# Robust Vision-Based Workout Analysis Using Diversified Deep Latent Variable Model

Hao Xiong, Shlomo Berkovsky, Roneel V. Sharan, *Senior Member, IEEE*, Sidong Liu and Enrico Coiera

***Abstract*— Exercising has various health benefits and it has become an integral part of the contemporary lifestyle. However, some workouts are complex and require a trainer to demonstrate their steps. Thus, there are various workout video tutorials available online. Having access to these, people are able to independently learn to perform these workouts by imitating the poses of the trainer in the tutorial. However, people may injure themselves if not performing the workout steps accurately. Therefore, previous work suggested to provide visual feedback to users by detecting 2D skeletons of both the trainer and the learner, and then using the detected skeletons for pose accuracy estimation. Using 2D skeletons for comparison may be unreliable, due to the highly variable body shapes, which complicate their alignment and pose accuracy estimation. To address this challenge, we propose to estimate 3D rather than 2D skeletons and then measure the differences between the joint angles of the 3D skeletons. Leveraging recent advancements in deep latent variable models, we are able to estimate 3D skeletons from videos. Furthermore, a positive-definite kernel based on diversity-encouraging prior is introduced to provide a more accurate pose estimation. Experimental results show the superiority of our proposed 3D pose estimation over the state-of-the-art baselines.**

## I. Introduction

Exercising and physical activity have many benefits for physical and mental health [1]. It is commonplace these days that more and more people take part in fitness workouts. This is mainly due to the contemporary sedentary lifestyle, which subtly motivates people to improve their health by exercising. Due to time and financial limitations, many prefer to perform their workouts independently at home. As such, numerous workout video tutorials have been published online, to illustrate the steps of doing the exercises correctly [2]. However, improperly following these workout tutorials may cause harm and body injuries.

To this end, Nagarkoti *et al.* developed a system providing visual feedback to people learning a workout from online video tutorials [3]. Firstly, the system required the learners to capture the video of themselves performing the workout. Then, dynamic time warping was deployed [4], to align the learner's and trainer's videos and match video frames for pose accuracy measurement. Upon selecting the matching frames, a 2D pose estimation method was applied to detect the 2D skeleton in the frames [5]. Finally, affine transformation was applied to align the 2D skeletons, detect the differences between them, and provide feedback to the learner. However, the main drawbacks of this approach were:

- The affine transformation could not accurately align the 2D skeletons due to the body shape variations across learners. The variability of the body shape affected the size of the detected skeletons, so that the alignment with the affine transformation could be inaccurate even if the poses of the 2D skeletons were identical.
- The affine transformation struggled to align 2D skeletons when the poses were not identical, so that the pose measurements were inaccurate. As a result, the system provided biased feedback to learners.

Rather than using 2D skeletons, we propose to estimate 3D skeletons based on images and then directly compare the joint angles between the 3D skeletons for the pose accuracy measurements, without relying on the affine transformation. To achieve this, we propose a diversified deep latent variable model with a diversity-encouraging (DE) prior. The DE prior ensures a greater diversity in the created latent variables, to span all the possible types of poses and body sizes. As a result, the model allows to achieve more accurate 3D pose estimations.

We experimentally evaluated the accuracy of the 3D skeleton estimations. The Human 3.6M dataset containing 15 types of poses with ground truth was utilized to evaluate the estimations [10]. We compared the performance of our method – with and without the DE prior – to a state-of-the-art baseline. We evaluated the accuracy of the detection of all the 15 poses in Human 3.6M and computed the mean error. The results show that the proposed method with DE prior outperformed the evaluated baseline methods.

## II. System and Methods

### A. Pre-processing and Overview

Our method aims to perform vision-based workout analysis. The workflow of the method is shown in Fig. 1. Given two videos of a trainer and a learner, dynamic time warping is applied to align the video images [3]. Upon matching the two, our algorithm estimates the pose accuracy in order to provide visual feedback to the learner. The trainer's and learner's images serve as the inputs. The 2D skeletons are detected first and their key points are used to fit to 3D skeletons. The differences between the two 3D skeletons are then identified and visually highlighted, to help the learner.

