A PCA based bio-motion generator to synthesize new patterns of human running

José María Baydal-Bertomeu [1] , **Juan Vicente Juan V Durá-Gil** [Corresp., 1] , **Ana Piérola-Orcero** [1] , **Eduardo Parrilla Bernabé** [1] , **Alfredo Ballester** [1] , **Sandra Alemany-Munt** [1]

[1] Instituto de Biomecánica de Valencia, Universitat Politècnica de València, Valencia, Spain

Corresponding Author: Juan Vicente Juan V Durá-Gil
Email address: juan.dura@ibv.upv.es

Synthesizing human movement is useful for most applications where the use of avatars is required. These movements should be as realistic as possible and thus must take into account anthropometric characteristics (weight, height, etc.), gender, and the performance of the activity being developed. The aim of this study is to develop a new methodology based on the combination of principal component analysis and partial least squares regression model that can generate realistic motion from a set of data (gender, anthropometry and performance). 18 volunteer runners have participated in the study. The joint angles of the main body joints were recorded in an experimental study using 3D motion tracking technology. A five-step methodology has been employed to develop a model capable of generating a realistic running motion. The described model has been validated for running motion, showing a highly realistic motion which fits properly with the real movements measured. The described methodology could be applied to synthesize any type of motion: walking, going up and down stairs, etc. As future work, we want to integrate the motion in realistic body shapes, generated with a similar methodology and from the same simple original dat.

1   # A PCA based bio-motion generator to synthesize new

2   # patterns of human running

3   José María Baydal-Bertomeu, Juan Vicente Durá-Gil*, Ana Piérola-Orcero,

4   Eduardo Parrilla-Bernabé, Alfredo Ballester-Fernández, Sandra Alemany-Munt.

6   Instituto de Biomecánica de Valencia, Universidad Politécnica de Valencia, Valencia, Spain.

8   Corresponding author:

9   Juan Vicente Durá-Gil

10   juan.dura@ibv.upv.es

Abstract

Synthesizing human movement is useful for most applications where the use of avatars is required. These movements should be as realistic as possible and thus must take into account anthropometric characteristics (weight, height, etc.), gender, and the performance of the activity being developed. The aim of this study is to develop a new methodology based on the combination of principal component analysis and partial least squares regression model that can generate realistic motion from a set of data (gender, anthropometry and performance). 18 volunteer runners have participated in the study. The joint angles of the main body joints were recorded in an experimental study using 3D motion tracking technology. A five-step methodology has been employed to develop a model capable of generating a realistic running motion. The described model has been validated for running motion, showing a highly realistic motion which fits properly with the real movements measured. The described methodology could be applied to synthesize any type of motion: walking, going up and down stairs, etc. As future work, we want to integrate the motion in realistic body shapes, generated with a similar methodology and from the same simple original data.

Introduction

It is well known that there is a large degree of information contained in the kinematics of a moving body which is influenced by parameters such as: gender, age, anthropometrical features, emotional state, personality traits, etc.(Troje 2008) A number of studies demonstrate the capability of the human visual system to detect, recognize and interpret the information encoded in the biological motion (Johansson 1973). There are also many attempts to analyse this information encrypted in human motion. Some researchers use discrete kinematics parameters such as ranges, speed, etc.(Dvorak et al. 1992) Others focus their studies on the sequence of movement along time, instead of recording simple parameters. In these cases, they analyse the complete function of time f(t) (Feipel et al. 1999). A number of kinematical models are based on frequency domain manipulations (Davis, Bobick, and Richards 2000) and multiresolution filtering (Bruderlin and Williams 1995). Nevertheless, the most common objective of these studies is to model and to classify the movement pattern of the person being measured, rather than creating new motions from the extracted information.

In this regard, motion synthesis is currently attracting a great deal of attention within the computer graphics community as a means of animating three dimensional realistic characters and avatars; and in the robotic field to provide controlled real-time dynamic motion for the locomotion and other activities (Kajita et al. 2002). With the computational resources available today, large-scale models of the body [i.e. models that have many degrees of freedom and are actuated by many muscles] may be used to perform realistic simulations (Pandy 2001). Nevertheless, it is necessary to perform lab experiments to track the positions and orientations of body segments executing the task aimed to be synthesized. Recording motion data directly from real actors and mapping them to computer characters is a common technique used to generate high quality motion (Li, Wang,

52 and Shum 2002). However, this technique requires a high effort in experimental work. Besides,

53 new measures are needed to include changes in the pattern of movement, such as age, weight,

54 gender or speed. In this sense, it would be useful to create a methodology based on biomechanical

55 models constructed from a database of motions, instead of a single actor, able to generate realistic

56 motions of individuals with different anthropometrical characteristics, with sufficient accuracy and

57 without the need to perform laboratory measurements.

58   Several authors have addressed the motion modelling and synthesis for biped walking,

59 jumping, pedalling (Troje 2002), or even stair-ascending (Vasilescu 2002). Classically the

60 mathematical approach of the synthesis of movement has been the dynamic optimization of

61 biomechanical body structures (Pandy 2001). These models provide detailed information of the

62 functioning of some structures, such as the description of muscle function during normal gait.

63 However, this approach becomes an unworkable problem when a greater number of body

64 structures are included in the model. A new approach based on Principal Components Analysis

65 (PCA) can facilitate the understanding of the information contained in the kinematics of a moving

66 human body and avoids the inclusion of the dynamics in the model. PCA can extract depth

67 information contained in the mathematical function and its derivatives not normally available

68 through traditional statistical methods (Ullah and Finch 2013). In this way PCA can be used on

69 different levels. For instance, Troje (2002) used PCA in two steps for the purpose of analysing and

70 synthesizing human gait patterns. In the first one, they extracted the main components from the

71 entire database, in order to eliminate redundancy and to reduce the dimensionality. In the second

72 step, PCA was applied particularly for each walker in order to retain the encoded information of

73 each walker-specific variation. In our research, we will also use a model (based on PCA) to extract

74 the most relevant information from the pattern of running. This information will be used to develop

75   a bio-motion generator which will solve the opposite problem of synthesizing new realistic

76   movements.

77   In addition, existing literature focused on synthesizing motion does not correlate the generated

78   movement to age, gender, performance parameters such as velocity or anthropometrical features.

79   In this sense our research has three goals. The first one is to generate a database of running

80   movements of a full human model. The second is to extract the signature of each motion, by means

81   of PCA technique and to correlate the distinctive styles of each runner with their anthropometrical

82   characteristics, age, gender, and performance parameters such as the velocity of the action. The

83   third is to develop a bio-motion generator based on a statistical model capable of synthesizing new

84   realistic running motion from a set of desired data: age, gender, height, body mass index (BMI)

85   and velocity.

86   Nowadays there exists a line of research developed in the field of anthropometry for the

87   purpose of obtaining a model of human body shape from a database of processed raw scans (Vinué

88   et al. 2014). The methodology followed in that line of research provides sufficient resolution to

89   synthesize accurate realistic representations of body shapes from a set of simple anthropometrical

90   parameters. Ballester et al (2014) describe a method based on the harmonization of body scan data

91   followed by a Shape Analysis procedure using Principal Component Analysis. The combination

92   of these techniques allows the generation of human 3D shape models from anthropometric

93   measurement data (age, height, weight, BMI, waist girth, hip girth, bust/chest girth, etc.). Our

94   hypothesis is that the use of a similarly based methodology to generate human motion instead of

95   human body shapes is possible, valid and reliable. The novelty in our approach is the generation

96   of running data from a set of easily measurable anthropometric parameters and a desired value of

97   running speed.

98

Materials and Methods

Data Gathering

101   An experimental phase was carried out with the aim of gathering a database of the running

102   movements that we needed to include in the biomechanical model. The data consisted of the 3D

103   joint angles of the main body joints.

104   The articulated human body model used in our study comprised 21 segments and 20 joints

105   distributed throughout the body. Lower limb: hip, knee, ankle and metatarsophalangeal joint; upper

106   limb: shoulder, elbow and wrist; trunk: pelvis, L5, L3, T12, T8 and Neck.

107   The positions of the joints of the human body model were defined in a recursive mode with

108   respect to the origin joint (father) of the related segment. This methodology is based on the

109   BioVision Hierarchical data (BVH) format (Meredith and Maddock 2001).

110   Each subject was kinematically characterised by means of 64 variables, defined as follows:

111   1- Vertical (Z) position of the root segment, in our case the hips.

112   2- Tri-dimensional orientation (X, Y, Z) of the total amount of segments with respect the

113   root. 21 x 3 = 63 variables.

