Motion Evaluation by Means of Joint Filtering for Assisted Physical Therapy

Julia Richter, Christian Wiede, Lars Lehmann, Gangolf Hirtz
Digital Signal Processing and Circuit Technology, Department of Electrical Engineering and Information Technology
Technische Universität Chemnitz
Reichenhainer Str. 70, 09126 Chemnitz, Germany
julia.richter@etit.tu-chemnitz.de

***Abstract*—The supervision of rehabilitation exercises is crucial for a successful therapy. Due to a lack of therapists, technical assistance systems have recently come into focus to assist patients during their exercises. Latest research proved that characteristic motion errors can be detected by using the Kinect skeleton joints in connection with Incremental Dynamic Time Warping (IDTW) and machine learning. However, the processed joints were manually selected and the classifier predicts in a frame-wise manner. In order to facilitate an extension with more exercises, a central issue of this paper is to realize an automatic joint selection with optimal classification accuracy. Moreover, we propose an algorithm that post processes the frame-wise prediction. The results for both joint selection and post processing are of high quality and therefore make a significant contribution to an efficient, perceptible and user-friendly feedback generation.**

***Index Terms*—Health Care, E-Rehabilitation, Machine Learning, IDTW.**

I. Introduction

Traditionally, patients perform their rehabilitation under the guidance of a supervising therapist, for example in rehabilitation facilities. Recent evidence has shown, however, that due to the increasing number of patients, a therapist supervises several patients at the same time [1]. Consequently, the therapist is not able to continuously check each patient's performance during the exercise execution. One way to tackle this problem is to control the patient's performance by means of technical assistance systems. In this context, e-rehabilitation has risen considerable research interest. Recent studies, which are based on the Kinect version 1.0 skeleton, proposed principles and solutions that determine the similarity between a pre-recorded reference and the motion performed without supervision, e. g. [2]–[9]. Latest research demonstrated that even exercise-specific motion errors could be detected by applying machine learning techniques [10]. The following section gives an overview about existing approaches, their achievements as well as their limitations. Thereupon, we introduce and evaluate our method, which is based on automatic joint filtering and a repetition-wise motion error evaluation. The paper closes with an outlook on future research.

II. Related Work

Driven by the launch of the Kinect version 1.0 in 2010, researchers have investigated assessment methods that utilize the Kinect skeleton model [11]. Recent work examined motions during sports, such as dancing, ballet training and Tai Chi exercises, using the Kinect skeleton [2]–[5]. These works introduced several distance metrics in order to compare a performed motion against a reference. Such metrics were, for example, angles between joints [2], angles between body motion vectors [3] or the mean Euclidean distance between joint trajectories [5].

In the field of physical rehabilitation, similar principles can be found. In order to assess motions, several of the presented works successfully demonstrated the use of Dynamic Time Warping (DTW) [3], [6], [8] or variants of DTW, such as IDTW [7] and Subsequence DTW [9]. In general, these studies aim at evaluating a performed motion against a pre-recorded template. The principle is to align the two sequences and to calculate similarity measures or scores. These measures were then used to evaluate the performed motions and to give feedback. Antunes et al., for example, calculated and visualized feedback vectors that show the difference between the performed and the template motion of body parts [8]. Baptista et al. extended this approach by investigating the alignment of the template and the performed exercise to provide real-time feedback [9]. Therefore, they compared the alignment performance of Subsequence DTW and Temporal Commonality Discovery. Another work designed exercise-specific rules in XML format in collaboration with clinicians [12]. In order to assess an exercise, joint distances, joint angles, body segment orientations and repetition velocities were taken into consideration. In contrast to our approach, these methods do not aim at detecting and classifying pre-defined error classes. At this point, extant literature gives insights into approaches that aim at assessing motions by employing machine learning techniques. Nevertheless, existing systems still do not aim at classifying pre-defined error motions. Leightley et al., for example, analyzed and quantified human mobility to be good or poor by training separate models for groups of joints [13]. Capecci et al. investigated Hidden Semi-Markov Models (HSMM) to provide assessment scores [14]. They compared the HSMM scores as well as scores obtained by DTW against clinical scores that were specified by clinicians. Thereupon, the results demonstrated that HSMM scores correlated better with the clinical scores.

