learning Models by Taxonomy.

Taxonomy	ResNet50	RNSt	ViT	CvT	XGB
1: Sedentary vs. non-sedentary	51.0	88.9	90.1	**92.1**	90.8
2: Activity type	N/A	47.4	**72.3**	71.4	65.8
3: Activity intensity	N/A	71.6	70.1	71.9	**73.4**

Note: Bold font indicates best performing model for each taxonomy. ResNet50 is the benchmark model that was only applied to taxonomy 1, so results are N/A for other taxonomies.

3.2. Taxonomy-Specific Model Performance

To break down the performance of each model on a class-by-class basis, we show confusion matrices in Table 3 for the final model selected for each taxonomy, along with class-specific precision and recall (Table 4).

Table 3. Confusion Matrices by Taxonomy for Evaluation Phase.

Taxonomy 1: Sedentary vs. Non-Sedentary							
Label\Predicted	**Sedentary**	**Active**			**Precision**	**Recall**	**F1**
Sedentary	**11,926**	2692			0.92	0.82	0.87
Active	1026	**13,968**			0.84	0.93	0.88
Taxonomy 2: Activity Type							
Label\Predicted	**Sedentary**	**Mixed Movement**	**Walking**	**Running**	**Precision**	**Recall**	**F1**
Sedentary	**2338**	0	283	6	0.65	0.89	0.75
Mixed movement	307	**758**	381	7	0.53	0.52	0.53
Walking	928	0	**1166**	5	0.34	0.56	0.42
Running	2	663	1580	**2861**	0.99	0.56	0.72
Taxonomy 3: Activity Intensity							
Label\Predicted	**Sedentary**	**Light**	**Moderate**	**Vigorous**	**Precision**	**Recall**	**F1**
Sedentary	**13,259**	4915	323	22	0.90	0.72	0.80
Light	197	**939**	115	14	0.11	0.74	0.20
Moderate	1253	2344	**6352**	60	0.92	0.63	0.75
Vigorous	2	83	110	**72**	0.43	0.27	0.33

Note: Values are video frames. The running results come from our testing fold and are heavily overweighted, making up 49% of the confusion matrix values despite being a class weight of only 3% of the total dataset. Therefore, we have suppressed the running AUC weight by 94%.

Table 4. Weighted Average Results for Each Taxonomy for Best-Performing Models.

	Precision	Recall	F1	Accuracy
Taxonomy 1: Sedentary/not: Benchmark model	0.50	0.66	0.31	56.3%
Taxonomy 1: Sedentary/not	0.88	0.91	0.87	87.4%
Taxonomy 2: Activity type	0.73	0.63	0.65	63.1%
Taxonomy 3: Activity intensity	0.87	0.69	0.75	68.6%

Note: The benchmark model was ResNet50 and was trained only on our images. The results for the other three taxonomies were the best-performing of the four trained models. For Taxonomy 1, this was CvT. For Taxonomy 2, this was ViT. For Taxonomy 3, this was XGB.

Taxonomy 1 (Sedentary vs. Non-sedentary) was the simplest taxonomy, and overall performance was high, with an overall AUC of 92.1 and 87.4% accuracy. Weighed precision was 0.88, recall was 0.91, and F1 score was 0.87 (Table 4). There was a slight bias toward the active class despite it being the minority class.

Taxonomy 2 was the activity type, and performance for this four-class model was lower than Taxonomy 1, with an overall AUC of 72.3 and accuracy of 63.1% (Table 4). Similar to Taxonomy 1, there were misclassification errors where we classified sedentary

activity as active or active as sedentary (Table 3). Differentiation between mixed movement versus walking was strong (walking was never misclassified as mixed movement, though about 26% of mixed movement does get misclassified as walking). Running, however, often gets misclassified as the other movement classes, though that is not surprising given its small class representation (1.8% of the testing dataset). Weighted precision was 0.73, recall was 0.63, and F1 was 0.65 (Table 4).

For Taxonomy 3 (activity intensity), model performance was 73.4 overall AUC and an overall accuracy of 68.6%. Weighted precision was 0.87, recall was 0.68, and F1 score was 0.75 (Table 4). There was a mostly natural error curve where egregious errors (mistaking vigorous as sedentary or vice-versa) are rare, but misclassifying as an adjacent class happens more often. We note a similar pattern as Taxonomy 2 where we tend to confuse vigorous activity (such as running) with more moderate activity (such as walking; Table 3).