114

Study sample

116   Eighteen people composed the study sample, with the same number of male and female. Their

117   ages ranged from 21 to 44 years (average age, 31 years). They were selected according to some

118   specific parameters, trying to cover a wide range in the anthropometric characteristics of height

119   and body mass index (Table 1). Ethical approval was obtained from the ethics committee of the

120   Universitat Politècnica València. All participants gave written informed consent.

121   **Table 1. Description of the anthropometrical parameters**

Gender	Parameter	Mean	Std.	Max.	Min.
Male	Age	32.0	5.1	40	26
	Weight	86.712	16.813	118.000	59.000
	Height	1.779	.086	1.910	1.630
	BMI	27.263	4.118	33.412	21.914
Female	Age	30.5	8.472	44	21
	Weight	74.596	17.535	108.000	50.000
	Height	1.675	.087	1.780	1.550
	BMI	26.362	4.747	34.473	19.257

122

123

124

Measurement and Protocol

126   The measurements were performed using commercial equipment based on 17 inertial sensors:

127   MOVEN studio. The commercial system has been validated by previous studies (Zhang et al.

128   2013; Thies et al. 2007). A sampling frequency of 120Hz was used. This system showed a very

129   high sensitivity to electromagnetic fields. For this reason, the measurements of running trials were

130   done outdoors in a location free of electromagnetic pollution.

Experimental procedure

132   For the purpose of controlling the pace of running, a 20-metre-long corridor, delimited with

133   cones every 5 meters was set up. Thus, we obtained four areas, one area of acceleration, two of

134   constant speed and a final deceleration area. Running at constant speed presents a periodic

135   timing in which the period depends on the velocity (Novacheck 1998). Nevertheless, acceleration

136 and deceleration periods are out of phase and the duration of cycles is variable. Therefore, the

137 running cycles used to create our model were selected within the area of constant speed.

138   In the case of running, the pattern of the movement changes with velocity (e.g. stride length,

139 maximum joint angles, etc.). For this reason, each runner completed six running trials at different

140 speeds. Initially, subjects started running at normal speed. In the second measurement subjects ran

141 at their maximum speed. The third and fourth trial were performed at a pace between normal and

142 maximum speed. The fifth trial was performed at the minimum speed at which each runner was

143 able to run, on the edge between walking and running. Running defined in this case as when there

144 is no phase of bipodal support (Biewener et al. 2004). The last trial was performed at an average

145 speed between the lower and the normal speed.

146   This procedure allowed us to obtain six observations representing the whole range of speeds

147 that each subject could execute.

148 **Mathematical procedure**

149   The methodology used in our study comprised 5 steps:

150   1-**Reduction of intra-personal variability**: joint angles are periodic by nature. We took the

151 most representative single stride for the purpose of reducing variability and dismiss the phases

152 with no consistent speed, such as acceleration and deceleration steps. The selected stride was

153 picked in the middle of the running sequence, guaranteeing constant speed.

154   2-**Time normalization**: time normalization is usually employed for the temporal alignment

155 of cyclic data obtained from different trials with different duration. In our approach, the number

156 of samples for each stride depended on the velocity of the running. At this point we normalised

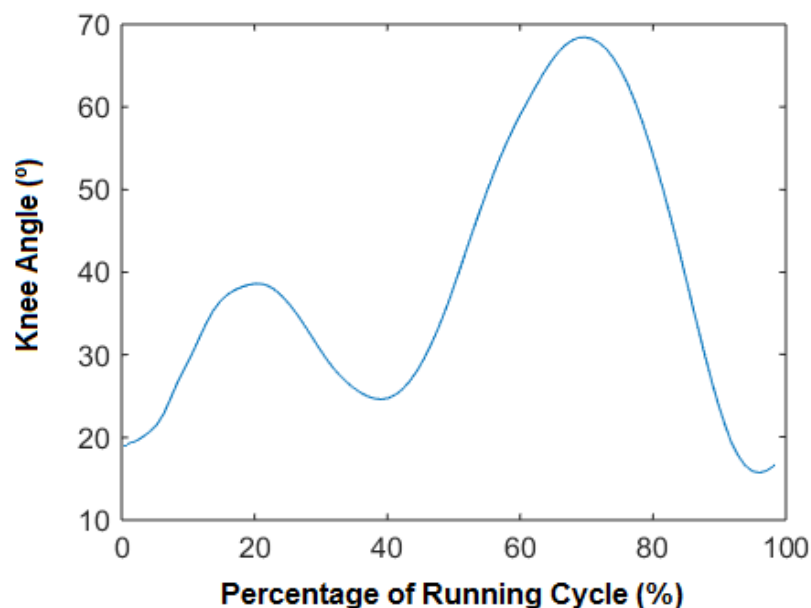
157 the variable "time" applying an interpolation technique based on cubic splines through the

158 measured values of the whole sequence of samples. This technique enables the normalization of

159   all the measurements to the percentage of the running cycle. The cubic spline was applied to

160   normalize at 50 equispaced time intervals per each variable (see the example in Fig. 1). The

161   application of the cubic spline to the 64 kinematical variables makes a total amount of 50 x 64

162   (3200) data per each subject

163

164   **Fig. 1: Knee angle vs % of the cycle**

165   3-**Data cleaning**: at this point a detailed checking and cleaning of inaccuracies of the

166   kinematical data was conducted. These type of inaccuracies were caused mostly by the

167   measurement system. The prevention of errors at this point is preferable to their later correction

168   once the model has been created. All the measurements have been manually analysed thoroughly

169   by an experienced examiner. The identified inaccuracies were treated as follows:

170   a) **Angular offsets:** this common error usually appears during the standing posture and can

171   affect the later joint angles registers during the trial (running) (Mills et al. 2007). Offsets

172   have been corrected manually eliminating (adding or subtracting) the difference between

173   the initial angle observed and the expected angle of the body segment at this position.

174   b) **Positioning error:** due to the fact that our measurement system, based on inertial sensors,

175   uses the earth's magnetic field to determine the reference position of each subject, it is

176   quite common to find subjects with slight differences in their initial reference positions.

177   In this case, we have proceeded by correcting the reference system aligning it with the

178   direction of running forward. Thus, we can guarantee that all measurements are equally

179   oriented.

180   c) **Non-physiological angles:** some errors in the registration of joint angles were detected in

181   the database. These inaccuracies came from errors of the inertial sensors. In this case, it

182   was not possible to correct the error effectively, thus we proceed by eliminating these

183   observations from the database and repeating the measurement.

184   4-**Dimensionality reduction**: the database of all measurements was combined in a single

185   matrix W. The initial number of observations is 108 (18 subjects x 6 velocities = 108). But three

186   measurements fail, therefore W has 105 rows (observations) and 3200 columns (50 equispaced

187   time intervals x 64 kinematical variables). Motion data of each observation is enclosed in the rows

188   of the matrix $W = (w_i)$, $i = 1,..., 105$.

189   Before the creation of the bio-motion generator, by means of a regression model, it was needed

190   to reduce the dimensionality of the motion data. Computing a PCA on the running data (contained

191   in matrix W), resulted in a decomposition of the data matrix W into an average running vector w_0

192   and 3200 weighed components, arranged in a 3200 x3200 matrix V:

$$W = W_0 + \alpha \cdot V \qquad (1)$$

194   where W_0 is a 105x3200 matrix with all rows equal to w_0 and $\alpha = (\alpha_i)$ with $i = 1,...,105$ is a

195   105x3200 matrix of PCA scores. Each observation w_i was thus expressed by a linear combination

196   of scores α_i and PCA components (columns of matrix V). Components represented factors related

197   to gender, anthropometrical traits and running speed. And scores represented individual

198   characteristics of each runner and performance of the running trial related to the previous factors.

199   PCA components are arranged in descendent order of explained variance of the original matrix

200   data. Thus the first columns in matrix V retained most of the information in the data sample and it

201   was possible to select a reduced number of components c.

202 $$w_{ic} = w_0 + \alpha_{ic} \cdot V_c \qquad\qquad (2)$$

203   Above w_0 denoted the average of all the running samples. The matrix V_c contained the first

204   components. α_{ic} represents the c scores of each observation of the database in the reduced

205   dimension space formed by the selected components. As score values change from negative to

206   positive values, the movement of the runner changes from men to women; high BMI to low BMI;

207   high speed to low speed, etc.

208   The decision of how many components to retain was a critical issue in the exploratory factor

209   analysis. To perform this decision we used the methodology of Parallel Analysis (PA) (Hayton,

210   Allen, and Scarpello 2004). PA is a Monte-Carlo based simulation method that compares the

211   observed eigenvalues (components) with those obtained from randomized normal variables. A

212   component is retained if its explained variance or information is higher than the information

213   provided by the eigenvectors derived from the random data.

