Automated Fitness Tracking Using Machine Vision

Andrew Dworschak, Justin Kang[1], Jacob Budzis, Jonah Killam, Rahat Dhande and Ryan Cotsakis

***Abstract*— Regular exercise is key to maintaining good health, and keeping a log of one's exercises is shown to improve both progress and dedication to the gym. While most workout logging technologies such as smart phone applications require an extensive amount of input from the user, we propose a scalable solution to track one's exercises automatically.**

Limiting the scope of this research to the classification of free-weight and bodyweight exercises, we develop a three-step system to classify exercises, repetitions, and weights in real-time before uploading this information to a cloud database. Using a recurrent neural network, we classify exercises based on skeletal vectors from a Microsoft Kinect camera. The classified exercise windows are then filtered through principle component analysis and a spectrograph to detect peaks that define repetitions. The three-dimensional skeletal data is then mapped to a two-dimensional image that can be scanned for the most probable colour-coded dumbbell weight.

On a validation set of 41 exercises, the exercise classification achieves 97% accuracy, the repetition counting achieves 84% accuracy, and the weight detection achieves 100% accuracy. These results demonstrate the viability of machine vision techniques to automate fitness tracking.

I. INTRODUCTION

Physical exercise is one of the most important contributors to maintaining good health. Despite this, 2 out of 3 Americans fail to meet the recommended minimum amount of weekly exercise. Among reasons for this lack of exercise is the high attrition rate faced by fitness facilities. According to the International Health and Racquet Sports Association, the difficulty in receiving immediate, structured feedback on one's workout is one of the principal drivers of gym attrition [1].

One way to address the problem of limited feedback in the gym is to create a workout log that allows users to easily observe trends in their exercise history. However, current products in the market are all either overly burdensome to the user or severely limited in their scope of available exercises.

As an example of an overly burdensome technology, BodySpace is a smart phone application that keeps track of a user's exercise history provided that the user manually types each of their exercises into the application during their workout. As a rough estimate, if the average user performs 25 sets in their workout and takes 1 minute to enter the information into their phone, then 25 minutes, or half of the duration of their workout, is spent entering information into their phone.

As an example of a limiting technology, eGym is a company specializing in sensor-enabled workout machines. They can only track the user's workout on several specific machines manufactured by eGym and are unable to track any free-weight exercises, the largest class of exercises. With eGym's model, the fitness facility is restricted to a single machine supplier, and the user is limited to a small set of machine-based exercises when they wish to enjoy the automatic logging of their workout.

Outside the market, there have been many recent efforts in academic research to develop exercise classification and repetition counting capabilities [2], [3]. Most approaches in the literature focus on the analysis of accelerometer data [4], [5], [6], [7], [8]. However, given that most gym users are reluctant to wear accelerometers during their workout along with recent advances in machine vision, we anticipate that machine vision will become the technology that facilitates broad exercise monitoring capabilities.

We propose a system that uses machine vision to automate the classification of exercises, repetitions, and weights, uploading this workout data in real-time to the appropriate user's fitness profile. This system gathers structured feedback without burden for the user by creating a record of the user's workout history without requiring any user input. Further, the system has the potential to accommodate any exercise, making it more versatile than current market alternatives.

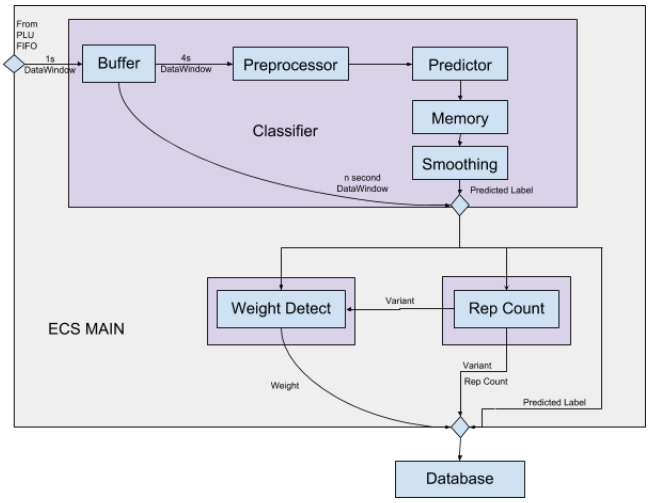

Fig. 1. Block Diagram of the System

The system is illustrated in a block diagram in Figure 1. It is composed of three subsystems: and Exercise Classification Subsystem, a Repetition Counting Subsystem and a Weight Detection Subsystem.

The remainder of this paper is structured as follows. Technology used and simplifications to the problem for this iteration of the system are described in the Setup section.

