

Biomechanical Features of Running Gait Data Associated with Iliotibial Band Syndrome: Discrete Variables Versus Principal Component Analysis

A. Phinyomark¹, S.T. Osis^{1,3}, D. Kobsar¹, B.A. Hettinga¹, R. Leigh¹, and R. Ferber^{1,2,3}

¹Faculty of Kinesiology, University of Calgary, Calgary, AB, Canada

²Faculty of Nursing, University of Calgary, Calgary, AB, Canada

³Running Injury Clinic, Calgary, AB, Canada

Abstract— The features associated with temporal gait biomechanical data are complex and multivariate and it is therefore necessary to identify methods that reduce the difficulty underlying the interpretation and identification of differences between groups of interest. Discrete variables and principal component analysis (PCA) are feature extraction methods that have been widely used. However, a comprehensive understanding of the relationship between discrete variables and PCA features has never been completed. The objectives of this study were to (1) determine the relationships between the two feature methods and (2) compare the performance of each for the identification and discrimination of between-group differences for injured and non-injured subjects. Running gait kinematic data of 48 patients experiencing iliotibial band syndrome (ITBS) were compared to a group of 48 asymptomatic control subjects for transverse plane hip and ankle joint and frontal plane hip joint waveform data. Twenty-two discrete variables and three to four PCA features were extracted from each waveform and divided into three subgroups: magnitude features, difference operator features, and phase shift features. The following key results were obtained: (1) strong correlations were found between discrete variables; (2) the first PCA feature captured the magnitude information and thus showed strong correlation with the discrete variables in the magnitude group; (3) there was no consistent result that showed all discrete variables were found in the first few principal components; (4) the performance of the PCA features in identifying between-group differences decreased (reduced the effect size) as compared to using the discrete variables, but this does not necessarily result in a decrease in the performance of the PCA features to discriminate between ITBS and controls using a support vector machine classifier. These results suggest care must be taken when selecting features of gait waveforms for both identification and discrimination of between-group differences for injured and non-injured runners.

Keywords— feature extraction, feature reduction, feature selection, gait analysis, kinematics.

I. INTRODUCTION

Both identification and discrimination of between-group differences are important in gait biomechanics research [1-3]. The three-dimensional (3D) aspect of gait kinematics makes the data complex and possibly correlated. Therefore, a smaller number of features extracted from a set of temporal waveforms representing joint angles during the gait cycle are necessary in order to improve the identification performance (the use of features to determine between-group differences; e.g. the multiple comparisons problem [1-2]) and/or the discrimination performance (the use of features to allow an unknown new subject to be correctly classified as belonging to one group or another [2-3]). Finally, identifying a smaller number of features can help reduce the difficulty of interpreting the 3D data into clinically meaningful information.

One of the most commonly used, and simplest feature extraction methods for analyzing gait data, is the use of discrete events and descriptive statistics of the gait waveform. For example, the joint angle at touchdown and toe-off, peak angles, ranges of motion, and time-to-peak angles are commonly extracted from waveform data. However, this method calls for the *a priori* selection of discrete features and relies on sufficient background knowledge [1] and/or subjective opinion. Moreover, the selection of these discrete features neglect the temporal information of the gait waveforms, and the strong correlations between untransformed features of gait waveforms may remain.

To analyze the entire gait waveforms, feature transformation techniques have been applied to gait kinematic data [4]. The most commonly used method is a principal component analysis (PCA) method. In brief, a PCA transforms a set of raw data into a set of linearly uncorrelated variables called principal components (PCs), and PCA-derived scores (PC-scores) are commonly used as the features of gait waveforms [2, 5-6]. While these features can capture temporal information, interpreting the biomechanical meaning of the PCA features is difficult and can be a subjective process [4]. The computational complexity for the PCA procedure also increases as compared to the discrete variables.

