Gait & Posture 76 (2020) 198–203

Contents lists available at ScienceDirect

Gait & Posture

journal homepage: www.elsevier.com/locate/gaitpost

Input representations and classification strategies for automated human gait analysis

Djordje Slijepcevic[a,*], Matthias Zeppelzauer[a], Caterine Schwab[b], Anna-Maria Raberger[b], Christian Breiteneder[c], Brian Horsak[b]

[a] St. Pölten University of Applied Sciences, Institute for Creative Media Technologies, St. Pölten, Austria
[b] St. Pölten University of Applied Sciences, Institute of Health Sciences, St. Pölten, Austria
[c] TU Wien, Institute of Visual Computing and Human-Centered Technology, Vienna, Austria

ARTICLE INFO

Keywords:
Ground reaction force
Gait classification
Machine learning
Gait disorders
Support vector machine

ABSTRACT

Background: Quantitative gait analysis produces a vast amount of data, which can be difficult to analyze. Automated gait classification based on machine learning techniques bear the potential to support clinicians in comprehending these complex data. Even though these techniques are already frequently used in the scientific community, there is no clear consensus on how the data need to be preprocessed and arranged to assure optimal classification accuracy outcomes.
Research question: Is there an optimal data aggregation and preprocessing workflow to optimize classification accuracy outcomes?
Methods: Based on our previous work on automated classification of ground reaction force (GRF) data, a sequential setup was followed: firstly, several aggregation methods – early fusion and late fusion – were compared, and secondly, based on the best aggregation method identified, the expressiveness of different combinations of signal representations was investigated. The employed dataset included data from 910 subjects, with four gait disorder classes and one healthy control group. The machine learning pipeline comprised principle component analysis (PCA), *z*-standardization and a support vector machine (SVM).
Results: The late fusion aggregation, i.e., utilizing majority voting on the classifier's predictions, performed best. In addition, the use of derived signal representations (relative changes and signal differences) seems to be advantageous as well.
Significance: Our results indicate that great caution is needed when data preprocessing and aggregation methods are selected, as these can have an impact on classification accuracies. These results shall serve future studies as a guideline for the choice of data aggregation and preprocessing techniques to be employed.

1. Introduction

Gait disorders can affect anyone, regardless of age, and often impede an individual's ability to participate in daily living activities such as walking and might even reduce movement efficiency in terms of energy consumption [1,2]. Gait analysis based on ground reaction force (GRF) assessment is a well-established method to diagnose the mechanisms that underlie gait disorders. The quantitative analysis of such data can provide relevant information for clinicians in diagnosing gait impairments, planning therapies and surgeries, supporting rehabilitation processes, or evaluating treatment outcomes [3]. However, quantitative gait analysis produces a vast amount of data, which are difficult to comprehend and analyze due to their high-dimensionality, temporal dependencies, strong variability, non-linear relationships, and intercorrelations [4]. Therefore, there is growing interest in employing machine learning techniques that allow for a cost-effective, fast and objective analysis of large amounts of gait measurements. Recently, automated gait classification has been successfully used for various patient groups [5] affected by stroke [6], Parkinson's disease [7], cerebral palsy [8], multiple sclerosis [9], osteoarthritis [10], or by age-related impairments [11].

Automated classification of gait is, however, a complex task consisting of many different processing steps which have to be carried out in a methodically correct way and for which various approaches exist. According to Figueiredo et al. [5] gait pattern recognition comprises the following main steps: (1) feature extraction, (2) feature normalization,

* Corresponding author at: St. Pölten University of Applied Sciences, Matthias Corvinus-Straße 15, 3100 St. Pölten, Austria.
E-mail address: djordje.slijepcevic@fhstp.ac.at (D. Slijepcevic).

https://doi.org/10.1016/j.gaitpost.2019.10.021
Received 22 December 2018; Received in revised form 8 October 2019; Accepted 14 October 2019

(3) feature selection, (4) forming a training and a testing dataset, (5) training a classification model, and (6) evaluating the performance. To date, there is no clear consensus on how to proceed in each of these steps. For tasks (2) to (6), the systematic review by Figueiredo et al. [5] might serve as a first guideline. For the first step of feature extraction, a variety of options can be found in the current literature, but so far no clear recommendation can be derived. However, different approaches in feature extraction might significantly effect classification accuracies.

