Skeletons distance evaluation for automated control of exercises in physical therapy

Parama Bagchi[1], Oleg S. Seredin[2], Andrey V. Kopylov[2], Egor E. Surkov[2], Nikita S. Mityugov[2], Alexei R. Tokarev[2], Debotosh Bhattacharjee[3]

[1] Department of CSE, RCC Institute of Information Technology, Beliaghata, Kolkata 700015, India – paramabagchi@gmail.com
[2] Tula State University, Lenin Ave. 92, 300012, Tula, Russia – oseredin@yandex.ru
[3] Department of CSE, Jadavpur University, Kolkata 700032, India – debotosh@ieee.org.

Keywords: Rehabilitation Assessment, Skeletal models, Motion Analysis, Weighted Euclidean distance

Abstract

The paper focuses on developing a method for evaluating the similarity between skeletal models of an instructor and a patient during physical therapy exercises. Unlike general-purpose measures, our approach considers the specific characteristics of therapeutic exercises, including initial positioning, predominant movements, and exercise pace. By incorporating the concept of informativeness for skeletal points and refining normalization techniques, we improve the accuracy of pairwise dissimilarity measures. Experimental results demonstrate improved separability of records based on the accuracy of exercise repetition, highlighting the potential of this approach for enhancing automated physical rehabilitation systems.

1. Introduction

Physical rehabilitation is a cornerstone of medical recovery, playing a pivotal role in restoring physical function, strength, and mobility of patients after serious illnesses such as trauma, heart attacks and strokes. However, methods of physical rehabilitation face significant challenges, including resource limitations, complex patient demands, and the need for continuous, personalized monitoring.

Recent advancements in computer vision and artificial intelligence offer promising solutions to these challenges. By leveraging depth sensors and intellectual algorithms, it is now possible to develop automated systems of physical rehabilitation that can objectively monitor, analyze, and provide real-time feedback on patient movements during exercise therapy.

Although numerous patient-centered systems have been developed for home rehabilitation, there is a notable lack of systems designed to support both the physiotherapist and the patient (Lam et al., 2016).

In modern physical rehabilitation programs, patients usually execute exercises while receiving intermittent feedback or guidance after the physiotherapist (instructor) demonstrates the movements. The structure of these exercises enables the synchronization of the movement phases between the physiotherapist and the patient, allowing for the subjective evaluation of pose similarity in corresponding frames (Fig. 1). Thus, the similarity measure plays a critical role in forming the overall evaluation of the exercise and in turn in the final assessment of the effectiveness of patient rehabilitation.

We use a skeletal description of a human pose to form the basis for measuring such compliance, because such an approach can greatly reduce the amount of personal data collected, allowing focus on movement analysis rather than identity and providing technical advantages in terms of efficiency and accuracy (Seredin et al., 2023).

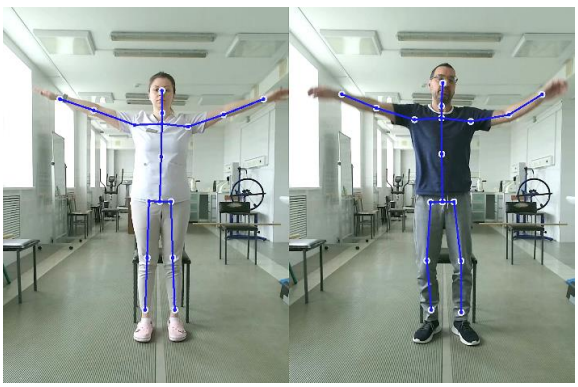


Figure 1. Frames from video records of the therapist and the patient with Microsoft Kinect v2 skeletal models