The original version of this chapter was inadvertently published with an incorrect chapter pagination 574–579 and DOI 10.1007/978-3-319-32703-7_112. The page range and the DOI has been re-assigned. The correct page range is 580–585 and the DOI is 10.1007/978-3-319-32703-7_113. The erratum to this chapter is available at DOI: [10.1007/978-3-319-32703-7_260](https://doi.org/10.1007/978-3-319-32703-7_260)

© Springer International Publishing Switzerland 2016

E. Kyriacou et al. (eds.), *XIV Mediterranean Conference on Medical and Biological Engineering and Computing 2016*,

IFMBE Proceedings 57,

DOI: 10.1007/978-3-319-32703-7_113

Despite the growing use of PCA in gait studies, and the need to detect significant differences and discriminate between two groups of interest, there remains a need for a comprehensive understanding of the relationship between discrete variables and PCA features. Therefore, the first purpose of this study was to determine the relationships between the two feature types by (1) dividing them into subgroups according to mathematical properties and information captured in discrete and waveform features, and (2) measuring the linear correlation between the two feature types based on the subgroups determined. Additionally, the second purpose of this study was to compare the performances of both feature types for the identification and discrimination of differences between groups of injured and non-injured subjects.

The injury examined was iliotibial band syndrome (ITBS), the second most common running-related injury and the most common cause of lateral knee pain [7]. Atypical running gait biomechanics are considered a primary factor in the aetiology of ITBS [8]. The current analysis focused on three running gait waveforms based on previous literature [7] demonstrating their relevance to ITBS: hip and ankle joint transverse plane angles and hip joint frontal plane angles.

II. METHODS

A. Data Collection and Processing

Transverse plane hip joint kinematic data were collected from 29 female runners with ITBS (34 ± 8 years) and 29 healthy female runners (35 ± 8 years), while transverse plane ankle and frontal plane hip angles were collected from 19 male runners with ITBS (39 ± 12 years) and 19 healthy male counterparts (39 ± 12 years). Injured runners were matched for sex, age, height, weight, and running speed, with healthy controls who had not experienced any musculoskeletal injuries over the six months prior to the time of testing. The University of Calgary's Conjoint Health Research Ethics Board approved the collection and the analysis of the data, and prior to collecting the data, all participants provided their written informed consent to participate.

Eight high-speed digital video cameras were used to film running gait at 200 Hz while the subjects ran on the treadmill at a comfortable self-selected speed for 20 s. Kinematic joint angles were calculated and normalized for the stance phase (1%-35%) and the swing phase (36%-100%) of running gait. Each feature was extracted from each stride, and the mean value of ten strides was used as the feature discrete value for each subject. Details of inclusion and exclusion criteria for the ITBS runners along with more details about data collection can be found in Phinyomark et al. [7].

B. Discrete Variables

In the current investigation, discrete gait biomechanical variables were divided into three groups based on mathematical

properties and the information captured in the features: (1) a magnitude feature group, (2) a difference operator feature group, and (3) a phase shift feature group. Specifically, a magnitude feature was defined as the amplitude value of the temporal waveform at a specific event of the gait cycle (i.e., a single-point magnitude feature) or an average amplitude value of the temporal waveform over a specific period of the gait cycle. Eight gait events were selected involving (F1) touchdown, (F2) mid-stance, (F3) toe-off, (F4) mid-swing, (F5) maximum peak during stance, (F6) minimum peak during stance, (F7) maximum peak during swing, and (F8) minimum peak during swing. In addition, meaningful joint angles determined from previous research [7] were used to determine the amplitude value of the temporal waveform over a specific period during (F9) stance, (F10) the swing phase, and (F11) the entire gait cycle.

A difference operator feature was defined as a relative change in the amplitude of two gait events and angular excursion and the range of motion were the two features of interest. Angular excursion was defined as the peak angle subtracted from the initial angle at the start of the gait cycle. These consisted of (F12-F13) the maximum peak excursion during the stance phase and the swing phase and (F14-F15) the minimum peak excursion during the stance phase and the swing phase. Range of motion (ROM) was defined as the difference between the maximum peak and the minimum peak during (F16) the stance phase, (F17) the swing phase, and (F18) the entire gait cycle.

A phase shift feature was defined as a relative change in the timing of two gait events, i.e., time-to-peak. Time-to-peak was defined as a time interval from the initial angle at the start of the gait phase to the peak angle. There were (F19-F20) time-to-maximum peak variables and (F21-F22) time-to-minimum peak variables during the stance and the swing phase. In total, there were 22 discrete variables or features extracted from each waveform.

