Data Driven Human Movement Assessment

Danny Dressler[1], Pavlo Liapota[2], and Welf Löwe[3]

[1] AIMO AB, https://www.aimo-health.com
[2] Softwerk AB, https://softwerk.se
[3] DISA, Linnaeus University, Welf.Lowe@lnu.se, https://lnu.se/disa

Abstract. Quality assessment of human movements has many of applications in diagnosis and therapy of musculoskeletal insufficiencies and high performance sport. We suggest five purely data driven assessment methods for arbitrary human movements using inexpensive 3D sensor technology. We evaluate their accuracy by comparing them against a validated digitalization of a standardized human-expert-based assessment method for deep squats. We suggest the data driven method that shows high agreement with this baseline method, requires little expertise in the human movement and no expertise in the assessment method itself. It allows for an effective and efficient, automatic and quantitative assessment of arbitrary human movements.

1 Introduction

Healthcare is in the middle of fundamental changes from fee-for-service to value-centred systems. Approaches for payment based on patient value (best possible health achieved) and system value (effective treatments at efficient costs) need to be able to measure clinical outcomes. This capability, until recently, was not part of most hospital, health, or enterprise-resource-planning systems [4]. With inexpensive sensor technologies and data analytics becoming increasingly available, it is nowadays possible to collect data on the clinical activities of healthcare, the health status of a patient and the change in this status after treatment. Our work contributes to modern healthcare with an automated objective method for the assessment of physical health of the human musculoskeletal system to help diagnose, predict or prevent related pain, injuries and long-lasting diseases.

Our approach supports the diagnosis of musculoskeletal issues based on commodity 3D motion capturing devices, such as the Kinect.[4] Unlike similar tools for physical therapists[5,6] adding little value to the caring or nursing process, our approach localizes issues and quantifies their severity.

Any effective and efficient method for the assessment of human movements should be independent of the actual movement. Moreover, it should be efficient to add new movements with little expertise in the movement and without any expertise in the method itself. This way new movements, e.g., for the assessment

[4] https://en.wikipedia.org/wiki/Kinect

[5] https://www.qinematic.com

[6] https://kinetisense.com

of insufficiencies or risks in specific professions or sports could be added swiftly by the medical and physiotherapist experts without any technical support.

Quite a few studies research the recognition of human movements using commodity 3D sensor technology [15, 17, 14, 5, 9]. While these approaches are similar to ours in their feature extraction and preprocessing steps, their goal is the classification of different movements, not their quality assessment. The same technology has also been used in movement quality assessment [16, 6, 8, 12]. However, these studies aim at *qualitative* assessments or at identifying different abnormal movement patterns rather than at *quantitatively* scoring the quality of a movement. Moreover, while the assessment methods are generalized to different movements [16, 6] or can be generalized [8, 12], the effort for adding a new movement was not studied yet. Finally, for some of the studies [8, 12], preprocessing transformations such as dimensionality reduction using manifold learning [3] enable a fast online assessment but make it impossible to localize the impairments. Pirsiavash et al. suggest a learning framework for training models able to quantitatively assess the quality of human movements from 2D videos [10]. Their approach trains a regression model from spatiotemporal pose features to scores obtained from expert judges. Features are extracted using unsupervised feature learning directly from 2D video data. Therefore, a localization of issues is not possible. Dressler et al. [2] suggest an automated, quantitative assessment method allowing for issue localization. However, it is a digitalization of an overhead deep squat assessment and, hence, tightly coupled to this movement.

The paper contributes with (i) a data driven, automated, quantitative methods for the assessment of issues with the human musculoskeletal system and with an evaluation of their (ii) accuracy and (iii) efficiency in adopting new movements. More specifically, five different assessment methods are suggested and their accuracy compared to a baseline method whose accuracy is validated by human experts. The effort and expertise needed to adapt new movements is discussed and empirically evaluated. Therefore, Section 2 introduces common preprocessing steps of the different data driven assessment methods that are then defined in Section 3. Section 4 evaluates them and proposes a champion. Section 5 concludes the paper and shows directions of future work.