214   5-**Regression model**: one of the objectives of our work was to generate a statistical model

215   capable of synthetizing new realistic running motion from a set of desired data: age, gender, height,

216   weight and velocity, also called 1D data. Accordingly, to devise the bio-motion generator, we

217   needed to establish the correlation between the 1D data and PCA scores, which provide the

218   signature of each motion.

219   The correlation was obtained as a regression model, combining a partial least squares (PLS)

220   regression model as a first step and a linear regression model (LRM) as a second step.

221   PLS methodology is explained in Wold (2006) and Geladi (1986). . This type of regression

222   model is suitable for the kind of data involved in the bio-motion generator since the input data of

223   the model is strongly correlated (anthropometrical information).

224   The PLS regression model takes the 1D data – age, height, weight and velocity– as input

225   information and produces a set of PCA scores as output. The LRM model was applied to these

226   output PCA scores to reflect the influence of gender in the PCA scores.

227   In the first step, we estimated a PLS model considering anthropometrical data and velocity of

228   the movement as independent variables and the PCA scores as dependent variables. The general

229   formula of a PLS model is:

230
$$Y - Y_0 = B \cdot (X - X_0) + E \quad (3)$$

231   where Y is the matrix of dependent variables, X is the matrix of independent variables, X_0 and

232   Y_0 are the matrices of mean data, B is the coefficient matrix of the PLS model and E is the

233   prediction error matrix.

234   Having 4 different 1D measurements for each subject $x_i = [age, height, weight, velocity]$, we

235   built a 105x4 matrix X formed by the concatenation of the vectors x_i. Notice that since matrix $= \boldsymbol{\alpha}$

236   , it is constituted by PCA scores and the mean matrix Y_0 is the zero matrix. Finally, we estimated

237   the coefficient matrix B from sample data X and $\boldsymbol{\alpha}$ with the PLS algorithm and substituted in the

238   equation above, obtaining the following model:

239
$$\boldsymbol{\alpha} = B \cdot (X - X_0) + E \qquad (4)$$

240

241   Where X_0 is the matrix of mean anthropometrical data and velocity, and E is the prediction

242   error matrix. PLS decomposes the independent and dependent variables in component spaces in

243   order to obtain their correlation. The number of significant PLS components in the model was

244   selected in a leave-one-out procedure and according to the explained variance (R^2) criteria.

245   Secondly, the influence of gender was modelled with a LRM of the prediction error matrix $\boldsymbol{E}$

246   with coefficients $\boldsymbol{a} = (a_1,...,a_c)$ and $\boldsymbol{b} = (b_1,...,b_c)$, where c is the number of retained PCA

247   components:

$$\hat{E} = \boldsymbol{a} + \boldsymbol{b} \cdot gender \qquad (5)$$

249   where $gender = 0$ for men and $gender = 1$ for women.

250   This way, the motion information related to gender which is part of the PLS error matrix E ,

251   and uncorrelated with the prediction derived from the PLS regression, was modelled. Notice that

252   a_j and b_j were considered zero whenever their F-value was below a desired level of statistical

253   significance of 95%.

254   Once we have obtained B, a and b, whenever we want to synthesize a running motion from

255   new anthropometrical data and velocity, we obtain the corresponding scores $\hat{\alpha}$ of the new realistic

256   running motion by the following formula:

$$\hat{\alpha} = [\boldsymbol{a} - B \cdot X_0] + \begin{bmatrix} B \\ \boldsymbol{b} \end{bmatrix} \cdot [X \quad gender] \qquad (6)$$

258   where X_0 is the matrix of mean anthropometrical data and velocity.

259   **Validation methodology of the bio-motion generator**

260   To validate the five-step methodology described to develop the bio-motion generator we

261   propose a comparison between each recorded observation and the prediction of running motion

262   generated by the model by means of the 'leave-one-out' procedure. The recorded observation is

263   considered the true angle curve of the running motion. The predicted motion is estimated using the

264   'leave-one-out' validation technique. That is, not including that observation in the bio-motion

265   generator. We wish to determine if both curves are reproducible and sufficiently similar to consider

266   that they represent the same motion. For this purpose, we use the Intraclass Correlation Coefficient

267   (ICC) as a measurement of the reliability, and the Standard Error of Measurement (SEM) as a

268   direct measurement of the global error between true and predicted angles. Theoretically the ICC

269   is defined as the ratio between the true variance and the predicted variance. The ICC varies

270   between 0 and 1 and can be interpreted as the proportion of variance due to the methodology (true

271   versus predicted data) in the total variance. An ICC greater than 0.8 is generally considered to be

272   good (Fleiss 2011). The ICC is determined between the measured or true curve (T_c) and the

273   estimated curve (E_c) provided by the bio-motion generator. The ICC is determined from the

274   variance of both curves (T_c) and (E_c) following the next equation:

275

$$ICC = \frac{\sigma_{E_c}^2}{(\sigma_{T_c}^2 + \sigma_{E_c}^2)} \; ; (7)$$

276   On the other hand, SEM represents the existent difference between observed (T_c) and

277   estimated curves (E_c) determined with the bio-motion generator, and provide an indication of the

278   real magnitude of the error.

279

$$SEM = \sigma_c \cdot \sqrt{(1 - ICC)}; \quad (8)$$

280   Where σ_c is the combined standard deviation of the true scores (T_c) and observed scores (E_c).

281   And S_E is the combined standard deviation of the true scores and observed scores.

282   We have obtained the SEM for each pair of true and predicted angles for the three spatial

283   directions in all the joints that form the human model. For that reason, we have represented the

284   SEM by its descriptive statistics (mean, std., 5-percentile and 95-percentile).

Results

Parallel analysis

287   The results of the PA (Fig. 2) have been obtained with the explained variance of the main

288   components extracted from the original data and the same obtained from randomized data. The

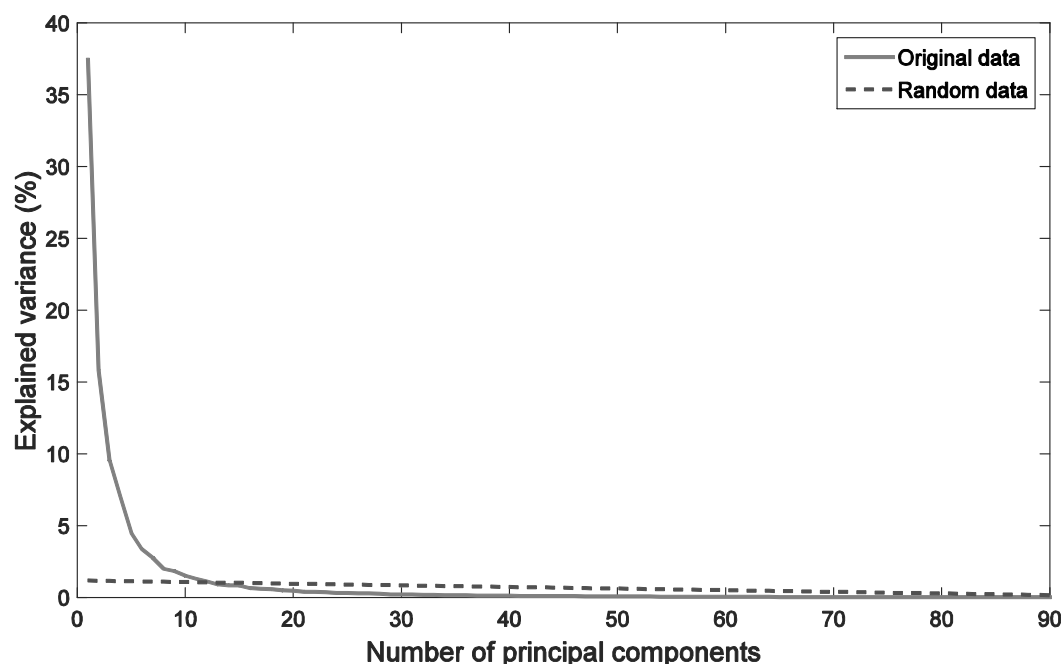
289   intersection point of both curves indicates the optimal number of components to extract from the

290   PCA. The original number of dimensions was 72 (3 related to the pelvis translation + 69 related to

291   the body segments orientation). The results of the PA recommend retaining the first 12

292   eigenvalues, which explain the 88.16% of the total variance. Thus the PCA allowed a percentage

293   of data reduction of 83%, from 72 variables to 12 weighed components.

294

295   **Fig. 2: Relation between the explained variance and the number of principal components**

296   **extracted from the database and from randomized data**

Regression model

298   As it has been explained in the methodology, the regression model consists of two parts, the

299   first including the anthropometrical data (PLS) and the second the gender (LRM). The dependent

300   variables of the PLS are the scores of the first 12 principal components (PC) of the kinematical

301   running motion. Therefore, they are uncorrelated and the optimal number of PLS components are

302   separately determined for each PC score (PC 1... PC 12) according to its adjusted R^2 plot (Fig. 3).

