Using Learnable Physics for Real-Time Exercise Form Recommendations

Abhishek Jaiswal[1]**, Gautam Chauhan**[1]**, Nisheeth Srivastava**[1]

[1]Indian Institute of Technology, Kanpur
Kalyanpur, Kanpur 208016 India
{abhi.jaiswal44, gautamchauhan0412, nisheeths}@gmail.com

Abstract

Good posture and form are essential for safe and productive exercise. Even in physical gym settings, a trainer may not always be present to monitor an exerciser's form. Rehabilitation therapies as well as fitness workouts can therefore benefit from recommender systems that can provide real-time evaluation of exercise form and technique. In this paper, we present an algorithmic pipeline that can diagnose problems in exercises technique, and offer corrective recommendations for exercisers, with high sensitivity and specificity, in real-time. We use MediaPipe for pose recognition, count exercise repetitions(reps) using peak prominence detection and use a learnable physics engine to track motion evolution for each exercise. A test video is diagnosed based on deviations from the prototypical learned motion for a particular exercise using statistical learning. We evaluate our approach on six exercises to measure its effectiveness and find that it is considerable. These real-time interactive suggestions, counseled via low-cost equipment like smartphones, will allow exercisers to rectify potential mistakes in real-time making self-practice feasible while reducing the risk of workout injuries.

Introduction

Sedentary lifestyles and physical inactivity are prominent risk factors for cardiovascular diseases worldwide. Evidence also suggests that physical activity has dipped considerably over time (Ozemek, Lavie, and Rognmo 2019). Exercise improves life expectancy and has an effective therapeutic impact on physical and mental health (Jiménez-Pavón, Carbonell-Baeza, and Lavie 2020). Assistance in performing physical exercises (Fletcher et al. 2018) along with the promotion of the beneficial effects of physical activity (Lavie et al. 2019), therefore, has a vital role to play in improving health on a global scale.

Expert supervision in performing exercises is a scarce resource in the physical world. Even those under the supervision of a personal trainer need assistance while doing self-practice. Consequently, digital assistive technologies have emerged, playing an eminent role in improving accessibility to expert supervision for exercises. In gym settings, pose estimation has been extensively used to detect and assist in exercises (Ng 2020; Khurana et al. 2018). Similarly, many other projects have developed approaches for exercise recognition and repetition counting to benefit personal training (Chapron et al. 2018; Alatiah and Chen 2020; Spina et al. 2013).

Sensor-based methods (Velloso et al. 2013; Spina et al. 2013) have been prolifically used for pose assessment focusing on exercise detection (Chang, Chen, and Canny 2007; Šeketa et al. 2015; Seeger, Buchmann, and Van Laerhoven 2011), rep counting (Soro et al. 2019; Spina et al. 2013), incorrect pose diagnosis (Yurtman and Barshan 2014; Giggins, Sweeney, and Caulfield 2014; Lee et al. 2020; Kowsar et al. 2016) and recommendations (Zhao et al. 2014; Velloso et al. 2013; Spina et al. 2013). However, sensors can be obtrusive, expensive, and difficult to calibrate correctly, and so may be best suited for high-performance settings (Khurana et al. 2018). More appropriate for less intensive settings, vision-based methods (Wang et al. 2019; Liu and Chu 2020; Gharasuie, Jennings, and Jain 2021; Wang, Chen, and Duan 2021) have recently gained prominence due to advancement in deep learning techniques and mobile camera technology. This direction of research is very promising because it allows for the possibility of entirely sensor-free tracking of exercise performance benefiting a sizable audience.

At present, however, such proposals face considerable difficulties. Most vision-based approaches to exercise tracking work with predetermined heuristic parameters which vary across exercises and participants, requiring considerable hand-crafting (Liu and Chu 2020; Chen and Yang 2020). While vision-based approaches to exercise type recognition and rep-counting are plentiful, approaches that seek to track exercise form are limited to simple upper body exercises with relatively little body movement (Liu and Chu 2020; Kowsar et al. 2016; Chen and Yang 2020). Further, most such approaches offer exercise diagnoses retrospectively after processing entire recorded exercise sessions (Khurana et al. 2018; Soro et al. 2019).

We identify the over-general nature of the deep learning architectures used in vision-based exercise tracking pipelines as a key problem blocking progress in this area. Rather than use generic neural network architectures, we propose using a specific variety of neural networks, specifically designed to learn relationships between physical objects, as the base inference engine in such recommender systems. Using one such architecture - Interaction Network (Battaglia et al. 2016) - this paper describes a novel