Tables 2 and 4 also show the performance of the benchmark model (ResNet50 trained only on our dataset and not pre-trained on any external datasets). On the two-class sedentary-non-sedentary taxonomy, the benchmark model exhibited a 56.25% validation accuracy, which is only slightly better than random. The micro-averaged f1 score was 0.53; as this was a binary classifier, its precision and recall were also 0.53. The AUC metric (Table 2) was 51%—not much better than random choice. After observing these results, we elected not to proceed with training the benchmark model on the remaining two taxonomies, as it is clear from comparing the benchmark model to our four other models on Taxonomy 1, the simplest to detect taxonomy, that the benchmark is severely underperforming, and is doing barely better than random choice. In our experiments, which typically consider per-class and multiclass values, these values will differ. Of note, the specialized models performed significantly better for the same task.

4. Discussion

This study successfully uses deep learning models trained on continuously recorded third-person videos to predict details about body posture, activity type, and intensity of physical behavior using ground-truth classifications relevant to physical activity and health [22,24]. This is the first known study to apply machine learning methods for computer vision to consensus labels in physical activity research, including the Compendium of Physical Activities and Sedentary Behavior Research Network Consensus Terminology Project. We found a simple classification of sedentary versus non-sedentary had high accuracy (87%) that was comparable to the inter-rater agreements for the gold-standard of human annotation [13,26]. As the complexity of the classification problem increased (i.e., to four activity types and four intensity categories), the overall accuracy declined, but overall, this proof-of-concept application of computer vision to automatically annotate video-recorded direct observation is promising and warrants future research to examine methods to improve accuracy.

The results from the present study are comparable to those of Carlson et al. [26], who used computer vision to estimate measures of a direct observation tool System for Observing Play and Recreation in Communities (SOPARC), which is a momentary direct observation system that includes periodic environmental scans of a particular setting (e.g., park) and documenting how many people are in the setting at a given time and the current activity level (sedentary, walking, or vigorous) [45–47]. They found concordance correlation coefficients that were considered moderate to good, ranging from 0.55 for the number active in a scene to 0.88 for the number of people in the scene, as compared to manual annotation [26]. Renzai et al. estimated sit-to-stand transitions and sitting, standing, and walking in a sample of Parkinson's Disease patients completing scripted taxonomies in a laboratory setting. They reported an overall weighted accuracy of 84.0% in their test fold [34]. In the present study, under less controlled conditions and more detailed aspects of physical behavior, we observed accuracies that align with previous studies, with results ranging from 63% to 87%.

Other applications of computer vision in physical activity and health research have focused on quantifying aspects of the built environment that are important for physical activity promotion [48]. Adams et al. found that a deep learning approach applied to Google Street View was able to classify microscale features of pedestrian streetscapes (e.g., sidewalks, walk signals) with >84% accuracy compared to human annotation [25]. Yue et al. [49] also applied convolutional neural networks to Google Street View and found validation accuracy of >82% for environmental features and subsequently linked those predicted built environment features with chronic conditions and mental health [48]. Given the diminishing cost of image/video data collection and storage, image-based analyses may provide important contextual information for obesity-related research, including nutrition and physical activity [48]. Cameras have been used to quantify the extent to which an environment is "obesogenic" [50], objectively assess travel patterns [51], active transport to school policies [52], and conduct environmental audits of the built environment [28]. Additionally, computer vision has been used to count repetitions in body-weight training exercises (i.e., squatting, burpees, push-ups, sit-ups, and jumping jacks) with high accuracy (>85%) [35]. Collectively, the present study and the previous research suggest a need to continue to develop machine learning methods to automatically annotate images/videos. To date, advancement in algorithm development for wearable devices to predict physical behavior has been hindered by a lack of commonly labeled data and the excessive cost of collecting and manually training direct observation data. The potential utility of computer vision to automate aspects of data annotation would dramatically increase the feasibility, scale, and efficiencies of future research in this area.

