Training Pose

Dr. Gil Ben-Artzi.
Ester Reznikov, Yaara Goldenberg.
Department of Computer Science.
Ariel University Center of Samaria.

Abstract

Fitness exercises are very beneficial to personal health and fitness; however, they can also be ineffective and potentially dangerous if performed incorrectly by the user. Exercise mistakes are made when the user does not use the proper form or pose. The system is planned to detect the exercise pose of trainer and provide a feedback on the user's form, without using any sensor except a simple camera. In this paper, we propose a system that provides visual feedback on the way that the user preforms training exercise using HMR and SMPL models. Our goal is to help prevent injuries and improve the quality of people's workouts without a trainer.

1 INTRODUCTION

Exercises are beneficial to health and fitness, however, it is difficult for an individual to achieve the correct pose in training, it can also be very dangerous if performed incorrectly. Many people work out and perform exercises regularly but do not maintain the proper form (pose). This could be due to a lack of formal training through classes or a personal trainer. The main goal in this paper is to aid people in performing the correct posture for exercises by building a model that detects the user's exercise pose and provides useful feedback on the user's form, using a combination of the latest advances in pose estimation, and by that help to prevent injuries and improve the quality of people's workouts with just a computer and a webcam. The first step of Training Pose uses human pose estimation, a difficult but highly applicable domain of computer vision. For pose estimation we use the Skinned Multi-Person Linear model (SMPL) which is a skinned vertex based model that accurately represents a wide variety of body shapes in natural human poses. The second part is to match two different videos, the user form and the correct form. The user may perform the exercise in different pace from how it performs in the target video. In this paper we emphasize on the human pose and not the pace of the exercise, so we use Dynamic Time Warping (DTW) to match the stages of the exercises between the source and the target videos. Third part is presenting a useful visual feedback for the user by applying arrows on the HMR model to compere the main joints for the current exercises between source and target.

2 RELATED WORK

2.1 OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields

OpenPose[1] can provide real-time 2D pose estimation using a nonparametric representation to learn the body parts in an image dataset. This is the first open source library available real time system for multi-person 2D pose detection, including body, foot, hand, and facial key points. The method takes the entire image as the input for a CNN to jointly predict confidence map for body part detection and PAFs for part association. The parsing step performs a set of bipartite matchings to associate body part candidates. Finally assemble them into full body poses for all people in the image. They present the first bottom up representation of association scores via Part Affinity Fields (PAFs), a set of 2D vector fields that encode the location and orientation of limbs over the image domain. Real-time multi-person 2D pose estimation is a critical component in enabling machines to visually understand and interpret humans and their interactions. In their paper, they present an explicit nonparametric representation of the key point association that encodes both position and orientation of human limbs. Second, they design an architecture that jointly learns part detection and association. Third, they demonstrate that a greedy parsing algorithm is sufficient to produce high-quality parses of body poses, and preserves efficiency regardless of the number of people. Fourth, they prove that PAF refinement is far more

important than combined PAF and body part location refinement, leading to a substantial increase in both runtime performance and accuracy. They have open-sourced this work as OpenPose, the first real-time system for body, foot, hand, and facial key point detection.

