ESTIMATING ENERGY COST OF PHYSICAL ACTIVITIES FROM VIDEO USING 3D-CNN NETWORKS

by

Pragya Shrestha Chansi

A Thesis Submitted in

Partial Fulfillment of the

Requirements for the Degree of

Master of Science

in Computer Science

at

The University of Wisconsin-Milwaukee

May 2023

ABSTRACT

ESTIMATING ENERGY COST OF PHYSICAL ACTIVITIES
FROM VIDEO USING 3D-CNN NETWORKS

by

Pragya Shrestha Chansi

The University of Wisconsin-Milwaukee, 2023
Under the Supervision of Professor Rohit J. Kate

This research proposes a machine learning model that can estimate the energy cost of physical activities from video input. Currently, wearable sensors are commonly used for this purpose, but they have limitations in terms of practicality and accuracy. A deep learning model using three dimensional convolutional neural network (3D-CNN) architecture was used to process the video data and predict the energy cost in terms of metabolic equivalents (METs). The proposed model was evaluated on a dataset of physical activity videos and achieved an average accuracy of 71% on energy category prediction task and an root mean squared error (RMSE) of 1.14 on energy cost prediction task. The findings suggest that this approach has the potential for practical applications in physical activity surveillance, health interventions, and at-home activity monitoring.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

CNN	Convolutional Neural Network
LSTM	Long Short-Term Memory
MSE	Mean Squared Error
NLP	Natural Language Processing
RMSE	Root Mean Squared Error
RNN	Recurrent Neural Network

ACKNOWLEDGEMENTS

First, I would like to express my gratitude to my advisor Prof. Rohit J. Kate, who has been helping through my thesis research. I am thankful towards his tireless efforts, patience, and encouragement in supporting my research journey. Without his support and motivation, I would not have been able to complete my research.

Besides my advisor, I would also like to thank Prof. Scott Strath and the Department of Kinesiology at UWM (University of Wisconsin Milwaukee) for providing the experimental data used in this study. I would also like to thank Prof. Jun Zhang and Prof. Scott Strath for agreeing to be my thesis committee members.

Finally, I would like to thank my parents for their unwavering love, support, and encouragement throughout my academic journey. Thank you for your guidance, faith, and everything you have done for me.

Chapter 1

1 Introduction

1.1 Background and Problem Statement

In general, wearable sensors are used for estimating the energy cost of physical activities. When we think of the energy cost to perform an activity, one common metric to represent the energy cost of physical activities is metabolic equivalents (METs). One metabolic equivalent (MET) is defined as the amount of oxygen consumed while sitting at rest and is equal to 3.5 ml O2 per kg body weight and minute [14]. Usually, when a body is at rest, the energy cost is equal to 1 MET. However, MET usually ranges between 1.5-3.0 when we perform light activity. Anything greater than 3.0 MET would reflect moderate to vigorous physical activity [14]. Several studies have used wearable sensors to estimate energy expenditure and METs during physical activities. For example, a study by Kate et al. (2016) used a body-worn accelerometer to estimate METs while the participants of a study performed various physical activities. Wearable sensors are an accurate and convenient method for estimating energy costs. But it might not be convenient to wear these devices all time. With the rapid development in Artificial Intelligence and deep learning models, the study aims to research the possibility of utilizing deep learning methodologies for estimating energy cost through Video Surveillance.

1.2 Motivation and Objectives

Wearable Accelerometers and sensors can be used to derive the estimates of energy cost through METs. But this requires us to continually wear these devices for accurate measurement which may not be realistic or practical in all situations. In this study, we research a machine learning model

that can derive the approximate energy cost through video as an alternative solution. Accurate assessment of physical activity has several applications from Physical Activity Surveillance to evaluation of the effectiveness of health interventions in monitoring hospitalized patients and older adults in nursing facilities [2]. Furthermore, the predictive model could be applied in smart gyms, or virtual reality (VR) gaming that mainly relies on camera sensors [1] or even at-home activity monitoring. The primary objective of this study is to research potential deep learning networks that are capable of capturing meaningful patterns from video to accurately estimate the energy cost and intensity of physical activity.

Chapter 2

2 Literature Review

This chapter evaluates published literature on estimating energy and intensity of physical activities from Video. There has been ongoing research on improving energy cost estimation and physical activity intelligence from wearable devices. But there has not been much progress on energy cost estimation through video surveillance. Having said that, recently there have been studies on video processing for energy cost estimation and physical activity detection.

In 2022, Perrett et al conducted a study to determine what activities were most suited to calibrate a vision-based personalized energy expenditure model without a calorimeter [3]. The video data consisted of 32 seconds of continuous action/activities. A wide range of activities were included in the experiment such as standing, sitting, walking, wiping, vacuuming, sweeping, lying, exercising, stretching, cleaning, and reading. As stated in the study, the participants were filmed using an off-the-shelf RGB-D (Red, Green, Blue plus Depth) sensor. For privacy reasons, the video footage was pseudonymized by extracting silhouettes of the participants. For each action and pair of actions, the average METs values amongst different participants were analyzed and later used for fine-tuning the personalized energy expenditure model. The goal of the paper was to fine-tune personalized calorie expenditure data, and it was able to do so with an average mean-squared error of approximately 2. Data from 10 participants were used and a total of 4 hour video was used to train the model in the experiment. The paper focuses more on action detection rather than energy cost estimation. This study focuses more on energy cost estimation and improving the performance of energy cost estimation from video analysis.

