Classification of Movement Quality in a Weight-Shifting Exercise

Vonstad, Elise Klæbo[1], **Su, Xiaomeng**[1], **Vereijken, Beatrix**[2], **Nilsen, Jan Harald**[1], **Bach, Kerstin**[1],
[1] Norwegian University of Science and Technology, Department of Computer Science
[2] Norwegian University of Science and Technology, Department of Neuromedicine and Movement Science

Abstract

In exercise games, it is often possible to gain rewards, i.e. points, by only partly completing an intended movement, which can undermine the effect of using such games for exercise. To ensure usability and reliability of exergames, correct movements must be accurately identified. Aim of the current study was to evaluate performance of machine learning models in classifying weight-shifting movements as correct or incorrect. Eleven healthy elderly (6 F) performed a stepping exercise in a correct (with weight shift) and an incorrect (without weight shift) version. A 3D Motion Capture (3DMoCap) system calculated joint center positions (JCPs); 2270 repetitions (1133 correct) were recorded. Random Forest (RF), k-Nearest Neighbor (k-NN) and Support Vector Machine (SVM) classification models were built. Evaluation: 10fold leave-one-group-out cross validation (CV), repeated for all persons. Results showed high accuracy and recall in all classifiers. Average accuracy and recall was RF = 0.989, k-NN = 0.949, SVM = 0.958. Highest was RF on all JCPs, and SVM on shoulder JCPs (both 0.996). Lowest was k-NN on ankle JCPs (0.879). This study shows that all three models can distinguish correct and incorrect repetitions with high accuracy and recall, also by using selected JCPs. RF consistently outperformed the other models.

1 Introduction

Exercise games, or exergames, are games played on a computer screen that use bodily movements as input to interact with the game. This form of exercising is gaining popularity and attention from both researchers and therapists. In recent years, it has been shown that doing exercises elicited by games is a more motivating and fun way of exercising than conventional exercise programs, while being as effective as conventional exercise when used in cooperation with therapists [Nicholson *et al.*, 2015], [Skjaeret *et al.*, 2016]. This is encouraging with respect to the increasing number of elderly in the population, as we might utilize exergames as a tool to promote self-management of exercise in people of older age. Exergames for elderly might decrease the load on the health care system in the coming years in two ways: by preventing or reducing loss of independence due to reduced physical function, and by empowering elderly to effectively exercise without having to travel to a therapist or training center for supervised exercise. Exergames are fun and motivating partially because they provide additional, extrinsic motivation to complete a movement – points or score in the game. Because people have differences in their body shapes and sizes, the game system needs to accept a wide variety in movements to allow for different players to play the game. This also means that in many situations, it is possible to gain points without doing the complete exercise movement intended, or just doing a small version of the movement, as reported in e.g. [Pasch *et al.*, 2009]. People quickly catch that this is possible: they learn how to cheat. Such incorrectly performed exercise repetitions undermine the effect of exergaming, as it might make the quality of the exercise performed poorer and give lower gains in skill or function than could be expected if the exercise was performed correctly. Apart from being less effective, this can also be dangerous as over-estimation of one's own skill is related to increased fall risk in elderly [Sakurai *et al.*, 2013]. For exergames to be effective and useful, it is vital that they can accurately identify the performance of an exercise repetition as being correct or incorrect. To enable such classification, accurate tracking movement while exergaming is a prerequisite. As the usability and accuracy of different measurement devices varies, finding a trade-off that gives a good enough measurement accuracy while being user friendly is especially challenging. The gold standard for motion capture accuracy, marker-based 3D Motion Capture (e.g. Vicon Motion Systems Ltd) camera systems give very accurate measurements of body movements, but are expensive, require a fixed (laboratory) setting and expert users. Currently, the most promising alternative measurement methods are the marker-less time-of-flight (ToF)/depth camera systems such as the Kinect v2 (Microsoft Inc), and inertial measurement unit (IMU) systems such as the Xsens (Xsens Technologies B.V.). These are easy to use, portable and low-cost, but do not give as accurate full-body measurements as the 3DMoCap systems, especially when measuring hands and feet [van Diest *et al.*, 2014]. ToF camera systems usually utilize a skeleton model based on the 3D cloud mapping of a person to analyze movements, where joint center positions (JCPs) are calculated and used in analyses. Using JCPs, it is possible to represent the person being tracked with enough information to identify different activities [Gaglio *et al.*, 2015], analyze postural stability [Dehbandi *et al.*, 2017] or use the positions as input to a video-based game [Shih *et al.*, 2016]. The ToF based systems show promising results regarding accuracy of measuring torso/upper body movements, as their discrepancy from a 3DMoCap system are reported to be within acceptable ranges [Bonnechère *et al.*, 2014], [Matsen *et al.*, 2016]. Still, others warn about limitations in measurements of shoulder movements when comparing to goniometers [Huber *et al.*, 2015].