Algorithm 1 summarizes the proposed 3D skeleton reconstruction algorithm. As can be seen, the algorithm relies on a pre-trained deep latent variable model and the 2D skeleton detection method proposed by [5]. Having detected the 2D skeleton, the algorithm iteratively optimizes the

All the authors are affiliated with the Australian Institute of Health Innovation, Macquarie University, Sydney, NSW 2109, Australia (e-mail: {hao.xiong,shlomo.berkovsky,roneel.sharan,sidong.liu,enrico.coiera}@mq.edu.au).

978-1-7281-1990-8/20/$31.00 ©2020 IEEE

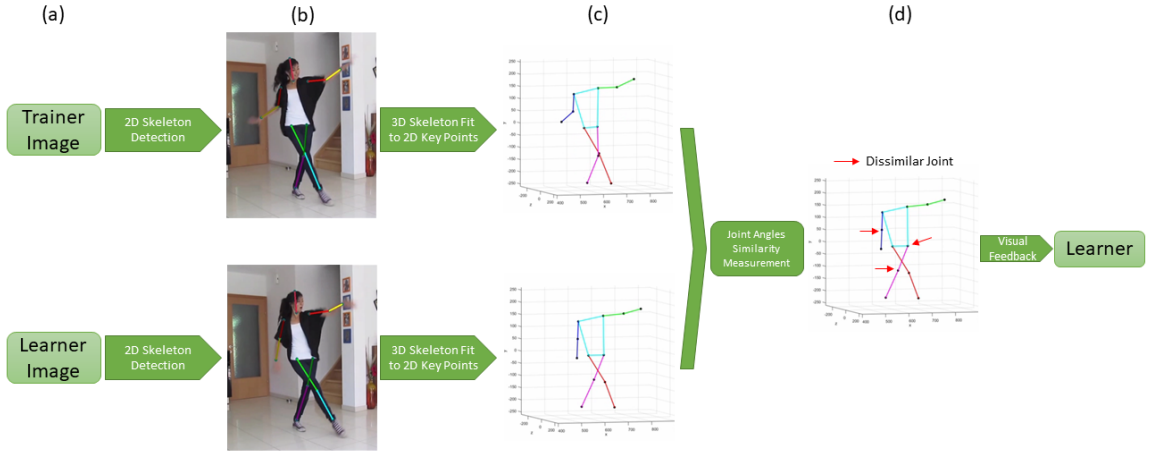


Fig. 1. **Workout analysis workflow**: (a) Trainer and learner images, (b) 2D skeletons, (c) 3D skeletons, (d) Dissimilarities between the 3D skeletons.

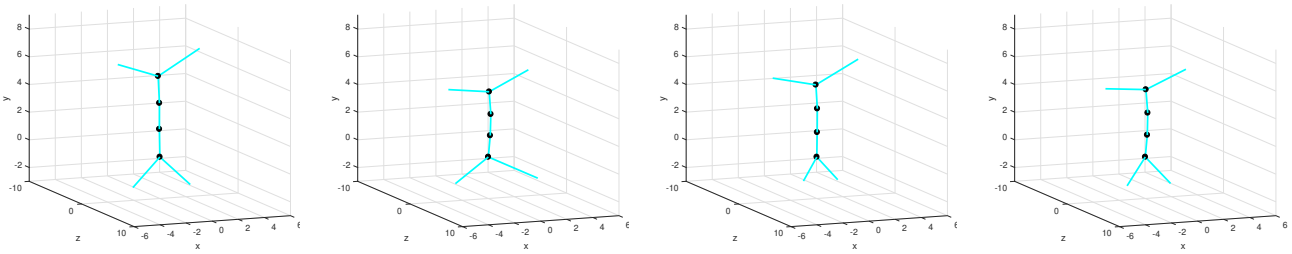

Fig. 2. Sampling the latent space of the upper body skeleton.