C. Principal Component Analysis

The PCA attempts to account for as much of the variability in the original data within the first few PCs. In the current investigation, the first PCs that collectively explained at least 90% of the cumulative variance in the original data, and had an eigenvalue >1 [5] were retained. The PC-scores were computed by multiplying the standardized data matrix (zero mean, unit variance) by the eigenvector matrix and were used as the PCA features. Subsequently, each PCA feature was assigned to one of the three feature groups as defined by the discrete variables: the magnitude feature group, the difference operator feature group, and the phase shift feature group. This step was either done by inspecting the shape of the PC loading vector [4, 9] or by examining

the differences between two representative original [9] or reconstructed [4] waveforms chosen from some quantiles of the data. Hence, the number of PCA features in each group was dependent on each waveform.

D. Data Analysis

First, all features extracted were assigned to their respective feature groups. Next, one representative feature from each group was determined using Cohen's effect size, d , for both discrete variables and PCA features. An effect size between 0.20-0.49 was defined as a small effect, 0.50-0.79 a medium effect, and ≥ 0.80 a large effect [10]. Third, the correlation coefficients between the representative features were computed for each waveform and correlation coefficients between features within the same group were also calculated. A correlation coefficient, r , of ≤ 0.35 was defined as a weak correlation, 0.36 to 0.67 a moderate correlation, and 0.68 to 1.0 a strong correlation [11]. Finally, the performance of both feature types for the identification and the discrimination of differences between the ITBS and healthy runners were respectively examined by a one-way analysis of variance (ANOVA) and a support vector machine (SVM) classifier with a linear kernel [2, 12]. Significance was set at $P < 0.05$, and a leave-one-out cross validation method was applied to obtain classification rates.

III. RESULTS

All proposed discrete variables were computed, except two variables from the hip frontal plane waveform (F9 and F11) because of no meaningful joint angles during stance. For the PCA, three PCs (PC1-PC3), which explained 92.20% of the variance in the data, were retained for the hip transverse plane waveform. Four PCs (PC1-PC4), which explained 91.39% and 93.82% of the variance in the data were retained for the ankle transverse plane and hip frontal plane waveforms, respectively. Effect size (range of values) and correlation coefficients for magnitude, difference operator, phase shift, and PCA features in the same group are shown in Table 1.

The representative discrete features for the hip transverse plane waveforms for the female ITBS and healthy groups were F6 (the minimum peak angle during stance; a magnitude feature), F11 (the mean value of joint angles during 26%-39% and 51%-65% of gait cycle; a magnitude feature), F14 (the minimum peak excursion during stance; a difference operator feature), and F19 (the time-to-maximum peak during stance; a phase shift feature). The PC1-score was a magnitude feature while PC2- and PC3-scores were a difference operator feature and a phase shift feature, respectively.

Table 1 Effect size, d , values of features and correlation coefficients, r , between features in the same group: magnitude discrete variables, difference operator discrete variables, phase shift discrete variables and PCA features.

Waveform		Magnitude	Difference	Phase shift	PCA
Hip Transverse – Female	d	0.20-0.95	0.21-0.84	0.02-0.24	0.22-0.70
	r	0.57-0.99	0.18-0.99	0.01-0.75	< 0.01
Ankle Transverse – Male	d	0.79-1.20	0.03-0.24	0.12-0.40	0.08-1.04
	r	0.65-0.99	0.03-0.84	0.09-0.44	< 0.33
Hip Frontal – Male	d	0.13-1.05	0.06-0.87	0.07-0.36	0.01-0.48
	r	0.04-0.95	0.01-0.93	0.17-0.74	< 0.08