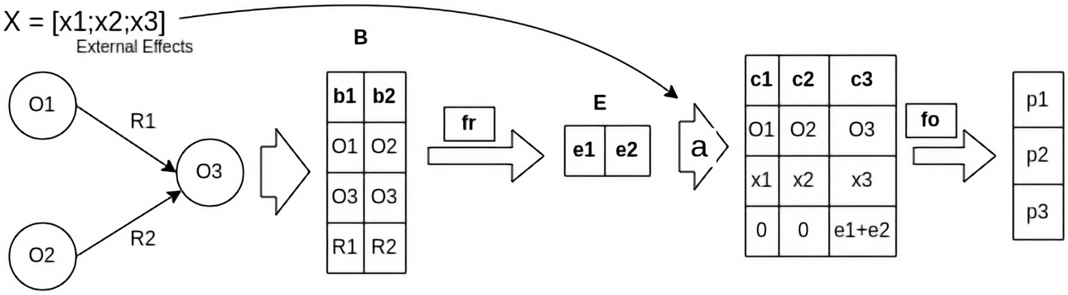


Figure 1: Example showing working of Interaction Network. b1,b2 are input to relation-centric function and c1,c2,c3 are input to object-centric function

recommender system for real-time exercise form correction. We show that our solution works with very high sensitivity and specificity for various full-body and upper-body exercises and provides recommendations early enough for the exerciser to make corrections.

We explain the working of our physics inference model in Physics prediction using Interaction Network section and the components of our recommender system in the Recommending Pose Corrections section. The value of our system is demonstrated through experiments on different exercises in the Empirical Evaluation section. Finally, we conclude by placing our proposal within the context of existing approaches towards exercise-relevant prediction and recommender systems, and identifying directions for future work, in the Discussion section.

Physics prediction using Interaction Network

Interaction Network (also referred to as IN) (Battaglia et al. 2016) is a general-purpose physics engine simulator that can reason about objects and their relationships using Graph Networks. The graph nodes represent the objects, and the edges represent their relations. The IN (Figure 1) takes the properties of the objects (mass, shape, position, velocity) and their relational properties (spring constant, restitution coefficient, internal forces) at the current time and predicts the next step dynamics. All this information is passed as matrices. The standard architecture builds upon a relation-centric function f_R and an object-centric function f_O. The f_R takes objects, and their relation attributes as input and predicts the effect of these interactions. The f_O predicts the next step dynamics for an object using its current state and aggregation of all the interaction effects it receives. They can accurately predict object dynamics over long trajectories. For more information, readers are requested to refer to the original IN paper (Battaglia et al. 2016).

For our use, we exploit the IN's rigid-body dynamics learning capability to model human biomechanical signals. Our graph network has N_O number of selected body landmarks (see Table 1 for some exercise-specific examples) and N_R relations between them. R_r and R_s are binary receiver and sender matrices of size N_O x N_R which index kinematic relationships between different body landmarks various exercises. D_S is the dimension of object attributes. O is an object matrix of size D_S x N_O. O_S is a matrix of sender landmarks of size D_S x N_R. Similarly, O_R is a matrix of receiver landmarks of size D_S x N_R and R_a is matrix of relation attributes of size D_R x N_R.

We feed sender-receiver landmark properties, along with their analogous relational attributes to the relation-centric function f_R, which predicts the associated effect matrix E of size D_E x N_R where D_E is the effect dimension.

$$E = f_R(O_S; O_R; R_a) \qquad \textit{where ; means concatenation}$$

Product of the effect matrix with the binary Receiver Matrix yields $\bar{E} = ER_r^T$ assimilating the net effects on each receiving objects in its columns. This, along with the object matrix O, is fed to the object centric function f_O to predict the next step dynamics P for each landmark.

$$P = f_O(O; \bar{E})$$

Recommending Pose Corrections

We first outline the overall methodology of our pipeline, followed by a detailed description of its sub-components. To begin with, we feed a recorded or a live video to our pipeline, which predicts frame by frame keypoints for 25 joints through Mediapipe Pose detection API (Bazarevsky et al. 2020). Depending on motion evolution, we select predetermined exercise specific landmark points (Table 1) followed by normalization and smoothing for physics modelling. The ML model predicts the motion rollouts for all the landmarks with visibility of only the initial rep state. Using these predictions, we calculate the Mean Squared Error (MSE) for individual landmarks. These errors are transformed to the frequency domain for further processing, as described in subsequent sections. Essentially, we use frequency domain information from the MSE signals to classify exercise reps as either correct or incorrect (in one of the multiple predefined modes of failure) using a Random Forest multi-class classifier.

Rep Counting using Peak Prominence

Each exercise consists of cyclic movements, which we exploit for repetition segregation. We track a landmark and find peaks in its periodic displacement plot. All peaks found need not necessarily be one separating a rep from another. To detect actual peaks, we find peaks' importance using peak prominence and use its standard deviation as a cutoff for our

	Landmarks					
Squats	Nose	Left Hip	Right Hip	Left Knee	Right Knee	-
Sit-ups	Nose	Shoulder	Hip	Knee	-	-
Push-ups	Shoulder	Hand	Hip	Toe	-	-
Lunges	Shoulder	Hip	Front Knee	Back Knee	Back Toe	Front Heel

Table 1: Body Landmarks for four full body exercises

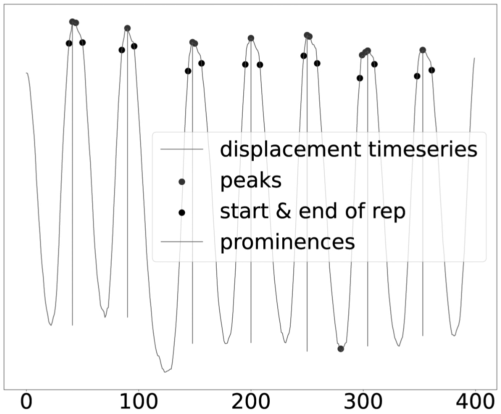


Figure 2: Peak prominence plot showing peaks in landmark's vertical displacement time series for rep counting. Timestamps are on X-axis and vertical displacement on y.