**Algorithm 1** 3D Skeleton Reconstruction Algorithm

**Input:** trained diversified deep latent variable model with latent variables $z_{pose}$ and $z_{skel}$
**Input:** reference trainer image $I_{ref}$, input learner image $I_{inp}$
**Output:** reconstructed 3D skeletons of $I_{ref}$ and $I_{inp}$

1: **2D skeleton:** apply the method of [5] to $I_{ref}$ and $I_{inp}$
2: **while** not converged **do**
3: optimize rigid transformation parameters $s$, $R$ and $t$ for fixed latent variables $z_{pose}$ and $z_{skel}$
4: optimize latent variables $z_{pose}$ and $z_{skel}$ for fixed rigid transformation parameters $s$, $R$ and $t$
5: calculate the error of objective function (6)
6: **end while**

rigid transformation parameters and the skeleton/pose latent variables, until convergence.

### B. CMU Motion Capture Database

To train our method, we used the CMU motion capture database[1], which provides video recordings of more than 100 humans. The CMU database also contains hundreds of motions grouped into categories, which renders it perfect for training. Since our pose estimation algorithm is 3D-oriented, it fits the CMU 3D human skeleton to an image, to estimate the 3D skeleton. To estimate the 3D pose, our algorithm learns the latent variables of human motions. Then, specific 3D skeleton is estimated by optimizing the learned latent space.

[1] http://mocap.cs.cmu.edu/

### C. Diversified Deep Latent Variable Model

*1) Deep Gaussian Process:* In a deep Gaussian process [6], we define output variable $Y \in R^{N\times D}$, intermediate layer latent variable $X \in R^{N\times Q_x}$, and bottom layer latent variable $Z \in R^{N\times Q_z}$. Thus, the output $Y$ of the deep Gaussian process is associated with two stacked latent variables $X$ and $Z$, and the objective function is defined as:

$$F(\theta) = \log \int_{X,Z} p(Y|X)p(X|Z)p(Z)\mathrm{d}X\mathrm{d}Z. \qquad (1)$$

The model parameters $\theta$ are learned by variational inference.

In our case, $Y$ refers to the estimated 3D skeleton. With the learned model parameter $\theta$, a new 3D human skeleton can be generated by sampling the bottom layer's latent space $Z$, as shown in Fig. 2.

*2) Kernel-Based Diversity Prior:* In our model, the kernel $K_\phi$ consists of a positive-definite correlation function $R(\phi_i, \phi_j)$ and the "prior kernel" $\sqrt{\pi(\phi_i)\pi(\phi_j)}$ [7]. Then, the kernel can be expressed as:

$$K(\phi_i, \phi_j) = R(\phi_i, \phi_j)\sqrt{\pi(\phi_i)\pi(\phi_j)}, \qquad (2)$$

where $R(\phi_i, \phi_i) = 1$.

Here, the kernel-based diversity prior allows for repulsion, where we use the probability product kernel to construct each

TABLE I

3D EUCLIDIAN JOINT DETECTION ERROR OF THE BASELINE METHOD OF [9] VS THE PROPOSED METHOD WITH AND WITHOUT THE DIVERSITY-ENCOURAGING PRIOR. EVALUATED ON THE 15 POSES OF THE HUMAN3.6M DATASET.

| Method | Direction | Discuss | Eat | Greet | Phone | Pose | Purchase | Sit | SitDown | Smoke | Photo | Wait | Walk | WalkDog | WalkPair | Mean |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Yasin *et al.* [9] | **67.8** | 58.7 | 90.3 | 72.1 | **78.2** | 75.7 | **71.9** | 103.2 | 132.8 | **91.3** | 91.6 | 84.7 | 70.9 | 81.2 | 76.7 | 83.7 |
| w/o DE prior | 71.6 | 49.5 | 86.1 | **53.2** | 89 | 74.9 | 79.3 | 107.4 | 126.5 | 97.7 | 74.3 | 88.2 | 65.8 | 86.4 | 71.9 | 81.4 |
| with DE prior | 73.4 | **43.2** | **82.1** | 60.7 | 84.8 | **70.4** | 76.8 | **97.1** | **114.6** | 102.4 | **68.9** | **79.4** | **64.3** | **78.5** | **66.2** | **77.5** |

