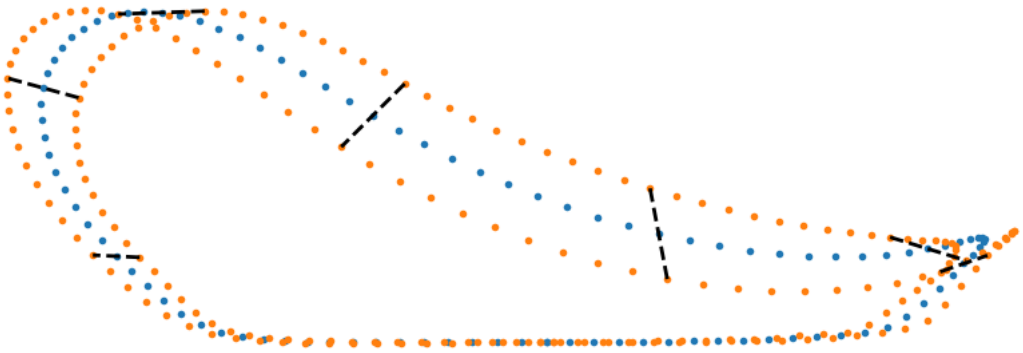


Data driven running technique identification

Master's thesis in Computer science and engineering

JOHAN LAMM, JESSICA PREIMAN

Department of Computer Science and Engineering
CHALMERS UNIVERSITY OF TECHNOLOGY
UNIVERSITY OF GOTHENBURG
Gothenburg, Sweden 2023

Master's thesis 2023

Data driven running technique identification

A comparison between inertial measurement unit and marker-based systems

JOHAN LAMM, JESSICA PREIMAN

UNIVERSITY OF
GOTHENBURG

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Computer Science and Engineering
Chalmers University of Technology
University of Gothenburg
Gothenburg, Sweden 2023

Data drive running technique identification
JOHAN LAMM
JESSICA PREIMAN

Supervisor: Moa Johansson, Department of Computer Science and Engineering
Examiner: Devdatt Dubhashi, Department of Computer Science and Engineering

Master's Thesis 2023
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Cover: The third principal component direction on running stride positional data.

Typeset in LaTeX
Gothenburg, Sweden 2023

Data drive running technique identification
JOHAN LAMM, JESSICA PREIMAN
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg

Abstract

We evaluate if acceleration and rotational IMU data and marker-based positional data can be used to quantify the running technique. We also investigate patterns between the runners' anatomy, fitness level and technique, and how common instructions impact their running technique.

Running technique data, consisting of rotational velocity and acceleration from a foot mounted IMU and position from a marker-based motion capture system, is collected from 47 participants, together with data on anatomy and fitness level. Participants perform a testing protocol containing treadmill running at different velocities while receiving different technique instructions. Data is processed to extract a representative stride cycle for each data source for every participant at every velocity and technique instruction.

We evaluate three methods to quantify the technique using dimensionality reduction and reconstruction: sequential feature selection using multivaritate linear regression, principal component analysis, and autoencoder. Best performance is obtained for principal component analysis on all data sources. Information loss is significantly larger on rotational and acceleration data from the IMU than for positional data from the marker-based system.

Limited patterns between the anatomy and fitness level of runners and their technique were observed, and the found patterns are generally on parameters that are not related to technique, such as ground contact time and contact to flight-time ratio.

Most technique instructions are shown to impact technique, but the effect diminishes as velocity increases. A larger impact is seen when runners are asked to increase back-kick height, knee lift, or frequency, and a smaller impact is seen when asked to land further back with the foot or push the hip forward.

Keywords: Running technique, Inertial Measurement Unit, Marker-based, Motion capture.

Acknowledgements

Thanks to:

- Our supervisor Moa Johansson for valuable input during the project.
- TRACKS learning environment for the possibility of conducting the project.
- Dan Kuylenstierna for supporting us in using the physiology lab.
- Alexander Nilsson for information regarding running technique.
- Our examiner Devdatt Dubhashi
- All test participants who volunteered for the data collection.

Jessica Preiman and Johan Lamm, Gothenburg, 2023-06-17

Contents

Abbreviations

IMU:	Inertial measurement unit
RE:	Running economy
QTM:	Qualisys track manager
AIM:	Automatic identification of markers
PC:	Principal component
FS:	Future selection
PCA:	Principal component analysis
AE:	Autoencoder
CDF:	Cumulative distribution function
KPI:	Key performance indicator
GCT:	Ground contact time
SVD:	Singular value decomposition

List of Figures

List of Tables

1

Introduction

The decreased cost, size, battery consumption, and improved performance of sensors have enabled new possibilities in several fields. One such field is sports biomechanics. Our thesis will explore the opportunity to use new sensor technologies for running technique analysis. We aim to develop a system of high availability that could be used in the future to reduce the risk or prevalence of injuries and improve sports performance.

Running technique analysis has traditionally been performed using expensive marker-based camera systems in lab conditions and video or real-time visual analysis in the field. These methods impose several challenges: Current motion capture systems are expensive and in high demand. Traditional video analysis also requires an experienced coach and lacks objectivity as it relies on the beholder's opinion.

We will evaluate the use of an Inertial Measurement Unit (IMU), a cheap and lightweight sensor consisting of accelerometers and gyroscopes that measure acceleration and angular rate to perform running technique identification. Evaluation will be done by comparing performance with a marker-based system. If successful, this will decrease cost, increase availability compared to marker-based systems, and increase objectivity over a human coach.

The study design also allows us to study the relationship between technique, anatomy, and fitness level, how instructions impact the technique, and how the running technique is best quantified.

1.1 Thesis scope and research questions

The main scope of the thesis is determined by the research questions (RQ) stated below. In addition to the RQs, the thesis has an applied component of how and what data should be collected.

RQ1: Can IMU sensor data and marker-based motion capture data from running be quantified into key performance indicators (KPIs) describing the technique?

We aim to obtain a low-dimension representation of the stride data whilst preserving interpretability. We want to find a limited number of KPIs that, when combined, describe the stride data with as little information loss as possible while also having a meaningful interpretation from a coaching standpoint.

RQ2: What patterns can be identified between technique and runner attributes?

Data will be analyzed to observe how the level and anatomy of the runners affect the stride at different velocities. What, if any, correlations and patterns can be found? Do faster runners have different techniques compared to slower? Do taller runners differ compared to shorter ones? Do body mass index influence technique?

RQ3: How is the technique affected by common instructions?

The testing protocol will include several standard instructions aimed at improving running technique, giving us objective data on the impact of instructions on the stride. This will allow us to study the actual impact of instructions on the runner's technique.

2

Theory

The theory chapter provides an outline of the algorithms and methods used and background on running technique, camera-based motion capture, and inertial measurement units.

2.1 Running technique

There exists a large amount of research done on running economy (RE), biomechanics in running, and optimal running technique. Therefore, some of them are presented in this chapter to give an overview of findings in this field. In addition to this, we conducted an interview with a professional running coach to support the literature and provide input on what is done in the field.

2.1.1 Litterature survey

This literature survey contains a broad variety of articles in the area of running technique, RE, motion capture, and general biomechanics when running.

I. Moore [1] examines whether there is an economical running technique or any difference in oxygen consumption depending on the biomechanics while running. This review concluded that several biomechanical factors affect RE, such as a shorter stride length, larger stride angles, and less leg extension at toe-off. In another similar article, written by Folland JP et al. [2], they also studied the relationship between running technique and RE. After the participants performed their test protocol, the study concluded that some aspects of stride parameters and lower limb angles had an essential impact on RE. In addition, they also found correlations between trunk forward lean and RE.

Fukuchi RK et al. [3] wrote an article where the goal was to examine some gait-biomechanic variables and what effects different velocities had on them, with the perspective of both injuries and RE. The data collection they performed and then made public contains the data of 28 runners that were captured using a 3D motion-capture system in three different velocities: 2.5 m/s, 3.5 m/s, and 4.5 m/s. The 3D motion-capture system consisted of 12 cameras and each runner had 48 markers during the test. These were placed to identify both technical and anatomical variables. From the data collection, they found the most correlations with the angles and increased speed in the hip extension, hip flexion, and knee flexion. Figure 2.1 illustrates knee

extension and flexion. Other findings were that the increase in stride length was greater than stride frequency, which also had a noticeable increase. Contrary to the hypothesis, they found that the foot-strike pattern was not altered.

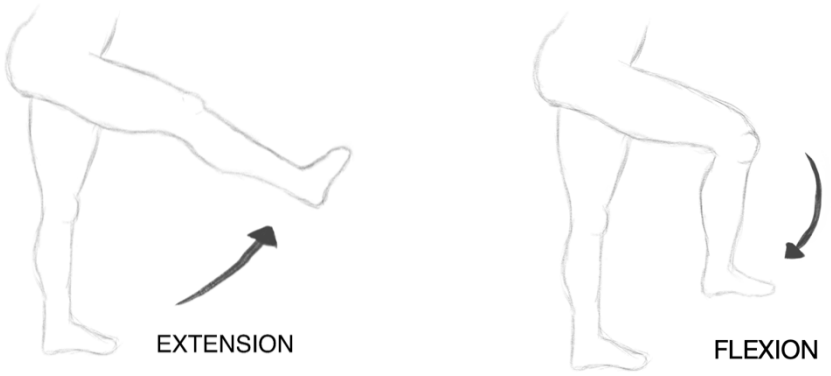


Figure 2.1: Illustration of knee extension and knee flexion.

Instead of specified velocities, article [4] by T. F. Novacheck, studied several biomechanic parameters during walking, running, and sprinting. The main biomechanic parameters were different angles, forces, timing, and the main differences between walking, running, and sprinting. The way they studied these where by taking the average of 15 stride cycles for each parameter. A step cycle was defined as the movement from the first contact with the ground until the next contact with the ground for the same foot. One of their findings was that when the speed increased, the initial contact in the step cycle changed from the hindfoot to the forefoot. Another parameter that changed when the speed increased was that both the pelvis and trunk tilted more forward, and the angle between the runner's ankle and tibia increased. In addition to this, they investigated the effects of pronation, defined as the angle between your foot and tibia. The conclusion was that larger pronation was related to several severe injuries.

2.1.2 Interview with a running coach

To complement the research, we also interviewed Alexander Nilsson, an experienced runner and running coach at the track and field club Mölndals AIK. In addition to his experience, Alexander has a bachelor's degree in sports coaching from the University of Gothenburg. In the interview, which is summarized in Appendix A, Nilsson said that runners' common mistakes are: overstriding, which is taking too long steps; low step frequency; flat steps, which come from low knee drive and back kick; leaning back with the upper body and a sitting position for the hips. Nilsson concludes that even good runners usually make some of these mistakes to a smaller extent than amateur runners.

2.2 Motion capture

This section provides background on how a marker-based camera system and an inertial measurement unit work.

2.2.1 Marker-based systems

Marker-based motion capture systems are a technology used to track the movement and position of objects or people using reflective markers and multiple cameras. By combining images from multiple cameras, the position of markers can be calculated in the three-dimensional space [5].

This process uses geometric relations between the projections in the camera to estimate the 3D position. An example of such estimation, under the pinhole camera model, is shown in Figure 2.2.

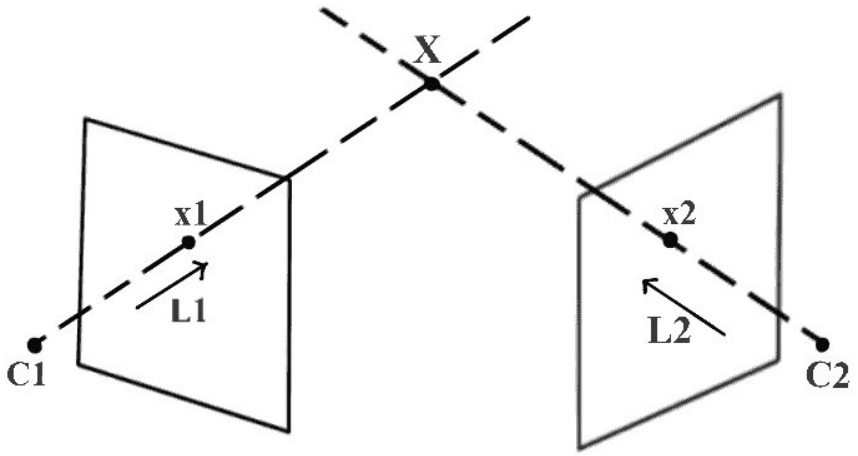


Figure 2.2: 3D point estimation using two 2D images.

By having known camera positions and orientations, the image planes shown as squares in Figure 2.2 and thus the image plane points $x1$, and $x2$ are also known. The lines $L1$ and $L2$ are then given from the camera centers $C1$ and $C2$, and the image plane points and the 3D point, X can be directly estimated as the intersection between $L1$ and $L2$, or the closest distance between lines if measurement errors are present [6]. This principle can be extended to multiple cameras, lens distortion, and projection errors, but the idea remains the same [6].

Marker-based systems are considered the gold standard in motion capture and are commonly used as a reference when evaluating other motion capture systems because of their superior accuracy [7]. However, several issues, such as complexity, time consumption, and cost, still need to be addressed. A typical marker-based motion capture system includes around ten high-performance cameras. Such a system has a

cost range from approximately 50.000$ and upwards [8]. They also generally require specially trained staff, placing the markers can take several minutes, and the subject is limited to the lab area [7].

2.2.2 Inertial measurement unit

An Inertial Measurement Unit (IMU) is an electronic device that measures and reports acceleration and angular rate using accelerometers and gyroscopes [5]. In addition to measuring acceleration and rotations, an IMU can be used for motion capturing by integrating the acceleration data from the accelerometers and rotational velocity from gyroscopes to obtain position and orientation [9]. A disadvantage with IMUs for position estimation is that they suffer from drift, as any measurement errors are accumulated over time when integrating [9].

IMUs have several advantages over traditional marker-based motion capture systems. An IMU is cheaper compared to marker-based camera systems, with a price in the 100$ range [10]. IMUs are lightweight and portable, making them ideal for use in applications where it is not practical or desirable to use a marker-based camera system, such as tests in the field[11]. IMUs can also provide real-time data, making them useful for monitoring movement in dynamic and fast-moving environments [11].

