Forensic Interpretation Framework for Body and Gait Analysis: Feature Extraction, Frequency, and Distinctiveness

Dilan Seckiner[1,2], Xanthé Mallett[3] , Claude Roux[1], Simone Gittelson[1], Philip Maynard[1], Didier Meuwly[3,4]

[1] Centre for Forensic Science, University of Technology Sydney, 15 Broadway, Ultimo New South Wales 2007

[2] Institute of Forensic Medicine, University of Zurich, Winterthurerstrasse, 8057 Zurich, Switzerland: Dilan.Seckiner@irm.uzh.ch

[3]School of Law, University of Newcastle, New South Wales, 2308;
Honorary Associate in Centre for Forensic Science, University of Technology Sydney

[4]Netherlands Forensic Institute, Laan van Ypenburg 6, The Hague, The Netherlands, [5]University of Twente, Enschede, The Netherlands

ABSTRACT

Surveillance is ubiquitous in modern society, allowing continuous monitoring of areas that results in capturing criminal (or suspicious) activity as footage. This type of trace is usually examined, assessed, and evaluated by a forensic examiner to ultimately help the court make inferences about who was on the footage. The purpose of this study was to develop an analytical model that ensures applicability of morphometric (both anthropometric and morphological) techniques for photo-comparative analyses of body and gait of individuals in CCTV images, and then to assign a likelihood ratio. This is the first paper of a series: This paper will contain feature extraction to observe repeatability procedures from a single observer, in turn, producing the frequency and distinctiveness of the feature set within the given population. To achieve this, an Australian population database of 383 subjects (stance) and 268 subjects (gait) from both sexes, all ages above 18, and ancestries was generated. Features were extracted, defined, and their rarity viewed among the developed database. Repeatability studies were completed in which stance and gait (static and dynamic) features contained low levels of repeatability error (0.2% – 1.5 TEM%). For morphological examination, finger flexion, and feet placement were observed to have high observer performance.

1. INTRODUCTION

Gait analysis can be described as the manner in which a person undertakes a locomotor activity (walking or running) (Birch *et al.*, 2020). The gait cycle is the time between two consecutive occurrences of one repetitive event involved in walking (Birch *et al.*, 2020). The two major phases within the gait cycle are the stance phase (foot has ground contact) and the swing phase (foot is in the air) (*ibid*). The four stages within the stance phase includes: [1] loading response, [2] mid-stance, [3] terminal stance, and [4] pre-swing. The swing phase comprises three stages: [1] initial swing, [2] mid-swing and [3] terminal swing (Birch *et al.*, 2020). The assessment of gait from surveillance footage is considered a forensic tool that can potentially contribute to all stages of an investigation, including intelligence gathering (Macoveciuc, *et al.*, 2019). Forensic gait examination is the combination of forensic image analysis and photographic comparison of trace and reference materials (Seckiner *et al.*, 2018; Seckiner *et al.*, 2019). The examination of such materials ultimately aims to evaluate the strength of evidence at source and activity levels, and this strength is evaluated based on the trace (obtained in the form of CCTV footage) and the comparison material (obtained from the person of interest) (*ibid*). However, Seckiner *et al.*, (2019) highlighted that the assessment of gait materials lacks scientific validity through failure for experts to use of the ACE-V (Analysis, Comparison, Evaluation, and Verification) protocol and logical inference models (such as models that assign likelihood ratios) for the evaluation step.

The ACE-V protocol needs to be implemented within this forensic examination of body and gait for scientific validity. ACE-V has become the general widespread stepwise approach to all forensic pattern evidence types that guides examiners throughout the examination process (Langenburg, 2012). The practitioner independently analyses and compares the trace and reference materials, to which a strength is assigned, that supports one of two propositions (prosecution or defence) with respect to the other (Langenburg, 2012). From here, a second practitioner completes the final stage of the examination process through verification, critically assessing the forensic findings of the first practitioner (*ibid*).