There have also been a few studies where calorie expenditure was predicted through video analysis. Peng et al. attempted to explore the problem of automatically inferring the number of kilocalories used by humans during physical activity from observing videos [11]. For the study, an Omni-source benchmark Vid2Burn was used to estimate calorie expenditure from video data. The dataset featured a range of high and low-intensity activities. However, it didn't have actual energy expenditure data. Therefore, the annotations and results inferred in the experiment were merely approximations and didn't reflect exact measurements. For our study, the experimental data was collected from a well-calibrated environment where a subject was in a specially constructed air-tight chamber which enabled directly measuring the oxygen consumption by the subject. We could say the data used in this experiment reflect well with actual energy costs.

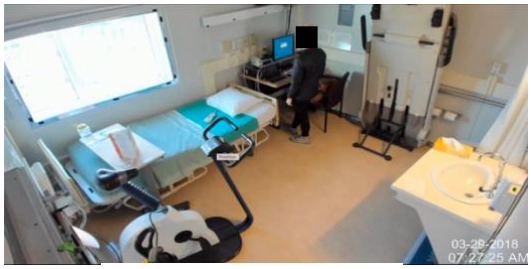
Hypothetically, one create a model that perfectly modeled the energy expenditure of a particular person. But it would not perform well for a different person. Based on the metabolism and weight of the person, for the same activity two people could have varying energy expenditure costs. In 2019, Saponaro et al. performed a study to estimate physical activity strength and energy expenditure using age, gender, speed, and activity cues [1]. They were able to reach an overall accuracy of 89.5% for physical activity strength estimation and an average Energy expenditure difference of 1.96 kCal/min. For our study, we will not be using age, gender, speed, or activity data. We want the machine learning model to make an accurate representation of energy cost and category simply from the video of the person performing a task.

Chapter 3

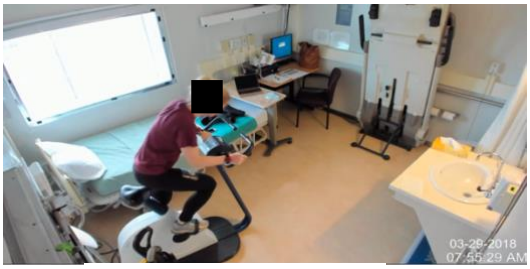
3 Materials and Methods

3.1 Data Collection

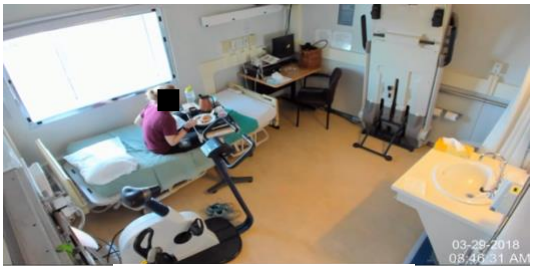
For this study, the dataset was obtained from the Department of Kinesiology at UWM (University of Wisconsin-Milwaukee). The video data were collected on multiple subjects in a specially-designed environment to monitor the physical activity of the person. In this study, data from 4 subjects were used for training and testing purposes. For each subject, there were approximately 12 hrs. of video, where the subject performed a range of physical activities including, sitting, standing, reading, walking, walking on a treadmill, stretching exercises, working on a computer, etc. Snapshots of a subject performing various physical activities are shown in Figure 3.1.