The aim of the current study was to assess the performance of ML classifiers. In order to capture the participants' full-body movements as accurately as possible, we used a 3DMoCap system to measure high-quality movement data to ensure that the classification was performed on the actual movements the participants performed. Furthermore, as JCPs is commonly used in more user-friendly measurement devices, we chose to use this as input to the classification model in the current study, possibly allowing insight into whether data from ToF/depth cameras could be used as input to classification models in the future.

As there are several ways to successfully classify the type of movement being performed using machine learning, we hypothesized that it is feasible to use learning algorithms to analyze whole-body movement patterns to classify if a detected movement was performed correctly or not. Thus, this paper aims to investigate the classification performance of three common classification algorithms on JCP 3DMoCap data from a weight-shifting balance task in correct or incorrect performances.

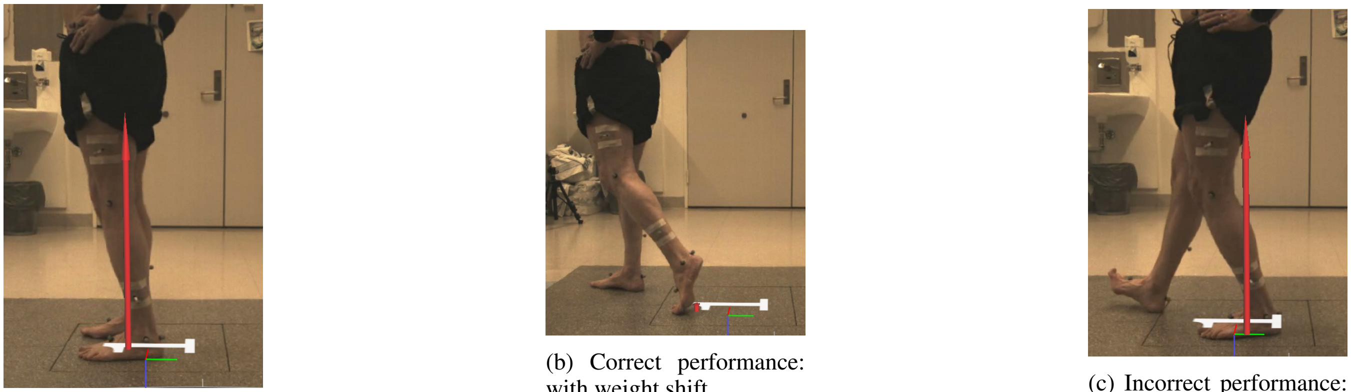
2 Related Work

In movement analysis, machine learning has been used mostly on data from sensors that track persons outside of the lab, as data from e.g. inertial measurement units is challenging to analyze with traditional methods. ML analysis methods have been used in for example activity recognition [Mukhopadhyay, 2014], [Lara and Labrador, 2013], and in identification of falls [Aziz *et al.*, 2012] using data from IMUs. Furthermore, IMUs have been used in classification of movement performance in adults [Giggins *et al.*, 2014], although in this paper it only reached medium-to-good classification accuracy. In [Yurtman and Barshan, 2014] a complete system of movement detection and error classification concerning movement amplitude was implemented using wired IMUs to record movement during physiotherapy exercises, with good results. One study used machine learning to evaluate movement quality in exercises performed by children, using smart-phone IMU sensors to measure movements and using natural fatigue as a mechanism to produce wrong performances [Carvalho and Furtado, 2016]. Lo Presti et al [Lo Presti and La Cascia, 2016] showed a wide range of ML methods being used on identification of human actions using ToF/depth cameras, with good results, however not reporting any studies that aimed to classify the quality of detected movements. The use of ML methods on data from 3DMoCap measurement systems has also increased in recent years, but is mostly used to identify human actions and not to assess the quality of movements. For example, ML models were successfully used to discriminate between f.e. jumping and walking in a continuous stream of MoCap data [Kapsouras and Nikolaidis, 2014]. To our knowledge, research is scarce on automatic classification of movement quality measured using high-quality JCP data obtained from 3DMoCap systems.