2.2 SMPL: A Skinned Multi-Person Linear Model

The authors present a learned model of human body shape and pose dependent shape variation that is more accurate than previous models and is compatible with existing graphics pipelines. The Skinned Multi-Person Linear model (SMPL) [2] is a skinned vertex based model that accurately represents a wide variety of body shapes in natural human poses. The parameters of the model are learned from data including the rest pose template, blend weights, pose-dependent blend shapes, identity-dependent blend shapes, and a regressor from vertices to joint locations. The goal is to automatically learn a model of the body that is both realistic and compatible with existing graphics software. To that end, they describe a "Skinned Multi-Person Linear" (SMPL) model of the human body that can realistically represent a wide range of human body shapes, can be posed with natural pose-dependent deformations, exhibits soft-tissue dynamics, is efficient to animate, and is compatible with existing rendering engines. The SMPL model decomposes body shape into identity-dependent shape and non-rigid pose-dependent shape, and take a vertex-based skinning approach that uses corrective blend shapes. A single blend shape is represented as a vector of concatenated vertex offsets. They begin with an artist created mesh with 6890 vertices and 23 joints. The mesh has the same topology for men and women, spatially varying resolution, a clean quad structure, a segmentation into parts, initial blend weights, and a skeletal rig. Because their model decomposes shape and pose, they train these separately, simplifying optimization, and there are separate models for men and women. The model is working because the good quality training data is important. Here they use thousands of high-quality registered template meshes. Importantly, the pose training data spans a range of body shapes enabling us to learn a good predictor of joint locations. Additionally, training all the parameters (template shape, blend weights, joint regressor, shape/pose/dynamic blend shapes) to minimize vertex reconstruction error is important to obtain a good model. Here the simplicity of the model is an advantage as it enables training everything with large amounts of data. SMPL uses standard skinning equations and defines body shape and pose blend shapes that modify the base mesh. They train the model on thousands of aligned scans of different people in different poses. The form of the model makes it possible to learn the parameters from large amounts of data while directly minimizing vertex reconstruction error. Specifically, they learn the rest template, joint regressor, body shape model, pose blend shapes, and dynamic blend shapes.

2.3 End-to-end Recovery of Human Shape and Pose [3]

The authors present an end-to-end framework for recovering a full 3D mesh of a human body from a single RGB image. They use the generative human body model, SMPL, which parameterizes the mesh by 3D joint angles and a low dimensional linear shape space. Estimating a 3D mesh opens the door to a wide range of applications such as foreground and part segmentation, which is beyond what is practical with a simple skeleton. The output mesh can be immediately used by animators, modified, measured, manipulated and retargeted. Their output is also holistic – they always infer the full 3D body even in cases of occlusion and truncation.

An overview of the proposed framework- an image is passed through a convolutional encoder, this is sent to an iterative 3D regression module that infers the latent 3D representation of the human that minimizes the joint reprojection error, the 3D parameters are also sent to the discriminator, whose goal is to tell if these parameters come from a real human shape and pose. Their approach is similar to 3D interpreter networks in the use of reprojection loss and the more recent adversarial inverse graphics networks for the use of the adversarial prior. They go beyond the existing techniques in multiple ways:

1. They infer 3D mesh parameters directly from image features, while previous approaches infer them from 2D key points. This avoids the need for two stage training and also avoids throwing away valuable information in the image such as context.
2. Going beyond skeletons, they output meshes, which are more complex and more appropriate for many applications. Again, no additional inference step is needed.
3. Their framework is trained in an end-to-end manner. They out-perform previous approaches that output 3D meshes in terms of 3D joint error and run time.
4. They show results with and without paired 2D-to-3D data. Even without using any paired 2D-to-3D supervision, their approach produces reasonable 3D reconstructions.

2.3.1 Iterative 3D Regression with Feedback

The goal of the 3D regression module is to output θ given an image. however, directly regressing θ in one go is a challenging task, particularly because θ includes rotation parameters. In this work, they take inspiration from previous works and regress θ in an iterative error feedback (IEF) loop, where progressive changes are made recurrently to the current estimate.

2.3.2 Factorized Adversarial Prior

The re-projection loss encourages the network to produce a 3D body that explains the 2D joint locations, however anthropometrically implausible 3D bodies or bodies with gross self-intersections may still minimize the re-projection loss. To regularize this, they use a discriminator network that is trained to tell whether SMPL parameters correspond to a real body or not. They refer to this as an adversarial prior as in since the discriminator acts as a data-driven prior that guides the 3D inference. The results of this paper without using any paired 3D data are promising since they suggest that we can keep on improving our model using more images with 2D labels, which are relatively easy to acquire, instead of ground truth 3D, which is considerably more challenging to acquire in a natural setting.