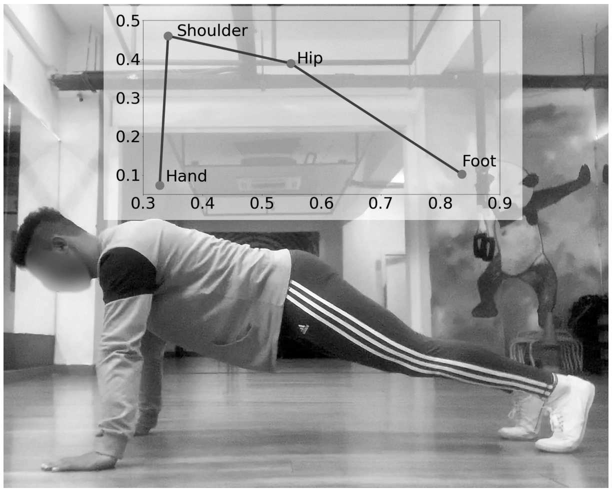


Figure 3: Push-ups - stick figure and corresponding video frame.

high pass filter. In real-life scenarios, discontinuous exercise reps are common as performers feel tired and distracted, many times having considerable break between successive reps. We remove that extraneous motion data between reps by evaluating a cutoff. Displacement values above the cutoff mark the start and stop of a valid rep. Figure 2 shows the result of rep counting for a single lunges video.

$$cutoff = peak - (prominence * 0.1)$$

Preprocessing

MediaPipe Pose Detection API provides 3D positional time series data for 25 body landmarks for each exercise. We transform these coordinates for unidirectional facing and use the resulting view along with the landmark's displacement amplitude to fix the representative landmarks for each exercise. These representative landmarks remain fixed for all computations of the exercise. We apply Locally Weighted Scatterplot Smoothing (LOWESS) to each time series (Cleveland 1981) to reduce noise, discard reps with significant pose estimation errors and Min-Max normalize the x,y coordinates to induce translational invariance. This preprocessing stage outputs stick figure representations for each exercise rep (Figure 3).

Learning exercise dynamics

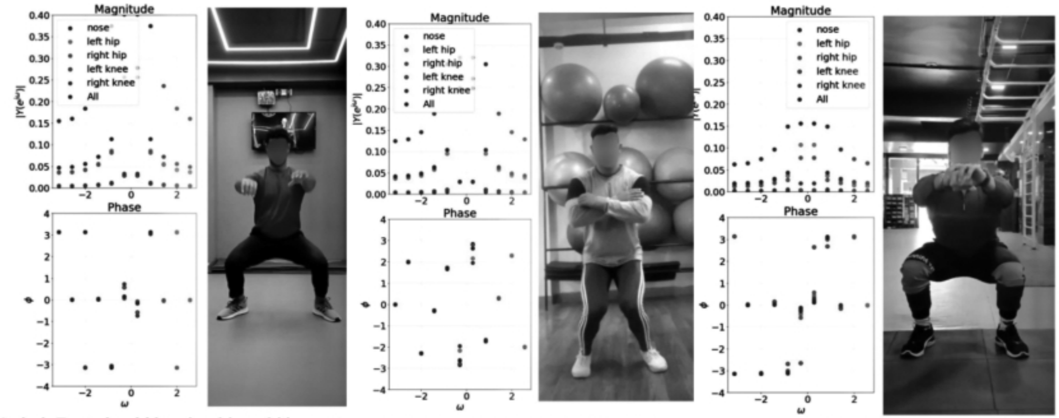
We represent body landmarks as nodes of a graph and the connections between them as the edges. There are two edges between each pair of joints, in the forward and backward direction. We consider only position and velocity as object attributes. For the relational attributes, we feed the joint to joint distances and angles, with a value of plus or minus one indicating the forward or the backward edge direction.

The objects' current position, velocity, and the relational-attribute matrix are fed to the IN to predict the next step's position and velocity using only the correct rep videos such that the model learns the physics of the exercise. For a test exercise rep, we feed the initial state of the representative landmarks as input and let the model predict trajectory roll-outs while repeatedly feeding actual relational attributes

Error Analysis and Rep Classification

The MSE time series emitted by the Interaction Network informs the next stage of our pipeline. We hypothesize that our physics engine predicts the correct method of performing an exercise, such that considerable deviation from it would hint at an incorrect rep. Further, the specific combination of MSE from different body components would hint at the specific way in which the exerciser is failing to perform the rep correctly.

