AIFit: Automatic 3D Human-Interpretable Feedback Models for Fitness Training

Mihai Fieraru[1] Mihai Zanfir[1] Silviu Cristian Pirlea[1]
Vlad Olaru[1] Cristian Sminchisescu[2,1]

[1]Institute of Mathematics of the Romanian Academy, [2]Lund University
[1]{firstname.lastname}@imar.ro, [2]cristian.sminchisescu@math.lth.se

Abstract

I went to the gym today, but how well did I do? And where should I improve? Ah, my back hurts slightly... *User engagement can be sustained and injuries avoided by being able to reconstruct 3d human pose and motion, relate it to good training practices, identify errors, and provide early, real-time feedback. In this paper we introduce the first automatic system,* AIFit*, that performs 3d human sensing for fitness training. The system can be used at home, outdoors, or at the gym.* AIFit *is able to reconstruct 3d human pose, shape, and motion, reliably segment exercise repetitions, and identify in real-time the deviations between standards learnt from trainers, and the execution of a trainee. As a result, localized, quantitative feedback for correct execution of exercises, reduced risk of injury, and continuous improvement is possible. To support research and evaluation, we introduce the first large scale dataset,* Fit3D*, containing over 3 million images and corresponding 3d human shape and motion capture ground truth configurations, with over 37 repeated exercises, covering all the major muscle groups, performed by instructors and trainees. Our statistical coach is governed by a global parameter that captures how critical it should be of a trainee's performance. This is an important aspect that helps adapt to a student's level of fitness (i.e. beginner vs. advanced vs. expert), or to the expected accuracy of a 3d pose reconstruction method. We show that, for different values of the global parameter, our feedback system based on 3d pose estimates achieves good accuracy compared to the one based on ground-truth motion capture. Our statistical coach offers feedback in natural language, and with spatio-temporal visual grounding.*

1. Introduction

In nowadays busy, high pressure working environments, fitness is essential in order to stay in shape, maintain balance, enhance the immune system, and prevent the emergence of chronic diseases. It is also critical for the elderly in order to maintain mobility, combat anxiety, and slow-down aging. This has increasingly resonated with the broad public. Besides the growing number of standard gym subscriptions, there are emergent online services (e.g. Peloton, Mirror or ClassPass, among others) that aim to bring fitness at home. Some trainers run popular Youtube fitness channels or apps (e.g. Athlean-X, The Fitness Marshall, Blogilates, etc.), and public interest spurs billions of searches and views of such instructional video each year. However, whether at the gym, at home, or outdoors, fitness enthusiasts face some of the same outstanding challenges: making sure they exercise correctly, avoid injury, gain insights into their progress, maintain motivation to get the job done, and ultimately have fun. Even when personal trainers are available, their engagement is typically limited to the time spent with the trainee at the gym. In practice, personal trainers may need to joggle between different clients, making it difficult to provide the continuous observation, feedback, and encouragement their clients sometimes need in order to progress. This naturally raises the question whether personal experience can be improved by leveraging recent advances in 3d human sensing and AI. To complement human trainers, in this paper we propose *AIFit*, the first AI-enhanced training system for fitness. The system is able to reconstruct 3d human pose over time, count repetitions, and automatically provide localized feedback, visually grounded in images of the trainee, and phrased in natural language displayed on a screen. In order to support research and evaluation, we introduce *Fit3D*, a large-scale dataset of over 3 million images and ground truth 3d motion capture poses, collected from 13 subjects (including one licensed fitness instructor and one advanced fitness subject), observed by 4 different RGB cameras, together with 3d scans of each subject. The dataset features 37 exercises consisting of simple and compound motions, covering all major muscle groups and articulation types, including, among many others, warm-ups,

barbells, dumbbells, push-ups, or yoga.

Our proposed methodology includes large-scale monocular and multi-view evaluation of 3d human pose reconstruction for fitness training using *Fit3D*, models for automatic identification of exercise repetitions, as well as methods to compare instructors' and trainees' performances according to statistical policies defined over mined features (passive and active) defining the exercise, and carrying most of its motion energy. Our statistical coach is governed by a global parameter ranging between 0 and 1 that models how critical it is in regard to a student's performance: 0 - very critical, 1 - very relaxed. In practice, the parameter helps the coach adapt to a student's level of fitness (i.e. beginner vs advanced vs expert) or to the expected accuracy of the underlying 3d pose reconstruction method. We show that, for different values of this parameter, our feedback system based on 3d pose estimates achieves high accuracy when compared to one based on ground-truth motion capture 3d poses. Finally and importantly, our statistical coach provides easy to understand, visually grounded spatio-temporal feedback, in natural language. A system overview is shown in fig. 1.

2. Related Work

Visual human sensing has been extensively studied[24, 17, 16, 36, 22, 18, 37, 38, 14, 31, 3, 12, 7, 8]. Applications exist in many domains such as automotive industry [21, 25], fashion industry [9, 30], activity recognition [28] and many others. One particular area which was less considered by research is reconstruction and activity analysis for fitness training, as also addressed in this work.

AI Fitness Training It has been well established [15, 1, 23, 19, 32] that fitness and exercising have a high impact on the physical and mental health of humans, motivating the need for methodologies and scientific studies to evaluate the correctness of physical exercises and to provide feedback.

Several prior studies focus on this topic, but lack a detailed 3d (temporal) analysis in terms of all major body joints and muscle groups, as well as feedback. Most previous work [11, 27, 39, 2] operates on real-time sensory data collected e.g. from IMUs or Microsoft Kinect [29]. [27] uses a wearable IMU device to monitor leg activity and posture, and generates a report at the end of each day. [2] introduce a dataset of physical activities captured using Microsoft Kinect. However, these are limited to basic motions such as walking, up-and-go and step exercises, without a broader coverage of motions and muscle groups.