2.3 Dimensionality reduction and reconstruction

Dimensionality reduction is a data transformation from a high dimensional space into a lower dimensional space [12]. The goal of such transformation is to retain information while reducing complexity. Several methods for dimensionality reduction of different complexity exist. Three of the most common methods are feature selection, principal component analysis, and autoencoder [12]-[13]. In this thesis, we restrict ourselves to these methods. However, numerous other methods exist [12]. We will analyze the use of these methods for stride data compression to hopefully obtain useful KPIs describing the stride.

2.3.1 Feature selection

In feature selection (FS), the goal is to reduce dimensions by choosing a subset of existing features using manual or algorithm-guided selection. Several such methods exist based on filter or wrapper strategies [14]. A benefit of feature selection is that the compressed representation has intrinsic meaning as it directly corresponds to features in the original data. However, information loss can be more significant than with more advanced methods.

In our thesis, we use sequential feature selection combined with multivaritate linear regression. In sequential feature selection, features are added one at a time in a greedy fashion until the desired number of features is reached [15]. An outline of the algorithm is given below:

```
selected features = empty set
remaining features = set of all features
while number of selected features < number of desired features:
    best score = -inf
    for feature in remaining features:
        score = score(selected features + feature, all features)
        if score > best score
            best score = score
            best feature = feature
    add best feature to selected features
    remove best feature from remaining features
```

As we are using the algorithm for unsupervised dimensionality reduction, the scoring is done according to how well the subset can predict the full set of features. In this step, multivariate linear regression is used to predict the full set of features as linear combinations of the chosen subset by minimizing the squared errors. This also allows us to use the regression model for reconstruction to judge how large the information loss is.

2.3.2 Principal component analysis

Principal component analysis (PCA) is a transformation technique for high-dimensional data analysis. It transforms the data into a new coordinate basis based on the variance in the original dataset [16]. This transformation is done such that the first coordinate in the transformed basis, the principal component (PC), is chosen as the direction with the highest variance. Subsequent PCs are chosen as the directions that maximize variance and are orthogonal to the existing components [16]. A visualization of the principal components for a set of 2D data is shown in Figure 2.3.

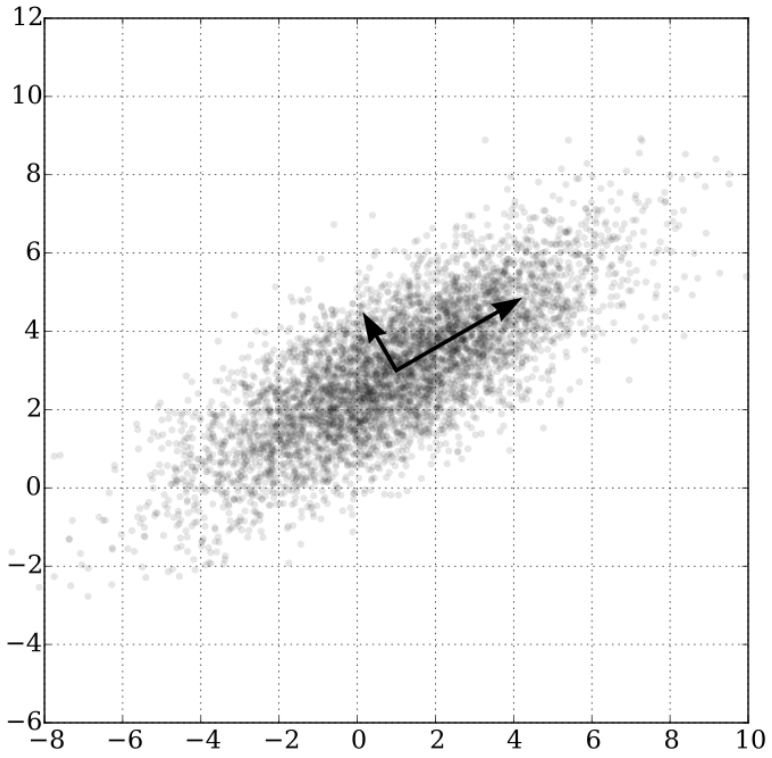


Figure 2.3: A visualization of 2D data and its principal component vectors, with the larger vector being the direction of maximum variance.[1]

Principal component analysis can be implemented in several different ways but is commonly implemented using singular value decomposition (SVD), as it allows the calculation of all principal components simultaneously [16]. A mathematical description of the procedure is described below.

SVD is a factorization of a matrix M into the form $U\Sigma V^T$ where U is an $m \times m$ unitary matrix, Σ is an $m \times n$ diagonal matrix with non-negative real numbers on the diagonal, and V^T is the transpose of an $n \times n$ unitary matrix V. By constructing a data matrix X with samples on rows and features as columns and centering data by subtracting the mean of each feature from its values, the principal components are given as the columns of the V matrix after performing SVD on X. The components are then ordered by the singular values in the Σ matrix.

PCA is used for dimensionality reduction by choosing the PCs with the largest singular values, as these will correspond to directions with the most variance, and discarding the PCs with the lowest variance by setting them to 0.

[1] Nicoguaro (https://commons.wikimedia.org/wiki/File:GaussianScatterPCA.svg), https://creativecommons.org/licenses/by/4.0/legalcode

2.3.3 Autoencoder

An autoencoder (AE) is a machine learning model capable of learning compression and decompression from data. The model is often said to consist of two parts, an encoder, and a decoder. The encoder transforms data from input to the latent space, and the decoder tries to reconstruct the input from the latent space representation[17]. A visualization of an autoencoder structure is shown in Figure 2.4.

Autoencoders are implemented as a feed-forward artificial neural network and can be trained by minimizing the reconstruction error between input and output. Using one or more hidden layers in the encoding and decoding part, together with non-linear activation functions, the autoencoder can learn non-linear mappings of data. Furthermore, by limiting the size of the bottleneck layer, the autoencoder can be used for dimensionality reduction by forcing the network to learn efficient compression and decompression of data [17].

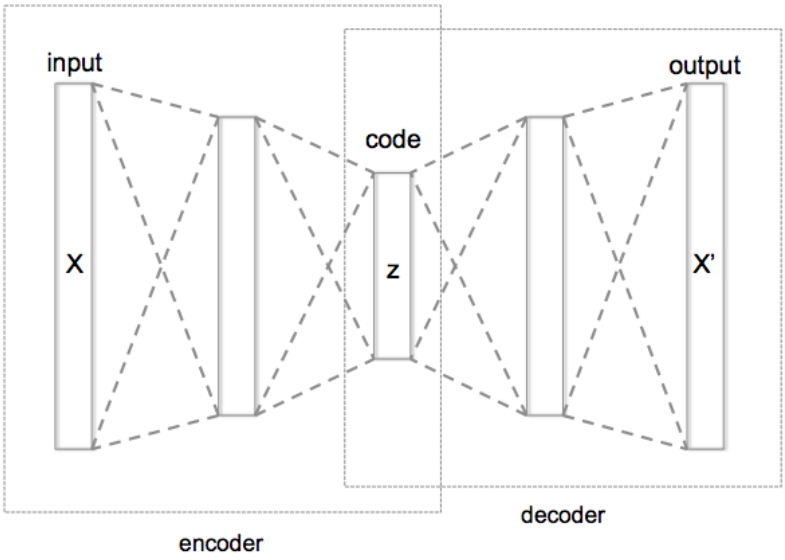


Figure 2.4: Schematic structure of an autoencoder with one hidden layer in the decoder and encoder.[2]

2.3.4 Comparison of chosen dimensionality reduction models

In Table 2.1, a brief comparison of the chosen dimensionality reduction models is given. Figure 2.5 shows a Visualization of the selected dimensionality reduction methods for compression and reconstruction in the 2D to 1D case for sample data. The figure highlights the strengths and weaknesses of all models. As can be seen, the feature selection retains one feature (x-axis), and the error is only present in feature

[2]Chervinskii https://commons.wikimedia.org/wiki/File:Autoencoderstructure.png), https://creativecommons.org/licenses/by-sa/4.0/legalcode

2 (y-axis). The principal component analysis finds the direction that minimizes the variance, and errors are thus lower than for feature selection. The autoencoder can learn non-linear patterns and has the lowest errors of all models, but could also easily overfit and find patterns that do not exist.

Table 2.1: Comparison of chosen dimensionality reduction models.

Model:	FS	PCA	AE
Complexity	Lowest	Middle	Highest
Variability of output	Deterministic	Deterministic	Stochastic
Explainability	Highest	Middle	Lowest

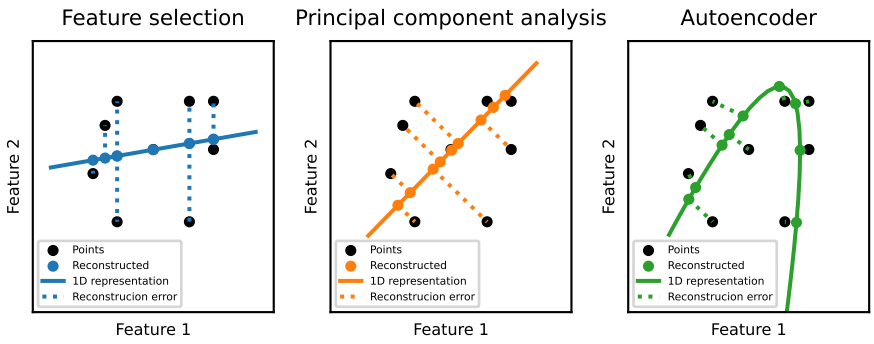


Figure 2.5: Visualization of selected dimensionality reduction methods for the 2D-1D case showing the compression and reconstruction of the three models for sample data.

2.4 Interpolation

In engineering applications, data is often obtained from sampling, where the dependent variable is observed at a discrete number of values for the independent variable. Interpolation is the process of generating new data points between an existing discrete set of data points, allowing us to estimate intermediate values. Several interpolation methods exist, but we restrict ourselves to cubic spline interpolation, which uses piecewise 3rd-degree polynomials between observed data points [18].

The cubic spline interpolations take two known data points and create a 3rd-degree polynomial between the data points [19]. Assume the data points (x_i, y_i) and (x_{i+1}, y_{i+1}), then between these points we have the following 3rd-degree equation:

$$S_i(x) = a_i x^3 + b_i x^2 + c_i x + d_i \tag{2.1}$$

That is valid for $x_i \leq x \leq x_{i+1}$. The four coefficients, a_i, b_i, c_i, and d_i must be calculated to solve this equation. This is done by using the fact that the endpoint

for the first spline has the same value, y, as the second spline starting point. Also, by setting the condition that the splines should be smoothly joined, the splines will have continuous first and second derivatives:

$$S_i'(x_{i+1}) = S_{i+1}'(x_{i+1}) \quad \& \quad S_i''(x_{i+1}) = S_{i+1}''(x_{i+1}) \tag{2.2}$$

By setting up these equations for each spline and setting arbitrary conditions to the endpoints, all coefficients are solvable by creating a system of linear equations that can be solved by solving the matrix equation Ax=b. How a interpolation and the data points can look like is shown in Figure 3.8.

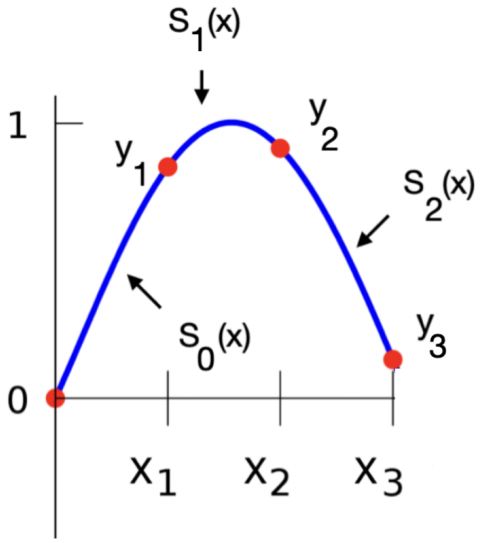


Figure 2.6: Example of the polynomial interpolation variables used to calculate a new polynomial.

2.5 Statistical analysis

This section contains a brief summary on statistical methods, such as Pearsons correlation, probability value and z-test.

2.5.1 Pearson Correlation

The Pearson correlation coefficient, r, is a value used to define the linear correlation between two variables [20]. The coefficient r is between -1 $\leq$ r $\leq$ 1 and indicates how strong the correlation or negative correlation is between the variables.

The Pearson coefficient r is the coefficient for a linear function that minimizes the distance to all data points in the data set. The formula to calculate the Pearson

coefficient, r, is defined in Equation 2.3, where n is the sample size, x_i and y_i are the individual samples indexed by i, and the sums run over all sample indices.

$$r = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}} \tag{2.3}$$

2.5.2 Probability value

Using the probability value, p is beneficial during hypothesis testing where $0 \leq$ p ≤ 1 [21]. The p-value expresses the probability, under the assumption of the null hypothesis, of obtaining a result equal to or more extreme than what is observed. The null hypothesis is a guess on what the statistical analysis should have as a result.

To determine when the null hypothesis should be rejected, a significance level is set, α. A typical value for α is 0.05, which means that when p $< \alpha$, the obtained result is with 95% certainty, not random chance. This indicates that the null hypothesis is unlikely and should be rejected. In the case where p $> \alpha$, the p-value indicates evidence for the null hypothesis [22].

2.5.3 Z-test

Z-test is a type of statistical test that uses random samples from populations to conclude, where a population in statistics refers to a pool of individuals [23], [24]. Z-test is used instead of t-test when the sample size is larger than 30 samples. There are two types of z-test, 1-sample and 2-sample analysis [24]. 1-sample analysis determines if the population mean and the hypothesized value differs. In contrast, 2-sample analysis determines if two populations' mean significance level null hypothesis (H_0) and alternative hypothesis (H_A) that are compared to the mean (μ) of the population. In 1-sample analysis, the null hypothesis H_0 equals a value μ_0, which is correct if $\mu_0 = \mu$. If $\mu_0 \neq \mu$, then the null hypothesis is rejected, and the alternative hypothesis is correct because the definition of the alternative hypothesis is that $\mu_0 \neq \mu$. However, in the 2-sample analysis, the null hypothesis H_0 is that the two different population means are the same, $\mu_1 = \mu_2$. If this is not true, the alternative hypothesis H_A, $\mu_1 \neq \mu_2$, will be true.