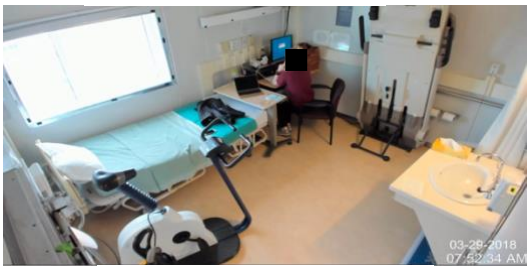
Standing



Cycling

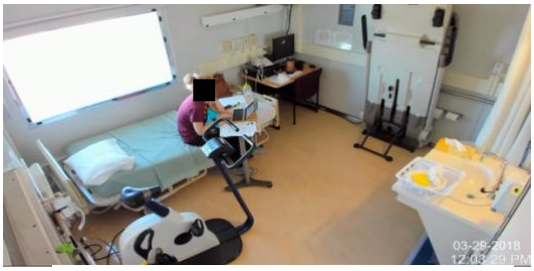


Eating



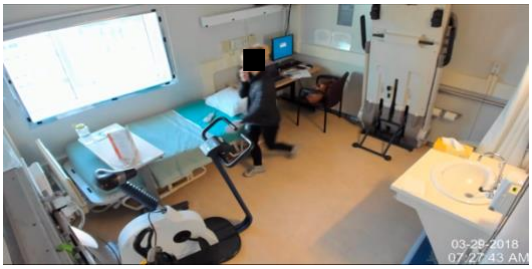
Sitting

Walking on Treadmill



Working on Computer

Sitting on Bed



Walking

Figure 3.1 Snapshots of a subject performing various physical activities

3.2 Machine Learning Techniques

3.2.1 Regression

In machine learning, regression is a type of supervised learning to predict a continuous output variable based on one or more input variables. The goal of regression is to find the best-fit line or curve that describes the relationship between the input and output variables. Based on the best line of fit, we could predict the output variable on unseen input. Regression is a powerful tool for statistical inference and has also been used to try to predict future outcomes based on past observations [12].

3.2.2 Classification

Classification is a supervised learning task that involves categorizing input data into predefined classes or categories based on the features/attributes of the data. In essence, classification algorithms intend to establish a relationship between input variables and output variables. The correlation between the input and output variables helps us predict the class categories of novel input. There are various classification algorithms available in machine learning, including logistic regression, decision trees, support vector machines, and neural networks. These algorithms differ in their assumptions, model complexity, and ability to perform well on different types of datasets. Classification has numerous real-world applications, such as image recognition, spam filtering, sentiment analysis, and credit risk analysis. [15]

3.2.3 Neural Networks

Neural Networks are a massive parallel combination of simple processing units which can acquire knowledge from the environment through a learning process and store the knowledge in its connections as stated by Haykin [4]. Neural networks are loosely based on the idea of mimicking the learning process of the human brain. A neural network consists of an input layer, one or more hidden layer(s), and an output layer. An input layer takes the input for the network and the output layers generate the output. There may be several hidden layers in a neural network based on the complexity of the network. During training, a neural network typically learns predict output from inputs by iteratively updating the weights of its nodes to minimize the error in the output it generates. The weights are tuned by an iterative procedure using the training data to approximate the relation between input and output [2]. It can be used for both classification and regression tasks.

3.2.4 Convolution Neural Networks (CNNs)

Convolution Neural Networks (CNNs) are a type of neural network architecture [5] and is one of the most popular machine learning method for image processing. The CNN gives exceptional performance in machine learning problems that deal with image classification, image segmentation and other computer vision tasks [6]. In a typical CNN architecture, we have a convolutional layer, a pooling layer, and fully connected layers. In the convolutional layers, a filter slides across the input while extracting local features. The extracted features are then fed to another layer which allows the network to gradually learn complex features and distinguish the features/patterns that will lead to a better performance of the model. The number of layers in the neural networks depends on how complex we want our network to be. More layers also mean more computational

time and resources. So, there is a trade-off between the complexity of the network and computational time.

3.2.5 3D CNN

A 3D CNN is an extension of a CNN network that takes three-dimensional data as input. Therefore, we could provide a sequence of images as the input of the network to perform video analysis. A 3D CNN model is capable of extracting features from both spatial and temporal dimensions by performing 3D convolutions to capture the temporal motion information encoded in the sequence of frames [7].

3.2.6 CNN-RNN architecture

An recurrent neural network (RNN) combines the state information from the previous timestamp with the input from the current timestamp to generate the state information and output for the current timestamp. [8] RNNs are a type of neural network architecture that is useful in dealing with time series data. For video analysis, the CNN-RNN model is one of the popular neural network architectures. We first extract the features from the images using the CNN model and then pass it to the RNN model to capture temporal information from the sequence of the features from the frames.

3.2.7 Model Evaluation

In this study, we performed both regression and classification tasks. In machine learning, there are various metrics available to evaluate the performance of the machine learning model based on the machine learning tasks. Some of the performance metrics used are Accuracy and F1-Score for

classification, and Root Mean Squared Error (RMSE) and R-Squared (coefficient of determination) for regression.

Accuracy

Accuracy is one metric for evaluating classification models which specify that performance metric is the ratio between the number of correctly classified samples and the overall number of samples [9].

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

For binary classification, accuracy can also be calculated in terms of positives and negatives as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

where TP = True Positives, TN = True Negatives, FP = False Positives, and FN = False Negatives.

F1 Score

F1 Score, also known as F-Score or F-measure, is a common metric used in the evaluation of classification models. F1 Score aims to measure the balance between precision and recall. Precision is the ratio of the true positive results to the sum of true positive and false positives. On the other hand, recall is the ratio of true positive results to the sum of true positives and true negatives in the test set. The F1 Score is a weighted harmonic mean of precision and recall, where the weight of the precision and recall can be adjusted based on the specific needs of the application [10]. The formula to compute F1 Score is:

$$F1\ Score\ = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision}\ +\ \text{Recall}}$$

Root Mean Squared Error (RMSE)

RMSE is the standard deviation of the errors made by a model that predicts the target variable [17]. RMSE stands for Root Mean Squared Error, which is a popular metric used for evaluating the accuracy of a regression model. It measures the square root of the average squared differences between the predicted and actual values of the target variable.

