UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

Deep Learning Approach to Skeletal Performance Evaluation of Physical Therapy Exercises

Permalink

https://escholarship.org/uc/item/9f28d94s

Author

Garg, Bhanu

Publication Date

2023

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA SAN DIEGO

Deep Learning Approach to Skeletal Performance Evaluation of Physical Therapy Exercises

A thesis submitted in partial satisfaction of the
requirements for the degree Master of Science

in

Electrical Engineering (Machine Learning and Data Science)

by

Bhanu Garg

Committee in charge:

Professor Sujit Dey, Chair
Professor Pamela Cosman
Professor Pengtao Xie

2023

Copyright

Bhanu Garg, 2023

All rights reserved.

The Thesis of Bhanu Garg is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2023

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to express my heartfelt gratitude to all those who have contributed to the completion of this thesis.

First and foremost, I am deeply grateful to my supervisors, Dr Sujit Dey and Dr Pamela Cosman, for their unwavering guidance, support, and invaluable insights throughout this research journey. Their expertise, patience, and encouragement have been instrumental in shaping this work and refining my understanding of the subject matter.

I would also like to acknowledge Alexander Postlmayr, without whom my research would have no doubt taken five times as long. It is his support that helped me in an immeasureable way.

The manuscript, in full, is a reprint of the material as it appears in:

B Garg, A Postlmayr, P Cosman, S Dey. ”Short: Deep Learning Approach to Skeletal Performance Evaluation of Physical Therapy Exercises”, IEEE International Conference on Connected Health: Applications, Systems and Engineering Technologies - 2023.

The thesis author was the primary investigator and author of this paper.

ABSTRACT OF THE Thesis

Deep Learning Approach to Skeletal Performance Evaluation of Physical Therapy Exercises

by

Bhanu Garg

Master of Science in Electrical Engineering (Machine Learning and Data Science)

University of California San Diego, 2023

Professor Sujit Dey, Chair

At-home exercising strongly predicts physical therapy patient outcomes, underscoring the need for analyzing patient behaviors at-home via remote patient monitoring. Contemporary methods for remote patient monitoring rely on specialized sensors, i.e., Inertial Measurement Units, RGB-Depth cameras, motion capture systems, or stereo vision which are costly and not scalable to all physical therapy patients. Here, we observe a lack of literature using only a monocular RGB camera. In this thesis, we demonstrate a skeletal feedback model for at-home exercises using only video acquired from a smartphone camera. We propose models for (i) Patient Performance Evaluation - which classifies the correctness of exercises, and (ii) Guidance - which identifies why the exercise went wrong so the patient can correct themselves. We use these models on our dataset of four common physical therapy exercises labeled by a physical therapist. Our results demonstrate the feasibility of using skeletal data from state-of-the-art 3D human pose estimation models for physical rehabilitation exercise evaluation and guidance. Thus, we enable remote patient monitoring and guidance from a single camera - making it highly cost-effective and scalable.

Introduction

The field of human action evaluation (HAE) is broad in its applications, ranging from gait analysis [30] to judging Olympic performance [25]. Physical rehabilitation is one of many applications which can significantly benefit from using HAE technologies [4]. Using skeleton features in HAE was shown to be promising in [11]; however, these studies are limited to low fidelity exercises, i.e., large amplitude movements.

Using specialized sensors, such as Inertial Measurement Units (IMUs) [20], RGB-D cameras [17] ,[22],[34],[23], or motion capture systems [8] for HAE have shown promising results in displaying accurate assessment and quantification of rehabilitation and strength exercises. While these specialized sensor-based assessment technologies provide high accuracy, they are limited in applicability due to the inherent cost and complex nature of obtaining specialized hardware for action evaluation. Moreover, standardizing skeleton data and deep feature representation methods from the sensors is another key issue in developing reliable quality assessment algorithms for HAE [14]. Thus, relying on specialized hardware limits practical applications of HAE and prevents the creation of standard datasets required to advance the technology.

Recent advances in 3D Human Pose Estimation (HPE) such as [24], [19] have allowed for the feature extraction of skeletal key points from a monocular RGB camera. These 3D HPEs are popularly used to predict key joint positions on the body [13]. In this thesis, we show that classical machine learning methods such as Dynamic Time Warping (DTW) are limited for monocular RGB HAE due to the inherent noise associated with predicting the depth dimension from a monocular RGB camera. Using $2D$ skeletal information in conjunction with deep learning has shown promising results for regression analysis [15]. In HAE for physical therapy (PT), getting 3*D* skeletons is necessary to clinically assess range of motion in key joints. Furthermore, physical rehabilitation requires corrective feedback for improving patient outcomes [10] [12]. Although monocular RGB 3D HPE is promising, we observe no HAE with high-fidelity feedback in the literature.

In our work, we propose a framework for skeleton-based HAE for PT exercises from a single camera, that evaluates patient repetitions as correct or not, and offers explainability for correcting the incorrect movements. For this thesis, we narrow the scope to human Patient Performance Evaluation and guidance, assuming 3D skeletal features are obtained from a 3D human pose estimator.