Other methods operate directly on raw RGB images, without the use of additional sensors. [34] proposes a fitness feedback method based on SMPL model fitting. Given a single frame from a fitness exercise, the subject's fitted SMPL body shape is compared to a correct reference shape. This approach neither operates in the temporal domain nor does it provide interpretable feedback. [33] proposes a deep learning framework trained on annotated data, which predicts 2d human body poses in outdoor sport videos, associates them temporally, and provides training suggestions for incorrect poses. Since the analysis is performed in 2d (as opposed to 3d), precise feedback is not always possible.

Repetition Segmentation Segmenting a video into repetition intervals is a well-studied problem. It is usually applied to class-agnostic actions and is split into periodicity detection (determining if a frame is part of a repeating action or not) and repetition counting (predicting the count number of an action in a video). Differently from previous work, our approach offers a precise segmentation of each repetition in a video, without any particular assumption on the actual length of each interval. Several methods have been proposed, exploiting auto-correlation [5] or taking use of optical flow features under Wavelet transforms [26]. More recently, two methods [13, 6] introduce deep learning models that directly predict the period of a repetition. Several class-agnostic video repetition datasets are also introduced: QUVA [26] for repetition counting, PERTUBE [20] for periodicity detection and Countix [6] for both tasks. Note that, as opposed to Fit3D, Countix does not contain the bounds of each repetition interval, but only the bounds of the periodic subsequence and its repetition count.

3. Fit3D Dataset

To assist fitness training and to stimulate research in the area, we record a 3d motion capture dataset featuring 13 human subjects performing fitness exercises. Among them, there is one licensed fitness instructor (considered the reference for correctness in exercise execution), while the rest are considered trainees of various levels of skill.

We use a VICON motion capture system consisting of 12 motion cameras, synchronized with 4 RGB cameras. The capture process involves reflective markers which are affixed to either the subject's skin or clothing. All subjects are dressed in gym-like attires which usually fit tightly on the body. During the recordings, the subjects use several typical gym objects: 2 dumbbells, a barbell, and a rubber band. A low-height, footless table is used to ease the difficulty of the exercises involving lifting the barbell.

The fitness exercises target the major body muscle groups: arms, legs, back and abdomen. We split them into two groups: simple (involving basic repetitions such as *push-up*, *squat* or *dumbbell biceps raise*) and compound (assuming more complex routines involving multiple body regions, such as *burpees*, entailing a push-up and a jump, or *clean and press*, assuming certain trajectories of the arms). Each subject is asked to perform each type of exercise for a minimum of 5 repetitions.

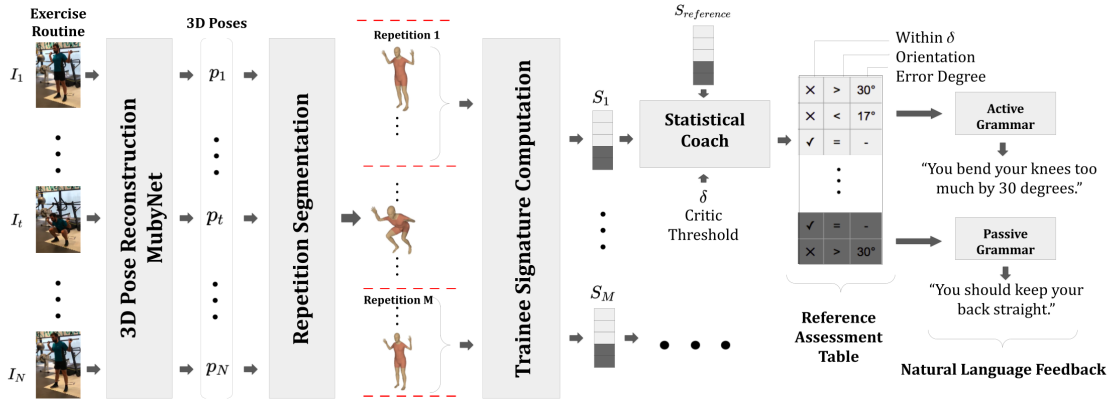

Figure 1. *AIFit* overview. Given a video of a trainee performing an exercise, (a) the system performs **3d pose reconstruction** in each frame and then (b) applies **repetition segmentation** to automatically count the number of 3d pose repetitions and determine each repetition interval. Next, **exercise modelling** (c) computes an *exercise signature* using the angular features of each repetition of the trainee (see fig. 3 for a detailed view). (d) The **statistical coach** compares each repetition signature against the instructor reference signature under a critic threshold that allows for different degree of error. The results of the comparison are populated into a **reference assessment table** specifying which deviations are greater than the critic threshold, the sign of the deviation and the degree of error. Finally, based on the table, e) *AIFit* produces **natural language feedback** for the trainee, using either an active or a passive grammar.

The subject height varies between 1.55-1.9m and the weight between 60-110kg, the dataset covering a mix of fit subjects (persons who exert a high degree of physical activity) as well as less trained ones. *Fit3D* consists of $2,964,236$ unique MOCAP 3d skeletons synchronized with RGB images. Each skeleton is also accompanied by the GHUM [35] human body model parameters, obtained by fitting the body model to the markers. Each subject is also 3d scanned (please see our Sup. Mat. for examples). We split the dataset into training and validation (10 subjects - $2,278,572$ images), and testing (3 subjects - $685,664$ images), with all exercise types available in both subsets. In addition, we manually segment each video into repetitions, annotating a total of $2,964$ timestamps. We define the subset of recordings for the trainees as *Trainees3D* and the subset of recordings for the instructor as *Reference3D*.