Strengths of this study include a protocol that included third-person image data and ground-truth measures that are labeled consistent with consensus terminology in the field of physical behavior assessment [22–24]. Previous work in this area has focused on first-person or "ego-centric" cameras or video recordings, which have many strengths but provide less detail on specific body posture and movements (e.g., steps) that may be important for algorithm development metrics like steps or body posture that are needed to train wearable sensor algorithms [28–30,51–56]. Limitations include data from one geographical region and, although efforts were made to ensure a range of activities, the class distributions were unbalanced. Future research should collect additional data to ensure more balanced distributions and that more detailed taxonomies can be evaluated. The ground truth is human-labeled data, and although the inter-rater reliability was high (ICC = 0.95), there is the possibility of human errors or inconsistencies. This work was conducted on data collected with the explicit permission of the study subjects and for the purpose of analyzing similarly collected videos generated and used in research contexts with collection and use protocols approved by the respective Institutional Review Boards (IRBs). There are established ethical frameworks for the collection of camera data in research contexts [57], but this type of research is not without risk [58], and we advise that future use of these and derivative modeling techniques should always be subject to the appropriate rules and regulations and IRB approval.

There are several clear avenues of improvement: first and foremost, collecting additional annotated data will enhance the robustness and generalizability of our results. We attribute at least part of the errors we reported to the underrepresentation of certain classes in our training data (see Table 1). Additional data collection is currently underway and involves videos from multiple research groups. This additional data collection will allow us to diversify the training data both in terms of physical activity representation but also in terms of the demographics of the subjects and the variety of contexts in which they are filmed (observations conducted in coastal California are unable to represent snow as the setting, for example). Future research will consider refinement of feature selection. Our current work uses image pixels and AlphaPose "stick figure" representations of the subject in the video as the only types of features. We are planning to explore additional pose detection feature extraction options, such as Google's MediaPipe [59]. Other feature extensions discussed below will involve the use of more than one still frame per observation. There

are several improvements to the machine learning models that we are considering. First, for the more complex classification in Taxonomies 2 and 3 (activity type and intensity), merging the AlphaPose features into the Vision Transformers as a secondary input seems promising and has been suggested in similar video analysis problems [60]. Second, we are planning to investigate the effect of analyzing short video fragments instead of individual still images. This will entail making the input recurrent: that is, rather than only using a single frame, we plan to use a collection of frames: for example, the previous 4 s and 1 s into the future (if used in a non-real-time application). This could be done by using a (i) fixed convolution on the input being fed into the existing pipelines, or (ii) using a ConvLSTM [61] model that was designed for pooling together sequences of similar images that are taken at different time stamps, or (iii) using a ViViT [60] which is a Video Vision Transformer with a single label, or (iv) having a sequence of convolutional layers in addition to the existing Transformer models that are eventually merged into the final output layer. With more information available, models trained on short video fragments may show improved accuracy, potentially at the cost of runtime performance (both for training and for using a trained model to annotate a full video), as such models ingest significantly more input. Finally, with multiple trained models that use different inputs, we plan to investigate a variety of ensemble techniques to see if we can improve model accuracy by using multiple classifiers (we note parenthetically that some of our methods like XGBoost can already be considered a form of an ensemble classifier, but in this particular case, we are talking about combining results of multiple different trained models).

There are many advantages to using computer vision to label video recordings of physical behavior, including increased scalability and potentially more consistency compared to manual annotation by human observers. The present study was the first to apply computer vision to established taxonomies in physical behavior research within naturalistic environments. Although the accuracy for Taxonomy 1 (Sedentary/Non-sedentary) was comparable to human annotation, future research that includes additional model development and validation on larger and more diverse datasets is needed to enhance the accuracy and generalizability of the models for the more complex taxonomies. For optimal use of computer visions as ground-truth measures, researchers will likely need some way to evaluate the performance of the machine learning models on a frame-by-frame basis and the ability for researchers to "correct" errors in the predictions, such as within an annotation software. This application of computer vision to classify aspects of posture, activity type, and intensity from a third-person video demonstrates a promising proof of concept that should be explored in future research.