To successfully use the z-test, these basic assumptions are needed for good results: random samples, continuous data, normal distribution, or large sample size; in 2-sample, the data sets must be independent, and the standard deviation must be known for the population. If these demands are achieved, the statistical significance can be calculated with these z-test formulas:

$$Z = \frac{\overline{\mathrm{x}} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \tag{2.4}$$

$$Z = \frac{\overline{\mathrm{x}_1} - \overline{\mathrm{x}_1}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \tag{2.5}$$

Where $\overline{x}$ is the population's mean, μ_0 is the hypothesized value, σ is the standard deviation, σ^2 is the variance, and n is the sample size. The first equation, Eq. (2.4), is for 1-sample analysis, and the second equation, eq. (2.5), is for a 2-sample analysis. Both 1-sample and 2-sample analysis has three subcategories of z-test: two-tailed, right one-tailed, and left one-tailed test.

The null hypothesis represents the z-score being equal to 0; the further the z-value is from 0, the more significant the effect is, therefore, more likely to have a more considerable statistical significance. A tail represents the significance level; when z-scores reach the critical values, the significance level covers them and will be classified as statistically significant [25].

3

Methods

The methods chapter outlines the approach and methodologies used in the project, containing data collection, data processing and analysis methods.

3.1 Data collection

This section describes the data collection. This includes the recruitment of participants, test protocol, test setup, sensor, and system configurations.

3.1.1 Participants

Participants were recruited by contacting several different student groups on Chalmers and track and field groups in Gothenburg. In total, 47 people participated in the data collection. The participants consisted of 14 female and 33 male runners, aged 18-56, with fitness levels between elite and not exercising. Each attribute with the mean value and standard deviation for male and female runners is found in table 3.1.

Table 3.1: The value n represents the number of participants of each gender (n=47 in total). Each attribute that we collected together with the consent form is listed with the average ± and the standard deviation together with (min-max) values.

Attribute	**Male** (n=33)	**Female** (n=14)
Age	25.9 ± 8.3 (19-56)	24.9 ± 4.8 (18-35)
Weight	73.7 ± 8.5 (53-92)	59.9 ± 5.7 (45-72)
Length	180.8 ± 6.3 (169-196)	170.2 ± 5.9 (163-186)
Leg length (hip to knee)	47.8 ± 2.1 (44-53)	44.6 ± 3.0 (41-50)
Leg length (knee to floor)	58.5 ± 2.4 (53-63)	54.8 ± 2.9 (47-60)

3.1.2 Testing protocol

The testing protocol was developed using the most common technique mistakes and instructions described in the theory chapter. In addition, velocities were chosen to represent the full spectrum most runners train at. Finally, time at each velocity was chosen to allow amateur runners to perform the test without fatigue. The test protocol is shown in Table 3.2.

Table 3.2: The list of technique instructions given during the test protocol and in which order they were given.

Number	Technique instruction
1	Natural running, run as you are used to
2	Try to land further back, not to overstride
3	Lift your knees a little higher
4	A little higher back kick
5	Push the hip forward, straighten the upper body, and slightly lean forward
6	Increase the step frequency a little

All participants in this specific order were given the technique instructions specified in table 3.2. Each technique instruction was performed during a 1.5 min or 2 min interval. Each interval consisted of 1 min at 12 km/h and 30 sec at 15 km/h. Additionally, a 30 sec on 18 km/h was added if the participant felt comfortable with the velocity. All tests were conducted at the Chalmers Physiology lab using a Rodby treadmill, a picture of the lab is shown in Figure 3.1.

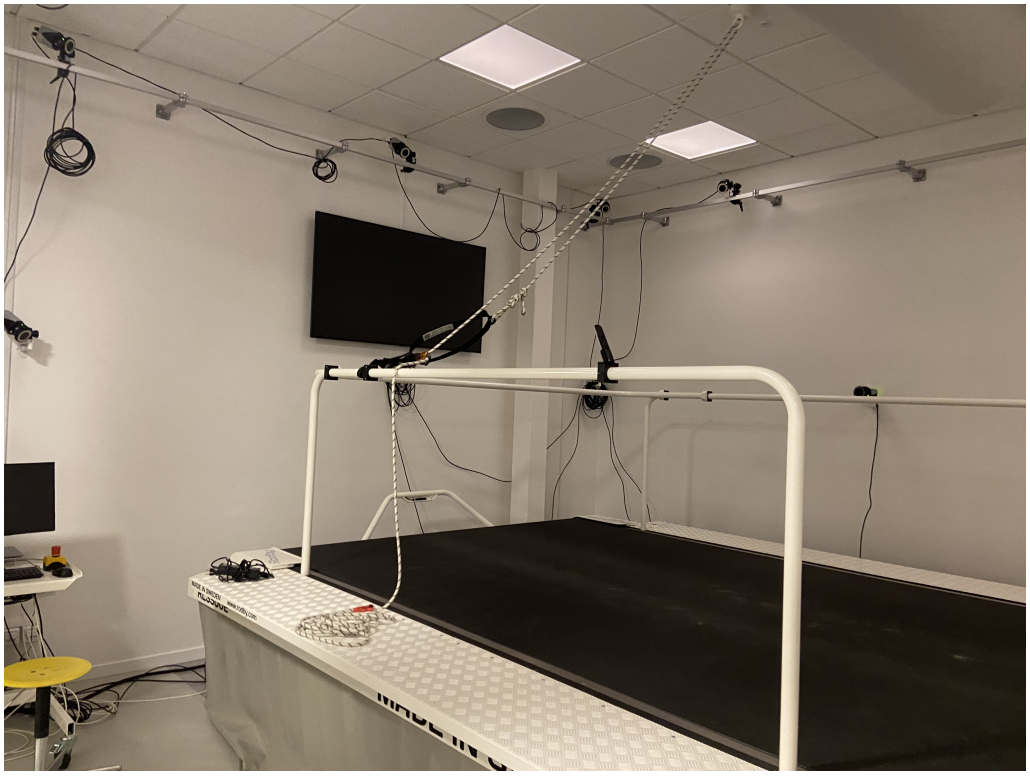


Figure 3.1: Chalmers TRACKS physiology lab.

In addition to the running part, we collected primary data from the participants on age, weight, height, gender, leg length, and athletic history. In the same documentation, referred to as a consent form and found in Appendix B, each participant consented to save this information and the data collected from the test.

3.1.3 Placement of the sensor and markers

The IMU was placed in the middle of the foot, where the foot is at its widest. Four reflective markers were placed with the intention of forming a foot. One marker at the toe, one on each side of the IMU (left and right), and one at the heel. The placements are illustrated in figure 3.2. Finally, the IMU and the markers were attached with tape to the shoe.

Figure 3.2: The placement of the markers that are represented with black dots in the figure, and the placement of the IMU sensor that is represented with a blue square.

3.1.4 IMU data

IMU data was collected using the MetaMotionS sensor from MBIENTLAB [11]. The device includes, in addition to sensors, a microcontroller with memory and Bluetooth transmission. The data that the sensor gives consists of 3-axis acceleration and rotation data. The data was streamed over Bluetooth using a sampling frequency of 100Hz and with maximum sensitivities of $\pm$16G for acceleration and 2000°/s for rotation.

Due to a large amount of noise and measurement errors, coupled with difficulties in finding initial conditions, we used the raw data of acceleration and rotation instead of attempting to integrate to obtain position and orientation. We also choose to use total acceleration and rotation values as the IMU reports values in its own relative coordinate system, which would differ depending on how the IMU was mounted to the shoe.

3.1.5 Marker based data

Marker-based data was collected using Qualisys Motion capture system [26] at Chalmers TRACKS Physiology lab. The setup has 12 marker-based cameras and 2 video cameras angeled at the treadmill. The videos that the Qualisys system captures during the tests will not be used as a part of the analysis in this project. The Qualisys system uses a software called Qualisys Track Manager (QTM) to capture the markers [27].

To collect the data from the markers, we created an Automatic Identification of Markers (AIM) model in QTM using the placement described in chapter 3.1.3. Also, the same sample frequency was used during the measurements, 100 Hz, as for the IMU. The benefit of using an AIM model is that it gives an example of how the markers are placed compared to the rest of the markers so that the motion-capture cameras can easier find all the markers during each test. This is particularly useful when having multiple test participants with different shoe sizes, among other differences [28]. The AIM model that was created and used is shown in Figure 3.3. In Figure 3.4 we can see what a couple of steps look like during a trial in the QTM software. The higher curve represents how the heel moves and the lower one is how the toes move.

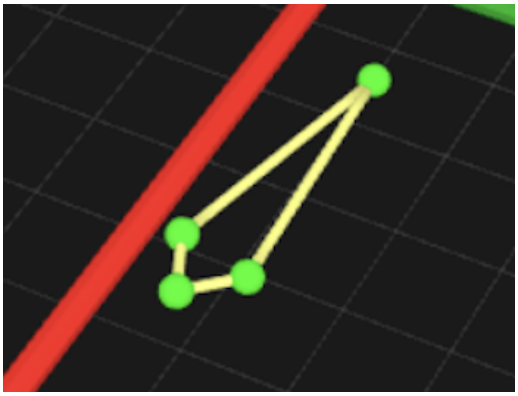


Figure 3.3: AIM model formed as a foot that shows the 4 markers placement during each test.

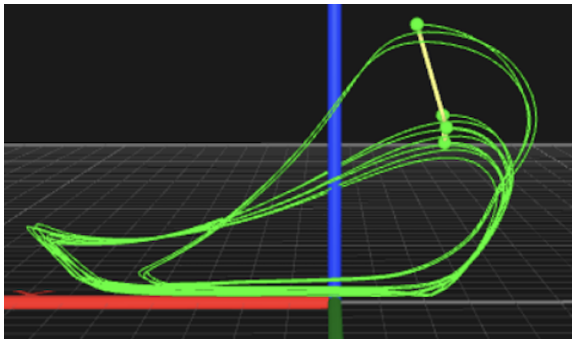


Figure 3.4: A couple of step cycles from QTM, the higher curve represents the heel and the lower curve represents the markers on the toes.

3.2 Data processing

This section describes the steps taken to obtain a useful representation of data for further analysis from the cleaned data. This includes:

- Data cleaning, such as handling missing data and errors in data.
- Sensor fusion
- Generating a sample stride from all runners during all technique instructions for IMU and marker data.
- Transforming data by interpolation to have the same dimensions, as sample frequency was constant during all tests, but runners have different step frequencies resulting in different dimensions for their sample stride.
- Perform and evaulate dimensionality reduction methods to obtain a lower dimension representation suitable for further analysis.

3.2.1 Missing data

During a measurement done with the Qualisys system, it is not always guaranteed that all markers are found at every time sample. However, the QTM adds the ones it finds but is insecure of, under its own category: unidentified trajectories. An example of how it looks in QTM is shown in Figure 3.5. This makes it possible, to some extent, correct the QTM mistakes manually. There are unfortunately also cases when the motion cameras cannot find a marker at all, for example, because of where the marker is and the camera angles are not optimal for that marker and time sample. After the measurement, the software summarizes the measurement by showing each marker's fill level, which is a percentage of how much of each marker the system found. When overlooking all 272 files (47 participants and 6 intervals each), each has 4 markers with 9000 or 12000-time samples. We decided to define a data set as good enough when each marker has at least a fill level of 90 %.

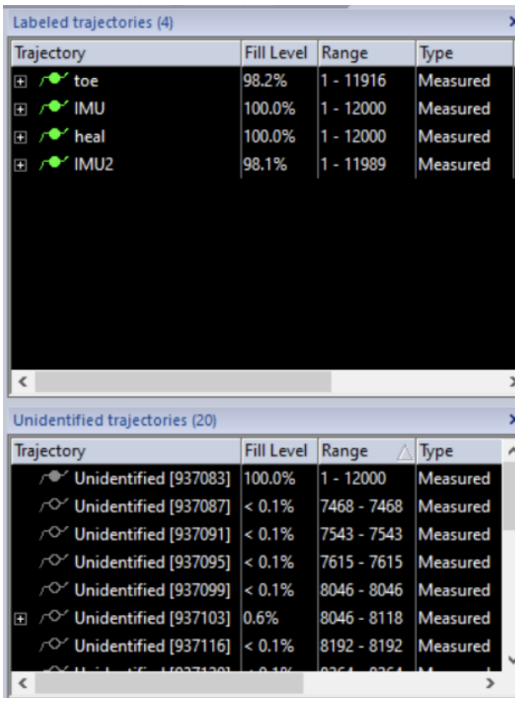


Figure 3.5: An example from QTM on fill level and unidentified trajectories

3.2.2 Incorrect data

The data included several incorrect data points, such as misidentification by the Qualisys track manager software, markers falling off during trials, or incorrect sensor values from the IMU. Unfortunately, due to the volume of data, it was infeasible to correct all these errors manually. Apart from the manual trajectory identification described above, errors were handled automatically using outlier detection and discarding low-quality data when processing.

3.2.3 Sensor fusion

The IMU data sampling was separate for gyroscope and accelerometer sensors, with signals sometimes suffering from time shifts compared to each other. For example, the accelerometer data could be sampled 1 ms after the gyroscope data. The sampling frequency was also not always consistent and samples could be delayed or early by $\pm$ 2 ms.

Due to this, the merging of data was done by matching on the nearest timestamp instead of equal keys using the merge-as-of method in pandas [29]. A tolerance of 5ms was used to obtain desired behavior, that shifted samples should be merged together, but periods with missing samples should not be incorrectly merged.

It should be noted that the data collected from the Qualisys system is not synchronized with the IMU data collected. It is, therefore not possible to compare specific steps from the IMU and marker-based to each other.