4. Methodology

Our proposed automatic system for fitness training (*AIFit*) takes as input a video of a person performing fitness exercises and outputs human-interpretable language feedback. The main components of the system execute the following tasks: a) 3d pose estimation to compute angular features; b) segmentation of the 3d pose sequence into individual repetitions; c) an exercise modelling extracts angular features and applies statistical operators in order to obtain an exercise signature for each of the repetitions; d) statistical coach identifies errors in exercise signatures with respect to a precomputed exercise signature of an instructor, and e) provides natural language feedback with visual grounding. An overview of our methodology is shown in fig. 1.

4.1. Segmentation of Repetitions

Given a sequence of N frames of a given routine, our goal is to extract the temporal intervals $T = \{T_i | T_i = [t_i, t_{i+1}], i = 1, 2, \ldots, k\}$ corresponding to all k repetitions of the given exercise. We propose using estimated 3d poses in each frame $P = \{p_1, p_2, \ldots, p_N\}$ as an intermediate representation of the motion. Our method should be robust to the quality of 3d poses, to the motion variation in each repetition and to the number of repetitions each subject chooses to execute (we assume a minimum of k_{min} repetitions, i.e., $k \geq k_{min}$). We introduce a two-stage algorithm, where we first assume the length of the repetitions within a video is fixed, and then use this estimate as an initialization for refinement using constrained continuous optimization.

Initialization. To obtain a first estimate of the segmentation, we assume a fixed-period of the pose signal, $T_{\text{init}} = T(t_{\text{start}}, \tau) = \{T_i | t_i = t_{\text{start}} + (i-1)\tau\}$, where τ is a period and t_{start} represents the starting point of the repetitions. We define the affinity between two 3d poses $A(p_m, p_n)$ as the negative mean per joint position error (MPJPE) between the p_m and p_n poses. To determine τ^*, the initial estimate of the period, we define the auto-correlation of the signal as:

$$R_{PP}(\tau, s) = \frac{1}{N - 2s - \tau} \sum_{t=s}^{N-s-\tau} A(p_t, p_{t+\tau}) \quad (1)$$

where s is the size by which the signal is shrunk at both ends, to account for noisy components not part of any repetition (in theory, one could use two s values, one for each end, but in practice we didn't notice significant differences). We iterate over s and then τ and select the first period τ^* of the signal to be the smallest τ for which $R_{PP}(\tau, s)$ reaches

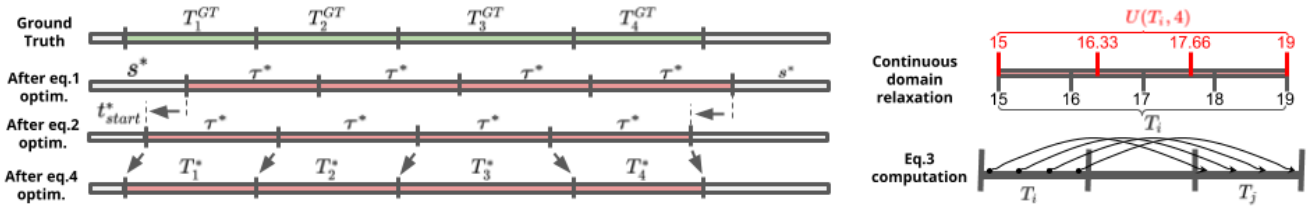

Figure 2. (***Left***) Example of the effect of each equation in the repetition segmentation algorithm. (***Top Right***) Example of sampling a fixed number of 4 timestamps from T_i using U. (***Bottom Right***) Affinity computation between two intervals T_i and T_j, as described in eq. 3.

a local maximum point. The corresponding s for this local maximum is s^*. In short, τ^* is the period that maximizes the auto-correlation of the signal, where noise outside repetitions is taken into account.

Once the period τ^* is estimated, we search for the beginning of the first repetition t_{start} maximizing the average affinity A_{avg} of $T(t_{\text{start}}, \tau^*)$, which we define as:

$$A_{\text{avg}}(T) = \frac{1}{k_{\text{min}}^2} \sum_{i=1}^{k_{\text{min}}} \sum_{j=1}^{k_{\text{min}}} A_{\text{seq}}(T_i, T_j) \tag{2}$$

where:

$$A_{\text{seq}}(T_i, T_j) = \frac{1}{\tau^*} \sum_{l=1}^{\tau^*} A(p_{t_i+l}, p_{t_j+l}) \tag{3}$$

Eq. 3 computes the similarity between two repetitions of equal period τ^* (intervals T_i and T_j), as shown in fig. 2 bottom right. Eq. 2 averages similarities between all possible pairs of intervals, being a global affinity of the repetition segmentation T, parameterized at this stage only by t_{start}, since τ^* is already found in eq. 1. We select t^*_{start} as the smallest value for which $A_{\text{avg}}(T(t_{\text{start}}, \tau^*))$ has a local maximum, as it provides the highest similarity between repetitions. We select the smallest such maximum to prevent solutions such as the beginning of the 2nd/3rd/etc. interval, which are also local maxima.

Optimization. Next, we drop the fixed period assumption and use $T(t^*_{\text{start}}, \tau^*) = \{T^*_i | t^*_i = t^*_{\text{start}} + (i-1)\tau^*\}$ as initialization for a nonlinear constraint optimization in the T domain.