Validation is required prior to using the model in casework. The testing of a biometric system involves the algorithm and matching of scores for verification purposes and for determining similarity scores (ISO/IEC, 2006). Within the testing of a biometric system, there are three types of evaluations: [1] technology evaluation; [2] scenario evaluation; and [3] operational evaluation (*ibid*).

Technology evaluation involves the standardised testing of all algorithms to view their performance (for both the environment they are being used in and the collected population) (ISO/IEC, 2006). This stage involves the development of the model where the data testing is completed on data not previously used and the results should be repeatable (*ibid*). An example in relation to forensic gait analysis would be the data collection for the morphological and anthropometric examinations for forensic gait analysis, and a model is developed.

Scenario evaluation requires the testing on a complete system, in which the environment simulates that of a real-world target application of interest (ISO/IEC, 2006). Each tested system will involve a combination of various comparisons using the same population. Test results will be repeatable to the modelled scenario in controlled conditions (*ibid*). Regarding forensic gait analysis, it is about the development of an analytical method for forensic scenarios, such as using morphometric techniques for forensic evaluation. This consists of assigning a likelihood ratio (LR) to measure the probative value.

Operational evaluation focuses on the implementation of the method in an operational workflow and will not be repeatable. It includes the education of the practitioners, in order for them to be able to use the method, to integrate it in their practice, and to report about it and describe it in court. As there are unknown and undocumented differences between operational environments, ground truth[1] can be difficult to determine, particularly within the unsupervised and uncontrolled environments (ISO/IEC, 2021). Within forensic gait analysis, this will be the implementation of the method in casework, which involves the evaluation.

Forensic gait analysis for this paper will be restricted to considering the technology evaluation, whereas future papers will cover the scenario evaluation, on the basis of the results of the analytical method (morphometric assessment). The operational evaluation is not within the scope of this research. First the development of an evaluation framework is necessary. Once developed and tested, the validation and performance of the technology as well as the data can then be thoroughly examined within this framework.

Development of automated tools for forensic pattern evidence examination (such as gait analysis) requires the examination process to be formalised (Montani *et al.*, 2019). In this way, the technological developments can be integrated optimally with transparency (*ibid*). During this developmental stage, the reliability of the tools and processes need to be addressed and tested.

Reliability in forensic science has been defined in terms of a measure of validity[2], which includes the classification of error rates (false positives and false negatives) (Morrison *et al.*, 2010). Error is known as the variance between a measurement and the true value (or ground truth), which does not include practitioner error (Christensen *et al.*, 2013). The practitioner error comes in the form of repeatability and reproducibility, both of which are attributed to variations of precision. Further, as Roux *et al.*, (2022) highlighted, the reliability is reliant on the methodology and logical reasoning, which does not include the uncertainties linked to the

[1] Ground truth by Cardoso *et al.*, 2014 I is defined as 'the reference values used as standard for comparison purposes'.

[2] Validity is defined by Meuwly *et al.*, (2017) as 'range of conditions for which the method has been tested'.

evaluation of the trace itself. In a forensic context, repeatability, describes the variations within constant conditions within the same operator and/or instrument, whereas reproducibility is demonstrated with different operators or instances. They are known as the closeness of the agreement between the results of successive measurements of the same measure carried out under the same conditions. It must be noted that, to obtain the best possible measurements, accuracy and precision are both required. This paper focuses on the repeatability component, to attain the best measurement for the analysis. This repeatability data will be used in the logical framework.

Previous studies that explored validity, repeatability and reproducibility were completed by Birch *et al.*, (2019) and Birch *et al.*, (2021) using the Sheffield Features of Gait tool. Birch *et al.*, (2019; 2021) studies tested the contribution of 14 participants on 18 pieces of footage, to complete observational gait analysis in 3D (*ibid*). Within these studies, however, the examination was completed on an avatar – a model to characterise a human – but did not represent a true depiction of an individual from footage, or surveillance materials, highlighting a gap within the applicability in scenario or operational based evaluations. Traces captured by surveillance cameras in forensic settings are generally in poor camera conditions, and the avatars are not representative of those conditions. Therefore, although the studies by Birch *et al.,* (2019) and Birch *et al.,* (2021) were a preliminary study aimed to address a gap within the literature, the repetition of this study on surveillance footage would increase its applicability to determine the values of validity, repeatability, and reproducibility in an operational condition. This study will explore the repeatability component on the examination of both trace and reference footages.