3 Approach

3.1 Data set

As there are no open data sets containing labelled weight-shifting balance exercises, we conducted a data collection to obtain a labelled training data set. Collection of time series data was conducted November 2017 using a 10-camera, 100Hz, 3DMoCap system (Vicon Motion Systems Ltd). Simultaneous ground reaction force (GRF) data was collected using a 1000Hz force plate (Kistler Inc) embedded in the floor, and digital video in sagittal view was recorded for quality control purposes. Reflective markers were placed according to the Plug-in-Gait full-body biomechanical model, with head and hand markers excluded. Eleven participants were recruited from local exercise groups for elderly. There were 6 females and 5 males, and mean age was 69.3 years (1SD 4.0). Participants performed two versions of a balance exercise movement common in stroke rehabilitation (as seen in e.g. [Okubo *et al.*, 2016]). Both versions had the same starting position (Figure 1a), with both feet placed on the force plate. The red arrow originating at the feet of the participant represents the 3D ground reaction force (GRF). In the "correct" performance of the movement, the right foot was placed in front of the person, off the force plate, and body weight was shifted over to the right foot while keeping the left foot in contact with the force plate (as seen in Figure 1b, where the remaining GRF on the left foot is small), before moving the right foot back to the force plate. In the "incorrect" version of the movement, the same step was performed, but the person did not shift body weight over to the right foot when they took the step (as seen in figure 1c, where the GRF on the left foot is large). This movement pattern was chosen as they are typical ways of performing this weight-shifting exercise correctly and incorrectly, as described and demonstrated by a physical therapist experienced in stroke rehabilitation. Participants were instructed orally on how to perform these movements with and without weight shift, but were encouraged to move in a way that was natural to them. One repetition was one completion of such a movement: from the moment the person was standing in the starting position, through taking the step, until the person had the right foot back in the starting position. During one trial, 10 repetitions were completed in sequence. Each round of 10 repetitions was performed three times, producing a 3x10 block of repetitions to mimic a normal sequence of exercising. To reduce risk of fatigue from repeating the same movement many times during the test session, test persons first performed two 3x10 block of repetitions in the correct version of the movement, then had a 5-minute break and completed two 3x10 blocks of the incorrect version. This was then repeated so that each person com-


(a) Start and end position

(b) Correct performance: with weight shift

(c) Incorrect performance: without weight shift

Figure 1: a) Shows the start and end position of the movement. b) Shows the correct performance, and c) an example of an incorrect performance.

pleted 240 repetitions in total: 120 repetitions of each version of the movement. Data from 11 persons were collected, with one person only completing half of the test protocol. This resulted in 2520 recorded repetitions.

3.2 Pre-processing and feature extraction

Figure 2 shows the data processing model used to analyze the data. Marker data was first quality checked in the Vicon Nexus software, and missing position data from markers were gap-filled using the built-in algorithms. JCP time series data was extracted from the Plug-in-Gait biomechanical model. Some repetitions were not included due to participants doing a different movement (e.g. loss of balance, side-stepping), or due to partial capture of repetitions at the beginning or end of a trial. This resulted in JCP time series data from 2270 repetitions being included for further analysis, 1133 correct and 1137 incorrect. Statistical features from each JCP time series were computed: these included mean, median, standard deviation, sum, variance, minimum and maximum values.