The formula for calculating RMSE is:

$$RMSE\ = \sqrt{\frac{\sum(actual\ -\ predicted)^2}{n}}$$

where n is the number of data points, actual is the actual value of the target variable, and predicted is the predicted value of the target variable.

R-Squared

The statistical measure R-squared (R^2) is used to indicate the amount of variance in the dependent variable that can be explained by the independent variable(s) in a regression model. This essentially measures how well a regression model fits the data, or the model's goodness of fit. The value of R-squared ranges from 0 to 1. A value of 1 would indicate a perfect fit and 0 would indicate the model's inability to explain the variance of the dependent variable. However, it's important to note that a high R-squared value doesn't necessarily imply that the model is the optimal fit for the data, as there may be other unaccounted factors. [13]

R-squared can be calculated using the following formula:

$R^2 = 1 - (SS\_res / SS\_tot)$

where SS_res is the sum of squares of the residuals (the difference between the actual and predicted values), and SS_tot is the total sum of squares (the difference between the actual values and the mean of the dependent variable).

K-Fold Cross Validation

K-Fold Cross Validation is a statistical method in machine learning that is commonly used to compare the performance of the model on data points that it has never seen before. In *k*-fold cross-validation, the original sample is randomly partitioned into *k* equal-sized subsamples. Of the *k* subsamples, a single subsample is retained as the validation data for testing the model, and the remaining $k - 1$ subsamples are used as training data. The cross-validation process is then repeated *k* times, with each of the *k* subsamples used exactly once as the validation data. The *k* results can then be averaged to produce a single estimation. The advantage of this method over repeated random sub-sampling (see below) is that all observations are used for both training and validation, and each observation is used for validation exactly once. 10-fold cross-validation is commonly used but in general, *k* remains an unfixed parameter [16].

3.3 Methodology

3.3.1 Data Preprocessing

Each subject had 12 hrs. of continuous video data. Creating a machine learning model that takes hours of video information as input would be computationally very expensive. In the annotated data for the subject corresponding to the video, the METs were reported every minute. Therefore, to synchronize the video to the METs data, 1-minute video intervals corresponding to their respective MET values were extracted. The aspect ratio of each video was 1280x720 pixels. Video

Data from 4 subjects were preprocessed. Each video file was approximately 12GB, this means a total of 48hrs. of video about 48GB of data was processed and used in this study. Although the annotated data during data collection, had a total of 21 subjects, data from 4 subjects were used due to computational limits. In future work, we could use data from all the subjects to enhance the performance of the predictive model.

3.3.2 Processing Frames from Video

For each video interval, the frames were extracted as a two-dimensional array consisting of pixel values. The video data was approximately 12 GB per subject, which would be difficult to process on a normal computer. Often, machine learning models do not need high-quality images for image processing and depending on the task, lower quality images could be sufficient. In Figure 3.2, we compare the same image with different aspect ratios.

Figure 3.2 Comparison of Original and Resized Images

When we resize images, the level of detail in the picture decreases as we make it smaller. Each pixel information is considered a feature in the machine learning model so higher quality images would mean more features. High-quality images lead to complex features and high computational costs. If computational costs are a concern, experimenting with lower-quality images and monitoring the performance of the model would be one way to address this issue. Therefore, during

the initial phase, images with different aspect ratios were fed to the machine learning model to observe the time taken to train the model and its capability to correctly predict on new images. We could say that the aspect ratio of the frames was fine-tuned to find the perfect balance between the cost of computation and the performance of the model. There was no notable change in the performance between the image with an aspect ratio of 80 X 45 and 160 X 90. The lower aspect ratio was chosen simply to reduce the storage and computational cost. Compared to the original aspect ratio of 1280 X 720, the selected aspect ratio for training the model was reduced significantly. If we carefully look at the resized image with an aspect ratio of 80 X 45, it is pixelated i.e., it does not have many details and it is difficult for the human eye to interpret this image. This may seem counter-intuitive, but for machine learning models, depending on the type of task, less detailed pictures could lead to more accurate predictions. For example, in our case, we want the model to learn and interpret how energy is expended when there is no movement, light movement, or extensive movement. The reduced details in image size will help the model focus on the larger picture. The quality of frames can be increased or decreased based on the type of image processing. In our case, the predictive model was able to generate meaningful results with reduced pixel size, so a higher aspect ratio was not used.

Even with the lower aspect ratio, the training time was considerably higher. This made it difficult to fine-tune the model and experiment with different neural network architectures. Therefore, 20 frames were extracted from each video to reduce the size of the input data. The basic idea was to collect 1 frame every 3 seconds from the video. This means, there were 20 frames from each 1-minute video that were passed as a sequence to the predictive model. The visual illustration of how the frames were picked from each video for machine learning is shown in Figure 3.3.