Chapter 1

Related Work

We review various approaches proposed for general HAE and for PT applications, and highlight the major differences between them and our work. Although specialized sensor-based systems for physical rehabilitation are popular in the literature [18], we observe limited approaches based on a single RGB camera. To our knowledge, there are no 3D skeletal feature, monocular-based approaches. The literature highlights the importance of 3D skeletal features as a constituent to HAE in physical therapy; therefore, our work complements existing work that quantifies physiotherapy metrics such as range-of-motion and joint angle-based success criteria implicitly, by learning from data.

1.1 HAE based on complex sensors

Rooted in classical signal processing [6], matching techniques have seen some success in RGB-D applications [32]. Another classical approach [9] was used as a screening tool by performing anomaly detection on several activities of daily living (ADL). This system allows for health service provider intervention for given neuromusculoskeletal conditions but does not address rehabilitation. Moreover, we observe classical analytical methods are insufficient for datasets with low signal-to-noise ratio, i.e., collected using monocular RGB with skeletal features engineered using state-of-the-art HPE models.

Deep learning approaches have been popular for regression analysis [27] [33] on RGB-D, optical tracking system [17], and IMU [29] datasets. These regression models do not provide any patient feedback. The ability to use deep learning to diagnose and track the progression of Alzheimer's Disease was shown in [34].

1.2 HAE based on a single RGB camera

Outside of physical therapy, approaches based on a single RGB camera for HAE have been shown by [15], which utilized 2D skeletal features extracted from monocular RGB to train deep learning regression models on the UNLV Olympic Diving and MIT Olympic Ice Skating Scoring datasets. Pseudo3D was used to extract spatiotemporal features from video on the UNLV Diving dataset [7]. The feature engineering here [26] is limited to pseudo representations of human pose estimation.

For physical rehabilitation, [28] uses a regression-based quantitative scoring model trained on the KIMORE dataset using monocular RGB. A lack of accurate determination of joint angles from 2D skeleton data limits the extent of evaluation of exercises. For example, indicators based on range of motion of joints, a prevalent metric used in PT, cannot be modeled (even implicitly) by 2D joint data. Feedback on static exercises only was shown in [21]; these comprise a small fraction of rehabilitative exercises.

Chapter 2

Methodology

We begin by defining key terms and notation in Section 2.1. In Section 2.2, we explain our data collection method, while in Section 2.3, we build the case to show that classical methods such as Dynamic Time Warping are not suitable for evaluating repetitions and providing guidance results for data collected using monocular RGB camera videos. We then describe our proposed methods in Sections 2.4 and 2.5.

2.1 Terminology

We focus on HAE using a monocular RGB camera which is available in most smartphones and tablet computers. For simplicity, any data is from a monocular RGB camera unless otherwise specified.

For Patient Performance Evaluation (PPE), we formulate and define the following tasks in our processing and data pipeline.

- **Skeletonization:** Given an RGB frame containing a human, skeletonization refers to the process of extracting the 3D (x,y,z) coordinates of key joint positions and constructing a stick figure-like model of the human. The term **skeleton** refers to the constructed stick figure. This thesis uses a custom skeletonizer that gives results similar to [24].

- **Segmentation:** Given a sequence of skeletons from the video feed, segmentation is the process of extracting exercise repetitions from the entire exercise session. Each exercise

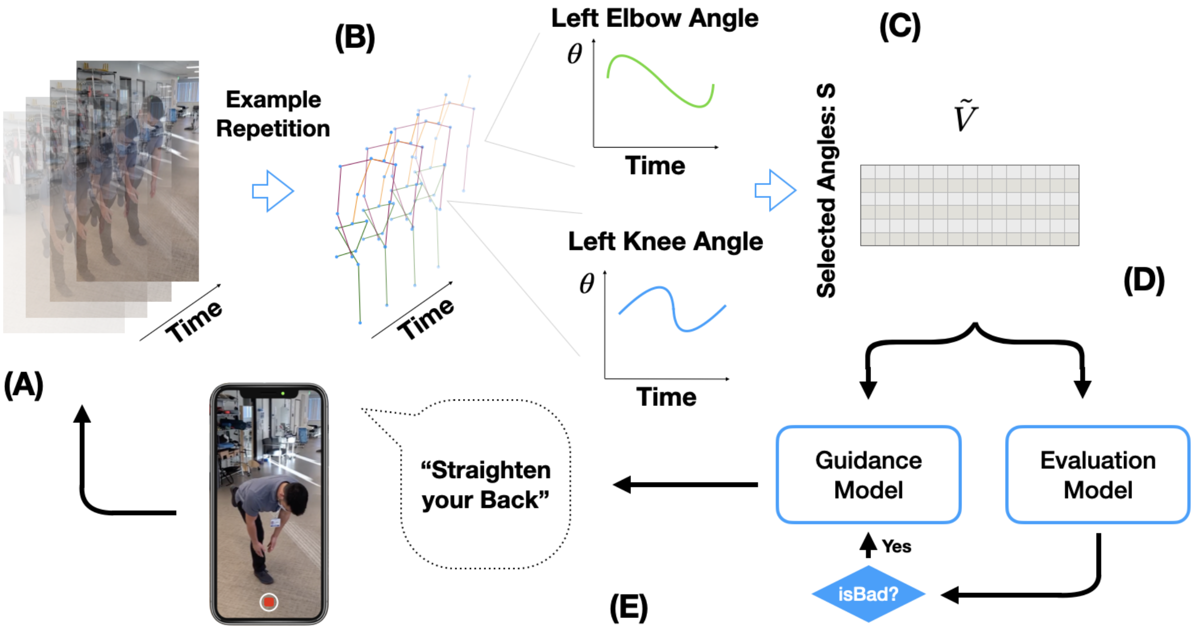

Figure 2.1. Process Flow: (A) Patient records a video performing an exercise, (B) each frame is skeletonized, (C) the skeletons are transformed to an angular domain and post-processed, (D) the resulting vector is fed to PPE and Guidance models, (E) patient receives feedback.