Skeleton-based methods for measuring the similarity or dissimilarity between human figures, postures, and actions can be categorized into four main groups. The first group focuses on the 3D positions of skeleton vertices, which correspond to joints, using pairwise relative positions or covariance matrices to describe poses, though these are insufficient for accurate activity detection like fall detection (Hussein et al., 2013; Wang et al., 2012; Yan et al., 2018). The second group considers general geometric characteristics of the skeleton, such as bounding rectangles, geometric moments, and distances from specific points like the head or center of mass of the human body, offering robustness to estimation errors but limited flexibility in complex environments (Chen et al., 2011; Zhang et al., 2017). The third group emphasizes the relationship between the skeleton and human body parts (Bian et al., 2015), recognizing that movements are more effectively observed through the shapes, lengths, and locations of bones rather than joints (Zhang et al., 2020), and may also consider the relationships between neighboring body parts (Du et al., 2015). Finally, the fourth group adopts a featureless approach to pattern recognition, representing objects through pairwise similarity measures or differences, with

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLVIII-2-W9-2025-33-2025 | © Author(s) 2025. CC BY 4.0 License.

distance metric learning being a further development of this approach (Chen et al., 2018; Kaya and Bilge, 2019).

The exercises in physical rehabilitation have their own specific movement patterns, which include the initial positioning of the patient, the predominant types of movements used, and the pace of the exercises. The similarity measure proposed in this work, unlike general-purpose skeleton-based measures, allows us to take these characteristics into account to improve the accuracy and reliability of assessing the patient's rehabilitation progress.

The primary contribution of this work is the introduction of the concept of informativeness of skeletal points during the execution of physiotherapy exercises. The incorporation of informativeness enabled us to adjust the pairwise dissimilarity function between skeletal models, considering the specifics of each particular exercise and varying attention to movements of certain directions.

Additionally, we propose a more robust normalization mechanism for skeletal models aimed at aligning the coordinate systems of different sensors and accounting for the peculiarities of software libraries used for generating skeletal pose descriptions.

The experimental results demonstrate an improvement in the efficiency of distance estimation. This enhancement is attributed to the refined dissimilarity function and the robust normalization mechanism, which together ensure more accurate and consistent comparisons of skeletal models across different exercises and sensor systems. The proposed approach thus contributes to more reliable assessment and feedback in physiotherapy exercise analysis.

2. The problem of measuring dissimilarities between skeleton descriptions in physical rehabilitation

Basic rehabilitation programs usually include the predefined set of physical exercises with specially designed movement patterns. Such exercises are rather formal and consist of simple cyclic movements of the body parts. The physiotherapist provides patients with a demonstration of the exercise. The patient's task is to repeat movements after the physiotherapist as accurately as possible. Besides the therapeutic effect, the ability of the patient to mimic certain movement patterns characterizes the effectiveness of patient rehabilitation (Lam et al., 2016).

In most cases assessment of the patient's progress is performed by the physician based on subjective impression and experience. The development of an automated system of physical rehabilitation control let to obtain more unbiased and accurate results for comparison of patients and physiotherapist movements and provides the precise daily profile of patient’s activity.

A predefined sequence of actions, repetition of exercises after a physiotherapist either live or using a pre-recorded video, as well as the primarily cyclical nature of these exercises, allows for the identification and comparison of similar phases within the exercises. In this context, a critical aspect in both the temporal alignment of movement sequences recorded during therapy sessions, and in determining the differences between poses of the instructor and the patient, is provided by a two-argument function. This function takes descriptions of the compared poses as its arguments and outputs a measure of their difference. A method for constructing such a function is proposed in this work.

The skeleton-based model of a human figure, adopted here, allows us to reduce the problem of measuring the dissimilarity between physiotherapist and patient execution of an exercise to the problem of evaluation of distances between two skeleton sequences, recorded during the therapeutic process. We do not address here the process of skeleton model construction, it can be obtained by the special sensors like Microsoft Kinect or Intel RealSence or by using a special neural network-based solutions that build skeleton models from images produced by a conventional RGB camera like YOLO11-pose, Alphapose (Fang et al., 2023), Google MediaPipe Pose Landmarker and others.

Regardless of the specific source used to obtain the skeletal pose model, all skeletal models can be standardized through the methods proposed in (Surkov et al., 2024). These methods involve a series of transformations and normalizations that ensure consistency across different models, enabling direct comparison and integration of data. Thus, we will assume that the skeletal model is represented in the standard form shown in Figure 2.