978-1-5090-4014-8/17/$31.00 ©2017 IEEE

Since all the presented approaches cannot recognize exercise-specific errors and by the majority require personalized references, a novel approach was suggested in [10] by Richter et al. This work extended the IDTW approach by a frame-wise motion error identification with the aim of detecting typical errors of the exercise hip abduction. For this purpose, classifiers were trained with one global, manually selected set of joints. Additionally, a normalization method was introduced in [10] to avoid the recording of personalized references. By doing this, a new patient can directly start with the exercise and the therapist's limited time is not required for recording the reference. In their approach, Richter et al. solved the previously mentioned alignment problem by detecting global minima in the exercise sequence. In this way, the starting points of every repetition could be detected and compared against the template in real-time using IDTW.

For this approach, however, there remains need for a method that automatically selects skeleton joints to obtain optimal classification results. In the work of Khan et al., only joints that meet the condition $\sigma_\mathrm{x} + \sigma_\mathrm{y} + \sigma_\mathrm{z} \leq \gamma$ were used for a particular exercise [7]. σ_x, σ_y and σ_z are the variances for a certain joint over a whole recording, and γ was a threshold value that was set to 0.005. Next to this method, several other approaches were used for joint selection, especially with regard to applications in the field of action recognition: Wang et al. determined the information content of joints by means of their potential and kinetic energy and used this energy information to weight the influence of the joints [15]. A further example for joint selection is presented in [16]. In addition to automatic joint selection for the detection of error motions in [10], more work is needed to prepare the presently frame-wise output to a feedback that the user is able to grasp and to handle while performing physical training.

In this paper, we therefore present a joint filtering technique that aims at an automatic optimization of the classification accuracy that has been manually achieved so far. We furthermore propose an evaluation method that, based on the frame-wise output in [10], generates a statistical output after every performed repetition of the exercise. This post-processing is sensible because the frame-wise output results in fast feedback changes since our system acquires skeleton data with approximately 30 Hz. Consequently, the user cannot perceive this fast changing information, whereas a summarized feedback representing repetitive errors would be more perceptible. Moreover, we introduce the designed feedback concept that facilitates feedback generation in phases when no therapist is present.

III. Method

Similar to [10], we want to detect typical motion errors that occur during the exercise hip abduction using non-personalized data. At this point, we employed the Kinect version 1.0 as well. The Kinect sensor and a processing unit is called *sensor unit* in this paper. The classes for this exercise are listed and explained in Table I. L represents the identifier for each class.

TABLE I
Classes with their description for the exercise hip abduction.

L	Class	Description
C	Correct	The exercise is performed correctly.
BK	Bent Knee	The knee of the abducted leg is flexed.
FO	Foot Outside	The foot is rotated outwards.
UB	Upper Body	The torso is bending sideways in the opposite abduction direction.
WP	Wrong Plane	The abducted leg does not move strictly sideways, but has a backward component.

A. Joint Filtering

While [10] used a manually defined, global joint combination, we propose to use class-specific joint combinations that maximize the accuracy η_n of each single class. These accuracies are calculated as follows:

$$\eta_n = (\mathrm{TPR}_n + \mathrm{TNR}_n) \cdot 0.5\,,\, n \in \{1, ..., N\}\,, \qquad (1)$$

whereas TPR_n and TNR_n are the true positive and the true negative rates for each of the N classes. These accuracies were determined for combinations of $J = 8$ joints that are possibly relevant for hip abduction. They are the joints belonging to the abducted leg, the upper torso and the head. The combinations consist of a number of 1 to 8 joints. This results in M combinations, while

$$M = \sum_{k=1}^{J} \binom{J}{k} = 255\,. \qquad (2)$$

The combinations with the highest corresponding class accuracies were chosen to train the single classifiers. The feature vector generation for training is the same as described in [10]. For both the manually selected, global and the automatically determined joint combinations, confusion matrices were calculated. Moreover, the overall accuracy for all classes N was calculated according to Equation 3.

$$\eta_{all} = \frac{1}{N} \cdot \sum_{n=1}^{N} \eta_n\,, \qquad (3)$$

Both the confusion matrices and the accuracies are presented and compared in Section IV-A.