2 Common Preprocessing of the Assessment Approaches

A *sequence* is a list of frames from a recording of a human movement. It is called a *master* sequence if the movement is executed perfectly. Any sequence that is to be scored is called a *user* sequence. Each *frame* is a record of features and describes the body posture at a specific point in time during a movement. Each *feature* describes an aspect of the body posture at a specific point in time. A feature is called *direct* if it is directly measured by the 3D camera or *indirect* if it is computed from the direct or other indirect features. The direct features include the x, y, and z coordinates of skeleton joints. Indirect features include angles between limbs and the axes of the 3D coordinate system. An *aggregated sequence* is a list of aggregated frames aggregating one or more sequences into one. An *aggregated frame* aggregates two or more frames into one. It is a vector of sample distributions of the feature values of each feature. For scoring, we compare a user sequence with an aggregated master sequence.

All assessment approaches perform four steps: (1) *Building the aggregated master sequence* is only performed once while the following steps are performed for each user sequence. (2) *Preparing the sequences* mitigates noisy feature values due to random camera and skeleton recognition errors. (3) *Matching* aligns a user sequence with the aggregated master sequence. (4) *Scoring* computes indicators for individual features and an overall movement score. Details are introduced in the following paragraphs and precisely defined in [2].

Step (1) Building the aggregated master sequence assumes a set of master sequences, builds an initial aggregated master sequence and aggregates it with the other master sequences. All but the first two steps are done automatically: (1.1) Select the best master sequence. It should have a constant movement speed without any delays and stops. (1.2) Cut off leading and trailing frames of postures that do not belong to the movement. (1.3) Prepare this master sequence, i.e., apply all sub-steps of Step (2) before matching and all but the first sub-step after matching. (1.4) Group subsequent frames. Each group contributes to a separate aggregated frame. (1.5) Separately for each group and for each feature, compute a numerical sample distribution of the feature values.

For the remaining master sequences, (1.6) Prepare the master sequence: Step (2) before matching. (1.7) Match the master sequence with current aggregated master sequence: Step (3). Matching maps each frame of the master sequence to an aggregated frame of the aggregated master sequence. (1.8) Prepare the master sequence: Step (2) after matching. (1.9) For each frame and each feature of the master sequence, add the feature value to the sample distribution of the respective feature of the mapped aggregated frame.

Step (2) Preparing the sequences, before matching (2.1) *Floor clip plane alignment*: for each frame, the joint position vectors are rotated such that the floor clip plane is parallel to the x, z plane. (2.2) *Smoothening*: for all *direct* features, a sliding average of feature values is computed. (2.3) *Interpolate*: if a joint was not visible for $< k$ consecutive frames its position is interpolated. If a joint was not visible for $\geq k$ consecutive frames, the joint is considered not tracked and an error is reported. **After matching** (2.4) *Cut* leading and trailing frames of postures not belonging to the movement. (2.5) Compute a *Scaling* transformation (procrustes analysis) that moves each joint of the first frame of a sequence to the corresponding mean joint position of the first aggregated frame of the aggregated frame sequence. Then apply this scaling to all other frames of the sequence. (2.6) Compute a *Hip rotation* transformation for the first frame of a sequence that lets skeletons "look" towards the camera. Then apply this rotation to all other frames of the sequence. (2.7) Compute the *indirect features* from the direct features for each frame.

Step (3) Matching Let N be the number of aggregated master sequence frames and M be the number of user sequence frames. A *matching* $\mathcal{M}$ is a relation $\subseteq [1 \dots N] \times [1 \dots M]$. A matching $\mathcal{M}$ is *correct iff*

(i) $\forall n \in [1 \ldots N] : (n, \_) \in \mathcal{M}$

(ii) $\forall m \in [1 \ldots M] : \begin{cases} (\_, m) \in \mathcal{M} \vee & \text{(matched)} \\ \forall m' \leq m : (\_, m') \notin \mathcal{M} \vee & \text{(unmatched leading)} \\ \forall m' \geq m : (\_, m') \notin \mathcal{M} & \text{(unmatched trailing)} \end{cases}$

(iii) $(n, m) \in \mathcal{M} \Rightarrow \nexists (n', m') \in \mathcal{M} \wedge n' < n \wedge m' > m$

For a matching to be correct, all aggregated master sequence frames are matched (i), all user sequence frames are either matched, leading or trailing (ii), and the matching must obey the order of frames in the sequences (iii).