element of the kernel matrix and define the repulsion. Thus, every kernel element is expressed as an inner product of two probability distributions:

$$K(\phi_i, \phi_j; \rho) = \int_{\chi} f(x|\phi_i)^{\rho} f(x|\phi_j)^{\rho} dx, \tag{3}$$

where $\rho > 0$. For simplicity, let $\rho = 1$. The normalized variant $R(\phi_i, \phi_j)$ can be derived as follows:

$$R(f_1, f_2) \quad = K(f_1, f_2)/\sqrt{K(f_1, f_1)K(f_2, f_2)}. \tag{4}$$

*3) Diversity-Encouraging Objective Function:* In the proposed model, we maximize $F(\theta)$

$$F(\theta) = \log \int_{X,Z} p(Y|X)p(X|Z)p(Z)\mathrm{d}X\mathrm{d}Z + \log |K_{\phi}^{X}|^{\lambda_1} + \log |K_{\phi}^{Z}|^{\lambda_2}. \tag{5}$$

Here, $|K_{\phi}^{X}|$ and $|K_{\phi}^{Z}|$ are the diversity encouraging (DE) priors on the intermediate and bottom layers, respectively.

### D. Objective Function and Optimization

With the detected 2D joint positions $J_{est,k}$, we can fit the 3D skeleton. To achieve this, an objective function

$$E = \sum_{joint\ k} ||sRF(p(z_{skel}), p(z_{pose}))_k + t - J_{est,k}||^2, \tag{6}$$

is minimized, where $s$, $R$, $t$ refer to the scaling, rotation and translation parameters of a rigid transformation, and $z_{skel}$ and $z_{pose}$ are the latent variables of the skeleton and pose, respectively.

Given $z_{skel}$ and $z_{pose}$, the output of $F(p(z_{skel}), p(z_{pose}))$ is the 3D skeleton in the coordinates of the CMU training dataset. Thus, the reconstructed skeleton produced by $F()$ is fit onto the estimated joint points $J_{est,k}$ with the rigid transformation. Here, 2D joint points of ankles, knees, hips, elbows, wrists, and shoulders are estimated.

To solve (6), we initially estimate the parameters of the rigid transformation with least squares using the mean shape of the upper body. Then, the latent variables of the pose and skeleton are optimized using the trust-region-reflective algorithm [8]. The optimization is performed iteratively until convergence.

## III. EVALUATION

In the experiments, we compared our method with [9], which integrated the annotated 2D pose images and the 3D motion capture data from the CMU dataset to achieve simultaneous 2D pose estimation and 3D pose recovery. To the best of our knowledge, this is the state-of-the-art pose estimation method. It was trained on the same CMU data, which allows us to use it as a comparative baseline.

We evaluated the proposed method on the Human 3.6M dataset [10], which contains several subjects at 15 poses such as eating, sitting, smoking, and more. The skeleton data in Human 3.6M consists of 12 body joints including ankles, knees, hips, wrists, elbows, and shoulders on both sides of the body. In similar to the evaluation protocol of [9], we drew one in every 64 frames from the sequences of Subject 11 in Human 3.6M for testing. Rigid transformation with Procrustes analysis were used to align the reconstructed and the ground truth 3D poses. After alignment, the average 3D Euclidean joint error was measured [11].

Table I shows the error of the joint detection for the proposed method (with and without the DE prior) and the baseline method of [9]. The results show that for 11 out of the 15 poses in Human 3.6M, our method outperforms the baseline. Notably, for 10 out of the 11 poses, the DE prior allowed to achieve a lower joint detection error. Overall, the mean error achieved by the proposed method with the DE prior is 4.8% lower than without the DE prior and 7.4% lower than the baseline method.