repetition has start and stop frame numbers in the video. We use **r** to denote a repetition.

- **Patient Performance Evaluation (PPE):** Given an exercise repetition, PPE assigns the exercise repetition a label from the set $\{good, bad\}$. Ground truth labels are determined by an expert, i.e., a licensed physical therapist.
- **Guidance:** For a repetition that has been tagged *bad*, guidance is the suggestion to correct the form of that repetition. For example, while doing a deadlift, possible guidance could be ”keep back straight” if the patient has a rounded back.

Fig. 2.1 shows the overall flow of our proposed process. Given the abundance of literature on human pose estimation and segmentation, we narrow our attention to PPE and guidance methods.

2.2 Feature Engineering

We record student volunteers performing exercises on phone cameras. We provide more details on data collection in Section 4.1. We use a custom skeletonization model as described in

Section 2.2.1 , which outputs the 3D positions of 17 key joints as in [24]. The videos are then segmented into individual repetitions, and skeletons are processed as described in Section 2.2.2. The output after the above processing is fed to downstream models.

2.2.1 Skeletonization

Our custom 3D human pose estimation (HPE) pipeline incorporates YOLOv4 [1] for human detection, Cascaded Pyramid Network (CPN) [5] for 2D pose estimation, and EvoSkeleton [16] for lifting the poses from 2D to 3D. This pipeline aims to accurately estimate the three-dimensional poses of humans from input images.

The first step of the pipeline involves using YOLOv4 [1], a state-of-the-art object detection model, to detect and localize humans in the input images or frames of a video. YOLOv4 [1] provides tight bounding box coordinates around each detected human.

Once the humans are detected, the pipeline proceeds to the second step, which employs CPN [5] for 2D pose estimation. CPN is a deep learning-based model that estimates the 2D joint locations (e.g., elbows, knees, etc.) from the detected humans in the bounding boxes from the previos step. It leverages cascaded pyramid networks to iteratively refine the estimated joint positions, resulting in more accurate 2D poses.

The final step of the pipeline utilizes EvoSkeleton [16], a technique for lifting the 2D poses to their corresponding 3D representations. EvoSkeleton [16] employs a learned 3D human pose model to estimate the depth and orientation of each joint. By considering geometric constraints and prior knowledge, EvoSkeleton [16] generates a 3D human pose estimation from the previously obtained 2D joint positions.

The above pipeline outputs the 3D positions of 17 key joints as in [24]. Based on our research, we found the above pipeline to comprise of open source components, and meeting our needs of near real time inference speed. Further analysis on accuracy and time complexity during inference is beyond the scope of the thesis.

2.2.2 Feature Extraction

We begin by selecting joint angles relevant to the exercise E as determined by an expert. We note that the selected joint angles a_j could be those that move during the exercise as well as those that might be required to stay stationary. Let $\mathscr{A}_E = \{a_j\}$ be the set of such selected joint angles for exercise E.

To formalize, for a repetition $\mathbf{r}$ belonging to exercise E of time length T video frames, we skeletonize to obtain a tensor of shape $(T, 17, 3)$, for 17 joints and 3 spatial dimensions. For each frame $i \in [1, .., T]$, we compute the angles for the joints in $\mathscr{A}_E$. Let $v_{(j,i)}$ be the angle for joint a_j at time i. The 2D tensor $\mathbf{V} = [[v_{(j,i)}]]$ of dimensions $(|\mathscr{A}_E|, T)$ is then smoothed and subsampled as described in the paragraph below to get $\tilde{\mathbf{V}}$. This $\tilde{\mathbf{V}}$ is used as an input to the models described in Sections 2.4.1 and 2.4.2.

Smoothing filter: We use an averaging filter with window size 5 on the time series of individual joint angles to smooth out outliers.

Subsampling: As different subjects perform exercise repetitions at different rates, the $\mathbf{V}$ vary in width. This can cause problems in training deep neural networks including difficulties in batch training and data preprocessing. Based on our discussions with the physical therapist, the exercise rate is seldom useful in classifying a repetition as *good* or *bad*. More importance is given to the form, which does not depend on the rate. Hence, we subsample each repetition to 20 equidistant frames.