In order to compare any two intervals (of possible different lengths), we need to uniformly sample the same number of frames from each interval. The function $U(T_i, n_S)$ does this, by simply uniformly collecting n_S continuous frame coordinates between the start and end frame of an interval T_i. For an e.g., please see fig. 2 top right, where, when sampling 4 frames from the interval $[15, 19]$ consisting of 5 frames (in black), we obtain: $U([15, 19], 4) = \{15, 16.33, 17.66, 19\}$ (in red). The pose at a non-discrete timestamp is obtained via linear interpolation $\hat{p}_x = p_{\lfloor x \rfloor} \cdot (1 - (x - \lfloor x \rfloor)) + p_{\lceil x \rceil} \cdot (x - \lfloor x \rfloor)$. This allows generalizing eq. 3 to the case where T is parametrized by $t_i, i = 1, 2, \ldots, k_{min}$:

$$\hat{A}_{\text{seq}}(T_i, T_j) = \frac{1}{n_S} \sum_{\substack{u \in U(T_i, n_S) \\ v \in U(T_j, n_S)}} A(\hat{p}_u, \hat{p}_v) \tag{4}$$

and derive $\hat{A}_{\text{avg}}(T)$ from eq. 2 by replacing $A_{\text{seq}}(T_i, T_j)$ with $\hat{A}_{\text{seq}}(T_i, T_j)$.

Our objective becomes maximizing $\hat{A}_{\text{avg}}(T)$ over t_i, $i = 1, 2, \ldots, k_{\text{min}}$, with the following constraints:

$$t_{i+1} - t_i > \delta \text{ for } i = 1, 2, \ldots, k_{\text{min}} \text{ and a small } \delta \tag{5}$$

to ensure we do not obtain a trivial solution due to the similarity of 3d poses in consecutive frames and

$$|t_{k_{\text{min}}+1} - t^*_{k_{\text{min}}+1}| < \tau^* \tag{6}$$

to ensure the solution we find does not overlap with possible subsequent repetitions (beyond k_{min}).

Note that due to the general MPJPE distance function we use to define pose affinity, our method generalizes to any type of physical exercise and no hand-crafted features need to be designed for particular motions. The choice of hyper-parameters n_S and δ is validated on the training set.

4.2. Exercise Modelling

The input to our *AIFit* system is an untrimmed sequence of 3d poses of a trainee performing an exercise. We apply the temporal repetition detector in §4.1 to segment the sequence into individual units. The aim is to provide visual and textual feedback on the correctness of the exercise that is easy to interpret by the trainee, so that errors can be easily understood and corrected. We decide to use only angular features as they are robust to a person's scale, build (i.e. different bone lengths) and global orientation. A general set of angular feature functions is defined around the major articulations of the human body. In fig. 4 we show examples for a few angular feature signals computed on two exercises performed by both a trainee and the instructor. On the top, we illustrate the spine angle during a squat exercise. Ideally, as can be seen for the instructor, it should be kept straight (i.e. around π), with minimal jitter during the

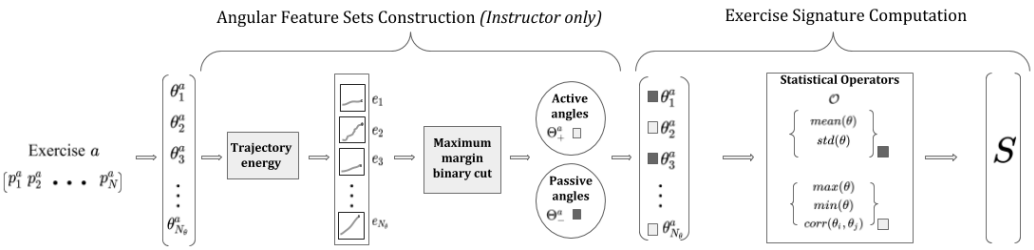

Figure 3. *Exercise Modelling:* **(Left) Active and passive angular feature sets construction** (instructor only). For an *exercise* a and for each angular feature function, we integrate its motion trajectory over the instructor's sequence of 3d poses, and get the motion energy of each feature function. We cluster the energies into two sets, active Θ^a_+ (associated with high energy) and passive Θ^a_- (associated with low energy) by using a maximum margin binary cut. **(Right) Exercise signature computation**. Both for trainees and instructor exercises, a signature is produced from the computed angular features, corresponding cluster assignments (derived from instructor exercises) and predefined statistical operators (applied to each of the two sets of angular features).

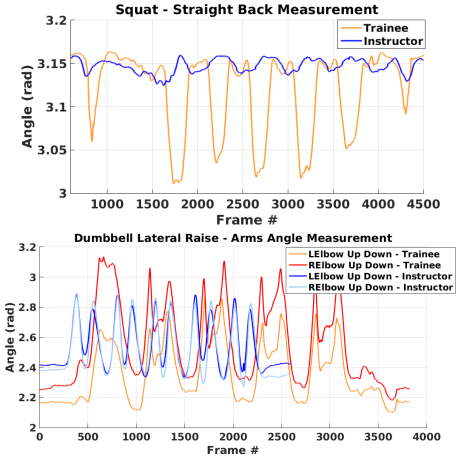

Figure 4. Measurement differences between a trainee and the instructor. (***Top***) For a squat exercise we measure the angle formed by the pelvis, mid-spine and neck. Ideally, it should be π (180°, straight line). The instructor performance comes close to that value. (***Bottom***) For a lateral dumbbell raise exercise we measure the angle between the upper arms and the spine. Ideally, the angle phase between left and right should be the same (see the instructor performance) and the movement frequency should be constant throughout the exercise (the trainee movement is more uneven).

exercise. However, it is clearly noticeable that the trainee does not manage to keep their back straight. Furthermore, the spine bends synchronously with the repetitions of the squat. On the bottom side of fig. 4, we analyze a dumbbell lateral rise exercise. We measure the angle at the left and right shoulders between elbow articulations and the spine, for both the instructor and trainee. It can be observed that the instructor has a nearly perfect phase angle with almost constant frequency for both his arms, whereas the trainee has chaotic arm movement. Such measurements offer us insight for the design of the proposed methodology.

We make the observation that, for a given fitness exercise, a low number of features define the motion, whereas the others are kept constant. We propose to automatically detect the two different categories of features, which we call *active* and *passive*, given instructor's demonstrations. As the two categories also entail different types of signal statistics, we therefore create separate distinct policies with corresponding signal operators (e.g. 'mean' angle) that aggregate over the temporal domain.