Firstly, in the literature on gait classification, several recorded trials of a subject are usually either averaged to a single waveform [12–14] or all available trials are provided to a classifier [15,16,10]. To date, it is unclear which of these data aggregation strategies serves best for gait classification. Recently, a statistical method based on the notion of depth was suggested which identifies the most representative trial [17]. This approach, however, has not been employed in the gait classification community yet.

Secondly, there is no clear consensus on how the raw signals should be preprocessed and transformed to form an appropriate input feature vector for the machine learning algorithm. Regarding the available input data (ground reaction force (GRF) and center of pressure (COP) components), it is still unclear which form of representation (i.e. raw data, relative changes, or signal differences) is best suited for machine learning. Based on our earlier work [18] the two primary aims of this article are: to (i) evaluate the effects of different data aggregation methods on gait classification performance and to (ii) investigate which input representations and combinations of representations perform best for automated gait classification. To facilitate the comparability of machine learning approaches and to optimize performance, it is critical to identify best practice procedures for the individual steps of gait classification. The results of this article shall serve future studies as a guideline on machine learning for gait analysis.

2. Methods

2.1. Patients and dataset

The anonymized data used in this study are part of an existing clinical gait database maintained by a rehabilitation center of the Austrian Workers' Compensation Board (AUVA). The AUVA is the social insurance for occupational risks for more than 3.3 million employees and 1.4 million pupils and students in Austria. This retrospective study was approved by the local Ethics Committee of Lower Austria (GS1-EK-4/299-2014).

The dataset utilized comprises GRF measurements from 728 patients with gait disorders (GD) and data from 182 healthy controls, both of various physical composition and gender (see Table 1).The dataset is balanced regarding the number of persons per class, the number of recorded sessions per person and the number of trials per person. The dataset includes gait disorders associated with the calcaneus (n = 182), ankle (n = 182), knee (n = 182), and hip (n = 182). A well-experienced physical therapist (with more than a decade of clinical experience) has manually labeled the dataset based on the available medical diagnosis of each patient. The individual GD classes include patients after joint replacement surgery, fractures, ligament ruptures, and related disorders associated with the above-mentioned anatomical areas. The most common injuries present in the hip class are fractures of the pelvis and thigh as well as luxation of the hip joint, coxarthrosis, and total hip replacement. The knee class comprises patients after patella, femur or tibia fractures, ruptures of the cruciate or collateral ligaments or the meniscus, and total knee replacements. The ankle class includes patients after fractures of the malleoli, talus, tibia or lower leg, and ruptures of ligaments or the Achilles tendon. The calcaneus class comprises patients after calcaneus fractures or ankle fusion surgery. All of the injuries mentioned above may occur individually or in combinations within each class.

Table 1
Details on the dataset employed, the demography of the participants and the pre-defined classes.

		Age (yrs.)	Body Mass (kg)	Sex	
Class	n	Mean (SD)	Mean (SD)	(m/f)	Num. trials
Healthy controls	182	34.3 (14.0)	74.6 (15.8)	94/88	1,456
Calcaneus	182	44.3 (10.5)	86.3 (16.4)	167/15	1,456
Ankle	182	40.6 (10.9)	88.3 (18.2)	151/31	1,456
Knee	182	40.4 (12.3)	86.2 (20.3)	133/49	1,456
Hip	182	40.6 (12.8)	81.5 (15.0)	153/29	1,456
Total	**910**	**40.0 (12.1)**	**83.4 (17.1)**	**698/212**	**7,280**

2.2. Data recording and preprocessing

Gait analysis was performed on a 10 m walkway with two centrally embedded force plates (Kistler, Type 9281B12). The force plates were placed in consecutive order, allowing a person to walk across by placing one foot on each plate. Both plates were flush with the ground and covered with the same walkway surface material, so that targeting was not an issue. During a session, participants walked unassisted and without walking aid at self-selected walking speed until a minimum of eight valid recordings were available.

All processing steps and subsequent analyses were performed in Matlab 2017b (The MathWorks Inc., Natick, MA, USA). The three analog GRF signals, as well as the two COP signals, were converted to digital signals using a sampling rate of 2000 Hz and a 12-bit analog-digital converter (DT3010, Data Translation Incorporation, Marlboro, MA, USA) with a signal input range of ± 10 V. A threshold of 10 N was used for step detection and 30 N for COP calculation. Raw signals were filtered using a 2nd order low-pass Butterworth filter with a cut-off frequency of 20 Hz. All gait measurements were time-normalized to 1000 points (100% stance). Amplitude values of the three force components, i.e., vertical (V), medio-lateral (ML), and anterior–posterior (AP), were expressed as a multiple of body weight by dividing the force by the product of body mass times acceleration due to gravity.