Author Contributions: Conceptualization, S.K.K., A.D., S.E. and V.K.; methodology, S.K.K., K.Y., S.J.S., J.M., S.E., A.D. and V.K.; validation, S.E. and A.D.; formal analysis, S.E.; data collection, J.M., K.Y. and S.K.K.; writing—original draft preparation, S.K.K. and S.E.; writing—review and editing, S.K.K., K.Y., S.J.S., J.M., S.E., A.D. and V.K.; funding acquisition, S.K.K., A.D. and V.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Institutes of Health, Small Business Innovation Research contract 75N91023C00031.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Review Board of California Polytechnic State University (protocol code 2018-184 and July 2017).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Due to the participant being identifiable in the videos, a limited subset of data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results. Authors Mr. Kagan and Mx. Eglowski are employed by the company SentiMetrix Inc. The remaining

authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Piercy, K.L.; Troiano, R.P.; Ballard, R.M.; Carlson, S.A.; Fulton, J.E.; Galuska, D.A.; George, S.M.; Olson, R.D. The Physical Activity Guidelines for Americans. *JAMA* **2018**, *320*, 2020–2028. [CrossRef] [PubMed]
2. Ekelund, U.; Tarp, J.; Fagerland, M.W.; Johannessen, J.S.; Hansen, B.H.; Jefferis, B.J.; Whincup, P.H.; Diaz, K.M.; Hooker, S.; Howard, V.J.; et al. Joint associations of accelero-meter measured physical activity and sedentary time with all-cause mortality: A harmonised meta-analysis in more than 44 000 middle-aged and older individuals. *Br. J. Sports Med.* **2020**, *54*, 1499–1506. [CrossRef] [PubMed]
3. Paluch, A.E.; Bajpai, S.; Bassett, D.R.; Carnethon, M.R.; Ekelund, U.; Evenson, K.R.; Galuska, D.A.; Jefferis, B.J.; Kraus, W.E.; Lee, I.M.; et al. Daily steps and all-cause mortality: A meta-analysis of 15 international cohorts. *Lancet Public Health* **2022**, *7*, e219–e228. [CrossRef] [PubMed]
4. Evenson, K.R.; Wen, F.; Herring, A.H. Associations of Accelerometry-Assessed and Self-Reported Physical Activity and Sedentary Behavior with All-Cause and Cardiovascular Mortality among US Adults. *Am. J. Epidemiol.* **2016**, *184*, 621–632. [CrossRef] [PubMed]
5. Matthews, C.E.; Keadle, S.K.; Berrigan, D.; Lyden, K.; Troiano, R.P. Influence of Accelerometer Calibration Approach on Moderate-Vigorous Physical Activity Estimates for Adults. *Med. Sci. Sports Exerc.* **2018**, *50*, 2285–2291. [CrossRef] [PubMed]
6. Migueles, J.H.; Cadenas-Sanchez, C.; Rowlands, A.V.; Henriksson, P.; Shiroma, E.J.; Acosta, F.M.; Rodriguez-Ayllon, M.; Esteban-Cornejo, I.; Plaza-Florido, A.; Gil-Cosano, J.J.; et al. Comparability of accelerometer signal aggregation metrics across placements and dominant wrist cut points for the assessment of physical activity in adults. *Sci. Rep.* **2019**, *9*, 18235. [CrossRef] [PubMed]
7. Keadle, S.K.; Lyden, K.A.; Strath, S.J.; Staudenmayer, J.W.; Freedson, P.S. A Framework to Evaluate Devices that Assess Physical Behavior. *Exerc. Sport. Sci. Rev.* **2019**, *47*, 206–214. [CrossRef] [PubMed]
8. Toth, L.P.; Park, S.; Springer, C.M.; Feyerabend, M.D.; Steeves, J.A.; Bassett, D.R. Video-Recorded Validation of Wearable Step Counters under Free-living Conditions. *Med. Sci. Sports Exerc.* **2018**, *50*, 1315–1322. [CrossRef] [PubMed]