3.2.4 Generation of sample strides

The first step toward obtaining a sample stride is detecting individual strides from the data. This has previously been done using peak acceleration to determine footstrikes [30]. However, we found that at high velocity the push-off could have higher peak acceleration than the footstrikes making it difficult to use for automatic detection. The automatic step detection was performed using the find peaks algorithm with the idea outlined below.

```
function find peaks(samples,distance)
    peaks = empty list
    for sample in samples:
        if sample > max(other samples within distance from sample)
            add sample index to peaks
            skip to first sample outside distance
    return peaks
```

We used the scipy [31] implementation, which is optimized to further improve runtime, but the idea remains the same. To determine the best data source and filter length for stride start/stop detection, we evaluated based on the standard deviation of the distance between peaks. As the step frequency is very consistent when running, we can automatically evaluate how well the data source and filter work as errors would increase the standard deviation. We tested maxima and minima detection for rotational and accelerational data on raw data using moving averages of different filter lengths.

The best performance on IMU data was found using rotational data, detecting minimal, with a moving average with a filter length of 20 samples. An example of such detection is shown in Figure 3.6.

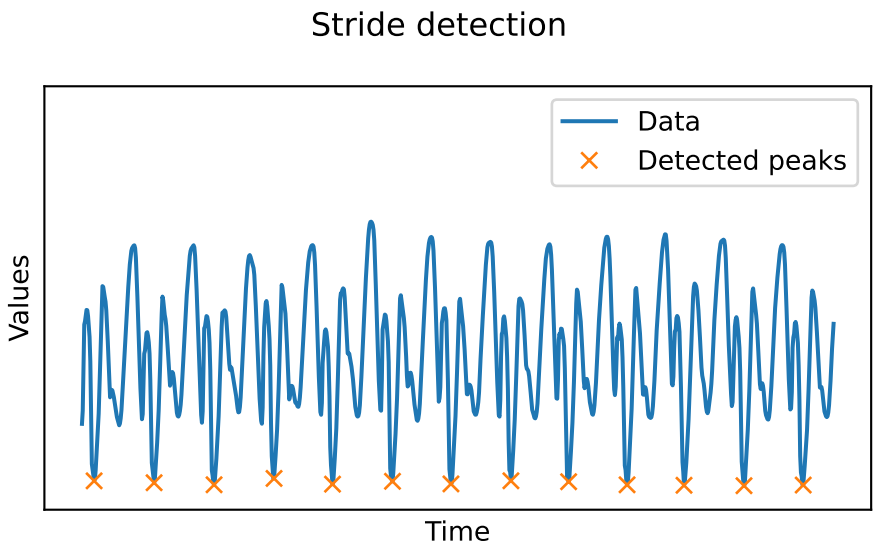


Figure 3.6: Stride detection using find peaks on the negative running sum of rotational IMU data.

This detection was much easier on marker-based data as data came in as coordinates, and it was more a matter of which point in the stride cycle to use as start/end. We used minimization over the Z coordinate with a moving average filter of length 20. This was chosen as it corresponds to the mid-phase of the step between foot-strike and push-off.

After detecting the start/endpoints of strides, these were aggregated into a single representative stride (for each runner, at each velocity, and for each technique instruction) for the IMU data using the median of all values at every timestamp from start to end, outlined below.

```
function aggregate(data,peak indices)
    stride times = difference between consecutive peak indices
    sample stride time = median(stride times)
    aggregated = empty list
    for i in range(0,sample stride time):
        temp = empty list
        for j in peak indices:
            add data[i+j] to temp
        add median(temp) to aggregated
    return aggregated
```

The use of median values where chosen as it automatically handles incorrect data or outliers, as long as these make up only a small portion of data while using mean could result in incorrect data affecting the result. For marker-based data, the representative stride was chosen as the individual stride, where the distance between the detected

start and endpoint was the smallest.

Different methods were used for several reasons. Firstly, the aggregation smoothed out differences due to sampling timing. As the IMU data was sampled, very different values could be obtained depending on the sampling timing relative to the stride. This phenomenon was especially present at landing and toe-off, and a visualization of the effect is shown in Figure 3.7.

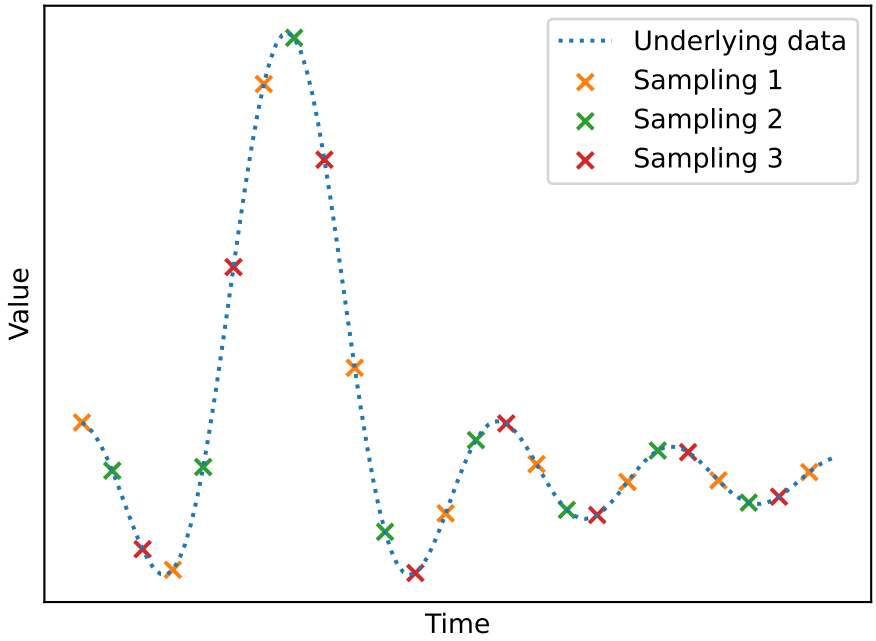


Figure 3.7: Visualization of how sampling timing can lead to different results when the underlying function varies quickly compared to the sampling frequency.

Secondly, issues with Bluetooth transmission meant that the IMU data had several missing samples, and the method allowed us to use as much of the available data as possible. Thirdly, some runners would move relative to the treadmill, and thus using aggregation, we would end up with a stride where the start and end position was not lined up for the marker-based positional data.

For IMU data, raw values were chosen as they contain absolute data. For marker-based data, we used position relative to the stride's start/stop detection point.

3.2.5 Frequency transformation

The next step was to transform the average stride data to the same size, as most common dimensionality reduction algorithms require data of the same dimensions. As runners have different step frequencies and we had a fixed sample rate, the data was of different dimensions.

The transformation was done using cubic spline interpolation to a size of 100-time units. This means that we change the time scale from absolute values to a scale

relative to one's own frequency. Figure 3.8 shows an example of the result of such interpolation on accelerational data from parts of a stride cycle.

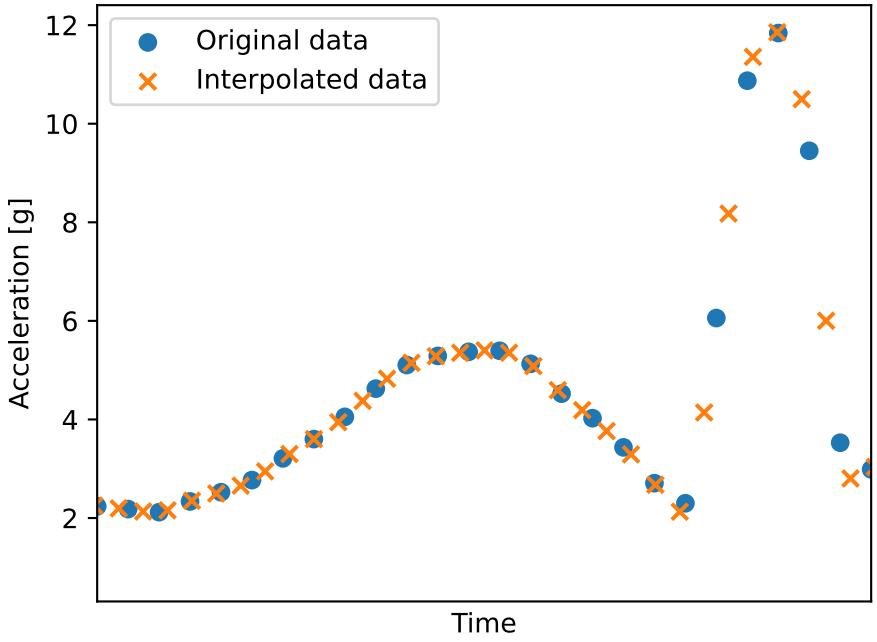


Figure 3.8: Example of interpolating acceleration data during a stride cycle.

3.2.6 Dimensionality reduction

Sequential feature selection, Principal component analysis, and autoencoder were tested to compare performance over different dimensions. The performance was evaluated based on the reconstruction error using K-fold cross-validation with five groups split on runners to ensure fair comparison and no data leaks.

Models were implemented using the scikit-learn API for machine learning in Python [32]. Feature selection and reconstruction were implemented using the SequentialFeatureSelector class with forward direction with performance and reconstruction evaluated and implemented using LinearRegression class. Principal component analysis was implemented using the PCA class, with the reconstruction done using the inverse_transform function. The autoencoder was implemented using the MLP class with an architecture of one hidden layer in the encoder and the decoder with ten neurons. The model was trained using mean squared error from input to output as loss, using the ADAM [33] optimizer, trained until no performance gains were seen for 100 epochs for a holdout validation set.

3.3 Data analysis

This section describes the data analysis done to answer RQ2 and RQ3.

3.3.1 Observed relationships

Observed relationships were analyzed using the Pearson correlation coefficient and p-value. The analysis was conducted using three attributes from the runners: Body mass index (BMI), leg length, and fitness level. These attributes were correlated against the compressed stride representations from all data sources. We also included several commonly defined stride parameters for analysis. These are outlined in Table 3.3.

Table 3.3: Definition of attributes and parameters in upcoming tables.

Attribute	
BMI	BMI, Body mass index, a relation between length and weight
LL	The total leg length (Leg length XX + YY)
FL	Fitness level, from reported personal best or estimations using athletic history stated in the consent form
Parameter	
Fr	Frequency, steps per minute
BK	Back-kick height, highest vertical position of the foot
Flatratio	Back kick height/stride length
KLAT	Knee lift at turn point, Vertical position of foot at the turn point before impact
GCT	GCT, Ground contact time, the time between foot strike and toe-off
CF-ratio	Contact to flight time ratio, the ratio of GCT to flight time, where the flight time is the time when the foot is in the air
MSM	Maximum sideways movement of foot, in positive Y direction
SMAT	Sideways movement of the foot at turning point before impact
LA	Angle of the foot at initial ground contact

3.3.2 Impact of technique instructions

The impact of technique instructions was evaluated using paired z-test, recommended over a t-test when the sample size is greater than 30 as stated in chapter 2.5.3. This analysis was conducted for each technique instruction at each velocity by comparing the normal (no technique) data to the technique instruction data, using the compressed stride representations from all data sources and the defined stride parameters.

4

Results

The results chapter presents the results. This includes analysis of dimensionality reduction methods to generate KPIs, analysis of the KPIs, the observed patterns between runner attribute and stride parameters, and the impact of the technique instructions.

4.1 RQ1, Dimensionality reduction

Figures 4.1-4.3 show the reconstruction error as a function of the number of dimensions for the selected dimensionality reduction and reconstruction methods on all data sources. This is used to judge the performance of reconstruction and determine how many dimensions are optimal.

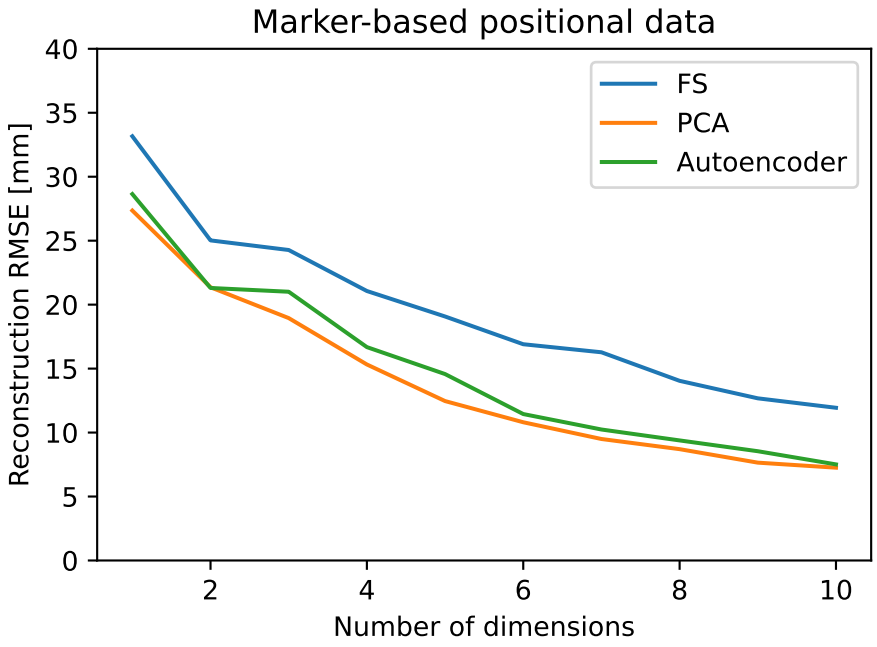


Figure 4.1: Reconstruction error as a function number of dimensions for FS, PCA, and Autoencoder on positional data from the marker-based system.