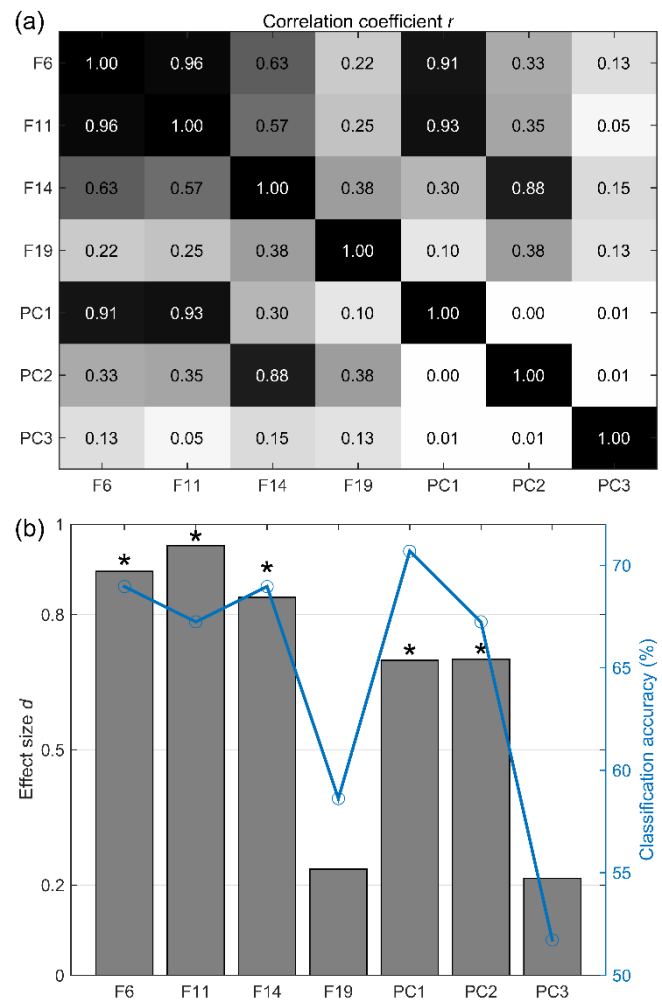


Fig. 1 Transverse plane gait waveform of the hip joint for female runners. (a) Correlation coefficients, r , between the representative features: F6, F11, F14, F19, PC1, PC2, and PC3. A strong correlation is presented in white font. (b) Effect size, d , (bars), statistical between-group differences (*) and classification accuracy (%) (line) for the differences between female ITBS and healthy runners.

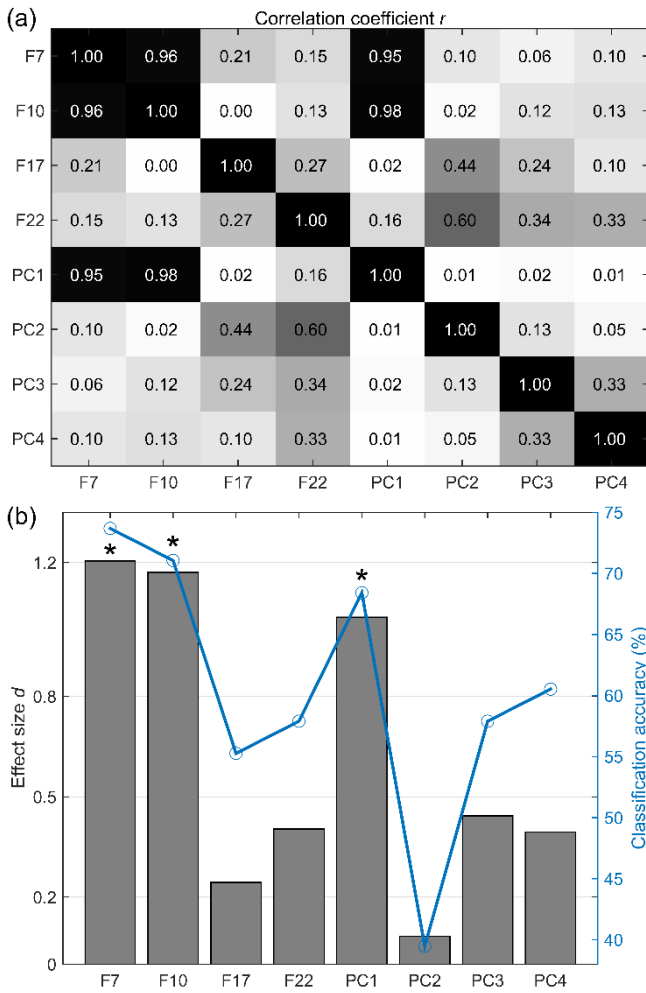


Fig. 2 Transverse plane gait waveform of the ankle joint for male runners. (a) Correlation coefficients, r , between the representative features: F7, F10, F17, F22, PC1, PC2, PC3, and PC4. A strong correlation is presented in white font. (b) Effect size, d , (bars), statistical between-group differences (*) and classification accuracy (line) for the differences between male ITBS and healthy runners.

The representative discrete features of ankle transverse plane waveforms for the male ITBS and healthy groups were F7 (the maximum peak angle during swing; a magnitude feature), F10 (the mean value of angles during 1%-17%, 31%-39%, and 62%-100% of gait cycle; a magnitude feature), F17 (ROM during the swing phase; a difference operator feature), and F22 (the time-to-minimum peak during the swing phase; a phase shift feature). The PC1-score was a magnitude feature while PC2- to PC4-scores were phase shift features.