Our matching algorithm *Sequence Alignment* [2] is a generalization of Dynamic Time Wrapping (DTW) [11]. It finds a matching with minimum costs among all *correct* matchings. The costs of a matching are defined as the deviations of the aggregated master sequence frames and the matched user sequence frames. More specifically, the deviation of a value v of feature f of a user sequence frame from a distribution $D_{f,n}$ of the corresponding feature values in an aggregated master sequence frame n with mean $\mu_{f,n}$ and standard deviation $\sigma_{f,n}$ is the z-score of v in $D_{f,n}$, i.e.,

$$d_{f,n}(v) = \frac{v - \mu_{f,n}}{\sigma_{f,n}}. \tag{1}$$

The deviation of a frame to an aggregated master sequence frame is the average of the deviations of all contained features. The deviation of an aggregated master sequence and a user sequence is the average deviation of the matched aggregated and user frames.

Step (4) Scoring is based on selected features and the deviation of their actual from the expected values weighted with the "difficulty" of a body posture. What "difficult" means is defined for each joint based on special indirect features, i.e., angles between limbs related/adjacent to that joint, e.g., the angle between tight and lower leg determines the difficulty of a posture for the knee. For each joint $j \in 1 \ldots J$, an angle a_j^0 for the relaxed posture and an angle a_j^1 for the difficult posture is defined offline by a physiotherapist, e.g., for the knee, $a_{knee}^0 \approx 180°, a_{knee}^1 \approx 20°$. W.l.o.g. we assume $a_j^0 > a_j^1$. These angles are defined once and independent of the movements to assess based on knowledge about the human musculoskeletal system. Offline, once for the aggregated master sequence, we calculate weights w_n between 0 and 1 for its N frames as follows. Let $\mu_{j,n}$ be the mean of the special angle j in the aggregated frame $n \in 1 \ldots N$, truncated if larger than a_j^0 and smaller than a_j^1. The weight w_n of the aggregated frame n is set to the length $||\cdot||_2$ of the vector $\boldsymbol{w}_n = [w_{1,n}, \ldots, w_{J,n}]$ of the J special angle weights $w_{j,n}$, one for each joint j, divided by length of the vector $\mathbf{1} \in \mathbb{R}^J$ of J ones to normalize the weights where $w_{j,n} = \left(\frac{\mu_{j,n} - a_j^1}{a_j^0 - a_j^1}\right)^4, j \in [1, J], n \in [1, N]$ and $w_n = \frac{||\boldsymbol{w}_n||_2}{||\mathbf{1}||_2}$. Let f be a feature selected for scoring and v_m be the feature value for frame $m \in 1 \ldots M$ matched to an aggregated frame $n \in 1 \ldots N$ and $d_{f,n}(v_m)$ be the deviation as calculated according to Equation (1). Set d_f to the average of the three largest values $d_{f,n}(v_m) \times w_n$. An overall movement quality score is calculated from these weighted averaged deviations d_f, cf. Section 3.

3 Data Driven Assessment Methods

Steps (1)–(4) are agnostic w.r.t. the actual movement to be assessed. They can be computed based on a set of *master sequences* of any movement. The only movement expertise necessary is to pick *master sequences* from sequences recorded for a movement and to select relevant features. For each suggested scoring method, we will discuss the additional expert knowledge required.