To extract this information, we transform the MSE time series from all the representative landmarks to the frequency domain using the discrete-time Fourier transform (DTFT). DTFT provides magnitude and phase values for each time series. Conversion to the frequency domain helps in two ways. It gives a fixed-sized representation of the variable-length time series. It also helps to extract the features of the



Label: Feet should be shoulder-width apart

Label: Knee should not bend inward

Label: Correct

Figure 4: DTFT representations of error signature of different squats labels. Note how the phase plots for the incorrect squats differ from each other as well as from a correct squat systematically.

time series. This output of the DTFT, called the error signature (Figure 4), is a vector representation of an exercise rep of variable duration. For our case, we take the principal 11 amplitudes and the corresponding phase values to build the rep error signature. At the final stage of our pipeline, we use a Random Forest classifier for classification, operated in a multi-class classification setting with the error signatures from correct and incorrect classes.

Real Time Assessment

The users record themselves doing an exercise through our mobile application. This camera feed is provided to the MediaPipe Android API, which outputs the coordinates for body landmarks as a time series. These time series constitute of multiple reps. As a rep is identified, its data is sent to the server where our pipeline classifies it as correct or diagnoses it as a mistake of a particular type. A corrective message specific to the estimated diagnosis is displayed to the user through our application. For exercisers operating at normal tempo, this feedback arrives before their next rep is halfway complete.

Empirical Evaluation

Data

For full body exercises, we used a proprietary dataset obtained from E-Trainer Analytics Wizard Pvt. Ltd to conduct the evaluations reported in this paper. This dataset contains the front and side view of individuals doing four exercises - squats, push-ups, lunges and sit-ups. There is one correct class for each exercise, whereas incorrectly done exercises could belong to multiple classes.Incorrect videos were annotated with corrective suggestions by expert physical trainers. Each video consists of a single person doing multiple reps of an exercise as the central object in the frame. The total data consists of at least 150 reps for each exercise performed by seven exercisers.

We used a train test split of 60%-40% for training the classifier on incorrect classes. For the IN and correct class classification, we chose either 60%-40% or 80%-20% splits depending on the count of correct reps to maintain class balance among all the correct-incorrect classes.

To compare our approach against existing models of posture and form prediction, we also experimented with a publicly available dataset (Ng 2020), which contains annotated data for three exercises - front raise, bicep curl, and shoulder press. We show results for - shoulder press and front raise, as they easily integrate with our existing pipeline.

Methods

Our IN input consists of x, y position and velocity data of pre-selected landmarks and two top and bottom stationary reference points for all the experiments. The choice of landmarks depended on our understanding of the biomechanics of each exercise. Velocity is approximated as the difference between current and previous coordinates. Both the reference points have x coordinate 0.5 and y values as 0 and 1.

Our physics learning engine has two feed-forward neural networks - the relation-centric model f_R and the object-centric model f_O. f_R consists of 4 layers, 256 neurons each, with ReLU activation and f_O consists of 3 layers of 256 dimensions and ReLU activation. The output layer for each model had linear activation and .5 dropout was applied to each hidden layer. These networks were trained for 2500 epochs with early stopping. We optimized the network parameters using AdamW optimizer (Loshchilov and Hutter 2017) with 1cycle learning rate policy (Smith 2018) and a learning rate of 0.0003. This physics engine emits MSE times series for all the referenced landmarks. These are transformed into DTFT based error signatures. Our classifier categorizes these error signatures of each rep into correct or one of the incorrect classes. We tested different classifiers and found Random Forest to be most consistent across all the exercises. All the classifiers' hyperparameters were tuned using randomized search cross validation.

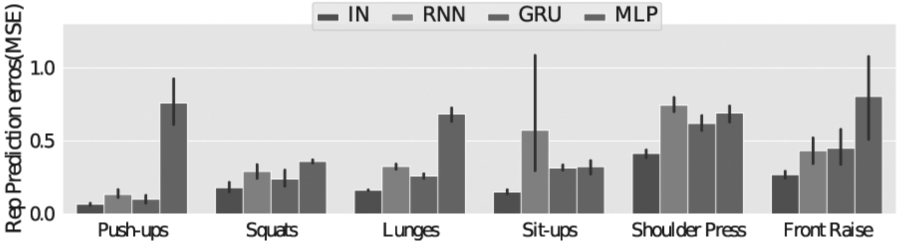


Figure 5: Average rollout prediction errors over exercise reps(MSE) for Baseline Models and Interaction Network. Even though MLP have good F1 scores (Table 3) in some cases, high prediction error makes their performance unreliable.

Baseline Comparisons

Looking at the previous works (Liu and Chu 2020; Chen and Yang 2020), we identify principle techniques applied for posture recommendation. We compare our physics learning pipeline against these well-known architectures by substituting them against the IN. Among available learnable dynamics predictors, we exploit Interaction Networks for their interpretability and simplicity. Our pipeline can also function with other motion predictors, to the degree they can accurately mimic the dynamics of the exercise. To test this hypothesis, we evaluate our model against several baselines.