In relation to surveillance footage, the European Network of Forensic Science Institutes (ENFSI) guidelines, state that following the examination from CCTV footage, the findings are usually evaluated against two mutually exclusive propositions: [1] the first is by the authorities;

[2] and the alternative by the defendant (ENFSI, 2015). The prosecution's proposition (denoted H_p) states the source of the trace material is the person of interest, or that the trace material originates from the same person as the reference material (e.g., the height of the person measured in the CCTV image and the height measurements of the person of interest describe the same person), whereas the defence's proposition (denoted H_d) states that the trace does not originate from the person of interest, or that the trace material and the reference material do not originate from the same person (e.g., the height of the person measured in the CCTV image and the height measurements of the person of interest describe two different people) (*ibid*). These propositions address the question of source, where the current approach is to convey a probative value expressed in terms of a LR. The LR is the probability of the observations, E, given the prosecution's proposition, H_p, divided by the probability of the observations, E, given the defence's proposition, H_d. The probabilities forming an LR may be based on empirically derived data (NIFS, 2017).

Studies have implemented likelihood ratio frameworks for various types of biometric trace materials. For example, a study by Champod and Meuwly (2000) developed an interpretation framework for speaker recognition, and studies by both Neumann *et al.*, (2012) and Meuwly and Veldhuis, (2012) developed interpretation frameworks for fingerprint recognition. To apply the LR approach in other biometric fields, it has been highlighted by Meuwly and Veldhuis, (2012) that a biometric LR-based system combines the use of biometric databases, technologies, and the likelihood ratio approach to probabilistically evaluate the evidential value of a trace and a reference material. The quality of the inference is dependent on the quantity and properties of the data that are used to assess the within and between-source variabilities (*ibid*). As stated by Meuwly and Veldhuis (2012), the classic 'forensic identification' disciplines rely primarily on personal probabilities for the assessment of the evidence. Likelihood ratio approaches are seen to be promising within forensic biometrics (Meuwly and

Veldhuis, 2012). Therefore, it is imperative for the implementation of likelihood ratios within the forensic gait analysis discipline as it develops.

This paper is Part 1 of the development and implementation of a forensic evaluation framework that uses a likelihood ratio. The ACE-V protocols (with focus on the ACE component for this study) were abided by, to improve the scientific approaches applicable to forensic gait analysis. The specific aims and objectives for the paper are as follows:

1. Present and describe statistically a wide set of possible gait and stance related features, the aim being to design a specific feature vector/set in each case, depending on the availability of the features on the questioned material.
2. Propose an empirical validation approach within the logical framework for the proposed and described features.

This paper comprises the development of an analytical model[3], showing distinctive features of body and gait in a forensic context, including the extraction of data, from data collection through to data analysis and data entry. Whilst establishing this, an Australian population database of 383 subjects for stance and 268 subjects for gait, including adult males and females of all ages and ancestries, was developed with the purpose of allowing both morphological and anthropometric (morphometric) assessment to examine features of the body during stance and gait. A manual Seckiner (2021) of the features that were extracted was developed to provide a step-by-step guide for the single observer for stance and gait analysis, thus allowing consistent results across the data, and in the future, will allow a study on reproducibility. The frequency of the features (and its variants), which will be used for assigning the denominator of the LR, was explored to highlight the features that were rarer within the sample population. Finally, the

[3] An analytical model in the context of this paper, is the use of a defined set of morphometric measurements that are robust to the forensic environment

repeatability results within a single observer, which will be used for assigning the numerator of the LR, were also completed and are presented within this paper. These two elements will allow the forensic practitioner to assign a robust strength of the evidence to body and gait observations.

