Automated Design of Robust Discriminant Analysis Classifier for Foot Pressure Lesions Using Kinematic Data

John Yannis Goulermas*, *Member, IEEE*, Andrew H. Findlow, Christopher J. Nester, David Howard, and Peter Bowker

***Abstract*—In the recent years, the use of motion tracking systems for acquisition of functional biomechanical gait data, has received increasing interest due to the richness and accuracy of the measured kinematic information. However, costs frequently restrict the number of subjects employed, and this makes the dimensionality of the collected data far higher than the available samples. This paper applies discriminant analysis algorithms to the classification of patients with different types of foot lesions, in order to establish an association between foot motion and lesion formation. With primary attention to small sample size situations, we compare different types of Bayesian classifiers and evaluate their performance with various dimensionality reduction techniques for feature extraction, as well as search methods for selection of raw kinematic variables. Finally, we propose a novel integrated method which fine-tunes the classifier parameters and selects the most relevant kinematic variables simultaneously. Performance comparisons are using robust resampling techniques such as Bootstrap $632+$ and k-fold cross-validation. Results from experimentations with lesion subjects suffering from pathological plantar hyperkeratosis, show that the proposed method can lead to $\sim 96\%$ correct classification rates with less than 10% of the original features.**

***Index Terms*—Bootstrap, classification, discriminant analysis, feature extraction/selection, foot kinematics, gait, genetic algorithm, hyperkeratosis, regularization.**

I. Introduction

THE analysis of gait is of paramount importance to the assessment of a multitude of human pathologies associated with kinesiological performance, the evaluation of their necessitated remedial procedures, and also the understanding of the intrinsic properties of the generative human biomechanical model. The use of increasingly sophisticated gait sensors for kinematic, kinetic, electromyographic and anthropometric measurements in the recent years, has initiated a commensurate sophistication in the possibilities and needs for tools to model and assess the human gait.

Manuscript received September 1, 2004; revised January 23, 2005. This paper is part of the REAL-PROF project (www.realprof.eu.com), supported by the European Commission, IST programme, eHealth Unit, contract number IST/2001/38429. *Asterisk indicates corresponding author.*

*J. Y. Goulermas is with the Department of Electrical Engineering and Electronics, Brownlow Hill, University of Liverpool, Liverpool L69 3GJ, U.K. (e-mail: j.y.goulermas@liverpool.ac.uk).

A. H. Findlow, C. Nester, and P. Bowker are with the Centre for Rehabilitation and Human Performance Research (CRHPR), Salford University, Salford M5 4WT, U.K. (e-mail: a.h.findlow@salford.ac.uk; c.j.nester@salford.ac.uk; p.bowker@salford.ac.uk).

D. Howard is with the CRHPR and the School of Computing, Science and Engineering, Salford University, Salford M5 4WT, U.K. (e-mail: d.howard@salford.ac.uk).

Digital Object Identifier 10.1109/TBME.2005.851519

Because traditional approaches cannot address complex issues, the fields of machine learning and intelligent pattern analysis [1] have received a increasing interest in gait science over the last decade. The recent reviewing texts of [2] and [3], for example, summarize a multitude of works using techniques such as fuzzy analysis, multivariate statistics, neural networks, and signal processing for modeling, classification and prediction of gait variables. Other representative examples include [4] who used support vector machines (SVMs) for the classification of age groups using kinetic and kinematic gait data. Feedforward neural networks were applied to the prediction of electromyographic (EMG) activity from kinematic patterns in [5], while in [6] they were used to estimate motor unit parameters from surface detected action potentials. A dynamic recurrent neural network was employed in [7] to predict kinematic variables from EMG data. Principal component analysis (PCA) was used in [8] to identify muscle activation patterns using surface EMG, and in [9] to assess gait normality in children using sagittal plane joint data. Additional examples include [10] who employed cluster analysis to identify abdominal and erector spinae muscle activity patterns, and [11] who used linear discriminant analysis (LDA) to differentiate between normal and flat foot subjects through the use of force measurements. Furthermore, wavelets coupled with a kinematic model were used in [12] for detecting postural and walking patterns in the elderly, while [13] applied fuzzy logic to the detection of gait events in functional electrical simulation.

This paper provides a thorough analysis for robust classification of dynamic gait data, with particular attention to a number of well-known issues. First, as pointed out in [2] the rich data that contemporary sensor arrays generate, yields a very high degree of dimensionality. Such a problem strongly dictates a dimensionality reduction, in order to improve clinical interpretability as well as reduce costs. Unfortunately, the high dimensionality is further aggravated by insufficient data due to the fact that gait data collection is costly, laborious and time consuming; this gives rise to the well-known statistical small sample size issue [14]. Another concern with small samples is that complicated machine learning algorithms may not be the most suitable ones. Neural networks [15] for instance, may have low generalization ability, often need fine-tuning of many parameters, are more cumbersome to implement, and prone to local minima. Classifiers with simple decision surfaces (see Section III-A) can lead to more robust performance and better decision interpretation.

We investigate and compare the performance of different types of Discriminant Analysis designed to operate on small data sam-

TABLE I
ELEMENTS OF THE CAPTURED AND PREPROCESSED MOTION DATA. (A) JOINTS AND THEIR CORRESPONDING RIGID BODY SEGMENTS USED TO CALCULATE THE EULER ANGLES. (B) THE SEVEN STANCE EVENTS SELECTED FROM THE MOTION ANGLES

(a)

Joint	Proximal Segment	Distal Segment
ankle joint complex(AC)	lower leg	heel
mid tarsal joint(MT)	heel	midfoot
1st ray(R1)	midfoot	1st metatarsal
1st metatarsal phalangeal(MP1)	1st metatarsal	hallux
rearfoot complex(RFC)	lower leg	midfoot
{metatarsal heel}(MH)	heel	1st metatarsal
{toe forefoot}(TF)	midfoot	hallux

(b)

Selected Timing Events
heel contact(HC)
foot flat(FF)
ankle neutral(AN)
heel off(HO)
max. ankle dorsiflexion(MAD)
max. toe dorsiflexion(MTD)
toe off(TO)

ples. Moreover, different dimensionality reduction techniques for generation as well as selection of kinematic features are examined. Finally, we propose a flexible method to perform simultaneous classifier fine-tuning and feature selection, in order to achieve robust performance on a subset of the raw kinematic variables. Despite the application of the proposed algorithms to a specific problem, they are not problem specific, but are rather generic and applicable to gait data afflicted by the aforementioned issues.

This paper aims to use foot motion data to predict which of two patterns of plantar pressure lesions a patient belongs to. Specifically, we apply the above techniques to the classification of kinematic gait patterns for pathological plantar hyperkeratosis (PPH), a debilitating pathology which causes thickening of the stratum corneum of the skin [16]. We have designed a novel system for automated *in vivo* kinematic data acquisition which, unlike previous motion tracking-based designs that use mostly frontal or sagittal plane projections, is capable of recording angular joint data in all three planes. While, previous works have modeled the foot statically [17], [18] and as a single segment [19], we have proposed a multisegment model of the lower leg and foot, which enables us to measure relative movements between segments and estimate three-dimensional (3-D) joint displacement angles during gait dynamically. Additionally, we employ mathematical pattern analysis methods in order to investigate the existence of intrinsic statistical dependencies between foot function and PPH, in the context of elementary kinematic behavior. Doing so also enables the assessment of the sensitivity and accuracy of the proposed multisegment foot model.

Two subject groups are employed, one (henceforth denoted by ω_1) with PPH under the metatarsal heads one and five, and another (ω_2) with PPH under metatarsal heads two, three and four (normal subjects cannot be used in this study, due to the uncertainty of currently undergoing PPH formation). By establishing a mathematical mapping denoted by ψ between a set of kinematic measurements x and the different PPH groups ω_i, we can verify the premise [20] that (confined or inordinate) variations of rearfoot motion are the primary cause of PPH. Such a claim can provide the basis for a potential patient screening and prevention plan prior to developing PPH. Since, to the best of our knowledge, there are no other previous works examining the link of kinematic foot function and PPH to use as a basis for comparison, we experiment with a large set of algorithms with different characteristics. The originality of this paper is summarized by the successful combination of a multisegment foot model and pattern recognition techniques to examine for any relationship between functional biomechanics and PPH.

This paper is structured as follows. Section II summarizes the data capture procedures and defines the problem at hand. Sections III-A–D outline the basics of Bayesian classification and describe techniques for dimension reduction and regularization, Section III-E presents the model error estimators we use, and Sections III-F and G discuss feature extraction and automated classifier design. The experimentation setups and their results are presented in Section IV, while Section V concludes the article.

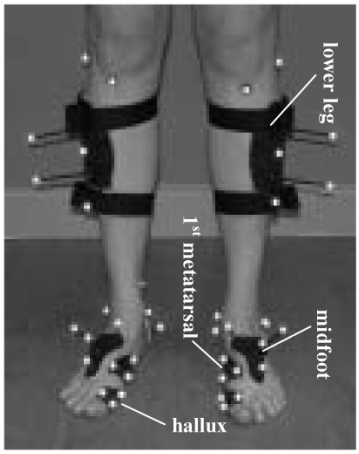
II. KINEMATIC MODELLING OF THE LOWER LEG AND FOOT AND DATA ACQUISITION

To describe the motion of the leg and foot, a multisegment foot/ankle model comprising the lower leg, heel, midfoot, first metatarsal and the proximal phalanx of the hallux [20] was used. Table I(a) summarizes the joints and their defining segment pairs.

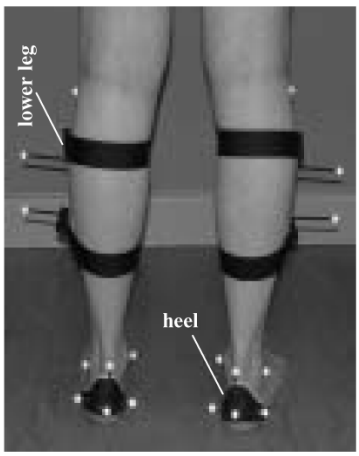
Reflective spheres (6-mm-diameter) were mounted on rigid plastic plates and attached to the leg, heel, midfoot, first metatarsal, and hallux (see Fig. 1). An infrared-camera-based motion tracking system was used to capture the 3-D motion of the reflective spheres during the gait cycle, at a sampling frequency of 100 Hz.

We used the calibrated anatomical systems technique [21], [22] to ensure that the description of foot/ankle joint motion from the reflective spheres related to the cardinal anatomical planes of motion. The kinematic variables measured were 3-D Euler angle displacements between segments in the Frontal (F), Transverse (T), and Sagittal (S) planes of motion. Seven clinically relevant kinematic events during the gait cycle were used as a means of extracting discrete kinematic data values from the time series data for each joint. These seven events are important characteristic events in the normal walking cycle. We developed [20], [23], [24] novel preprocessing algorithms to normalize the data to the stance boundaries of heel contact (HC) and toe off (TO), and subdivide the remaining data using five discrete timing events which occur between HC and TO [ordered in Table I(b)]. Finally, the captured data were referenced to the subjects' static relaxed anatomical position, so that zero degrees in the kinematic data relates to the position of that segment when the subject was standing relaxed. Fig. 2 exemplifies the acquired data for the joint of one subject. A 16th order 20-Hz cutoff Butterworth filter was used for noise reduction.