2.3 Classical Methods: Dynamic Time Warping

Dynamic Time Warping (DTW) is a method for measuring similarity between two temporal sequences. It was used successfully in the context of physical therapy exercises in [32] where data was collected with a Microsoft Kinect Camera. In our experiments, DTW could be applied to patient repetitions to compare them with a ”ground truth” coming from our physical therapist consultant. In concept, it could allow the cumulative difference across all key

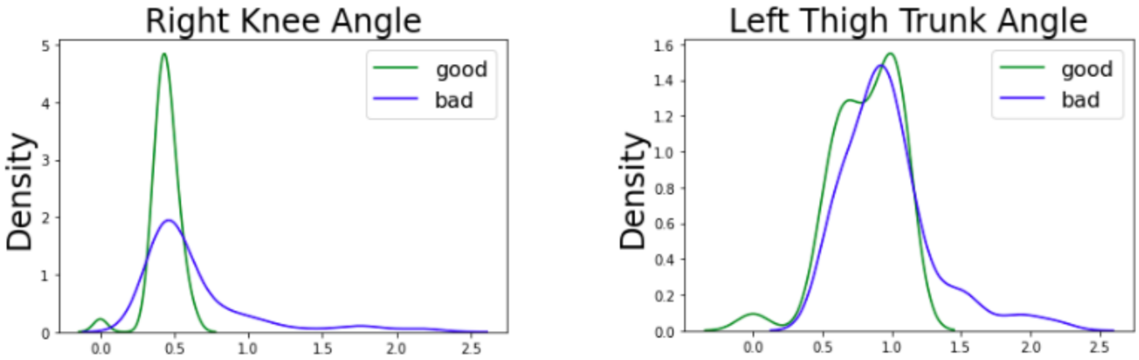

Figure 2.2. The distributions of similarity scores produced by DTW on good (green) and bad (blue) repetitions are highly overlapping.

angles to be used as a general threshold for PPE while using angle-specific thresholds to provide feedback on any particularly incorrect angles. Despite its efficacy with RGB-D based datasets, we observed poor results on our 3D HPE skeletal data. Fig. 2.2 plots the distribution of similarity scores per joint angle from DTW for *good* and *bad* repetitions; the substantial overlap between the two distributions makes the thresholding-based approach infeasible. Thus, we turn to deep learning-based methods that are more potent in finding fine patterns in the data.

2.4 Patient Performance Evaluation (PPE)

In this section, we describe our neural network-based approach to PPE. We first describe the data processing steps and then the model architectures.

2.4.1 CNN-based PPE

In this section, we describe the convolutional neural network (CNN) architecture used for classifying $\tilde{\mathbf{V}}$ to the PPE label set $Y = \{good, bad\}$. In the first layer, we apply a convolution filter on the time axis (temporal convolution), i.e., a kernel of shape $(3,1)$ and output channels 32. Then the next layer is convolution on the angles (spatial convolution), i.e., a kernel of shape $(1, |\mathscr{A}_E|)$ and output channels 64. After each of these two convolutions, a ReLU non-linearity is followed by a batch norm. Finally, the feature map from the last convolution is average-pooled,

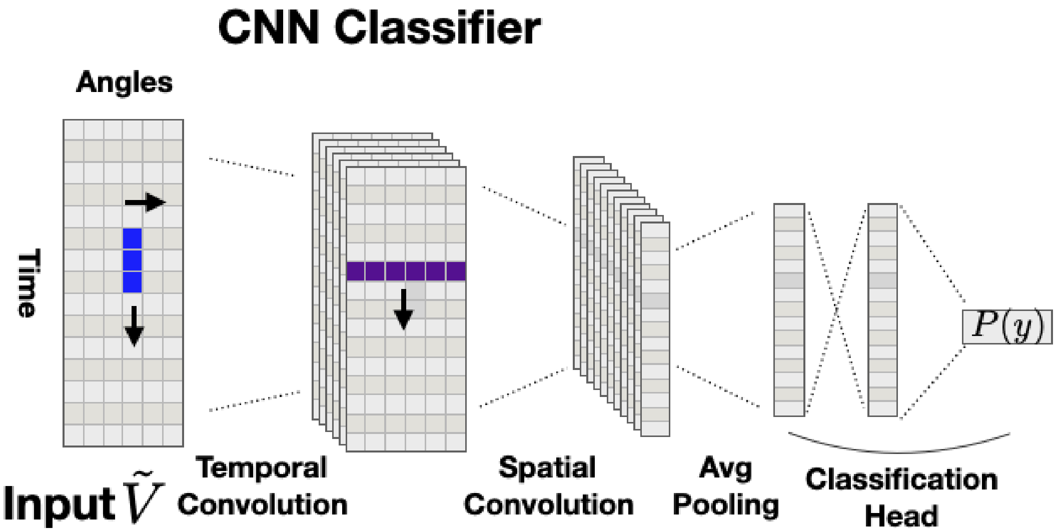

Figure 2.3. CNN based PPE: Temporal Convolution followed by Spatial Convolution, Avg Pooling and Classification Head

followed by a binary classification head. The intuition for choosing the above architecture is that the temporal convolution learns the temporal context at time t_i of the angle a_j. Once the context of the angle has been set, the model learns the interdependence relations among the joint angles. To the best of our knowledge, ours is the first method to view human Patient Performance Evaluation as a Spatio-Temporal convolution network.