[1]This article was written by Andrew Dworschak and Justin Kang, based on the research of all 6 authors dating July 2017 - April 2018.

Each of the three subsystems are described in the Method section. The performance of the system is quantified in the Results section. Finally, we discuss our ongoing work on this project in the Conclusion section.

II. SETUP

A. Technology used

In this project, we use the Microsoft Kinect V2 camera. This camera generates RGB images, depth-based images and three-dimensional skeletal vectors recognizing up to six people at a time. The two sets of images and skeletal vectors constitute inputs to the system.

B. Simplifications of the problem

Recall that the goals of an eventual version of the system are to be able to classify any exercise performed with any weights, anywhere in the gym with high accuracy and to upload this data to the appropriate user's profile. However, this iteration of the system addresses only a subset of this functionality, making several simplifications:

- The system is limited to classifying 11 free-weight and body-weight exercises and their variants. Classified exercises include bicep curl (simultaneous and alternating variants), squat, sitting shoulder press, lateral arm raise, lunge (left and right variants), sit-up, dumbbell row (left and right variants), and sitting tricep extension.
- The system is limited to classifying three sets of color-coded weights.
- The system assumes that the user will take at least a 5 seconds break between their exercises.
- The system only uses one camera, and thus has a limited field of vision.
- The system does not incorporate the identification of users after they leave the frame.
- The system does not handle obstructions to the field of vision.

We feel that despite these simplifications, the system demonstrates the viability of the machine vision approach to exercise classification. By addressing free-weight exercises, the current bottleneck of the industry, we demonstrate the capacity of the system to outperform existing solutions.

III. METHOD

The system consists of three subsystems. The Exercise Classification Subsystem receives a time series of skeletal vectors as input, determines which exercise is being performed and when that exercise has been completed. The Repetition Counting Subsystem receives data from a completed exercise and determines the number of repetitions performed during that exercise. The Weight Detection Subsystem receives data from a completed exercise and determines which weight, if any, is being lifted by the user.

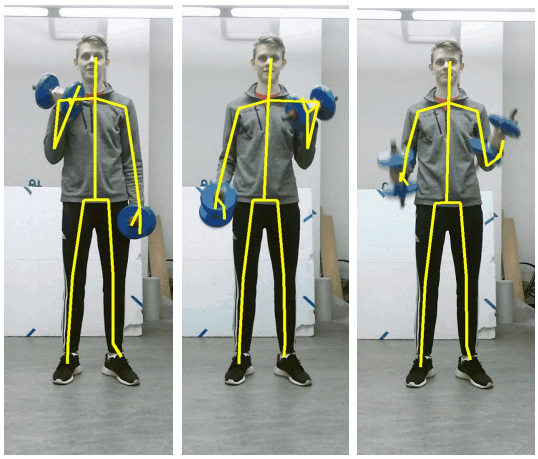

Fig. 2. Example of generated skeleton, and mapping onto color image

A. Exercise Classification

The Exercise Classification Subsystem consists of a preprocessing step that enforces translational and rotational invariance in the data, a predictor that takes a window of data as input and outputs a set of probabilities for which exercise is being performed, and a smoothing algorithm that determines at which time the exercises begin and end.

1) Preprocessing: Because we want our system to remain robust to users performing exercises while facing in different directions and standing in different parts of the frame, it is important to enforce translational and rotational invariance in our data. To accomplish this, the preprocessing step receives the raw skeletal vectors from the Kinect camera which consist of 75 points for each frame representing 25 joints of the human body in three-dimensional space. From this data, we calculate the angle between joints at 21 locations on the body and the normalized (mean of 1) distance between each of the 25 joints and the centroid of the body for each frame. By considering these 46 time series as opposed to the original 75 time series, our system becomes insensitive to translation and rotation in the camera frame, and disregards the absolute height difference between different users.

This processed data is then stored in a queue and delivered to the predictor to prevent the dropping of frames.

2) Predictor: The predictor employs a recurrent neural network (RNN) that accepts processed skeletal input vectors and outputs the most likely exercise classification from a given list. In addition to the 11 exercise variants, the RNN can output a 'NULL' exercise to indicate that the user is not performing any recognized exercise at that time. An RNN is the appropriate choice for this application due to the time series nature of the input data. Once trained, it exhibits an extremely quick computation time suitable for running the system in real-time.

The predictor outputs one list of probabilities per second representing the various exercises the user could be performing. This data is read by the smoothing algorithm to prevent erroneous classifications.