The main disadvantage of working with the above models [2]-[3] is the inaccuracy in the identification of the wrists and palms so the use on pose with an emphasis on the hand position is very limited, as shown in Figure 1. To overcome this limitation, there is a model called SMPL-X [4] which facilitate the analysis of human actions, interactions and emotions, they compute a 3D model of human body pose, hand pose, and facial expression from a single monocular image. To achieve this, they use thousands of 3D scans to train a new, unified, 3D model of the human body, SMPL-X, that extends SMPL with fully articulated hands and an expressive face. Learning to regress the parameters of SMPL-X directly from images is challenging without paired images and 3D ground truth. Consequently, they follow the approach of SMPLify, which estimates 2D features and then optimizes model parameters to fit the features. They improve on SMPLify in several significant ways: they detect 2D features corresponding to the face, hands, and feet and fit the full SMPL-X model to these; they train a new neural network pose prior using a large MoCap dataset; they define a new interpenetration penalty that is both fast and accurate; they automatically detect gender and the appropriate body models (male, female, or neutral); their PyTorch implementation achieves a speedup of more than $\times 8$ over Chumpy. They use the new method, SMPLify-X, to fit SMPL-X to both controlled images and images in the wild. They evaluate 3D accuracy on a new curated dataset comprising 100 images with pseudo ground-truth. In this work we didn't use the above model because of hardware limitations.

2.4 Visual Feedback

A visual feedback [5] system for core training using a monocular camera image. To support the user in maintaining the correct postures from target poses, by adopting 3D human shape estimation for both the target image and input camera video. Its aim to provide

Figure 1. An example of the HMR output for press exercise.

a user interface for visual feedback in sports training using state-of-the-art pose estimation approaches. In contrast to 3D sensing using laser sensors or depth cameras, the proposed system only requires a single image from a standard web camera. The framework consists of two main components: pose estimation and visual feedback. To obtain 3D human shape and pose from a single image, the proposed system adopts pose estimation based on OpenPose to realize the bounding box of the human region. It also employs human mesh recovery to estimate the 3D human pose. For the generation of 3D human mesh, an SMPL model is used. For visual feedback, a runtime user interface is proposed to guide the user on the differences between target pose and the current pose from the captured frame. Both target and current poses are estimated from a single image using the pose estimation method discussed above. The user can predefine the easy to see viewpoints on the pre-training user interface. The user can first, select the target image with a standard pose to be used for core training. Afterward, the system generates a 3D human model in two view windows using the proposed pose estimation approach. For each view window, the user can select the good viewpoints in his/her judgment. Finally, the system saves the predefined set of viewpoints. At the user interface the 3D human model that denotes the target 3D pose is represented by blue color while the 3D human model that represents the current pose from the camera image is represented by red color. The current pose model and the target pose model are displayed in a superimposed manner. Different colors of markers are used to indicate the differences between current and target models using for training guidance. They used red, orange, yellow and green-yellow colors in their implementation to represent distances ranging from far to close. For representing the differences between target and current 3D models, they adopt color visualization on predefined 10 marker positions on the human body. They set the marker positions at both hands and elbows, shoulders, knees, and ankles because the differences of the two models are more apparent at the end parts than the human trunk. They did not choose hip joints on

the body because the differences in hip points are not apparent due to coincided waist positions. They used shoulders, elbows, and hands positions for a good representation of the upper body postures, which are the most important parts in core training. To obtain the marker positions using SMPL-based human model, they superimposed the aforementioned joint positions with SMPL mesh model to determine the start and end points for each marker. To obtain the mesh vertices of markers, they substituted the transformation matrix of perspective projection from the 3D model to the output image. Finally, they calculate the difference, in vertex positions with four colors. To verify the efficiency and feasibility of the proposed Visual feedback system, the researchers compared the proposed system with the usual visual feedback on skeletal information. They asked users to use both systems and answer a questionnaire, the results was that the proposed guidance using 3D human model is easier for understanding target posture and that it's more effective in correcting postures when performing core training and requires less time to achieve the correct posture from the usual visual feedback on skeletal information. The current execution time of the proposed visual feedback system is about 2 seconds for one frame. Although it may be suitable for core training because the user has to maintain the pose over a long period, it will be bottle-neck for other sports such as gymnastics and ball sports where movement is in high speed.