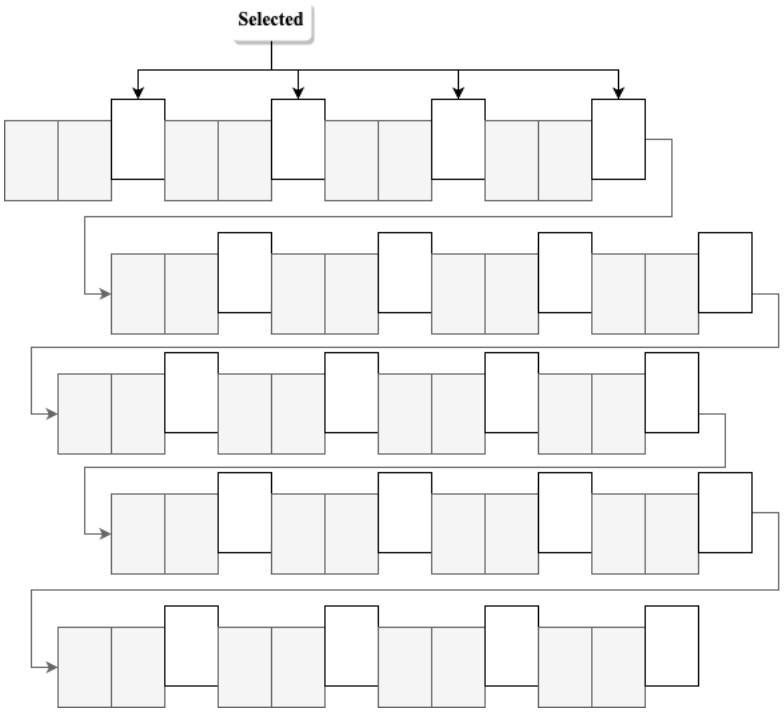


Figure 3.3 Visual Illustration on How frames were selected

The frames were then converted into two-dimensional arrays containing each pixel's information. The resulting arrays were then stored for further processing. The total size for data from 1 subject was approximately 600 MB which reduced the computational time for training the model significantly. Initially, the model was also tested at 60 frames per minute and 30 frames per minute respectively. However, there was not much change in the performance of the model. So, 20 frames per minute was decided in order to lower the training time of the machine learning model.

3.3.3 Converting METs value into categories

The original dataset did not contain categorical information on the physical intensity (Energy Category). Therefore, the majority of postures corresponding to the 1-minute interval based on which the METs were computed were used along with the METs value to determine the category

of the METs. For this study, we used three categories: Sedentary, Light, and MVPA (Moderate to Vigorous Physical Activity). The following logic was used to map the categories.

```
if METs < 1.5 or (METs < 2.0 and code=="Posture1"):
    r = "Sedentary"
elif code=="Posture2" or (METs >= 1.5 and METs < 3.0):
    r = "Light"
elif METs >= 3.0:
    r = "MVPA"
```

The “code” was determined by finding the majority of postures during the 1-minute interval. If most of the postures are sitting, lying, or crouching/kneeling/squatting then the code is “Posture1”. If most of the postures are standing, then the code is “Posture2”. The code is neither otherwise.

3.3.4 Data Statistics

After converting 12 hrs. video for four subjects 1001, 1002, 1003, and 1004, approximately 700 1 min video intervals were obtained. A total of 2868 videos were used for training and testing the predictive model. The distribution of METs value for individuals as well as all subjects is shown in Table 3.1.

Table 3.1 Data Distribution for Regression Analysis

SUBJECTS	INSTANCES	STANDARD DEVIATION	MEAN	MIN	MAX
1001	714	2.03	2.65	0.65	11.32
1002	718	1.24	2.13	0.39	6.22
1003	719	1.08	1.87	0.06	6.68
1004	717	1.32	1.92	0.04	7.82
TOTAL	2868	1.49	2.14	0.04	11.32

In the above table, we can see that subject 1001 represents greater variability in METs value compared to other subjects with a range of METs values between 0.65 and 11.32. The overall mean of subject 1001 is also higher compared to other subjects. Usually, when the body is at rest, it is expected to expend energy around 1 METs. But we can see that the min METs value for the subjects is less than 1. For Subjects 1003 and 1004, it's less than 0.04. This could mean that the dataset we are using has outliers and it can affect the way our regression model converges. Overall, the mean value throughout all the subjects is approximately 2.0 METs. This implies that most of the time the subjects were either at rest or performing light activity. It's also evident from the energy category distribution that the Sedentary Category makes up most of the data. The distribution of energy category for individuals as well as all subjects is shown in Figures 3.4 and 3.5 below.