9. Lyden, K.; Keadle, S.K.; Staudenmayer, J.; Freedson, P.S. The activPAL™ Accurately Classifies Activity Intensity Categories in Healthy Adults. *Med. Sci. Sports Exerc.* **2017**, *49*, 1022–1028. [CrossRef]
10. Lyden, K.; Keadle, S.L.K.; Staudenmayer, J.W.; Freedson, P.S. Validity of Two Wearable Monitors to Estimate Breaks from Sedentary Time. *Med. Sci. Sports Exerc.* **2012**, *44*, 2243–2252. [CrossRef]
11. Keadle, S.K.; Patel, S.; Berrigan, D.; Christopher, C.N.; Huang, J.; Saint-Maurice, P.F.; Loftfield, E.; Matthews, C.E. Validation of ACT24 Version 2.0 for Estimating Behavioral Domains, Active and Sedentary Time. *Med. Sci. Sports Exerc.* **2023**, *55*, 1054–1062. [CrossRef] [PubMed]
12. Cox, M.F.; Petrucci, G.J., Jr.; Marcotte, R.T.; Masteller, B.R.; Staudenmayer, J.; Freedson, P.S.; Sirard, J.R. A Novel Video-Based Direct Observation System for Assessing Physical Activity and Sedentary Behavior in Children and Young Adults. *J. Meas. Phys. Behav.* **2020**, *3*, 50–57. [CrossRef]
13. Keadle, S.K.; Martinez, J.; Strath, S.J.; Sirard, J.R.; John, D.; Intille, S.; Argello, D.; Amalbert-Birriel, M.; Barnett, R.K.; Thapa-Chhetry, B.; et al. Evaluation of within and between site agreement for direct observation of physical behavior across four research groups. *J. Meas. Phys. Behav.* **2023**, *6*, 176–184. [CrossRef]
14. Weinland, D.; Ronfard, R.; Boyer, E. A survey of vision-based methods for action representation, segmentation and recognition. *Comput. Vis. Image Underst.* **2011**, *115*, 224–241. [CrossRef]
15. Liang, M.; Hu, X. Recurrent convolutional neural network for object recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3367–3375.
16. Van Oord, A.; Kalchbrenner, N.; Kavukcuoglu, K. Pixel Recurrent Neural Networks. In Proceedings of the International Conference on Machine Learning, New York, NY, USA, 19–24 June 2016; pp. 1747–1756.
17. Herath, S.; Harandi, M.; Porikli, F. Going deeper into action recognition: A survey. *Image Vis. Comput.* **2017**, *60*, 4–21. [CrossRef]
18. Poppe, R. A survey on vision-based human action recognition. *Image Vis. Comput.* **2010**, *28*, 976–990. [CrossRef]
19. Moeslund, T.B.; Granum, E. A survey of advances in vision-based human motion capture and analysis. *Comput. Vis. Image Underst.* **2006**, *104*, 90–127. [CrossRef]
20. Turaga, P.; Chellappa, R.; Subrahmanian, V.S.; Udrea, O. Machine recognition of human activities: A survey. *IEEE Trans. Circuits Syst. Video Technol.* **2008**, *18*, 1472–1488. [CrossRef]
21. Chaaraoui, A.A.; Climent-Perez, P.; Francisco Florez-Revuelta, F. A review on vision techniques applied to human behaviour analysis for ambient-assisted living. *Expert. Syst. Appl.* **2012**, *39*, 10873–10888. [CrossRef]
22. Tremblay, M.S.; Aubert, S.; Barnes, J.D.; Saunders, T.J.; Carson, V.; Latimer-Cheung, A.E.; Chastin, S.F.M.; Altenburg, T.M.; Chinapaw, M.J.M.; Project, S.T.C. Sedentary Behavior Research Network (SBRN)—Terminology Consensus Project process and outcome. *Int. J. Behav. Nutr. Phy* **2017**, *14*, 75. [CrossRef]
23. Bureau of Labor and Statistics: American Time Use Survey. Available online: http://www.bls.gov/tus/tables.htm (accessed on 1 December 2023).