2.4.2 Attention based PPE

In this section, we describe the architecture of the attention-based transformer classifier used for PPE. Attention networks have gained popularity by learning to attend to values at different timestamps and have been successful in natural language processing and computer vision [31]. The number of heads is chosen to be 2, with a 512 embedding dimension for the encoder. A classification head follows the encoding head.

2.5 Guidance and Exercise Feedback

This section proposes our method for generating guidance and feedback mechanisms for the *bad* repetitions. An autoencoder is a model that encodes the input into a feature map and

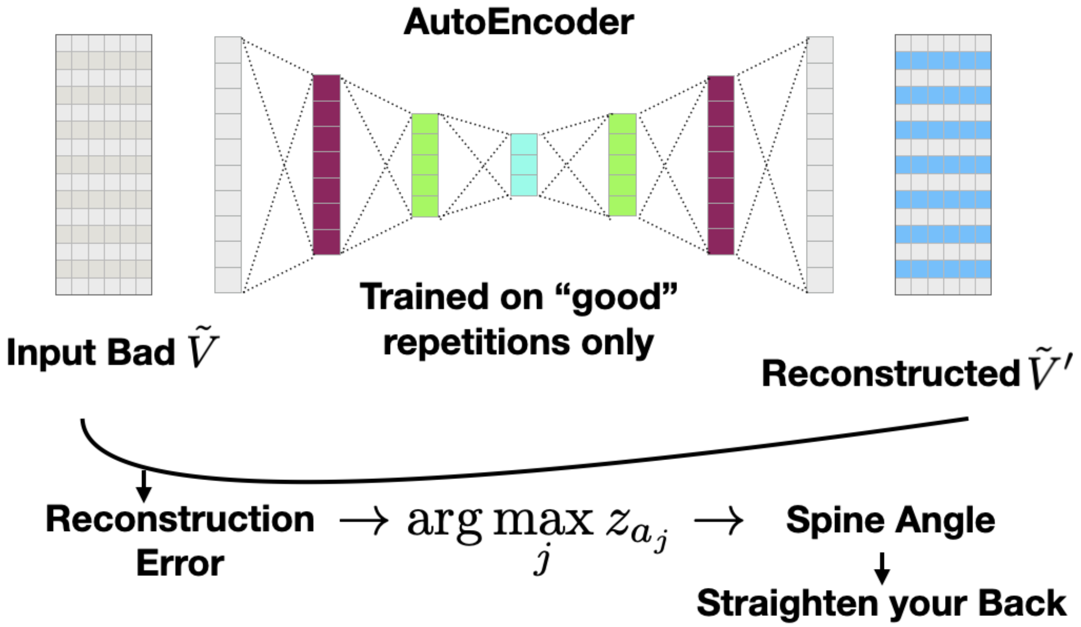
AutoEncoder
Trained on “good” repetitions only
Input Bad $\tilde{V}$
Reconstructed $\tilde{V}'$
Reconstruction Error
$\rightarrow \arg\max_{j} z_{a_j} \rightarrow$
Spine Angle
Straighten your Back

Figure 2.4. Guidance Process: Autoencoder trained on good reps learns the intrinsic features of exercises. Reconstruction error is used to get feedback cues.

then uses the encoding to decode it back to the input. It is used to train an underlying model distribution of the data. With this context, we first train an autoencoder using only the *good* repetitions to make the neural network learn a robust *good* template of the exercise. The mean and variance $(\mu_{a_j}, \sigma_{a_j})$ of the reconstruction error for each joint angle are collected. Then for a given *bad* repetition, the reconstruction error e_{a_j} from the autoencoder is computed. The angles with the highest z-score: $z_{a_j} = \frac{e_{a_j} - \mu_{a_j}}{\sigma_{a_j}}$ are selected as the ones responsible for the repetition being *bad*. The process is shown visually in Fig. 2.4.

AutoEncoder:

Following similar arguments on the efficacy of attention-based models on time series data as described in Section 2.4, we propose to use an attention based autoencoder. While implementing a corresponding CNN-based auto-encoder, we found maintaining the convolutional upsampling layers to consistently output the same size vector as the input for different exercises non-trivial, and adding unnecessary complexity to our model pipeline. In the attention-based autoencoder model, we use the default number of dimensions for the embeddings, with the number of heads equal to $|\mathscr{A}_E|$, and mean square error (MSE) as the loss function.

Chapter 3

Experiments

3.1 Data

In this section, we discuss the details of our dataset, which we use to demonstrate the efficacy of our proposed PPE and guidance models. As existing datasets [17] [3] [2] do not have corrective feedback and classification scores, we were required to create our own.

3.1.1 General Setup

Based on consultation with a physical therapist, we selected four exercises that are frequently prescribed in physical therapy, involve compound movements, and collectively cover multiple focus areas (shoulder, legs, hips). The exercises are: double leg Romanian dead-lift (DoubleRDL), single leg Romanian dead-lift (SingleRDL), single leg mini squat (SingleMS), and rotator cuff (RotatorCuff).