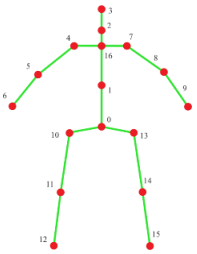


Figure 2. Standard skeletal model with point numbers.

Direct use of Euclidean distances between corresponding points of skeletons can lead to inaccurate or incorrect results due to several key factors. First, skeletal representations often involve landmarks that may vary significantly in their spatial configuration depending on the body proportions, or perspective distortions caused by camera angles. Second, Euclidean distances are sensitive to global transformations such as rotation, and scaling. This lack of invariance makes direct distance measures unreliable for comparing poses across diverse scenarios. Finally, human motion involves complex articulations and hierarchical dependencies between joints, which cannot be adequately captured by simple point-to-point distance metrics. A more robust approach requires incorporating domain-specific knowledge, to ensure accurate and meaningful comparisons. Thus, relying solely on Euclidean distances without considering these factors can lead to misleading conclusions in pose analysis and movement comparison tasks.

Besides the commonly used normalizations, the main idea proposed here, is that points of the skeleton that are more mobile during the exercise contribute more significantly to the similarity assessment, whereas stationary points may be disregarded since the differences in these points are primarily due to anatomical variations between the instructor and the patient and have minimal impact on the evaluation of exercise performance quality. The instructor's recording can naturally serve as a basis for determining the mobility of the skeletal model points during the exercise.

3. The proposed method of skeletons distance evaluation

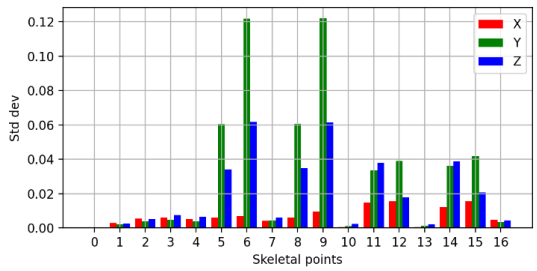
To address the above-mentioned problems, the estimation of the dissimilarity measure between skeletal models is performed in several key stages.

1. To ensure independence from the type of cameras used, the skeletal models obtained from the sensor are standardized (Fig. 2) using transformations proposed by the authors in (Surkov et al., 2024).
2. Normalization for height and preliminary translation to match base points (point 0 in Fig. 2) are carried out (Seredin et al., 2023).
3. The weights of importance for skeleton points are evaluated based on the standard deviations, calculated using the instructor record.
4. The transformation matrix is estimated based on four points that are least susceptible to relative changes during movement.
5. A correspondent transformation is applied to patient skeletons.
6. The weighted average Euclidean distance between the corresponding points of two skeletons is then computed.

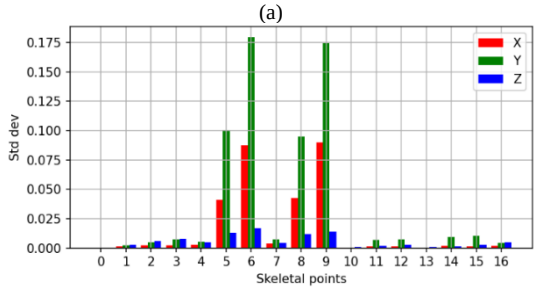
The first stage is carried out according to (Surkov et al., 2024).

On the second stage the human height estimation calculated (Seredin et al., 2023) as the geodesic distance between points 3 and 15 and between points 3 and 19 (Fig. 2). The value of height is averaged between the ten largest values obtained after a certain time of a person's staying in the sensor's field of view. Then, the coordinates of the points in the compared skeletal models are shifted such that the reference point numbered 0 is positioned at the origin of the coordinate system.

Using the recording of the instructor's exercises as a reference, it is possible to assess the degree of mobility of the skeleton points by calculating their standard deviation during the execution of movements. The greater the value of the standard deviation, the larger the contribution that a particular skeleton point should make to the final dissimilarity score (Fig. 3).



