A Temporal Similarity Matrix Based Approach for 3D Pose Comparison in Exercise Videos

Sathvik Inteti
sinteti@umd.edu
Mohammed Nayeem Teli
nayeem@umd.edu

University of Maryland, College Park
Maryland, USA

Abstract

The benefits of an exercise rely on proper form. However, it is challenging to be aware of the correct form without some external assistance. People often record themselves or consult a personal trainer to analyze their form. It is impractical to have access to an expert trainer at all times. We propose a robust system that compares a sample video of an exercise with a video of an expert demonstrating the correct form. For pose comparison, we extract 3D human pose from both videos. The preprocessing is done by leveraging RepNet [4] and 4D Humans [5]. RepNet is used to extract frames containing repetitions and 4D Humans is used to predict 3D joint locations. The goal of this novel approach is to evaluate the 3D pose of an exercise video against that of an expert. The core of our method involves constructing a temporal pose self-similarity matrix (TPSM). This matrix is used to find the saddle frames within each exercise video to enable easy frame-to-frame pose comparison. The preliminary results suggest that we can accurately predict the saddle frames.

1 Introduction

When exercising, it is hard to be aware of your form because we cannot see ourselves from a third person point of view. Many people will hire a personal trainer to pinpoint exactly where their form is wrong. However, it is impractical to have a personal trainer available at all times. That's why it's very common to see people in the gym recording themselves to analyze their form. It is not likely that the users will have a ideal setup at which they can record themselves performing movements. These videos have multiple challenges involving various viewpoints, contain extraneous people and objects, and different resolutions. Exercises are also usually performed in repetitions. Maintaining overall proper form boils down to maintaining it in each repetition. It is imperative that any approach focused on form analysis can compare each rep individually. Dwibedi et al [4] propose RepNet which is a neural network architecture that can predict repetitions within videos but it does not address the pose in those repetitions.

In this paper, we propose a robust approach that can estimate and evaluate poses from in-the-wild exercise videos. Our approach compares exercise videos using a temporal similarity matrix. We estimate 3D human pose from these videos to provide a unique visual analysis of people's form.

The foundation of our approach is a strong monocular 3D human pose estimation model. There has been tremendous progress on monocular 3D human pose estimations within the last couple of years [5, 9]. The emergence of transformers [7] in deep learning architectures have led to new state-of-the-art models. Goel et al [5] propose a novel architecture to leverage Vision Transformers [3] with a Transformer decoder to estimate an SMPL model for humans inside a video. Zhu et al [9] introduce Spatial and Temporal Transformers to predict 3D skeletons from 2D joint positions. Both approaches achieve state-of-the-art performance on quality benchmarks, Human3.6M [6] and 3DPW [8].

We propose a pipeline that compares the form a sample video of an exercise with a video of an expert demonstrating the correct form. We leverage RepNet and 4D Humans to preprocess our videos. RepNet is used to partition our video by repetitions. 4D Humans is used to predict 3D joint locations for each partition of the original video. We identify several challenges that prevent us from taking a naive approaches to form comparison. To account for these challenges, we introduce a novel temporal pose self-similarity matrix (TPSM). Our approach works by selecting the most relevant frames for the exercise in the videos. We then do simple frame-to-frame pose comparison between the videos.

We introduce a new dataset, FitFrames. Our dataset builds off of an open-source dataset available on Kaggle [1]. Our dataset consists of video frames of several different exercise movements. We annotate each rep of an exercise with a range of frames that would be valid saddle frames. The meaning of saddle frames are discussed in subsequent sections. We also annotate each rep with a binary score representing whether the form is good or bad. We validate our method on this FitFrames dataset.