Approach (0) is not data driven but used for assessing the accuracy of the others. It is actually highly dependent on expertise in the NASM overhead deep squat, an exercise standardized by the National Academy of Sports Medicine (NASM)[7] that comes with movement execution and scoring specifications. It is medically validated in the sense that a low NASM score is a good indicator of mobility and stability insufficiencies that, in turn, indicate current or future problems with the musculoskeletal system. The NASM suggests assessing different potential weak links of a body from feet to head and to score them individually. An overall NASM score is then set based on the scores for these weak links contributing to the overall score with different weight factors [7].

For each weak link feature f, the weighted averaged deviations d_f is expected to be 0, but a deviation $d_f^0 > 0$ may still be ignorable. Offline, once for each weak link, we define d_f^0 along with a value d_f^1 marking a clear deviation from the expectation. These thresholds are different for the different weak links. Each actual deviation d_f is linearly mapped to an indicator such that $ind_f(d_f^0) = 0$ and $ind_f(d_f^1) = 1$ by $ind_f(d_f) = \frac{d_f - d_f^0}{d_f^1 - d_f^0}$. Then $ind_f(d_f)$ is updated such that negative values are set to 0. Then $ind_f(d_f)$ is softened: $ind_f(d_f) := ind_f(d_f)^2$ if $ind_f(d_f) < 1$ and $ind_f(d_f) := \sqrt{ind_f(d_f)}$, otherwise. Finally, $ind_f(d_f)$ is updated such that values above 1.5 are cut off. The baseline score S_0 is a weighted sum Ind of 14 weak link indicators ind_f. We pick the maximum of the (left-right) symmetric indicators. This leaves 7 indicators. We select different feature weights 1, 2, and 4 according to the NASM specification; the sum of these 7 weights is 13. Then score $S_0 = \max\left(1 - \frac{Ind}{10}; 0\right)$ with $Ind \in [0; 19.5 = 13 \times 1.5]$.

Approach (1) computes the extremeness of the deviations of a value d_f, $f \in F$ the set of features selected for scoring, in terms of the probability that the same or even higher deviating feature values can be observed. It is similar to the approach in [13] suggested for data driven assessment of software quality. Let X_f be a random variable describing observable deviations in feature f for all user sequences. The indicator $ind_f(d_f)$ of an actually observed deviation d_f is

$$ind_f(d_f) = Pr(X_f \leq d_f), ind_f(d_f) \in [0, 1]. \quad (2)$$

An $ind_f(d_f) = 1$ is worst. It means that it is certain that all deviations are smaller than or equal to d_f. An $ind_f(d_f) = 0$ is the best. It means that it is certain that all deviations are larger than d_f. As the distribution of X_f is unknown, we use a sample distribution of (sufficiently many) observations d_f instead and approximate $Pr(X_f \leq d_f)$ numerically. Approach (1) defines the score $S_1 \in [0, 1]$

[7] https://en.wikipedia.org/wiki/National_Academy_of_Sports_Medicine

as the complementary of the joint probability of the indicator probabilities, also approximated numerically: $S_1(\boldsymbol{d}) = 1 - Pr(X_1 \leq d_1 \wedge \ldots \wedge X_{|F|} \leq d_{|F|}), \boldsymbol{d} = [d_1, \ldots, d_{|F|}]$, i.e., a score $S_1(\boldsymbol{d}) = 1$ is best while a score $S_1(\boldsymbol{d}) = 0$ is worst. This continues to hold for all scoring approaches defined below. Approach (1) is purely data driven; no additional expertise in the movement is needed.

Approach (2) is based on the indicator probabilities $ind_f(d_f)$ as defined in Equation (2) and sets the score based on a weighted sum of these probabilities. Let $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_{|F|}]$ be a column vector of positive weights and $\boldsymbol{ind}(\boldsymbol{d}) = [ind_1(d_1), \ldots, ind_{|F|}(d_{|F|})]$ a row vector of feature indicators according to Equation (2). Let $\bullet$ be the dot product and $|| \cdot ||_1$ the sum of elements of vectors. $S_2(\boldsymbol{d}) = 1 - \frac{\boldsymbol{ind}(\boldsymbol{d}) \bullet \boldsymbol{\beta}}{||\boldsymbol{\beta}||_1}$. Approach (2) is not purely data driven anymore as it requires additional expertise in the movement, namely the definition of weights of features selected for scoring.