2. MATERIALS AND METHODS

In general, surveillance footage provides poor quality materials, where some trace materials are rejected if features are not visible for examination, and other accepted if features are visible. This component can be seen as part of Figure 1, which demonstrates the overarching aim of the papers of implementing an interpretation framework for body and gait data. This paper will focus on the extraction of features, frequency of its variants, and repeatability studies. In this study, the majority of data collection (including photography and filming of volunteers) was completed in a room with filming area dimensions of 9.05m by 5.23m. There was a source of artificial lighting in the room with no windows to provide shifts in natural light – as the preferred conditions because over-lighting can cause loss of information within the images.

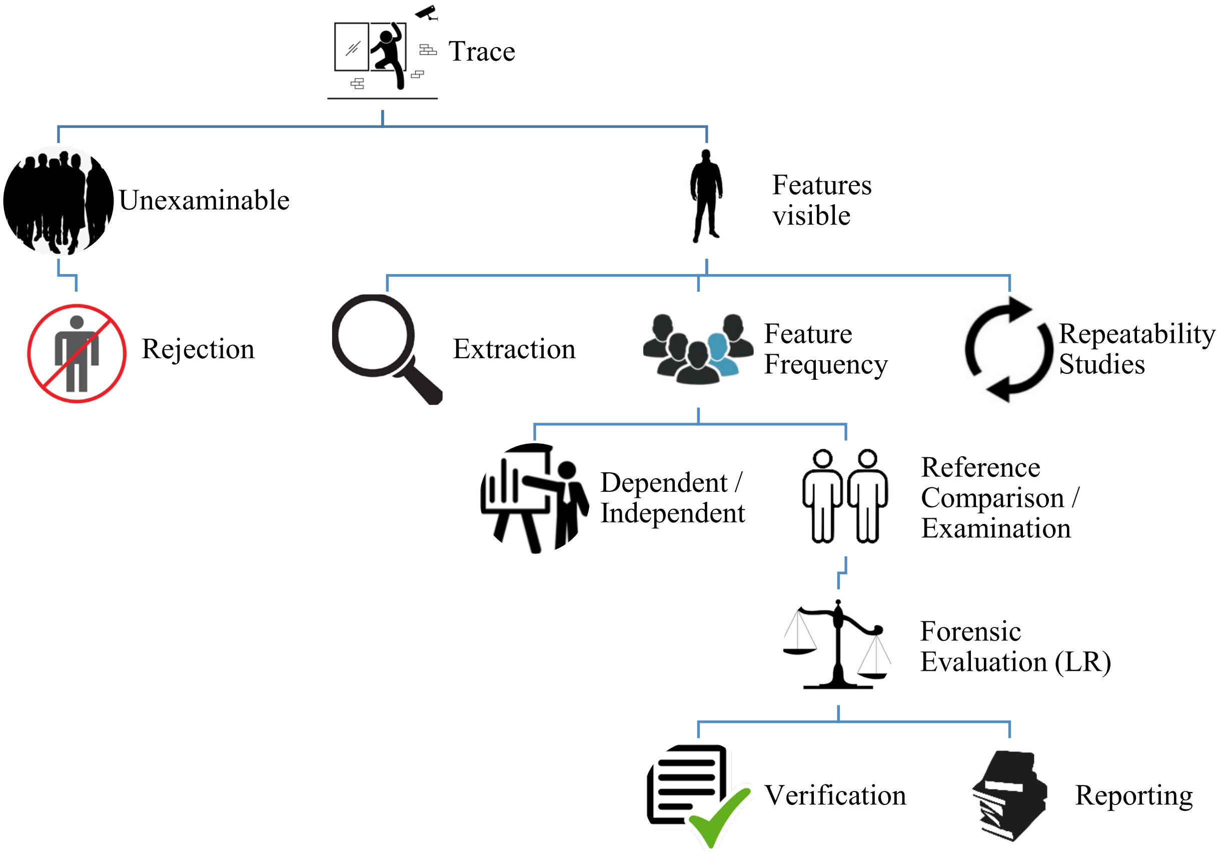

Figure 1. Overarching Interpretation Framework for Assigning the Strength of the Evidence

In attempt to eliminate perspective distortion, the camera height was adjusted to the umbilicus level of the subject (situated at approximately 1.1m from ground surface). Motion blurring of the distal appendicular anatomy was observed within gait subjects, but was of no consequence as it was minor. Finally, a combination of videos and high shutter speed photography was captured from each subject using a Canon 70D camera, where features can then be extracted.