B. Post Processing

In a next step, the patient's performance was evaluated prediction-wise. For this purpose, a concept was developed that exploits the TNR and the TPR of the single classifiers. In addition to this, the occurrence of the classes within one repetition was considered as well.

1) Weighting: Since the predictions have different influences at different stages within one repetition, we introduced weighting functions for each class n. This weighting is sensible, because, for instance, the error WP normally occurs in the middle of the repetition, so that the prediction can be more

trusted in this phase. The weighting functions were Gaussian density functions $G_{\mathrm{n}}(t)$ with different standard deviations σ_{n}:

$$G_{\mathrm{n}}(t) = \frac{1}{c_{\mathrm{n}}} \cdot e^{\frac{(t-\mu)^2}{2{\sigma_{\mathrm{n}}}^2}} , \tag{4}$$

while c_{n} is used for normalization and calculated as follows:

$$c_{\mathrm{n}} = \sum_{m=1}^{T} e^{\frac{(m-\mu)^2}{2{\sigma_{\mathrm{n}}}^2}} . \tag{5}$$

T is the number of frames per repetition, t is the frame index and $\mu = 0.5 \cdot T$. σ_{n} was calculated as follows:

$$\sigma_{\mathrm{n}} = f_{\mathrm{n}} \cdot T . \tag{6}$$

f_{n} is a factor between 0 and 1. The optimal values for σ_{n} were determined by finding the factors that maximize η_{all}. For our dataset of 10 persons, the factors f_{n} for each class with index n that gave optimal results were $(0.6, 0.3, 0.3, 0.3, 0.1)$. The resulting functions for $G_{\mathrm{n}}(t)$ are illustrated in Figure 1.

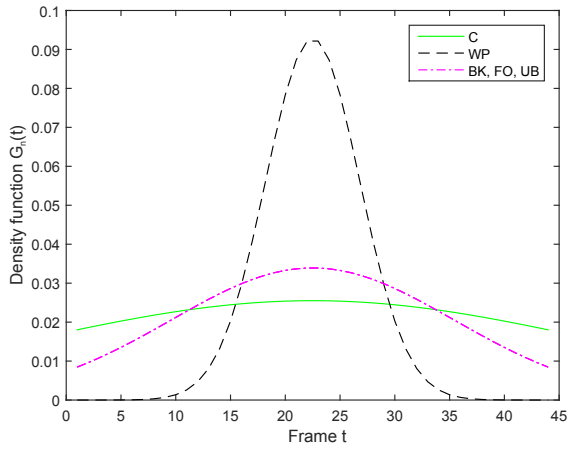

Fig. 1. Chosen weight functions for the classes C, BK, FO, UB and WP within one example repetition.

For every frame t, the prediction $P_{\mathrm{n}}(t)$ was then multiplied by the corresponding value of $G_{\mathrm{n}}(t)$. Since the prediction only has values of 0 or 1, it serves as a mask for $G_{\mathrm{n}}(t)$. The products are summed up for one repetition according to Equation 7 and result in a measurement for the occurrence of each class occ_{n} within one repetition.

$$occ_{\mathrm{n}} = \sum_{t=1}^{T} P_{\mathrm{n}}(t) \cdot G_{\mathrm{n}}(t) . \tag{7}$$

2) Error checking: Now, that we have determined the weighted occurrence occ_{n} for each class, we checked every error class, i.e. BK, FO, UB and WP against the class C. As an error class and the class C can be predicted simultaneously, we determined whether a certain error outweighs the class C. This check proportionally includes the $\mathrm{TPR_n}$ and the $\mathrm{TNR_n}$ of the single classifiers, so that a classifier with higher accuracy predominates a classifier with a lower accuracy. In this way, the performance of each single classifier is automatically