The exercises selected are commonly prescribed by physical therapists to target different muscle groups and improve overall strength and stability. The double leg Romanian dead-lift (DoubleRDL) is a bilateral exercise that primarily targets the posterior chain, including the hamstrings, glutes, and lower back. It involves hinging at the hips while keeping the back straight and lowering the weight down towards the ground. The single leg Romanian dead-lift (SingleRDL) is a unilateral variation of the exercise, where the focus shifts to improving balance, stability, and targeting each leg individually. It helps strengthen the same muscle groups as the

DoubleRDL but with an added emphasis on stability and symmetry. The single leg mini squat (SingleMS) is another unilateral exercise that targets the quadriceps, hamstrings, and glutes. It involves squatting on one leg while keeping the other leg elevated, helping to enhance stability, balance, and strength. Lastly, the rotator cuff (RotatorCuff) exercises specifically target the muscles responsible for shoulder stabilization and mobility. These exercises usually involve the use of resistance bands or light weights and aim to strengthen the rotator cuff muscles, which can help prevent shoulder injuries and improve overall shoulder function.

Students working on the project volunteered to be recorded performing exercises. A total of 10 subjects were selected to partake in the study, with ages ranging from 21 to 27 years. The participant pool consisted of 8 males and 2 females. Before recording each exercise, the subjects were shown a demonstration video of approximately one minute, created by the physical therapist that showed several repetitions of the exercise with proper form, with instructions to do so. The subjects were asked to recreate 10 repetitions seen in the video to the best of their ability. For unilateral exercises, subjects were asked to perform 5 repetitions on each side. After the first 10 repetitions, subjects were informed about common mistakes seen in physical rehabilitation. For example, a DoubleRDL is often performed with the subject's spine being too rounded with too much scapular protraction or the subject's knees being too bent or locked out. These subtle mistakes have a profound impact on muscle activation during the exercise. The subjects were instructed to incorporate these incorrect postures into an additional 10 repetitions. The duration of the recordings ranged from 2 to 3.5 minutes.

To test our proposed methods' robustness to different viewing angles and camera types, we recorded subjects using five different smartphones (iPhone Models X, 11, 7, 8, and Google Pixel 3A) placed on tripods approximately 4 feet high and at five different angles (30, 60, 90, 120, and 150 degrees) to the subject. To summarize, data consists of 1000 repetitions for each exercise from 10 subjects, and 5 camera angles.

Table 3.1. Results: F1 scores of the PPE models, and top-2 accuracy of the Guidance models. The Baseline used is the DTW method proposed in [32].

Exercise	Evaluation (F1 Scores)			Guidance (top-2)	
	Baseline	Ours		Baseline	Ours
	DTW	CNN	Attention	DTW	Attention
DoubleRDL	0.197	**0.262**	0.255	0.846	**0.884**
SingleRDL	0.052	0.222	**0.274**	0.652	**0.679**
SingleMS	0.435	**0.544**	0.519	0.805	**0.858**
RotatorCuff	0.569	**0.78**	0.769	0.652	**0.758**

3.1.2 Labeling

For each video recording, each exercise was segmented manually into repetitions which were classified as being *good* or *bad*, and the two most erroneous movements that needed correction were noted. Our final dataset includes 19% *good* repetitions for DoubleRDL, 15% for SingleRDL, 42% for SingleMS, and 50% for RotatorCuff. This variation is due to the fact that subjects were more knowledgeable about proper exercise form for some exercises compared to others.

3.2 Results

We split the 10 subjects into 3 folds (with 4, 3, and 3 subjects), and train the model on 2 of the folds, and evaluate the model on the remaining fold. We show the mean F1 score of the 3-fold cross validation.

Patient Performance Evaluation For PPE, we use negative log likelihood loss for training. We use the *Adam* optimizer for training with learning rate $1e-4$ and default weight parameters $\beta = (0.9, 0.98)$, and batch size 16. We report the F1 score of the binary classification model. We see that both the CNN and Attention-based models consistently outperform DTW (see Table 3.1). Between the CNN and Attention-based models, the Attention-based model performs slightly better than the CNN-based model.

Our results show promise that our novel solutions can classify exercises based on single camera input.

Table 3.2. Guidance cues defined by PT.

Exercise	Guidance Criteria
DoubleRDL	Knees Too Bent, Knees Locked, Back Too Round, Feet Too Far Apart
SingleRDL	Knees Too Bent, Knee Locked, Back Too Round, Leg Not In-Line
SingleMS	Hips Not Level, Squat Too Low, Twisting Torso
RotatorCuff	Twisting Torso, Arm Too Extended, Lifting Arm Too High/Low

Guidance We map each joint angle to an action item that the subject could do to correct that angle based on the output of the autoencoder. For guidance models, we report the $top-2$ accuracy. Specifically, it is considered correct if either of the two most erroneous predicted angles match the PT judgment. Human body movements have constraints and the joint angles do not operate in isolation from one another. Correcting one angle could affect the correctness of the other angle - for example, in DoubleRDL, making the back straight would still be conducive to correcting locked knees. Hence, the top-2 accuracy metric is justified.

For training the Attention autoencoder, we use the Adam optimizer with learning rate $1e-3$ and default weight parameters $\beta = (0.9, 0.98)$ with batch size 32. In Table 3.1, we see that the Attention-based autoencoder method outperforms DTW. Ours is the first deep learning-based approach to PT guidance.