3) Smoothing: Reading the classification probabilities from the predictor, the smoothing algorithm applies a low pass filter that ignores exercises done for only several seconds. Then, by searching for transitions from the 'NULL' classification to that of a valid exercise, the algorithm ascertains a beginning and ending time for each exercise window, determines the most probable exercise over the span of each window, and labels each window with an appropriate exercise label. This procedure would be problematic if a user transitioned quickly from one exercise to another, as temporal filters may prevent the 'NULL' classification from being asserted. However, compared to a system that did not enforce a period of 'NULL' classification between each exercise, this method is observed to improve the overall accuracy of the Exercise Classification Subsystem significantly.

B. Repetition Counting

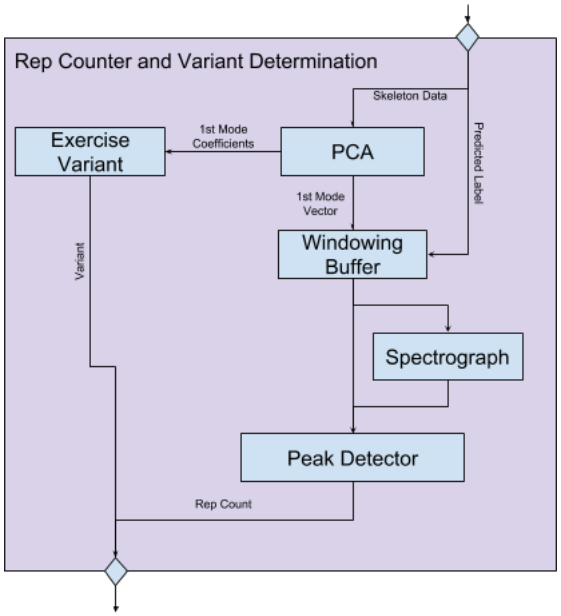

Fig. 3. Block diagram of the subsystem responsible for Repetition Counting and Variant determination

When given a labelled exercise, the Repetition Counting Subsystem selects the skeletal data associated with that exercise. This skeletal data consists of a three-dimensional set of coordinates for each of the 25 joints recorded. First, the non-vertical components of the data are discarded, reducing the vectors to contain only the vertical component of each joint. This simplification is appropriate because all of the exercises in our sample set are limited to free-weight and body-weight exercises that rely on the force of gravity to provide resistance. Thus, by discarding the non-vertical components of the vectors, the system reduces noise that could influence the accuracy of classifications.

Next, we project the remaining 25-dimensional joint data $X(t)$ defined in the interval $[t_{start}, t_{end}]$ into a one-dimensional signal space $y(t)$. To accomplish this, we first select the middle subset of the data by constraining $t_{start} + \tau < t < t_{end} - \tau$ for a fixed constant τ. This ignores the most uncertain portion of the data at the beginning and end of the exercise where large movements that are not part of the exercise can occur.

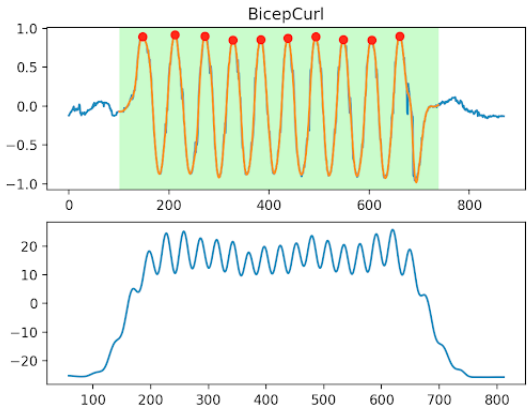

Fig. 4. Upper plot depicts the principle component of motion for the captured skeleton. Lower plot is part of a spectrogram, showing the intensity of the dominant frequency in the upper plot

Using prepackaged software, we decompose $X(t) = U\Sigma V^*$ using Singular Value Decomposition and select the first column of U corresponding to the eigenvector with the most variance in the data, denoted w. Then, considering the full range of t:

$$y(t) = w^T X(t), t \in [t_{start}, t_{end}]$$

The upper plot of Figure 4 depicts a signal $y(t)$ generated from skeletal data of a user performing the bicep-curl exercise. Due to the limitation of the Exercise Classification Subsystem's ability to determine precisely when an exercises has ended, small windows of skeletal data from before and after the exercises are included in the signal $y(t)$.