2.3. Gait classification

The present paper builds upon the general gait classification pipeline established by Slijepcevic et al. [18] and uses it as a baseline for the performed experiments. A schematic illustration of the pipeline is shown in Fig. 1. In a first step, Principal Component Analysis (PCA) is applied to the raw input data, i.e. to each input representation separately (feature extraction).[1] Next, the resulting features, i.e., the principal components that retain 98% of the overall variance in the input data, are concatenated and z-standardized (feature normalization). The features are provided to a classifier which is trained and evaluated in a cross-validation manner. For the best parameters found during cross-validation the model is trained on the entire training set. To account for generalizability of this model we evaluated it on a completely independent and unseen dataset (see Figure S1 in the supplementary material). As demonstrated in [18], Support Vector Machines (SVM) are a suitable classifier for gait data outperforming several competitors, e.g., multi-layer perceptrons and the k-nearest neighbors algorithm. The SVM is trained in a multi-class fashion using a one-vs-one strategy.

2.3.1. Data aggregation methods

Usually, several trials per person are recorded during gait analysis. Thus, the question arises whether and how the information from these different trials can be aggregated. Such an aggregation step could be

[1] For each original input signal and derived representation a PCA is performed on a matrix of size 334 (samples) × t (trials), where t depends on the considered dataset.

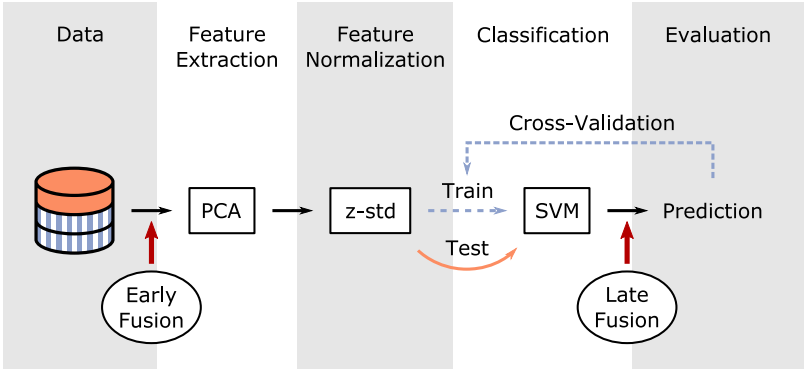

Fig. 1. Illustration of the employed gait classification framework. The dataset consisted of a training set (blue, dashed) and an independent test set (orange, solid). The latter was used to evaluate the generalizability of our classification.

implemented in an *early fusion* or a *late fusion* manner (see Fig. 1). The former directly affects the input data and thus precedes the feature extraction step, whereas the latter is directly applied to the classifier's predictions and affects mostly the classification scheme. Popular early fusion approaches include: (i) mean waveform, (ii) median waveform, and (iii) the most representative trial.

The *mean waveform* approach consists of averaging each measurement from a session (in this case, eight trials) pointwise. The resulting waveform should result in a more robust representation than the original signals by removing inter-trial variations and retaining the overall characteristic shape. The *median waveform* approach is similar but utilizes the point-wise median instead. It is more robust to outliers but may generate less smooth waveforms than the mean waveform approach. Both approaches could diminish informative waveform characteristics, or even cause artifacts that provide a distorted representation [19]. To overcome this problem, Sangeux et al. [17] proposed a statistical method to determine the *most representative* trial. Thereby, this approach assures that original measurement data is used. For machine learning, however, performance might be affected by the fact that not all available and potentially essential information is considered. A schematic illustration of the early fusion approaches is given in Fig. 2.

The late fusion approach utilizes all available original trials for the training of the model. As a result, the classifier returns one prediction per trial. These predictions are considered weak because they are based on individual measurements. The late fusion approach combines these weak predictions into a strong prediction. A robust approach for the combination of several predictions is majority voting. The majority vote is calculated based on the statistical mode, which returns the element (class label) that occurs most often in a set of predictions. For majority voting, only predictions with a likelihood of more than 40% for one of the five classes are used. Thereby, the negative influence of ambiguous trials is reduced. A schematic illustration of the late fusion approach is presented in Fig. 2.