Multi layer perceptron (MLP) Baseline: The MLP Baseline, fed with the same input data as a flattened vector, predicts the same future state dynamics as the IN. It has three 256-length hidden layers with Rectified Linear Unit (ReLU) activation. In principle, it has all the information available to learn the interaction dynamics, necessitating assimilation of relation indices implicitly without any scene factorization.

Recurrent Neural Network (RNN) and Gated Recurrent Unit (GRU) Baselines: The sequential modeling capacity of these networks equip them to model posture evolution. Both RNN and GRU architectures have three recurrent units with three features in the hidden state. They take the same flattened input vector as the MLP baseline. The hidden state output is fed to a fully connected layer to predict future dynamics. GRU is a more recent type of RNN capable of modeling longer-term dependencies, which can improve performance for many tasks.

Apart from the above deep learning based methods, several heuristics based methods have been traditionally used for exercise diagnosis. We compare our pipeline against Ng (2020) and Chen and Yang (2020), which utilize geometric thresholds, over features extracted from body landmarks, to detect incongruity in exercise reps. We also tested for variations of IN to estimate factors affecting its performance. The models below partly modify the IN architecture or its input for comparison.

Attribute Hidden IN: This model is an ablation modification of the IN. It has the same architecture, but the relation attribute matrix is a null matrix. It has enough information to calculate interaction attributes from position data, demanding them to calculate complex distance and inclination functions.

Independent object IN: This network has the relation-centric component f_R removed (The interaction effect vector is set to 0). It cannot model object-object interactions but can infer repeated cyclic motions of exercise.

Fully Connected(FC) and Global Connection(GC) IN: We also explore the variations of object-object connectivity to model different levels of interactions. FC IN connects each joint with every other joint for all the joints of a given exercise. It has the same capacity as the Interaction Network but takes additional irrelevant input. For GC IN, apart from the local relation dynamics, all the landmark points are connected to the top and bottom stationary points facilitating modelling of both local and global interactions, which may improve information propagation.

Results

We evaluate our diagnostic system on four vertical motion full body exercises (squats, push-ups, sit-ups, and lunges) and two upper body exercises (shoulder press and front raise) on three criteria: Rep Counting , Posture diagnosis, and Real-time prediction. Our peak-prominence based algorithm counted all reps in the full-body exercises with 100% accuracy. For each rep detected, we measured recommendation accuracy using weighted F1 scores in a multi-class classification setting.

Posture diagnosis Our results show that a pipeline endowed with physics learning capability can effectively differentiate between correct and incorrect exercise reps (Table 2 and Table 3). Physics-based design either beats all other baselines or gives a comparable performance. All the models perform equally well for fewer incorrect classes (e.g., Sit-ups, Shoulder Press) or for trivial variations in joints' motion range (e.g., Push-ups) i.e., when the exercise is relatively simple. Front Raise shows the most performance improvement (Table 2), which also has the most incorrect classes

Model	Shoulder Press*	Front Raise
Ng	0.90	0.77
Pose Trainer	0.49	0.76
MLP	**0.99 ± 0.01**	0.84 ± 0.04
RNN	0.99 ± .01	0.79 ± .05
GRU	0.95 ± .06	.80 ± .04
IN	.98 ± 0.01	**0.88 ± 0.03**

Table 2: Baseline comparisons for two upper body exercises(*Shoulder Press analysis for two incorrect classes).

Model	Squats	Push-ups	Lunges	Sit-ups
MLP	0.91 ± 0.02	0.98 ± 0.03	0.95 ± 0.03	**0.99** ± **0.01**
RNN	0.85 ± 0.04	0.98 ± 0.01	0.94 ± 0.01	0.98 ± 0.02
GRU	0.87 ± 0.03	0.98 ± 0.01	0.93 ± 0.02	0.94 ± 0.04
IN	**0.94** ± **0.02**	**0.98** ± **0.01**	**0.97** ± **0.01**	0.98 ± 0.01

Table 3: Baseline comparisons for four full body exercises (left). Classification results reported using weighted F1 score with standard deviations over five train-test runs.

Front Raise	NN	RNN	GRU	MLP
2 Classes	0.96 ± 0.03	0.93 ± 0.02	0.93 ± 0.06	0.96 ± 0.03
4 Classes	0.91 ± 0.03	0.90 ± 0.01	0.89 ± 0.05	0.91 ± 0.01
6 Classes	0.82 ± 0.04	0.79 ± 0.05	0.80 ± 0.04	.88 ± 0.03

Table 4: Baseline Comparison with exercise Complexity for Front Raise. Complexity of Classification increases with number of classes in a multi class classification setting. Methods that do not model exercise dynamics show significant performance drop as the number of classes increases.

(five) among all the tested exercises.