The representative discrete features for the hip frontal plane waveforms for male runners were F7 (the maximum peak angle during swing; a magnitude feature), F10 (the mean value of angles during 36%-45% of gait cycle; a

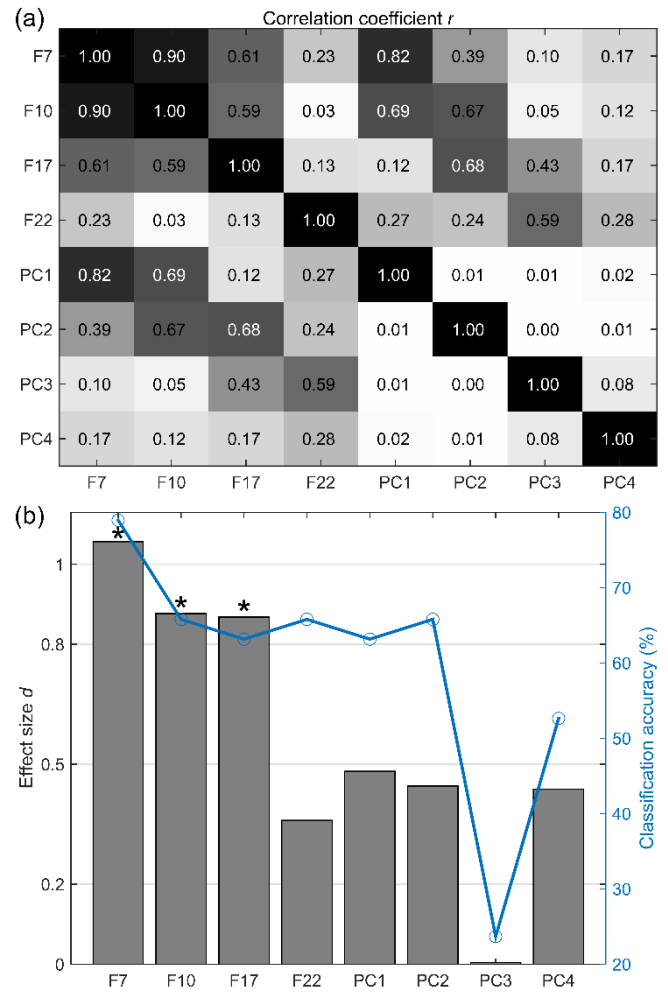


Fig. 3 Frontal plane gait waveform of the hip joint for male runners. (a) Correlation coefficients, r , between the representative features: F7, F10, F17, F22, PC1, PC2, PC3, and PC4. A strong correlation is presented in white font. (b) Effect size, d , (bars), statistical between-group differences (*) and classification accuracy (line) for the differences between male ITBS and healthy runners.

magnitude feature), F17 (ROM during the swing phase; a difference operator feature), and F22 (the time-to-minimum peak during the swing phase; a phase shift feature). The PC1-score was a magnitude feature while PC2- and PC4-scores were difference operator features and PC3-score was a phase shift feature.

Correlation coefficients between the representative features for the three waveforms are shown in Fig. 1(a), 2(a), and 3(a). Effect sizes for the representative features and their statistical significance for the three waveforms are shown in Fig. 1(b), 2(b), and 3(b), together with classification accuracies of discrimination between ITBS and healthy groups.

IV. DISCUSSION AND FUTURE WORKS

Strong linear relationships were found between most of the discrete variables in the same group, particularly in the magnitude feature group (Table 1). These results are consistent with previous literature showing that strong correlations can be found between gait waveforms [13]. Thus, excluding the remaining set of highly correlated discrete-time variables should not effectively reduce discrimination performance. These results are similar to the classification of microarray gene expression data [14] as well as surface EMG data [15] and suggest that adding or eliminating the correlated variables from a feature vector does not significantly increase or decrease the classification accuracy of a classifier. Future studies, however, should further investigate the issue of “redundancy” of discrete variables from gait waveform data using other correlation measures (when the relationship is not linear) and as well the effect on classification performance (for both linear and nonlinear classifiers).