To provide a baseline without aggregation of the available data, we employ all eight trials per person individually during the training and testing. Thus, each trial was predicted separately, and the information about the membership of the trial to a specific person was not utilized.

2.3.2. Input representations

We further investigate the expressiveness and suitability of different input representations for gait classification. Two different types of input representations are distinguished here: original *input signals* and *derived signals*.

Original input signals comprise the time and body weight

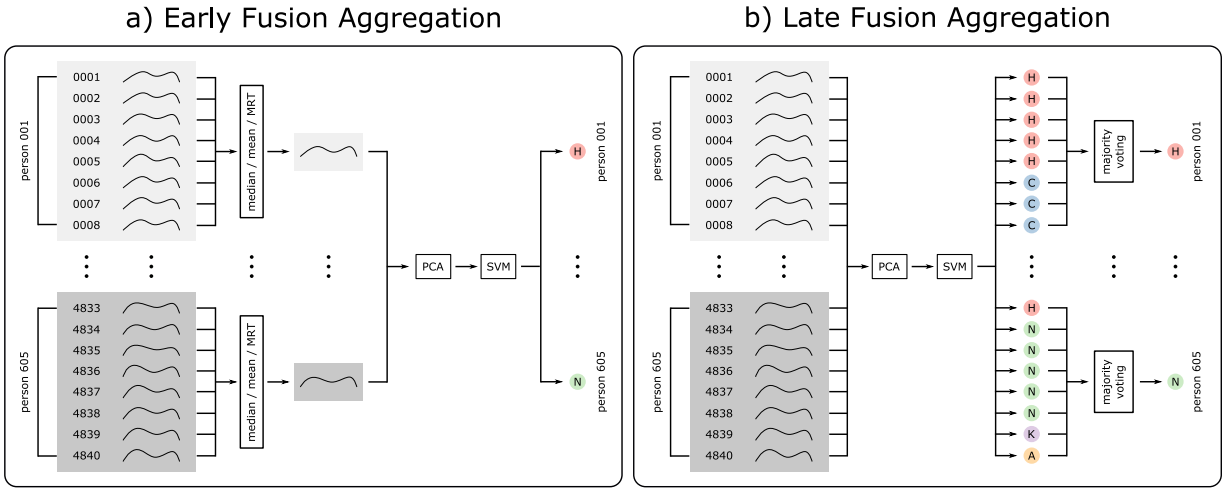

Fig. 2. (a) Schematic for the early fusion aggregation, i.e., mean, median, and most representative trial (MRT) approaches. Prior to training, the eight signals of one subject are aggregated by calculating a mean or median waveform, respectively or one trial is selected by MRT. (b) Schematic for the late fusion aggregation, which employs majority voting. For the training of an SVM, all recorded trials of the subjects are used. For the actual prediction of the test set, majority voting is applied to obtain a decision at subject level.

Table 2
Classification results (%) of the experiment investigating different aggregation methods over several trials (RB: 20%). Highest achieved results are highlighted bold.

Trial selection	Five-fold cross-validation on training set				Independent test set			
	Acc	P	R	F1	Acc	P	R	F1
Baseline without aggregation	52.0 (2.0)	51.8 (1.8)	52.7 (2.6)	51.4 (1.7)	56.5 (2.3)	56.2 (2.7)	56.6 (2.3)	56.0 (2.6)
Mean waveform approach	53.9 (5.2)	53.5 (6.5)	54.4 (5.9)	52.7 (5.9)	58.9 (2.8)	59.5 (1.0)	58.6 (2.2)	58.5 (1.8)
Median waveform approach	51.4 (3.3)	51.1 (3.7)	52.0 (3.7)	50.3 (3.8)	56.7 (3.5)	57.9 (2.7)	56.9 (2.9)	56.4 (2.6)
Most representative trial (MRT)	50.1 (2.0)	50.4 (2.1)	50.7 (2.4)	49.0 (2.0)	56.9 (5.3)	57.7 (4.6)	57.0 (4.3)	56.5 (4.8)
Majority voting	**55.5 (2.4)**	**54.4 (2.5)**	**56.6 (2.3)**	**54.3 (2.5)**	**61.0 (2.4)**	**60.9 (2.9)**	**61.1 (2.4)**	**60.1 (2.7)**

normalized waveforms, i.e., F_V, F_{AP}, F_{ML}, COP_{AP}, and COP_{ML} components of the affected (A) and unaffected (U) lower extremity. The affected and unaffected body side were defined by the physical therapist during data annotation. In case of healthy controls or bi-laterally affected patients the affected side was chosen randomly to avoid a bias.