303   PLS components are retained until their R^2 curve exhibits a decrease or a non-significant increase.

304   Thus, for instance, two PLS components are retained for PC 1, whereas no components are

305   considered for PC 7 and PC 9. Notice that for those PC with 0 retained components, the PLS model

306   provides their mean value as output. This way, the motion information associated to those PC

307   which is provided by the PLS model is the average motion.

308

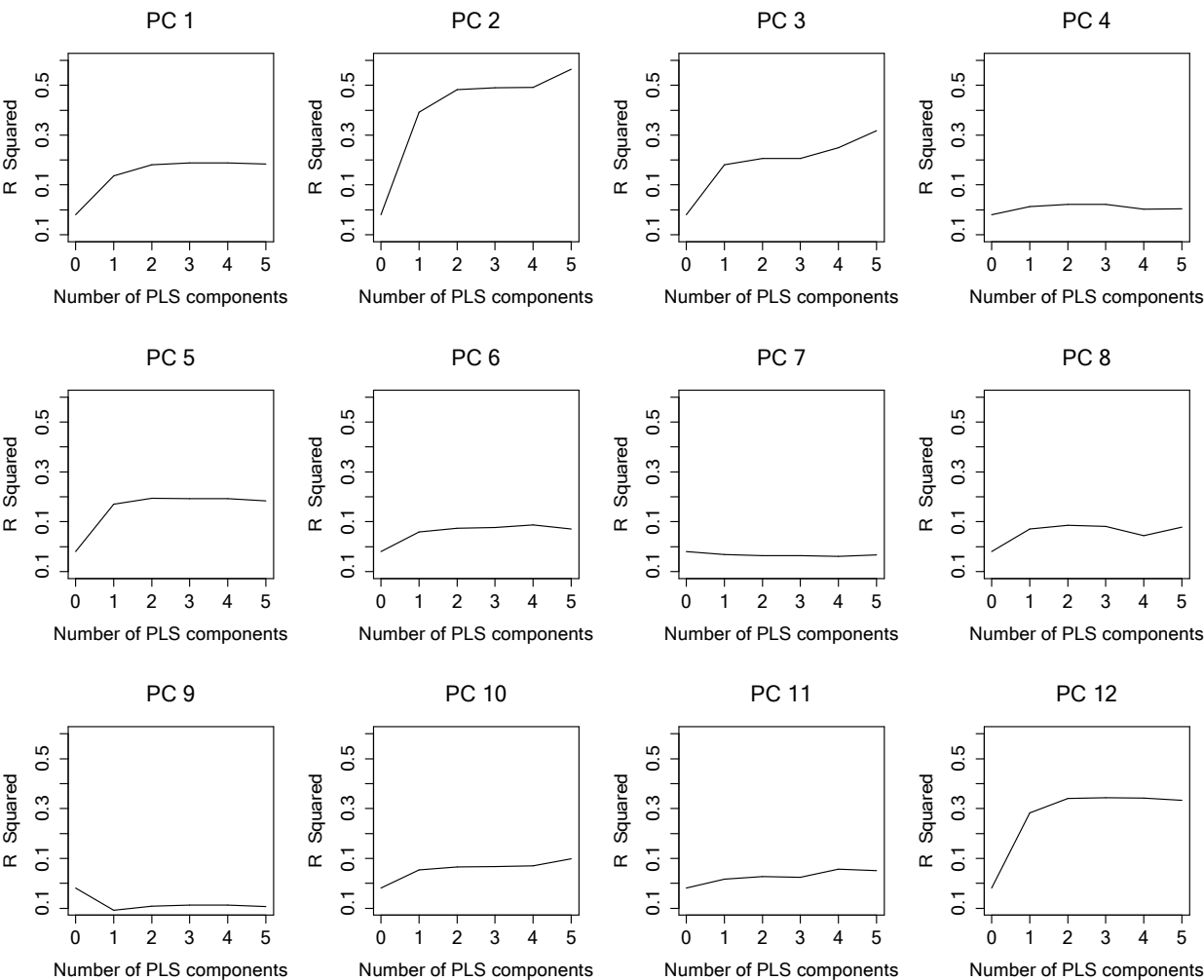



Fig. 3: Leave-one-out R^2 estimation plots for the PLS model.

310   With regard to the LRM, which analyses the influence of gender in the kinematics of running,

311   the PCA scores which are significantly affected by gender are PC3, PC7, PC8, PC9 and PC11

312   (Table 2). The prediction obtained in the first step of the model is improved by the influence of

313   gender on these PC. PC 7 and PC 9 are only affected by gender, since their number of retained

314   PLS components was 0.

315   **Table 2. ANOVA table for the linear models**

		Df	Sum Sq	Mean Sq	F-value	Pr (>F)	

PC1	Gender	1	214117	214117	1.092	0.298	
	Residuals	103	20188135	196001			
PC2	Gender	1	140256	140256	2.687	0.104	
	Residuals	103	5377216	52206			
PC3	Gender	1	259962	259962	5.592	0.0199	*
	Residuals	103	4788029	46486			
PC4	Gender	1	6242	6242	0.14	0.709	
	Residuals	103	4591853	44581			
PC5	Gender	1	79045	79045	3.434	0.0667	
	Residuals	103	2370796	23017			
PC6	Gender	1	2470	2470	0.125	0.724	
	Residuals	103	2028601	19695			
PC7	Gender	1	81973	81973	4.361	0.0392	*
	Residuals	103	1935909	18795			
PC8	Gender	1	88710	88710	7.979	0.00568	**
	Residuals	103	1145086	11117			
PC9	Gender	1	62331	62331	4.962	0.0281	*
	Residuals	103	1293890	12562			
PC10	Gender	1	3246	3246	0.355	0.552	
	Residuals	103	940560	9132			
PC11	Gender	1	111297	111297	13.31	0.000417	***
	Residuals	103	861467	8364			
PC12	Gender	1	1373	1373	0.29	0.591	
	Residuals	103	487860	4737			

316   *Note*: Signification codes 0.001 '**'; 0.01 '**'; 0.05 '*'

317

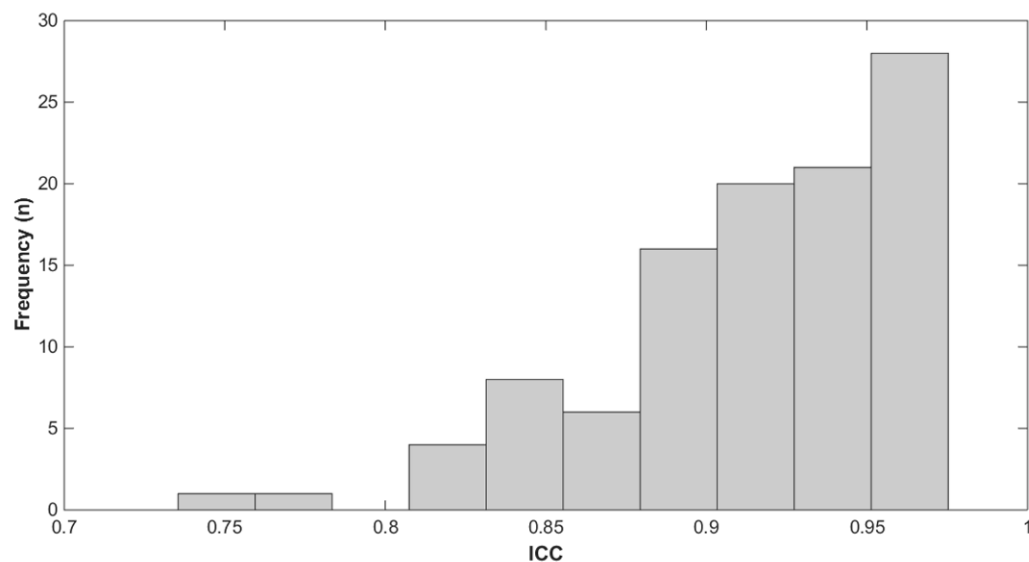
318   **Validation of the bio-motion generator**

319   The results of the reliability study, computed from the 90 observations and the same calculated

320   by means of the leave-one-out technique, showed that the mean and standard deviation of ICC,

321   was 0.91(0.04) with a 5 percentile of 0.829 and 95 percentile of 0.971. Only one subject exhibit

322   an ICC lower than 0.8 in 2 observations (Fig. 4).

323

324   **Fig. 4: Frequency histogram of the ICC.**

325   The SEM between the real and the predicted angles determined with the leave-one-out model

326   showed a mean (std.) of 4.16º (6.80º), with a 5 percentile of 0.41º and 95 percentile of 14.23º (Fig.

327   5).

