Research Proposal: Workout Recognition Using 2D and 3D Convolutional Neural Networks

James Peralta, Jeffrey Boyd
Department of Computer Science
University of Calgary
Calgary, Canada
{james.peralta, jboyd}@ucalgary.ca

***Abstract*—Physical activity has proven to be effective in the primary and secondary prevention of a variety of chronic diseases including cancer and cardiovascular disease. Although exercise provides many incentives, people have trouble motivating themselves to maintain their routine. Workout loggers are used to track progression, and have shown to motivate users by showing them their improvements before concrete results were visible in the body. Current workout loggers require users to manually enter data into their phones, or write it down into a notebook. To automate this process, our system will use 2D and 3D Convolutional Neural Networks to analyze gym footage, recognize the workouts being performed, and count repetitions.**

Keywords—Workout Recognition, Machine Learning, Transfer Learning, Convolutional Neural Networks

I. Introduction

In 2017, nearly 50% of deaths in Canada were caused by cancer and heart disease [1]. In fact, 44% of adults over the age of 20 have at least 1 of 10 common chronic diseases [2]. To help reduce these chronic diseases, we must invest in preventive measures such as empowering our society into a healthy lifestyle. However, creating the habits to routinely engage in physical activity is a long-term process that requires a lot of personal motivation. Only 20% of people succeed in long-term weight loss maintenance [3].

A promising area of technology - persuasive technology, is designed to influence and change human behaviors. Emerging persuasive technology in the fitness sector has already shown to improve the health and fitness practices of its users [4]. Within the space of persuasive technology, activity tracking is one of the most prevalent strategies. Users of activity trackers that monitor their workout progress commented on how this technology had motivated them and helped them make durable changes in their lifestyle [5]. Activity Trackers can be used for more simple data, such as counting steps but fail to capture advanced data, such as repetitions of workouts. In this paper, we focus on a specific type of activity tracking, workout logging.

Although workout loggers offer plenty of value to their users, they have shown very poor retention rates. [6] found that within the first two weeks, 62% of users that downloaded a workout logging app stopped using it. This was because manual entry of data can be too repetitive and tedious, eventually leading to these technologies to go unused [4]. There have been attempts at removing the step of manually entering data through wearables [7] [8] [9], computer vision [10], and adding sensors onto equipment [11].

Computer vision is a promising solution because it can capture a richer set of spatial features. Wearables lose accuracy depending on where the sensor is placed. For example, a sensor placed on the wrist will not detect leg exercises such as a leg press because the hand is stationary during this exercise. Furthermore, adding sensors to equipment will not be able to track bodyweight exercises such as pushups where the user is not using any equipment. Not only can computer vision avoid these pitfalls, it is also the only solution that can be implemented as a single-point solution, which will avoid instrumenting many users or many machines.

II. Related work

A. *Using optical flow*

The current state of the art for this problem is Gymcam [10]. GymCam used optical flow to generate a feature vector of 27 features. This feature vector was passed along a pipeline of three separate multilayer perceptrons. With this procedure they were able to obtain an accuracy of 84.6% for exercise segmentation, 93.6% exercise recognition and count the number of repetitions within +/-1.7 on average.

B. *Feature based versus neural networks*

Patsadu et al. evaluated feature-based and neural network classifiers for human gesture recognition [12]. In their study, they used a Kinect camera to extract a body-joint position vector containing the locations of 20 joints including the head, hands, and feet. Using this vector as input into a classifier, they classified between three actions; standing, sitting down, and laying down. They observed the performance of multiple classifiers such as Back Propagation Neural Networks, SVMs, Decision Trees, and Naive Bayes. Out of every classifier, neural networks performed the best.

C. *Using 2D Convolutional Neural Networks*

Karpathy et al. surveyed the performance of two dimensional convolutional neural networks (2D CNNs) for video classification [13]. 2D CNNs avoid the feature extraction phase required by Gymcam and feeds raw frames as input into the network. This simplified the pipeline by removing the need to generate custom features and was shown to have significant performance gains over the feature-based approaches. Since 2D

CNNs take in a single frame as input, it can only encode spatial features. This can be a problem due to the temporal nature of videos. As part of their contribution, they proposed three ways of encoding temporal information, shown in Figure 1.