2 Related Works

2.1 Monocular 3D Human Pose Estimation

Monocular 3D Human Pose Estimation is the task of estimating joint locations in 3D space from a single RGB camera. New approaches have emerged which leverage Transformers to push the state-of-the-art. Goel et al [5] propose a method to estimate the SMPL models for humans in an image inside called HMR 2.0. This approach relies on a Vision Transformer and a Transformer Decoder at the heart of its approach. Goel et al leverage PHALP to track humans throughout videos. Together, Goel et al propose a pipeline to fit a SMPL model on humans throughout videos. Zhu et al [9] propose MotionBert, an alternative transformer based approach. Zhu et al propose the use of spatial and temporal transformers in conjunction to predict 3D joint locations from 2D skeletons. Unlike Goel et al, MotionBert is a temporally consistent approach. However, MotionBert requires a preprocessing step of extracting 2D skeletons from another work, AlphaPose. Both approaches have shown state-of-the-art performance on many quality benchmarks such as Humans3.6M [6] and 3DPW [8].

2.2 Video Repetition Counting

RepNet [4] is the most prominent work on counting repetitions of an action in videos. RepNet incorporates a layer in its architecture that constructs a Temporal Self-similarity Matrix

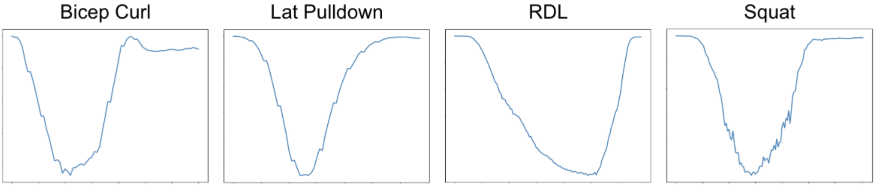

Figure 1: **Pose Similarity of Every Frame to Starting Frame in One Rep Exercise Videos:** We plot the pose similarity of every frame with the starting frame. A convex pattern is noticeable for all exercise classes. This aligns with our intuition as most exercises the starting and ending positions are the same. Therefore, there must be one position where we are the most dissimilar to the starting position and ending position (e.g. the point at which we begin to return to the start/end position). We call the frame corresponding to this position, the saddle frame.

(TSM). First, the frames are embedded using 2D and 3D convolutions. TSM is constructed by finding the cosine similarity between all pairs of embedding. TSM visibly shows multiple diagonal lines corresponding to repetitions in a video. Repnet uses a deep learning architecture to process the TSM and output periodicity scores and period counts for each frame.

3 Methodology

3.1 Priors for Input Data

To constrain this task, we leverage a few priors about the input data. Most exercises start and end with the same pose (e.g, Figure 1). We make this assumption for all exercise videos. These exercises can also be performed in repetitions successively. Therefore, we assume that such an exercise will be present somewhere in the video input. Our approach "crops" the video to remove extraneous information and contain only frames relevant to the movement in the preprocessing steps. For the current version of our approach, we work with data where the focus of the camera is on the subject of interest. We plan to add additional support to choose between subjects in the future.

3.2 Preprocessing Steps

We first want to preprocess our input videos by partitioning the video based on the reps of the exercise and extracting 3D joint locations. Figure 2 illustrates the pipeline for the preprocessing steps. To preprocess the videos, we leverage RepNet [4] to extract period counts and periodicity scores for each frame. We discard any frames which have a periodicity score under 0.5. We assign an integer label to each frame in the video corresponding to the repetition it is part of. This step is only useful when the exercise is performed more than once within the video. In cases where RepNet detects no repeated action, we assign the same label to all the frames in the video. We group the frames by label, thus successfully partitioning the video by repetitions. We are able to significantly reduce the cost of 3D human pose estimation by only computing on relevant partitions of the video.