As aforementioned, to allow the examination of subjects and the subsequent extraction of features, the initial step involved the recruitment of participants to form a database. Subjects were asked to complete a questionnaire that included collection of their sex, age, and existing pathologies that may have an influence on their stance and/or gait. The total number of subjects recruited and assessed were 383 for stance, of these, 268 volunteers' data also provided information suitable for the analysis of gait. Subjects within this study were instructed to wear their usual daily attire and walk at a comfortable pace. The strategy implemented, was the use of a fixed set of features that were evaluated across all subjects as they were able to be extracted from all subjects. For stance/body, males (182) and females (201) were relatively even in numbers. The age groups: 18 – 29 group (217) was the largest, followed by 50+ (113), and the smallest group was 30 – 49 (53). For ancestry, Caucasians (311) were the largest group, followed by Asian (54) and 'Other' (18). For gait, males (130) and females (138) were relatively even in numbers; whereas for ancestry, Caucasians (229) were the largest group, followed by Asian (30) and finally 'Other' (9). In age, group 18 – 29 (129) was the highest, followed by 50+ (99), and the smallest group was 30 – 49 (40) (see Table 1 in Seckiner *et al.*, (2022)).

Table 1 Features for Stance and Gait. The following table indicates morphometric variables produced within this research project. The highlighted features in blue are those of which further tests were completed and LR's were assigned, which will be featured in future papers. The definitions for each of these features are given in (see Tables 6 and 7 in Seckiner et al.*, (2022)).*

Stance - Morphological Feature	Gait - Morphological Feature	Gait Phase
1. Head Level	1. Lateral Placement of Upper Arm	Backward Arm Swing
2. Lateral Head Tilt	2. Lateral Placement of Forearm	
3. Projection of Head	3. Rotation of the Forearm	
4. Head Displacement	4. Level of Elbow Flexion	
5. Thoracic Projection	5. Rotation of Hand	
6. Abdominal Projection	6. Finger Flexion	
7. Upper Torso Shape	7. Lateral Placement of Upper Arm	

8. Torso Musculature	8. Lateral Placement of Forearm	Forward Arm Swing
9. Upper Thoracic Curvature	9. Rotation of the Forearm	
10. Thoracic Curvature	10. Level of Elbow Flexion	
11. Lumbar Curvature	11. Rotation of Hand	
12. Shoulder Level	12. Finger Flexion	
13. Position of Shoulder	13. Lateral Trunk Sway	Complete Cycle
14. Rotational Position Shoulder	14. Orientation of Lower Extremities	
15. Antero-Posterior Placement of Upper Arm	15. Head Level	Midstance
16. Lateral Placement of Upper Arm	16. Lateral Head Tilt	
17. Upper Arm Muscle Definition	17. Shoulder Level	
18. Antero-Posterior Placement of Forearm	18. Lateral Placement of Upper Arm	
19. Lateral Placement of Forearm	19. Lateral Placement of Forearm	
20. Lateral Rotation of the Forearm	20. Level of Elbow Flexion	
21. Lower Arm Muscle Definition	21. Rotation of Hand	
22. Antero-Posterior Placement of Hand	22. Finger Flexion	
23. Lateral Rotation of the Hand	23. Thoracic Projection	
24. Finger Flexion	24. Abdominal Projection	
25. Antero-Posterior Pelvic Tilt	25. Upper Thoracic Curvature	
26. Lateral Pelvic (Surface Anatomy) Asymmetry	26. Thoracic Curvature	
27. Gluteal Projection	27. Lumbar Curvature	
28. Gluteal Shape	28. Gluteal Shape	
29. Antero-Posterior Hip Deviation	32. Lateral Placement of Upper Leg	
30. Lateral Hip Deviation	32. Lateral Placement of Lower Leg	