Twenty-seven subjects with PPH were recruited for this study (Table II summarizes the demographics) and their written informed consent was obtained. Each subject walked in a straight



(a)



(b)

Fig. 1. (a) Front and (b) back views of the constructed marker sets and bases.

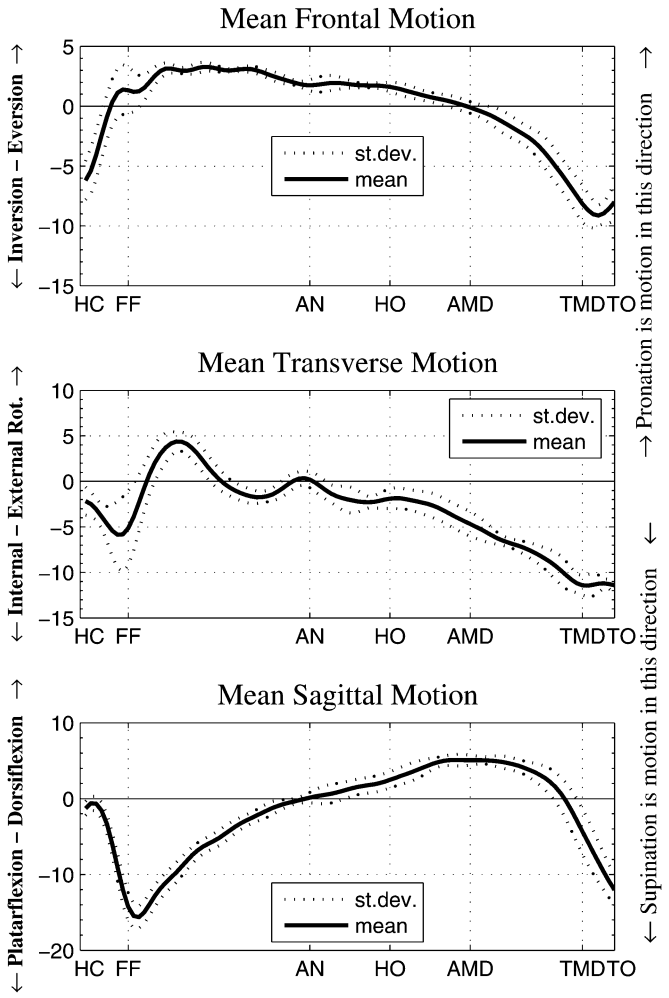


Fig. 2. Example of kinematic angular data showing the frontal, transverse and sagittal motions for the ankle joint complex of a 2-3-4 PPH subject. The mean refers to an averaging over six cycles. The seven stance events described in Table I(b) are indicated by vertical lines.

line and 6 full gait cycles were recorded for each limb. The kinematic data from these 6 trials were averaged to produce single measurements for each limb of each subject. As we have treated each foot as an individual measurement, there are a total of $n = 51$ measurements x_i with $i = 1, \ldots, n$. We have two PPH groups denoted by ω_1 (for PPH under metatarsal heads 1 and 5) and ω_2 (under 2, 3, and 4), with corresponding sizes $n_1 = 24$ and $n_2 = 27$ patterns (3 subjects had PPH only on one foot). The use of 7 events, for each of the 3 planes and for 7 joints, yields a total of 147 individual angle variables. We define each of the x_i patterns to consist of these 147 measurements together with the 5 stance events $\{\mathrm{FF}, \mathrm{AN}, \mathrm{HO}, \mathrm{MAD}, \mathrm{MTD}\}$ expressed in stance time percentage. Thus, we have $d = 152$ dimensions, and the sought classification mapping can be denoted by $\psi : \Re^d \rightarrow \Omega = \{\omega_1, \omega_2\}$, which maps a given measurement x_i to its true target label $t(x_i) \in \Omega$. For convenience, we also denote the entire data set as a collection of patterns and labels $D = \{(x_i, t(x_i)) : i = 1, \ldots, n\}$.

TABLE II
STATISTICS OF THE 27 PPH SUBJECTS EMPLOYED IN THE STUDY. $\mathrm{PPH_{R/L}}$ STANDS FOR THE FOOT SIDE WITH PPH, AND ω_1/ω_2 FOR GROUPS WITH PPH AT METATARSAL HEADS 1-5/2-3-4

	Age (years)	Subjects	Weight (kg)	Height (m)	$\mathrm{PPH_R}$		$\mathrm{PPH_L}$	
					ω_1	ω_2	ω_1	ω_2
Males	58.5	9	84.8	1.76	5	2	6	3
Females	44.8	18	75.7	1.61	8	9	8	10
Overall	49.4	27	78.7	1.66	13	11	14	13

III. Discriminant Analysis and Dimensionality Reduction

A. Gaussian Maximum-Likelihood Classification

Given a measurement $x \in \Re^d$, one way of establishing the means for classification is through the posterior probability $p(\omega_i \mid x)$ and the *Bayes* rule $p(\omega_i \mid x) = p(x \mid \omega_i) \cdot p(\omega_i)/p(x)$ where $p(x \mid \omega_i)$ is termed the conditional likelihood, $p(\omega_i)$ the prior for class ω_i and $p(x)$ the evidence. When these quantities are known for all classes, ψ can rely on the Bayes decision rule to minimize the average decision risk by classifying x as member of class ω_i if $p(\omega_i \mid x) > p(\omega_j \mid x)$ for all $j \neq i$ (assuming equal risks and benefits for incorrect and correct classifications). Typically, it is assumed that the likelihoods follow multivariate Gaussian distributions

$$p(x \mid \omega_i) = \left((2\pi)^d |\Sigma_i|\right)^{-1/2} \cdot e^{-(1/2)(x-\mu_i)' \Sigma_i^{-1} (x-\mu_i)} \quad (1)$$

with mean μ_i and covariance matrix Σ_i for each ω_i. This is because Gaussians have good computational tractability and can

model adequately a wide range of situations. Employing a maximum-likelihood formulation [15], [25], the unbiased estimates can (for the particular distribution) be defined as

$$\hat{\mu}_i = \frac{1}{n_i} \sum_{x:t(x)=\omega_i} x$$

and

$$\hat{\Sigma}_i = \frac{1}{n_i - 1} \sum_{x:t(x)=\omega_i} (x - \hat{\mu}_i)(x - \hat{\mu}_i)'. \quad (2)$$

The final decision rule can make use of the following discriminant function:

$$g_i(x) = -(x - \hat{\mu}_i)' \cdot \hat{\Sigma}_i^{-1} \cdot (x - \hat{\mu}_i) - \log(|\hat{\Sigma}_i|) + 2\log(p(\omega_i)) \quad (3)$$

which enables ψ to classify x as ω_i when $g_i(x) > g_j(x), \forall\, j \neq i$. It can be shown [25] that ψ forms linear pairwise decision surfaces when all covariances are equal to a pooled covariance matrix $\hat{\Sigma}_W = \sum_{i=1,\ldots,c} (n_i - 1)/(n - c)\hat{\Sigma}_i$, where $c = 2$ is the total number of classes. In this case, (3) refers to *linear discriminant analysis* (LDA). When the covariances are individual as in (2), the decision boundaries are hyperquadric surfaces. In this case, we obtain *quadratic discriminant analysis* (QDA). The (asymptotically optimal) estimate of (2) gives a $d \times d$ covariance matrix $\hat{\Sigma}_i$ of rank up to $\min(d, n_i - 1)$. The fact that PPH [16], [20] (refer to Sections I and II for background, problem definition, and data collection details) sample sizes n_i and n are much smaller than the number of kinematic variables $d = 152$, causes singularity to $\hat{\Sigma}_i$ and $\hat{\Sigma}_W$; this implies that the data lies in subspaces of $\Re^d$ and makes $g_i(x)$ non identifiable.

It should be noted that, although for large samples QDA may outperform LDA, for small samples LDA is controlled by fewer parameters and can be less susceptible to model violations. This effect occurs despite the fact that the classes may actually have significantly different covariances. Similar considerations may justify the use of other more restrictive arrangements (e.g., spherical or diagonal covariances). Analogously, compared to more complex classifiers, such as neural networks [15], the linear/quadratic decision surfaces of L/QDA can be more suitable for certain problems. This is because the particular data supports simpler decision boundaries to yield more stable model estimations, rather than the assumptions of Gaussian densities or restrictive covariance structures being correct [26]. This observation is also reflected in the well-known bias-variance tradeoff, where the increased bias of simpler classification boundaries can often be compensated by the lack of much higher variance exhibited by complicated classifiers.

B. Variance-Based Dimension Reduction

An effective method to tackle the singularity or ill-conditioning of the covariances, is to reduce dimensionality to $b < d$ via *principal component analysis* [27], [28]. PCA employs a linear transform of a feature $x \in \Re^d$ to a feature $y = M'x \in \Re^d$ (assuming zero-mean data). The principal idea is to choose an orthogonal matrix M such that the new covariance $\Sigma_Y = E[yy']$ is diagonal. If we define M to be the modal matrix of Σ_X (the matrix with columns the eigenvectors of Σ_X) and Λ its Spectral matrix (the diagonal matrix with the corresponding eigenvalues λ_i) we obtain

$$\Sigma_Y = M'\Sigma_X M = \Lambda. \quad (4)$$

Geometrically, the d eigenvectors represent the principal axes of the hyperellipsoidal data distribution. By keeping the b components with the largest eigenvalues, we ignore the directions with the smallest data variance, since as shown by (4) the variance along the i_{th} component is equal to λ_i. Thus, by substituting M with a $d \times b$ matrix $\tilde{M}$ with columns the first b eigenvectors (assuming an order of $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$), the information loss incurred by projecting x onto the subspace spanned by the b eigenvectors is minimal in the residual mean square error (r.m.s.e.) sense

$$E\left[\left\|x - \tilde{M}\tilde{M}'x\right\|^2\right] = \sum_{i=b+1}^{d} \lambda_i. \quad (5)$$

C. Separability-Based Dimension Reduction

The dimensionality reduction of PCA does not necessarily lead to retaining the most useful data directions. In fact, removing a low-variance subspace may result to loss of significant discriminatory information [15], [28]. An alternative method is to reduce dimensionality by maximizing an explicit measure of class separability. Such a measure can be expressed in terms of the total within-class S_W and between-class S_B scatter matrices

$$S_W = \sum_{i=1}^{c} \sum_{x:t(x)=\omega_i} (x - \mu_i)(x - \mu_i)'$$
$$S_B = \sum_{i=1}^{c} n_i(\mu_i - \mu)(\mu_i - \mu)' \quad (6)$$

with μ and μ_i being the global and the ith group's mean, respectively, as the ratio J_X of their determinants

$$J_X = \frac{|S_B|}{|S_W|}. \quad (7)$$

Since the determinant of a scatter matrix corresponds to the product of variances along the principal data directions, J_X takes higher values when the classes exhibit better mean separation and smaller overall variances. This method is called *linear fisher discriminant* (LFD) [14], [15], [25] and projects each x to $y = M'x \in \Re^b$ with $b \leq \min(d, c - 1)$ dimensions, in a way such that the so called Rayleigh quotient $J_Y(M) \equiv |M'S_BM|/|M'S_WM|$ is maximized for a sought projection M. It turns out that M is the $d \times b$ matrix consisting of the b generalized eigenvectors of $S_BM = S_WM\Lambda$ with the largest eigenvalues from $\Lambda = \text{diag}(\lambda_1, \ldots \lambda_d)$, and also, that this corresponds to an optimal linear discriminant in the mean square error sense.