2.5 Pose Trainer

Pose trainer [5], a software application that detects the user's exercise pose and provides useful feedback on the user's form. First, the user records a video of themselves performing a selected exercise. The video is recorded from a particular perspective (facing the camera, side to the camera, etc.) that allows the exercise to be seen. For pose estimation, they use OpenPose to label RGB images. After that, there is key point Normalization and Perspective Detection. There are two approach to evaluate the exercise posture given normalized key points:

- Geometry Evaluation by compute body vectors from key points of interest and use personal training guidelines and their own recorded videos to design geometric heuristics, evaluating on the body vectors.
- Use more data-driven, machine learning approach and DTW.

This application provides feedback only for 4 different exercises:

1. Bicep Curls.
2. Front Raise.
3. Shoulder press.
4. Shoulder shrug.

The disadvantages of this method are that it works on a very limited number of exercises, you must train the model for each exercise separately, and for the geometric approach you need to do a lot of pre-analysis of your data.

2.6 Everybody Dance Now [6]

This paper presents a simple method for "do as I do" motion transfer: Given two videos – one of a target person whose appearance we wish to synthesize, and the other of a source subject whose motion we wish to impose onto our target person – we transfer motion between these subjects by learning a simple video-to-video translation. Although their method is quite simple, it produces surprisingly compelling results. To transfer motion between two video subjects in a frame-by-frame manner, they must learn a mapping between images of the two individuals. Their goal is, therefore, to discover an image-to-image translation between the source and target sets. However, they do not have corresponding pairs of images of the two subjects performing the same motions to supervise learning this translation. Even if both subjects perform the same routine, it is still unlikely to have an exact frame to frame pose correspondence due to body shape and motion style unique to each subject. They observe that key point based pose preserves motion signatures over time while abstracting away as much subject identity as possible and can serve as an intermediate representation between any two subjects. To accomplish this task, they divide their pipeline into three stages – pose detection, global pose normalization, and mapping from normalized pose stick figures to the target subject.

2.6.1 Encoding Body Poses

To encode the body, pose of a subject image, they use a pre-trained pose detector which accurately estimates 2D x, y joint coordinates. Then they create a colored pose stick figure by plotting the key points and drawing lines between connected joints as shown in the figure.

2.6.2 Global Pose Normalization

In different videos, subjects may have different limb proportions or stand closer or farther to the camera than one another. Therefore, when retargeting motion between two subjects, it may be necessary to transform the pose key points of the source person so that they appear in accordance with the target person's body shape and location.

2.6.3 Temporal Smoothing

To create video sequences, we modify the single image generation setup to enforce temporal coherence be-

tween adjacent frames. Instead of generating individual frames, they predict two consecutive frames where the first output $G(x_t - 1)$ is conditioned on its corresponding pose stick figure $x_t - 1$ and a zero image Z (a placeholder since there is no previously generated frame at time t - 2). The second output $G(x_t)$ is conditioned on its corresponding pose stick figure x_t and the first output $G(x_t - 1)$. Consequently, the discriminator is now tasked with determining both the difference in realism and temporal coherence between the "fake" sequence $(x_t - 1, x_t, G(x_t - 1), G(x_t))$ and "real" sequence $(x_t - 1, x_t, y_t - 1, y_t)$.

2.6.4 Face GAN

They add a specialized GAN setup to add more detail and realism to the face region. After generating the full image of the scene with the main generator G, they input a smaller section of the image centered around the face (i.e. 128×128 patch centered around the nose key point), $G(X)_F$, and the input pose stick figure sectioned in the same fashion, X_F , to another generator G_F which outputs a residual $r = G_F(X_F, G(X)_F)$. The final synthesized face region is the addition of the residual with the face region of the main generator $r + G(X)_F$.