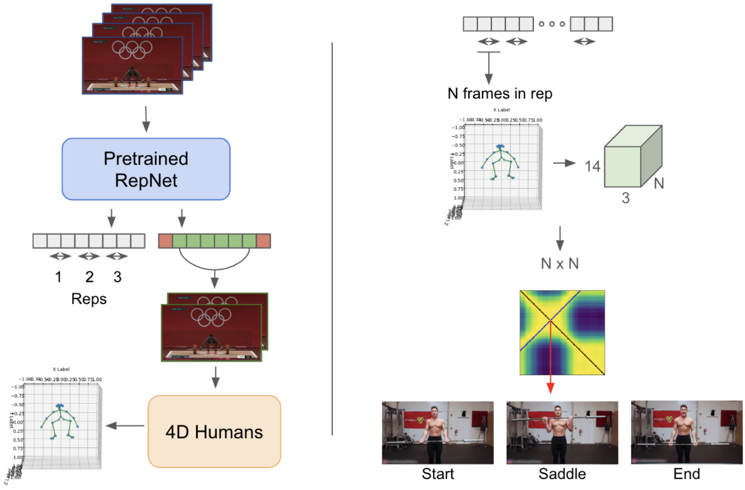

Figure 2: **(Left) Preprocessing Steps:** (1) We leverage RepNet [4] to partition videos into repetitions of the exercise. The green represents frames that are kept by RepNet. (2) 4D-Humans [5] is used on the frames kept by RepNet to predict 3D Skeletons. **(Right) Saddle Frame Detection:** For each repetition, we construct a pose self-similarity matrix. We diagonally traverse this matrix to detect the intersection point which corresponds to our saddle frame.

To extract 3D joint locations for each partition of the video, we follow the model architecture of 4D Humans [5]. We slightly modify the architecture to predict 3D joint locations rather than SMPL parameters. Due to time constraints, the 3D joint positions are extracted from the SMPL model rather than training a model to predict them directly. We utilize 25 OpenPose [2] 3D joint locations. 4D Humans is a per-frame estimation technique and can sometimes lead to temporal inconsistencies of 3D poses between frames. The result can be a slightly noisy outputs which are visible in Figure 1.

3.3 Saddle Frame Detection

There are several challenges for detecting saddle frames in exercise videos: (1) Different samples may perform the exercise with varied speeds, (2) Different samples may be smaller/bigger (i.e. in terms of height and weight). Our approach is resistant to all of these challenges.

To motivate our approach, consider a comparison between the 3D human poses of two images. To address concern (2), we make our pose similarity metric according compare limbs (the valid connections of two joints as specified by OpenPose [2]) rather than joints. For each limb (denoted i) in OpenPose, we take the sample limb (s_i) and the expert limb (e_i). We find the cosine similarity between the two vectors s_i and e_i. We omit limbs that do not hold much significance such as ears and toes. We take the average similarity of all limbs to be the similarity for two poses. This approach only considers angles and excludes the length of the limbs making it resistant to challenge (2). We refer to this similarity metric as pose similarity.

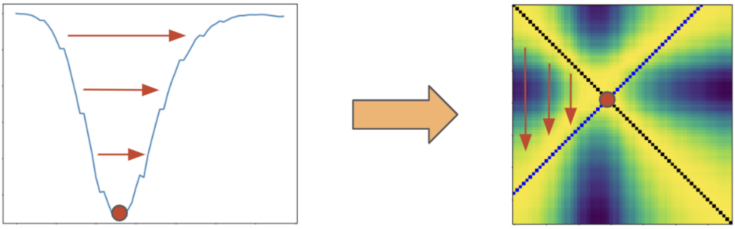

Figure 3: **Temporal Pose Self-Similarity Matrix:** The motivation for TPSM is to capture the pose similarity over time while being resistant to noisy outputs for a single frame. The arrows indicate the similar poses of non-neighboring frames. Visually, we can see that TPSM captures the information in the plots from Figure 1 well.

The intuition to address concern (1) is to select frames from both videos that encapsulate the essence of the movement. Consider a push-up exercise. What would be the frames that capture the movement? You would probably point towards the starting position, the position at which your chest is closest to the ground, and the ending position. We argue that this generalizes to most exercises. Figure 1 supports our claim. It is clear the starting and ending points are important. In between, we move towards a position where there is the muscle contraction and then return to the starting position. We want to find that position where we switch to moving back to the starting position. We define the saddle frame is the frame at which the pose is most different from the starting and ending pose. Simply, this is the position where we begin to return to the starting position (e.g. the end of the movement).