Unlike PCA, LFD utilizes effectively class information to locate the most discriminating directions, but the drastic dimensionality reduction (since $\text{rank}(S_B) \leq c - 1 = 1$ for our two-class problem) may lose important information. Nevertheless, at the presence of insufficient data, LFD is not always superior [29].

D. Covariance Regularization

Apart from covariance singularity the small sample sizes often cause ill-conditioning, and make the low variance subspace give rise to biased estimations. Using (4) the inverse of the i_{th} class covariance estimate can be written as $\hat{\Sigma}_i^{-1} = M_i \Lambda_i^{-1} M_i'$. This shows that small variances yield significant contributions to the discriminant $g_i(x)$ of (3). To remedy this, an alternative route to that of PCA and LFD is Regularization [15], [26], [30], [31], which attempts to increase the bias of the classifier while reducing its variance by changing the parameters away from their theoretical estimates.

Regularized discriminant analysis (RDA) [32] is a Bayesian classification method like LDA and QDA which replaces the covariance estimate of (2) with a regularized estimate $\hat{\Sigma}_i(\alpha, \gamma)$ controlled by a mixing parameter α and an eigenvalue shrinkage parameter γ both within $[0, 1]$. First, each individual $S_i = (n_i - 1)\hat{\Sigma}_i$ and the pooled $S_W = (n - c)\hat{\Sigma}_W$ scatter matrices are calculated. Then, each $\hat{\Sigma}_i(\alpha)$ is given by the convex combination

$$\hat{\Sigma}_i(\alpha) = \frac{(1-\alpha)S_i + \alpha S_W}{(1-\alpha)(n_i - 1) + \alpha(n - c)} \tag{8}$$

α controls the degree of regularization of the individual covariances toward the pooled estimate. $\alpha = 0$ yields QDA, $\alpha = 1$ LDA, while other values yield intermediate arrangements less strict than LDA. Subsequently, a shrinking of $\hat{\Sigma}_i(\alpha)$ toward a user-defined symmetric positive-definite matrix H is obtained via

$$\hat{\Sigma}_i(\alpha, \gamma) = (1-\gamma)\hat{\Sigma}_i(\alpha) + \gamma \frac{\text{Trace}(\hat{\Sigma}_i(\alpha))}{\text{Trace}(H)} H. \tag{9}$$

The effect of this step is to alter the eigenvalue profile of $\hat{\Sigma}_i(\alpha)$ toward the one dictated by H. A typical choice for H is the identity matrix I; in this case the second sum of the right-side of (9) corresponds to a spherical covariance of radius relative to the mean within-class ω_i variance. From other possible choices [32], experimentation showed better performance using the diagonal matrix of the global variances $H = \text{diag}(\sigma_1^2, \ldots, \sigma_d^2)$ (which is equivalent to a data standardization). Increasing values of γ strengthens the unbiasing effect of the sample estimates. The special case where $\gamma = \alpha = 1$ yields discriminants $g_i(x)$ acting as a minimum Euclidean distance classifier (for equal priors and standardized data), while for $\alpha = 0$ and $\gamma = 1$ a quadratic classifier based on covariances weighted with the average within-class variances.

E. Model Error Estimations

The error rate $\varepsilon(\psi)$ of a classifier ψ, which is defined as the probability of misclassification $P(\psi(x) \neq t(x))$, for some pattern x and its class label $t(x)$, is important for both *model selection* (i.e., choosing the best from a set of available ψ) and *model assessment* (i.e., evaluating generalization performance of the final ψ). Due to the finiteness of the data samples, various methods to calculate estimates $\hat{\varepsilon}(\psi)$ for the true error exist [1], [14], [26], [30].

The *Resubstitution* (or *apparent*) error $\hat{\varepsilon}_A$ which uses all available data D for both training and testing is far too optimistically biased for practical use with small samples. The *Hold-Out* method, which uses a portion of data for training and the rest for testing, is pessimistically biased as only part is utilized for training. The *k-fold Cross-Validation(kCV)* error estimate is calculated by first splitting the data set D to k folds D_{CV}^j, for $j = 1, \ldots, k$, of approximate size n/k and then repeating the experiment k times, each using $k - 1$ folds for training, while testing on the remaining one. Formally, if $\psi(x; D)$ denotes the decision of ψ designed with a set D and tested on x, then the error estimate is defined as

$$\hat{\varepsilon}_{k\text{CV}} = \frac{1}{n} \sum_{j=1}^{k} \sum_{x \in D_{\text{CV}}^j} L\left(\psi(x; D - D_{\text{CV}}^j), t(x)\right) \tag{10}$$

where the 0–1 loss function $L(\omega_u, \omega_v)$ is 1 iff $\omega_u \neq \omega_v$. Although *Leave-One-Out* (LOO) is a very popular instance of kCV (for $k = n$) due to its simplicity and its near zero prediction error bias, its large variance makes it inappropriate, particularly in small sample situations [33]. Moderate, however, values for k have been found in [33], [34] to exhibit acceptably low levels of bias and variance. In this paper, we employ 10CV for model selection. As it is not possible to run a complete kCV which requires all possible $\binom{n}{n/k}$ arrangements, in order to reduce the sensitivity, we use the average for 10 repetitions of 10CV using different fixed permutations.

Bootstrap is a resampling technique more suitable for small samples and of lower variance than kCV. It forms b different samples D_B^j each of size n, where data is chosen from D with replacement. Because the standard bootstrap $B0$ estimator, defined by

$$\hat{\varepsilon}_{B0} = \frac{\sum_{j=1}^{b} \sum_{x \in D - D_B^j} L\left(\psi(x; D_B^j), t(x)\right)}{\sum_{j=1}^{b} |D - D_B^j|} \tag{11}$$

is pessimistically biased, it can be combined with Resubstitution to produce the corrected B632 estimate $\hat{\varepsilon}_{B632} = 0.632\hat{\varepsilon}_{B0} + 0.368\hat{\varepsilon}_A$ [35]. Nevertheless, $\hat{\varepsilon}_{B632}$ can have largely optimistic bias associated to overfitting situations where $\hat{\varepsilon}_A \approx 0$ [33], [34]. Because in this work, due to the low n/d ratio and the model order optimization the classifiers used may overfit, we employ a robust bootstrap variation, the $B632+$ estimator, capable of low bias and variance, defined as

$$\hat{\varepsilon}_{B632+} = (1 - w) \cdot \hat{\varepsilon}_A + w \cdot \hat{\varepsilon}_{B0}. \tag{12}$$

Unlike B632, the combination weight $w = 0.632/(1 - 0.368\hat{R})$ now depends dynamically on an estimate $\hat{R} \in [0, 1]$ of the *relative overfitting rate*. Thus, adjustment between the two extremes of no overfitting and maximum relative overfitting, allows a variation of $\hat{\varepsilon}_{B632} \leq \hat{\varepsilon}_{B632+} \leq \hat{\varepsilon}_{B0}$. Full details of the method can be found in [36], while a recent comparison study for biomedical data in [37]. In this paper, we use $B632+$ for assessment of the final models, using $b = 500$ resamplings, but also report other estimators.

It should be noted that errors estimated for model selection, cannot be reported as final generalization errors, as they are almost certainly optimistic. The reason is that model selection is a methodical optimization procedure which unavoidably exploits the structure of training/testing data of the error estimator. For example, as discussed in [38], a common mistake is to report the minimum error of selecting the optimal number of hidden nodes

in a neural network as an estimate of its generalization accuracy. In this paper, since the small sample sizes do not allow for any training-testing-validation setups, we use 10CV for model selection and $B632+$ for model assessment in order to increase objectivity of comparisons.

F. Feature Selection

Dimensionality reduction can be achieved with either *feature extraction*, (i.e., the projection of the existing variables to a new hopefully more useful subspace, as in PCA and LFD discussed earlier), or *feature selection* [39]–[41] which refers to the selection of a subset of relevant attributes from the existing raw sensor measurements. Although feature extraction can be more powerful due to the possible incorporation of statistical problem properties within the applied projection, feature selection can allow for better problem understanding (simpler clinical gait analysis), sensor cost reduction (not all joints or planes or gait events will need processing) and faster execution times. It should be noted, that although theoretical analysis shows that the inclusion of additional independent features reduces the Bayes error rate, in practice it may lead to worse performances due to optimistic model assumptions and/or inadequate data samples [25].

If we define a subset of indexes of the existing d features $F \subseteq \{1, \ldots, d\}$ and denote with D_F a reduced form of the data set D with all the components of its elements x not contained in F removed, there are two critical factors in establishing a feature selection methodology. The former is the search procedure for finding F, which can be realized by algorithms such as exhaustive search, branch-and-bound, genetic and tree search, sequential and floating search methods [40], [42]. The latter factor concerns the quality measurement criterion for D_F. Two general methodologies exist [39], [40]; *wrappers*, which employ the classifier ψ as a black box so that the sought D_F minimizes an estimation of the misclassification error, and *filters*, which in order to be faster they assess the statistical properties of D_F such as class distance or separability which usually do not relate to ψ directly.

G. Automated Classifier Design

Although for the current problem, we could use a method such as the Sequential Forward Floating Selection (SFFS) method [43] which has been shown in [42] to be a rapid and robust selector, or its extension [44], we propose the use of *genetic algorithms* (GAs), due to their high flexibility, which is needed in the current work as will be explained later. GAs [45], [46] are simple, directed, stochastic optimizers, which exhibit considerable robustness in the exploration of complex solution spaces, without the need for restrictive assumptions of other optimization methods, such as continuity, differentiability, unimodality or low combinatorial properties. GAs require a memory structure, the population of individuals, which contains a finite quantity of potential solutions (phenotypes). Each solution is encoded by a well-defined data-structure (genotype) allowed to evolve through specific genetic operations, such as recombination and mutation. Simulated evolution based on the doctrine of survival of the fittest, forces the better solutions to disseminate their genes to the subsequent generations more frequently. The higher the fitness of an individual, the greater the chance it has to breed and pass on its sound genome, while the less fit individuals are progressively doomed to extinction.

GAs have been used successfully for feature processing in different ways. Recent works include [47] where features are extracted via optimal rotations preceded by PCA-based dimensionality reduction using nearest distance classification. In [48] an optimal feature scaling is sought together with a feature selection scheme, applied to a k-nearest neighbor (kNN) classifier. Concerning feature selection, [49] performed a comparison of different types of GAs using the Naive Bayes classifier, [50] successfully applied GAs to a colonography application using an SVM committee, while [51] applied GAs to EEG channel selection for brain computer interfaces with SVMs.