We also demonstrate examples of the 3D skeleton reconstructed by the proposed algorithm. We applied our method to a video sequence of a woman dancing (selected frames with the reconstructed 3D skeletons are shown in Fig. 3). Although these poses are reasonably complex and diverse, the estimated 3D skeletons closely resemble the poses of the dancer. We also demonstrate two examples of test subject 11 from the Human 3.6M dataset (see Fig. 4). As can be seen, the 3D skeletons generated by the proposed method are accurate in both cases. These examples indicate that our algorithm is applicable to both image- and video-based 3D skeleton estimation tasks.

## IV. DISCUSSION AND CONCLUSIONS

This work aimed to develop a 3D pose estimation tool, allowing for provision of visual feedback to people learning how to perform exercises. Prior methods, exploiting 2D skeleton estimation and affine transformation, often struggle with the varying human body sizes and the diversity of poses. To alleviate this, we proposed a deep latent variable model that was shown to accurately estimate 3D skeletons from videos and images alike, and to outperform the state-of-the-art methods. By integrating a diversity-encouraging prior into the latent variable model, we were able to better capture the characteristics of the poses and skeleton, and further improve the accuracy of the 3D skeleton estimations.

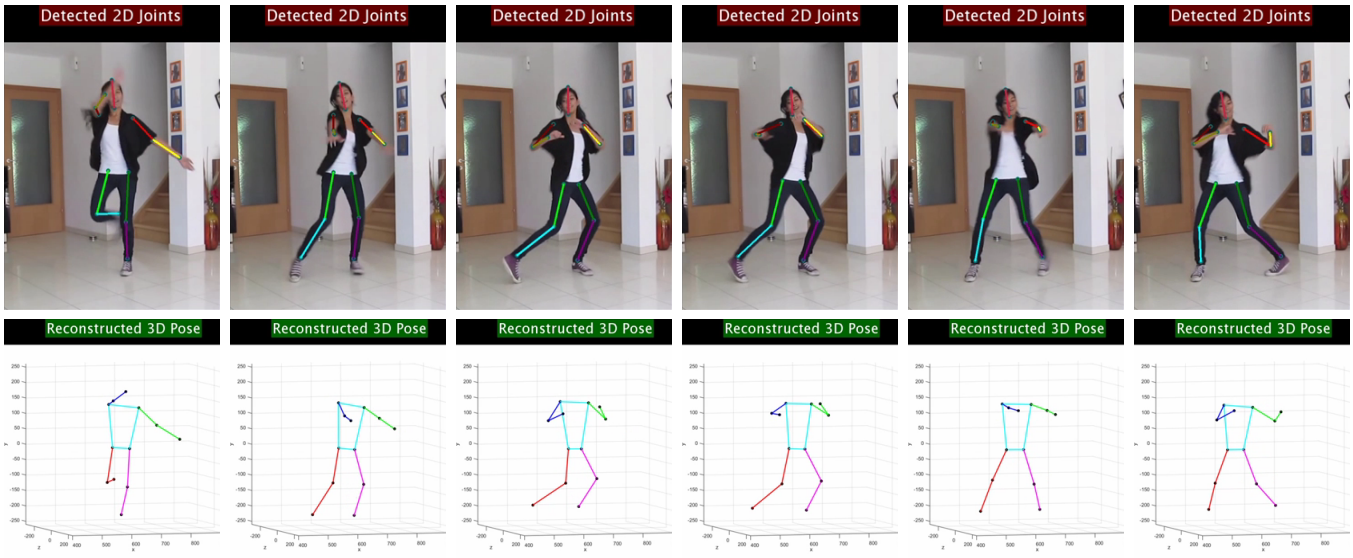


Fig. 3. **Video-based 3D Pose Estimation.** Top row: video frames with 2D skeletons detected with [5]. Bottom row: estimated 3D skeletons.