D. *Using 3D Convolutional Neural Networks*

Tran et al. introduced a three dimensional Convolutional Neural Network (3D CNN) they named, C3D [14]. The 3D CNN takes videos as input and encodes both spatial and temporal information without the need to perform workarounds such as early fusion. The difference between a 2D CNN and a 3D CNN is that the convolution operations use two dimensional and three dimensional kernels respectively, see Figure 2. They went on to show how 3D CNNs were better than 2D CNNs at preserving temporal information and outperformed 2D CNNs on various video analysis tasks. More interestingly, 3D CNNs we're shown to run 2x faster than optical flow solutions.

III. Proposed method

Our study will involve four phases; data collection, data labelling, building a classifier, and evaluating the classifier. We will use CNNs as our classifier due to the improvements in speed and performance over feature-based approaches. Our goal is to establish a baseline for neural networks for this task, analyze the benefits of using neural networks, and report the pitfalls of using neural networks.

A. *Data collection*

For data collection, we have arranged to book an athlete gym at the University of Calgary. We will strategically place cameras around the gym to capture each machine. Participants will be invited for a 30-90 minute workout session where they will perform 6 different exercises of their choice. Each camera will only store footage when it detects a participant in the field of view. This will create clips that start when a participant enters and ends when there is no longer any participants in the view. 60 participants, performing 3 sets per exercise will leave us with 1080 exercises captured on video.

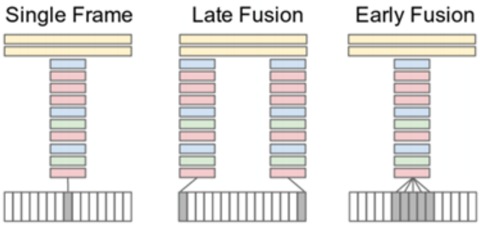

Fig. 1. The single frame approach only encodes information from a single frame, late fusion encodes temporal information from two frames which are 15 frames apart, and the early fusion encodes temporal information from several contiguous frames. Figure from Karpathy et al.

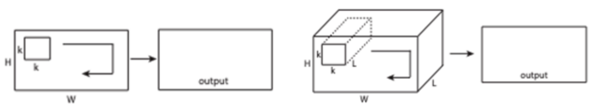

Fig. 2. 2D Convolutions (left) vs. 3D Convolutions (right). Figure from Tran et al.

B. *Data labelling*

After building up the dataset of exercises, we will have to label them before training our Neural Network. We will use GymCam's annotation tool which is optimized for labelling workouts and their repetitions. For each clip, we will draw a bounding box around a user when they begin performing a workout and remove it when they finish the set. Along with this bounding box we will label the exercise they performed along with the number of repetitions. This will create a dataset that can be used for exercise segmentation, exercise localization, exercise recognition and repetition counting.

C. *Building a classifier*

With this dataset we can now design a Neural Network that will locate the user, recognize the workout, and count repetitions. To perform these tasks we will create a pipeline of CNNs as follows. The raw frame is passed into an OpenPose network to detect humans in the field of view. Once a human is located we will crop a rectangle around them to focus in on the user. This frame will be fed into a Neural Network that will predict if the person in the frame is performing an exercise. If the person is performing an exercise, we will begin creating a clip using all subsequent frames until the user stops performing the exercise. Lastly, we will input this clip into a classifier that will predict the workout being performed and the number of repetitions. Each classifier mentioned above will either be a 2D or 3D CNN.

D. *Evaluating the classifier*

Similar to [10], we will separately evaluate the performance of exercise recognition and repetition counting. The performance of a predictor can be measured using metrics of accuracy, precision, and recall. There is often a trade-off between optimizing recall versus precision and in our case, we will optimize the recall metric. To generate these metrics with the least amount of variance and bias [15], we will perform stratified ten-fold cross validation on our collected dataset.

IV. Current Progress

We have developed the first stage of the data processing pipeline by creating the software that streams frames from a wired camera (Logitech C270) or a network camera (Ezviz Mini 0) into a server. To understand how we will use CNNs for this problem, we have experimented with 2D CNNs, OpenPose, and YoloV3 using a dataset of handpicked YouTube workout videos. We have also submitted an ethics review to the University of Calgary Conjoint Faculties Research Ethics Board (CFREB). Once our study is approved by the CFREB, we can begin phase 1 of our expected deliverables.