The derived signal representations are calculated based on the original input signals. Two types of derived signals are investigated: the approximate first derivative (D_A,D_U) of each original input signal and the absolute difference between the input signals of the affected and unaffected lower extremity (Δ).

Furthermore, the expressive power of different combinations of the individual signal representations is examined, i.e., the combination of the original input signals and the derived representations of the affected and unaffected sides.

2.4. Experimental setup

Prior to the experiments, the dataset was randomly divided into a training set (65%) and an independent test set (35%), see Fig. 1. This split remained unchanged for all experiments. The classification experiments utilized a probabilistic SVM with a linear kernel (provided by the LIBSVM library [20]). For hyper-parameter selection, a grid search over the regularization parameter $C \in [2^{-5};2^{10}]$ was employed. During the grid search, a five-fold cross-validation was performed on the training set. After hyper-parameter selection an SVM with the best parameters was trained on the entire training set. To assess the generalizability of the methods, the test set was divided into three equally large and balanced test splits, on which we evaluated the SVM. By using multiple splits, it was possible to estimate not only the generalization ability but also the expected variation in performance for different subsets of test samples. The evaluation was conducted by calculating four performance measures, i.e. classification accuracy (Acc), precision (P), recall (R), and F1-score (F1), defined in terms of number of true positives (*TP*), true negatives (*TN*), false positives (*FP*), and false negatives (*FN*) as follows:

$$\mathrm{Acc} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \quad R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \quad F1 = 2 \times \frac{P \times R}{P + R}$$

Furthermore, a sequential setup was followed: first, different aggregation methods were examined, and second, based on the best aggregation method the expressiveness of different (combinations of) signal representations was investigated. All results are reported as mean (SD), unless otherwise stated.

3. Results

The results of the first experiment investigating different aggregation methods over several trials are summarized in Table 2. The performances of the five-fold cross-validation and the evaluation on the independent test set showed similar trends. This demonstrates the generalization ability of our method. In the following, we discuss the results of the independent test set, which are more objective than the results on the training set. The baseline approach, where all available trials are employed (without aggregation), yielded an accuracy of 56.5% (2.3) (RB[2] : 20%). The use of the median waveform and MRT did not outperform the baseline performance. Within the group of early fusion approaches the mean waveform approach showed the greatest improvement with an accuracy of 58.9% (2.8). The late fusion approach, i.e., majority voting, achieved the highest absolute scores in all performance measures (although not statistically significant in this experiment).

The results of the second experiment in which we investigated the expressiveness of different signal representations on the independent test set are presented in Table 3 (further performance measures can be found in the supplementary material). The results obtained during the five-fold cross-validation follow a trend similar to that of the independent test set and are presented in the supplementary material. The first column in Table 3 indicates which components were used in each experiment: (1) each input signal separately (first five rows), (2) the combination of all three GRF components (row 6), (3) the combination of both COP components (row 7), and (4) the combination of all signals (GRF + COP, last row). For each of these selections, the columns show which (combinations of) derived representations were employed for both affected (A) and unaffected (U) sides. Most notably, column {A, Δ, D_A} shows the highest performance for most input configurations (in six of the eight rows), including also the overall best result with a classification accuracy of 62% (GRF + COP). For the F_{ML} component (row three), combination {A, D_A} provides the best result. The combination {A, D_A, U, D_U} provides the best results for the COP_{ML} component (row four) and the combination of both COP components (row seven). The comparison between the individual GRF and COP components (first five rows) and the three combinations, GRF, COP, and GRF + COP (last three rows) indicates that the combination of all components (last row) performed best. Furthermore, incorporating the information from both legs (via Δ) as well as using the first derivative (in particular of the affected leg) shows to be beneficial.