Approach (3) is based on the indicator probabilities $ind_f(d_f)$ as defined in Equation (2) and defines the score as a weighted sum of these probabilities. However, it computes the weights $\boldsymbol{\beta}$ using linear regression for a sample of k data points mapping deviation vectors to actual scores provided by human experts. Let $\boldsymbol{S}$ be a vector of such human expert scores $[S^1, \ldots S^k], S^i \in [0; 1]$ and let $\boldsymbol{Ind}$ be a $(k \times |F|)$-matrix with each row i containing a vector $\boldsymbol{ind}(\boldsymbol{d}^i) = [1, ind(d_1^i), \ldots, ind(d_{|F|}^i)]$ of probabilities of deviations $ind(d_f^i)$ of the features $f \in F$ in the user sequence i corresponding to the score S^i. Set the weights $\boldsymbol{\beta}$ such that the error $||\boldsymbol{\epsilon}||_2$ in $\boldsymbol{S}^T = \boldsymbol{Ind} \bullet \boldsymbol{\beta} + \boldsymbol{\epsilon}^T$ gets minimal. Using these weights that are negative for all $ind(d_f^i)$, define the score for a vector of deviations $\boldsymbol{d}$ in an unknown sequence: $S_3(\boldsymbol{d}) = \frac{\boldsymbol{ind}(\boldsymbol{d}) \bullet \boldsymbol{\beta}}{||\boldsymbol{\beta}||_1}$. Approach (3) requires an expert scoring of k movements. The weights are set in data driven supervised learning.

Approach (4) also computes the weights using linear regression but constrains them such that symmetric features get the same weight. Therefore, we compute a vector of sums $sym(\boldsymbol{d})$ of these symmetric features before regression. Let $\mathcal{F} = \{F_1, \ldots, F_s\} \subseteq 2^F$ be a set of sets of the symmetric features from the set of selected features F. Define $sym_{\mathcal{F}}(\boldsymbol{d}) = \left[\left(\sum_{f \in F_1} ind(d_f) \right), \ldots, \left(\sum_{f \in F_s} ind(d_f) \right) \right]$. For computing the s weights, we use a matrix $\boldsymbol{Ind}$ with each row i containing a vector $sym_{\mathcal{F}}(\boldsymbol{d}^i)$ and define $S_4(\boldsymbol{d}) = S_3(sym_{\mathcal{F}}(\boldsymbol{d}))$ using these weights.

In addition to the requirements of Approach (3), Approach (4) needs the definition of symmetric features. Since, Approach (4) sets weights as a special case of the weights set by Approach (3), both approaches should converge with the number k of training data points growing. However, for smaller k, Approach (4) could avoid overfitting the training data and could, therefore, require fewer user sequences to be scored by experts for the regression.

Approach (5) is a variant of Approach (4). However, it normalizes the deviation values d_f based on the sample data points before calculating their extremeness $ind(d_f)$. Whenever a human expert scores a movement as (almost) perfect, i.e., $S(\boldsymbol{d}) > 0.95$, we capture the corresponding deviation values $d_1, \ldots, d_{|F|}$. This way, we compute sample distributions $\hat{D}_1, \ldots, \hat{D}_{|F|}$ of ignorable deviations, one for each selected feature $f \in F$. Then we compute a vector of ignorable deviation thresholds $[\hat{d}_1, \ldots, \hat{d}_{|F|}]$ with $\hat{d}_i$ the 95-th percentiles of $\hat{D}_i$. Deviation value vectors are normalized using these ignorable deviation thresholds: $norm(\boldsymbol{d}) = \left[\frac{d_1 - \hat{d}_1}{\max(D_1) - \hat{d}_1}, \ldots, \frac{d_{|F|} - \hat{d}_{|F|}}{\max(D_{|F|}) - \hat{d}_{|F|}}\right]$, where $\max(D_i)$ is the maximum deviation in the i-th selected feature of all sample data points. For computing the weights, use a matrix $\boldsymbol{Ind}$ with each row i containing a vector $sym_{\mathcal{F}}(norm(\boldsymbol{d}^i))$ and define $S_5(\boldsymbol{d}) = S_3(sym_{\mathcal{F}}(norm(\boldsymbol{d})))$ using these weights. Approach (5) has the same requirements as Approach (4).