3 TECHNICAL APPROACH

3.1 Pose Estimator

For pose estimation we considered two different approach: human pose estimation in 3D using SMPL model which is a skinned vertex based model that accurately represents a wide variety of body shapes in natural human poses, or human pose estimation in 2D using OpenPose which can provide real-time 2D pose estimation using a nonparametric representation to learn the body parts in an image dataset. We examined the models by comparing different poses in both of them. Eventually we focused on the 3D approach due two main reasons:

1. 3D visual feedback is more efficient then 2D visual feedback and easier to compare.
2. In 3D approach it possible to compute the angles between joints which was used at the DTW stage 3.3.

3.2 Data Collection and Video Processing

After we choose the SMPL model we used the HMR model which is an end-to-end framework for recovering a full 3D mesh of a human body from a single RGB image. As we mention the main disadvantage of the HMR and SMPL is the inaccuracy in the recognition of the arms and wrist. We compered different poses in the HMR result and choose the exercises based on this results. We choose exercises with no emphasis on wrist and elbow angles, due to inaccuracy in the visual feedback and the inaccuracy of the angle calculates which effect on the DTW that described at 3.3). To analyzed the videos, we process it using HMR frame by frame without sampling.

3.3 DTW

The next part is to match two different videos. The user may perform the exercise in different pace from how it performs in the target video. In this paper we emphasize on the human pose and not the pace of the exercise, so we use Dynamic Time Warping (DTW) to match the stages of the exercises between the source and the target videos. In DTW, we try to dynamically identify the joint angels in the second sequence that corresponds to a given joint in the first. Given two sequences of joints angles we used DTW to find the optimal match between the sequences. To find the best match we refer only to the relevant joints angles in each exercise and the data in the sequence is their average. The selected joints angle for each exercise:

- Squat-
  - Right knee (computed by the joints- right ankle, right knee and right hip).
  - Left knee (computed by the joints- left ankle, left knee and left hip).
- Jumping Jacks
  - Right armpit (computed by the joints- right elbow, right shoulder and right hip).
  - Left armpit (computed by the joints- right elbow, right shoulder and right hip).
  - Groin (computed by the joints- right knee, left knee and the neck (which constitutes the center joint of the body)).
- Lateral Raises-
  - Right armpit (computed by the joints- right elbow, right shoulder and right hip).
  - Left armpit (computed by the joints- right elbow, right shoulder and right hip).

As shown in Figure 2.

After we found the optimal path between the sequences, we change the original videos such that the pace of both exercise it the target and the source will be similar. We did that by duplicate the frames which the pose is preform faster.

3.4 Normalization

To maintain proportion and increase accuracy we normalized the figures in two stages:

1. For each figure we found the joints of the edges of the body (head and foots) and cut the image so

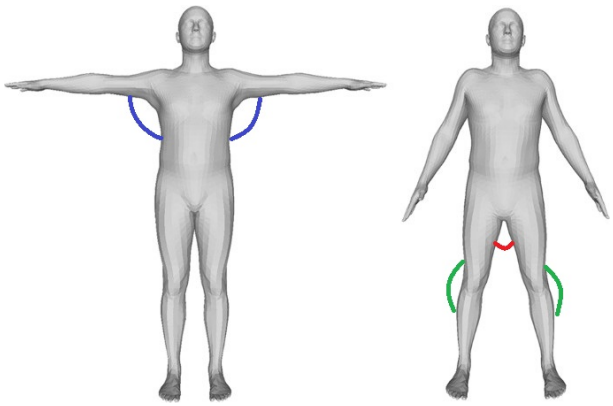

Figure 2. An visualization of the DTW angles.
In blue the armpit angle using in jumping jacks and lateral raises. In green the knee angles using in the squat. In red the groin angle using in jumping jacks.

that the image will include only the figure without unnecessary background.

2. We translate the target image on the source image according to the right ankle so that the two figures will "stand" in the same place in the image.

