GymCam: Detecting, Recognizing and Tracking Simultaneous Exercises in Unconstrained Scenes

RUSHIL KHURANA*, Carnegie Mellon University
KARAN AHUJA*, Carnegie Mellon University
ZAC YU, University of Pittsburgh
JENNIFER MANKOFF, University of Washington
CHRIS HARRISON, Carnegie Mellon University
MAYANK GOEL, Carnegie Mellon University

Worn sensors are popular for automatically tracking exercises. However, a wearable is usually attached to one part of the body, tracks only that location, and thus is inadequate for capturing a wide range of exercises, especially when other limbs are involved. Cameras, on the other hand, can fully track a user's body, but suffer from noise and occlusion. We present *GymCam*, a camera-based system for automatically detecting, recognizing and tracking multiple people and exercises simultaneously in unconstrained environments without any user intervention. We collected data in a varsity gym, correctly segmenting exercises from other activities with an accuracy of 84.6%, recognizing the type of exercise at 93.6% accuracy, and counting the number of repetitions to within ± 1.7 on average. GymCam advances the field of real-time exercise tracking by filling some crucial gaps, such as tracking whole body motion, handling occlusion, and enabling single-point sensing for a multitude of users.

CCS Concepts: • **Human-centered computing** → *Ubiquitous and mobile computing systems and tools*;

Additional Key Words and Phrases: exercise tracking; single-point sensing; health sensing; computer vision

ACM Reference Format:
Rushil Khurana, Karan Ahuja, Zac Yu, Jennifer Mankoff, Chris Harrison, and Mayank Goel. 2018. GymCam: Detecting, Recognizing and Tracking Simultaneous Exercises in Unconstrained Scenes. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 4, Article 185 (December 2018), 17 pages. https://doi.org/10.1145/3287063

1 INTRODUCTION

Regular physical workout improves well-being and reduces the risk of obesity, diabetes, and hypertension [12, 16, 26]. To maintain overall health and build strength, the Centers for Disease Control and Prevention (CDC) recommends adults to strength train at least twice a week[1]. However, despite the benefits of regular exercise,

*Both authors contributed equally to the paper.
[1]Centers for Disease Control and Prevention. Physical activity recommendations for adults: cdc.gov/physicalactivity/everyone/guidelines/adults.html

Authors' addresses: Rushil Khurana, Carnegie Mellon University, Pittsburgh, USA, rushil@cmu.edu; Karan Ahuja, Carnegie Mellon University, Pittsburgh, USA, kahuja@cs.cmu.edu; Zac Yu, University of Pittsburgh, Pittsburgh, USA, zac.yu@pitt.edu; Jennifer Mankoff, University of Washington, Seattle, USA, jmankoff@cs.washington.edu; Chris Harrison, Carnegie Mellon University, Pittsburgh, USA, chris.harrison@cs.cmu.edu; Mayank Goel, Carnegie Mellon University, Pittsburgh, USA, mayankgoel@cmu.edu.

2474-9567/2018/12-ART185 $15.00
https://doi.org/10.1145/3287063

Fig. 1. GymCam uses a camera to track exercises. ***(Top)*** Optical flow tracking motion trajectories of various points in the gym. Green showcases points classified as exercises and red showcases non-exercise points. ***(Bottom Left)*** Individual exercise points are clustered based on similarity to combine points belonging to the same exercise. ***(Bottom Right)*** For each exercise (cluster) GymCam infers the type of exercise and calculates the repetition count.

most people struggle to maintain steady progress. This failure is often attributed to lack of motivation and feedback [15, 17, 30].

One way to tackle lack of motivation is through gamification and tracking [6]. The ability to view personalized data enhances awareness and enables reflection of exercise regimens [23]. However, capturing and tracking a regimen is challenging. Manual tracking is most accurate, but this is tedious for end users. Thus, numerous commercial and academic efforts have focused on automatically tracking and quantifying physical activity, the most pervasive being step count captured by a worn device (*e.g.,* FitBit, Apple Watch, [27, 28]). Nowadays, consumer devices can track some cardio and strength-training exercises using special applications[2, 3]. These applications generally rely on a wearable's inertial measurement unit (IMU) to monitor *e.g.,* arm motion as users perform different exercises. Such techniques can be robust for some specific exercises, but fail for many others due to sensor placement where there is limited signal. For example, Maurer *et al.* found that detecting ascending

[2]Gymaholic: http://www.gymaholic.me
[3]Gymatic: https://itunes.apple.com/us/app/vimofit-auto-exercise-tracker/

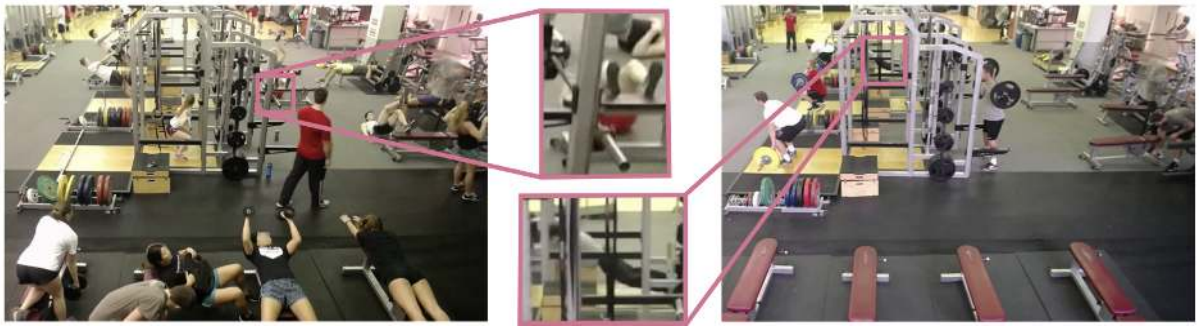

Fig. 2. In gym settings, user pose can be challenging to determine due to significant occlusion.

motion such as climbing stairs is more accurate when the IMU is attached to the bag than when attached to a person's shirt [19]. Similarly, data from a smartwatch is inadequate for exercises involving other parts of the body (*e.g.,* leg presses). An alternative is to instrument the exercise machine rather than the user, but that is too intrusive and also makes free-weight and body weight exercises harder to track. This presents a need for a method to robustly identify and track a wide range of exercises that a user might perform, while maintaining the seamlessness offered by wearable devices.

To this end, we present *GymCam* (Figure 1), a vision-based system that uses off-the-shelf cameras to automate exercise tracking and provide high-fidelity analytics, such as repetition count, without any user- or environment-specific training or intervention. Instead of requiring each user in the gym to wear a sensor on their body, *GymCam* is an external single-point sensing solution, *i.e.,* a single camera placed in a gym can track **all** people and exercises simultaneously. One camera-based approach would be to track body motions to detect user pose [25, 35]. However, these techniques are error-prone due to significant occlusion in gym settings (*e.g.,* Figure 2). Thus, instead of attempting to accurately estimate body keypoints (*i.e.,* skeletons), GymCam leverages the insight that *almost all repetitive motion in a gym represents some form of exercise.* Such motions can be readily captured by a camera, despite heavy occlusion, and used to segment and recognize various simultaneous exercises. We also found that it is extraordinarily rare for two separate people to exercise at the *exact* same rate and time, allowing for robust segmentation even when users are adjacent.