The reason for requiring the modeling flexibility and implementational simplicity of GAs, is that we design an integrated classifier, which in addition to feature selection simultaneously fine-tunes the various hyperparameters of the classifier (for example, the regularization parameters α and γ for RDA). Such a scheme allows a straightforward training procedure which exploits properties of the data set as well as properties of the classifier in a unified manner. An analogous approach was adopted successfully in [48] where for the kNN classifier the parameter k was encoded in the chromosome together with feature processing information.

In order to reflect the classification accuracy directly into the feature selection process, we use a wrapper style quality assessment for D_F. This makes use of (10) to minimize the tenfold CV error estimate $\hat{\varepsilon}_{k\mathrm{CV}}$ measured on the reduced data set D_F. Additionally, in order to drive the selector toward the fast recovery of a few relevant variables, we impose a second objective that maximizes feature sparsity. In essence, the implemented GA solves the following mixed-integer maximization problem:

$$\max_{\substack{F\subseteq\{1,\ldots,d\}\\ \alpha,\gamma\in[0,1]}} \underbrace{1-\hat{\varepsilon}_{k\mathrm{CV}}\big(\psi(\alpha,\gamma);D_F\big)}_{f_1} + w_1 \underbrace{\left(1-\frac{|F|}{d}\right)^{w_2}}_{f_2} \quad (13)$$

where $w_1 \geq 0$ is a user-defined weight to scalarise the two objectives of accuracy (f_1) and sparsity (f_2), and $w_2 \geq 1$ another weight to force a rapid nonlinear feature decrease. Each population solution is encoded in two parts. In the feature selection part d bits are used to signify the use or elimination of each feature. In the variable part, $v_l \times v_n$ bits are used, where v_l is the number of bits used per variable and v_n the number of classifier variables. For real-to-binary mapping we use a Gray-coding scheme [45] to have positional continuity between the two representations. For all real parameters we use $v_l = 10$ bits, which give discrete steps of $1/(2^{10}-1)$. The extremely high size of the combined solution space is evident, when, for $d = 152$ features and $v_n = 2$ variables, the algorithm has to scan efficiently $2^d - 1 + 2^{v_n \cdot v_l} \approx 10^{45}$ possible solutions.

The algorithm is implemented according to the *mod-GA* scheme [46], also found to work efficiently for more complicated type of problems [52]. At every generation, from a population of a total of g_{size} members, g_{child} offspring are created and $g_{\mathrm{size}} - g_{\mathrm{child}}$ survivors are selected probabilistically depending on their fitness values. Algorithm components include, linear ranking to control the selective pressure via a

parameter $\eta_{\max}$ [45], an elitism operator that copies the best member to the next generation, and a no-duplicates scheme for better allotment of population slots. The two operators of crossover and mutation are applied independently with probability p_c and $1 - p_c$, respectively. When mutation is chosen, each allele is flipped with probability p_m proportional to the chromosome length. We use a uniform crossover to randomly choose bits from either parent, and terminate evolution when the top member fitness evaluation from (13) stagnates for g_{term} generations.

The typical procedure is to initialize the $g_{\text{size}} \times (d + v_l \times v_n)$ population matrix in random. However, for the left $g_{\text{size}} \times d$ part, we make use of an informed scheme to assign a bit value of 1 to the ith gene of each chromosome according to a probability value p_i. This is calculated using a class separability information

$$J_i = \frac{(\mu_{i,1} - \mu_{i,2})^2}{\sigma_{i,1}^2 + \sigma_{i,2}^2} \tag{14}$$

where $\mu_{i,\omega}$ and $\sigma_{i,\omega}^2$ are the mean and variances of the ith feature for class ω; note, that this is a one-dimensional (1-D) analogue of (7). Subsequently, all J_i are sorted and p_i are assigned through a linear scaling within $[0.1, 0.9]$. This initialization scheme has been shown to accelerate feature selection in [53], where a divergence measure was used instead of J_i with a naive Bayes classifier.

IV. Experimentations and Results

A. Testing LDA and QDA

The first methods to evaluate are the LDA and QDA. As the entire set of $d = 152$ features is used, the associated covariances become singular. In these cases, a small number is added to the zero eigenvalues, large enough to permit stable inversion in (3). Equal priors $p(\omega_i) = 0.5$ are used in all the experiments of this section. As can be seen in Table III(a), both methods behave as random decision makers. This shows the need for dimensionality reduction to eliminate redundant and noisy features and increase numeric stability. For comparisons within Table III, we mainly use the $\hat{\varepsilon}_{B632+}$ estimation, while $\hat{\varepsilon}_{10\text{CV}}$ is typically used for model selection (where denoted by "$*$") or reported for completeness. The pessimistic $\hat{\varepsilon}_{B0}$ and the optimistic $\hat{\varepsilon}_{B632}$ are also reported for completeness. For reasons discussed in Section III-E we consider the $B632+$ to be the most objective one for comparisons, while use kCV for model selection. Note, that because they have different properties, such as the higher variance of kCV, there is not always high correlation between bootstrap and kCV errors, and this is accentuated by the small sample sizes. In any case, we keep model selection and model evaluation estimators separate in order to avoid introduction of bias (see also discussions [38], [40]). Alternatively, we could use different partition number k with kCV for the two tasks.

B. Evaluation of Feature Extraction

Table III(b) reports the error of LDA and QDA using PCA for feature extraction. However, as PCA can be used to eliminate more dimensions than the zero-variance ones, such as those considered redundant or noisy, we run a model search to select those $b < d$ that minimize the r.m.s.e. in (5). In this case, the new features become the coefficients of expressing the gait angles and times as linear combinations of the b principal directions. The search is depicted in Fig. 3(a), where the $\hat{\varepsilon}_{10\text{CV}}$ errors for the first $b = 1, \ldots, 45$ dimensions are evaluated. The optimum model, that is the one with the lowest $\hat{\varepsilon}_{10\text{CV}}$, is at $b = 32$ and 1 for the LDA and QDA respectively. It can be seen that PCA-based reduction is not adequate as accuracy is still less than 65%. This shows, that for the particular problem, the removal of low-variance subspace does not increase discriminatory information (see also Sections III-B and III-C).

The more promising dimensionality reduction method LFD, projects each PPH pattern to a 1-D space since we have $c = 2$ PPH groups. Subsequently, we subject these projections to a nearest (Euclidean) class mean classifier. However, because of the small sample size issue, the within-class scatter S_W defined in (6) is singular with a maximum rank of $\min(d, n - c)$ and prevents solution of the eigensystem $S_B M = S_W M \Lambda$ (see Section III-C). This well-known LFD drawback has received great attention to other research areas, such as face image recognition, and has resulted to the derivation of various correction schemes. To provide a comparison between such schemes, we have implemented ten variants that cope with S_W singularity. Table IV

TABLE III
Misclassification Rate (%) Estimations for all Implemented Methods in Six Groups. The First Three Columns Report Bootstrap Estimations for 500 Resamplings. The $B632+$ Errors (Boldfaced) Are Used for Final Model Comparisons. The Last Column Reports the 10-Fold CV Error Averaged Over 10 Distinct Sample Permutations. Asterisks Denote CV Errors Minimized in a Model Selection Search

	Algorithm	Errors			
		$\hat{\varepsilon}_{B0}$	$\hat{\varepsilon}_{B632}$	$\hat{\varepsilon}_{B632+}$	$\hat{\varepsilon}_{10CV}$
(a)	LDA	48.61	30.72	**47.92**	50.59
	QDA	51.55	49.18	**51.55**	50.98
(b)	LDA+PCA	39.92	25.24	**35.79**	12.75*
	QDA+PCA	37.17	37.20	**37.20**	37.45*
(c)	$\text{LFD}_{\text{shrink}}$	21.89	13.84	**16.51**	13.92*
	LFD_{svd}	36.82	23.27	**31.96**	25.69*
	$\text{LFD}_{\text{pseudo}}$	34.89	22.05	**29.70**	26.47
	LFD_{diag}	34.89	22.05	**29.70**	26.47
	$\text{LFD}_{\text{direct}}$	34.32	32.52	**32.82**	34.31
	$\text{LFD}_{\text{total}}$	22.00	13.91	**16.60**	19.22
	LFD_{null}	22.00	13.91	**16.60**	19.22
	$\text{LFD}_{\text{null2}}$	22.00	13.91	**16.60**	19.22
	LFD_{pca}	22.43	14.18	**16.99**	18.63
	$\text{LFD}_{\text{pcanull}}$	22.00	13.91	**16.60**	19.22
(d)	RDA	27.36	17.29	**21.67**	14.71*
	RDA+PCA	21.89	13.84	**16.50**	10.98*
	$\text{RDA+GA}_{\text{p2}}$	26.34	16.65	**20.67**	14.51*
	$\text{RDA+GA}_{\text{p4}}$	25.80	16.31	**20.15**	13.73*
(e)	LDA+ScFS	22.21	19.81	**20.12**	15.69*
	QDA+ScFS	22.40	19.93	**20.26**	15.88*
	LDA+SFS	18.01	12.82	**13.95**	05.88*
	QDA+SFS	34.24	21.65	**28.97**	07.06*
	LDA+SFFS	13.24	08.37	**09.27**	00.00*
	QDA+SFFS	33.07	20.90	**27.65**	03.73*
(f)	$\text{RDA+GA}^{\text{f4}}$	10.61	08.87	**08.99**	06.47*
	$\text{RDA+GA}^{\text{f7}}$	09.40	07.38	**07.54**	03.92*
	$\text{RDA+GA}^{\text{f9}}$	06.88	05.07	**05.19**	02.35*
	$\text{RDA+GA}^{\text{f12}}$	06.14	03.88	**04.07**	01.18*

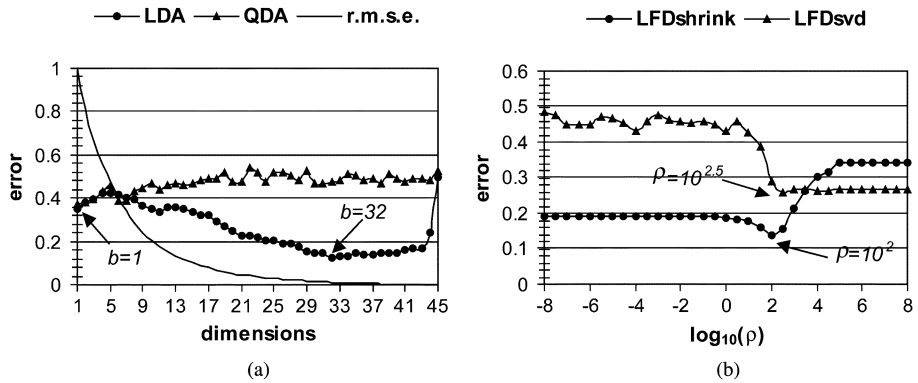


Fig. 3. $\hat{\varepsilon}_{10\mathrm{CV}}$ plots for varying model parameters, with the optimal ones pointed by arrows. (a) LDA and QDA classifiers for different feature cardinalities b extracted via PCA. The exponentially decreasing r.m.s.e. is also shown. (b) $\mathrm{LFD}_{\mathrm{shrink}}$ and $\mathrm{LFD}_{\mathrm{svd}}$ projection methods for different values of the regularizer ρ.