Angular Features. We use angular features to capture the motion statistics of the most important kinematic limb joints – knees, elbows, shoulders – and the spine. For the kinematic limb joints we compute several types of angular features: articulation angle (between the two limbs connecting to the child and parent joints in the kinematic tree), angle between the parent limb and the 'Up'/'Right'/'Forward' axes. The 'Up' axis is always considered to be the y axis, the 'Right' axis is given by the orientation of the shoulder joints, and the 'Forward' axis is constructed as the cross-product of the 'Up' and 'Right' axes. By defining the coordinate axes this way, we ensure that features are invariant to the global rotation of the human subject. For the spine, we only use the articulation angle. This feature is used to determine how straight the back is, an important aspect in many fitness routines. We denote the angular feature set $\Theta = \{\theta^i\}_{i=1...N_\theta}$, with N_θ the total number of angular features.

Construction of Active and Passive Feature Sets. For a given exercise $a \in \mathcal{A}$, where $\mathcal{A}$ is the set of all exercises, we are interested in partitioning the set of feature functions Θ in two subsets: the *active* set Θ^a_+ and the *passive* set Θ^a_-. Intuitively, the former includes the features that define the motion of the exercise (i.e. carry most of the *energy*; e.g. knees bending in a squat) and the latter includes all the other features (i.e. carry the least *energy*; e.g. back kept straight).

We use the 3d motion of the instructor $P^a_I = (p_1, \ldots, p_N)$ as reference for an exercise a. We compute the response for each different feature function over the se-

quence, $\theta^i(P_I^a) = (\theta^i(p_1), \ldots, \theta^i(p_N))$, with $\theta^i \in \Theta$. We define the energy of feature function θ^i as:

$$e^i = \sum_{j=1}^{N-1} |\theta^i(p_{j+1}) - \theta^i(p_j)| \quad (7)$$

We sort the energies $\{e\}_{i=1\ldots N_f}$ and find the cut with the largest margin separating them in two clusters – of high and low energy, respectively – and automatically gather the corresponding final feature function sets Θ_+^a and Θ_-^a. An overview of this process is shown in fig. 3 (left).

Exercise Signature Computation. In order to compare between trainee and instructor exercise routines, we propose to model them by an *exercise signature* (see fig. 3 (right)). First, for each type of feature function set (*active* Θ_+^a or *passive* Θ_-^a), we consider different statistical operators $\mathcal{O}$. We compute all the feature function responses on all the detected repetitions in a sequence and temporally aggregate responses in each repetition window using different operators, as follows: for the *active* feature functions we use $\mathcal{O}_+^u(\theta^i) = \{'max', 'min'\}$ as unary aggregation operators and $\mathcal{O}_+^p(\theta^i, \theta^j) = \{'correlation'\}$ as the pairwise aggregation operator, where $\theta^i, \theta^j \in \Theta_+^a$; for the *passive* feature functions we consider just the unary aggregation operators $\mathcal{O}_-^u(\theta^i) = \{'mean', 'std'\}$, where $\theta^i \in \Theta_-^a$. The unary operators are chosen to reflect the periodic nature of the *active* features and the stationary nature of the *passive* features. The pairwise operator for the *active* features reflects the correlation between angles in a synchronized exercise (e.g. both knees follow the same movement in a squat). The output of these operators on the angular features is the *exercise signature* and we denote it by S^a.

4.3. Statistical Coach

Given a repetition of an exercise routine by a trainee, we compute its signature. This trainee signature, S, is compared against a precomputed reference signature, $S_{\text{reference}}$, of the instructor from the *Reference3D* subset of our dataset. For each pair of entries in the signatures, if the absolute difference is larger than a threshold we consider it as an erroneous execution and also retain the sign of the difference (i.e. lower, higher or equal). The thresholds are set such that they account for the worst performing trainee seen in the *Trainees3D* subset of our dataset. All thresholds used are further scaled by a global parameter δ, with values between 0 and 1, that models the system's sensitivity to errors and can be interpreted as feedback of the statistical coach (the '*critic*') on the accuracy of the exercise execution. The statistical coach populates a *reference assessment table* that is further used to generate textual feedback.

4.4. Natural Language Feedback

AIFit provides human-interpretable visual and textual feedback for each repetition of an exercise routine. For the textual output, we use two different grammars, one providing feedback for the active angles and one for the passive angles (see production rules in listing 1). The non-terminals are dependent on the type of error, feature and aggregation operators. For each identified error in the trainee routine we also provide a visual output: an image where the error occurs and the corresponding reference image of the instructor showing the correct execution.

Listing 1. Production rules for the active (**top**) and passive (**bottom**) grammars.

```
You <verb> your <noun> too <adv1> [to the <adv2>] by
<numeral> degrees.
<verb> := lower|raise|extend|bend|move
<noun> := {joints}
<adv1> := much|little
<adv2> := front|back|right|left

You should keep your <noun> <adj2> | (<adv1> (<adj1>|<adv2>)).
<noun> := {joints}
<adj2> := higher|lower|still|straight
<adv1> := more|less
<adj1> := extended|bent
<adv2> := to the right|to the left
```

5. Experiments

We learn and validate all of our components on the training set of *Fit3D* and report results on the test set.