Because of the data redundancy, only the best discriminatory discrete variables for each feature type should therefore be used for further analysis. The current investigation, however, involved only three types of time-domain features. Future studies should involve other useful discrete variables in the frequency domain and time-frequency representation in an attempt to better understand the relationships between discrete variables of gait waveform data.

Since no linear relationship, or a weak relationship, existed for the PCA features (Table 1), all the PCA features were interpreted and assigned the feature types by inspecting the shape of the PC loading vector together with examining the differences between two representative extreme subject waveforms. For example, the PC1 loading vector exhibited a positive magnitude throughout the gait cycle for all the waveforms of interest. The extreme raw waveforms also showed a large difference in the amplitude during the entire gait cycle suggesting that PC1 captured between-subject variance as a magnitude feature throughout the gait cycle. The results of the current study also showed that the PC1-score had a strong linear relationship with the discrete variables in the magnitude feature group (Fig. 1(a), 2(a), and 3(a)). However, the performance of the PC1-score in identifying between-group differences decreased, along with a reduced effect size, as compared to using the discrete variables in the magnitude feature group (Fig. 1(b), 2(b), and 3(b)).

This current study also suggests that the *a priori* selection of discrete features, which relies on either relevant background knowledge [1] or search approaches combined with filter metrics [2, 3, 12], achieved good results and only one representative magnitude discrete variable (such as a peak angle) for each waveform was necessary, as compared to using the PC1-score. For instance, the effect size for the

PC1-score of the hip transverse plane waveform decreased as compared to using the minimum peak angle during stance (F6) and the mean value of the joint angles (F11) (Fig. 1(b)). However, the classification accuracy using the PC1-score increased as compared to only using the discrete variables. Moreover, the PC1-score was sufficient enough to determine between-group differences (Fig. 1(b), 2(b), and 3(b)), while the *a priori* selection of PCA features was unnecessary. Therefore, these results suggest care must be taken when selecting gait waveform features for either the identification or the discrimination between two groups of injured and non-injured runners. Moreover, future research is needed to better understand the relationship between identification and discrimination performances of biomechanical features.

Since a difference operator feature was defined as a relative change in the amplitude of two gait events, it is reasonable to expect that the representative discrete variables in this group, and in the magnitude feature group, exhibited a moderate correlation (e.g. F6, F11 and F14 in Fig. 1(a); F7, F10 and F17 in Fig. 3(a)). A strong correlation was also found between the representative discrete variable and the PC-score if the PC captured the similar relative change in the amplitude (e.g. F14 and PC2 in Fig. 1(a); F22 and PC2 in Fig. 3(a)). In the current investigation, PC2-score of the hip transverse plane waveform, as well as PC2- and PC4-scores of the hip frontal plane waveform, were associated with a difference operator feature. The loading vectors of these PCs had a positive or a negative peak aligned with one local peak in the mean raw waveform and then an opposite peak angle aligned with the consecutive local peak in the mean raw waveform. Specifically, PC2-scores of the hip transverse plane waveform for a female group and the hip frontal plane waveform for a male group captured a large change between the minimum peak and the maximum peak angles during the swing phase (F17). In addition, PC4-score of the hip frontal plane waveform captured a change between local peaks for both gait phases and also captured a change in the timing of the maximum peak angle during the swing phase. Further, the performances for identifying and discriminating between-group differences using the PC-scores in this group were similar to the results found in the group of magnitude features.

Phase shift features did not provide good results for both identification and discrimination of ITBS and healthy runners in comparison to other features types. However, phase shift features may be useful for identifying between-group differences for other running-related injuries such as patellofemoral pain (PFP) and future research is therefore needed to evaluate the usefulness of this feature type. On the other hand, due to weak linear relationships between features, the representative group discrete features may

improve the discrimination performance of the classifier [14-16].

In the current investigation, PC3-score for the hip transverse plane waveform, PC2- to PC4-scores for the ankle transverse plane waveform and the PC3-score for the hip frontal plane waveform were likely a phase shift feature. Specifically, PC3-score of the hip transverse plane waveform captured a difference in the timing of the maximum peak during the swing phase (F21). For the ankle transverse plane waveform, the PC2-score captured differences in the timing of the maximum peak angle during the stance phase and as well the swing phase as a combination of F19 and F21, while PC3- and PC4-scores only captured the timing of the maximum peak angle during the swing phase (F21). For the hip frontal plane waveform, the PC3-score captured differences in the timing of the maximum and minimum peak angles during the swing phase as a combination of F20 and F21.