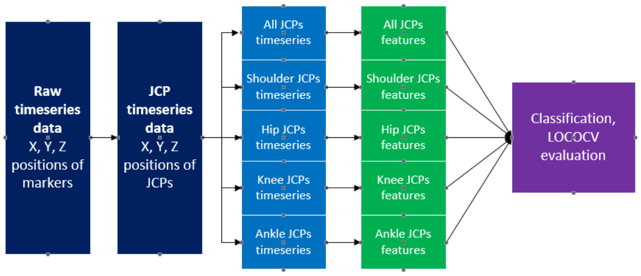


Figure 2: Data flow model

3.3 Test-train-split

Using the SciKit-Learn library [Pedregosa *et al.*, 2012], the data was split into training and test sets, where the Leave-One-Group-Out Cross-Validation (LOGOCV) method was used to exclude data from one person and use as the test set in each iteration. This is a suitable method in the exercise domain, where it is likely that a model would be trained on other people's data than data from the current player being evaluated for correct/incorrect repetitions.

3.4 Classification models

A random forest (RF, n estimators: 10) classifier, a k-nearest neighbor (k-NN, k = 10) classifier and a support vector machine (SVM, kernel = polynomial) classifier were trained and tested, using the SciKit-Learn library, in each iteration of the train-test-split. Hyperparameters were not tuned due to the success of the initial parameter settings. Results were obtained as confusion matrices, where accuracy and recall were reported. Recall was chosen as a primary outcome measure as it is vital in this setting, aside from overall accuracy.

4 Results

Table 1 shows average accuracy from all LOGOCV iterations for classification of incorrect and correct repetitions by the three classifiers. Overall, results show that all three classification models achieve very high accuracy of around 95 % in almost all classifications. The RF and SVM models achieved the highest accuracies, with 99.6 % on shoulder JCPs and all JCPs, respectively. Lowest accuracy was reached by the k-NN model on data from ankle JCP, 87.9 %. Recall results (Figure 3 & 4) showed that all three models achieved largely more than 90 % accuracy in both correct and incorrect repetitions. Figure 3 shows recall for correct repetitions by all classifiers, in each of the JCP selections. RF consistently achieved >95 % recall, being the most consistent in the different JCP selections of the three models. Average recall of correct repetitions was 98.9 % for RF, 94.4 % in k-NN and 96.0 % in SVM. The SVM model performed best of the three on recall of correct repetitions on data from all JCPs, but also had the most variable performance in the other JCP selections. K-NN reached around 95 % on all JCP selections except in ankle JCPs, where it was the overall worst performing model of the three. Figure 4 shows recall accuracy for incorrect repetitions by all classifiers, in each of the JCP selections. Again, RF is most consistent with an average of

99.0 %, while k-NN and SVM achieved 95.2 % and 95.6 %, respectively. k-NN had the lowest recall of all models in all JCPs for incorrect repetitions, with 85.8 % in data from ankle JCPs. All three models had the highest recall when using data from all JCPs, although recall from using JCP selections, especially shoulder JCPs, was also high.

	Random Forest	k-NN	SVM	Avg
All	99.0 %	96.8 %	99.6 %	98.5 %
SHO	99.6 %	96.4 %	96.2 %	97.4 %
HIP	99.2 %	96.8 %	92.1 %	96.0 %
KNE	97.5 %	96.6 %	94.1 %	96.1 %
ANK	99.3 %	87.9 %	96.8 %	94.7 %
Avg	98.9 %	94.9 %	95.8 %	96.5 %

Table 1: Accuracy of classifiers for the different joint centre positions.

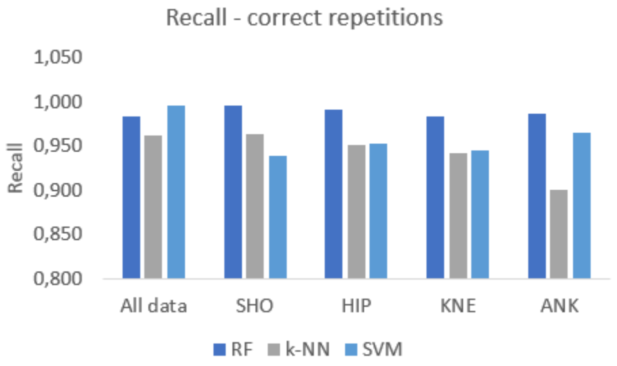


Figure 3: Recall for correct repetitions by all classifiers on all JCPs, shoulder (SHO), hip (HIP), knee (KNE) and ankle (ANK) JCPs.