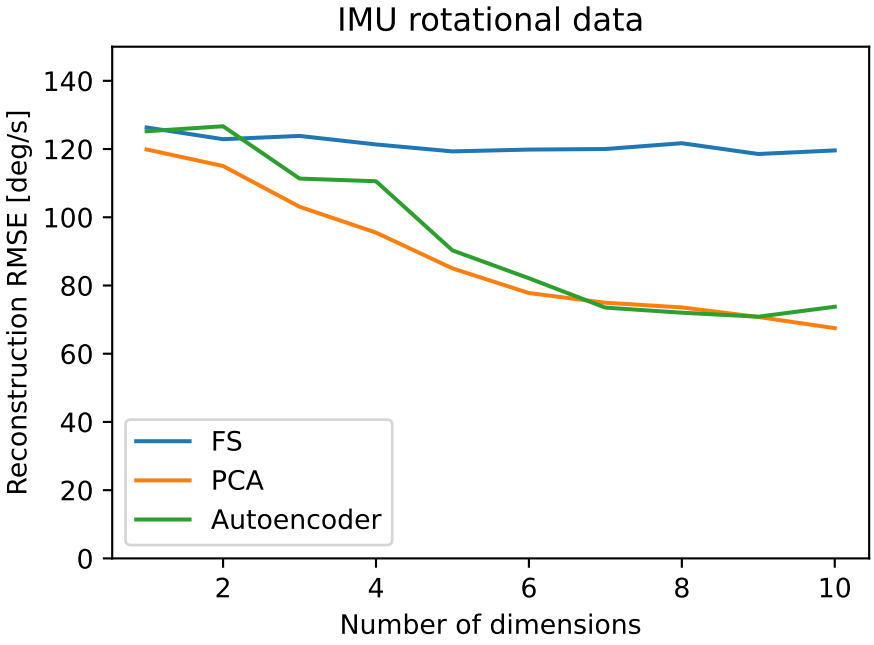


Figure 4.2: Reconstruction error as a function number of dimensions for FS, PCA, and Autoencoder on rotational data from the IMU.

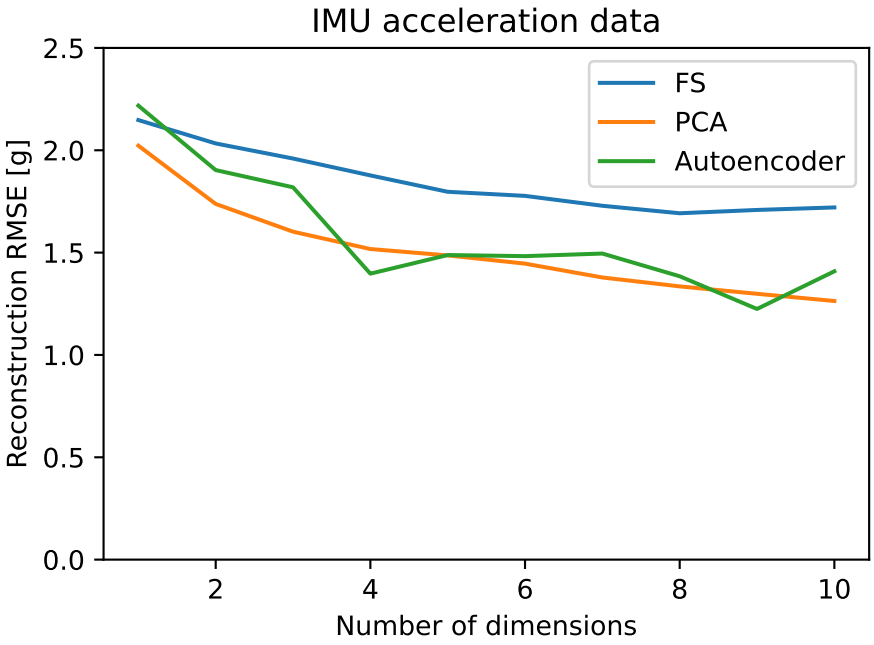


Figure 4.3: Reconstruction error as a function number of dimensions for FS, PCA, and Autoencoder on acceleration data from the IMU.

As can be observed, PCA performs best or close to best at all data sources and at all tested numbers of dimensions, with autoencoder having close performance and feature selection a larger performance loss. Based on performance and the fact that it is deterministic, unlike the autoencoder, PCA was selected as the dimensionality reduction method for further analysis.

PCA enables analysis of the explained variance from each individual principal component using the eigenvalues. Table 4.1 shows the explained variance ratio and cumulative sum of explained variance ratio after each additional principal component for the first 5 principal components on all data sources. The explained variance ratio reveals how much of the total variance in the data is explained by the principal component, and the cumulative sum reveals how much variance is explained by all principal components added so far.

Table 4.1: Explained variance ratio and cumulative sum for the first five principal components on all data sources.

Data source	Number of PCs:	1	2	3	4	5
Marker-based positional	Explained variance ratio	0.558	0.281	0.054	0.038	0.021
	Cumulative sum	0.558	0.840	0.894	0.932	0.953
IMU rotational	Explained variance ratio	0.416	0.135	0.093	0.077	0.047
	Cumulative sum	0.416	0.551	0.643	0.720	0.767
IMU acceleration	Explained variance ratio	0.246	0.184	0.155	0.062	0.049
	Cumulative sum	0.246	0.430	0.585	0.647	0.696

Visualizations of the first five principal components for the marker-based positional data are shown in Figures 4.4-4.8. The Figures display a side (XZ-plane), top (XY-plane), and front (YZ-plane) view of the average stride in blue and ± 2 standard deviations along the principal component direction in orange. As it can be hard to tell the timescale, the figures also include mappings between matching time indices to allow easier analysis of timing factors in the side view. The coordinate axes are the X-axis as the running direction, positive moving forward; the Y-axis as the left-right direction for the runner with positive as right; and the Z-axis vertical with positive up.

Figure 4.4 shows the first principal component, which accounts for 55% of variance in the strides. This component accounts primarily for stride size and frequency which can be seen best from the side view where the $\mu \pm 2\sigma$ strides are scaled, and the smaller has a higher frequency as judged from the matching time samples being further in the cycle in the smaller stride.

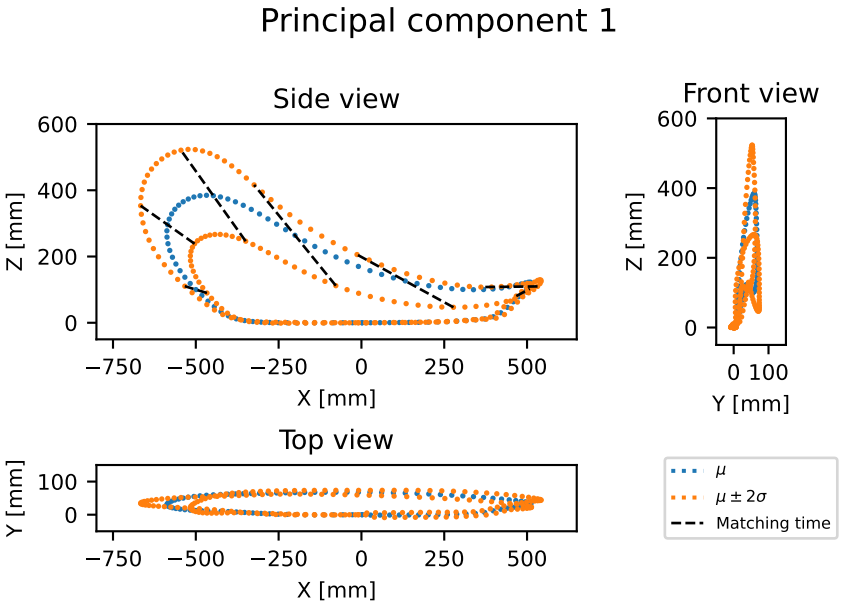


Figure 4.4: Principal component 1 on marker-based positional data.

Figure 4.5 shows the second principal component, which accounts for 28% of variance in the strides. It corresponds mostly to the vertical position during flight and the timing, which can be seen best in the side view with the $\mu \pm 2\sigma$ strides having different vertical positions in flight, and the larger stride having faster velocity, unlike in PC1.

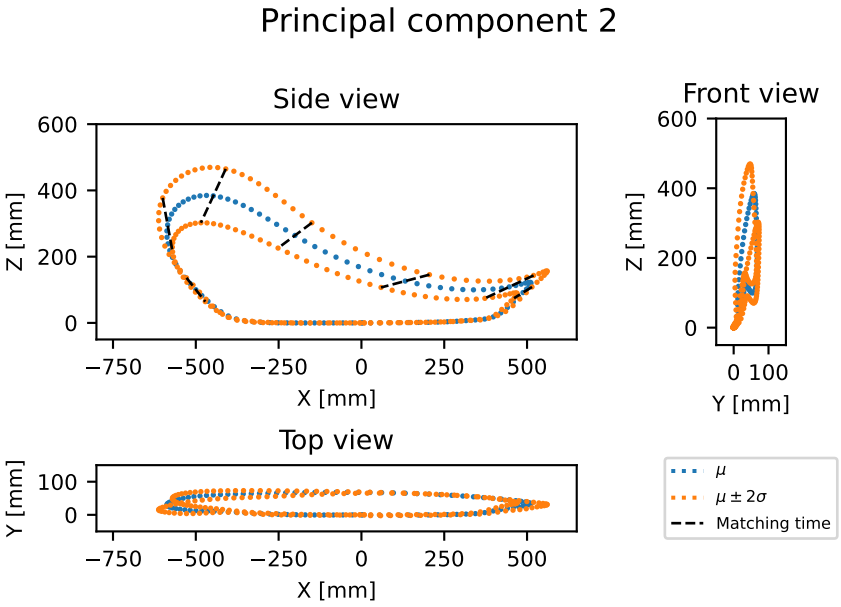


Figure 4.5: Principal component 2 on marker-based positional data.

Figure 4.6 shows the third principal component, which accounts for 5% of variance in strides. It correspond to a balance in vertical position and timing, with one direction

being back-kick further back, turning point further front, and flight in between faster and lower, and the other the reverse, back-kick and turn-point closer together and flight in between higher and slower.

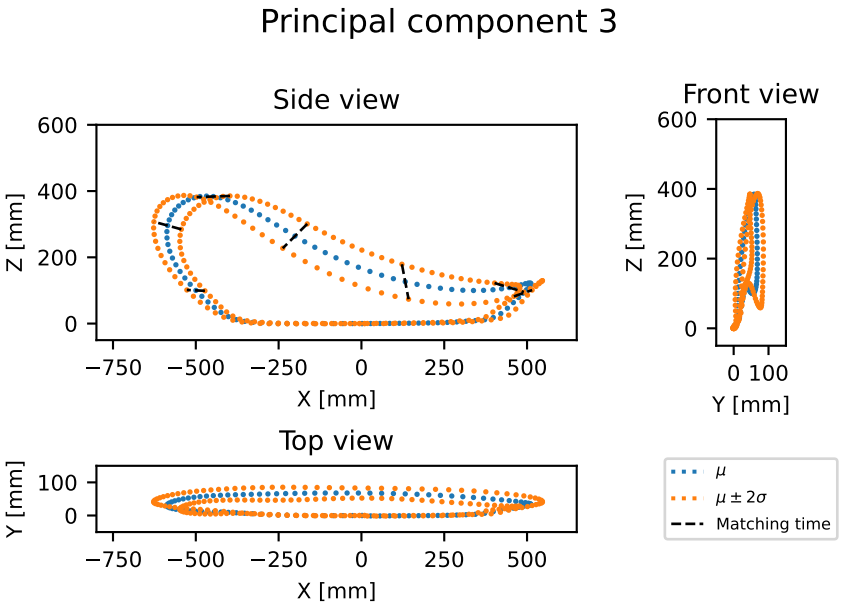


Figure 4.6: Principal component 3 on marker-based positional data.

Figure 4.7 shows the fourth principal component, which accounts for 4% of the variance in the strides. It corresponds almost entirely to the sideways movement of the foot and has almost no timing differences, which is best seen in the top or front views.

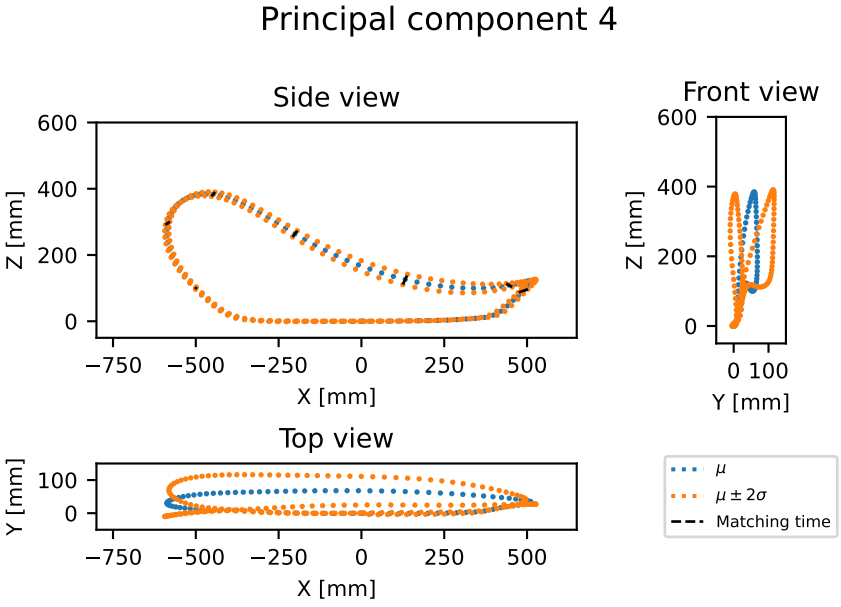


Figure 4.7: Principal component 4 on marker-based positional data.

Figure 4.8 shows the fifth principal component, which accounts for 2% of variance in the strides. It corresponds mostly to the position at the turning point before landing in both the running (X) and sideways (Y), best seen in the side and top view at X=500, and also has a minor difference during flight from back-kick to the turning point.