Our initial approach that we considered is to plot how similar every frame is to the starting frame of the repetition. Figure 1 illustrates the results for various exercises. We fit a least squares polynomial on the plot to make the plot differentiable. The x-value of the global minimum will be the saddle frame. This approach is the best for videos that are perfectly cropped by RepNet [4]. However, there will be inputs that have only one rep and lots of extraneous frames. These extraneous poses can be much very dissimilar to the poses in the rep and would be the global minimum as opposed to the saddle frame. In order to address this case, we propose a temporal pose self-similarity matrix approach.

To find such position, we introduce a novel pose self-similarity (TPSM) matrix. We were motivated by RepNet's self-similarity matrix which was used to find patterns in the frame embeddings which ultimately showed repetitions. The idea is to see if patterns emerged when it comes to pose inside one rep of an exercise. Consider a pose at frame i, p_i, Figure 1 implies that if i is a frame part of a repetition, then there will exist a pose at frame j p_j that is very similar to p_i. To find j given i, we can construct a TPSM matrix that demonstrates the similarity between every pair of frames in the video. To construct the SPS matrix, we take the pose similarity between the pose of one frame against all other frames in the rep. Figure 3 shows an example of a TPSM matrix and how it embeds the information of the plots from Figure 1. We can then iterate through the matrix column by column. Once we find a column that has high similarity with frames that are not neighbors, we can consider this the beginning of the movement. In similar a similar manner, we can iterate through the columns in reverse to find the ending position of the movement. The other columns are discarded. We can identify that most matrices now have a distinct 'X' shape. Figure 3 provides an example of a TPSM. The yellow indicates high similarity while blue indicates

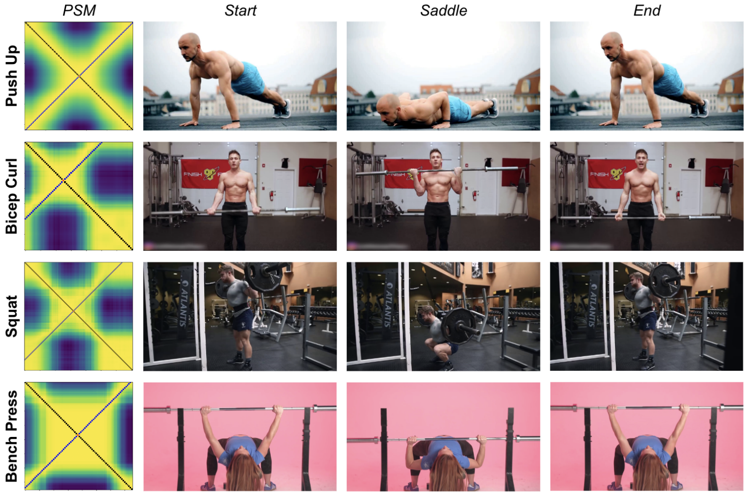

Figure 4: **Saddle Frame Detection on FitFrames Dataset Videos:** We demonstrate our capability to detect saddle frames when there is one subject in a video. Our approach is robust to a variety of exercise classes in FitFrames.

	Accuracy ↑		Accuracy ↑
Push Up	0.978	Pull Up	-
RDL	-	Deadlift	-
Bench Press	-	Squat	-
Bicep Curl	-	Lat Pulldown	-
	Total Accuracy	-	

Table 1: Accuracy of TPSM on FitFrames Classes

low similarity. The yellow along the diagonal of the matrix represents the similarity between a frame and its neighboring frames. The other diagonal represents the similarity between a frame i and some non-neighboring frames. These diagonals intersect at some point. The intersection region is a group of frames which can be considered a saddle frame. In order to find the saddle frame, we traverse the matrix diagonally. We consider the sum of the similarity scores on each diagonal. We select the highest sum with an average similarity (of all elements on the diagonal) above a threshold. The selection is the saddle frame for this repetition.