Since classification only indirectly measures performance, we also analyze MSE predictions(Figure 5). In all cases, a physics learning engine best describes motion dynamics. Surprisingly, MLP-based pipeline works well for multi-class classification, but their dynamics also significantly deviates from ground truth. Higher prediction errors cause more variation in its results, especially when the dynamics become complex and the number of classes increases (Table 4). Even for an ill-suited model, classification results can be good if the error signatures are separable, but such arbitrary performance gains do not scale well as the complexity of exercise increases.

To see what part of IN results in performance improvement, we perform an ablation study modifying several of its parts. The MSE predictions demonstrate (Figure 6) the utility of relational attributes for physics prediction. Variations in joint-to-joint modelling work equivalently well for the modifications we tried. Global connections support faster information propagation ensuing slightly better dynamics prediction but Stochastic interactions, with many intrinsic and extrinsic factors affecting the exercise (like fatigue, motivation, body pain, and distractions), contain the expected performance gain. Many hidden factors like joint-to-joint force and muscle tension also prevent the model from exploiting global propagation.

The Fully Connected IN model has more unnecessary information in irrelevant relations. Still, its performance is comparable, presumably because the IN learns to weigh the importance of critical relations for each exercise. This can also mitigate one of the limitations of our model i.e., obligation to an explicit relational matrix, provided that the pose estimation is accurate enough to detect all body landmarks. The independent object IN underperforms (Figure 6), owing to its incapability of modeling interactions.

Diagnosis latency In terms of latency, we find that the evaluation for the last rep is shown to the user approximately by the time the ongoing rep is half complete prompting the user to correct any mistakes in technique without much delay (see Table 5 for a quantitative summary of the latency for four full-body exercises across multiple videos and Table 6 for examples of real-time recommendations from our system).

Discussion

Self-training will become more prominent as people find less and less time to follow a dedicated gym routine. Our system works to help such users promote healthy living without compromising their daily activities. Our recommender system accurately detects whether a person is correctly performing an exercise or not, and offers real-time recommendations encouraging users to correct their exercise form. High F1 scores for all the six tested exercises support our basic contention that using a learnable physics engine for system inference permits high generalizability across a variety of exercises.

Our interactive system focuses on rep counting and diagnosis, assuming that the exercise performed is known (or is easily knowable). For instance, Moran et al. (2022) used MediaPipe Pose detection API (Bazarevsky et al. 2020) for pose recognition to detect the type of exercise someone is performing in real-time, a capability that could easily inform exercise type in our pipeline.

Several recently proposed systems (Ng 2020; Vyas 2019; Wang, Chen, and Duan 2021) have used state-of-the-art pose estimation techniques to craft heuristic joint angle thresholds for pose correction or feedback. Recently, real-time pose diagnosis was done by Alatiah and Chen (2020) using pre-

Exercise	Mean(sec)	Standard deviation(sec)
Squats	0.55	0.13
Sit-ups	0.39	0.07
Push-ups	0.36	0.11
Lunges	0.54	0.09

Table 5: Lag time(seconds) for new rep recognition.

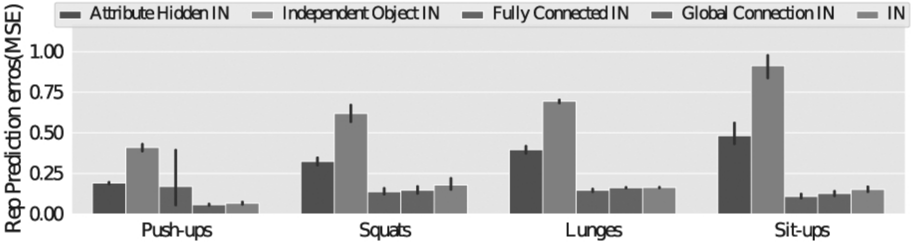


Figure 6: Average rollout prediction error over exercise reps(MSE) for Ablation Models and Interaction Network. Models without relation information experience significant drop in performance for dynamics prediction.

Data type	Link
IN rollouts	https://tinyurl.com/rolloutsDemo
Exercise Demo	https://tinyurl.com/realTimeDemo

Table 6: Video Simulations of Interaction Network predictions rolled out over time and of exercise sessions diagnosed using our system in real-time.

calculated parameters like a range of motion and major joint angles. Similarly, Ying et al. (2021) developed a personal training system that compares the real-time input with the features of pre-stored correct exercises to detect incorrect moves. In such systems, only binary correct/incorrect feedback is offered rather than corrective recommendations.

More granular diagnoses are possible in a system recently proposed by Liu and Chu (2020), who designed three domain-based joint angle indicators, modeled each rep using an RNN to learn these indicators and visually indicate the mistake location for two upper body dumbbell exercises. For these two simple exercises, they get classification accuracy above 90%. However, their approach requires per frame annotation for training. Similarly, Gharasuie, Jennings, and Jain (2021) developed a low-cost system using AlphaPose(Fang et al. 2017) based arm angles for upper-body exercises to count reps using smartphone cameras. They trained their system on data recorded in the gym and calculated various exercise phase parameters to estimate user fatigue levels indirectly. While such heuristic joint-angle based methods provide helpful textual feedback in some instances, they tend to work well only for isolation arm exercises (where only a few joints are involved) and do not achieve significant diagnostic accuracy without extensive frame-level annotation. Our system, in contrast, with a more sophisticated inference engine, works well for compound exercises using only video-level annotation.