Figure 3.4 METs Category Distribution for each subject

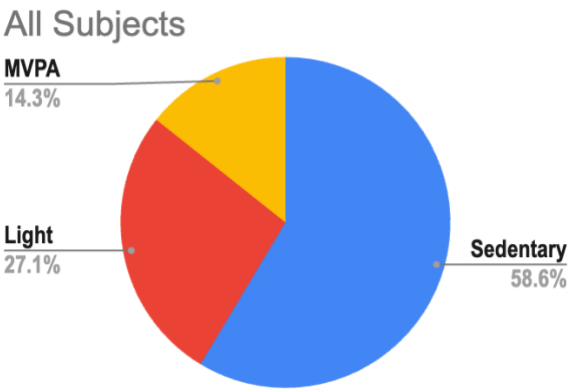


Figure 3.5 METs Category Distribution for all subjects

3.3.5 Predictive Models

In this study, a 3D CNN architecture was used for both classification and regression. The predictive models consisted of 3 convolutional layers with nodes 32, 64, and 128 respectively, followed by a pooling layer and batch normalization layer. The features from the convolutional layer were flattened and passed to a 512-node Dense layer for decision-making. The architecture of the 3D-CNN model is shown in Figure 3.6.

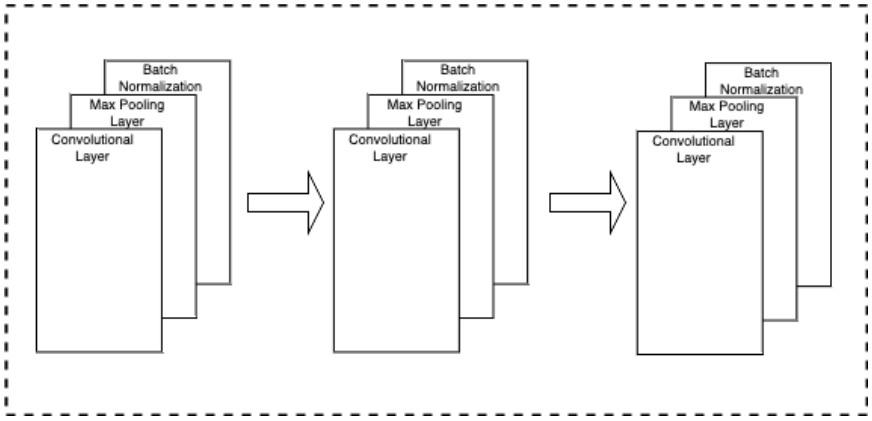


Figure 3.6 Architecture of 3D-CNN Model for Video Analysis

As shown in Table 3.1, Figures 3.4 and 3.5 in the Data Statistics section, the data distribution across various energy categories and METs value is not balanced. To avoid overfitting to the majority

class and range, regardless of training data size, a small batch size of 32 instances per training epoch was used. The metrics used to measure the performance of the regression model were the RMSE score and R-Squared coefficient. While the metrics for the classification model were accuracy and F1-Score. The precision and recall value for each categorical value were also computed while testing the model. All the models in the experiment were trained on 100 epochs. However, the model may reach a plateau in less than 100 epochs. To address this use, an early stopping call was added to each model that was triggered when the model didn't improve for 15 continuous epochs. Initially, the model was trained and fine-tuned using a single subject. An iterative approach was taken to fine-tune the different parameters of the model and the model architecture presented in this section is the one that was able to perform the best.

3.3.6 Evaluation of the Machine Learning Model

Machine Learning is all about generalization which means that model's performance can only be measured with data points that have never been used during the training process [17]. To make sure we train and test our model with different data points we always split our data into training and test sets.

Subject-Dependent evaluation approach

During the initial phase of the research, while training and fine-tuning the 3D-CNN model, the data from the same subject was used for both training and testing. To make sure, the model is capable of predicting the energy cost of unseen subject data, 5-fold cross-validation was used to evaluate the model.

Subject-Independent evaluation approach

The goal of our experiment is to build a predictive model that can correctly estimate the energy cost and category for everyone. Due to computational limitations, we used data from only 4 subjects. However, to ensure the robustness and generalizability of the machine learning model, a subject-independent evaluation approach was employed. Specifically, the model is being tested on a separate set of subjects from those used in training, allowing for an accurate assessment of its performance on novel data.

3.3.7 Test Bed and Experimental Setup

For the experiments, I utilized a MacBook Pro equipped with a RAM capacity of 16 gigabytes and an Apple M1 chip. This computational device is providing me with the necessary processing power to run the required algorithms and models for my research project. The machine learning models and data were processed using Python and machine learning libraries in Python such as Keras, NumPy, and pandas. The source code used for the study is open-source and can be found on the Github repository. *[https://github.com/pragyasresta29/video-regression]*.

Chapter 4

4 Results and Discussion

4.1 Subject-Dependent Evaluations Results

For evaluation of the model within the subject, 5-fold cross-validation was used to test the performance of the model's ability for estimating energy cost within the same subjects. The model was trained and tested using 1 subject, 2 subjects, and 3 subjects respectively. The average performance of all the trained models were recorded as shown in Tables 4.1 and 4.2.