TABLE IV
THE IMPLEMENTED LFD VARIANTS DESIGNED FOR HANDLING INSUFFICIENT DATA

LFD$_{\mathbf{shrink}}$ [31]	a popular way of covariance stabilisation using $S_W + \rho \cdot I$ (ρ>0), to shrink the covariance towards a spherical one and impose magnitude control on the columns of M.
LFD$_{\mathbf{svd}}$ [54]	using Singular Value Decomposition (SVD), S_W is replaced by $U \cdot diag(s_1,\ldots,s_{rank(S_W)},\rho,\ldots,\rho) \cdot V$, where s_i are the nonzero singular values of S_W, U and V the orthogonal matrices and ρ >0.
LFD$_{\mathbf{pseudo}}$ [55]	S_W^{-1} is replaced by its pseudoinverse S_W^{+} to obtain the eigensystem $S_W^{+}S_B M = M\Lambda$.
LFD$_{\mathbf{diag}}$ [14]	$\lvert M'S_B M\rvert$ is maximised subject to $M'S_W M = I$; the projection becomes $M = M_W \Lambda_W^{-1/2} M_B$, where M_W and M_B contain the eigenvectors of S_W and $\Lambda_W^{-1/2} M_W' S_B M_W \Lambda_W^{-1/2}$ of nonzero eigenvalues and Λ_W are the eigenvalues of S_W.
LFD$_{\mathbf{direct}}$ [56]	the reverse of $\mathrm{LFD}_{\mathrm{diag}}$ since it discards the nullspace of S_B first. $\lvert M'S_W M\rvert$ is minimised subject to $M'S_B M = I$.
LFD$_{\mathbf{total}}$ [57]	the same as $\mathrm{LFD}_{\mathrm{diag}}$, but S_W is replaced by the total scatter $S_T = S_W + S_B$ so that an equivalent quotient $\lvert M'S_B M\rvert / \lvert M'S_T M\rvert$ is maximised.
LFD$_{\mathbf{null}}$ [58]	the nullspace basis K_W of S_W is calculated and then the eigenvalues of $K_W K_W' S_B K_W K_W'$ are found; hence, $\lvert M'S_B M\rvert$ is maximised subject to $M'S_W M = 0$.
LFD$_{\mathbf{null2}}$ [59]	the nullspace of $S_T = S_W + S_B$ is removed so that the nullity of the new S_W becomes *rank(S_B)*; then the nullspace of the new S_W is used to form M so that $M'S_W M = 0$.
LFD$_{\mathbf{pca}}$ [60],[61]	first, PCA projects the data to the subspace spanned by the *rank(S_W)* eigenvectors of S_T with the highest eigenvalues and then proceeds with the eigensystem as normally.
LFD$_{\mathbf{pcanull}}$ [62]	PCA is used to remove the nullspace of S_T; then the nullspace and rangespace of the new S_W are processed separately to produce a composite projection M of typically *2·rank(S_B)* dimensions.

outlines these variants and their fundamental algorithmic characteristics.

From these variants, only the $\mathrm{LFD}_{\mathrm{shrink}}$ and $\mathrm{LFD}_{\mathrm{svd}}$ are parametric as they depend on a user-defined regularization parameter ρ. As before, we fine-tune the model by seeking the value that minimizes the misclassification rate $\hat{\varepsilon}_{10\mathrm{CV}}$. Fig. 3(b) plots the error for varying ρ within $10^{[-8:0.5:+8]}$. It can be seen that the best models are the ones corresponding to values of $\rho = 10^2$ and $10^{2.5}$ for $\mathrm{LFD}_{\mathrm{shrink}}$ and $\mathrm{LFD}_{\mathrm{svd}}$, respectively. Final errors for all LFD variants are shown in Table III(c). The fact that the error of $\mathrm{LFD}_{\mathrm{shrink}}(\hat{\varepsilon}_{B632+} = 16.51)$ is half the error of $\mathrm{LFD}_{\mathrm{svd}}(31.96)$, shows that for the particular problem, regularization toward a spherical covariance is more beneficial than regularization restricting the Frobenius norm.

$\mathrm{LFD}_{\mathrm{pseudo}}$ and $\mathrm{LFD}_{\mathrm{diag}}$ both give a high $\sim 30\%$ error, as both discard the nullspace of S_W, which is apparently important for discrimination [58], [59]. $\mathrm{LFD}_{\mathrm{direct}}$ is (as expected) the worst with an error of 32.82%, since the dimensionality loss from the S_B nullspace removal is very high as $\mathrm{rank}(S_B) \leq c - 1 = 1$. $\mathrm{LFD}_{\mathrm{total}}$, $\mathrm{LFD}_{\mathrm{null}}$, $\mathrm{LFD}_{\mathrm{null2}}$, and $\mathrm{LFD}_{\mathrm{pcanull}}$ produce equal low error rates of 16.6% as they are similar in the sense of taking into account the nullspace of S_W in various ways despite some fundamental algorithmic differences (equality of their error rates is owed to the small number of data classes). Finally, $\mathrm{LFD}_{\mathrm{PCA}}$ also produces a small error, as the existence of two classes does not leave margin for much loss of discriminatory information (one dimension reduced additionally to the nullspace of S_T). Overall, from this comparison, it can be seen that nullspace methods are better, and without the need to fine-tune any parameters. Overall, LFD methods achieve considerable dimensionality reduction from 152 to 1 dimension only along a direction that maximizes class separability, but this seems adequate for PPH discrimination with best accuracies of $\sim 83\%$.

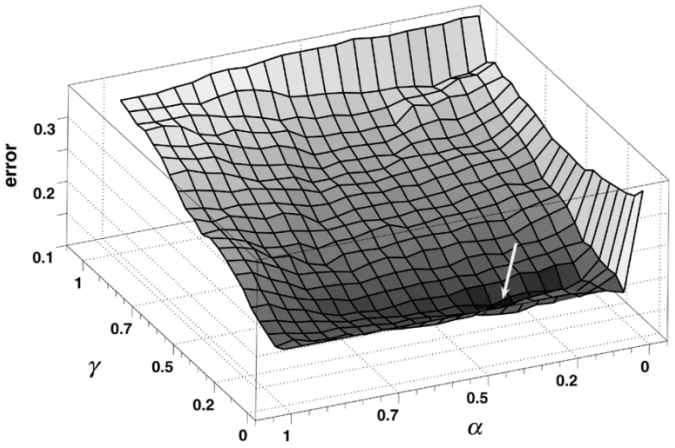


Fig. 4. Error surface of RDA using the entire $d = 152$ features with varying values of α and γ. The minimum occurring at $(\alpha, \gamma) = (0.3, 0.1)$ is pointed by the arrow.

C. Testing Regularization

The positive effect of tackling covariance singularity and ill-positioning via regularization, is visible in Table III(d), where despite the use of all $d = 152$ raw features the misclassification rate remains at a reasonably low level of 21.67%. Model selection is performed by error estimation on each (α, γ) point of a 21×21 uniformly spaced grid $[0 : 0.05 : 1]^2$ (when $\gamma = 0$ caused instability it was replaced by a small quantity of 10^{-10}). Fig. 4 shows the entire error surface with its lowest elevation $\hat{\varepsilon}_{10\text{CV}} = 14.71$ at $(\alpha, \gamma) = (0.3, 0.1)$.

We also test the potentially beneficial effect of combining reduced feature extraction and regularization. To do this, we add to the previous setup an external loop that preapplies PCA with varying number of eigenvectors $b = 1, \ldots, 45$. To accelerate computation, a smaller 11×11 grid $[0 : 0.1 : 1]^2$ is used for (α, γ). RDA + PCA in Table III(d) has an error of 16.50% which is lower than RDA. Fig. 5(a) plots the best error and the associated α and γ values, for each slice b of the search space, where the overall minimum occurs at $(b, \alpha, \gamma) = (33, 0.6, 0.1)$. Fig. 5(b) shows the entire surface at slice $b = 33$.

Finally, we test two more RDA implementations, which employ the GA in the outer loop to estimate α and γ. In the first one, $\text{RDA} + \text{GA}_{p2}$, the two parameters take resolution much higher than the previously used grids. The 10-bits encoding (see Section III-G) allows resolution at $1/(2^{10} - 1)$. As shown in Table III(d) both $\hat{\varepsilon}_{B632+}$ and $\hat{\varepsilon}_{10\text{CV}}$ are slightly lower, as the GA has a larger search space for minimization. The optimum parameters now are $(\alpha, \gamma) = (0.3529, 0.1457)$ which are near the previously found RDA ones. The second version, $\text{RDA} + \text{GA}_{p4}$, uses the same resolution, but in addition we have allowed separate α_i and γ_i values for each ith class. Traditionally, RDA uses common regularization parameters for efficiency [32]. The optimum values for $(\alpha_1, \alpha_2, \gamma_1, \gamma_2)$ are (0.5445,0.2972, 0.1505,0.1574). The fact that the error of 20.15% is slightly better than $\text{RDA} + \text{GA}_{p2}$ and RDA shows that decoupling the parameters may give slightly better performance for the particular problem. To allow for maximum flexibility, in all GA-based searches used henceforth we encode 4 parameters for RDA.

D. Evaluation of Feature Selection

As mentioned in Section III-F, due to the high combinatorial complexity of selecting a subset F from the d features various fast search techniques exist. Here, we have implemented for comparison three different search methods [28], [42], [43]. ScFS is the simplest one and performs scalar filter search by sorting all features according to a class separability measure. Equation (14) is used for this measure, and the features with the highest scores are taken to form each subset. The second method, the sequential forward search (SFS) one, iteratively adds the next feature which together with all previously selected ones maximizes a score. For the evaluation of this score, we have used LDA and QDA in a wrapper approach. To alleviate the nesting effect of SFS (once a feature is added it cannot be removed), we also implemented the sequential floating forward search (SFFS) method, which employs backtracking to reassess and replace past decisions.

Fig. 6(a) plots the class separability scores for all features individually, where only a small fraction of the features has relatively high scores. When the output of ScFS for an increasing portion of selected features F is subjected to LDA and QDA in Fig. 6(b), it can be seen that the addition of more than one features increases misclassification rate. This shows that features cannot be examined independently and that a filter-based criterion cannot reflect the classification performance. However, one feature only can manage $\sim 80\%$ classification accuracy in Table III(e); this is the angle of the Rearfoot Complex joint at Foot Flat event on the Sagittal plane (see Table I). The results for LDA/QDA + SFS and LDA/QDA + SFFS are shown in Table III(e). The optimum number of selected features are shown by the arrows in Fig. 6(b). Overall SFFS gives better results, which is expected as its search is more thorough than SFS due to nesting avoidance. The lowest error so far is that of LDA + SFFS with $\hat{\varepsilon}_{B632+} = 9.27\%$, for $|F| = 19$ features. Despite the low $\hat{\varepsilon}_{10\text{CV}}$ of QDA the $\hat{\varepsilon}_{B632+}$ errors remain high. Further tests showed that the correlation between the two estimators is very poor for LDA and especially for QDA which is more numerically susceptible to small samples due to the separate parameterization each class requires.