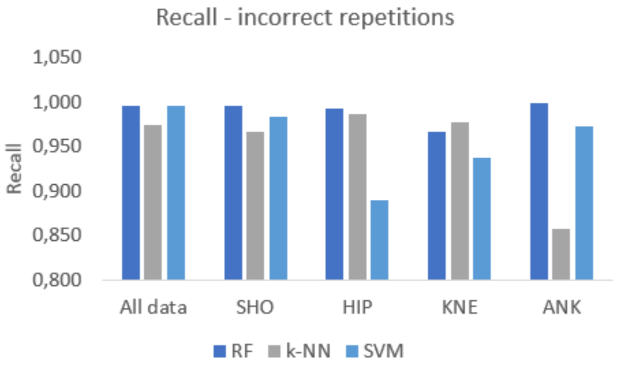


Figure 4: Recall for incorrect repetitions by all classifiers on all JCPs, shoulder (SHO), hip (HIP), knee (KNE) and ankle (ANK) JCPs.

5 Discussion

This paper aimed to evaluate the performance of three ML classification models in classifying correctly and incorrectly performed repetitions of a weight-shifting exercise, using JCPs measured with a 3DMoCap system. Performance of Random Forest, K-Nearest Neighbor and a Support Vector Machine was evaluated. Results indicated that all three models are able to distinguish between incorrect and correct repetitions with high accuracy and recall (with an average accuracy of 98.9 %, 94.9 % and 95.5 %, respectively). Results from the current study are similar to those seen in [Gaglio *et al.*, 2015] and in [Liu *et al.*, 2017], where novel methods were used to classify activities using JCPs from Kinect, outperforming other approaches on the same data set. However, these results are not directly comparable to results in the current study, as the mentioned studies are not concerned with movement quality but with movement type. Compared to other studies on movement quality (e.g. [Giggins *et al.*, 2014], [Yurtman and Barshan, 2014]), which are based on data from IMUs, the achieved accuracy in the current study is higher. This is possibly an effect of the movements in this study being instructed, and that the movements in these other studies are more complex and varied. Also, the IMU data might not represent the movements as accurately as the 3DMoCap data does. Using all JCPs in the classification reached marginally higher accuracy than using any of the JCP selections, as seen in Table 1. The RF model was consistently slightly more accurate than the other two models, for both accuracy and recall. In light of the issue of avoiding in-game rewards for incorrect performance, recall of incorrect repetitions is a vital score here. The RF model achieved >95 % recall in all JCP selections. The k-NN and SVM models also achieved high recall, but were not as consistent in JCP selections as the RF model. Other studies using JCPs typically use all joints, or only joints that are tracked with good accuracy during the whole capture, as seen in [Gaglio *et al.*, 2015]. Therefore, the results from classification of movement quality using JCP selections in the current study might not be comparable to results from selected JCPs in other studies. Results also reflect that the data from incorrect and correct repetitions were very different, as all three models accurately distinguished between them. The oral instructions might have contributed to this, as the instructions probably influenced the movement patterns. Spontaneous, natural movements might be more variable than what was seen in this data set. Also, the correct movements were performed with more upper-body movement towards the stepping foot, and the heel of the stance foot was also lifted from the force plate. Furthermore, data from only the ankle JCPs were also classified with >80 % accuracy and recall by all models, which was not expected as both movements include similar stepping movements in the feet. The movements of the feet alone were different enough in the correct and incorrect repetitions to enable accurate classification, which might be a result of the aforementioned heel-lifts seen in only the correct trials. This probably resulted in more variable JCP's during correct repetitions, enabling the ML models to accurately identify them. Using ML-models for the purpose of evaluating movement quality using data from ToF/depth cameras seems feasible given the very good performance achieved here. Furthermore, the good performance achieved in this study indicates that the models possibly can reach acceptable accuracy and recall also with lower-quality data. This can facilitate implementation of ML models into more user-friendly exergaming contexts. Recall results in classification of both correct and incorrect repetitions are very encouraging for applying ML in analysis of movements during exergaming, as this could make it harder for the player to receive rewards without performing the intended movement correctly. However, as the current movements were not elicited by an actual exergame, it remains to be determined whether a similar level of accuracy can be achieved in more realistic exergaming movements. Furthermore, the high accuracy in all JCP selections suggests that it might be feasible to use only the more accurate measurements of shoulder or hip JCPs from using ToF/depth cameras, and still accurately identify correct and incorrect repetitions of a weight-shifting exercise. This could provide a way of using ML in exergames to more accurately reward movements during play, thus ensuring movement quality to a greater extent than the existing systems do. Future work will focus on the use of ML models in actual exergame situations, as this possibly elicits movements that are noisier than in the current study, hence making the repetitions difficult to classify as being incorrect or correct. Using motion capture systems with lower accuracy, and only using e.g. shoulder JCPs as input to the classification models would also be interesting to test in an actual exergaming setting, to see if the movements are still different enough to be classified as being correctly or incorrectly performed with similar accuracy to this study.