Table 4.1 Regression Results for 5-fold Cross Validation

SUBJECTS	AVG RMSE	AVG R-SQUARE	RMSE	R-SQUARE
1001	0.74	0.87	0.63	0.88
1001, 1002	0.74	0.80	0.63	0.86
1001, 1002, 1003	0.71	0.78	0.60	0.82

Table 4.2 Classification Results for 5-fold Cross Validation

SUBJECTS	DISTRIBUTION		AVG ACCURACY	ACCURACY
1001	*Light*	34.87%	0.80	0.81
	Sedentary	47.34%		
	MVPA	17.79%		
1001, 1002	*Light*	32.05%	0.80	0.84
	Sedentary	50.98%		
	MVPA	16.97%		
1001, 1002, 1003	*Light*	27.75%	0.78	0.83
	Sedentary	56.90%		
	MVPA	15.34%		

From the evaluation results, it is evident that the models presented in this research are capable of producing the same result regardless of the number of subjects. However, the model with data from 3 subjects performed the best in terms of estimating the cost of energy with an average RMSE of 0.71. For classifying intensity of physical activity, the model with 1 subject performed the best with an average accuracy of 80%. The reason for this could be how the data was distributed among all the categories. For classification with 1 subject, we used subject 1001 which had the most balanced data distribution among the classes: Light, Sedentary, and MVPA compared to other subjects which lead to higher accuracy. Overall, the performance of the machine learning model was pretty good with an average RMSE of 0.73 and an average accuracy of 79%.

4.2 Subject-Independent Evaluation Results

To ensure that the predictive model can generalize the features in the video regardless of the subject/participant, the model was tested between subjects, i.e., the model was trained and tested between different subjects. This section covers the evaluation results for both subject-independent validations for energy cost and category estimation.

Predicting Energy Cost

For energy cost estimation i.e., estimation of METs value from the video, RMSE score and R-SQUARED (coefficient of determination) were used to evaluate the performance of the regression model. The results of the cross-subject validations are shown in Table 4.3.

Table 4.3 Regression Subject-Independent Evaluation Results between subjects

TRAIN SUBJECTS	TEST SUBJECTS	RMSE	R-SQUARED
1001, 1002, 1003	1004	0.89	0.53
1001, 1002, 1004	1003	0.99	0.43
1001, 1003, 1004	1002	0.92	0.27
1002, 1003, 1004	1001	1.73	0.27
AVG		**1.14**	**0.37**

Compared to the results when the machine learning model was tested within the same subject, the performance of the subject-wise cross validation decreased. The average RMSE when trained on 3 subjects increased to 1.14 from 0.73. This means, on average the difference between the energy cost predicted by our model and the actual energy cost is 1.14. The model's performance is not accurate but it's not bad given the average energy cost among our subjects was around 2.14 METs.

Predicting Physical Intensity Category from Video

For predicting the intensity of physical activity, accuracy, and F1-score were used to evaluate the performance of the classification model. The classification report of the model across multiple subjects is shown in the Tables 4.5 and 4.6.

Table 4.4 Classification Subject-Independent Evaluation Results between Subjects

MODEL	TRAIN SUBJECTS	TEST SUBJECTS	ACCURACY
1	1001, 1002, 1003	1004	0.75
2	1001, 1002, 1004	1003	0.83
3	1001, 1003, 1004	1002	0.63
4	1002, 1003, 1004	1001	0.63
AVG			**0.71**

Table 4.5 Classification Report Across Category between subjects for Table 4.4

MODEL	CATEGORY	INSTANCES	PRECISION	RECALL	F1-SCORE
1	*Light*	179	0.56	0.48	0.52
	Sedentary	457	0.80	0.97	0.88
	MVPA	81	1.00	0.10	0.10
AVG			**0.79**	**0.52**	**0.50**
2	*Light*	138	0.70	0.59	0.64
	Sedentary	494	0.88	0.93	0.91
	MVPA	87	0.71	0.66	0.68
AVG			**0.76**	**0.73**	**0.74**
3	*Light*	210	0.78	0.15	0.25
	Sedentary	392	0.60	0.98	0.75
	MVPA	116	0.93	0.33	0.48
AVG			**0.77**	**0.49**	**0.49**
4	*Light*	249	0.49	0.62	0.55
	Sedentary	338	0.71	0.70	0.70
	MVPA	127	0.90	0.44	0.59
AVG			**0.70**	**0.59**	**0.61**
TOTAL			**0.75**	**0.58**	**0.58**

When the model was trained and tested with the same subjects the average accuracy was around 79%. The accuracy decreased to 71% when we tested the model with subjects that were not used to train the model. The precision and recall score was comparatively higher for the Sedentary category compared to others. This implies that the machine learning model was able to learn the sedentary category very well. If we closely monitor the classification report presented in Table

4.6, we can infer that our data was not balanced i.e., there were more training and testing instances for the Sedentary category compared to the Light and MVPA category. On careful analysis of the test results, it was found on average the training set consisting of subject 1001 performed slightly better than compared to when other training subjects were used. When we look at the statistics of the energy cost and category values in the Data Statistics Section, it is evident that the distribution of data is comparatively even over the range of possible values. This could have allowed the machine learning model to prevent bias towards the majority class of the input data and accurately represent the underlying relationship between the video data and energy expenditure.