E. Results of Feature Selection With Regularization

Here, we test the proposed RDA + GA method (Section III-G) which uses RDA to regularise the covariance estimates, and a GA to select the optimal feature subset F and the four RDA parameters. Table V contains the GA parameters used. The tradeoff between maximizing classification rate f_1 and sparsity f_2 within (13) is regulated by w_1. Table III(f) presents four distinct Pareto solutions for different values of w_1 with 4, 7, 9 and 12 features (x in $\text{RDA} + \text{GA}^{fx}$ denotes the final $|F|$). The general observation from our experiments is that lower w_1 values obtain more features but with better accuracy. However, this holds for small $|F|$ as the existence of many features reduce the generalization rates.

Overall the resulting classifiers outperform the previous ones. Even with 4 features, $\text{RDA} + \text{GA}^{f4}$ has an accuracy of 91.01%, which is better than the previously best LDA + SFFS with 90.73% accuracy and 19 features. Experimentation

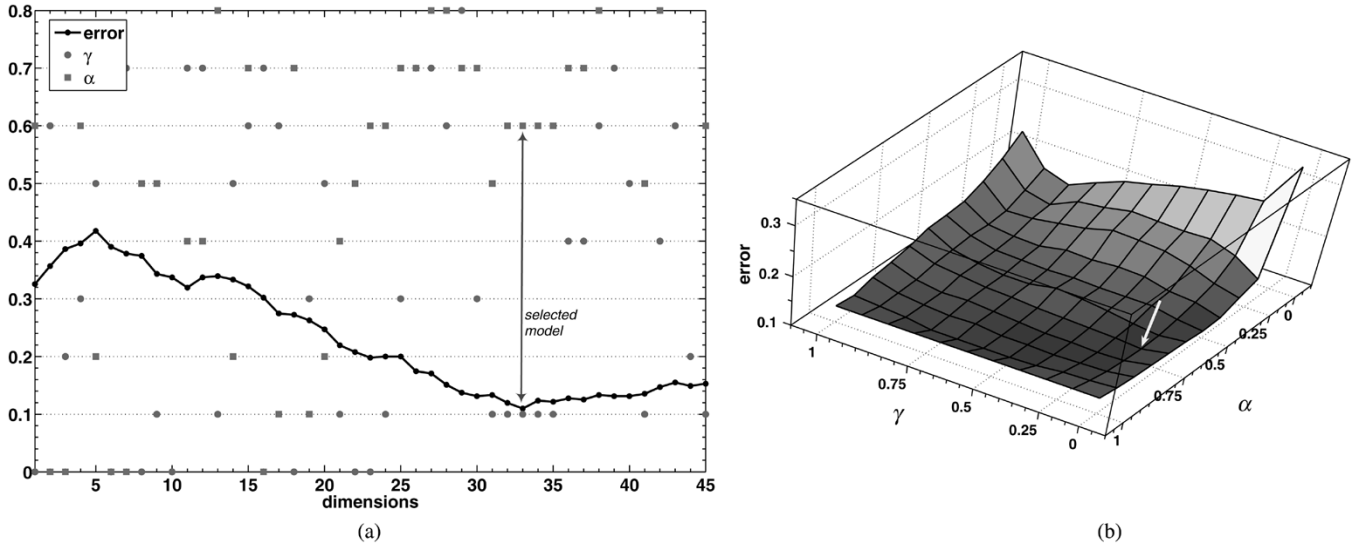


Fig. 5. (a) Plot of the minimum $\hat{\varepsilon}_{10\mathrm{CV}}$ and its associated α and γ values of RDA preceded by PCA. The optimum model is shown by the arrow at $b = 33$ dimensions with $(\alpha, \gamma) = (0.6, 0.1)$ and an error of 10.98%. (b) Error surface plot at slice $b = 33$, where the arrow points at the optimum $(0.6, 0.1)$.

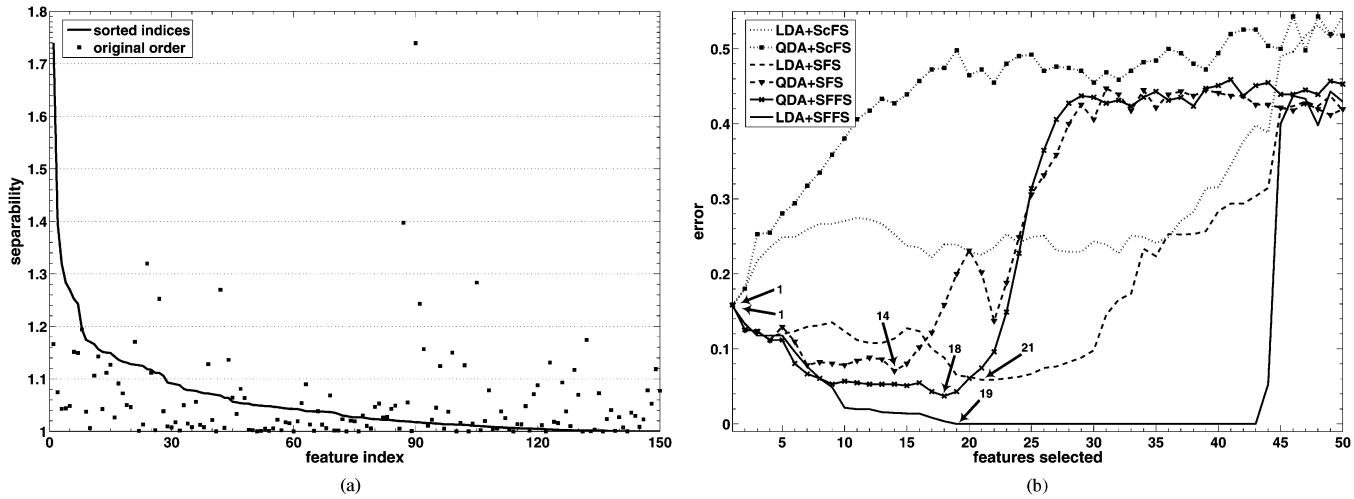


Fig. 6. (a) Measures of class separability for each feature individually. (b) Error plots for LDA and QDA preceded by feature selection via ScFS, SFS, and SFFS for varying number of selected features. The best models are shown by arrows annotated by the number of optimal dimensions.

also showed the appealing fact that the use of small feature subsets in conjunction with regularization increases the correlation between bootstrap and CV error estimators, and also decreases overfitting ratios as the difference between B632 and $B632+$ in Table III(f) becomes small. The best classifier, the $\mathrm{RDA} + \mathrm{GA}^{f12}$ with 12 features and an accuracy of 95.93%, uses a smaller $w_1 = 0.2$ to give lower optimization weight to sparsity f_2 and more to accuracy f_1. Table VI contains the exact features for the four solutions, their regularization parameters α_1, α_2, γ_1, γ_2, and the w_1 weights.

Fig. 7 shows statistics from the gradual reduction of features and errors over the optimization course for $\mathrm{RDA} + \mathrm{GA}^{f9}$ experiments, where the speed of the convergence in under 100 generations is apparent. Fig. 8 shows the population matrices of the best individuals from all generations and the initial population. It can be seen that the initial sparsity achieved by the heuristic of initializing the genes using (14) manages a dramatic speed acceleration. The figure also illustrates the different types of search deployed in the left and right part of the population matrices, where the 152 feature bits and the 40 bits for α and γ, respectively, are encoded.

TABLE V
THE GA PARAMETERS USED FOR ALL EXPERIMENTS

g_{size}	g_{child}	η_{max}	p_c	p_m	g_{term}	w_2
100	40	1.6	0.6	$3/(d+v_n \cdot v_l)$	50	1.5

V. CONCLUSION

This paper has tested an extensive range of dimensionality reduction and robust classification techniques for linking PPH and functional biomechanical foot data. As expected, the classical Discriminant Analysis techniques LDA and QDA performed very poorly because of the large ratio of measurement variables over available data samples; hence, they do not seem very suitable for such situations. The PCA and LFD methods that reduce dimensions by extracting new linear features performed better, since they eliminate redundant information and make the parameterization of the classifiers more tractable. The fact that

TABLE VI
EXAMPLES OF FEATURES AND PARAMETERS FOUND BY THE PROPOSED RDA + GA METHOD FOR DIFFERENT VALUES OF w_1. EACH SELECTED FEATURE IS IDENTIFIABLE BY ONE OF THE 147 *JOINT-EVENT-PLANE* TRIPLETS (SEE TABLE I) OR ONE OF THE FIVE EVENT TIME PERCENTAGES

RDA+ GA^{f4} (w_1=0.8)	**Joint**	MT	RFC	TF	TF	$\alpha_{1,2}$	$\gamma_{1,2}$
	Event	FF	FF	FF	FF	0.56	0.85
	Plane	F	S	F	S	0.02	0.37

RDA+ GA^{f7} (w_1=0.6)	**Joint**	MT	R1	MP1	RFC	MH	TF	TF	$\alpha_{1,2}$	$\gamma_{1,2}$
	Event	FF	HC	TO	FF	MAD	FF	FF	0.459	0.643
	Plane	F	T	T	S	T	F	S	0.837	0.456

RDA+ GA^{f9} (w_1=0.4)	**Joint**	AC	MT	MT	R1	RFC	RFC	MH	TF	MAD	$\alpha_{1,2}$	$\gamma_{1,2}$
	Event	AN	FF	MTD	HC	FF	AN	AN	FF	time	0.641	0.179
	Plane	T	F	S	T	S	F	S	S	%	0.289	0.108

RDA+ GA^{f12} (w_1=0.2)	**Joint**	AC	R1	R1	MP1	MP1	RFC	RFC	RFC	RFC	MH	TF	MAD	$\alpha_{1,2}$	$\gamma_{1,2}$
	Event	AN	HC	AN	FF	TO	FF	AN	MAD	MTD	AN	FF	time	0.413	0.097
	Plane	S	T	S	T	T	S	F	S	S	T	S	%	0.973	0.067

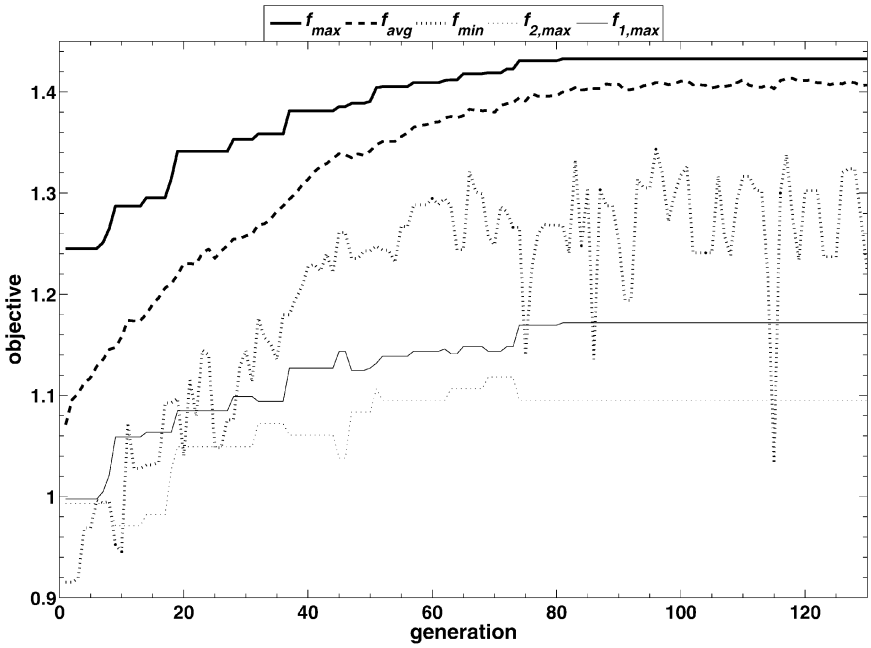


Fig. 7. Plots of the best, average, and worst values of the maximizing objective $f = f_1 + w_1 \cdot f_2$ for each of the 130 generations of $\mathrm{RDA} + \mathrm{GA}^{f9}$. The (scaled by 1.2 for clearer presentation) accuracy (f_1) and sparsity (f_2) objectives of the top member are also shown.