4 Evaluation

Method. We evaluate the accuracy of the different approaches as follows. The baseline approach (0) is the digitalization of the NASM overhead deep squat assessment and scoring, i.e., it implements the NASM scoring specification. The two approaches—the manual human-expert-based NASM overhead deep squat assessment and its digitalization—show a high agreement [2]. We assess the agreement of the scores $S_i, i \in [1, 5]$ with the scores S_0 of the baseline approach for an NASM overhead deep squat. As S_0 shows high agreement with the scoring of NASM experts, a high agreement of S_i and S_0 is considered a high accuracy.

We calculate the Pearson coefficient r for the correlation of scores coming from different approaches with the baseline approach. Altman and Bland argue that a high correlation is necessary but not sufficient for a good agreement between two approaches and suggest an additional analysis of differences [1]. Therefore, we additionally assess the differences of scores for the approach with the highest correlation to baseline. For this champion approach, we finally check the convergence of the correlation when trained on fewer data points in order to answer the question of how much training data is needed in order to stably get a high correlation.

Data. The assessment is based on a sample of 2094 user sequences recorded in 2018, the bulk during so-called company health days in Germany where AIMO recorded deep squats of health interested employees. The sample data contains scans of female and male persons, 25–65 years old, working in office jobs.

Implementation. The evaluation uses a Matlab implementation of the approaches (1) to (5) and the AIMO production code implementation of the baseline approach (0). All implementations use the same (single) master sequence and the same deviations of NASM features computed in Steps (1)–(3), cf. Section 2. For approach (2), we use the same weights as the baseline approach. For the approaches (1) and (2), the Pearson correlation coefficient r is deterministic for the sample. For the regression based approaches (3) to (5), the result depends on the splitting of the sample in training and test data sets, each containing 50% of the sample data. For these approaches, we therefore repeat the experiments and report the average r of 100 runs with fair random splitting of the sample in training and test data. Finally, we train the champion approach with $k = 2^4 \ldots 2^{10}$ data points. For each k , we select $2k$ data points randomly from the full set a 100 times. We split each set of size $2k$ 100 times randomly in k training and k test data points and capture r of these 100×100 runs.

Table 1. Correlation of scores of approaches (1)–(5) with baseline approach (0)

Approaches	S_1	S_2	S_3	S_4	S_5
Pearson coefficients r of correlation with S_0	0.235	0.569	0.738	0.791	**0.875**

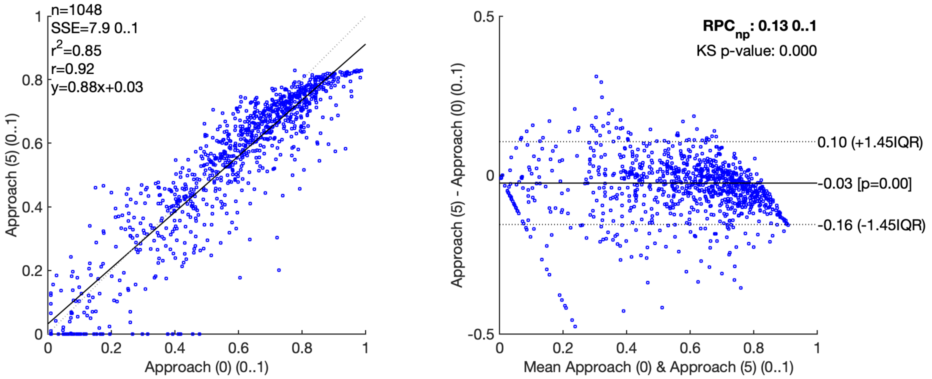

Fig. 1. Scores S_5' of approach (5) vs. scores S_0 of the baseline approach (0)