Chapter 4

Conclusion and Future Work

Our proposal entails the development of deep-learning models for Patient Performance Evaluation (PPE) and guidance in physical therapy rehabilitation, utilizing only a smartphone camera. By employing Convolutional Neural Networks (CNN) and Attention networks, our approach demonstrates enhanced results across all four exercises. Notably, we emphasize the significance of a robust dataset, which emerges as a key priority for future endeavors. Conversely, Dynamic Time Warping (DTW) does not exhibit similar benefits from such a dataset. To the best of our knowledge, our work represents the first deep learning-based approach for PT guidance. Moving forward, it is recommended to explore additional deep learning architectures, specifically autoencoders, within the guidance models. Furthermore, our future plans encompass conducting clinical trials to evaluate the efficacy of our method on patient outcomes. Additionally, we aim to expand our exercise library to encompass a wider range of exercises, thereby broadening the scope and applicability of our approach.

Through an extensive industry review, it has been observed that the field of physical therapy encompasses diverse schools of thought, leading to variations in exercise forms and evaluation methods. Furthermore, patients arriving from different medical backgrounds introduce further complexity, as the criteria for evaluating their progress may differ. Deep learning models thrive on standardized data to generate generalized predictions across individuals, whereas the industry demands a personalized approach. Hence, reconciling the need for standardization with the requirement for individualization poses a significant challenge in leveraging deep learning models effectively within the physical therapy domain. By carefully considering these factors, future work can pave the way for effective and commercially valuable solutions that balance standardization with the indispensable aspect of personalized care in physical therapy.

ACKNOWLEDGEMENTS

I would like to express my heartfelt gratitude to all those who have contributed to the completion of this thesis.

First and foremost, I am deeply grateful to my supervisors, Dr Sujit Dey and Dr Pamela Cosman, for their unwavering guidance, support, and invaluable insights throughout this research journey. Their expertise, patience, and encouragement have been instrumental in shaping this work and refining my understanding of the subject matter.

I would also like to acknowledge Alexander Postlmayr, without whom my research would have no doubt taken fives times as long. It is his support that helped me in an immeasureable way.

The manuscript, in full, is a reprint of the material as it appears in:
B Garg, A Postlmayr, P Cosman, S Dey. ”Short: Deep Learning Approach to Skeletal Performance Evaluation of Physical Therapy Exercises”, IEEE International Conference on Connected Health: Applications, Systems and Engineering Technologies - 2023.

The thesis author was the primary investigator and author of this paper.

Bibliography

[1] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *CoRR*, abs/2004.10934, 2020.

[2] Massimo Camplani, Adeline Paiement, L Tao, Sion Hannuna, Dima Damen (Aldamen), Majid Mirmehdi, and Tilo Burghardt. Depth video and skeleton of people walking up stairs, 2014.

[3] Marianna Capecci, Maria Gabriella Ceravolo, Francesco Ferracuti, Sabrina Iarlori, Andrea Monteriu, Luca Romeo, and Federica Verdini. The KIMORE Dataset: KInematic Assessment of MOvement and Clinical Scores for Remote Monitoring of Physical REhabilitation. *IEEE transactions on neural systems and rehabilitation engineering: a publication of the IEEE Engineering in Medicine and Biology Society*, 27(7):1436–1448, July 2019.

[4] Oya Celiktutan, Ceyhun Burak Akgul, Christian Wolf, and Bülent Sankur. Graph-Based Analysis of Physical Exercise Actions. In *Proceedings of the 1st ACM International Workshop on Multimedia Indexing and Information Retrieval for Healthcare*, MIIRH '13, pages 23–32, New York, NY, USA, 2013. Association for Computing Machinery. event-place: Barcelona, Spain.

[5] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. *CoRR*, abs/1711.07319, 2017.

[6] Yinpeng Chen, Margaret Duff, Nicole Lehrer, Hari Sundaram, Jiping He, Steven Wolf, and Thanassis Rikakis. A computational framework for quantitative evaluation of movement during rehabilitation. *AIP Conference Proceedings*, 1371, 06 2011.

[7] Li-Jia Dong, Hong-Bo Zhang, Qinghongya Shi, Qing Lei, Ji-Xiang Du, and Shangce Gao. Learning and fusing multiple hidden substages for action quality assessment. *Knowledge-Based Systems*, 229:107388, 2021.

[8] Chen Du, Sarah Graham, Colin Depp, and Truong Nguyen. Assessing Physical Rehabilitation Exercises using Graph Convolutional Network with Self-supervised regularization. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 281–285, 2021.

[9] Amr Elkholy, Mohamed E. Hussein, Walid Gomaa, Dima Damen, and Emmanuel Saba. Efficient and robust skeleton-based quality assessment and abnormality detection in human action performance. *IEEE Journal of Biomedical and Health Informatics*, 24(1):280–291, 2020.

[10] Marc P. Gruner, Nathan Hogaboom, Ike Hasley, Jared Hoffman, Karina Gonzalez-Carta, Andrea L. Cheville, Zhuo Li, and Jacob L. Sellon. Prospective, Single-blind, Randomized Controlled Trial to Evaluate the Effectiveness of a Digital Exercise Therapy Application Compared With Conventional Physical Therapy for the Treatment of Nonoperative Knee Conditions. *Archives of Rehabilitation Research and Clinical Translation*, 3(4):100151, 2021.