Once the signal $y(t)$ is successfully generated, the exact window in which the exercises took place is determined. First, the signal is filtered using a third order Butterworth low pass filter. The orange data on upper plot of Figure 4 shows the output of that filter. The exact window is then identified by finding the dominant frequency ω_D in the signal using a Fourier Transform:

$$Y(\omega) = DFT(y(t))$$
$$\omega_D = argmax\,(Y(\omega))$$

A spectrogram is then taken on the signal $y(t)$ using a window length that corresponds to two periods of the dominant frequency. The lower plot on Figure 4 shows the output of the spectrogram corresponding to ω_D. By observing the first time at which the power of the dominant frequency exceeds 50% of its maximum value and the last time at which the power returns below 50%, we identify the exact window when the exercise is being performed.

With this window identified, the number of repetitions can be counted using a simple peak detector. This is implemented by subtracting the mean from the signal and counting the number of peaks above zero, thereby ignoring small oscillations around the troughs.

1) Determining the Exercise Variant: Principal Component Analysis serves as a useful tool for determining which variant of an exercise is being performed. The dumbbell row, for instance, is often performed with either the left or the right hand. Looking at the magnitude of each dimension in the first principal component, we determine which joints contribute most the the primary variational mode of the signal $y(t)$. If it is determined, for instance, that the components of w corresponding to left arm-related joints have the greatest magnitude, it can be inferred that the left hand variant of the exercise is being performed.

C. Weight Detection

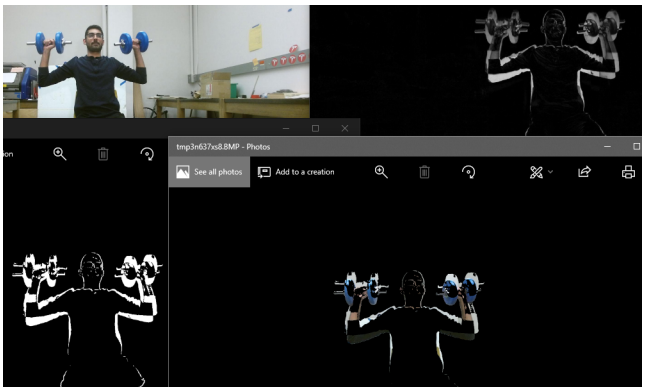

Fig. 5. Procedure for the image subtraction algorithm for weight detection

When given a labelled exercise, the Weight Detection Subsystem selects the skeletal vectors and RGB images associated with the time-window. In this project, the scope of weight detection was limited to the classification of dumbbells of three colours: blue, aqua, and purple, or the absence of any dumbbells.

First, the system determines whether dumbbell colour classification is appropriate for the given exercise. For example, it is appropriate for a user to be using dumbbells during bicep-curls and shoulder-press, but not during lunges and sit-ups. If the user is performing an exercise where dumbbell classification is not appropriate, then a weight of 0 lbs is classified.

If instead it is appropriate for the user to be lifting dumbbells, then the system follows three steps to classify the weight: skeletal mapping, image subtraction, and concentration thresholding.

1) Skeletal Mapping: First, we want to establish a mapping from the three-dimensional skeletal vectors to the two-dimensional RGB image. To accomplish this, we define a function $f : R^3 \rightarrow R^2$ such that $f(\alpha_i) = \beta_i$ for $i = 1, 2, ..., n$ where n is the number of joints analyzed. Then for each $\alpha_i = (1, x_i, y_i, z_i)$ we construct two matrices A and B, such that

$$A\alpha_i^T \alpha_i B \approx \beta_i, \forall i$$

To do so, we iteratively solve for A and B using linear projection. Starting with a guess for B, we solve for A^T in the over-constrained problem $(\alpha_i^T \alpha_i B)^T A^T = \beta_i^T$. The linear projection formula to solve for x in $Mx = b$ is

$$x = (M^T M)^{-1} M^T b$$

To encapsulate the data for all values of i in the original problem, let $b = (\beta_1^T, \beta_2^T, ...\beta_n^T)^T$, and $M = ((\alpha_1^T \alpha_1 B)^T, (\alpha_2^T \alpha_2 B)^T, ...(\alpha_n^T \alpha_n B)^T)^T$. With this new value for A, solve for a new B in the same way. Continuing this process until A and B converge, we find the closest projection of the three-dimensional skeletal vectors to the two-dimensional image.

Next, we locate the three-dimensional joints associated with each hand in the two-dimensional image. Because the dumbbells being lifted are very near the hands, we discard all parts of the image except a small rectangle around the position of the hands in our subsequent weight detection analysis.

2) Image Subtraction: By subtracting the RGB values of successive images from one another, we identify areas with the largest change in colour between successive frames. We pass these image differences through a binary thresholding filter to establish which areas change colour significantly. By softening the edges around the thresholded images using a moving average filter, we establish a continuous weighting of areas with the largest and most recent colour changes. This procedure is depicted in Figure 5.