To develop and evaluate our machine learning algorithms, we collected data in our university's gym for five days. In total, we recorded 42 hours of video and annotated 597 different exercises. We did not record the number of gym users because our protocol required immediate anonymization of the data (*i.e.,* faces blurred). Users of the gym were informed that a research team was recording video, but there was no other interaction with participants, minimizing observer effects (*e.g.,* intentional or unintentional changes to their routine). We note this problem often affects research studies where users are aware they are part of an exercise tracking research study, and the evaluation setting is constrained [21]. We believe this paper presents the first truly unconstrained evaluation of exercise tracking.

The overall process of GymCam is as follows:

(1) Detect all exercise activities in the scene (acc. = 99.6%), then
(2) Disambiguate between simultaneous exercises (acc. = 84.6%), then
(3) Estimate repetition counts (± 1.7 counts)
(4) Recognize common exercise types (acc. = 93.6% for 5 most common exercise types).

2 RELATED WORK

Previous approaches for exercise recognition and tracking include using wearables, instrumenting equipment, and using computer vision. In this section, we discuss past work related to each of these key approaches.

2.1 Using Wearables

To help users improve their posture and form, IMUs have been used to track the user's exercise movements, and visually compare against optimal posture and movements for guidance [29]. Guidance relies only on tracking user movements, but our goals include recognizing exercises as well. Data from multiple IMUs has been used to build Hidden Markov Models that can identify up to 9 exercises with 90% accuracy [8, 9]. Similarly, myHealthAssistant [27, 28] used a subject-specific Bayesian classifier to recognize 13 exercises with 92% accuracy, and count repetitions. Such systems typically use noise-free exercise datasets and demonstrate initial feasibility, but often do not address the problem of reliably segmenting exercises from other activities, and recognizing exercises in noisy, real-world data.

In recent years, there has been a shift in focus to build a system that could work in-the-wild. Crema *et al.* used Linear Discriminant Analysis (LDA) on a stream of data captured by a wearable IMU to classify and count the exercises performed by a user [13]. The most relevant prior work that uses inertial sensors is Recofit [21], which builds upon Muelbauher *et al.*'s work [22]. Both approaches leverage the repetitive nature of strength training exercises and use features based on autocorrelation to detect and recognize exercise segments from a stream of inertial data. Recofit uses dimensionality reduction for orientation invariance to combat some issues associated with sensor placement, yet like the others, it is still not possible to track exercises that do not involve the limb where the sensor is worn. Furthermore, as these approaches require dedicated worn sensors per user, they are not as scalable as a single-point sensing solution.

2.2 Using Computer Vision

There have been numerous computer vision-based systems that have focused on rehabilitation. Approaches include using a depth camera to recognize in-home physiotherapy exercises [3], or to track exercises for tele-rehabilitation [2]. For such use cases, feedback on exercise movements is vital. YouMove pairs a Microsoft Kinect with a projector to build an interactive mirror that allows users to record and learn physical movements [1]. More recently, systems such as OpenPose [7, 35], V NECT [20] and Adversarial PoseNet [11] have employed deep learning based approaches to track pose from a RGB feed. However, such techniques typically rely on estimating the skeleton of a user, which is unreliable when users are far from cameras and there is significant occlusion, as one is likely to encounter in a multi-user gym setting. To the best of our knowledge, no prior vision-based exercise recognition and tracking system has been tested in-the-wild.

We draw our motivation from repetition based systems. Levy *et al.* used deep learning on synthetically generated data to count repetitive actions such as strumming a guitar or a bird flapping its wings [18]. A similar approach taken by Papoutsakis *et al.* identifies common subsequences (*e.g.,* exercising) in a sequence of actions (*e.g.,* exercising and walking), and counts the number of repetitions of the common subsequences [24]. However, we found that these approaches do not work well with multiple repetitive exercises in the same video.

2.3 Instrumentation of Equipment

Gym equipment can also be instrumented to monitor its use. For instance, Velloso *et al.* instrumented both the users and the equipment with IMUs [31, 32] and used the system to train a novice user. A combination of IMU data and motion data from Kinect was used to record an expert's exercise movements. A novice using the same system can then compare their and the expert's movements and assess their progress through a workout routine. Ding *et al.* attached RFID tags to dumbbells and used the Doppler shift of the backscatter signal to

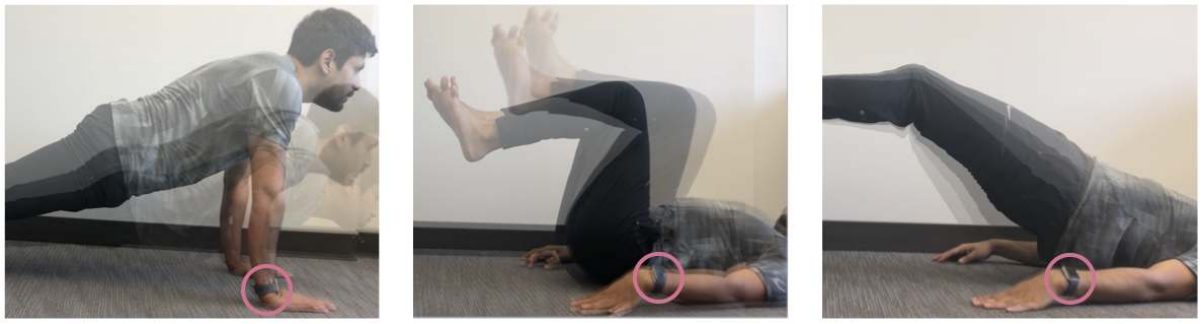

Fig. 3. A wristworn IMU (circled in the photos) is not ideally positioned to monitor many exercises.

recognize exercises, assess their quality and provide feedback to the users [14]. The system was able to achieve 90% accuracy in recognizing ten different gym activities. There are also many commercial products, such as Bazifit[4], which provides a sensor that integrates with equipment such as free weights and suspension trainers; the product provides repetition count, with posture and weight recommendations. Other products such as Bowflex[5] instrument dumbbells and other home gym products that sync with a user's phone via Bluetooth to log repetition counts.

3 THEORY OF OPERATION

We now discuss the underlying premise behind GymCam that allows it to: (1) detect motion, (2) cluster motions into separate exercises, and (3) identify and track individual exercises. GymCam leverages the insight that almost all repetitive motion in a gym represents some form of exercise. Even if a camera cannot see an entire person, it is still often able to see a small part of the body exhibiting repetitive motion, and can track that body part, linking it to an exercise later. However, when multiple users are exercising and potentially overlap in a video, it can be hard for camera-based systems to delineate the exact boundaries between the exercises – an issue worn sensors do not have to handle. Fortunately, we found it is extremely rare for two users to perform their exercises at exactly same time, speed, and phase (Figure 4). Thus, by calculating features that capture these dimensions, GymCam is able to differentiate between simultaneous exercises without any supervised training data.