6 Conclusion

In order to use exergames effectively as a training and rehabilitation tool, it is crucial that the exergame system can identify correct and incorrect exercise repetitions accurately. This paper shows that it is feasible to use ML models in the automatic classification of correctly and incorrectly performed weight-shifts in balance exercises. Applying ML models on high-quality JCP movement data from a weight-shifting exercise yielded accurate classification of correct and incorrect exercise repetitions. Results encourage the testing of such models on JCP data obtained while elderly are playing actual exergames, to investigate whether the models are equally accurate in a more natural and possibly noisier setting. However, this was done in a setting where the performance of repetitions was instructed, and the movements performed (for example the movement pattern of an incorrectly performed repetition) might differ from the movements performed here. The study also shows that using only selected JCPs yields accurate results as well, which is promising with regard to possible use of ML models on data from data capture methods that are lower cost and more user friendly.

References

[Aziz *et al.*, 2012] O. Aziz, E. J Park, G. Mori, and S. N Robinovitch. Distinguishing near-falls from daily activities with wearable accelerometers and gyroscopes using Support Vector Machines. *Conference proceedings: IEEE Engineering in Medicine and Biology Society. Annual Conference*, 2012:5837–5840, 2012.

[Bonnechère *et al.*, 2014] B. Bonnechère, B. Jansen, P. Salvia, H. Bouzahouene, L. Omelina, F. Moiseev, V. Sholukha, J. Cornelis, M. Rooze, and S. Van Sint Jan. Validity and reliability of the Kinect within functional assessment activities: Comparison with standard stereophotogrammetry. *Gait and Posture*, 39(1):593–598, 2014.

[Carvalho and Furtado, 2016] L. D. Carvalho and V. Furtado. Using machine learning for evaluating the quality of exercises in a mobile exergame for tackling obesity in children. *Proceedings of SAI Intelligent Systems Conference (IntelliSys)*, 15, 2016.

[Dehbandi *et al.*, 2017] B. Dehbandi, A. Barachant, A. H Smeragliuolo, J. D. Long, S. J. Bumanlag, V. He, A. Lampe, and D. Putrino. Using data from the Microsoft Kinect 2 to determine postural stability in healthy subjects: A feasibility trial. *PloS one*, 12(2):e0170890, 2017.

[Gaglio *et al.*, 2015] S. Gaglio, G. Lo Re, and M. Morana. Human Activity Recognition Process Using 3-D Posture Data. *IEEE Transactions on Human-Machine Systems*, 45(5):586–597, 2015.

[Giggins *et al.*, 2014] O. M Giggins, K. T. Sweeney, and B. Caulfield. Rehabilitation exercise assessment using inertial sensors: a cross-sectional analytical study. *Journal of NeuroEngineering and Rehabilitation*, pages 1–10, 2014.

[Huber *et al.*, 2015] M. E. Huber, A. L. Seitz, M. Leeser, and D. Sternad. Validity and reliability of Kinect skeleton for measuring shoulder joint angles: A feasibility study. *Physiotherapy (United Kingdom)*, 101(4):389–393, 2015.