328

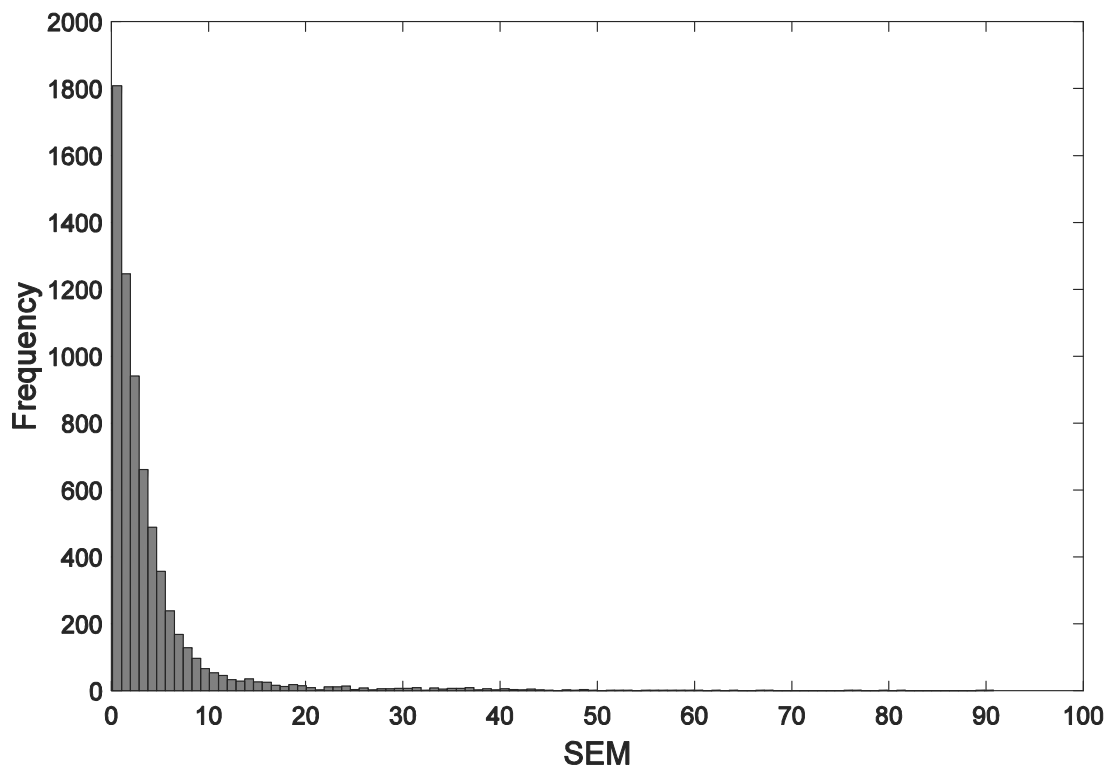



Fig. 5: Frequency histogram of the SEM.

Discussion

In this paper we have demonstrated that the five-step methodology on which the bio-motion generator is based provides running motion models closely resembling the measurements obtained with real subjects. However, while the SEM study shows that the vast majority of errors detected between actual and predicted data of the bio-motion generator are less than 10°, there are a percentage of observations (8%) in which greater errors are observed. This can be explained because the model has been obtained from a small number of subjects --only 18-- and therefore the bio-motion generator is not able to adjust the running specific characteristics of each corridor.

338   Future work in this line of research must be done to increase the database of real subjects measured

339   and incorporate greater variability in anthropometric and performance characteristics.

340   The bio-motion generator is based on a methodology which comprises 5 steps. In the fourth

341   step we tackle a dimensionality reduction based on PCA. This step is similar to that performed by

342   Troje (2002). However, there are some differences, as he obtained four main components that

343   explain more than 98% of the variance and we have obtained 12 components explaining 88.16%

344   of variance. The greater variability of our study is explained partly by the greater variability of the

345   running against walking and on the other hand by the greater speed range in our study in relation

346   to Troje, in which each subject could select a single comfortable walking speed. On the other hand,

347   Troje made a second reduction of the dimensionality based on the simplicity of temporal behaviour

348   of the walking components which could be modelled with pure sine functions with a scaled

349   fundamental frequency. This approach was not valid for the motion of running, due to the fact that

350   the 12 PCs of running cannot be modelled with a proportional frequency. This suggests that

351   running is a more complex motion than walking in the sense that there does not exist a proportion

352   between the frequency of oscillation of the different body segments.

353   The fifth step of the methodology consists of a two-step linear regression which correlates a

354   given list of 1D measurements with the PCA scores of movement. A linear regression technique

355   has been used before to approximate motion models from a reduced marker set and estimate the

356   remaining markers (Liu et al. 2005) or to model the motion-style and the spatio-temporal

357   movement (Torresani, Hackney, and Bregler 2006). However, it has not been used before to

358   synthesize new human motion directly from a set of anthropometrical and performance data. In

359   this sense, it can be considered a real breakthrough in the field of synthesis of human motion.

Conclusions

360

361   The major contribution of this paper is a novel statiscal methodology for modelling human

362   movements. The method described in this article has been developed and validated for running

363   motion, but this same methodology could be used to synthesize other types of motion: walking,

364   going up and down stairs, or even for sport movements such as: jumping, pedalling, golf swing

365   and putting, etc.

366   Our work aims to provide a realistic motion to body shapes that can be developed with the

367   methodology described in the work of Ballester et al. (2014). Those body shapes could include an

368   adjusted skeleton formed by a hierarchical set of interconnected joints and can be used to move

369   the body shape with the required or desired motion provided by our methodology (Fig. 6). The

370   integration of both methods will allow generating realistic avatars supplied with realistic motion

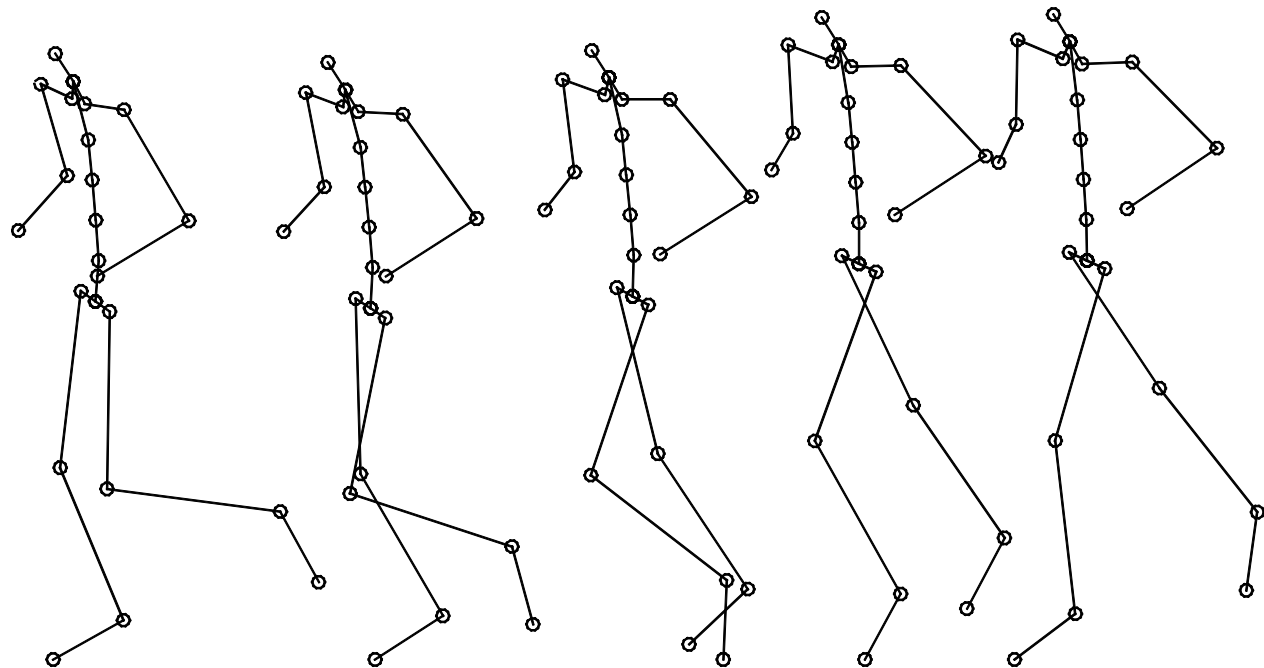
371   from a set of adjustable and simple anthropometrical and performance data and without the need

372   of the realization of new measurements.

373   A limitation of this study is the sample size. Further work needs to be done in order to validate

374   with a broader sample of people. Notwithstanding this limitations, the findings suggest that the

375   model is valid.