Apart from distinguishing different users, there are other challenges when relying solely on repetitive motion tracking. Foremost, periodicity can be exhibited by a user's gait or warm up before starting an exercise. Secondly, when placed in an unconstrained environment, users tends to be less deliberate with between-exercise moments (*e.g.* fidgeting, stretching, walking). These interludes can be quite periodic, and thus indistinct from exercises. Moreover, in the unconstrained environment of a gym, users may challenge themselves (*e.g.* lift challenging weights). Morris *et al.* [21] observed that "*self-similarity [or periodicity] may break down in intensive strength-training scenarios. For this reason, more validation of intensive weightlifting is important future work.*" We believe that the only viable approach to solve the problem of variations in exercise and noisy human behavior, is to collect extensive training data in the user's actual workout environment without significant observer effect.

4 DATA COLLECTION

We collected data in the Carnegie Mellon University's varsity gym (seen in Figure 5) over a five-day period.

[4] https://bazifit.com/
[5] www.bowflex.com

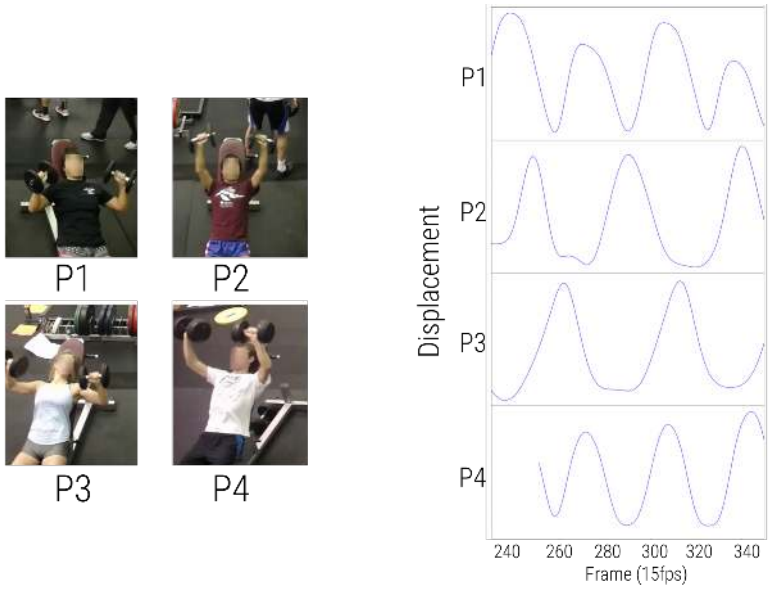

Fig. 4. Four users performing same exercise at different phase and frequency. *y*-axis shows displacement in terms of pixels.

4.1 Participants and Protocol

To ensure a wide, unobstructed view, we placed one camera on a wall at a height of approximately 4 meters. This placement was also inconspicuous, aiming to minimize observer effects (*e.g.*, users altering their warm-up or stretching routine, lifting usual weights). The university's Institutional Review Board and Department of Athletics officials agreed that as long as videos were immediately anonymized, we did not need signed consent from participants. Nonetheless, gym users were informed that a research team was recording anonymized videos and any questions, comments or objections should be raised to the gym staff (though none did). Thus, gym users were given no instructions regarding exercises, repetitions, breaks, etc., and is as close to unconstrained data collection as practically possible.

We used a Logitech C922 camera at a resolution of1920×1080 to record 15 frames per second (fps) video. We used a state-of-the-art face detection algorithm [36] to blur the faces of gym users and anonymize the videos. After dropping periods when the gym was empty, we had 42 hours of data spanning 5 days. We hand annotated 15 hours of this video, which contained 597 exercise instances.

4.2 Labeling

To annotate our captured videos, we tested several tools, but found that it was hard to efficiently annotate multiple exercises in the same frame while simultaneously recording their location, the repetition count and the type of exercise performed. In response, we built a custom annotation software (see Figure 5).

This software allows an annotator to load a video and use a mouse to draw bounding boxes around an exercise. The annotator can then edit the start and end frame for the annotation to correspond to the start and end of the exercise. The software also provides basic functions such as play/pause and fast forward at different rates. When the annotator indicates an end frame for a bounding box (*i.e.* an exercise), the software requests the repetition

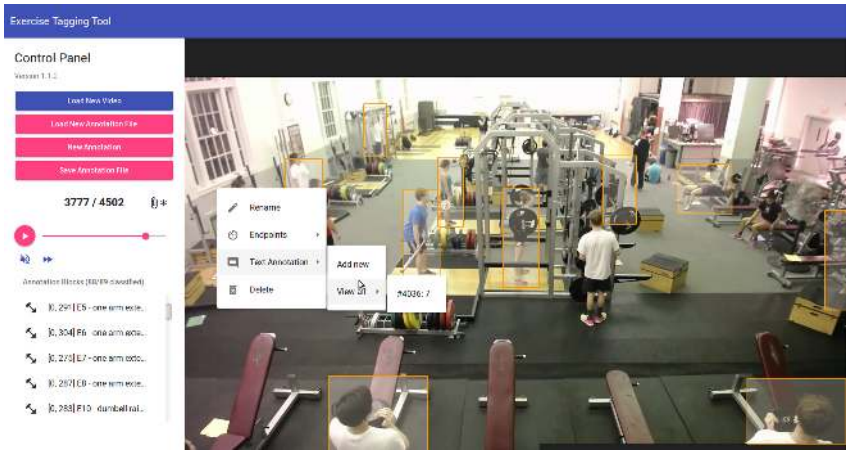

Fig. 5. Custom video annotation tool. Annotators drew the bounding boxes, and marked the start and end time for each exercise. They used the text annotation option to add exercise type and repetition count.

count. The annotator also has the ability to add text annotations to each bounding box, recording any notes of interest. We open-sourced our tool[6], and researchers can easily customize it for their specific use.

We recruited and trained two student annotators. They were instructed to not count any exercise with fewer than 3 repetitions. For exercises such as running on a treadmill, elliptical or other cardio machines, the annotators were simply asked to label "cardio" when asked for a repetition count. The annotators did not annotate ill-defined periods as exercises, but well-defined warm ups were labeled appropriately.

To allow multiple annotators to work simultaneously, we split each recorded video into 5 minute segments and the annotators processed these fragments in batches. If any exercises got split across two video segments, we counted them as two different exercises. This would never occur in a practical scenario since the input would be a continuous video stream. However, it was a procedural decision for us to ensure efficient parallelization of effort. The annotation software recorded all annotations as a JSON file. These files could be reloaded, along with their corresponding videos, to make any post-hoc changes to the annotations if needed. After all exercise start times, end times, and repetition counts were annotated, a single annotator labeled the exercise types. As there are several variants of each exercise, and different individuals may call one exercise by different names, this process ensured quality and consistent labels.

5 ALGORITHM

The goal of our work is to detect, identify, and track exercises, including when people are only partially visible. In fact, the real test of our approach is when the user is *barely* visible, but the camera can merely see a weight or a handlebar moving. Thus, GymCam starts by identifying all movements, and classifying them as repetitive or not. There could be several movements in a video that belong to the same exercise (*e.g.,* movement of different limbs, weights, and handlebar), so we combine similar repetitive exercise movements into exercise clusters. Next, for all motion trajectories in each identified cluster, we derive a combined trajectory to recognize the exercise type and estimate repetition count for that exercise (cluster). We will now describe our pipeline (Figure 6) in detail.