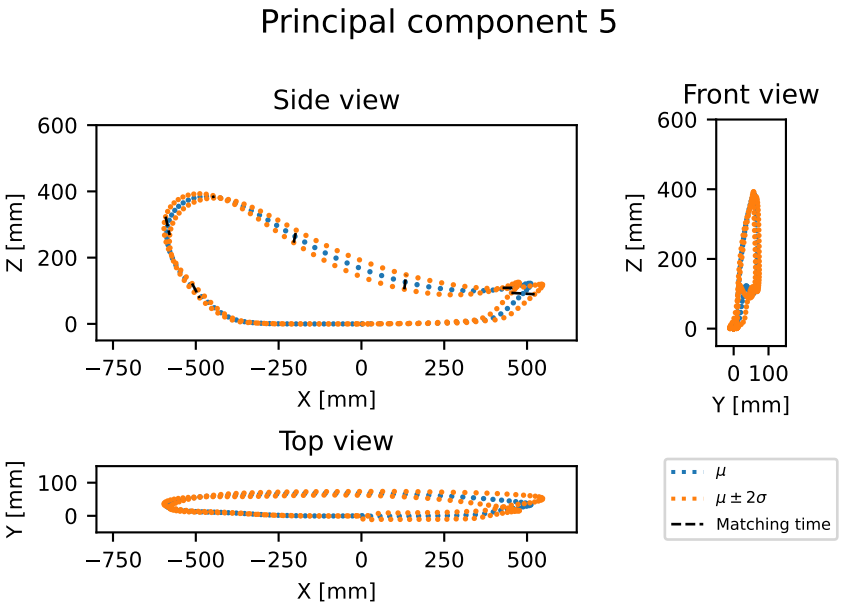


Figure 4.8: Principal component 5 on marker-based positional data.

4.2 RQ2, Pattern identification

Table 4.2 shows the observed relationships between the selected runner attributes and the defined stride parameters (as described in Table 3.3). Results are calculated between all combinations of attributes and parameters using Pearson correlation (r) and probability (p) and are presented separately for each velocity. P-values below 0.05 are highlighted with **bold** text to easily single out statistically significant correlations.

Table 4.2: Correlation between runner attributes and defined stride parameters.

Attri-		**12km/h**		**15km/h**		**18km/h**	
bute	**Parameter**	**r**	**p**	**r**	**p**	**r**	**p**
BMI	Fr [step/min]	-0.163	0.275	-0.107	0.490	-0.187	0.370
BMI	BK [mm]	-0.040	0.790	0.058	0.708	0.245	0.238
BMI	Flatratio	0.058	0.700	-0.043	0.782	-0.178	0.395
BMI	KLAT [mm]	0.250	0.090	0.269	0.077	0.099	0.636
BMI	GCT [ms]	0.232	0.116	0.149	0.334	0.085	0.686
BMI	CF-ratio	0.150	0.313	0.069	0.655	-0.010	0.960
BMI	MSM [mm]	-0.054	0.719	-0.201	0.190	-0.083	0.695
BMI	SMAT [mm]	-0.138	0.357	-0.115	0.458	-0.350	0.087
BMI	LA [°]	0.103	0.489	0.147	0.340	-0.172	0.412
LL	Fr [step/min]	-0.192	0.196	-0.377	**0.012**	-0.220	0.291
LL	BK [mm]	0.054	0.717	0.249	0.103	0.138	0.510
LL	Flatratio	-0.059	0.694	-0.201	0.192	-0.100	0.636
LL	KLAT [mm]	-0.268	0.068	-0.166	0.283	-0.226	0.277
LL	GCT [ms]	-0.008	0.956	0.057	0.712	0.123	0.559
LL	CF-ratio	-0.082	0.584	-0.097	0.529	0.014	0.948
LL	MSM [mm]	-0.155	0.297	0.074	0.634	0.370	0.069
LL	SMAT [mm]	-0.191	0.199	-0.035	0.823	0.045	0.830
LL	LA [°]	-0.194	0.192	-0.106	0.493	-0.062	0.769
FL	Fr [step/min]	-0.011	0.943	-0.154	0.318	-0.035	0.868
FL	BK [mm]	0.200	0.177	0.111	0.473	-0.029	0.889
FL	Flatratio	-0.268	0.069	-0.166	0.283	-0.033	0.876
FL	KLAT [mm]	-0.332	**0.023**	-0.195	0.205	-0.056	0.789
FL	GCT [ms]	-0.457	**0.001**	-0.322	**0.033**	-0.113	0.591
FL	CF-ratio	-0.441	**0.002**	-0.336	**0.026**	-0.105	0.617
FL	MSM [mm]	-0.125	0.403	-0.036	0.818	0.284	0.169
FL	SMAT [mm]	0.227	0.124	0.300	**0.048**	0.747	**0.000**
FL	LA [°]	-0.500	**0.000**	-0.418	**0.005**	0.024	0.911

BMI had no statistically significant correlations with any of the defined parameters. Leg length had a weak negative correlation with frequency but it was only significant at 15km/h, neither at 12km/h nor 18km/h.

More statistically significant correlations were found for fitness level and defined stride parameters. We observed a weak negative correlation with knee lift at the

turning point, but only at 12km/h.

For ground contact time and GCT to flight ratio, we observed a moderate negative correlation at 12km/h and a weak but significant negative correlation at 15km/h but no correlation at 18km/h. The better runners had lower GCTs and a larger proportion of the cycle spent in the air at slower velocities, but the pattern got weaker and disappeared as velocity increased. For side movement at the turning point, we found a strong correlation at 18km/h and a weak correlation at 15km/h. For the landing angle, we observe a moderate negative correlation at 12km/h and 15km/h but none at 18km/h.

4.3 RQ3, Technique instruction impact

In the tables 4.3-4.7 we have summarized the impact of each of the technique instructions (as defined in Table 3.2) on the defined stride parameters. Results were calculated for all technique instructions on all defined parameters using paired z-test and the change in mean ($\Delta\mu$) and probability (p) is presented for each velocity. P-values below 0.05 are highlighted with **bold** text to easily single out statistically significant changes.

In figures 4.9-4.13 we present the average normal (no technique instruction) stride and the average stride under each technique instruction. The figures follow the same convention as figures 4.4-4.8, showing side, front, and top views with and including mappings (lines) between matching time indices in the side view.

4.3.1 T2: Land further back, reduce overstriding

The instruction to land further back and try to reduce overstriding had a limited impact on the technique. Table 4.3 presents the changes in the mean value ($\Delta\mu$) for the defined parameters and the corresponding p-values. We observed a statistically significant reduction in knee lift at turning point 15 and 18 km/h and a small increase in frequency at 15km/h. Suprisingingly, we did not find any major changes in the landing angle, which is what this instruction serves to change. A visualization of how the instruction alters the stride is shown in Figure 4.9. As can be seen, the movement pattern looks almost identical, with only a small difference in timing.

Table 4.3: Impact of technique instruction 2 on defined stride parameters.

Instru-		**12km/h**		**15km/h**		**18km/h**	
ction	**Parameter**	$\Delta\mu$	p	$\Delta\mu$	p	$\Delta\mu$	p
T2	Fr [step/min]	-0.323	0.665	1.937	**0.005**	1.852	0.161
T2	BK [mm]	10.246	0.227	2.109	0.804	-17.675	0.131
T2	Flatratio	0.032	0.779	0.068	0.462	0.113	0.099
T2	KLAT [mm]	-3.308	0.097	-7.699	**0.000**	-10.863	**0.004**
T2	GCT [ms]	1.064	0.721	-0.465	0.856	8.636	0.213
T2	CF-ratio	0.005	0.725	0.013	0.221	0.046	0.093
T2	MSM [mm]	1.152	0.744	3.362	0.255	-2.934	0.514
T2	SMAT [mm]	-0.147	0.955	1.770	0.477	-0.561	0.896
T2	LA [°]	-0.727	0.323	-0.690	0.390	0.989	0.708

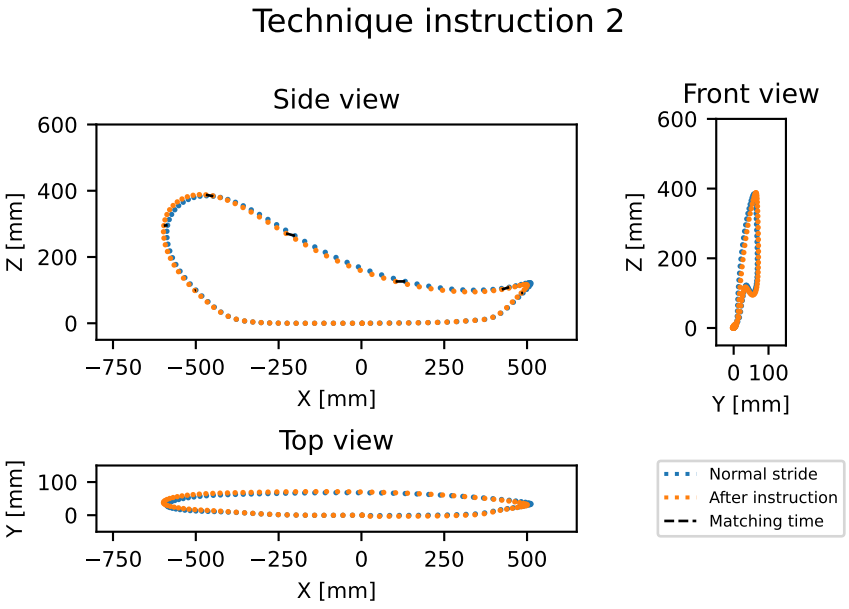


Figure 4.9: Average impact of technique instruction 2 on stride.

4.3.2 T3: Higher knee lift

The instruction to increase knee lift had a large impact on the technique, with significant increases in back-kick height, knee lift at the turning point, and reductions in flat-ratio, GCT, CF-ratio, maximum sideways movement, and landing angle. Table 4.4 presents the changes in the mean value ($\Delta\mu$) for the defined parameters and the corresponding p-values. We found significant changes in all parameters except sideways movement at the turning point at 12km/h and on many parameters at 15km/h. However, the impact was smaller at 18km/h, with the only significant change observed for the flat ratio. A visualization of how the instruction alters the stride is shown in Figure 4.10. As can be seen from the side view, the main difference is in the vertical position from the back-kick to the turning point.

Table 4.4: Impact of technique instruction 3 on defined stride parameters.

Instruction	**Parameter**	**12km/h** $\Delta\mu$	p	**15km/h** $\Delta\mu$	p	**18km/h** $\Delta\mu$	p
T3	Fr [step/min]	-1.995	**0.011**	-0.096	0.892	0.623	0.485
T3	BK [mm]	71.043	**0.000**	33.526	**0.000**	18.918	0.106
T3	Flatratio	-0.824	**0.000**	-0.247	**0.000**	-0.141	**0.034**
T3	KLAT [mm]	9.342	**0.001**	7.218	**0.028**	2.414	0.532
T3	GCT [ms]	-22.174	**0.000**	-17.143	**0.000**	-9.444	0.123
T3	CF-ratio	-0.090	**0.000**	-0.056	**0.000**	-0.021	0.346
T3	MSM [mm]	-7.900	**0.032**	-0.787	0.831	-8.702	0.134
T3	SMAT [mm]	-1.701	0.575	-1.737	0.534	-6.374	0.168
T3	LA [°]	-6.122	**0.000**	-6.221	**0.000**	-2.479	0.364

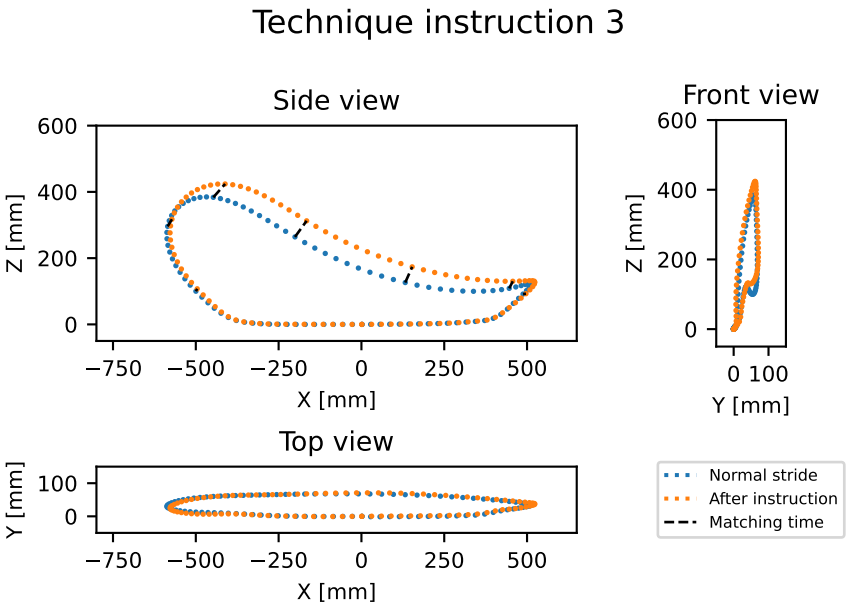


Figure 4.10: Average impact of technique instruction 3 on stride.

4.3.3 T4: Higher back kick

The instruction to increase back-kick height greatly impacted the technique, with significant increases to back-kick height, and reductions to frequency, flat-ratio, GCT, CF-ratio, and landing angle. Table 4.5 presents the changes in the mean value ($\Delta\mu$) for the defined parameters and the corresponding p-values. Changes in frequency, back kick, and flat-ratio, were present at all velocities but with the amount decreasing as velocity increased. For GCT, CF-ratio, and landing angle, we observed diminishing but significant effects from 12 to 15km/h but none or a reverse effect at 18km/h. A visualization of how the instruction alters the stride is shown in Figure 4.11. As can be seen from the side view, the main difference is the larger back-kick size and shape, with differences decreasing from the back-kick to the turning point.

Table 4.5: Impact of technique instruction 4 on defined stride parameters.