V. Expected deliverables

A. *Phase 1: 5 weeks*

- Set up the infrastructure in the gym: 1 week
- Invite participants into the gym to collect data: 4 weeks

B. *Phase 2: 2 weeks*

- Label the data collected: 1.5 weeks

- Aggregate visual and descriptive statistics of the dataset: 0.5 weeks

C. Phase 3: 11 weeks

- Experiment with OpenPose and YoloV3 for localizing users: 2 weeks
- Develop, train, evaluate, and reiterate on a 2D CNN: 6 weeks
- Experiment transfer learning using a pre-trained 2D CNN: 3 weeks

D. Phase 4: 11 weeks

- Develop, train, evaluate, and reiterate on a 3D CNN: 8 weeks
- Experiment transfer learning using a pre-trained 3D CNN: 3 weeks

Total: 29 weeks

REFERENCES

[1] Statistics Canada, "Leading causes of death, total population, by age group," *https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1310039401,* vol. , no. , p. , 2017.

[2] Goverment of Canda, "Prevalence of Chronic Diseases Among Canadian Adults," *https://www.canada.ca/en/public-health/services/chronic-diseases/prevalence-canadian-adults-infographic-2019.html,* vol. , no. , p. , 2019.

[3] R. R. Wing and S. Phelan, "Long-term weight loss maintenance.," *The American journal of clinical nutrition,* vol. 82, no. 1 Suppl, pp. 222S-225S, 2005.

[4] A. Ahtinen, E. Mattila, A. Väätänen, L. Hynninen, J. Salminen, E. Koskinen and K. Laine, "User experiences of mobile wellness applications in health promotion: User study of wellness diary, mobile coach and SeltRelax," in *2009 3rd International Conference on Pervasive Computing Technologies for Healthcare - Pervasive Health 2009, PCTHealth 2009*, 2009.

[5] T. Fritz, E. M. Huang, G. C. Murphy and T. Zimmermann, "Persuasive technology in the real world: A study of long-term use of activity sensing devices for fitness," in *Conference on Human Factors in Computing Systems - Proceedings*, 2014.

[6] D. Ledger, D.; McCaffrey, "How The Science of Human Behavior Change Offers The Secret to Long-Term Engagement," *http://endeavourpartners.net/assets/Endeavour-Partners-WearablesWhite-Paper-20141.pdf ,* 2016.

[7] D. Morris, T. S. Saponas, A. Guillory and I. Kelner, "RecoFit: Using a wearable sensor to find, recognize, and count repetitive exercises," in *Conference on Human Factors in Computing Systems - Proceedings*, 2014.

[8] M. Muehlbauer, G. Bahle and P. Lukowicz, "What can an arm holster worn smart phone do for activity recognition?," in *Proceedings - International Symposium on Wearable Computers, ISWC*, 2011.

[9] F. J. Ordóñez and D. Roggen, "Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition," *Sensors (Switzerland),* vol. 16, no. 1, 18 1 2016.

[10] R. Khurana, K. Ahuja, Z. Yu, J. Mankoff, C. Harrison and M. Goel, "GymCam," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies,* vol. 2, no. 4, pp. 1-17, 27 12 2018.

[11] H. Ding, L. Shangguan, Z. Yang, J. Han, Z. Zhou, P. Yang, W. Xi and J. Zhao, "FEMO: A platform for free-weight exercise monitoring with RFIDs," in *SenSys 2015 - Proceedings of the 13th ACM Conference on Embedded Networked Sensor Systems*, 2015.

[12] O. Patsadu, C. Nukoolkit and B. Watanapa, "Human gesture recognition using Kinect camera," in *JCSSE 2012 - 9th International Joint Conference on Computer Science and Software Engineering*, 2012.

[13] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar and F. F. Li, "Large-scale video classification with convolutional neural networks," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2014.

[14] D. Tran, L. Bourdev, R. Fergus, L. Torresani and M. Paluri, "Learning Spatiotemporal Features with 3D Convolutional Networks," 1 12 2014.

[15] R. Kohavi, "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection," 1995.

[16] D. E. R. Warburton, C. W. Nicol and S. S. D. Bredin, *Health benefits of physical activity: The evidence,* vol. 174, 2006, pp. 801-809.