[6] https://github.com/zacyu/exercise-annotation-tool

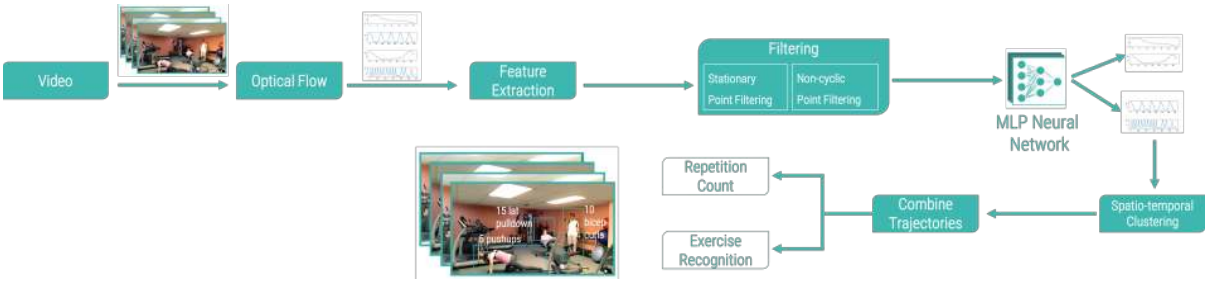

Fig. 6. GymCam system architecture.

5.1 Detecting Exercise Trajectories

To detect movement, we start by extracting optical flow trajectories from our video. We initially investigated OpenCV's implementation of Lucas-Kanade sparse optical flow [5]. However, the algorithm failed to track large, sudden movements and we switched to Wang *et al.*'s [33] dense optical trajectory extraction method to process all motion captured by the camera. For every video frame, the algorithm generates new keypoints, which are tracked continuously across frames to produce a motion trajectory. We found a keypoint max lifespan of 11 seconds was ideal for capturing several exercise repetitions, while also managing the processing time needed to track thousands of points in a video stream.

These motion trajectories are then converted into features and passed to a classifier. To limit the number of data points, we trim motion trajectories by removing stationary points (*i.e.*, any keypoint that moved less than 4-pixels between frames). We then normalize motion trajectories by their maximum translation and calculate a feature vector over an (empirically determined) sliding window of five seconds, with a stride of one second. Our feature vector consists of 27 features, a subset of which have also been used in prior work (see [4, 21]).

- **Frequency-based features:** Our working principle is that exercises are more periodic than non-exercises. We use frequency-based features to encode this property:
  - **Number of zero crossings:** We calculate the number of zero crossings of the keypoint motion trajectory, only in the x-axis, and only in the y-axis.
  - **Variance in zero crossings:** Exercises will be more periodic and have a lower variance in zero crossings than non-exercises.
  - **Dominant Frequency:** The dominant frequency of the signal calculated by frequency transformation.
  - **Autocorrelation:** Autocorrelation characterizes the periodicity of a signal.
  - **Maximum autocorrelation peak:** Higher value indicates higher periodicity.
  - **Frequency via autocorrelation:** The dominant frequency of signal determined via autocorrelation.
  - **Number of autocorrelation peaks:** Unusually high number of peaks indicate noisy signals, which are more likely to be non-exercises.
  - **Number of prominent peaks:** Represents the number of peaks higher than their neighboring peaks by a threshold (25%). A greater number of prominent peaks indicates higher periodicity.
  - **Number of weak peaks:** Similarly, we calculate the number of peaks smaller than their neighboring peaks by a threshold (25%). A greater number of weak peaks represents noisy and less periodic motion.
  - **Height of first autocorrelation peak after first zero crossing.** The height of the first peak after a zerolag provides an estimate of the signal's periodicity.
- **Non-frequency-based features:** Apart from the frequency-based features, we also calculate some non-frequency based features:

- **RMS:** The root-mean-square amplitude of the signal.
- **Span:** The span of the motion helps to characterize the intensity of the motion. We use overall span, and span in both x- and y-axes as features.
- **Displacement Vector:** Displacement helps us distinguish between exercises and other periodic motions such as walking. Non-exercise motions (such as walking) often have a higher displacement than exercise motions. We use the coefficients of the overall displacement vector, and displacement in both x- and y-axes, for a total of 9 features.
- **Decay:** Decay signifies the loss of intensity over time, a characteristic of exercise motions. We fit a line to the observed trajectory and use its coefficients as features.

We filter motion trajectories to bias our classifier to minimize false positives, at the cost of lower precision. This is because when an person is exercising, not all body parts may be involved in the motion. For example, legs do not move during a bicep curl, so a keypoint on a person's leg may be inside the bounding box created by an annotator, but would not be periodic. Similarly, improper form may cause a point to move while performing an exercise. Thus, not every motion trajectory inside an "exercise" bounding box is indicative of actual exercise motion. To protect the classifier from inaccurate training data, we filter motion trajectories with aggressive thresholds on frequency-based features. By filtering, we only provide the strongest and most representative examples of exercise trajectories to train our classifier. However, we do not perform any such filtering while validating the algorithm.

We use a multilayer perceptron (Scikit-Learn implementation with default hyperparameters) to classify every 5 second window segment of each keypoint trajectory as an exercise or not. The neural network optimizes the log-loss function using stochastic gradient descent. To smooth the output, we take a majority vote of three consecutive classifications and assign that as the output for each of those three classifications. Finally, we combine all consecutive *positive* classifications to construct a motion trajectory that was predicted as an exercise.

5.2 Clustering Points for Each Exercise

Exercises are often captured by many keypoint motion trajectories. Thus, our next step is to cluster keypoint motion trajectories into exercise groups. We perform clustering in two steps: (1) use spatio-temporal distribution of motion trajectories, and (2) use phase-differences between motion trajectories (Figure 4).

Given an exercise, the motion trajectories of its encapsulating keypoints will likely be close to one another in space *and* time. For space, we bootstrap the clustering algorithm by drawing bounding boxes next to each workout machine and station. Note, this only needs to happen once at the start of the system deployment (assuming machine and stations do not move). These boxes are non-overlapping and are representative of the exercise areas of the gym. Figure 7 shows these bounding boxes and also a distribution of exercises in our dataset.

Apart from spatial distribution, we also investigated the temporal separation between exercises. The exercise keypoints that overlap temporally as well as spatially are assigned to the same cluster. However, there is still a chance that exercises that are close to one another and occur together will be wrongly combined. To seperate such clusters, we also use phase information. For each cluster, we compute a phase-based similarity score between each trajectory-pair. For a pair of points that are not temporally co-located, the similarity is set to zero, and for others, the similarity is equal to the phase difference. We then threshold the phase difference (15 degrees) to assign a binary similarity score. In the end, we have a complete $N \times N$ adjacency matrix, where N denotes the number of motion trajectory points classified as an exercise. Given such a matrix, we calculate all connected graphs. Each graph denotes one exercise cluster associated with the nearest bounding box.