3D Pose Reconstruction. For experiments, we use either the ground-truth 3d pose reconstruction or the predicted 3d pose reconstructions. For the latter, we adapt a state-of-the-art 3d pose reconstruction network MubyNet[38], which was pre-trained on the Human3.6M[10], a large-scale 3d dataset capturing everyday activities. We define its variants as follows: MubyNet-SV – single view reconstruction, MubyNet-MV – multi-view view reconstruction, while MubyNet-{*}-FT denote the fine-tuned versions of the network trained on Fit3D for 5 epochs. For the multi-view reconstruction, we first run MubyNet in all available cameras. Next, we transform all the 3d pose reconstructions from camera space to world space (assuming known camera parameters) and apply a median operation to obtain a single 3d pose reconstruction estimate. In table 1 we show the reconstruction errors for the different variants considered on the test split of *Fit3D*. Both fine-tuned variants achieve significantly lower reconstruction errors compared to the original ones, and the multi-view approaches are also consistently better than single view ones. We also test against SPIN[12] another state-of-the-art 3d pose and shape reconstruction method. The errors are on a par with MubyNet but higher than our fine-tuned methods. This shows that our proposed *Fit3D* dataset covers novel 3d poses, that are outside the distribution of current 3d pose datasets used in the literature, such as Human3.6M.

Segmentation of Repetitions. We validate and test our

Method	w/o Procrustes SV	w/o Procrustes MV	w. Procrustes SV	w. Procrustes MV
SPIN	89.5	67.6	61.0	50.7
MubyNet	90.4	71.9	67.7	57.9
MubyNet-FT	52.4	45.4	41.1	35.7

Table 1. MPJPE errors (in mm) of different reconstruction methods, with/without Procrustes Alignment. Multi-view (MV) versions consistently outperform the single-view (SV), while fine-tuned ones (FT) outperform their regular counterparts.

Input	GT joints	MubyNet SV	MubyNet MV	MubyNet SV-FT	MubyNet MV-FT
Acc.	0.731	0.661	0.690	0.708	0.730

Table 2. Accuracy (IoU) performance for temporal segmentation of an exercise into repetitions, shown by using both ground-truth and estimated 3d input joints. Segmentation accuracy increases with the quality of the input poses. It saturates for the MubyNet-MV-FT predictions, where the performance is similar to that obtained using GT poses.

repetition segmentation method on sequences from *Fit3D*, using the annotated timestamps. We validate the hyper-parameters of the method on the training set using the ground truth 3d joints as input and report the performance of the method on the 3 subjects in the test set. Here we set $k_{\text{min}} = 5$ as the minimum number of repetitions executed (and annotated) in a video. Since we are interested in repetition segmentation, we use the intersection-over-union (IoU) to measure the accuracy of segmentation, averaged over all k_{min} repetitions of the test sequences.

Table 2 shows the performance of our repetition segmentation method using as input 3d pose sequences obtained from different reconstruction methods or the ground truth. As expected, the accuracy of our repetition segmentation method is positively correlated with that of the 3d pose estimation method MubyNet used to generate the input of the method (the smaller the reconstruction error, the better the segmentation accuracy of repetitions). Yet, when the reconstruction error is not so high (as in MubyNet-MV-FT), the repetition segmentation method performs on a par with using ground truth 3d poses (0.730 vs. 0.731 IoU), which proves the system's robustness to variation in input quality.

Repetition Counting. Although designed for fine-grained repetition segmentation, we also evaluate our repetition algorithm on the task of repetition counting (only estimating the period with which an action is repeated in a video, not the extent of each repetition unit). Our approach requires that the repetition is performed by a human, so we restrict our evaluation only to these types of videos. We report results on our *Fit3D* dataset and a subset of the Countix dataset [6] which we call CountixFitness. This subset is selected to consist of only videos of humans performing fitness exercises (e.g. 'front raises', 'lunge', 'jumping jacks', 'pull ups', 'push up', 'rope pushdown'). CountixFitness is a subset of only the train and validation set of Countix, since the test set does not contain action tags which we require to filter out actions not involving humans. The protocol in the literature assumes input videos are fully-periodic, so we trim both datasets to be periodic from start to end.

We set $k_{\text{min}} = 2$ and use the initial period τ^* obtained with the MubyNet SV-FT 3d pose representation as the AIFit estimated period. The number of counts is obtained by dividing the length of the video to the period and rounding it up. We also evaluate RepNet [6] on both datasets and report the two existing evaluation metrics used in the literature: the Off-By-One (OBO) error (the misclassification rate, where a video is classified correctly if the predicted count is within one count of the ground-truth) and the Mean Absolute Error (MAE) of count (where the absolute error is the absolute count difference between the ground truth and prediction, normalized by ground truth count). Results are computed on the 3 subjects in the Fit3D test set (444 videos) and on the CountixFitness validation set (267 videos).

Dataset	Fit3D OBO ↓	Fit3D MAE ↓	CountixFitness OBO ↓	CountixFitness MAE ↓
RepNet	0.520	0.740	**0.292**	**0.468**
AIFit	**0.140**	**0.253**	**0.292**	0.604

Table 3. Comparison of RepNet and our AIFit on the *Fit3D* and CountixFitness datasets.

Table 3 compares our method with RepNet on the two datasets. On Fit3D, AIFit significantly outperforms RepNet, while on CountixFitness the two algorithms perform similarly, with AIFit drifting a little more than RepNet when being wrong. Note that while RepNet is trained directly for repetition counting on the entire in-the-wild Countix dataset, our AIFit approach generalizes well on all datasets, but specializes on human actions. This is because our approach does not require any training and can easily be integrated on top of any existing 3d pose reconstruction method. We also experiment with using an alternative representation of the person in each frame. Instead of the 3d pose, we use the estimated 2d pose predicted using [4]. To define the dissimilarity measure between two 2d poses, we compute the Euclidean distances between corresponding keypoints, weight them by the product of their confidences and average them. Our results confirm that 3d pose is a stronger representation than the 2d pose, both in terms of the OBO error (0.140 vs. 0.920) and MAE (2.250 vs. 0.253).

Active and Passive Features. We conduct an ablation study to measure the similarity (IoU score) between the performances of the instructor and each of the trainees, in terms of mined *active* feature sets. We illustrate this study in fig. 5.