Closer technically to our approach, Pose Trainer (Chen and Yang 2020) uses OpenPose (Cao et al. 2021) based pose estimation on dumbbell exercises along with Dynamic Time Warping against template moves for rep diagnosis. They use angular heuristics for exercise improvement feedback. Similarly, AI Coach (Wang et al. 2019) analyses sports trajectories based on angles between joint key points to match against bad poses pre-annotated by experts. In these systems, for all identified bad pose frames, an exemplar-based video is recommended by the system. This is in contrast with our system, wherein holistic, body-focused textual feedback is offered to users.

Thus, to summarize, this paper presents a novel system for recommending form corrections to people performing rep-based exercises in real-time with high precision. We introduce the use of learnable physics engines to model body physics, a task for which they are very well-suited. The success of our physics model permits downstream classifiers to accurately diagnose modes of failure of exercises using differential prediction error residuals between the model prediction and actual observations. Empirical evaluations show that our system diagnoses defective techniques in complex full-body exercises with high sensitivity and specificity. We expect the adoption of such interactive systems to help healthcare providers scale up access to supervised physical exercise.

We conclude with a brief exploration of the limitations of our system, and possible directions for future work. The most critical technical limitation of the present system is its reliance on pre-defined relational attributes for each exercise's Interaction Network. These attributes depend on the nature of human biomechanics and must be decided beforehand. Learning relational attributes from data could improve this performance even further, a clear direction for future work. Our system is currently tested only for exercises with significant vertical periodicity, an artifact of our peak-prominence based rep-counting scheme, though vertical periodicity also exists in many other exercises. Replacing this with a more sophisticated rep-counting method could extend our system's capabilities to a more general set of exercises. In particular, given the known diagnostic value of gait analysis in predicting health outcomes for the elderly (Cesari et al. 2005; Verghese et al. 2009), extending this system's digital diagnostic capabilities to monitoring and diagnosing gait-related problems presents a very promising direction for future work.

References

Alatiah, T.; and Chen, C. 2020. Recognizing exercises and counting repetitions in real time. *arXiv preprint arXiv:2005.03194*.

Battaglia, P.; Pascanu, R.; Lai, M.; Jimenez Rezende, D.; et al. 2016. Interaction networks for learning about objects, relations and physics. *Advances in neural information processing systems*, 29.

Bazarevsky, V.; Grishchenko, I.; Raveendran, K.; Zhu, T.; Zhang, F.; and Grundmann, M. 2020. Blazepose: On-device real-time body pose tracking. *arXiv preprint arXiv:2006.10204*.

Cao, Z.; Hidalgo, G.; Simon, T.; Wei, S.-E.; and Sheikh, Y. 2021. OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 43(01): 172–186.

Cesari, M.; Kritchevsky, S. B.; Penninx, B. W.; Nicklas, B. J.; Simonsick, E. M.; Newman, A. B.; Tylavsky, F. A.; Brach, J. S.; Satterfield, S.; Bauer, D. C.; et al. 2005. Prognostic value of usual gait speed in well-functioning older people—results from the Health, Aging and Body Composition Study. *Journal of the American Geriatrics Society*, 53(10): 1675–1680.

Chang, K.-h.; Chen, M. Y.; and Canny, J. 2007. Tracking free-weight exercises. In *International Conference on Ubiquitous Computing*, 19–37. Springer.

Chapron, K.; Plantevin, V.; Thullier, F.; Bouchard, K.; Duchesne, E.; and Gaboury, S. 2018. A more efficient transportable and scalable system for real-time activities and exercises recognition. *Sensors*, 18(1): 268.

Chen, S.; and Yang, R. R. 2020. Pose Trainer: correcting exercise posture using pose estimation. *arXiv preprint arXiv:2006.11718*.

Cleveland, W. S. 1981. LOWESS: A program for smoothing scatterplots by robust locally weighted regression. *American Statistician*, 35(1): 54.

Fang, H.-S.; Xie, S.; Tai, Y.-W.; and Lu, C. 2017. Rmpe: Regional multi-person pose estimation. In *Proceedings of the IEEE international conference on computer vision*, 2334–2343.

Fletcher, G. F.; Landolfo, C.; Niebauer, J.; Ozemek, C.; Arena, R.; and Lavie, C. J. 2018. Promoting physical activity and exercise: JACC health promotion series. *Journal of the American College of Cardiology*, 72(14): 1622–1639.

Gharasuie, M. M.; Jennings, N.; and Jain, S. 2021. Performance Monitoring for Exercise Movements using Mobile Cameras. In *Proceedings of the Workshop on Body-Centric Computing Systems*, 1–6.