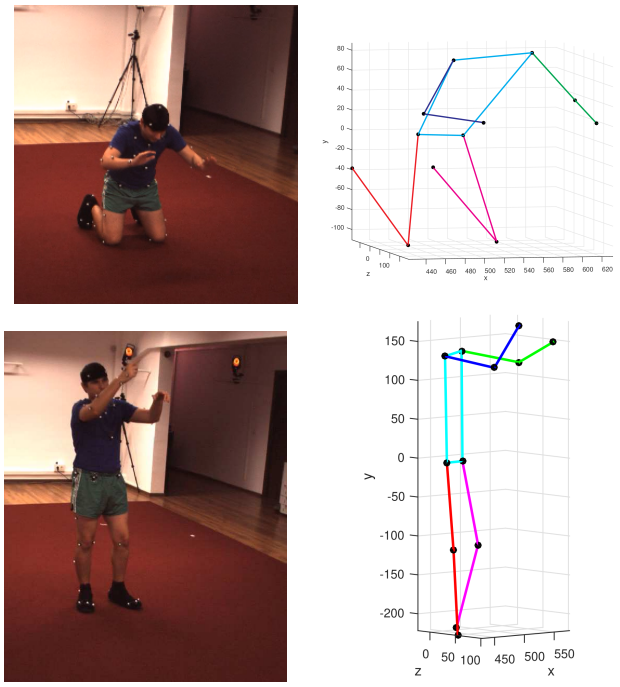

Fig. 4. **Human 3.6M Dataset Examples:** (left) input images, (right) reconstructed 3D skeletons.

While our work achieved low error rates, it needs to be integrated with video tutorials. It is important not only to identify incorrect motions, but also the timing of these motions, and be able to provide timely feedback to learners. Our work paves the way to future adaptive applications offering personalised feedback to learners. A range of intelligent interfaces and persuasive technologies can benefit from the developed pose estimation methods [12], although their usability and efficacy is yet to be validated.

## ACKNOWLEDGMENT

The authors acknowledge the support of the NHMRC Partnership Centre for Health System Sustainability (grant ID 9100002) administered by the Australian Institute of Health Innovation, Macquarie University.

## REFERENCES

[1] F. J. Penedo and J. R. Dahn, "Exercise and well-being: A review of mental and physical health benefits associated with physical activity," *Current Opinion in Psychiatry*, vol. 18, no. 2, pp. 189-193, 2005.

[2] Samsung Electronics America, "Samsung Health," Internet: *https://www.samsung.com/us/support/owners/app/samsung-health*, 2018.

[3] A. Nagarkoti, R. Teotia, A. K. Mahale and P. K. Das, "Realtime indoor workout analysis using machine learning and computer vision," *41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2019, pp. 1440-1443.

[4] M. Muller, " Dynamic Time Warping," *Information Retrieval for Music and Motion*, Berlin: Springer, 2007, pp. 69-84.

[5] Z. Cao, T. Simon, S. E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *CVPR*, 2017, pp. 7291-7299.

[6] A. C. Damianou and N. D. Lawrence, "Deep Gaussian processes," in *International Conference on Artificial Intelligence and Statistics*, 2013, pp. 207-215.

[7] J. Zou and R. Adams, "Priors for diversity in generative latent variable models," in *NIPS*, 2012, pp. 2996–3004.

[8] R. H. Byrd, R. B. Schnabel, and G. A. Schultz, "A trust region algorithm for nonlinearly constrained optimization," *SIAM Journal on Numerical Analysis*, vol. 24, no. 5, pp. 1152–1170, 1987.

[9] H. Yasin, U. Iqbal, B. Krüger, A. Weber, and J. Gall, "A dual-source approach for 3D pose estimation from a single image," in *CVPR*, 2016, pp. 4948-4956.

[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," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 7, pp. 1325-1339, 2014.

[11] E. Simo-Serra, A. Ramisa, G. Alenyà, C. Torras, and F. Moreno-Noguer, "Single image 3D human pose estimation from noisy observations," in *CVPR*, 2012.

[12] S. Berkovsky, J. Freyne, and H. Oinas-Kukkonen, "Influencing individually: Fusing personalization and persuasion, *ACM Trans. on Interactive Intelligent Systems*, vol. 2(2), e. 9, 2012.