(a)



(b)

Figure 3. Standard deviations for coordinates of instructor skeletal model points: (a) – hand movement in sagittal plane (standing position) (b) – hand movement in frontal plane (sitting position

Thus, the standard deviation values of the instructor's skeleton points characterize the informativeness or attention given to each specific point.

Let $S_{k,i}^{m}$, $m \in \{x, y, z\}$, $k = 1, \ldots, K$, $i = 0, \ldots, 16$, where m represents a coordinate axis, k is the frame number and i is the number of skeleton point, be a set of skeletal models, estimated from the therapist or instructor video sequence. Each skeleton model corresponds to the k-th frame, consists of 17 points and each point has three coordinates. All skeletons are normalized as described in the first two stages.

The standard deviation of skeleton points could be computed in the following way:

$$\sigma_i^m = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\left(S_{k,i}^{m} - \bar{S}_i^m\right)^2}\ ,\ m \in \{x, y, z\}\ ,\ i = 0, \ldots, 16\ , \quad (1)$$

where $\bar{S}_i^m$ is the mean value of the i-th point on the coordinate axis m for the entire video.

The final weights are normalized as:

$$w_i^m = \frac{\sigma_i^m}{\sum_{m \in \{x,y,z\}} \sum_{i=0,\ldots,16} \sigma_i^m}, \quad (2)$$

Note that the weight calculation is performed only once for each exercise.

While normalizing skeletal models by height and translating the base point to the origin reduces some variability, it remains insufficient for accurately measuring dissimilarity due to systematic errors arising from differences in pose orientation, body proportions, and sensor-specific biases (Fig. 4). To overcome these limitations, advanced normalization techniques are applied in stage four to ensure robust and meaningful dissimilarity measurements.

We applied here the simplified version of Kabsch algorithm (Kabsch, 1976) to compute the optimal rotation and scaling matrix $\mathbf{R}$ that aligns two sets of 3D points, $\mathbf{x}$ (source) and $\mathbf{y}$ (transformed), under the assumption that translation has already been applied and the centroids of both point sets are at the origin. Such assumption based on the previous normalization stage with special selection of points, and let us to slightly reduse the computetional cost. The Kabsch algorithm minimizes the root-mean-square deviation (RMSD) between corresponding points in the two sets by solving the orthogonal Procrustes problem:

$$\min_{\mathbf{R}} \| \mathbf{Y} - \mathbf{X}\mathbf{R} \|_F^2,$$

where $\mathbf{X} \in \mathbb{R}^{n \times 3}$ and $\mathbf{Y} \in \mathbb{R}^{n \times 3}$ are matrices formed by stacking the coordinates of the source and transformed points, respectively, and $\mathbf{R} \in \mathbb{R}^{3 \times 3}$ is the rotation and scaling matrix to be determined.

The algorithm begins by computing the covariance matrix $\mathbf{H}$, which captures the correlation structure between the two point sets:

$$\mathbf{H} = \frac{1}{n}\mathbf{X}^{\top}\mathbf{Y}.$$

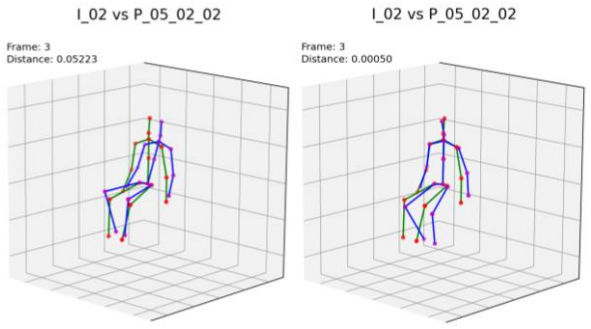
This step assumes that the data is already centered, meaning the centroids of both point sets are zero. The covariance matrix is then decomposed using singular value decomposition (SVD):

$$\mathbf{H} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\top},$$

where $\mathbf{U}$ and $\mathbf{V}$ are orthogonal matrices containing the left and right singular vectors, respectively, and $\mathbf{\Sigma}$ is a diagonal matrix of singular values. The optimal rotation and scaling matrix $\mathbf{R}$ is computed as:

$$\mathbf{R} = \mathbf{V}\mathbf{U}^{\top}.$$

A critical step in the algorithm is to check for improper rotations by evaluating the determinant of $\mathbf{R}$. If $\det(\mathbf{R}) < 0$, the transformation represents a reflection, which is corrected by flipping the sign of the last row of $\mathbf{V}^{\top}$ and recomputing $\mathbf{R}$. This ensures that the resulting transformation corresponds to a proper rotation ($\det(\mathbf{R}) = 1$).