376

377   **Fig. 6: Reconstructed virtual biomechanical model (skeleton+motion).**

378

References

380 Ballester, Alfredo, Eduardo Parrilla, Jordi Uriel, Ana Pierola, Sandra Alemany, Beatriz Nacher,
381   Jorge Gonzalez, and Juan Carlos Gonzalez. 2014. '3D-Based Resources Fostering the
382   Analysis, Use, and Exploitation of Available Body Anthropometric Data'. In *5th*
383   *International Conference on 3D Body Scanning Technologies*.

384 Biewener, Andrew A., Claire T. Farley, Thomas J. Roberts, and Marco Temaner. 2004. 'Muscle
385   Mechanical Advantage of Human Walking and Running: Implications for Energy Cost'.
386   *Journal of Applied Physiology* 97 (6): 2266–74. doi:10.1152/japplphysiol.00003.2004.

387 Bruderlin, Armin, and Lance Williams. 1995. 'Motion Signal Processing'. In *Proceedings of the*
388   *22nd Annual Conference on Computer Graphics and Interactive Techniques*, 97–104.
389   ACM.

390 Davis, J., A. Bobick, and W. Richards. 2000. 'Categorical Representation and Recognition of
391   Oscillatory Motion Patterns'. In *IEEE Conference on Computer Vision and Pattern*
392   *Recognition, 2000. Proceedings*, 1:628–35 vol.1. doi:10.1109/CVPR.2000.855878.

393 Dvorak, J., J. A. Antinnes, M. Panjabi, D. Loustalot, and M. Bonomo. 1992. 'Age and Gender
394   Related Normal Motion of the Cervical Spine'. *Spine* 17 (10 Suppl): S393-398.

395 Feipel, V., B. Rondelet, J. P. LePallec, O. DeWitte, and M. Rooze. 1999. 'The Use of Disharmonic
396   Motion Curves in Problems of the Cervical Spine'. *International Orthopaedics* 23 (4):
397   205–9. doi:10.1007/s002640050351.

398 Fleiss, Joseph L. 2011. *Design and Analysis of Clinical Experiments*. Vol. 73. John Wiley & Sons.

399 Geladi, Paul, and Bruce R. Kowalski. 1986. 'Partial Least-Squares Regression: A Tutorial'.
400   *Analytica Chimica Acta* 185 (January): 1–17. doi:10.1016/0003-2670(86)80028-9.

401 Hayton, James C., David G. Allen, and Vida Scarpello. 2004. 'Factor Retention Decisions in
402   Exploratory Factor Analysis: A Tutorial on Parallel Analysis'. *Organizational Research*
403   *Methods* 7 (2): 191–205. doi:10.1177/1094428104263675.

404 Johansson, Gunnar. 1973. 'Visual Perception of Biological Motion and a Model for Its Analysis'.
405   *Perception & Psychophysics* 14 (2): 201–11. doi:10.3758/BF03212378.

406 Kajita, Sliuuji, Funiio Kanehiro, Kenji Kaneko, Kiyoslii Fujiwara, Kazuliito Yokoi, and Hirohisa
407   Hirukawa. 2002. 'A Realtime Pattern Generator for Biped Walking'. In *Robotics and*
408   *Automation, 2002. Proceedings. ICRA'02. IEEE International Conference on*, 1:31–37.
409   IEEE.

410 Li, Yan, Tianshu Wang, and Heung-Yeung Shum. 2002. 'Motion Texture: A Two-Level Statistical
411   Model for Character Motion Synthesis'. In *ACM Transactions on Graphics (ToG)*, 21:465–
412   472. ACM.

413 Liu, Guodong, Jingdan Zhang, Wei Wang, and Leonard McMillan. 2005. 'A System for Analyzing
414   and Indexing Human-Motion Databases'. In *Proceedings of the 2005 ACM SIGMOD*
415   *International Conference on Management of Data*, 924–926. ACM.

416 Meredith, Maddock, and S. Maddock. 2001. 'Motion Capture File Formats Explained'.
417   *Department of Computer Science, University of Sheffield* 211: 241–244.

418 Mills, Peter M., Steven Morrison, David G. Lloyd, and Rod S. Barrett. 2007. 'Repeatability of 3D
419   Gait Kinematics Obtained from an Electromagnetic Tracking System during Treadmill
420   Locomotion'. *Journal   of   Biomechanics*   40   (7):   1504–11.
421   doi:10.1016/j.jbiomech.2006.06.017.

422 Novacheck, Tom F. 1998. 'The Biomechanics of Running'. *Gait & Posture* 7 (1): 77–95.
423   doi:10.1016/S0966-6362(97)00038-6.

424 Pandy, Marcus G. 2001. 'Computer Modeling and Simulation of Human Movement'. *Annual*
425   *Review of Biomedical Engineering* 3 (1): 245–273.

426 Thies, S. B., P. Tresadern, L. Kenney, D. Howard, J. Y. Goulermas, C. Smith, and J. Rigby. 2007.
427   'Comparison of Linear Accelerations from Three Measurement Systems during "reach &
428   Grasp"'. *Medical   Engineering   and   Physics*   29   (9):   967–72.
429   doi:10.1016/j.medengphy.2006.10.012.

430 Torresani, Lorenzo, Peggy Hackney, and Christoph Bregler. 2006. 'Learning Motion Style
431   Synthesis from Perceptual Observations'. In *Advances in Neural Information Processing*
432   *Systems*,   1393–1400.
433   http://machinelearning.wustl.edu/mlpapers/paper_files/NIPS2006_458.pdf.

434   Troje, Nikolaus. 2002. 'Decomposing Biological Motion: A Framework for Analysis and
435   Synthesis of Human Gait Patterns'. *Journal of Vision* 2 (5): 2. doi:10.1167/2.5.2.

436   Troje, Nikolaus. 2008. 'Retrieving Information from Human Movement Patterns'. In
437   *Understanding Events*, edited by Thomas F. Shipley and Jeffrey M. Zacks, 308–35. Oxford
438   University Press.

439   Ullah, Shahid, and Caroline F. Finch. 2013. 'Applications of Functional Data Analysis: A
440   Systematic Review'. *BMC Medical Research Methodology* 13: 43. doi:10.1186/1471-
441   2288-13-43.

442   Vasilescu, M. Alex O. 2002. 'Human Motion Signatures: Analysis, Synthesis, Recognition'. In
443   *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, 3:456–460.
444   IEEE.

445   Vinué, G., T. León, S. Alemany, and G. Ayala. 2014. 'Looking for Representative Fit Models for
446   Apparel   Sizing'.   *Decision   Support   Systems*   57   (January):   22–33.
447   doi:10.1016/j.dss.2013.07.007.

448   Wold, H. 2006. 'Partial Least Squares'. In *Encyclopedia of Statistical Sciences*, edited by Samuel
449   Kotz, Campbell B. Read, N. Balakrishnan, Brani Vidakovic, and Norman L. Johnson.
450   Hoboken,   NJ,   USA: John   Wiley   &   Sons,   Inc.
451   http://doi.wiley.com/10.1002/0471667196.ess1914.pub2.

452   Zhang, Jun-Tian, Alison C. Novak, Brenda Brouwer, and Qingguo Li. 2013. 'Concurrent
453   Validation of Xsens MVN Measurement of Lower Limb Joint Angular Kinematics'.
454   *Physiological Measurement* 34 (8): N63. doi:10.1088/0967-3334/34/8/N63.

455

Description of the anthropometrical parameters

PeerJ Computer Science   Manuscript to be reviewed

Gender	Parameter	Mean	Std.	Max.	Min.
Male	Age	32.0	5.1	40	26
	Weight	86.712	16.813	118.000	59.000
	Height	1.779	.086	1.910	1.630
	BMI	27.263	4.118	33.412	21.914
Female	Age	30.5	8.472	44	21
	Weight	74.596	17.535	108.000	50.000
	Height	1.675	.087	1.780	1.550
	BMI	26.362	4.747	34.473	19.257

ANOVA table for the linear models.