Results. Table 1 shows the correlation of the scores of approaches (1)–(5) with the scores of the baseline approach (0). Approach (5) has the highest correlation. Based on the regression equation of S_5 and S_0 analyzed for the training data, we adjust the final scoring for approach (5) to $S_5'(\boldsymbol{d}) = \max(9.09 \times S_5(\boldsymbol{d}) - 5; 0)$. This gives an even higher correlation with an average of $r = \mathbf{0.924}$ for 100 runs. Fig. 1 shows the Bland-Altman plots of approaches (0) and (5) for analyzing their differences for one random run. The mean of the differences of the two approaches is -0.03 score points. 95% of the differences of the two approaches are between 0.1 and -0.16 score points. Fig. 2 shows the correlation *convergence* of the approaches (5) and (0). For each training set size $k = 2^4 \ldots 2^{10}$ (log scale) it displays the intervals containing 95% (outer) and 68% (inner) of the

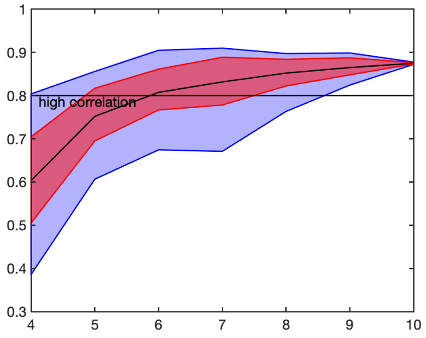

Fig. 2. Convergence of correlation (S_5, S_0) for $2^4 \ldots 2^{10}$ training data points

Pearson correlation coefficients r and their mean. For approach (5), the average correlation to baseline converges from $\mu_{r_4} = 0.603$ (with 95% of the observed values of $r_4 \in [0.386, 0.804]$) to $\mu_{r_{10}} = 0.875$ (with 95% of $r_{10} \in [0.872, 0.877]$).
Discussion. The correlation of approach (5) with baseline is very high (Pearson coefficient $r = 0.924$). The Bland-Altman differences analysis shows a bias towards the baseline of -0.03 score points, which can be neglected. This only means that approach (5) is more conservative compared to the baseline approach. Almost all differences between the two approaches are in the interval of $[0.1, -0.16]$ score points, which is an acceptably low difference interval with a range of 0.26 score points, only 13% of the maximum possible range of 2. We conclude that the two approaches show indeed a high agreement and that we can use the fully data driven approach instead of experts or digitalized expert assessment instructions. As for the question of how much training data is needed in order to stably get a high correlation: on average the correlation is high ($r > 0.8$) even with $\approx 2^6 = 64$ training data points. However, to be confident to 95% that a high correlation is achieved, we need $\approx 2^9 = 512$ training data points. This means that human experts have to manually score ca. 500 movements. For new kinds of movements, we need to bias the costs of a digitalization of an expert assessment against the expert costs for scoring the sample movements. Our experience with the deep squat shows that we get the sample data in one health day and expert-score them in another five person days. Also a digitialization would need ca. 100 scored movements for testing but, additionally, weeks of software development. This makes the data driven approach cheaper and faster.
Threads to validity. There are some limitations to the above conclusions. We simulate a high number of human expert scores for deep squat movements with a digitalized assessment of that movement validated against fewer expert scores. While the digitalized and expert assessments highly agree [2], this method might introduce a systematic error. There is nothing in the data driven approach tied to the deep squat. However, the validation needs to be confirmed for other movements, as well. Also, no extra effort has yet been made in selecting a representative sample, which is needed to generalize results to the whole population.