At the end of clustering, we combine the trajectories of all keypoints within a cluster to create a representative, average trajectory for further analysis (Figure 8). More specifically, we take the average of all the points within

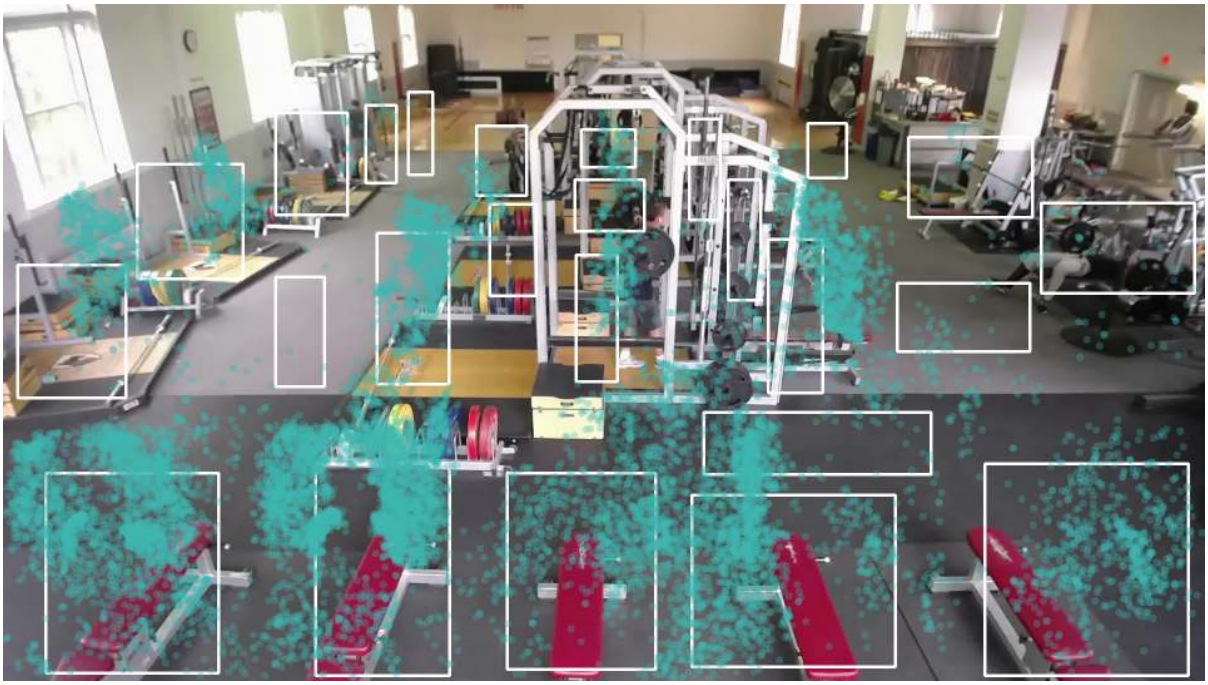

Fig. 7. An image of the gym with a distribution of all the exercise motions observed across all videos. The white boxes are the manually-drawn boxes to aid in clustering.

the cluster, accounting for the duration of each point, and smoothing it with a Hann window (size=1 second). This trajectory is used in our next process: exercise recognition and repetition count.

5.3 Repetition Count

Once a representative, average trajectory for each cluster is obtained from the previous step, we calculate the repetition count. To objectively disambiguate actual exercises from warm ups, we disregard any exercises that have less than five repetitions in the ground truth annotations. We train a multilayer perceptron regressor (Scikit-Learn; default hyperparameters) that uses our frequency-based features for each combined trajectory (as detailed in section 5.1), and outputs an estimate for the repetition count.

5.4 Exercise Recognition

Similar to repetition count, we leverage the cluster-average trajectory to infer the exercise type. We first quantize the trajectory into fixed-length segments as input to our classifier. We then run a sliding window (length five seconds, stride of one second) over this motion trajectory. Each window is passed to a multi-layer perceptron classifier (Scikit-Learn; default hyperparameters) to predict the exercise label, and we take a probability-based majority vote over all windows in the trajectory.

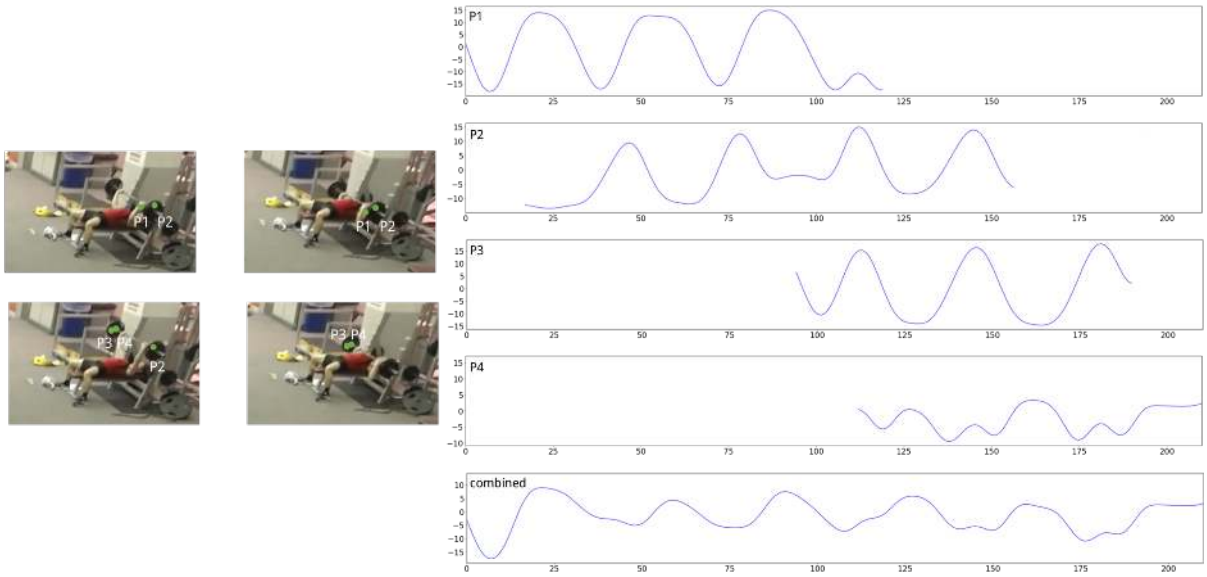

Fig. 8. Individual and combined trajectories for an exercise. The *x*-axis is frame numbers (15 fps).

6 RESULTS

6.1 Detecting Exercise Trajectories

We first report the results for distinguishing keypoint motion trajectories as *exercise* or *non-exercise*. For this, we performed a leave-one-day-out-cross-validation, which yielded a per-day, mean cross-validation exercise detection accuracy of 99.86%, with a mean false positive rate of 0.001% and precision of 23%. Again, we optimized our algorithm to reduce false positives at the expense of precision.

6.2 Clustering Points for Each Exercise

There are 597 distinct exercises in our ground truth annotated data. GymCam was able to accurately track 84.6% of these exercises. It also had a false positive rate of 13.5%, with most errors due to miscellaneous cyclic non-exercise motion such as warm-ups, rocking while seated, and walking.