LFD techniques outperformed PCA-based discrimination, was also expected since LFD reduce the data by choosing more discriminant subspaces, while as mentioned in Section II, PCA simply retains high variance subspaces. Overall, from all the tested nonparametric LFD variations, the nullspace ones perform better which is a reasonable behavior since nullspace information has to be exploited while selecting the discriminant subspaces (see LFD references for theory details). However, the availability of only two classes cannot distinguish further between them as they give similar results. On the other hand, RDA (using either a grid search or a GA for model selection) also performed better than its unregularized counterparts LDA and QDA, which is sensible as RDA deviates from the theoretical covariance estimates of LDA and QDA in an attempt to improve classification. Additionally, RDA is shown to be further improving with a preapplication of PCA. This is owed to the combined benefit of the two and shows that regularization on reduced subspaces is more responsive. In this way, RDA becomes equivalent to the best LFD methods. Overall, the feature selection techniques ScFS, SFS and SFFS which retain specific variables, are very promising since they present a much simpler problem to the classifier without the need for manipulation of all available variables. For this reason, the best one, LDA + SFFS, shows improved performance over either feature extraction or regularization both of which process in different ways all available variables. All feature selection methods work better with LDA, probably because QDA provides unnecessarily more complex decision boundaries than LDA. The SFFS is better than ScFS and SFS since it treats the available features more globally. The RDA + GA method we proposed achieved the highest classification accuracy. It was shown to manage an $\sim 96\%$ accuracy with less than 10% of the available features. The reason for its higher accuracy is that its search space is much broader than other methods, such as LDA + SFFS. On its own, RDA is a general case of LDA or QDA, and the GA a general case of SFS or SFFS with the additional capability of fine-tuning the RDA parameters. Thus, theoretically RDA + GA has more chances of producing better results by combining algorithmically the power of corrected covariance estimates via regularization and pattern space simplification via variable reduction.

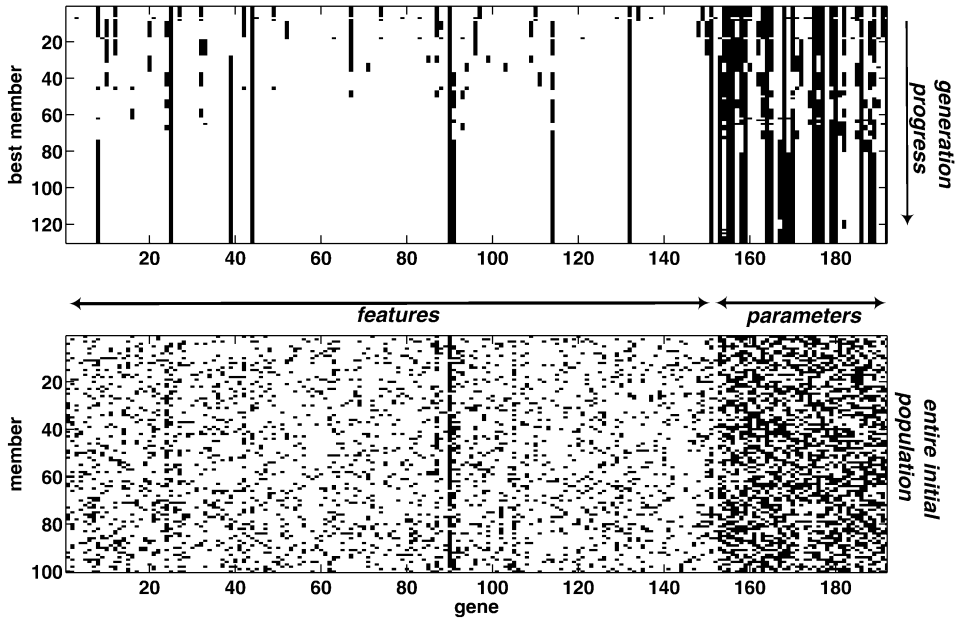


Fig. 8. Population matrices where black pixels denote genes set to 1. Solutions are represented by 152 bits for the feature set F and 40 for the RDA parameters $\alpha_{1,2}$ and $\gamma_{1,2}$. (top) the 130×192 matrix with the ith row containing the best chromosome of the ith generation. (bottom) the 100×192 matrix with the ith row containing the ith member of the sorted initial (generation 0) population.

The experimented and proposed methods are generic and applicable to gait investigations, other than the lesion classification and the type of kinematic data employed here. This is because the work was focused toward providing a robust classification analysis for small sample size situations, which is common in biomechanical and gait sciences. Additionally, the employed feature selection incurs a sensor reduction which is important, not only because it reduces the costs and accelerates the data collection procedures, but also because it provides a simpler clinical model for examining and understanding the interconnections between gait and pathology, through the specification of a few key variables.

In the context of PPH, we have provided some original findings, that establish a correlation between the foot pressures induced by irregularities of the foot motion and PPH formation. This has the following benefits. First, the link of theoretical interest between foot kinematics and PPH is demonstrated; this reinforces the claim of causal relationship between motion and PPH formation. Second, from the biomechanical point of view, the proposed multisegment foot model is ascertained for its sensitivity and richness of kinematic information, which may be valuable for other types of gait analyses. Third, concerning the clinical factors, it provides promising directions for constructing automated patient screening systems which detect lesions prior to their observable formations. This can have significant impact to the danger assessment and prevention of other types of debilitating lesions, such as neuropathic diabetic foot ulcerations which can lead to limb amputation.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their comments and suggestions.

REFERENCES

[1] A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical pattern recognition: A review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 1, pp. 4–37, Jan. 2000.

[2] T. Chau, "A review of analytical techniques for gait data. Part 1: Fuzzy, statistical and fractal methods," *Gait Posture*, vol. 13, pp. 49–66, 2001.

[3] ——, "A review of analytical techniques for gait data. Part 2: Neural network and wavelet methods," *Gait Posture*, vol. 13, pp. 102–120, 2001.

[4] R. Begg and J. Kamruzzaman, "A machine learning approach for automated recognition of movement patterns using basic, kinetic and kinematic gait data," *J. Biomech.*, vol. 38, pp. 401–408, 2005.

[5] S. D. Prentice, A. E. Patla, and D. A. Stacey, "Artificial neural network model for the generation of muscle activation patterns for human locomotion," *J. Electromyogr. Kinesiol.*, vol. 11, pp. 19–30, 2001.

[6] R. E. Bekka, S. Boudaoud, and D. Chikouche, "The use of neural network system in the identification of motor unit characteristics from surface detected action potentials: A simulation study," *J. Neurosci. Methods*, vol. 116, pp. 89–98, 2002.

[7] G. Cheron, F. Leurs, A. Bengoetxea, J. P. Draye, M. Destree, and B. Dan, "A dynamic recurrent neural network for multiple muscles electromyographic mapping to elevation angles of the lower limb in human locomotion," *J. Neurosci. Methods*, vol. 129, pp. 95–104, 2003.

[8] M. A. Perez and M. A. Nussbaum, "Principal components analysis as an evaluation and classification tool for lower torso sEMG data," *J. Biomech.*, vol. 36, pp. 1225–1229, 2003.

[9] M. Tingley, C. Wilson, E. Biden, and W. R. Knight, "An index to quantify normality of gait in young children," *Gait Posture*, vol. 16, pp. 149–158, 2002.

[10] S. G. White and P. J. McNair, "Abdominal and erector spinae muscle activity during gait: The use of cluster analysis to identify patterns of activity," *Clin. Biomech.*, vol. 17, pp. 177–184, 2002.

[11] A. Bertani, A. Cappello, M. G. Benedetti, L. Simoncini, and F. Catani, "Flat foot functional evaluation using pattern recognition of ground reaction data," *Clin. Biomech.*, vol. 14, pp. 484–493, 1999.

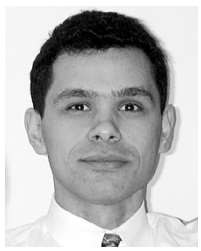
[12] B. Najafi, K. Aminian, A. Paraschiv-Ionescu, F. Loew, C. J. Bulla, and P. Robert, "Ambulatory system for human motion analysis using a kinematic sensor: Monitoring of daily physical activity in the elderly," *IEEE Trans. Biomed. Eng.*, vol. 50, no. 6, pp. 711–723, Jun. 2003.

[13] M. M. Skelly and H. J. Chizeck, "Real-time gait event detection for paraplegic FES walking," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 9, no. Jan., pp. 59–68, 2001.

[14] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed. New York: Academic, 1990.

[15] C. M. Bishop, *Neural Networks for Pattern Recognition*. Oxford, U.K.: Oxford Univ. Press, 1996.