24. Ainsworth, B.E.; Haskell, W.L.; Herrmann, S.D.; Meckes, N.; Bassett, D.R., Jr.; Tudor-Locke, C.; Greer, J.L.; Vezina, J.; Whitt-Glover, M.C.; Leon, A.S. 2011 Compendium of Physical Activities: A second update of codes and MET values. *Med. Sci. Sports Exerc.* **2011**, *43*, 1575–1581. [CrossRef] [PubMed]
25. Adams, M.A.; Phillips, C.B.; Patel, A.; Middel, A. Training Computers to See the Built Environment Related to Physical Activity: Detection of Microscale Walkability Features Using Computer Vision. *Int. J. Environ. Res. Public Health* **2022**, *19*, 4548. [CrossRef] [PubMed]
26. Carlson, J.A.; Liu, B.O.; Sallis, J.F.; Hipp, J.A.; Staggs, V.S.; Kerr, J.; Papa, A.; Dean, K.; Vasconcelos, N.M. Automated High-Frequency Observations of Physical Activity Using Computer Vision. *Med. Sci. Sports Exerc.* **2020**, *52*, 2029–2036. [CrossRef] [PubMed]
27. Cleland, C.; Reis, R.S.; Ferreira Hino, A.A.; Hunter, R.; Fermino, R.C.; Koller de Paiva, H.; Czestschuk, B.; Ellis, G. Built environment correlates of physical activity and sedentary behaviour in older adults: A comparative review between high and low-middle income countries. *Health Place* **2019**, *57*, 277–304. [CrossRef] [PubMed]
28. Oliver, M.; Doherty, A.R.; Kelly, P.; Badland, H.M.; Mavoa, S.; Shepherd, J.; Kerr, J.; Marshall, S.; Hamilton, A.; Foster, C. Utility of passive photography to objectively audit built environment features of active transport journeys: An observational study. *Int. J. Health Geogr.* **2013**, *12*, 20. [CrossRef] [PubMed]
29. Doherty, A.R.; Kelly, P.; Kerr, J.; Marshall, S.; Oliver, M.; Badland, H.; Hamilton, A.; Foster, C. Using wearable cameras to categorise type and context of accelerometer-identified episodes of physical activity. *Int. J. Behav. Nutr. Phys. Act.* **2013**, *10*, 22. [CrossRef] [PubMed]
30. Doherty, A.R.; Moulin, C.J.; Smeaton, A.F. Automatically assisting human memory: A SenseCam browser. *Memory* **2011**, *19*, 785–795. [CrossRef] [PubMed]
31. Kozey-Keadle, S.; Libertine, A.; Lyden, K.; Staudenmayer, J.; Freedson, P.S. Validation of Wearable Monitors for Assessing Sedentary Behavior. *Med. Sci. Sports Exerc.* **2011**, *43*, 1561–1567. [CrossRef] [PubMed]
32. Piercy, K.L.; Troiano, R.P. Physical Activity Guidelines for Americans from the US Department of Health and Human Services. *Circ. Cardiovasc. Qual. Outcomes* **2018**, *11*, e005263. [CrossRef]
33. Dodge, Y. Chi-Square Distance. In *The Concise Encyclopedia of Statistics*; Springer: New York, NY, USA, 2008; pp. 68–70.
34. Rezaei, B.; Christakis, Y.; Ho, B.; Thomas, K.; Erb, K.; Ostadabbas, S.; Patel, S. Target-Specific Action Classification for Automated Assessment of Human Motor Behavior from Video. *Sensors* **2019**, *19*, 4266. [CrossRef]
35. Ferreira, B.; Ferreira, P.M.; Pinheiro, G.; Figueiredo, N.; Carvalho, F.; Menezes, P.; Batista, J. Deep learning approaches for workout repetition counting and validation. *Pattern Recognit. Lett.* **2021**, *151*, 259–266. [CrossRef]
36. Zhang, H.; Wu, C.; Zhang, Z.; Zhu, Y.; Lin, H.; Zhang, Z.; Sun, Y.; He, T.; Mueller, J.; Manmatha, R.; et al. ResNeSt: Split-Attention Networks. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), New Orleans, LA, USA, 19–20 June 2022; pp. 2735–2745.
37. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S. An image is worth 16 × 16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
38. Wu, H.; Xiao, B.; Codella, N.; Liu, M.; Dai, X.; Yuan, L.; Zhang, L. Cvt: Introducing convolutions to vision transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual, 11–17 October 2021; pp. 22–31.
39. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
40. Teichmann, M.; Thoma, M.; Lee, J. Tensorvision. Available online: https://github.com/TensorVision/TensorVision (accessed on 1 October 2023).
41. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
42. Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014; Part V 13. pp. 740–755.
43. Ridnik, T.; Ben-Baruch, E.; Noy, A.; Zelnik-Manor, L. Imagenet-21k pretraining for the masses. *arXiv* **2021**, arXiv:2104.10972.
44. Fang, H.-S.; Li, J.; Tang, H.; Xu, C.; Zhu, H.; Xiu, Y.; Li, Y.-L.; Lu, C. Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 7157–7173. [CrossRef]
45. McKenzie, T.L. Context Matters: Systematic Observation of Place-Based Physical Activity. *Res. Q. Exerc. Sport.* **2016**, *87*, 334–341. [CrossRef]
46. Cohen, D.A.; Setodji, C.; Evenson, K.R.; Ward, P.; Lapham, S.; Hillier, A.; McKenzie, T.L. How much observation is enough? Refining the administration of SOPARC. *J. Phys. Act. Health* **2011**, *8*, 1117–1123. [CrossRef] [PubMed]
47. Ward, P.; McKenzie, T.L.; Cohen, D.; Evenson, K.R.; Golinelli, D.; Hillier, A.; Lapham, S.C.; Williamson, S. Physical activity surveillance in parks using direct observation. *Prev. Chronic Dis.* **2014**, *11*, 130147. [CrossRef] [PubMed]
48. Carlson, J.A.; Hipp, J.A.; Kerr, J.; Horowitz, T.S.; Berrigan, D. Unique Views on Obesity-Related Behaviors and Environments: Research Using Still and Video Images. *J. Meas. Phys. Behav.* **2018**, *1*, 143–154. [CrossRef]
49. Yue, X.; Antonietti, A.; Alirezaei, M.; Tasdizen, T.; Li, D.; Nguyen, L.; Mane, H.; Sun, A.; Hu, M.; Whitaker, R.T.; et al. Using Convolutional Neural Networks to Derive Neighborhood Built Environments from Google Street View Images and Examine Their Associations with Health Outcomes. *Int. J. Environ. Res. Public Health* **2022**, *19*, 12095. [CrossRef]