3.5 Visual Feedback

To present a useful visual feedback for the user we applied arrows on the HMR model to compere the main joints for the current exercises between source and target. As mention, for each exercise there are main joints that effects on the way that the user preform the pose. The selected joints for each exercise:

- Squat-
  - Right knee.
  - Left knee.

  As shown is Figure 3.
- Jumping Jacks-
  - Right knee.
  - Left knee.
  - Right elbow.
  - Left elbow.

  As shown is Figure 4.
- Lateral Raises-
  - Right elbow.
  - Left elbow.

  As shown is Figure 5.

4 CONCLUSION AND FUTURE WORK

In this paper we aid people in performing the correct posture for exercises by building a model that detects the user's exercise pose and provides useful feedback on the user's form, using a combination of the latest

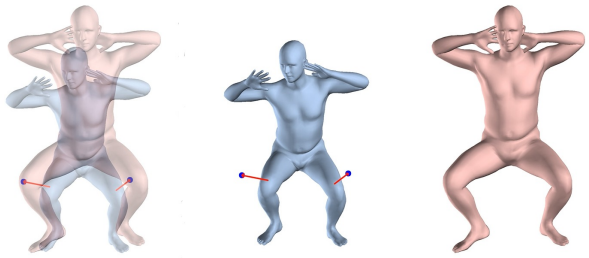

Figure 3. The visual feedback that is given for the squat exercise.

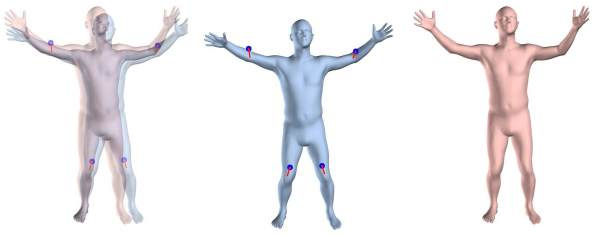

Figure 4. The visual feedback that is given for the jumping jacks exercise.

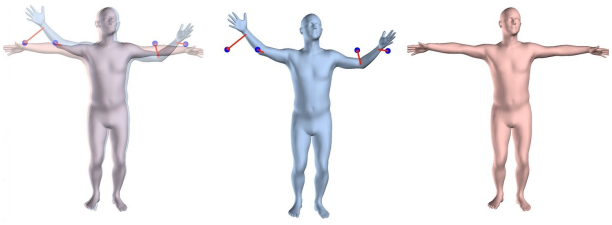

Figure 5. The visual feedback that is given for the lateral raises exercise.

advances in pose estimation, and by that help to prevent injuries and improve the quality of people's workouts. Our main improvement is that after we reconstructed Visual Feedback model that only works on core exercises we use DTW and by that allowing visual feedback for other exercises where movement is in high speed, such as powerlifting, weightlifting and CrossFit.

We have identified several extensions for this work. The first improve would be to expending the model with more exercises, and improve accuracy by using better human pose dedication, and extend it to an application that allows users to record a video and get pose feedback at any place or time. Another direction would be to improve the visual feedback by using "do as I do" method that described in Everybody Dance Now.

REFERENCES

[1] Z. Cao, G. Hidalgo, T. Simon, S. Wei and Y. Sheikh. *OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields*. 2019.

[2] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll and M.J. Black. *SMPL: A Skinned Multi-Person Linear Model*. 2015.

[3] A. Kanazawa, M.J. Black, D.W. Jacobs and J. Malik. *End-to-end Recovery of Human Shape and Pose (HMR)*. 2018.

[4] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. Osman, D. Tzionas and M.J. Black. *Expressive Body Capture: 3D Hands, Face, and Body from a Single Image (SMPL-X))*. 2019.

[5] H. Xie, A. Watatani and K. Miyata *Visual Feedback for Core Training with 3D Human Shape and Pose*. 2019.

[6] Chen and R.R. Yang *Pose Trainer: Correcting Exercise Posture using Pose Estimation* 2020.

[7] C. Chan, S. Ginosar, T. Zhou, A. Efros. 2020. *EveryBody Dance Now* 2019.