[16] S. E. Thomas, P. J. Dykes, and R. Marks, "Plantar hyperkeratosis: A study of callosities and normal plantar skin," *J.f Investigat. Dermatol.*, vol. 85, no. 5, pp. 394–397, 1985.
[17] E. J. Van Langelaan, "A kinematical analysis of the tarsal joints. An X-ray photogrammetric study," *Acta Orthopaedica Scandinavica Supplement*, vol. 54, no. 204, pp. 1–269, 1983.
[18] A. Lundberg and O. K. Svensson, "The axes of rotation of the talocalcaneal and talonavicular joints," *Foot*, vol. 3, no. 2, pp. 65–70, 1993.
[19] G. Wu and P. R. Cavanagh, "ISB recommendations for standardization in the reporting of kinematic data," *J. Biomech.*, vol. 28, no. 10, pp. 1257–1261, 1995.
[20] A. H. Findlow, "Kinematics of the lower leg and foot during gait in subjects with pathological plantar hyperkeratosis," Ph.D. dissertation, Salford Univ., Centre Rehabil. Human Performance Res., Manchester, U.K., 2004.
[21] A. Cappello, A. Cappozzo, C. U. Della, and A. Leardini, "Bone position and orientation reconstruction using external markers," in *Three-Dimensional Analysis of Human Locomotion*, P. Allard, A. Cappozzo, A. Lundberg, and C. L. Vaughan, Eds. Chichester, U.K.: Wiley, 1997, pp. 147–172.
[22] A. Cappozzo, A. Cappello, C. U. Della, and F. Pensalfini, "Surface-marker cluster design criteria for 3-D bone movement reconstruction," *IEEE Trans. Biomed. Eng.*, vol. 44, no. 12, pp. 1165–1174, Dec. 1997.
[23] A. H. Findlow, C. J. Nester, and P. Bowker, "Repeatability of kinematic data for five segment foot/ankle model and validation of its use in the definition of key gait events," in *Proc. Int. Soc. Biomechanics XIXth Congr.*, 2003, pp. 105–105.
[24] ——, "Deriving gait temporal events from foot kinematic data," in *Proc. Int. Conf. Biomechanics of the Lower Limb in Health, Decease and Rehabilitation*, 2003, pp. 142–143.
[25] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York: Wiley, 2001.
[26] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning*. Berlin, Germany: Springer-Verlag, 2001.
[27] J. O. Ramsay and B. W. Silverman, *Functional Data Analysis*. Berlin, Germany: Springer-Verlag, 1997.
[28] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, 2nd ed. New York: Academic, 2003.
[29] A. M. Martínez and A. C. Kak, "PCA versus LDA," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 2, pp. 228–233, Feb 2001.
[30] G. J. McLachlan, *Discriminant Analysis and Statistical Pattern Recognition*. New York: Wiley Interscience, 1992.
[31] B. D. Ripley, *Pattern Recognition and Neural Networks*. Cambridge, U.K.: Cambridge Univ. Press, 1996.
[32] J. H. Friedman, "Regularized discriminant analysis," *J. Am. Statist. Assoc.*, vol. 84, no. 405, pp. 165–175, 1989.
[33] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proc. 14th Int. Joint Conf. on Artificial Intelligence (IJCAI)*, vol. 2, 1995, pp. 1137–1143.
[34] J. K. Martin and D. S. Hirschberg, "Small Sample Statistics for Classification Error Rates I: Error Rate Measurements," Dept. Inf. Comput. Sci., Univ. California, Irvine, CA, Tech. Rep. 96-21, 1996.
[35] B. Efron, "Estimating the error rate of a prediction rule: Improvement on cross-validation," *J. Am. Statist. Assoc.*, vol. 78, no. 382, pp. 316–331, 1983.
[36] B. Efron and R. Tibshirani, "Improvements on cross-validation: The $632+$ bootstrap method," *J. Am. Statist. Assoc.*, vol. 92, no. 438, pp. 548–560, 1997.
[37] S. Wehberg and M. Schumacher, "A comparison of nonparametric error rate estimation methods in classification problems," *Biometrical J.*, vol. 46, pp. 35–47, 2004.
[38] G. Schwarzer, W. Vach, and M. Schumacher, "On the misuses of artificial neural networks for prognostic and diagnostic classification in oncology," *Statist. Med.*, vol. 19, pp. 541–561, 2000.
[39] A. L. Blum and P. Langley, "Selection of relevant features and examples in machine learning," *Artif. Intell.*, vol. 97, no. 1–2, pp. 245–271, 1997.
[40] R. Kohavi and G. H. John, "Wrappers for feature selection," *Artif. Intell.*, vol. 97, pp. 273–324, 1997.
[41] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learning Res.*, vol. 3, pp. 1157–1182, 2003.
[42] A. Jain and D. Zongker, "Feature selection: Evaluation, application and small sample performance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 2, pp. 153–158, Feb. 1997.
[43] P. Pudil, F. J. Ferri, J. Novovičová, and J. Kittler, "Floating search methods for feature selection with nonmonotonic criterion functions," in *Proc. IEEE Int. Conf. Pattern Recognition*, 1994, pp. 279–283.
[44] P. Somol, P. Pudil, J. Novovičová, and P. Paclík, "Adaptive floating search methods in feature selection," *Pattern Recogn. Lett.*, vol. 20, pp. 1157–1163, 1999.
[45] T. Back, *Evolutionary Algorithms in Theory and Practice*. Oxford, U.K.: Oxford Univ. Press, 1996.
[46] Z. Michalewicz, *Genetic Algorithms + Structures = Evolution Programs*, 3rd ed. Berlin, Germany: Springer-Verlag, 1996.
[47] C. Liu and H. Wechsler, "Evolutionary pursuit and its application to face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 6, pp. 570–582, Jun. 2000.
[48] M. L. Raymer, W. F. Punch, E. D. Goodman, L. A. Kuhn, and A. K. Jain, "Dimensionality reduction using genetic algorithms," *IEEE Trans. Evol. Comput.*, vol. 4, no. 2, pp. 164–171, Jul. 2000.
[49] E. Cantú-Paz, "Feature subset selection by estimation of distribution algorithms," in *Proc. Genetic and Evolutionary Comp. Conf. (GECCO)*, 2002, pp. 303–310.
[50] M. T. Miller, A. K. Jerebko, J. D. Malley, and R. M. Summers, "Feature selection for computer-aided polyp detection using genetic algorithms," *Proc. SPIE (Medical Imaging)*, vol. 5031, pp. 102–110, 2003.
[51] M. Schröder, M. Bogdan, T. Hinterberger, and N. Birbaumer, "Automated EEG feature selection for brain computer interfaces," in *Proc. 1st IEEE EMBS Conf. Neural Engineering*, 2003, pp. 626–629.
[52] J. Y. Goulermas and P. Liatsis, "A collective-based adaptive symbiotic model for surface reconstruction in area-based stereo," *IEEE Trans. Evol. Comput.*, vol. 7, no. 5, pp. 1–21, Oct. 2003.
[53] E. Cantú-Paz, "Feature subset selection, class separability, and genetic algorithms," in *Proc. Genetic and Evolutionary Comp. Conf. (GECCO)*, 2004, pp. 957–970.
[54] Z. Q. Hong and J. Y. Yang, "Optimal discriminant plane for a small number of samples and design method of classifier on the plane," *Pattern Recogn.*, vol. 24, pp. 317–324, 1991.
[55] Y. Q. Cheng, Y. M. Zhuang, and J. Y. Yang, "Optimal fisher discriminant analysis using the rank decomposition," *Pattern Recogn.*, vol. 25, pp. 101–111, 1992.
[56] H. Yu and J. Yang, "A direct LDA algorithm for high-dimensional data—With application top face recognition," *Pattern Recogn.*, vol. 34, pp. 2067–2070, 2001.
[57] K. Liu, Y. Q. Cheng, and J. Y. Yang, "A generalized optimal set of discriminant vectors," *Pattern Recogn.*, vol. 25, no. 7, pp. 731–739, 1992.
[58] L. F. Chen, H. Y. M. Liao, M. T. Ko, J. C. Lin, and G. J. Yu, "A new LDA-based face recognition system which can solve the small sample size problem," *Pattern Recogn.*, vol. 33, pp. 1713–1726, 2000.
[59] W. Liu, Y. Wang, S. Z. Li, and T. Tan, "Null space-based kernel fisher discriminant analysis for face recognition," in *Proc. 6th Int. Conf. Automatic Face and Gesture Recognition*, 2004, pp. 369–374.
[60] D. L. Swets and J. J. Weng, "Using discriminant eigenfeatures for image retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 8, pp. 831–836, Aug. 1996.
[61] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, Jul. 1997.
[62] J. Yang and J. Y. Yang, "Why can LDA be performed in PCA transformed space?," *Pattern Recogn.*, vol. 36, no. 2, pp. 563–566, 2003.



John Yannis Goulermas (M'98) was born in Greece in 1970. He received the B.Sc. degree (Hons, Class I) in computation from the University of Manchester (UMIST), Manchester, U.K., in 1994. In 1996 and 2000, he received the M.Sc. degree by research and the Ph.D. degree from the Control Systems Centre, Department of Electrical Engineering and Electronics (EE&E) at UMIST working in the area of Machine Vision. He has worked for two years in industry in the area of financial/pricing modeling and optimization, and for three years in the Centre for Virtual Environments and the Centre for Rehabilitation and Human Performance Research of the University of Salford, Manchester, as a Senior Research Fellow in the area of biomechanics and intelligent gait analysis. He is currently a lecturer in the EE&E department at the University of Liverpool, Liverpool, U.K. His main research interests include pattern recognition, data analysis, artificial intelligence, machine vision and optimization.

Andrew H. Findlow received the Diploma in podiatry medicine from the Northern School of Podiatry, Manchester, U.K., the B.Sc. (Hons) degree in Health Care and Welfare from the University of Salford, Manchester, U.K., in 1991 and 1996, respectively. He is currently working towards the Ph.D. degree at the University of Salford.

He is currently a Lecturer in the Department of Podiatry, University of Salford. His main research interests are biomechanics and human gait analysis, particularly of the lower limb and foot, aetiology of neuropathic foot ulcers in diabetes, and real-world monitoring of kinematics and kinetics.

Dr. Findlow is a member of the Society of Chiropodist and Podiatrist, and IHS Diabetes and Obesity Research Network.

Christopher J. Nester received the B.Sc. (Hons) degree in podiatry in 1995 and the Ph.D. degree in 1999, from the University of Salford, Manchester, U.K.

He is a Senior Research Fellow in the Centre for Rehabilitation and Human Performance Research at the University of Salford. His research to date includes work on characterising the motion at the rear, mid, and forefoot joints, the effect of foot orthoses on motion and forces at the ankle, knee, and hip, as well as the development of an improved experimental kinematic model of the foot. He is also involved in a range of other foot, footwear, orthoses, and gait research and the strategic growth of the Research Centre at Salford University.

David Howard received the B.Sc. degree in mechanical and production engineering from Brunel University, Runnymede, U.K., in 1980 and the Ph.D. from Bath University, Bath, U.K., in 1987.

He is currently based in the School of Computing, Science and Engineering at the University of Salford, Manchester, U.K. Until 1997, his research interests were in the areas of robotics, theoretical mechanics, and design. However, since then, he has moved over to biomedical engineering, and now works with Salford's Centre for Rehabilitation and Human Performance Research (http://www.healthcare.salford.ac.uk/crhpr/). His current research interests include the use of gait simulation in the design of assistive devices, the control of FES, biomechanical modeling of the foot and lower limb, real-world monitoring of patients using assistive devices, and the myokinemetric signal as an alternative to EMG.

Peter Bowker received the B.Sc. degree in mechanical engineering and the Ph.D. in engineering metallurgy from the University of Salford, Manchester, U.K. Following periods in academic posts first in the University of Newcastle-upon-Tyne and then in the University of Aberdeen, he returned to Salford where he led the establishment of a new undergraduate degree in prosthetics and orthotics before becoming Dean of the Faculty of Health and Social Care and then Pro-Vice-Chancellor. His research interests lie principally in the biomechanics of the normal and pathological lower limb and its orthotic management, and he is currently Director of the Centre for Rehabilitation and Human Performance Research.