Only the first few PCs, or lower-order PCs [5], were retained as these PCs are generally applied in gait biomechanics research [6, 13]. However, we can observe that the first few PCs (PC1-PC4) of the ankle transverse plane waveform did not capture the three features types. These results support previous research [5] and suggest that excluding the remaining set of intermediate- and higher-order PCs may remove valuable information.

The following concluding remarks can be drawn from the current investigation: (1) strong linear relationships were found between discrete variables while no linear relationships, or weak linear relationships, were found between PCA features; (2) PC1 for each waveform captured the magnitude information and thus showed a strong linear relationship with the discrete variables in the magnitude group; (3) there is no guarantee that the first few PCs (PC2-PC4) for each waveform will capture all features types; (4) the ability of the PCA to identify between-group differences decreased as compared to using the discrete variables, but this does not necessarily mean that the performance of the PCA to discriminate between ITBS and controls using a support vector machine classifier will also decrease.

ACKNOWLEDGMENT

The Canadian Institutes of Health Research (CIHR) Fellowship and the Alberta Innovates-Health Solutions (AIHS) Postgraduate Fellowship partially funded this research.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

REFERENCES

1. Ferber R, Davis I M, Williams III D S (2003) Gender differences in lower extremity mechanics during running. *Clin Biomech* 18:350-357
2. Phinyomark A, Hettinga B A, Osis S T et al (2014) Gender and age-related differences in bilateral lower extremity mechanics during treadmill running. *PLoS One* 9:e105246
3. Eskofier B M, Kraus M, Worobets J T et al (2012) Pattern classification of kinematic and kinetic running data to distinguish gender, shod/barefoot and injury groups with feature ranking. *Comput Methods Biomech Biomed Eng* 15:467-474
4. Brandon S C E, Graham R B, Almosnino S et al (2013) Interpreting principal components in biomechanics: Representative extremes and single component reconstruction. *J Electromyogr Kinesiol* 23:1304-1310
5. Phinyomark A, Hettinga B A, Osis S et al (2015) Do intermediate- and higher-order principal components contain useful information to detect subtle changes in lower extremity biomechanics during running? *Hum Mov Sci* 44:91-101
6. Foch E, Milner C E (2014) The influence of iliotibial band syndrome history on running biomechanics examined via principal component analysis. *J Biomech* 47:81-86
7. Phinyomark A, Osis S, Hettinga B A et al (2015) Gender differences in gait kinematics in runners with iliotibial band syndrome. *Scand J Med Sci Sports* 25:744-753
8. Louw M, Deary C (2014) The biomechanical variables involved in the aetiology of iliotibial band syndrome in distance runners - a systematic review of the literature. *Phys Ther Sport* 15:64-75
9. Jones MC, Rice JA (1992) Displaying the important features of large collections of similar curves. *Am Stat* 46:140-145
10. Taylor R (1990) Interpretation of the correlation coefficient: a basic review. *JDMS* 1:35-39
11. Cohen J (1988) *Statistical power analysis for the behavioral sciences*. Lawrence Erlbaum Associate, Hillsdale, NJ
12. Fukuchi R K, Eskofier B M, Duarte M et al (2011) Support vector machines for detecting age-related changes in running kinematics. *J Biomech* 44:540-542
13. Deluzio KJ, Wyss UP, Zee B et al (1997) Principal component models of knee kinematics and kinetics: normal vs. pathological gait patterns. *Hum Mov Sci* 16:201-217
14. Ding C, Peng H (2005) Minimum redundancy feature selection from microarray gene expression data. *J Bioinf Comput Biol* 3:185-205
15. Phinyomark A, Phukpattaranont P, Limsakul C (2012) Feature reduction and selection for EMG signal classification. *Expert Syst Appl* 39:7420-7431
16. Phinyomark A, Phukpattaranont P, Limsakul C (2012) Fractal analysis features for weak and single-channel upper-limb EMG signal. *Expert Syst Appl* 39:11156-11163.

Address of the corresponding author:

Author: Angkoon Phinyomark
 Institute: Faculty of Kinesiology, University of Calgary
 Street: 2500 University Dr NW
 City: Calgary, Alberta
 Country: Canada
 Email: aphinyom@ucalgary.ca, angkoon.p@hotmail.com