4. Discussion

From the first experiment (see Table 2) we observe that achieved performances of all approaches are higher for the test set than for the training set. The reason for this is that for experiments on the test set the SVM was trained on the entire training data with the optimal parameters determined during cross-validation and grid search. The improved results on the test set show that additional training data are beneficial for the classifier. Furthermore, the first experiment indicates that the inclusion of membership information can be beneficial. Two aggregation methods, i.e. the mean waveform approach and majority

[2] RB refers to the analytical "random baseline" and represents the theoretical accuracy obtained when assigning class labels randomly, i.e. the case where nothing is learned from the data. For a balanced dataset the analytical RB is the reciprocal of the number of classes, i.e. 20% in our case. The empirically estimated RB according to [21] which further takes the sample size into account is approximately 26% in our case. Every increase over the RB means that the underlying model has learned something from the data.

Table 3
Classification accuracies (%) for different combinations of input signals and derived representations (RB: 20%). Highest achieved results are highlighted bold.

Signals	{A}	{U}	{A,U}	{A,Δ}	{U,Δ}	{A,D_A}	{U,D_U}	{A,D_A,U,D_U}	{A,Δ,D_A}	{U,Δ,D_U}
F_V	42.6	38.7	47.2	44.9	47.5	47.5	37.1	46.6	**48.9**	44.3
F_{AP}	44.3	40.7	45.6	42.0	42.3	42.6	40.3	44.3	**46.6**	45.3
F_{ML}	44.3	32.5	44.6	43.3	34.8	**45.6**	37.7	44.6	44.3	38.4
COP_{ML}	28.2	26.9	31.2	26.6	25.3	43.6	34.1	**44.9**	44.6	35.4
COP_{AP}	36.4	26.9	35.1	40.0	33.1	45.3	30.8	45.3	**46.2**	35.1
GRF	56.7	45.6	54.4	55.7	46.9	55.1	45.6	55.4	**60.0**	48.2
COP	37.1	30.8	43.0	41.6	32.1	48.2	34.1	**52.8**	51.8	36.1
GRF + COP	*61.0*	*47.2*	*58.7*	*60.3*	*49.8*	*59.3*	*49.8*	*61.3*	***62.0***	*51.2*

voting, achieved an improvement compared to the baseline where no aggregation was performed. Specifically, the late fusion approach, i.e., majority voting, achieved better results in absolute scores than the early fusion approaches.

To evaluate the robustness of our approach in more detail, we repeated the experiment with 10 different (randomly selected) train-test splits. The results are presented in the supplementary material in Table S1 due to space limitations. For all 10 repetitions, the previously determined optimal SVM parameters (obtained from grid search on the original train-test split) remained unchanged to avoid overfitting. Additional statistical comparisons on the F1-scores from Table S1 in the supplementary material revealed that majority voting and the mean waveform approach significantly outperformed all other methods (see supplementary material for details). In total numbers and on average, majority voting showed the best performance results. We assume that this is because in early fusion large parts of the available input information are removed at an early stage and are not available during the training process. For late fusion, this is not the case. Furthermore, a comparison of the baseline method and the late fusion approach revealed that the aggregation of weak predictions by majority voting allows for a more accurate prediction at subject level. Majority voting adds a layer of abstraction to the outputs of the classifier, which seems to increase robustness. The performance level from the results in Table S1 (supplementary material) are equivalent to that of Table 2. This shows that the employed training-test split does not bias the test result in Table 2.

The conclusion from the first experiment is that as much information as possible should be retained during the classification process and thus late fusion is recommended. Aggregation of information at later stages of the process seems to be superior to aggregation at an early stage, as relevant information of the individual trials is lost. The second experiment suggests that using only the original input signals might not always be the best choice. In most of our experiments, a combined representation of input signals and derived representations was advantageous, especially the combinations {A, Δ, D_A} and {A, D_A, U, D_U} in Table 3. Considerably lower accuracy was achieved when only the individual signals (first five rows in Table 3) were used. The use of a single COP signal (rows 4 and 5) lead to degeneration of the classifier in some cases, i.e., one class could not be modeled at all by the classifier. The combination of the three GRF components is considerably more expressive than the combination of the COP components. The best choice seems to be a combination of all signals (GRF + COP). This also supports our previous findings [16,18,22].

We further observed that the signals of the affected side are more expressive than those of the unaffected side ({A} vs. {U} in Table 3). This observation contradicts the findings of Williams et al. [23]. The combination of affected and unaffected input signals improved the results in five out of eight cases.