4 Experiments

4.1 FitFrames Dataset

FitFrames is dataset of video frames of a single rep of an exercise. We built our dataset off of a open source exercise video dataset from Kaggle [▯]. The original dataset is a compilation of non-copyrighted exercise videos from YouTube. These videos contained multiple repetitions of the exercise being performed. We leveraged RepNet to extract a single repetition from each video. There were several cases where RepNet failed to detect the correct repetition. We handled the following failure cases: (1) Any video that RepNet detected that did not contain a full repetition was scrapped, (2) Videos that RepNet detect the wrong starting frame but contains a movement similar that of an exercise were kept, (3) Videos that contain multiple repetitions that were not detected by RepNet were manually cropped. FitFrames contains annotations to setup the following two experiments: (1) Accuracy of Saddle Frame Detection, (2) Binary Classification of Exercise Form. We annotate each rep of an exercise with a range of frames that would be valid saddle frames. We also annotate each rep with a binary score representing whether the form is good or bad.

4.2 Accuracy of Saddle Frame Detection

We evaluate TPSM on saddle frame detection accuracy. Each class of exercise is evaluated separately in order to detect whether TPSM struggles with one exercise over another. We show the results of this experiment in Table 1. Preliminary results show that our method is able to accurately predict the saddle frames for push ups. TPSM achieves around 97% accuracy on Push Ups. We show the predicted saddle frames in Figure 4 for a few of FitFrames classes. As of 5/2/24, the annotations for other classes of FitFrames are not complete so accuracy results in Table 1 are not available currently.

4.3 Binary Classification of Exercise Form

We want to evaluate our approach on whether we can classify correct movements vs incorrect movements in exercise videos. We have assigned one video to be the expert video for each class in FitFrames. The expert's form will be compared to that of the sample video. We assign a similarity score between 0 and 1 for the sample. If this score is above a threshold then we classify it as a correct movement otherwise it is classified as incorrect. We evaluate our method at different threshold values to determine the best one. As of 5/2/24, the binary classification annotations for FitFrames are not yet complete.

5 Limitations

There much work to do to make this approach production ready. RepNet frequently fails to detect the correct starting frame. It may instead predict the saddle frame as the starting frame which us humans know doesn't make sense in the context of an exercise. There is an opening to work on RepNet predicting the correct starting frame for each exercise. In addition, 4D Humans is a per-frame optimization which can lead to temporal inconsistencies in the predicted 3D joint locations. MotionBert promises temporal consistency but requires a preprocessing inputs to extract 2D skeletons. We intend to explore an architecture that has the best of both worlds. One that is temporally consistent and directly predicts from input videos.

6 Conclusion

We have shown that we can build an approach to analyze people's exercise form given only exercise videos. Our approach is robust to numerous variations in input data. We are hopeful that our proposed methods (i.e. Figure 1 and Figure 3) of analyzing 3D human pose can be applicable in other contexts as well. Our proposed pose similarity metric is resistant to changes in shape and can be used as a loss function for model optimization. We are hopeful to turn our approach into an application which can be a great aid to people around the world.

References

[1] H Abdillah. Workout/exercises video. Open Source Dataset on Kaggle.

[2] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):172–186, January 2021. ISSN 1939-3539. doi: 10.1109/tpami.2019.2929257. URL http://dx.doi.org/10.1109/TPAMI.2019.2929257.

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2020.

[4] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. Counting out time: Class agnostic video repetition counting in the wild. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[5] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4d: Reconstructing and tracking humans with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, October 2023. doi: 10.1109/iccv51070.2023.01358. URL http://dx.doi.org/10.1109/ICCV51070.2023.01358.

[6] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, jul 2014.

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

[8] Timo von Marcard, Roberto Henschel, Michael J. Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[9] Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. Motionbert: A unified perspective on learning human motion representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.