3) Concentration Thresholding: We superimpose the weighting generated using image subtraction with the rectangle of interest around the position of the hands to establish the relative weight with which to consider each pixel in the rectangles. In this way, pixels that experience more recent colour change are weighted higher than those that do not.

By taking the weighted sum of each pixel's proximity to blue, aqua and purple hues, we determine the most likely colour of the dumbbells. The weight associated with this dumbbell colour is then uploaded to the cloud database along with the exercise and repetition information.

IV. RESULTS

To verify the efficacy of our system, we use 29 minutes of verification data that the neural net is not trained on. This includes 41 sets of 10 repetitions, spanning all 11 exercises and variants. Overall, the exercises are classified correctly with 97% accuracy, the weights are identified correctly with 100% accuracy, and the repetitions are counted with 84% accuracy. We provide several specifications to these results in the list below:

- Of the 41 sets, only one set of kneeling-row is misclassified as lateral-arm-raise.
- Every time the user is not exercising, the Exercise Classification Subsystem correctly classifies the current exercise as 'NULL'.
- 28 of the 41 exercises use dumbbells and the remaining 13 exercises do not use dumbbells. The weight is correctly classified in all 41 cases.
- As each exercise has 10 repetitions, all 10 repetitions are correctly counted in 20 of the 41 exercises.

- Many of the exercises in which the repetitions are not counted correctly count either 9 or 11 repetitions, only 1 repetition away from the correct value. The additional or omitted repetition occurs exclusively at the beginning and end of the exercise window.
- The sum of absolute error for each exercise in repetition counting is 65 repetitions out of the 410, giving the metric that 84% of the repetitions are correctly counted.

V. CONCLUSION

This research shows that machine vision can be used to effectively classify user exercises, weights and repetitions with high accuracy. As neither academia nor industry currently have the ability to track user workouts at scale, this project demonstrates the viability of machine vision as a long term solution to workout classification. Importantly, this research addresses the tracking of free-weight and bodyweight exercises which constitute the current bottleneck of fitness tracking.

However, a key input to this classification system is the skeletal vector data of users. Before this system can be implemented at scale, it is paramount that systems for accurate skeletal generation be developed for use in larger areas.

Our ongoing research focuses on the precursor to a large-scale skeletal tracking system. Namely, by installing several depth sensing cameras in a room, we are developing a computationally efficient algorithm to provide a real-time stream of three-dimensional information about the room. In this research, it is critical to address the stitching together of multiple depth and RGB images to create an accurate representation of objects in the room. From this stream of three-dimensional information, a neural network can be used to track human movement and generate skeletal representations of their bodies.

Once this skeletal tracking system is ready to be deployed at scale, the results from this research could be applied to real fitness facilities to fully automate the fitness tracking process.

ACKNOWLEDGMENT

We thank the Bycast Award for Entrepreneurship committee for their generosity in funding this research.

Part of this project was completed in accordance with the Capstone course requirements for Engineering Physics at the University of British Columbia, an accredited engineering program. We thank the program's staff for their feedback and their role in shaping our work.

REFERENCES

[1] John McCarthy. *IHRSA's Guide to Membership Retention*. International Health and Racquet Sports Association, 2007.

[2] Mathias Sundholm, Jingyuan Cheng, Bo Zhou, Akash Sethi, and Paul Lukowicz. Smart-mat: Recognizing and counting gym exercises with low-cost resistive pressure sensing matrix. pages 373–382, 09 2014.

[3] Natalia A Daz Rodrguez. *Semantic and Fuzzy Modelling for Human Behaviour Recognition in Smart Spaces*. IOS Press, AKA Verlag, 2016.

[4] Igor Pernek, Karin Anna Hummel, and Peter Kokol. Exercise repetition detection for resistance training based on smartphones. *Personal Ubiquitous Comput.*, 17(4):771–782, April 2013.

[5] Chuanjiang Li, Minrui Fei, Huosheng Hu, and Ziming Qi. Free weight exercises recognition based on dynamic time warping of acceleration data. pages 178–185, 2013.

[6] Keng-hao Chang, Mike Y. Chen, and John Canny. Tracking free-weight exercises. pages 19–37, 2007.

[7] Hristo Novatchkov and Arnold Baca. Machine learning methods for the automatic evaluation of exercises on sensor-equipped weight training machines. *Procedia Engineering*, 34:562 – 567, 2012. ENGINEERING OF SPORT CONFERENCE 2012.

[8] Eduardo Velloso, Andreas Bulling, Hans Gellersen, Wallace Ugulino, and Hugo Fuks. Qualitative activity recognition of weight lifting exercises. 2013.