The Δ-waveform represents the difference between the affected and the unaffected side and thus explicitly captures the symmetry between both sides. When combined with the signals of the affected side, a moderate increase in accuracy was present in three of eight cases ({A} vs. {A, Δ}). This result suggests that the classifier is able to derive symmetry-related information also from the raw input signals and does not necessarily need it to be explicitly provided. For the unaffected side, the Δ-waveform provides an improvement in seven of eight cases ({U} vs. {U, Δ}). Therefore, the Δ-waveform seems to carry important information. Adding the first derivative as an additional input representation to the signals of the affected or unaffected side showed improvements in 30 out of 40 cases (evident by comparing the first five columns with the last five columns in Table 3).

To obtain additional indicators for the usefulness of the representations, we conducted further experiments with the overall best input representation (GRF + COP, last row in Table 3). We have calculated all ($2^5 - 1 = 31$) possible combinations of A, D_A,U,D_U and Δ for the case GRF + COP and examined how often each representation occurs within the best 10 results. The most useful representations seem to be D_A (contained in 8 of the 10 best results) as well as Δ and A (each contained in 6 of the 10 best results). D_U (5/10 results) and U (4/10 results) seem less important.

The overall recommendation that can be derived from these experiments is that the combination of *more* input signals and input representations (even when they contain redundant information) can lead to better results. This is especially true for combining GRF and COP components but also for using the derivatives of the affected and unaffected sides. Even though the derivatives represent redundant information to the original signals, they might still help the classifier to better grasp class differences. Furthermore, the combination of the affected and unaffected side (either explicitly or implicitly trough Δ) seems to be beneficial as well. The results of our study provide a first indication of which signals to use and how to fuse them. Further investigations with alternative datasets are required to corroborate these findings.

5. Conclusions

The presented work aims at clarifying which aggregation method and which signal representations are best suited for the classification of data obtained from gait analysis (based on GRF assessment).The results show that the aggregation of several trials of one subject is beneficial especially when late fusion or mean waveform is used. Furthermore, the results indicate that the combination of the original signals with derived representations increases the expressive power of the data during feature extraction and classification. The combination of GRF and COP components with derived representations, even though they may be partially redundant, improved classification performance on our data.

Future research will investigate adaptively-learned feature representations as well as the modeling of relationships within a gait cycle to derive more expressive representations.

Acknowledgments

This work was partly funded by the NFB – Lower Austrian Research and Education Company (NFB) and the Provincial Government of Lower Austria, Department of Science and Research (LSC14-005 and FTI17-014) and the Austrian Research Promotion Agency (FFG) and the BMDW within the COIN-program (866855). We want to thank Marianne Worisch and Szava Zoltán for their great assistance in data preparation and their great support in clinical and technical questions.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.gaitpost.2019.10.021.

References

[1] W. Pirker, R. Katzenschlager, Gait disorders in adults and the elderly, Wiener Klinische Wochenschrift 129 (3–4) (2017) 81–95.

[2] P. Mahlknecht, S. Kiechl, B.R. Bloem, J. Willeit, C. Scherfler, A. Gasperi, G. Rungger, W. Poewe, K. Seppi, Prevalence and burden of gait disorders in elderly men and women aged 60–97 years: a population-based study, PLoS ONE 8 (7) (2013) e69627.

[3] R. Baker, Measuring Walking: A Handbook of Clinical Gait Analysis, Mac Keith Press, London, 2013.

[4] T. Chau, A review of analytical techniques for gait data. Part 1: Fuzzy, statistical and fractal methods, Gait Posture 13 (1) (2001) 49–66, https://doi.org/10.1016/S0966-6362(00)00094-1.

[5] J. Figueiredo, C.P. Santos, J.C. Moreno, Automatic recognition of gait patterns in human motor disorders using machine learning: a review, Med. Eng. Phys. (2018), https://doi.org/10.1016/j.medengphy.2017.12.006.

[6] H. Lau, K. Tong, H. Zhu, Support vector machine for classification of walking conditions of persons after stroke with dropped foot, Hum. Mov. Sci. 28 (4) (2009) 504–514, https://doi.org/10.1016/j.humov.2008.12.003.