6.3 Repetition Count

Repetition count accuracy helps in objective assessment of the time overlap between a predicted cluster and its corresponding ground truth match. We used 5-minute folds for cross-validation and achieved an accuracy of ±1.7 for counting repetitions with a standard deviation of 2.64.

6.4 Exercise Recognition

As discussed previously, our data was collected in an uncontrolled environment where participants were not instructed to perform a specific set of exercises, and so the distribution of exercise types was not uniform. Participants performed numerous atypical exercises and curating a balanced training set of conventional exercises from our data was challenging. We identified 18 common gym exercises and annotated their instances in our dataset (Table 1). We decided to disregard warm-up exercises because the annotator labeled many different

Table 1. Count of different exercise types

Exercise Type	Count
Squats	126
Deadlift	124
Benchpress	55
Arm Extension	56
Dumbell Benchpress	51
Shoulder Press	27
Pullups	24
Dumbell raises	18
Pushups	10
Cardio	9
Lat Pulldown	8
Dumbell Flies	8
Lying Dumbell Flies	4
Bicep Curl	3
Dumbell Raises	2
Tricep Extension	2
Dumbell Press	1
~~Warmup~~	~~69~~

exercises as "warm-up". The remaining 17 exercise types were classified with an accuracy of 80.6% with cross-validation across 5-minutes folds (Confusion Matrix: Figure 9). The five most frequently performed exercise types constituted roughly 69% of our data. We noticed that a lack of training data caused the less frequently seen exercises to be misclassified. Thus, if we only focus on the most frequent exercises, GymCam recognition accuracy increases to 93.6% (Confusion Matrix: Figure 10). Figure 11 shows the average identification accuracy as the number of recognized exercise types increase. This result indicates that our approach has the potential to differentiate between exercises based on our feature set, but a larger annotated dataset is needed.

7 DISCUSSION AND LIMITATIONS

GymCam provides an end-to-end pipeline to detect, track, and recognize multiple people and exercises in real-world settings, overcoming issues such as noisy data and visual occlusion. Based on our observations and experiences from building the system, we now discuss limitations to our approach. Besides completely missing an exercise, there are two types of major failure modes when an exercise is not recognized properly: (1) two exercises get clustered as a single exercise; and (2) one exercise gets split into two separate exercises as shown in Figure 12.

7.1 Reliance on Motion Differences for Clustering

When two or more individuals are exercising close to each other, and exhibit similar motion features such as phase and frequency, the individual exercise motion keypoints for each exercise may get combined into a single cluster. For example, it may occur during a group workout, when many people are roughly synchronized. Such cases are unavoidable, and should be expected to occur in uncontrolled environments. A potential solution is to augment GymCam with depth or body pose information to improve spatial clustering of nearby keypoints.

17 exercises	Arm extension	D. Flies	Benchpress	Squats	D. raises	Tricep extension	Deadlift	Pullup	Pushup	Lying D. flies	D. Benchpress	Bicep curl	Lat pulldown	D. raises	Shoulder press	D. press	Cardio
Arm extension	77	0	0	2	0	0	4	0	0	0	18	0	0	0	0	0	0
D. Flies	13	13	13	13	13	0	0	0	0	0	13	0	0	25	0	0	0
Benchpress	0	0	88	3	0	0	0	0	0	0	3	0	0	5	0	0	0
Squats	0	0	1	97	0	0	0	1	1	0	0	0	0	1	0	0	0
D. raises	0	0	0	0	0	0	0	0	0	50	50	0	0	0	0	0	0
Tricep extension	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0
Deadlift	0	0	1	0	0	0	96	0	0	0	0	0	0	0	3	0	0
Pullup	0	0	0	13	0	0	33	21	13	0	0	0	0	0	21	0	0
Pushup	0	0	0	70	0	0	10	0	20	0	0	0	0	0	0	0	0
Lying D. flies	50	0	0	0	0	0	0	0	0	50	0	0	0	0	0	0	0
D. Benchpress	10	0	2	0	2	0	0	0	0	0	84	0	0	2	0	0	0
Bicep curl	0	0	0	33	0	0	67	0	0	0	0	0	0	0	0	0	0
Lat pulldown	0	0	0	0	0	0	50	0	0	0	0	0	25	0	25	0	0
D. raises	0	0	15	0	0	0	5	0	0	0	10	0	0	65	5	0	0
Shoulder press	0	0	0	0	0	0	57	3	0	0	0	0	0	0	40	0	0
D. press	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0
Cardio	0	0	0	11	0	0	0	0	0	0	0	0	0	0	11	0	78

Fig. 9. Confusion matrix for exercise identification across 17 exercises.

5 exercises	Arm extension	D. Benchpress	Squats	Deadlift	Benchpress
Arm extension	77	16	4	4	0
D. Benchpress	14	84	0	0	2
Squats	0	0	99	0	1
Deadlift	0	0	0	99	1
Benchpress	0	3	5	0	92

Fig. 10. Confusion matrix for exercise identification across 5 exercises.

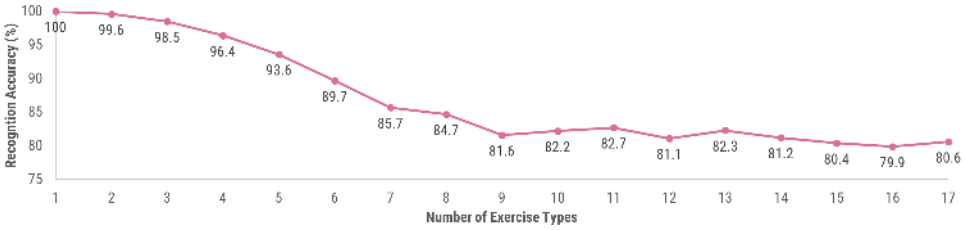

Fig. 11. Plot showing the accuracy of exercise recognition *vs.* number of exercise types

Additionally, higher framerate cameras could offer finer-grained phase information (*i.e.,* at 15 fps, participants synchronized to within $\tilde{7}0$ ms are indistinguishable).

Secondly, fatigue or improper exercise form may cause a person to show high variance between repetitions of the same exercise. Such irregular movements affects GymCam's performance as the algorithm relies on phase-based similarity of repetitions of the same exercise. In cases with irregular movements, the similarity

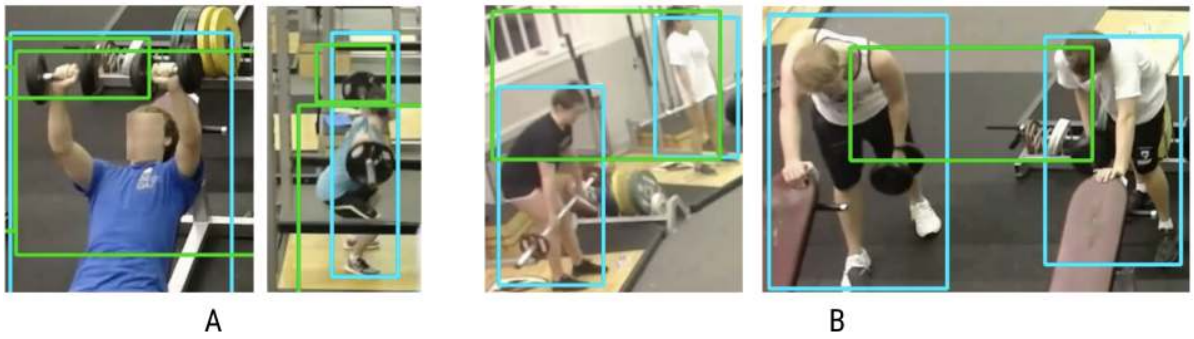

Fig. 12. The blue boxes represent ground truth, and the green boxes represent predicted exercises (clusters) in each image. **Left enclosed in red:** Examples of exercises where one exercise gets broken into two separate clusters. **Right enclosed in green:** Examples of exercises where two exercises get combined into a single cluster.

between trajectories decreases, which may introduce clustering errors and cause one exercise to be incorrectly broken into separate clusters.