(a) (b)

Figure 4. Illustration of systematic error arising from differences in initial joint position (a) and its correction (b)

To estimate the transformation matrix, we selected four points of the skeletal model that, due to anatomical features of the human body, are least susceptible to mutual coordinate changes during various movements. These are points numbered 0, 1, 10, and 13 in Figure 2.

After applying the found transformation, all that remains is to calculate the weighted Euclidean distance between the sets of points of the compared models. The distance between pairs of skeletons I (instructor) and P (patient) could be determined by the following expression:

$$R(I,P) = \frac{1}{N}\sum_{i=0}^{N-1}\sqrt{\sum_{m\in\{x,y,z\}} w_i^m \left(I_i^m - P_i^m\right)^2}. \qquad (3)$$

where N – number of used points of skeletons, I_i^m – m-th coordinate of i-th point of skeletal model I, P_i^m – m-th coordinate of i-th point of skeletal model P, w_i^m - weights of importance, represented the degree of attention to the m-th coordinate of i-th point (2).

4. Dataset for quality assessment

A special dataset was created for experimental evaluation of the proposed dissimilarity measure. We collected video records of instructor (physiotherapist) and 22 persons in laboratory environment. There were two kinds of exercises – in sitting position and in standing position. Each person repeated the instructor's exercises three times. The first time, they did it as accurately as possible. The second time, they did it less precisely, simulating a patient with minor motor impairments. The third time, they did it even less accurately, simulating a patient with severe motor impairments. We will denote these three classes of repetition as “good”, “intermediate” and “bad”. Thus, the total number of records in the dataset (including instructor) is equal to 134.

All data were obtained using Microsoft Kinect v2 and include the coordinates of 17 points of the standard skeletal model for each frame of the video recordings.

In the preprocessing stage the dynamic time warping (DTW) was applied to each instructor – patient records pair and optimal alignment was obtained. In this work, we do not consider the aforementioned transformation, assuming that when determining the difference between two video sequences, skeletal models were compared at corresponding frames.

5. Experimental results

In the experiments conducted, we investigated the influence of both the refined normalization process and the consideration of varying informativeness of skeletal points during exercise execution. The distance between the instructor's and patients' sequences was computed based on the averaging of dissimilarity scores between corresponding skeletal models in matched frames. The matching between two sequences was obtained by DTW with correspondent dissimilarity measure.

On one hand, in our collected data, the similarity of a patient's exercise performance to that of the instructor is measured on an ordinal scale. On the other hand, from the perspective of rehabilitation after illness, the numerical assessment of progress may vary significantly among different patients. Comparing the scores of various patients in terms of recovery degree and the condition of specific bodily systems remains a subject for further research. In this work, it was important for us to maintain correspondence between the increase in dissimilarity and the levels of “good”, “intermediate”, and “bad” execution.

To this end, for the overall evaluation of a series of exercises performed by patients, we used Spearman's rank correlation coefficient (Spearman, 1904) with averaging across all dataset. This approach ensured a robust alignment between qualitative rankings and quantitative differences in performance.

In the study, four variants of distance estimation between skeletal models were compared: weighted Euclidean distance (3) with normalization based on an affine transformation, weighted Euclidean distance with height-based normalization, average Euclidean distance with normalization based on an affine transformation, and average Euclidean distance with height-based normalization (Seredin et al., 2023).