Instru-ction	**Parameter**	**12km/h** $\Delta\mu$	p	**15km/h** $\Delta\mu$	p	**18km/h** $\Delta\mu$	p
T4	Fr [step/min]	-5.290	**0.000**	-3.739	**0.000**	-4.239	**0.005**
T4	BK [mm]	161.460	**0.000**	119.898	**0.000**	92.084	**0.000**
T4	Flatratio	-1.297	**0.000**	-0.656	**0.000**	-0.413	**0.000**
T4	KLAT [mm]	1.450	0.513	-1.957	0.414	-4.946	0.255
T4	GCT [ms]	-19.783	**0.000**	-11.081	**0.000**	2.667	0.727
T4	CF-ratio	-0.098	**0.000**	-0.055	**0.000**	0.003	0.933
T4	MSM [mm]	-5.814	0.110	2.128	0.594	-2.347	0.706
T4	SMAT [mm]	-1.969	0.468	-3.048	0.419	0.730	0.940
T4	LA [°]	-3.711	**0.000**	-2.187	**0.033**	1.364	0.638

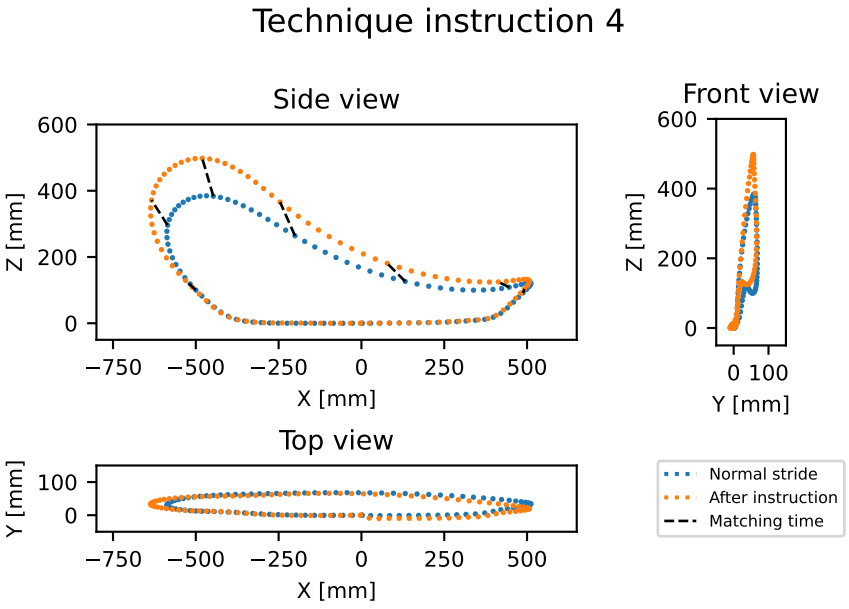


Figure 4.11: Average impact of technique instruction 4 on stride.

4.3.4 T5: Hips forward, forward lean

The instruction to push hips forward and lean forward had a moderate impact on the technique, with small but significant increases in back-kick height and reductions in knee lift at the turning point, GCT, CF-ratio, and landing angle. Table 4.6 presents the changes in the mean value ($\Delta\mu$) for the defined parameters and the corresponding p-values. Effects were spread between velocities, with no parameter seeing significant changes at all velocities. For maximum sideways movement we observed statistically significant changes in different directions at 15 and 18km/h, which is unexpected. We also saw the largest reduction in side movement at the turning point at 18km/h out of all instructions. A visualization of how the instruction alters the stride is shown in Figure 4.12. As can be seen, the movement pattern looks almost identical.

Table 4.6: Impact of technique instruction 5 on defined stride parameters.

Instru-		**12km/h**		**15km/h**		**18km/h**	
ction	**Parameter**	$\Delta\mu$	p	$\Delta\mu$	p	$\Delta\mu$	p
T5	Fr [step/min]	0.442	0.423	1.059	0.156	-0.943	0.296
T5	BK [mm]	20.236	**0.008**	14.982	0.115	13.026	0.227
T5	Flatratio	-0.165	0.242	-0.069	0.276	-0.030	0.552
T5	KLAT [mm]	-3.035	0.094	-6.926	**0.000**	-7.147	**0.037**
T5	GCT [ms]	-8.936	**0.002**	-5.000	**0.042**	-4.500	0.149
T5	CF-ratio	-0.027	**0.023**	-0.009	0.308	-0.014	0.125
T5	MSM [mm]	5.243	0.198	6.948	**0.003**	-11.443	**0.011**
T5	SMAT [mm]	-1.830	0.512	2.369	0.351	-12.311	**0.009**
T5	LA [°]	-2.239	**0.001**	-2.182	**0.009**	-0.723	0.570

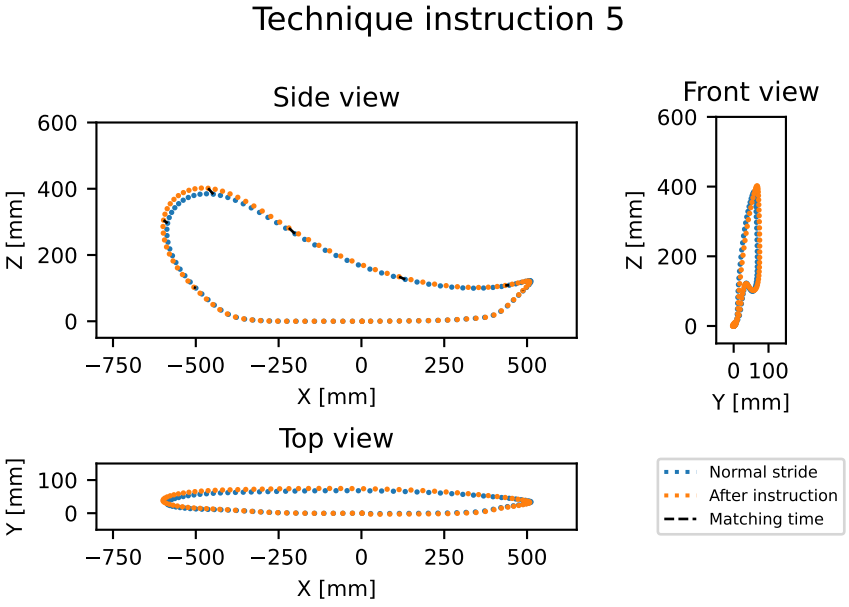


Figure 4.12: Average impact of technique instruction 5 on stride.

4.3.5 T6: Increase frequency

The instruction to increase frequency had a large impact on the technique, with significant increases in frequency, flat-ratio, and CF-ratio, and reductions in back-kick height, knee lift at the turning point. Table 4.7 presents the changes in the mean value ($\Delta\mu$) for the defined parameters and the corresponding p-values. The effects on frequency, back-kick height, flat-ratio, knee lift at the turning point, and CF-ratio were present all velocities. We also observed a significant reduction to GCT at 12km/h, sideways movement at 18km/h, and landing angle at 12km/h. A visualization of how the instruction alters the stride is shown in Figure 4.13. As can be seen from the side view, the main difference is a smaller back-kick size, with differences decreasing from the back-kick to the turning point.

Table 4.7: Impact of technique instruction 5 on defined stride parameters.

Instru-		**12km/h**		**15km/h**		**18km/h**	
ction	**Parameter**	$\Delta\mu$	p	$\Delta\mu$	p	$\Delta\mu$	p
T6	Fr [step/min]	12.174	**0.000**	13.526	**0.000**	13.653	**0.000**
T6	BK [mm]	-45.554	**0.000**	-69.157	**0.000**	-93.603	**0.000**
T6	Flatratio	0.602	**0.000**	0.564	**0.000**	0.518	**0.000**
T6	KLAT [mm]	-12.207	**0.000**	-14.597	**0.000**	-21.092	**0.000**
T6	GCT [ms]	-10.213	**0.001**	-3.409	0.138	2.778	0.670
T6	CF-ratio	0.033	**0.009**	0.057	**0.000**	0.070	**0.009**
T6	MSM [mm]	2.302	0.565	5.337	0.117	-5.321	0.356
T6	SMAT [mm]	-0.503	0.846	1.837	0.444	-9.886	**0.026**
T6	LA [°]	-3.416	**0.000**	-0.600	0.542	1.801	0.478

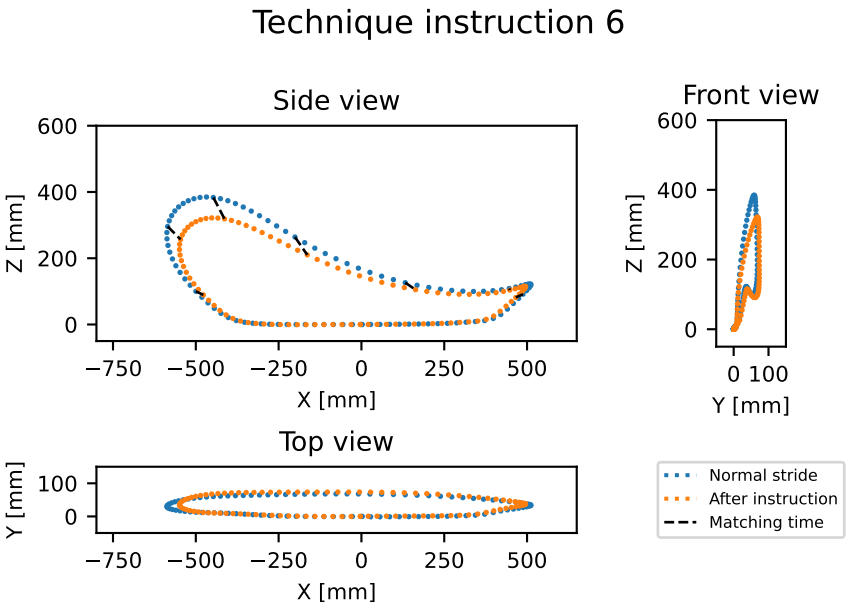


Figure 4.13: Average impact of technique instruction 6 on stride.

5

Discussion

5.1 RQ1, Dimensionality reduction

Principal component analysis obtained the best performance on all data sources. This is somewhat surprising, given that the Autoencoder should be a more capable model, but the limited data likely explains it, as it becomes difficult to draw the line between under and overfitting. Performance was evaluated on an unseen test set, but we also tested how non-linear the data was by letting the Autoencoder fit the full dataset. This obtained similar results as the PCA, and we can assume that even with a larger sample size, the Autoencoder would not significantly outperform PCA.

One can also note that our feature selection implementation performs pretty well compared to expectations on marker based data. However, convergance is not steady, which could be due to the greedy algorithm adding one feature at a time. It is possible that there exists other combinations of features that could have performed better at these number of dimensions, but searching for them would be infeasible as the runtime grows exponentially.

The information loss was significantly larger on both IMU data sources than on marker-based data for the same dimensions. There was no obvious choice of dimensions, with each additional dimension adding a little bit more information.

All dimensionality reduction and reconstruction were done by flattening the data and treating each timestamp (and coordinate in positional data) as separate dimensions. We did experiment with other data transformations, such as Fourier transform and polar coordinates, but found no increase in performance.

As described in 4.1 the first five principal components corresponded to factors that could be easily identified. However, the benefit is not certain compared to using, for example, the defined parameters. It is interesting to note that 84% of the variance in positional data is explained by only 2 KPIs that seem to correspond mostly to stride size, frequency, and timing.

Figure 5.1 shows the average impact of each technique instruction on the first four principal components. As seen, the instruction vectors span the plane and one should be able to control these PCs by combining technique instructions. However, from a coaching standpoint, it would have been better to have the compressed representation in directions that directly corresponded to technique instructions, such as the T4-T6

direction in the PC1 & 2 plane. Doing this would be perfectly possible by just expressing the representations as linear combinations of the PCs.

5.2 RQ2, Pattern identification

Overall we found limited patterns and correlations between runner anatomy and the defined stride parameters. Especially for leg length, this is a very surprising result as one would believe that the stride size would scale with leg length, and many of the defined parameters depend largely on stride size. For fitness level, we found patterns in ground contact time and contact-to-flight ratio, which are already studied and known to correlate with running economy [34], and as such, are of limited importance as they are more dependent on ability than technique. The largest correlation was for sideways movement at the turning point, where the better runners had lower sideways dispersion. This is a stride parameter that is not well-known or studied compared to others and could call for further study.

There is a legitimate possibility that most participants were too good runners due to selection bias, as a non-runner might not choose to participate. The majority of participants had some athletic background and might thus have developed good running techniques despite not currently being in shape. It would thus have been interesting to perform the study on more sedentary individuals with no athletic background, which could possibly have led to more patterns being seen.

5.3 RQ3, Technique instruction impact

As described in section 4.3, the instructions with the most impact on the defined stride parameters were increased knee lift, back-kick height, and frequency. Figure 5.1 shows a visualization of the average impact of each technique instruction on the first four principal components from the positional data. Not surprisingly, these instructions also have the largest impact on the principal components, as can be told from the vector lengths. We can also notice that the high back-kick and increased frequency instructions have opposite effects in the PC 1 & 2 plane (which makes up 84% of variance). Higher knee lift has a similar effect to back-kick on PC2 but without impacting PC1 as much, and is also the technique with the largest impact on PC3.

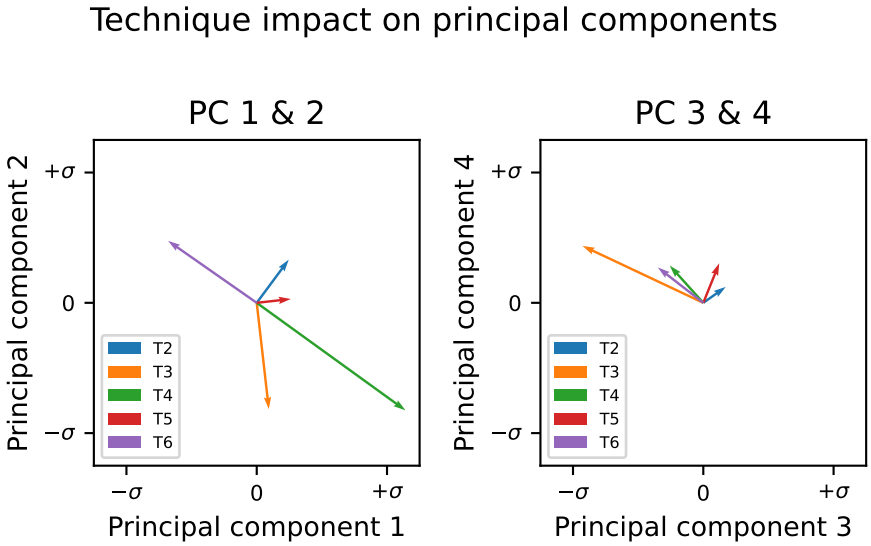


Figure 5.1: Average impact from technique instructions on principal components

The instruction to push the hip forward and lean forward had a smaller effect on most parameters compared to the three most impactful. However, we can observe that the effects correspond very well to the observed correlations between parameters and fitness level. For example, it had the largest reduction in side movement at the turning point, which had the highest correlation to fitness level. It could thus be the best instruction in terms of replicating what the best runners are doing. The instruction to land further back to reduce overstriding also had a limited impact on the stride. This could also be due to most participants not having a problem with overstriding, thus rendering the instruction useless.