		Df	Sum Sq	Mean Sq	F-value	Pr (>F)	
PC1	Gender	1	214117	214117	1.092	0.298	
	Residuals	103	20188135	196001			
PC2	Gender	1	140256	140256	2.687	0.104	
	Residuals	103	5377216	52206			
PC3	Gender	1	259962	259962	5.592	0.0199	*
	Residuals	103	4788029	46486			
PC4	Gender	1	6242	6242	0.14	0.709	
	Residuals	103	4591853	44581			
PC5	Gender	1	79045	79045	3.434	0.0667	
	Residuals	103	2370796	23017			
PC6	Gender	1	2470	2470	0.125	0.724	
	Residuals	103	2028601	19695			
PC7	Gender	1	81973	81973	4.361	0.0392	*
	Residuals	103	1935909	18795			
PC8	Gender	1	88710	88710	7.979	0.00568	**
	Residuals	103	1145086	11117			
PC9	Gender	1	62331	62331	4.962	0.0281	*
	Residuals	103	1293890	12562			
PC10	Gender	1	3246	3246	0.355	0.552	
	Residuals	103	940560	9132			
PC11	Gender	1	111297	111297	13.31	0.000417	***
	Residuals	103	861467	8364			
PC12	Gender	1	1373	1373	0.29	0.591	
	Residuals	103	487860	4737			

1   *Note*: Signification codes 0.001 '**'; 0.01 '**'; 0.05 '*'

Figure 1

Knee angle vs % of the cycle

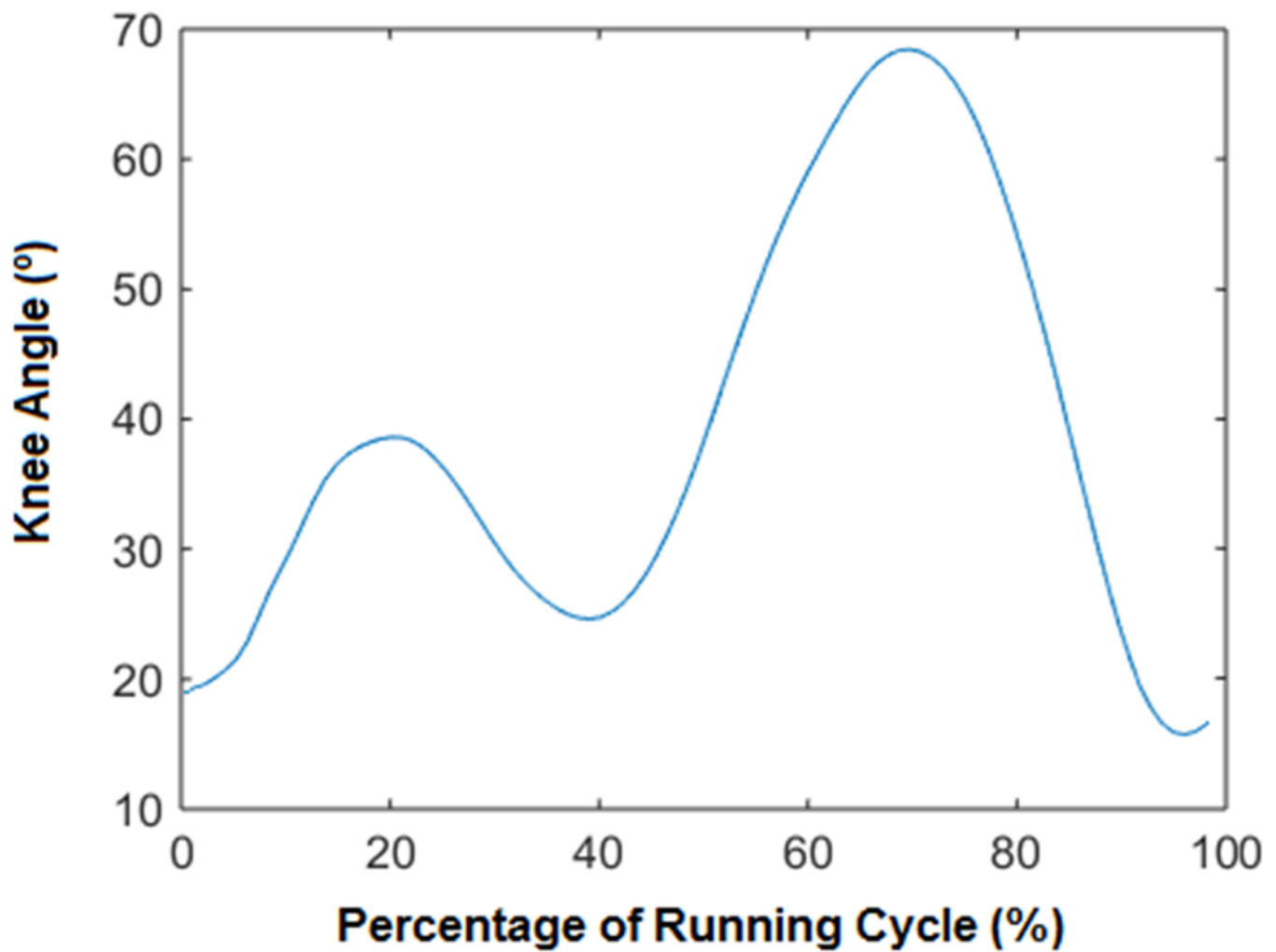



Figure 2

Relation between the explained variance and the number of principal components extracted from the database and from randomized dat.

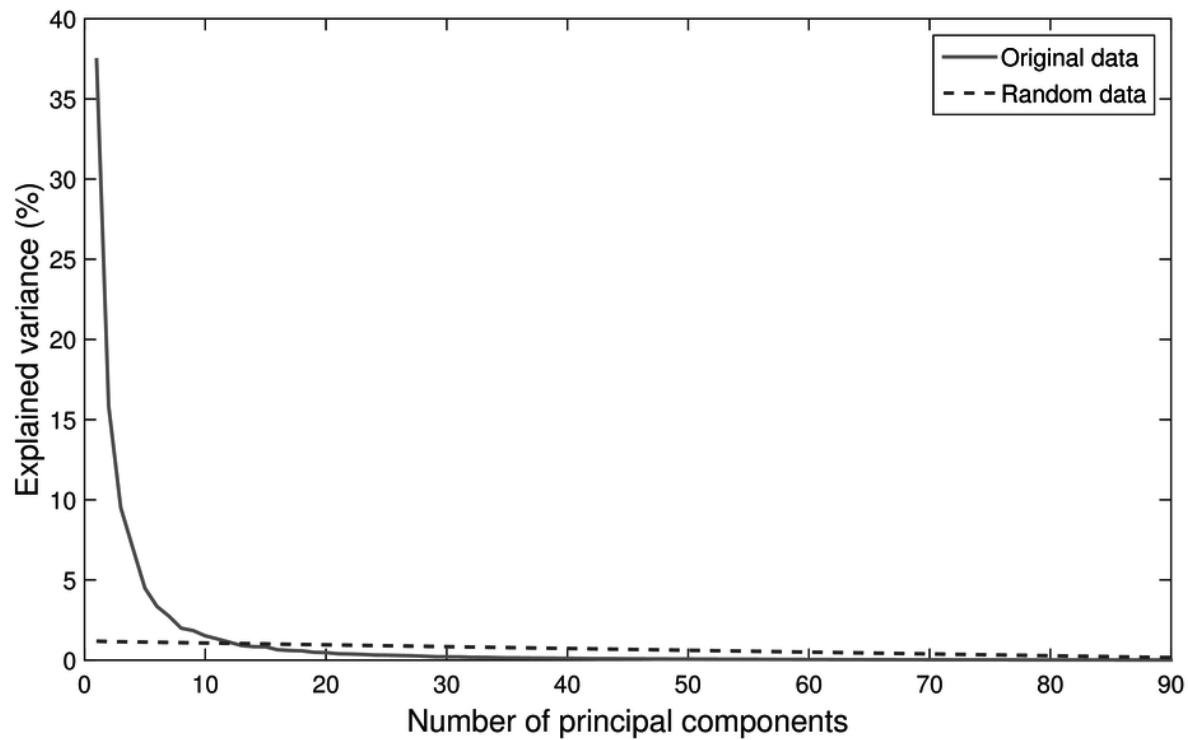



Figure 3

Leave-one-out R^2 estimation plots for the PLS mode.

Fig - 4 Frequency histogram of the SEM.