The results of the study are illustrated in Fig. 5. Each line corresponds to the records of certain patient (P_patient number) in dataset, 01 corresponds to standing position, 02 corresponds to sitting position. The color bars reflect the exercise execution levels: green color for “good”, blue color for “intermediate”, and red color for “bad”. The proper sequence should be green, blue red. Different order signals for error. The column (a) presents weighted Euclidean distance (3) with normalization based on an

affine transformation, column (b) – weighted Euclidean distance with height-based normalization, column (c) – average Euclidean distance with normalization based on an affine transformation, and column (d) – average Euclidean distance with height-based normalization.

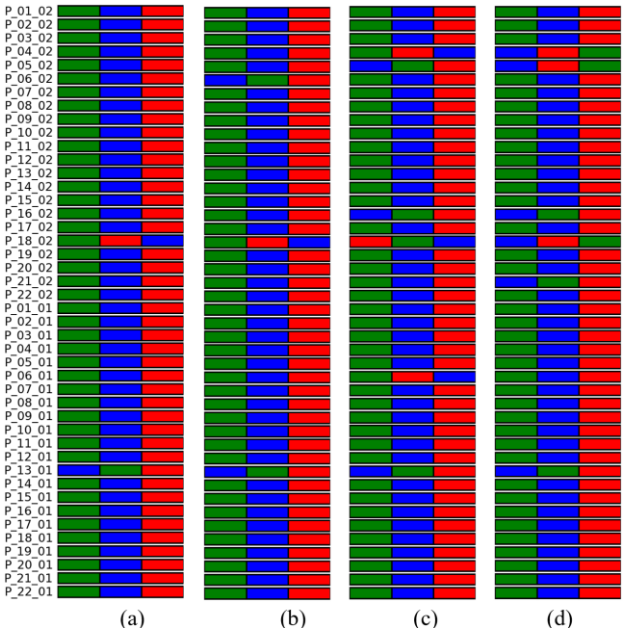


Figure 5. Results of the ablation study: (a) weighted Euclidean distance with normalization based on an affine transformation, (b) weighted Euclidean distance with height-based normalization, (c) average Euclidean distance with normalization based on an affine transformation, (d) average Euclidean distance with height-based normalization

Average Spearman's rank correlation coefficients ρ_{avr} for all four measures:

(a): ρ_{avr} = 0.987;

(b): ρ_{avr} = 0.966;

(c): ρ_{avr} = 0.909;

(d): ρ_{avr} = 0.864.

Experiments show a clear improvement in the separability of the dataset records by the degree of accuracy of repetition when using the proposed measure of difference between skeletal models. The most significant improvements are associated with the use of the proposed method for considering the varying informativeness of skeleton points for different movements.

6. Conclusion

In this study, we developed and evaluated a method for assessing the similarity between skeletal models of an instructor and patients during physical therapy exercises. The proposed approach addresses key challenges in traditional rehabilitation programs by leveraging advancements in computer vision and artificial intelligence to provide automated, objective feedback on exercise performance. Our contributions can be summarized as follows. We introduced the concept of informativeness for skeletal points, allowing us to adjust the pairwise dissimilarity function based on the specific characteristics of each exercise. This ensures that more mobile points, which are critical for evaluating movement quality, contribute significantly to the similarity assessment, while stationary points, influenced primarily by anatomical differences, are de-emphasized. A refined normalization process was proposed to align coordinate systems across different sensors and account for variations in pose orientation, body proportions, and sensor-specific biases. By applying affine transformations based on four stable skeletal points, we achieved improved consistency in comparing skeletal models. Through experiments with a custom dataset comprising video records of one instructor and 22 participants performing sitting and standing exercises, we demonstrated significant improvements in the separability of records based on the degree of exercise accuracy. The proposed method outperformed existing approaches, achieving high Spearman's rank correlation coefficients (up to 0.987).

While our work provides a robust framework for objectively assessing the effectiveness of patient physical rehabilitation, further research is needed to compare patient scores in terms of recovery degree and motor disorders. Additionally, extending the method to diverse exercises and populations could enhance its applicability in clinical settings.

Overall, this study highlights the potential of automated systems in objective rehabilitation assessments, offering a new tool for both physiotherapists and patients in monitoring progress of physical rehabilitation. Future work will focus on expanding the scope of the method and integrating it into therapeutic physical training.