Fig. 2. System with sensor and feedback unit assists a patient during his exercises in rehabilitation center. Left: Patient performing an exercise. Right: Feedback display.

represented in this check. If a specific error was detected, the corresponding ϵ_{n} is set to 1, otherwise to 0:

$$\epsilon_{\mathrm{n}} = \begin{cases} 1, & \text{if } \mathrm{TPR_C} \cdot \mathrm{TNR_C} \cdot occ_{\mathrm{C}} < \mathrm{TPR_n} \cdot \mathrm{TNR_n} \cdot occ_{\mathrm{n}} \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

This check was performed for all error classes. If none of the errors outweighed the class C, the repetition was assumed to be correct. Section IV-B presents the performance results for the repetition-wise error determination.

C. Feedback Concept

In order to give live-feedback to patients, we developed a concept including the sensor and a feedback unit. The feedback unit receives information about detected errors from the sensor unit. This information is displayed in the form of a textual output and a colored avatar, as presented in Figure 2.

IV. Results and Discussion

A. Joint Filtering

This section compares the performance of the manually selected, global joint combination of [10] with the results achieved by the automatically determined, class-specific combinations. To compare the classifier performance, we have, in contrast to [10], evaluated the predictions independently from each other, so that the results in Table II are slightly different from those in [10].

TABLE II
Global, frame-wise confusion matrices. η_{all} is 0.84.

	L	$\overline{\mathrm{L}}$	η_n
C	0.84	0.16	0.80
$\overline{\mathrm{C}}$	0.24	0.76	
BK	0.95	0.05	0.95
$\overline{\mathrm{BK}}$	0.05	0.95	
FO	0.83	0.17	0.70
$\overline{\mathrm{FO}}$	0.44	0.56	
UB	0.89	0.11	0.88
$\overline{\mathrm{UB}}$	0.13	0.87	
WP	0.80	0.20	0.87
$\overline{\mathrm{WP}}$	0.07	0.93	

TABLE III
Class-specific, frame-wise confusion matrices. η_{all} is 0.86.

	L	$\overline{\mathrm{L}}$	η_n
C	0.85	0.15	0.80
$\overline{\mathrm{C}}$	0.24	0.76	
BK	0.95	0.05	0.95
$\overline{\mathrm{BK}}$	0.05	0.95	
FO	0.81	0.19	0.73
$\overline{\mathrm{FO}}$	0.35	0.65	
UB	0.95	0.05	0.94
$\overline{\mathrm{UB}}$	0.07	0.93	
WP	0.85	0.15	0.89
$\overline{\mathrm{WP}}$	0.07	0.93	

In Tables II and III, L represents samples that belong to the classes C, BK, FO, UB or WP. $\overline{\mathrm{L}}$ denotes samples that were

classified to belong to the rest of the corresponding classifier, i. e. $\overline{\mathrm{C}}$, $\overline{\mathrm{BK}}$, $\overline{\mathrm{FO}}$, $\overline{\mathrm{UB}}$ or $\overline{\mathrm{WP}}$.

The results for the automatically determined, class-specific combinations are slightly better than the results for the global combination selected in [10]. By intuition, the optimal global combination was already chosen in [10], which already results in high accuracies in Table II. Nevertheless, an intuitive choice is not guaranteed for more complex exercises. Therefore, the presented automatized selection is an important processing step to ensure an optimal joint selection for classification when the system is extended with more exercises. Moreover, we would like to point out that, when the classifiers are regarded separately as we did at this point, we obtain maximal possible accuracies. Since the class C and the error classes can occur simultaneously, the decision for either correct or incorrect will result in a decreasing accuracy. [10] favored C over the error classes and obtained an overall accuracy of 0.82.

B. Post Processing

This section evaluates the performance of the post processing algorithm when using the class-specific joint combinations. Table IV shows the confusion matrices that represent the capability of the algorithm to detect errors within a repetition.