[Kapsouras and Nikolaidis, 2014] I. Kapsouras and N. Nikolaidis. Action recognition on motion capture data using a dynemes and forward difference representation. *Proceedings - International Conference on Pattern Recognition*, 25:2649–2654, 2014.

[Lara and Labrador, 2013] Oscar D. Lara and Miguel A. Labrador. A Survey on Human Activity Recognition using Wearable Sensors. *IEEE Communications Surveys & Tutorials*, 15(3):1192–1209, 2013.

[Liu *et al.*, 2017] Jun Liu, Amir Shahroudy, Dong Xu, Alex Kot Chichung, and Gang Wang. Skeleton-Based Action Recognition Using Spatio-Temporal LSTM Network with Trust Gates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[Lo Presti and La Cascia, 2016] Liliana Lo Presti and Marco La Cascia. 3D skeleton-based human action classification: A survey. *Pattern Recognition*, 53:130–147, 2016.

[Matsen *et al.*, 2016] F. A. Matsen, Al. Lauder, K. Rector, P. Keeling, and A. L. Cherones. Measurement of active shoulder motion using the Kinect, a commercially available infrared position detection system. *Journal of Shoulder and Elbow Surgery*, 25(2):216–223, 2016.

[Mukhopadhyay, 2014] S C Mukhopadhyay. Wearable sensors for human activity monitoring: A review. *IEEE Sensors Journal*, 15(3):1321–1330, 2014.

[Nicholson *et al.*, 2015] V. P. Nicholson, M. McKean, J. Lowe, C. Fawcett, and B. Burkett. Six weeks of unsupervised Nintendo Wii Fit gaming is effective at improving balance in independent older adults. *Journal of Aging and Physical Activity*, 23(1):153–158, 2015.

[Okubo *et al.*, 2016] Y. Okubo, D. Schoene, and S. R Lord. Step training improves reaction time, gait and balance and reduces falls in older people: a systematic review and meta-analysis. *British Journal of Sports Medicine*, 2016.

[Pasch *et al.*, 2009] Marco Pasch, Nadia Bianchi-Berthouze, Betsy van Dijk, and Anton Nijholt. Movement-based sports video games: Investigating motivation and gaming experience. *Entertainment Computing*, 1(2):49–61, 2009.

[Pedregosa *et al.*, 2012] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2012.

[Sakurai *et al.*, 2013] R. Sakurai, Y. Fujiwara, M. Ishihara, T. Higuchi, H. Uchida, and K. Imanaka. Age-related self-overestimation of step-over ability in healthy older adults and its relationship to fall risk. *BMC Geriatrics*, 13(1):15–17, 2013.

[Shih *et al.*, 2016] Meng Che Shih, Ray Yau Wang, Shih Jung Cheng, and Yea Ru Yang. Effects of a balance-based exergaming intervention using the Kinect sensor on posture stability in individuals with Parkinson's disease: A single-blinded randomized controlled trial. *Journal of NeuroEngineering and Rehabilitation*, 13(1):1–9, 2016.

[Skjaeret *et al.*, 2016] Nina Skjaeret, Ather Nawaz, Tobias Morat, Daniel Schoene, Jorunn Laegdheim, and Beatrix Vereijken. Exercise and rehabilitation delivered through exergames in older adults : An integrative review of technologies, safety and efficacy. *International Journal of Medical Informatics*, 85(1):1–16, 2016.

[van Diest *et al.*, 2014] Mike van Diest, Jan Stegenga, Heinrich J. Wörtche, Klaas Postema, Gijsbertus J. Verkerke, and Claudine J.C. Lamoth. Suitability of Kinect for measuring whole body movement patterns during exergaming. *Journal of Biomechanics*, 47(12):2925–2932, 2014.

[Yurtman and Barshan, 2014] A. Yurtman and B. Barshan. Automated evaluation of physical therapy exercises using multi-template dynamic time warping on wearable sensor signals. *Computer Methods and Programs in Biomedicine*, 117(2):189–207, 2014.