Acknowledgements

This research is funded by the Ministry of Science and Higher Education of the Russian Federation within the framework of the state task FEWG-2024-0001.

References

Bian, Z.-P., Hou, J., Chau, L.-P., Magnenat-Thalmann, N., 2015. Fall Detection Based on Body Part Tracking Using a Depth Camera. IEEE J Biomed Health Inform 19, 430–439. https://doi.org/10.1109/JBHI.2014.2319372

Chen, C., Zhuang, Y., Nie, F., Yang, Y., Wu, F., Xiao, J., 2011. Learning a 3D human pose distance metric from geometric pose descriptor. IEEE Trans Vis Comput Graph 17, 1676–1689. https://doi.org/10.1109/TVCG.2010.272

Chen, Y., Wang, N., Zhang, Z., 2018. DarkRank: Accelerating Deep Metric Learning via Cross Sample Similarities Transfer, Thirty-Second AAAI Conference on Artificial Intelligence.

Du, Y., Wang, W., Wang, L., 2015. Hierarchical recurrent neural network for skeleton based action recognition. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 07-12-June, 1110–1118. https://doi.org/10.1109/CVPR.2015.7298714

Fang, H. S., Li, J., Tang, H., Xu, C., Zhu, H., Xiu, Y., Li, Y. L., & Lu, C. (2023). AlphaPose: Whole-Body Regional Multi-Person Pose Estimation and Tracking in Real-Time. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(6), 7157–7173. https://doi.org/10.1109/TPAMI.2022.3222784

Hussein, M.E., Torki, M., Gowayyed, M.A., El-Saban, M., 2013. Human action recognition using a temporal hierarchy of

covariance descriptors on 3D joint locations. IJCAI International Joint Conference on Artificial Intelligence 2466–2472.

Kabsch, W. (1976). A solution for the best rotation to relate two sets of vectors. Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography, 32(5), 922–923. https://doi.org/10.1107/S0567739476001873

Kaya, M., Bilge, H.S., 2019. Deep Metric Learning: A Survey. Symmetry (Basel) 11, 26. https://doi.org/10.3390/sym11091066

Lam, A.W.K., Varona-Marin, D., Li, Y., Fergenbaum, M., Kulić, D., 2016. Automated Rehabilitation System: Movement Measurement and Feedback for Patients and Physiotherapists in the Rehabilitation Clinic. Hum Comput Interact 31, 294–334. https://doi.org/10.1080/07370024.2015.1093419

Seredin, O.S., Kopylov, A. V., Surkov, E.E., Huang, S.C., 2023. The basic assembly of skeletal models in the fall detection problem. Computer Optics 47, 323–334. https://doi.org/10.18287/2412-6179-CO-1158

Spearman, C., 1904. The Proof and Measurement of Association between Two Things. The American Journal of Psychology. 15 (1): 72–101. doi:10.2307/1412159.

Surkov, E., Seredin, O., Kopylov, A., Kushnir, O., 2024. Unifying Human Pose Estimation in the Fall Detection Problem. Pattern Recognition and Image Analysis 34, 1061–1073. https://doi.org/10.1134/S1054661824701104

Wang, J., Liu, Z., Wu, Y., Yuan, J., 2012. Mining actionlet ensemble for action recognition with depth cameras, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 1290–1297. https://doi.org/10.1109/CVPR.2012.6247813

Yan, S., Xiong, Y., Lin, D., 2018. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. arXiv preprint.

Zhang, S., Liu, X., Xiao, J., 2017. On geometric features for skeleton-based action recognition using multilayer LSTM networks, in: Proceedings - 2017 IEEE Winter Conference on Applications of Computer Vision, WACV 2017. Institute of Electrical and Electronics Engineers Inc., pp. 148–157. https://doi.org/10.1109/WACV.2017.24

Zhang, X., Xu, C., Tian, X., Tao, D., 2020. Graph edge convolutional neural networks for skeleton-based action recognition. IEEE Trans Neural Netw Learn Syst 31, 3047–3060. https://doi.org/10.1109/TNNLS.2019.2935173