TABLE IV
REPETITION-WISE CONFUSION MATRICES. η_{all} IS 0.87.

	L	$\overline{\mathrm{L}}$	η_n
C	0.91	0.09	0.89
$\overline{\mathrm{C}}$	0.12	0.88	
BK	0.87	0.13	0.93
$\overline{\mathrm{BK}}$	0.02	0.98	
FO	0.70	0.30	0.74
$\overline{\mathrm{FO}}$	0.22	0.78	
UB	0.89	0.11	0.90
$\overline{\mathrm{UB}}$	0.08	0.92	
WP	0.78	0.22	0.87
$\overline{\mathrm{WP}}$	0.04	0.96	

First of all, it becomes obvious, that η_{all} with 0.87 is even higher that the frame-wise η_{all} in Table III. Moreover, we achieved to decrease $\mathrm{FNR_C}$ from 0.15 to 0.09 and to increase $\mathrm{TNR_C}$ from 0.76 to 0.88. This means that the patient will not be confused by detected errors that actually were not present, while at the same time 88 % of the incorrect repetition could be detected. We would like to stress that a main reason for the $\mathrm{FNR_C}$ of 0.09 is the class FO, which is the least accurate of the error classes. This is mainly due to the fact that the localization of the ankle as well as the foot joints of the Kinect were relatively unstable in our recordings.

A closer inspection of Table IV reveals that the $\mathrm{TPR_n}$ of the error classes decreased. However, we obtained higher $\mathrm{TNR_n}$ for these classes. As a consequence, the user does not always get feedback about an occurred error. Since the user will be performing several repetitions, this will not pose a problem, because he or she will see the error in the following repetitions if this error is replicable. But at the same time, the patient will not be getting confused by receiving false alarm due to the very low $\mathrm{FPR_n}$.

Finally, we would like to stress that the results were obtained by using a non-personalized reference. This is a reference of a different person than the patient, which is represented in normalized, hierarchical coordinates as described in [10]. Moreover, the system was trained with normalized, hierarchical data of three other persons. The accuracy would be higher if we would have used the patient's personalized data. We, however, intentionally refrained from using personalized data to reduce the temporal effort for the therapist. The achievement of such a high accuracy for non-personalized data is promising and should encourage further investigations.

V. CONCLUSIONS AND FUTURE WORK

In this paper, we proposed a joint filtering technique to optimize the classifier performances. Based on this, we introduced a post processing algorithm that provides repetition-wise feedback to the patient. To sum up, the obtained results demonstrate that the proposed algorithm provides reliable feedback that points out exercise-specific motion errors to the user, especially in periods when no therapist is present. For future work, we plan to further develop and evaluate the proposed feedback concept. In addition to this, more research is needed to improve the used skeleton model in order to obtain more accurate skeleton data. Moreover, we plan to extend the system with further exercises, such as hip extension and flexion. In this context, we aim at automatizing the exercise integration process.

For the first time, this study demonstrated the detection of motion errors in exercise repetitions using non-personalized data. For this reason, it is a relevant contribution to the field of e-rehabilitation.

REFERENCES

[1] J. Richter, C. Wiede, A. Apitzsch, N. Nitzsche, C. Lösch, M. Weigert, T. Kronfeld, S. Weisleder, and G. Hirtz, "Assisted Motion Control in Therapy Environments Using Smart Sensor Technology: Challenges and Opportunities," in *Ambient Assisted Living, 9. AAL-Kongress, Frankfurt/M, Germany, April 20 - 21, 2016*. Springer Verlag, 2017, pp. 119–132.

[2] N. Gal, D. Andrei, D. I. Neme, E. Ndan, and V. Stoicu-Tivadar, "A Kinect based intelligent e-rehabilitation system in physical therapy," *Digital Healthcare Empowering Europeans*, pp. 489–493, 2015.