AIFit Feedback. In fig. 7 we provide quantitative analysis of the the AiFit feedback ablating over the different variants of 3d pose estimation methods (MubyNet) on the

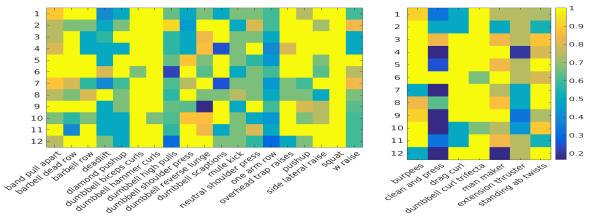

Figure 5. IoU score for active features between each of the twelve trainees and the instructor, for *simple* fitness exercises (***left***) and *compound* ones (***right***). I t can be clearly noticed that for compound exercise types, there are less common active features among the trainee and instructor. Also, for the simplest of exercises (i.e. squat, push-up, dumbbell biceps curls) we get the highest number of common active features between trainee and instructor.

Trainees3D subset. We consider the feedback of AIFit computed over the ground-truth 3d poses as the reference and compare against the feedback of AIFit using all other 3d pose reconstruction methods. We pose it as a classification problem, as feedback can be seen as a multi-class labeling (e.g. higher, lower, same) for different aggregation operators and feature types. We vary the global parameter controlling the system's sensitivity to errors, δ, from more restrictive to more permissive. A higher accuracy at higher critic thresholds is expected, as in this case the inaccuracies of the pose estimation methods have less impact on the feedback outcome. The more accurate 3d pose estimation methods also have lower reconstruction errors (table 1). At a critic threshold of $\delta = 0.5$, the feedback obtained with 3d pose estimation methods achieves around 80% accuracy for both active and passive policies. We also show examples of textual and visual feedback for different trainees exercising outdoors (fig. 6).

6. Conclusions

We have introduced *AIFit*, the first 3d human sensing-based automatic system for fitness training, together with *Fit3D*, a large-scale dataset of over 3 million images and corresponding 3d human shape and motion capture ground truth configurations, featuring 37 repeated exercises that cover all the major muscle groups, performed by certified trainers and trainees with different skill levels. *AIFit* is able to reconstruct 3d human pose and motion, reliably segment repetitions, mine critical active and passive features of the exercise, and identify deviations between correct execution models learnt from trainers, and the work of the trainee, in real-time. A statistical coach provides localized, quantitative feedback, in order to exercise correctly, reduce the risk of injury, and sustain continuous improvement. The statistical coach can operate over both a relaxed and a high-intensity training intensity regime, provides feedback in natural language, and with spatio-temporal visual grounding in trainee's execution, making it useful as an ubiquitous complement to less frequently available human trainers at the gym, outdoors, or at home. Models will be made available for research.[1]

Figure 6. Textual and visual feedback produced by our *AIFit* on real world videos, captured with a regular smartphone camera. We use MubyNet-FT to estimate the 3d pose of the trainee. For each example, we show the following: an image with the identified error of the trainee (***top row***), the 3d reconstruction of the trainee (***second row***), the corresponding image with the correct execution of the instructor (***third row***) and the textual feedback (***bottom row***). The two examples on the (***left***) show active features feedback, while the two on the (***right***) show passive features feedback. Notice generalization to various humans in different environments and camera viewpoints. *Please see Sup. Mat. for videos!*

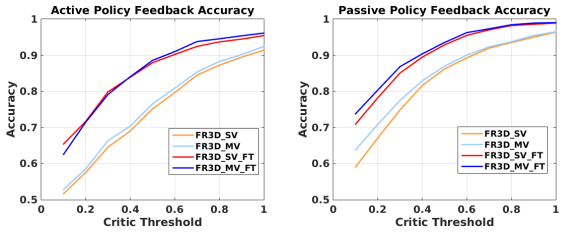

Figure 7. Accuracy at different critic thresholds when comparing the feedback produced by *AIFit* based on ground truth 3d poses against 3d poses estimated with different variants of MubyNet. We show the accuracy computed on active (***left***) and passive (***right***) feedback policies. For higher critic thresholds (more permissive system), the accuracy increases, as fewer errors are being reported.

Acknowledgments: This work was supported in part by the ERC Consolidator grant SEED, CNCS-UEFISCDI (PN-III-P4-ID-PCCF-2016-0180) and SSF.

[1] http://vision.imar.ro/fit3d

References

[1] Elizabeth H Anderson and Geetha Shivakumar. Effects of exercise and physical activity on anxiety. *Frontiers in psychiatry*, 4:27, 2013.

[2] João Antunes, Alexandre Bernardino, Asim Smailagic, and Daniel P. Siewiorek. Aha-3d: A labelled dataset for senior fitness exercise recognition and segmentation from 3d skeletal data. In *BMVC*, 2018.

[3] Abdallah Benzine, Bertrand Luvison, Quoc Cuong Pham, and Catherine Achard. Deep, robust and single shot 3d multi-person human pose estimation from monocular images. In *ICIP*, 2019.

[4] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017.

[5] Ross Cutler and Larry S. Davis. Robust real-time periodic motion detection, analysis, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):781–796, 2000.

[6] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. Counting out time: Class agnostic video repetition counting in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10387–10396, 2020.

[7] Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Three-dimensional reconstruction of human interactions. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[8] Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Learning complex 3d human self-contact. In *Thirty-Fifth AAAI Conf. on Artificial Intelligence (AAAI'21)*, 2021.

[9] Yuying Ge, Ruimao Zhang, Xiaogang Wang, Xiaoou Tang, and Ping Luo. Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. In *CVPR*, June 2019.

[10] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3d human sensing in natural environments. *PAMI*, 2014.