Giggins, O. M.; Sweeney, K. T.; and Caulfield, B. 2014. Rehabilitation exercise assessment using inertial sensors: a cross-sectional analytical study. *Journal of neuroengineering and rehabilitation*, 11(1): 1–10.

Jiménez-Pavón, D.; Carbonell-Baeza, A.; and Lavie, C. J. 2020. Physical exercise as therapy to fight against the mental and physical consequences of COVID-19 quarantine: Special focus in older people. *Progress in cardiovascular diseases*, 63(3): 386.

Khurana, R.; Ahuja, K.; Yu, Z.; Mankoff, J.; Harrison, C.; and Goel, M. 2018. GymCam: Detecting, recognizing and tracking simultaneous exercises in unconstrained scenes. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(4): 1–17.

Kowsar, Y.; Moshtaghi, M.; Velloso, E.; Kulik, L.; and Leckie, C. 2016. Detecting unseen anomalies in weight training exercises. In *Proceedings of the 28th Australian Conference on Computer-Human Interaction*, 517–526.

Lavie, C. J.; Ozemek, C.; Carbone, S.; Katzmarzyk, P. T.; and Blair, S. N. 2019. Sedentary behavior, exercise, and cardiovascular health. *Circulation research*, 124(5): 799–815.

Lee, J.; Joo, H.; Lee, J.; and Chee, Y. 2020. Automatic classification of squat posture using inertial sensors: Deep learning approach. *Sensors*, 20(2): 361.

Liu, A.-L.; and Chu, W.-T. 2020. A posture evaluation system for fitness videos based on recurrent neural network. In *2020 International Symposium on Computer, Consumer and Control (IS3C)*, 185–188. IEEE.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Moran, A.; Gebka, B.; Goldshteyn, J.; Beyer, A.; Johnson, N.; and Neuwirth, A. 2022. Muscle Vision: Real Time Keypoint Based Pose Classification of Physical Exercises. *arXiv preprint arXiv:2203.12111*.

Ng, J. 2020. *Posture evaluation for variants of weight-lifting workouts recognition*. Ph.D. thesis, UTAR.

Ozemek, C.; Lavie, C. J.; and Rognmo, Ø. 2019. Global physical activity levels-Need for intervention. *Progress in cardiovascular diseases*, 62(2): 102–107.

Seeger, C.; Buchmann, A. P.; and Van Laerhoven, K. 2011. myHealthAssistant: a phone-based body sensor network that captures the wearer's exercises throughout the day. In *BodyNets*, 1–7.

Šeketa, G.; Džaja, D.; Žulj, S.; Celić, L.; Lacković, I.; and Magjarević, R. 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*, 11–15. Springer.

Smith, L. N. 2018. A disciplined approach to neural network hyper-parameters: Part 1–learning rate, batch size, momentum, and weight decay. *arXiv preprint arXiv:1803.09820*.

Soro, A.; Brunner, G.; Tanner, S.; and Wattenhofer, R. 2019. Recognition and repetition counting for complex physical exercises with deep learning. *Sensors*, 19(3): 714.

Spina, G.; Huang, G.; Vaes, A.; Spruit, M.; and Amft, O. 2013. COPDTrainer: a smartphone-based motion rehabilitation training system with real-time acoustic feedback. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, 597–606.

Velloso, E.; Bulling, A.; Gellersen, H.; Ugulino, W.; and Fuks, H. 2013. Qualitative activity recognition of weight lifting exercises. In *Proceedings of the 4th Augmented Human International Conference*, 116–123.

Verghese, J.; Holtzer, R.; Lipton, R. B.; and Wang, C. 2009. Quantitative gait markers and incident fall risk in older adults. *The Journals of Gerontology: Series A*, 64(8): 896–901.

Vyas, P. 2019. *Pose estimation and action recognition in sports and fitness*. Master's thesis, San Jose State University.

Wang, J.; Qiu, K.; Peng, H.; Fu, J.; and Zhu, J. 2019. Ai coach: Deep human pose estimation and analysis for personalized athletic training assistance. In *Proceedings of the 27th ACM International Conference on Multimedia*, 374–382.

Wang, L.; Chen, Y.; and Duan, W. 2021. Monocular Keypoint based Pull-ups Measurement on Strict Pull-ups Benchmark. In *2021 4th International Conference on Computer Science and Software Engineering (CSSE 2021)*, 307–311.

Ying, H.; Liu, T.; Ai, M.; Ding, J.; and Shang, Y. 2021. AICoacher: A System Framework for Online Realtime Workout Coach. In *Proceedings of the 29th ACM International Conference on Multimedia*, 3787–3790.

Yurtman, A.; and Barshan, B. 2014. Automated evaluation of physical therapy exercises using multi-template dynamic time warping on wearable sensor signals. *Computer methods and programs in biomedicine*, 117(2): 189–207.

Zhao, W.; Lun, R.; Espy, D. D.; and Reinthal, M. A. 2014. Realtime motion assessment for rehabilitation exercises: Integration of kinematic modeling with fuzzy inference. *Journal of Artificial Intelligence and Soft Computing Research*, 4(4): 267–285.