[3] T.-C. Huang, Y.-C. Cheng, and C.-C. Chiang, "Automatic Dancing Assessment Using Kinect," in *Advances in Intelligent Systems and Applications-Volume 2*. Springer, 2013, pp. 511–520.

[4] P. Muneesawang, N. M. Khan, M. Kyan, R. B. Elder, N. Dong, G. Sun, H. Li, L. Zhong, and L. Guan, "A machine intelligence approach to virtual ballet training," *IEEE MultiMedia*, vol. 22, no. 4, pp. 80–92, 2015.

[5] T.-Y. Lin, C.-H. Hsieh, and J.-D. Lee, "A kinect-based system for physical rehabilitation: Utilizing tai chi exercises to improve movement disorders in patients with balance ability," in *2013 7th Asia Modelling Symposium*. IEEE, 2013, pp. 149–153.

[6] C.-J. Su, C.-Y. Chiang, and J.-Y. Huang, "Kinect-enabled home-based rehabilitation system using Dynamic Time Warping and fuzzy logic," *Applied Soft Computing*, vol. 22, pp. 652–666, 2014.

[7] N. M. Khan, S. Lin, L. Guan, and B. Guo, "A visual evaluation framework for in-home physical rehabilitation," in *Multimedia (ISM), 2014 IEEE International Symposium on Multimedia*. IEEE, 2014, pp. 237–240.

[8] M. Antunes, R. Baptista, G. Demisse, D. Aouada, and B. Ottersten, "Visual and human-interpretable feedback for assisting physical activity," in *European Conference on Computer Vision*. Springer, 2016, pp. 115–129.

[9] R. Baptista, M. Antunes, D. Aouada, and B. Ottersten, "Video-based feedback for assisting physical activity," in *Proceedings of the 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 5: VISAPP, (VISIGRAPP 2017)*, 2017, pp. 274–280.

[10] J. Richter, C. Wiede, B. Shinde, and G. Hirtz, "Motion Error Classification for Assisted Physical Therapy - A Novel Approach using Incremental Dynamic Time Warping and Normalised Hierarchical Skeleton Joint Data," in *Proceedings of the 6th International Conference on Pattern Recognition Applications and Methods - Volume 1: ICPRAM,*, 2017, pp. 281–288.

[11] J. Shotton, R. Girshick, A. Fitzgibbon, T. Sharp, M. Cook, M. Finocchio, R. Moore, P. Kohli, A. Criminisi, A. Kipman *et al.*, "Efficient human pose estimation from single depth images," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 12, pp. 2821–2840, 2013.

[12] W. Zhao, R. Lun, D. D. Espy, and M. A. Reinthal, "Rule based realtime motion assessment for rehabilitation exercises," in *IEEE Symposium on Computational Intelligence in Healthcare and e-health (CICARE), 2014*. IEEE, 2014, pp. 133–140.

[13] D. Leightley, J. S. McPhee, and M. H. Yap, "Automated analysis and quantification of human mobility using a depth sensor," *IEEE journal of biomedical and health informatics*, vol. 21, no. 4, pp. 939–948, 2017.

[14] M. Capecci, M. G. Ceravolo, F. Ferracuti, S. Iarlori, V. Kyrki, S. Longhi, L. Romeo, and F. Verdini, "Physical rehabilitation exercises assessment based on hidden semi-markov model by kinect v2," in *Biomedical and Health Informatics (BHI), 2016 IEEE-EMBS International Conference on Biomedical and Health Informatics*. IEEE, 2016, pp. 256–259.

[15] Y. Wang, S. Sun, and X. Ding, "A self-adaptive weighted affinity propagation clustering for key frames extraction on human action recognition," *Journal of Visual Communication and Image Representation*, vol. 33, pp. 193–202, 2015.

[16] M. Antunes, D. Aouada, and B. Ottersten, "A revisit to human action recognition from depth sequences: Guided svm-sampling for joint selection," in *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on Applications of Computer Vision*. IEEE, 2016, pp. 1–8.

978-1-5090-4014-8/17/$31.00 ©2017 IEEE