[11] Reza Haghighi Osgouei, David Soulsby, and Fernando Bello. Rehabilitation Exergames: Use of Motion Sensing and Machine Learning to Quantify Exercise Performance in Healthy Volunteers. *JMIR Rehabil Assist Technol*, 7(2):e17289, August 2020.

[12] Stephanie Hewitt, Ruth Sephton, and Gillian Yeowell. The Effectiveness of Digital Health Interventions in the Management of Musculoskeletal Conditions: Systematic Literature Review. *Journal of medical Internet research*, 22:e15617, June 2020.

[13] 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.

[14] Qing Lei, Ji-Xiang Du, Hong-Bo Zhang, Shuang Ye, and Duan-Sheng Chen. A survey of Vision-Based human action evaluation methods. *Sensors (Basel)*, 19(19), September 2019.

[15] Qing Lei, Hong-Bo Zhang, Ji-Xiang Du, Tsung-Chih Hsiao, and Chih-Cheng Chen. Learning Effective Skeletal Representations on RGB Video for Fine-Grained Human Action Quality Assessment. *Electronics*, 9(4), 2020.

[16] Shichao Li, Lei Ke, Kevin Pratama, Yu-Wing Tai, Chi-Keung Tang, and Kwang-Ting Cheng. Cascaded deep monocular 3d human pose estimation with evolutionary training data. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[17] Yalin Liao, Aleksandar Vakanski, and Min Xian. A Deep Learning Framework for Assessing Physical Rehabilitation Exercises. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 28(2):468–477, 2020.

[18] Yalin Liao, Aleksandar Vakanski, Min Xian, David Paul, and Russell Baker. A review of computational approaches for evaluation of rehabilitation exercises. *Computers in Biology and Medicine*, 119:103687, 2020.

[19] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. VNect:

Real-time 3D Human Pose Estimation with a Single RGB Camera. In *ACM Transactions on Graphics*, volume 36, July 2017. Issue: 4.

[20] Slobodan Milanko and Shubham Jain. Liftright: Quantifying strength training performance using a wearable sensor. *Smart Health*, 16:100115, 2020.

[21] Cristian Militaru, Maria-Denisa Militaru, and Kuderna-Iulian Benta. Physical exercise form correction using neural networks. In *Companion Publication of the 2020 International Conference on Multimodal Interaction*, ICMI '20 Companion, page 240–244, New York, NY, USA, 2020. Association for Computing Machinery.

[22] Ferda Ofli, Gregorij Kurillo, Štěpán Obdržálek, Ruzena Bajcsy, Holly Brugge Jimison, and Misha Pavel. Design and evaluation of an interactive exercise coaching system for older adults: Lessons learned. *IEEE J Biomed Health Inform*, 20(1):201–212, January 2015.

[23] Paritosh Parmar and Brendan Tran Morris. Measuring the quality of exercises. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 2241–2244, 2016.

[24] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. pages 7745–7754, 06 2019.

[25] Hamed Pirsiavash, Carl Vondrick, and Antonio Torralba. Assessing the quality of actions. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 556–571, Cham, 2014. Springer International Publishing.

[26] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *IEEE International Conference on Computer Vision (ICCV)*, October 2017.

[27] Noureddin Sadawi, Alina Miron, Waidah Ismail, Hafez Hussain, and Crina Grosan. Gesture correctness estimation with deep neural networks and rough path descriptors. In *2019 International Conference on Data Mining Workshops (ICDMW)*, pages 595–602, 2019.

[28] Faegheh Sardari, Adeline Paiement, Sion Hannuna, and Majid Mirmehdi. Vi-net—view-invariant quality of human movement assessment. *Sensors*, 20(18), 2020.

[29] Dapeng Tang. Hybridized hierarchical deep convolutional neural network for sports rehabilitation exercises. *IEEE Access*, 8:118969–118977, 2020.

[30] Lili Tao, Adeline Paiement, Dima Damen, Majid Mirmehdi, Sion Hannuna, Massimo Camplani, Tilo Burghardt, and Ian Craddock. A comparative study of pose representation and dynamics modelling for online motion quality assessment. *Computer Vision and Image Understanding*, 148:136–152, 2016.

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.

[32] Wenchuan Wei, Yao Lu, Catherine D. Printz, and Sujit Dey. Motion data alignment and real-time guidance in cloud-based virtual training system. In *Proceedings of the Conference on Wireless Health*, WH '15, New York, NY, USA, 2015. Association for Computing Machinery.

[33] Bruce X. B. Yu, Yan Liu, and Keith C. C. Chan. Skeleton-based detection of abnormalities in human actions using graph convolutional networks. In *2020 Second International Conference on Transdisciplinary AI (TransAI)*, pages 131–137, 2020.

[34] Bruce X. B. Yu, Yan Liu, Keith C. C. Chan, Qintai Yang, and Xiaoying Wang. Skeleton-based human action evaluation using graph convolutional network for monitoring Alzheimer's progression. *Pattern Recognition*, 119:108095, 2021.