7.2 Tracking Irregular Motions

The backbone of our algorithm is effective capture of motion trajectories across many keypoints. One of the most popular approaches to capture motion trajectories is Lucas-Kanade Optical Flow. While highly regarded and versatile, in our dataset, the algorithm failed to track exercise motion reliably. After investigating many failure cases, we realized that the algorithm failed to **continuously** track a keypoint if a person made sudden big movement, and instead initialized a new keypoint (not necessarily in the same location). Trajectories obtained from such keypoints do not contain sufficient information to classify repetitive and non-repetitive motions. To solve this problem, we used a variant of optical flow that is more resilient to irregular movements [33] and allows GymCam to classify individual motion trajectories as exercise/non-exercise with high accuracy (acc. = 99.6%). It highlights two potential points of failure in our approach: (1) choosing relevant keypoints to obtain motion trajectories in a frame; and (2) successfully capturing motion trajectories for the duration of the exercise.

7.3 User Identification

GymCam detects, recognizes, and tracks the exercise, but does not identify the user. Correlating the information between two sensors could be used to identify users. For example, Amazon Inc.'s Go Stores[7] combine information from users' phones, in-store cameras, and on-the-shelf sensors to track shoppers and their purchases. Similarly, practical deployment of GymCam could combine information from either a camera-based identification system or correlate data from a user's wearable device or use some form of manual identification step by the user (*e.g.,* using RFIDs [10] or QR codes [34] next to each workout station). It is arguable that relying only on pose-tracking might help in detecting the exercises *as well as* identifying the users. However, our pilot experiments showed that the current state-of-the-art pose tracking approaches were unable to handle the occlusion challenges.

7.4 Viewpoint Invariance

An ideal camera-based system would be viewpoint invariant and not require calibration for every camera position. Considering GymCam uses some spatial information, it is not entirely viewpoint invariant. For **exercise**

[7] http://www.wired.co.uk/article/amazon-go-seattle-uk-store-how-does-work

recognition, we use a simple bootstrapping and each area where users are likely to exercise is annotated. In a regular gym, where machines do not change positions, this annotation will be a one-time process. To detect **whether a user is exercising**, we only use time- and frequency-based features that do not change with position. Thus, it can be viewpoint invariant but we have not evaluated it formally.

7.5 Privacy

Using cameras enables accurate exercise tracking that is not limited to certain kinds of motion, but of course also raises privacy concerns. To mitigate this, the first step of our classification pipeline converts the raw video into optical flow trajectories. With this processed signal, GymCam can detect exercises, but sensitive user information is not easily recoverable. Indeed, with on-camera compute power, this could be the only data transmitted from the device, or perhaps the entire classification pipeline could be run locally.

7.6 Unconstrained Evaluation Environment

The primary insight of our algorithm is the periodicity of repetitions. However, as Morris *et al.* point out [21], periodicity is especially hampered during strength-training scenarios. When lifting challenging weights, for example, users often become fatigued and lose pace. Such issues are uncommon in cases where users participate in a user study, as the primary goal is not to challenge participants physically. In contrast, we developed and evaluated GymCam in a truly unconstrained environment, offering greater ecological validity and a more strenuous test. In addition to between-exercise movements and warm-ups, GymCam had to face an extra noise source: moving objects. For example, some participants in our dataset performed TRX (Total Body Resistance Exercise) workouts using ropes. In many cases, the suspended ropes kept oscillating after the exercise ended. Considering GymCam does not detect body pose and treats all repetitive motions equally, these rope oscillations resulted in a few falsely-recognized exercises.

8 CONCLUSION

With the surge in quantified self and health-focused wearable devices, the interest in exercise tracking is also rising. While tracking cardio exercises is relatively easy, tracking repetitive strength training exercise is still an outstanding problem. Most popular solutions involve using motion sensor-equipped wearables, but these devices are inadequate for capturing a wide range of exercises, especially ones involving other limbs. In this paper, we presented *GymCam*, a system that leverages a single camera to track a multitude of simultaneous exercises. GymCam relies on tracking motion and assumes that most repetitive motion in a gym are exercises in progress. To develop and evaluate our machine-learning algorithms, we collected data in CMU's varsity gym for five days. We segmented *all* concurrently occurring exercises from other activities in the video with an accuracy of 84.6%; recognized the type of exercise (acc.=92.6%) and counted the number of repetitions (± 1.7 counts). GymCam advances the field of real-time exercise tracking by filling some crucial gaps, such as tracking whole body motion, handling occlusion, and enabling single-point sensing for a multitude of users.

ACKNOWLEDGMENTS

We thank Priyanka Raja and Min Yan Beh for ideation and their help with early prototypes. We are also appreciative of the time, patience, and enthusiasm of Carnegie Mellon University's Department of Athletics staff.

REFERENCES

[1] Fraser Anderson, Tovi Grossman, Justin Matejka, and George Fitzmaurice. 2013. YouMove: enhancing movement training with an augmented reality mirror. In *Proceedings of the 26th annual ACM symposium on User interface software and technology*. ACM, 311–320.

[2] D Antón, A Goñi, A Illarramendi, et al. 2015. Exercise recognition for Kinect-based telerehabilitation. *Methods Inf Med* 54, 2 (2015).