It is also interesting to note that all instructions except landing further back resulted in significant decreases in ground contact times and landing angles at some velocities and altered PC4 in the same direction. From this, one can assume that just thinking about the technique lead to some implicit changes to the technique.

Overall the technique instructions work in such as way that they have consistent, similar effects for most runners, and we could use them to alter the stride in a desired way. The main problem is rather to determine how the stride should be changed as we found very limited patterns between fitness level and technique.

5.4 Future work

Several possibilities for extending the study exist, with some of them being:

- More variation in participants: As mentioned in 5.2, we believe that even the less fit participants had good running technique, and we would need more subjects with no or very limited athletic background to possibly find more patterns.

- Experiment with IMU data processing such as filter or integration. We used the raw data from the IMU as we found attempting to get positional data of sufficient quality from the IMU infeasible. Integration of IMU data for walking gait analysis is well studied with promising results [35]. However, extending these methods to running was not as easily done as we originally expected, most likely due to the much higher values of acceleration and rotation present when running compared to walking. One could also imagine using sequence-to-sequence prediction models such as recurrent neural networks but this was ruled out due to limited data amount.
- Orientation data. We only used positional and not orientational data from the marker-based system and it would be possible to extend the study to look at orientation data.
- Groupwise analysis of runners. One could extend RQ2 from looking at correlation to also trying to find the best technique. For example, one could imagine that there is an optimal stride pattern and that the better runners are more likely to be closer to it, and worse runners are more likely to be far off it. This would not be captured by correlation as the optimal is not at the extremes and would require a different type of analysis. One such method could be to split participants into groups, such as fast and slow, and analyze if the distributions differ between groups using, for example, using Kolmogorov–Smirnov test. We did some initial testing but found no differences between groups so perhaps this would also require more variation in participants to work.

6

Conclusion

We aimed to compare IMU data and marker-based data for data-driven running technique identification. To achieve this, we specified three research questions to cover the aim:

RQ1: Can IMU sensor data and marker-based motion capture data from running be quantified into key performance indicators (KPIs) describing the technique?

For marker-based positional data, we show that principal component analysis can be used to quantify running techniques with low information loss and high interpretability. The first two principal components correspond to 84 % of the variance in the data, and the first five 95%. Both acceleration and rotational data from IMU had significantly higher information loss.

RQ2: What patterns can be identified between technique and runner attributes?

We found limited correlations between anatomy and technique and moderate correlations between fitness level and technique, primarily on GCT, CCT to flight-time ratio, and sideways movement at the turning point before footstrike, which are parameters challenging to change by technique.

RQ3: How is the technique affected by common instructions?

We show the effect of five common technique instructions on the stride data, finding significant impacts when runners are asked to increase knee lift, back-kick height, and frequency, and less impact when they are asked to push their hip forward with lean and land further back.

Bibliography

[1] I. Moore, "Is there an economical running technique? a review of modifiable biomechanical factors affecting running economy.," *Sports Medicine*, no. 46, pp. 793–807, 2016. DOI: 10.1007/s40279-016-0474-4. [Online]. Available: https://doi.org/10.1007/s40279-016-0474-4.

[2] Folland JP, Allen SJ, Black MI, Handsaker JC, Forrester SE, "Running technique is an important component of running economy and performance.," *Med Sci Sports Exerc*, no. 49, pp. 1412–1423, 2017. DOI: 10.1249/MSS.0000000000001245. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5473370.

[3] Fukuchi RK, Fukuchi CA, Duarte M, "A public dataset of running biomechanics and the effects of running speed on lower extremity kinematics and kinetics," 2017. DOI: https://doi.org/10.7717/peerj.3298.

[4] T. F. Novacheck, "The biomechanics of running," *Gait Posture*, vol. 7, no. 1, pp. 77–95, 1998, ISSN: 0966-6362. DOI: https://doi.org/10.1016/S0966-6362(97)00038-6. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0966636297000386.

[5] E. van der Kruk and M. M. Reijne, "Accuracy of human motion capture systems for sport applications state-of-the-art review," *European Journal of Sport Science*, vol. 18, no. 6, pp. 806–819, May 2018. DOI: 10.1080/17461391.2018.1463397. [Online]. Available: https://doi.org/10.1080/17461391.2018.1463397.

[6] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[7] A. Chatzitofis, D. Zarpalas, P. Daras, and S. Kollias, "Democap: Low-cost marker-based motion capture," *International Journal of Computer Vision*, vol. 129, no. 12, pp. 3338–3366, 2021.

[8] *Compare motion capture systems*, https://www.comparesportstech.com/compare-motion-capture-systems.

[9] N. Ahmad, R. A. R. Ghazilla, N. M. Khairi, and V. Kasi, "Reviews on various inertial measurement unit (imu) sensor applications," *International Journal of Signal Processing Systems*, vol. 1, no. 2, pp. 256–262, 2013.

[10] L. Zhou, E. Fischer, C. Tunca, *et al.*, "How we found our imu: Guidelines to imu selection and a comparison of seven imus for pervasive healthcare applications," *Sensors*, vol. 20, no. 15, p. 4090, 2020.

[11] *Mbientlab research and development boards*. [Online]. Available: https://mbientlab.com/.

[12] L. van der Maaten, E. Postma, and H. Herik, "Dimensionality reduction: A comparative review," *Journal of Machine Learning Research - JMLR*, vol. 10, Jan. 2007.

[13] S. Khalid, T. Khalil, and S. Nasreen, "A survey of feature selection and feature extraction techniques in machine learning," in *2014 Science and Information Conference*, 2014, pp. 372–378. DOI: 10.1109/SAI.2014.6918213.

[14] J. Li, K. Cheng, S. Wang, *et al.*, "Feature selection," *ACM Computing Surveys*, vol. 50, no. 6, pp. 1–45, Dec. 2017. DOI: 10.1145/3136625. [Online]. Available: https://doi.org/10.1145/3136625.

[15] F. J. Ferri, P. Pudil, M. Hatef, and J. Kittler, "Comparative study of techniques for large-scale feature selection," in *Machine intelligence and pattern recognition*, vol. 16, Elsevier, 1994, pp. 403–413.

[16] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 4, pp. 433–459, Jun. 2010. DOI: 10.1002/wics.101. [Online]. Available: https://doi.org/10.1002/wics.101.

[17] Y. Wang, H. Yao, and S. Zhao, "Auto-encoder based dimensionality reduction," *Neurocomputing*, vol. 184, pp. 232–242, Apr. 2016. DOI: 10.1016/j.neucom.2015.08.104. [Online]. Available: https://doi.org/10.1016/j.neucom.2015.08.104.

[18] P. J. Davis, *Interpolation and approximation*. Courier Corporation, 1975.

[19] A. B. Qingkai Kong Timmy Siauw, *Python Programming and Numerical Methods*. Academic Press, 2020.

[20] S. Turney, "Pearson correlation coefficient (r)," vol. Guide Examples. Dec. 2022. [Online]. Available: https://www.scribbr.com/statistics/pearson-correlation-coefficient/.

[21] R. Bevans, "Understanding p-values," vol. Definition and Examples. Nov. 2022. [Online]. Available: https://www.scribbr.com/statistics/p-value/.

[22] "Understanding p-value in machine learining," Feb. 2023. [Online]. Available: https://www.askpython.com/python/examples/p-value-in-machine-learning.

[23] O. Momoh, "Population definition in statistics and how to measure it," 2023. [Online]. Available: https://www.investopedia.com/terms/p/population.asp.

[24] J. Frost, "Z test: Uses, formula examples," 2023. [Online]. Available: https://statisticsbyjim.com/hypothesis-testing/z-test/.

[25] J. Frost, "One-tailed and two-tailed hypothesis tests explained," 2023. [Online]. Available: https://statisticsbyjim.com/hypothesis-testing/one-tailed-two-tailed-hypothesis-tests/.

[26] *Qualisys motion capture systems*. [Online]. Available: https://www.qualisys.com/.

[27] *Qualisys track manager*. [Online]. Available: https://www.qualisys.com/software/qualisys-track-manager/.

[28] *Automatic identification of markers*. [Online]. Available: https://docs.qualisys.com/getting-started/content/getting_started/processing_your_data/using_aim_models/using_aim_models.htm?Highlight=AIM.

[29] T. pandas development team, *Pandas-dev/pandas: Pandas*, version latest, Feb. 2020. DOI: 10.5281/zenodo.3509134. [Online]. Available: https://pandas.pydata.org/docs/reference/api/pandas.merge_asof.html.

[30] K. Tumkur and S. Subbiah, "Modeling human walking for step detection and stride determination by 3-axis accelerometer readings in pedometer," in *2012 Fourth International Conference on Computational Intelligence, Modelling and Simulation*, IEEE, 2012, pp. 199–204.

[31] P. Virtanen, R. Gommers, T. E. Oliphant, *et al.*, "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," *Nature Methods*, vol. 17, pp. 261–272, 2020. DOI: 10.1038/s41592-019-0686-2.

[32] L. Buitinck, G. Louppe, M. Blondel, *et al.*, "API design for machine learning software: Experiences from the scikit-learn project," in *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, 2013, pp. 108–122.

[33] D. P. Kingma and J. Ba, *Adam: A method for stochastic optimization*, 2014. DOI: 10.48550/ARXIV.1412.6980. [Online]. Available: https://arxiv.org/abs/1412.6980.

[34] R. Di Michele and F. Merni, "The concurrent effects of strike pattern and ground-contact time on running economy," *Journal of science and medicine in sport*, vol. 17, no. 4, pp. 414–418, 2014.

[35] S. O. Madgwick, A. J. Harrison, and R. Vaidyanathan, "Estimation of imu and marg orientation using a gradient descent algorithm," in *2011 IEEE international conference on rehabilitation robotics*, IEEE, 2011, pp. 1–7.

A

Interview with Alexander Nilsson

Alexander Nilsson, kandidatexamen i Sportcoaching från Göteborgs Universitet och har haft flera uppdrag på Göteborgs Friidrottsförbund.

Jag skulle säga att de vanligaste felen hos motionärer är:

- Overstriding, dvs fotisättning för långt fram i steget
- För låg stegfrekvens
- För platt steg, dvs dåligt lågt knälyft och hälkick
- Bakåtlutad överkropp
- Sittande i höften

Sen finns det också mer individspecifika problem till exempel på grund av till exempel benlängdsskillnad, muskelimbalaser, dålig styrka eller smärta från skador som kan leda till osymmetriska rörelser som att knän eller höfter faller in. Många har dessutom en kombination av dessa och det är inte alltid uppenbart vad som är kärnan till problemet. Det finns naturligtvis också motionärer som har motsatsen till ovan nämnda problem men det är ovanligare. Generellt så skulle jag säga att löpare som varit aktiva från ung ålder, eller spelat fotboll etc oftast hittar en naturlig teknik och att det främst är personer som börjat i vuxen ålder på lägre nivå som har stora problem med tekniken.

Många av dessa teknikfel blir bättre med mer träning då kroppen är relativt bra på att hitta sin optimala teknik, men kan va bra med en medvetenhet till vad man ska jobba med för att inte fortsätta i samma hjulspår. Så genom att helt enkelt försöka tänka på att sätta i foten längre bak, öka frekvensen, lyfta på knä/hälkick och sträcka och luta fram kommer man i regel ganska långt.

Vad det gäller skillnaden mellan välträngade och riktig elit så är det i regel samma typer av problem fast i mindre omfattning. Riktigt duktiga löpare har också i regel en förmåga att hålla tekniken och springa avslappnat även när de blir trötta och därmed inte slösa energi. Sen finns det också faktorer som kan vara svåra att se med blotta ögat exempelvis lägre markkontakttid på grund av högre styvhet i exempelvis senor. Finns naturligtvis undantag även här, till exempel Timothy Cheruiyot som har helt annorlunda steg jämfört med resten av världseliten på 1500m.

B

Consent form

Grundläggande information inför testprotokollet

Detta är ett formulär för att samla in grundläggande data som vi behöver för att komplettera resultatet från testprotokollet, samt där du ger ditt skriftliga samtycke till att delta i testet.

Kom ihåg att du alltid har rätten att be om att få din data och dina personuppgifter borttagna.

Tack för att du medverkar i insamlingen av data till vårt examensarbete!
// Jessica Preiman och Johan Lamm

Namn?

_____

Mejladress?

_____

Födelseår?

_____

Kön?

_____

Vikt?

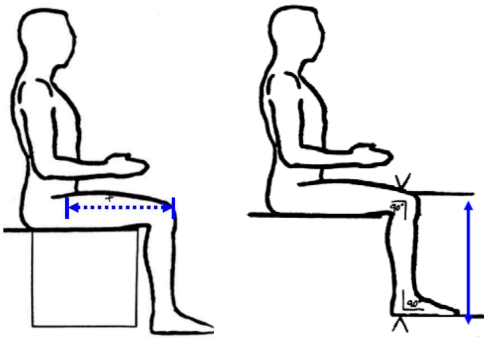
_____

Längd (i cm)?

_____

Benlängd XX, YY (i cm)?

_____



XX cm YY cm

Bild: *Int. J. Environ. Res. Public Health* 2010, *7*(3), 1047-1075; https://doi.org/10.3390/ijerph7031047

Idrottshistoria?
T.ex. hur ofta du tränar, typ av träning, mål, vad du har utövat tidigare, om du tävlar vad har du för PB osv.

_____

_____

_____

_____

_____

Genom att signera under godkänner du följande villkor:

- Att författarna till detta examensarbete inte är ansvariga för eventuella skador som sker under testet.
- Att min data kommer vara anonym i rapporten **men** att författarna till detta examensarbete kommer att lagrar ditt namn, den insamlade datan och informationen ovan.

_____ _____

Signatur Namnförtydligande