5 Conclusions and Future Work

The paper describes and evaluates data driven approaches to human movement assessment based on commodity 3D sensor technology. The goal was to define an automated, accurate, quantitative assessment method for general movements that allows for weakness localization. For the overhead deep squat movement, this goal was achieved. The suggested method requires only little expertise in the movement: experts need to score training recordings and select relevant features. For reducing the number of expert scores and for avoiding overfitting to the training data, a definition of movement symmetry axes and planes helps.

The data driven assessment approach was validated for the deep squat movement. It needs to be validated also for other movements before it is applicable in systems providing inexpensive and objective decision-support for the assessment of musculoskeletal insufficiencies. Moreover, the data driven approach itself could further be improved, e.g., features could be selected automatically based on their information entropy and non-linear and non-classic regressions approaches could beat the current champion. All this is matter of current and future work.

References

1. Altman, D.G., Bland, J.M.: Measurement in medicine: the analysis of method comparison studies. The Statistician **32**. (1983)
2. Dressler, D., Liapota, P., Löwe, W.: Towards an automated assessment of musculoskeletal insufficiencies. In: 11th Int. Conf. Innovation in Knowledge Based and Intelligent Engineering Systems: Data Selection in Machine Learning. (2019)
3. Elgammal, A.M., Lee, C.S.: The role of manifold learning in human motion analysis. In: Human Motion–Understanding, Modelling, Capture, and Animation. (2008)
4. Elton, J., O'Riordan, A.: Healthcare Disrupted: Next Generation Business Models and Strategies. Wiley (2016)
5. Jiang, M., Kong, J., Bebis, G., Huo, H.: Informative joints based human action recognition using skeleton contexts. Signal Processing: Image Comm. **33** (2015)
6. Khan, N.M., Lin, S., Guan, L., Guo, B.: A visual evaluation framework for in-home physical rehabilitation. In: IEEE Int. Symp. Multimedia. (2014)
7. National Academy of Sports Medicine: NASM Essentials of Personal Fitness Training. NASM, 6 edn. (2017)
8. Paiement, A., Tao, L., Camplani, M., Hannuna, S., Damen, D., Mirmehdi, M.: Online quality assessment of human motion from skeleton data. In: British Machine Vision Conf. BMVA Press (2014)
9. Pazhoumand-Dar, H., Lam, C.P., Masek, M.: Joint movement similarities for robust 3d action recognition using skeletal data. J. Visual Communication and Image Representation **30**. (2015)
10. Pirsiavash, H., Vondrick, C., Torralba, A.: Assessing the quality of actions. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) Computer Vision – ECCV. Springer (2014)
11. Sakoe, H., Chiba, S.: Dynamic programming algorithm optimization for spoken word recognition. IEEE Trans. Acoustics, Speech, Signal Processing **26**(1). (1978)
12. Tao, L., Paiement, A., Damen, D., Mirmehdi, M., Hannuna, S., Camplani, M., Burghardt, T., Craddock, I.: A comparative study of pose representation and dynamics modelling for online motion quality assessment. Computer Vision and Image Understanding **148**. (2016)
13. Ulan, M., Löwe, W., Ericsson, M., Wingkvist, A.: Introducing quality models based on joint probabilities. In: 40th International Conference on Software Engineering: Companion Proceedings. (2018)
14. Vemulapalli, R., Arrate, F., Chellappa, R.: Human action recognition by representing 3d skeletons as points in a lie group. IEEE Conf. Computer Vision and Pattern Recognition. (2014)
15. Wang, J., Liu, Z., Wu, Y., Yuan, J.: Mining actionlet ensemble for action recognition with depth cameras. In: IEEE Conf. Computer Vision and Pattern Recognition (2012)
16. Wang, R., Medioni, G., Winstein, C.J., Blanco, C.: Home monitoring musculoskeletal disorders with a single 3d sensor. In: IEEE Conf. Computer Vision and Pattern Recognition, Workshops. (2013)
17. Xia, L., Chen, C.C., Aggarwal, J.K.: View invariant human action recognition using histograms of 3d joints. In: IEEE Conf. Computer Vision and Pattern Recognition, Workshops. (2012)