[3] Ilktan Ar and Yusuf Sinan Akgul. 2014. A computerized recognition system for the home-based physiotherapy exercises using an RGBD camera. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 22, 6 (2014), 1160–1171.
[4] Abdelkareem Bedri, Richard Li, Malcolm Haynes, Raj Prateek Kosaraju, Ishaan Grover, Temiloluwa Prioleau, Min Yan Beh, Mayank Goel, Thad Starner, and Gregory Abowd. 2017. EarBit: Using Wearable Sensors to Detect Eating Episodes in Unconstrained Environments. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 37.
[5] Jean-Yves Bouguet. 2001. Pyramidal implementation of the affine lucas kanade feature tracker description of the algorithm. *Intel Corporation* 5, 1-10 (2001), 4.
[6] Dena M Bravata, Crystal Smith-Spangler, Vandana Sundaram, Allison L Gienger, Nancy Lin, Robyn Lewis, Christopher D Stave, Ingram Olkin, and John R Sirard. 2007. Using pedometers to increase physical activity and improve health: a systematic review. *Jama* 298, 19 (2007), 2296–2304.
[7] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In *CVPR*.
[8] Catherine B Chan, Daniel AJ Ryan, and Catrine Tudor-Locke. 2004. Health benefits of a pedometer-based physical activity intervention in sedentary workers. *Preventive medicine* 39, 6 (2004), 1215–1222.
[9] Keng-Hao Chang, Mike Y Chen, and John Canny. 2007. Tracking free-weight exercises. In *International Conference on Ubiquitous Computing*. Springer, 19–37.
[10] Rohit Chaudhri, Jonathan Lester, and Gaetano Borriello. 2008. An RFID based system for monitoring free weight exercises. In *SenSys*.
[11] Yu Chen, Chunhua Shen, Xiu-Shen Wei, Lingqiao Liu, and Jian Yang. 2017. Adversarial posenet: A structure-aware convolutional network for human pose estimation. *CoRR, abs/1705.00389* 2 (2017).
[12] Sheri R Colberg, Ronald J Sigal, Jane E Yardley, Michael C Riddell, David W Dunstan, Paddy C Dempsey, Edward S Horton, Kristin Castorino, and Deborah F Tate. 2016. Physical activity/exercise and diabetes: a position statement of the American Diabetes Association. *Diabetes Care* 39, 11 (2016), 2065–2079.
[13] C Crema, A Depari, A Flammini, E Sisinni, T Haslwanter, and S Salzmann. 2017. IMU-based solution for automatic detection and classification of exercises in the fitness scenario. In *Sensors Applications Symposium (SAS), 2017 IEEE*. IEEE, 1–6.
[14] Han Ding, Longfei Shangguan, Zheng Yang, Jinsong Han, Zimu Zhou, Panlong Yang, Wei Xi, and Jizhong Zhao. 2015. Femo: A platform for free-weight exercise monitoring with rfids. In *Proceedings of the 13th ACM Conference on Embedded Networked Sensor Systems*. ACM, 141–154.
[15] Gregory Heath, Elizabeth H Howze, Emily B Kahn, and Leigh Taylor Ramsey. 2001. Increasing physical activity. A report on recommendations of the Task Force on Community Preventive Services. *MMWR Recomm Rep* 50 (2001), 1–14.
[16] Ian Janssen and Allana G LeBlanc. 2010. Systematic review of the health benefits of physical activity and fitness in school-aged children and youth. *International journal of behavioral nutrition and physical activity* 7, 1 (2010), 40.
[17] Judy Kruger, Heidi Michels Blanck, and Cathleen Gillespie. 2006. Dietary and physical activity behaviors among adults successful at weight loss maintenance. *International Journal of Behavioral Nutrition and Physical Activity* 3, 1 (2006), 17.
[18] Ofir Levy and Lior Wolf. 2015. Live repetition counting. In *Proceedings of the IEEE International Conference on Computer Vision*. 3020–3028.
[19] Uwe Maurer, Asim Smailagic, Daniel P Siewiorek, and Michael Deisher. 2006. Activity recognition and monitoring using multiple sensors on different body positions. In *Wearable and Implantable Body Sensor Networks, 2006. BSN 2006. International Workshop on*. IEEE, 4–pp.
[20] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. 2017. VNect: Real-time 3D Human Pose Estimation with a Single RGB Camera. *ACM Transactions on Graphics* 36, 4, 14.
[21] Dan Morris, T Scott Saponas, Andrew Guillory, and Ilya Kelner. 2014. RecoFit: using a wearable sensor to find, recognize, and count repetitive exercises. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*. ACM, 3225–3234.
[22] Michael Muehlbauer, Gernot Bahle, and Paul Lukowicz. 2011. What can an arm holster worn smart phone do for activity recognition?. In *Wearable Computers (ISWC), 2011 15th Annual International Symposium on*. IEEE, 79–82.
[23] Rosemery O Nelson and Steven C Hayes. 1981. Theoretical explanations for reactivity in self-monitoring. *Behavior Modification* 5, 1 (1981), 3–14.
[24] Konstantinos Papoutsakis, Costas Panagiotakis, and Antonis A Argyros. 2017. Temporal Action Co-segmentation in 3D Motion Capture Data and Videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6827–6836.
[25] Kyle Rector, Cynthia L Bennett, and Julie A Kientz. 2013. Eyes-free yoga: an exergame using depth cameras for blind & low vision exercise. In *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility*. ACM, 12.
[26] Fernanda R Roque, Ana M Briones, Ana B García-Redondo, María Galán, Sonia Martínez-Revelles, Maria S Avendaño, Victoria Cachofeiro, Tiago Fernandes, Dalton V Vassallo, Edilamar M Oliveira, et al. 2013. Aerobic exercise reduces oxidative stress and improves vascular changes of small mesenteric and coronary arteries in hypertension. *British journal of pharmacology* 168, 3 (2013), 686–703.
[27] Christian Seeger, Alejandro Buchmann, and Kristof Van Laerhoven. 2011. Adaptive gym exercise counting for myHealthAssistant. In *Proceedings of the 6th International Conference on Body Area Networks*. ICST (Institute for Computer Sciences, Social-Informatics and

Telecommunications Engineering), 126–127.
[28] Christian Seeger, Alejandro Buchmann, and Kristof Van Laerhoven. 2011. myHealthAssistant: a phone-based body sensor network that captures the wearer's exercises throughout the day. In *Proceedings of the 6th International Conference on Body Area Networks*. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 1–7.
[29] Goran Šeketa, Dominik Džaja, Sara Žulj, Luka Celić, Igor Lacković, and Ratko Magjarević. 2015. Real-Time Evaluation of Repetitive Physical Exercise Using Orientation Estimation from Inertial and Magnetic Sensors. In *First European Biomedical Engineering Conference for Young Investigators*. Springer, 11–15.
[30] Martyn Standage, Joan L Duda, and Nikos Ntoumanis. 2003. A model of contextual motivation in physical education: Using constructs from self-determination and achievement goal theories to predict physical activity intentions. *Journal of educational psychology* 95, 1 (2003), 97.
[31] Eduardo Velloso, Andreas Bulling, and Hans Gellersen. 2013. MotionMA: motion modelling and analysis by demonstration. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1309–1318.
[32] Eduardo Velloso, Andreas Bulling, Hans Gellersen, Wallace Ugulino, and Hugo Fuks. 2013. Qualitative activity recognition of weight lifting exercises. In *Proceedings of the 4th Augmented Human International Conference*. ACM, 116–123.
[33] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu. 2013. Dense trajectories and motion boundary descriptors for action recognition. *International journal of computer vision* 103, 1 (2013), 60–79.
[34] Scott R Watterson, David Watterson, and Mark D Watterson. 2013. Systems and Methods to Generate a Customized Workout Routine. (Aug. 1 2013). US Patent App. 13/754,361.
[35] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. 2016. Convolutional pose machines. In *CVPR*.
[36] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters* 23, 10 (2016), 1499–1503.

Received May 2018; Revised August 2018; Accepted October 2018.