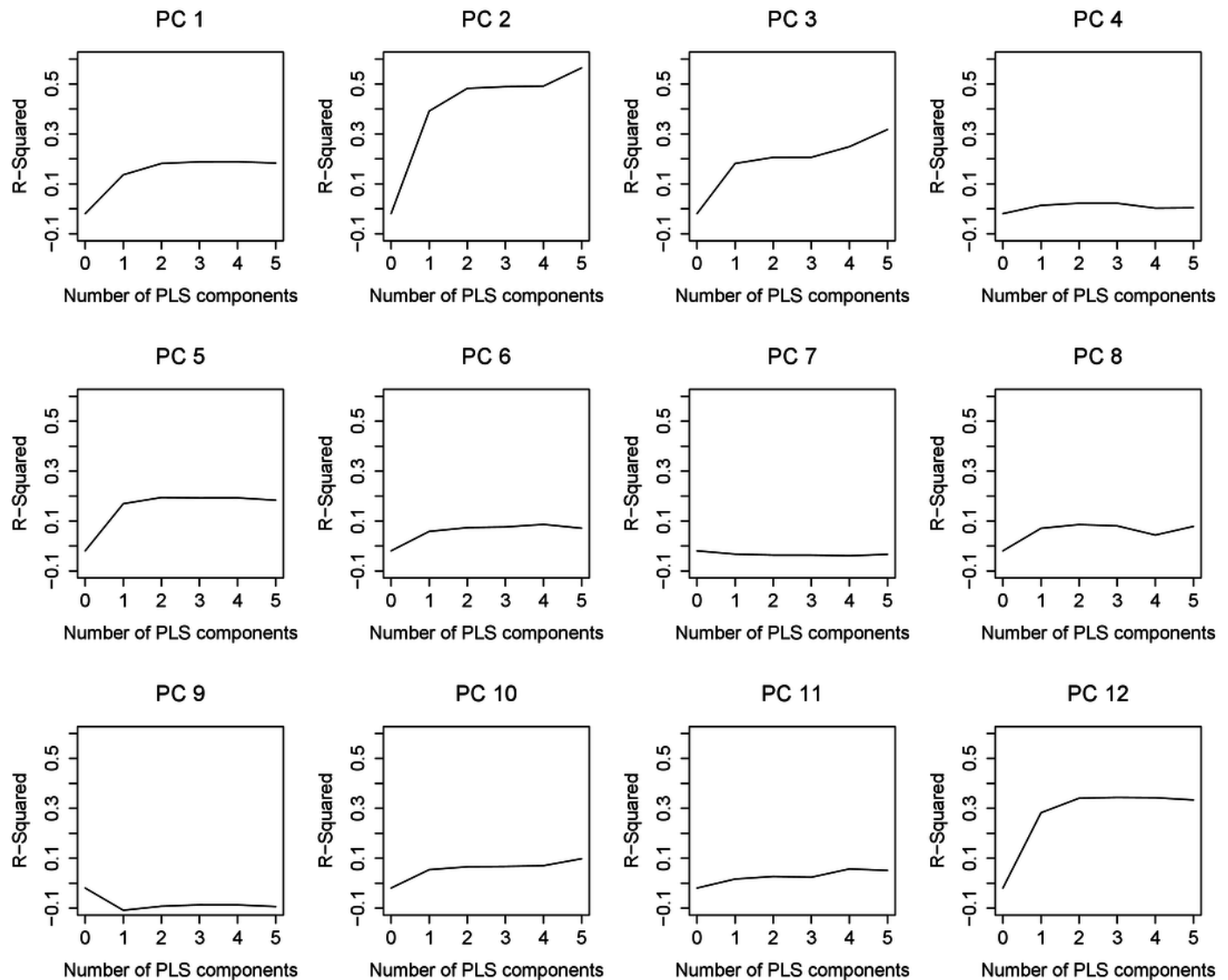



Figure 4

Frequency histogram of the IC.

Fig 5 - Reconstructed virtual biomechanical model (skeleton+motion).

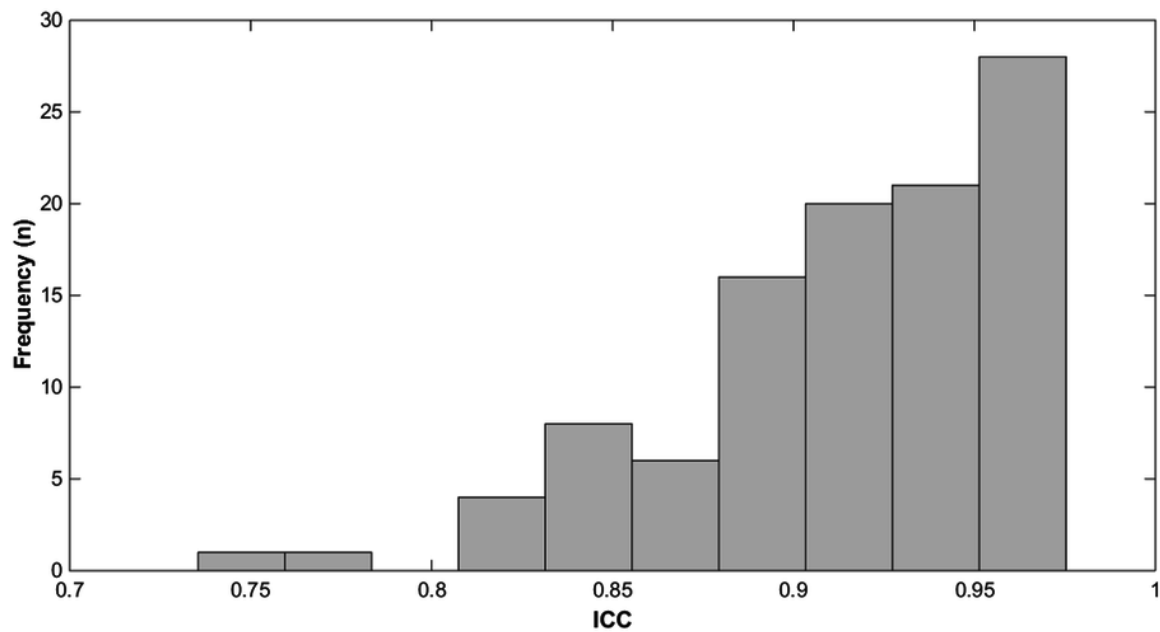



Figure 5

Frequency histogram of the SEM

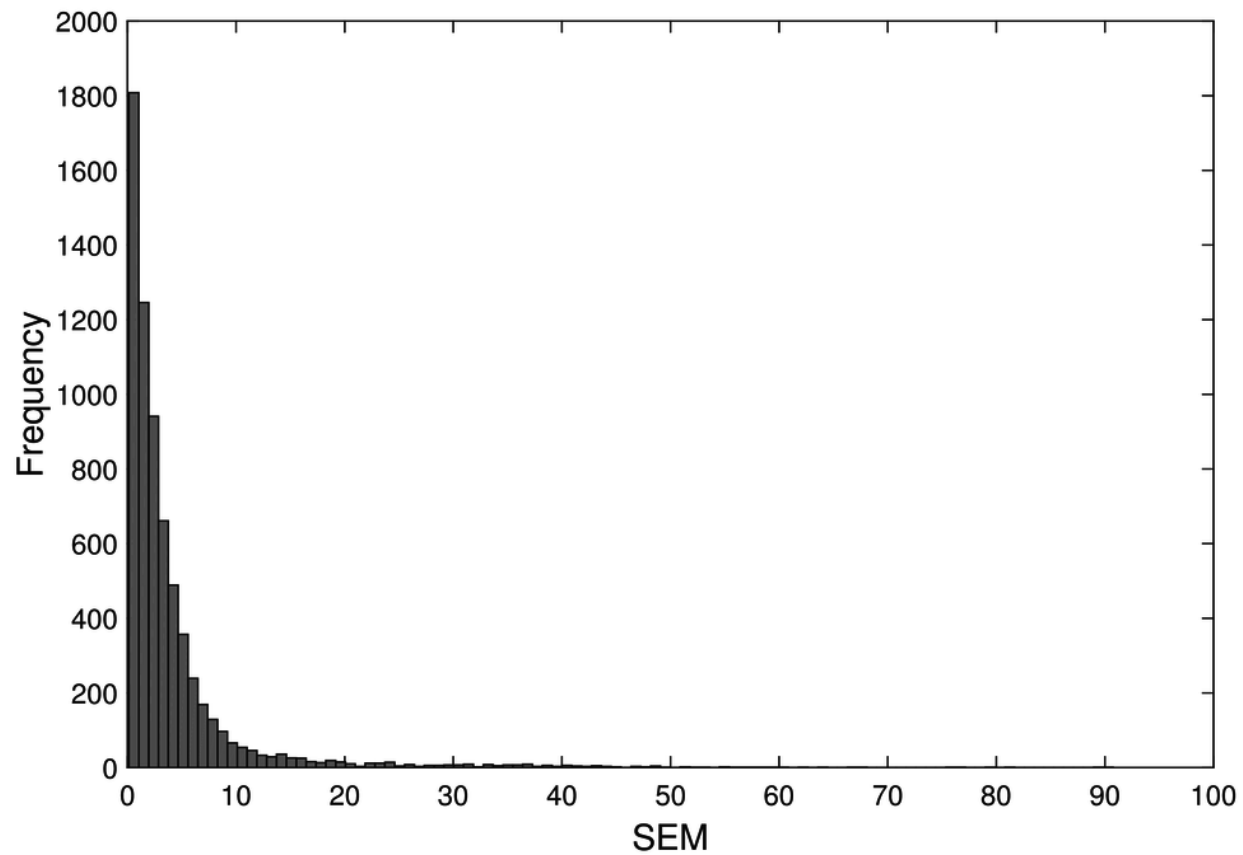



Figure 6

Reconstructed virtual biomechanical model (skeleton+motion)