50. Barr, M.; Signal, L.; Jenkin, G.; Smith, M. Capturing exposures: Using automated cameras to document environmental determinants of obesity. *Health Promot. Int.* **2015**, *30*, 56–63. [CrossRef]
51. Carlson, J.A.; Jankowska, M.M.; Meseck, K.; Godbole, S.; Natarajan, L.; Raab, F.; Demchak, B.; Patrick, K.; Kerr, J. Validity of PALMS GPS scoring of active and passive travel compared with SenseCam. *Med. Sci. Sports Exerc.* **2015**, *47*, 662–667. [CrossRef] [PubMed]
52. Kelly, P.; Doherty, A.R.; Hamilton, A.; Matthews, A.; Batterham, A.M.; Nelson, M.; Foster, C.; Cowburn, G. Evaluating the feasibility of measuring travel to school using a wearable camera. *Am. J. Prev. Med.* **2012**, *43*, 546–550. [CrossRef]
53. Harms, T.; Gershuny, J.; Doherty, A.; Thomas, E.; Milton, K.; Foster, C. A validation study of the Eurostat harmonised European time use study (HETUS) diary using wearable technology. *BMC Public Health* **2019**, *19*, 455. [CrossRef]
54. Willetts, M.; Hollowell, S.; Aslett, L.; Holmes, C.; Doherty, A. Statistical machine learning of sleep and physical activity phenotypes from sensor data in 96,220 UK Biobank participants. *Sci. Rep.* **2018**, *8*, 7961. [CrossRef] [PubMed]
55. Grauman, K.; Westbury, A.; Byrne, E.; Chavis, Z.; Furnari, A.; Girdhar, R.; Hamburger, J.; Jiang, H.; Liu, M.; Liu, X. Ego4d: Around the world in 3000 hours of egocentric video. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–20 June 2022; pp. 18995–19012.
56. Cartas, A.; Radeva, P.; Dimiccoli, M. Activities of daily living monitoring via a wearable camera: Toward real-world applications. *IEEE Access* **2020**, *8*, 77344–77363. [CrossRef]
57. Kelly, P.; Marshall, S.J.; Badland, H.; Kerr, J.; Oliver, M.; Doherty, A.R.; Foster, C. An ethical framework for automated, wearable cameras in health behavior research. *Am. J. Prev. Med.* **2013**, *44*, 314–319. [CrossRef]
58. Meyer, L.E.; Porter, L.; Reilly, M.E.; Johnson, C.; Safir, S.; Greenfield, S.F.; Silverman, B.C.; Hudson, J.I.; Javaras, K.N. Using Wearable Cameras to Investigate Health-Related Daily Life Experiences: A Literature Review of Precautions and Risks in Empirical Studies. *Res. Ethics* **2022**, *18*, 64–83. [CrossRef]
59. Google Developers. Media Pipe. Available online: https://developers.google.com/mediapipe (accessed on 1 October 2023).
60. Arnab, A.; Dehghani, M.; Heigold, G.; Sun, C.; Lučić, M.; Schmid, C. Vivit: A video vision transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 6836–6846.
61. Shi, X.; Chen, Z.; Wang, H.; Yeung, D.-Y.; Wong, W.-K.; Woo, W.-C. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 1–9.