[11] Xin Jin, Yuan Yao, Qiliang Jiang, Xingying Huang, Jianyi Zhang, Xiaokun Zhang, and Kejun Zhang. Virtual personal trainer via the kinect sensor. In *ICCT*, pages 460–463. IEEE, 2015.

[12] Nikos Kolotouros, Georgios Pavlakos, Michael J. Black, and Kostas Daniilidis. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In *ICCV*, 2019.

[13] Ofir Levy and Lior Wolf. Live repetition counting. In *Proceedings of the IEEE international conference on computer vision*, pages 3020–3028, 2015.

[14] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In *CVPR*, 2019.

[15] Xin Luan, Xiangyang Tian, Haixin Zhang, Rui Huang, Na Li, Peijie Chen, and Ru Wang. Exercise as a prescription for patients with various diseases. *Journal of sport and health science*, 8(5):422–441, 2019.

[16] Diogo C Luvizon, David Picard, and Hedi Tabia. 2d/3d pose estimation and action recognition using multitask deep learning. In *CVPR*, pages 5137–5146, 2018.

[17] Julieta Martinez, Rayat Hossain, Javier Romero, and James J. Little. A simple yet effective baseline for 3d human pose estimation. In *ICCV*, 2017.

[18] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3D pose estimation from monocular RGB. In *3DV*, 2018.

[19] Neville Owen, Phillip B Sparling, Geneviève N Healy, David W Dunstan, and Charles E Matthews. Sedentary behavior: emerging evidence for a new health risk. In *Mayo Clinic Proceedings*, volume 85, pages 1138–1141. Elsevier, 2010.

[20] Costas Panagiotakis, Giorgos Karvounas, and Antonis Argyros. Unsupervised detection of periodic segments in videos. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 923–927. IEEE, 2018.

[21] Yanwei Pang, Jin Xie, Muhammad Haris Khan, Rao Muhammad Anwer, Fahad Shahbaz Khan, and Ling Shao. Mask-guided attention network for occluded pedestrian detection. In *ICCV*, pages 4967–4975, 2019.

[22] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *CVPR*, 2019.

[23] Katrina L Piercy and Richard P Troiano. Physical activity guidelines for americans from the us department of health and human services: Cardiovascular benefits and recommendations. *Circulation: Cardiovascular Quality and Outcomes*, 11(11):e005263, 2018.

[24] A.-I. Popa, M. Zanfir, and C. Sminchisescu. Deep multitask architecture for integrated 2d and 3d human sensing. In *CVPR*, 2017.

[25] Amir Rasouli, Iuliia Kotseruba, Toni Kunic, and John K Tsotsos. Pie: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction. In *ICCV*, 2019.

[26] Tom F H Runia, Cees G M Snoek, and Arnold W M Smeulders. Real-world repetition estimation by div, grad and curl. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[27] Matthias Seuter, Lucien Opitz, Gernot Bauer, and David Hochmann. Live-feedback from the imus: Animated 3d visualization for everyday-exercising. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*, UbiComp '16, page 904–907, New York, NY, USA, 2016. Association for Computing Machinery.

[28] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *CVPR*, pages 1010–1019, 2016.

[29] Jamie Shotton, Toby Sharp, Alex Kipman, Andrew Fitzgibbon, Mark Finocchio, Andrew Blake, Mat Cook, and Richard Moore. Real-time human pose recognition in parts from single depth images. *Communications of the ACM*, 56(1):116–124, 2013.

[30] Alexey Sidnev, Alexey Trushkov, Maxim Kazakov, Ivan Korolev, and Vladislav Sorokin. Deepmark: One-shot clothing detection. In *ICCV Workshops*, Oct 2019.

[31] Kai Su, Dongdong Yu, Zhenqi Xu, Xin Geng, and Changhu Wang. Multi-person pose estimation with enhanced channel-wise and spatial information. In *CVPR*, 2019.

[32] Deborah F Tate, Elizabeth J Lyons, and Carmina G Valle. High-tech tools for exercise motivation: use and role of technologies such as the internet, mobile applications, social media, and video games. *Diabetes Spectrum*, 28(1):45–54, 2015.

[33] Jianbo Wang, Kai Qiu, Houwen Peng, Jianlong Fu, and Jianke Zhu. Ai coach: Deep human pose estimation and analysis for personalized athletic training assistance. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, page 374–382, New York, NY, USA, 2019. Association for Computing Machinery.

[34] Haoran Xie, Atsushi Watatani, and Kazunori Miyata. Visual feedback for core training with 3d human shape and pose. *2019 Nicograph International (NicoInt)*, pages 49–56, 2019.

[35] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, Bill Freeman, Rahul Sukthankar, and Cristian Sminchisescu. GHUM & GHUML: Generative 3D human shape and articulated pose models. *CVPR*, 2020.

[36] Wei Yang, Wanli Ouyang, Xiaolong Wang, Jimmy Ren, Hongsheng Li, and Xiaogang Wang. 3d human pose estimation in the wild by adversarial learning. In *CVPR*, 2018.

[37] Andrei Zanfir, Elisabeta Marinoiu, and Cristian Sminchisescu. Monocular 3d pose and shape estimation of multiple people in natural scenes - the importance of multiple scene constraints. In *CVPR*, 2018.

[38] Andrei Zanfir, Elisabeta Marinoiu, Mihai Zanfir, Alin-Ionut Popa, and Cristian Sminchisescu. Deep network for the integrated 3d sensing of multiple people in natural images. In *NIPS*, 2018.

[39] B. Zhou, M. Sundholm, J. Cheng, H. Cruz, and P. Lukowicz. Never skip leg day: A novel wearable approach to monitoring gym leg exercises. In *IEEE International Conference on Pervasive Computing and Communications (PerCom)*, 2016.