Chapter 4

5 Conclusion

5.1 Summary

In this thesis, we performed video analysis to predict the energy cost and intensity of physical activity using deep learning techniques. Using a 3D-CNN architecture for processing and analyzing video, the machine learning model was able to achieve an average accuracy of 71% and an RMSE of 1.14 when tested on novel subject data. When the model was tested on subjects that it was trained on it had better performance with an accuracy of 79% and RMSE of 0.73. The accuracy metric indicates the percentage of correct predictions made by the system, while the RMSE metric shows the average difference between the predicted and actual values of energy cost and category. This means the average error while estimating the energy cost was around 0.73 and 1.14 when tested on seen and unseen subjects respectively. The difference in performance indicates that the model was able to effectively learn the energy cost patterns of the subjects it was trained on but struggled with predicting energy cost for novel subjects. Overall, the results imply that the 3D-CNN model was effective in extracting features from the videos to predict energy cost and category. However, there is still room for improvement in terms of accuracy and RMSE.

5.2 Limitations and Future work

During the experiment, due to limited computational resources, the quality of the video data and the complexity of the models were significantly reduced to decrease the computational cost. For commercial or large-scale video surveillance applications, the quality of data could be upscaled which could improve the performance of the 3D-CNN model. Furthermore, data from more

subjects could be used which would help in better generalization of the machine learning model. Depending on the quality of data and the diversity of subjects, we could fine-tune the existing 3D-CNN model to improve it's predictive capabilities. During the initial phase of the experiment, various neural network architectures such as CNN-RNN, LSTM, and transformer were tested for both regression and classification, but the performance of these models was subpar. In the end, a 3D-CNN model was used as it was able to perform moderately. However, if we have the computational resources to use higher-quality videos we could experiment with other models or architectures that are widely used for video analysis such as CNN-RNN, CNN-LSTM, and transformer architecture.

Bibliography

[1] Saponaro, P., Wei, H., Dominick, G., & Kambhamettu, C. (2019, September). Estimating Physical Activity Intensity And Energy Expenditure Using Computer Vision On Videos. In *2019 IEEE International Conference on Image Processing (ICIP)* (pp. 3631-3635). IEEE.

[2] Kate, R. J., Swartz, A. M., Welch, W. A., & Strath, S. J. (2016). Comparative evaluation of features and techniques for identifying activity type and estimating energy cost from accelerometer data. *Physiological measurement*, *37*(3), 360.

[3] Perrett, T., Masullo, A., Damen, D., Burghardt, T., Craddock, I., & Mirmehdi, M. (2022). Personalized Energy Expenditure Estimation: Visual Sensing Approach With Deep Learning. *JMIR Formative Research*, *6*(9), e33606.

[4] S. Haykin, Neural Networks: A Comprehensive Foundation, Prentice Hall, New Jersey, 1999.

[5] Teuwen, J., & Moriakov, N. (2020). Convolutional neural networks. In *Handbook of medical image computing and computer assisted intervention* (pp. 481-501). Academic Press

[6] Albawi, S., Mohammed, T. A., & Al-Zawi, S. (2017, August). Understanding of a convolutional neural network. In *2017 international conference on engineering and technology (ICET)* (pp. 1-6). Ieee.

[7] Ji, S., Xu, W., Yang, M., & Yu, K. (2012). 3D convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, *35*(1), 221-231.

[8] Kamali, K. (1970). Deep Learning (Part 2)-Recurrent neural networks (RNN).

[9] Sokolova, M., Japkowicz, N., & Szpakowicz, S. (2006). Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation. In *AI 2006: Advances in*

Artificial Intelligence: 19th Australian Joint Conference on Artificial Intelligence, Hobart, Australia, December 4-8, 2006. Proceedings 19 (pp. 1015-1021). Springer Berlin Heidelberg.

[10] *F1 score in Machine Learning: Intro & Calculation*. V7. (n.d.). Retrieved April 26, 2023, from https://www.v7labs.com/blog/f1-score-guide

[11] Peng, K., Roitberg, A., Yang, K., Zhang, J., & Stiefelhagen, R. (2022). Should I take a walk? Estimating Energy Expenditure from Video Data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 2075-2085).

[12] Beers, B. (2021, October 30). What Regression Measures. Investopedia. Retrieved April 26, 2023, from https://www.investopedia.com/terms/r/regression.asp

[13] *R-squared*. Corporate Finance Institute. Retrieved April 26, 2023, from https://corporatefinanceinstitute.com/resources/data-science/r-squared

[14] Jetté, M., Sidney, K., & Blümchen, G. (1990). Metabolic equivalents (METS) in exercise testing, exercise prescription, and evaluation of functional capacity. *Clinical cardiology*, *13*(8), 555-565.

[15] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction (2nd ed.). New York: Springer.

[16] Pramoditha, R. (2020, December 20). *K-fold cross-validation explained in Plain English*. Medium. Retrieved April 26, 2023, from https://towardsdatascience.com/k-fold-cross-validation-explained-in-plain-english-659e33c0bc0

[17] Moody, J. (2019, September 6). *What does RMSE really mean?* Medium. Retrieved April 26, 2023, from https://towardsdatascience.com/what-does-rmse-really-mean-806b65f2e48e