[7] F. Wahid, R.K. Begg, C.J. Hass, S. Halgamuge, D.C. Ackland, Classification of Parkinson's disease gait using spatial-temporal gait features, IEEE J. Biomed. Health Inform. 19 (6) (2015) 1794–1802, https://doi.org/10.1109/JBHI.2015.2450232.

[8] L. Van Gestel, T. De Laet, E. Di Lello, H. Bruyninckx, G. Molenaers, A. Van Campenhout, E. Aertbelin, M. Schwartz, H. Wambacq, P. De Cock, K. Desloovere, Probabilistic gait classification in children with cerebral palsy: a Bayesian approach, Res. Dev. Disabilit. 32 (6) (2011) 2542–2552, https://doi.org/10.1016/j.ridd.2011.07.004.

[9] M. Alaqtash, T. Sarkodie-Gyan, H. Yu, O. Fuentes, R. Brower, A. Abdelgawad, Automatic classification of pathological gait patterns using ground reaction forces and machine learning algorithms, 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, IEEE, 2011, pp. 453–457.

[10] C. Nesch, V. Valderrabano, C. Huber, V. von Tscharner, G. Pagenstert, Gait patterns of asymmetric ankle osteoarthritis patients, Clin. Biomech. 27 (6) (2012) 613–618, https://doi.org/10.1016/j.clinbiomech.2011.12.016.

[11] J. Wu, J. Wang, PCA-based SVM for automatic recognition of gait patterns, J. Appl. Biomech. 24 (1) (2008) 83–87.

[12] D. Soares, M. de Castro, E. Mendes, L. Machado, Principal component analysis in ground reaction forces and center of pressure gait waveforms of people with transfemoral amputation, Prosthet. Orthot. Int. 40 (6) (2016) 729–738.

[13] J. Christian, J. Krll, G. Strutzenberger, N. Alexander, M. Ofner, H. Schwameder, Computer aided analysis of gait patterns in patients with acute anterior cruciate ligament injury, Clin. Biomech. 33 (2016) 55–60, https://doi.org/10.1016/j.clinbiomech.2016.02.008.

[14] B.M. Eskofier, P. Federolf, P.F. Kugler, B.M. Nigg, Marker-based classification of young-elderly gait pattern differences via direct PCA feature extraction and SVMs, Comput. Methods Biomech. Biomed. Eng. 16 (4) (2011) 435–442.

[15] P. Levinger, D. Lai, R. Begg, K. Webster, J. Feller, The application of support vector machines for detecting recovery from knee replacement surgery using spatio-temporal gait parameters, Gait Posture 29 (1) (2009) 91–96.

[16] D. Slijepcevic, B. Horsak, C. Schwab, A. Raberger, M. Schüller, A. Baca, C. Breiteneder, M. Zeppelzauer, Ground reaction force measurements for gait classification tasks: effects of different PCA-based representations, Gait Posture 57 (2017) 4–5.

[17] M. Sangeux, J. Polak, A simple method to choose the most representative stride and detect outliers, ResearchGate 41 (2) (2014), https://doi.org/10.1016/j.gaitpost.2014.12.004.

[18] D. Slijepcevic, M. Zeppelzauer, A.-M. Gorgas, C. Schwab, M. Schüller, A. Baca, C. Breiteneder, B. Horsak, Automatic classification of functional gait disorders, IEEE J. Biomed. Health Inform. 22 (5) (2018) 1653–1661.

[19] T. Chau, S. Young, S. Redekop, Managing variability in the summary and comparison of gait data, J. NeuroEng. Rehabil. 2 (1) (2005) 22, https://doi.org/10.1186/1743-0003-2-22.

[20] C. Chang, C. Lin, LIBSVM: a library for support vector machines, ACM Trans. Intell. Syst. Technol. 2 (2011) 27:1-27:27.

[21] E. Combrisson, K. Jerbi, Exceeding chance level by chance: the caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy, J. Neurosci. Methods 250 (2015) 126–136.

[22] D. Slijepcevic, M. Zeppelzauer, C. Schwab, A. Raberger, B. Dumphart, A. Baca, C. Breiteneder, B. Horsak, P 011-towards an optimal combination of input signals and derived representations for gait classification based on ground reaction force measurements, Gait Posture 65 (2018) 249.

[23] G. Williams, D. Lai, A. Schache, M. Morris, Classification of gait disorders following traumatic brain injury, J. Head Trauma Rehabil. 30 (2) (2015) E13–E23.