31. Orientation of Lower Extremities	33. Knee Flexion	
32. Lateral Placement of Upper Leg	34. Placement of Feet	
33. Upper Leg Muscle Definition	35. Lateral Weight Bearing Feet	
34. Antero-Posterior Knee Joint Position	36. Lateral Placement of Upper Leg	Swing
35. Position/Orientation of the Knee Joint	37. Lateral Placement of Lower Leg	
36. Patellar Level	38. Placement of Feet	
37. Level of Infrapatellar Folds	39. Somatotype	Full Body
38. Lateral Placement of Lower Leg		
39. Lower Leg Muscle Definition		
40. Antero-Posterior Ankle Deviation		
41. Lateral Ankle Deviation		
42. Placement of Feet		
43. Lateral Weight Bearing of the Feet		
44. Somatotype		

2.1 Data Processing

When captured randomly, people recorded on CCTV footage are not generally walking perpendicular or directly parallel to the camera, but rather in random directions, resulting in quarter views of the person being recorded. However, for the purpose of this study, subjects were viewed in full body-height from four directions (anterior, posterior, right profile, and left profile). Assessing varying ages, ancestries, and both sexes was important to view any possible dependencies; therefore, subjects were recruited from as wide a demographic as possible. Shoes were identical in model for subjects (unisex shoes) to reduce variances in footwear (joggers, boots, thongs etc.) that might be introduced. Following the recruitment and filming of volunteers, videos were cut into stills, cropped, resized and placed into templates within Photoshop. Although this may have potentially resulted in a loss of quality in the images, they were resized to allow consistency, particularly for gait footages, as individuals were walking to and from the camera, producing varying sizes. This allowed for the anthropometric assessment, where anatomical landmarks were determined, marked, and measured. Finally, a plumb line[4] was added onto the footage, where the correct and faulty alignments of a subject's body was assessed (Kendall *et al.*, 2005). In stance, a subject's feet placement are equidistant to that of the line of reference. Any deviations observed from the plumb line were categorised into 'slight', 'moderate', and 'marked', depending on the amount of deviation detected. As these categories are relative, to make them more repeatable and reproducible, they were quantified through measurements, angles, and alignment/deviation from the plumb line.

[4] A plumb line is a cord with a weighted plumb attached to provide a vertical line, dividing the body into two (coronal and/or sagittal) (Kendall *et al.*, 2005). A virtual plumb line was also added, which applied the same concepts to that with the weighted plumb, dividing the body into two.

2.2 Repeatability Studies and Assessment of Features

In this study, repeatability is meant to assess the level of single-observer repeatability of the measurements during their examination. This determines whether the features are kept within the pool for assessment, whether the refinement of the classification of features for improvement is required, or the elimination of features as a result of measurements that are unable to be repeated.

2.2.1 Anthropometric Measurements

Most anthropometric landmarks applied within stance and gait were adopted from various anthropometric studies which involve *in situ* measurements. As the limbs are constantly flexing and extending during locomotion, the selection of anthropometric landmarks was primarily joint related, thus permitting application of measurements to all phases of gait. Measurements were obtained during the mid-stance phase (specifically at feet adjacent) of gait (4 frames at anterior, posterior, left and right sides) as it is the closest to stance, and to apply dynamic (gait) measurements, the heel strike phase were assessed to determine the distance between the limbs during locomotion (leading to a total of 8 frames for the anthropometric assessment). As no ground truth data from the source can be collected from trace material, importance for this study was placed on consistency of measurements with low repeatability error. A total of 17 anthropometric measurements and 16 anthropometric landmarks were developed while subjects were in 'normal' position (Figure 2). For gait, 25 measurements with 20 anthropometric landmarks were developed (see Figures 1 - 3 and Tables 2 - 5 in Seckiner *et al.*, (2022)).

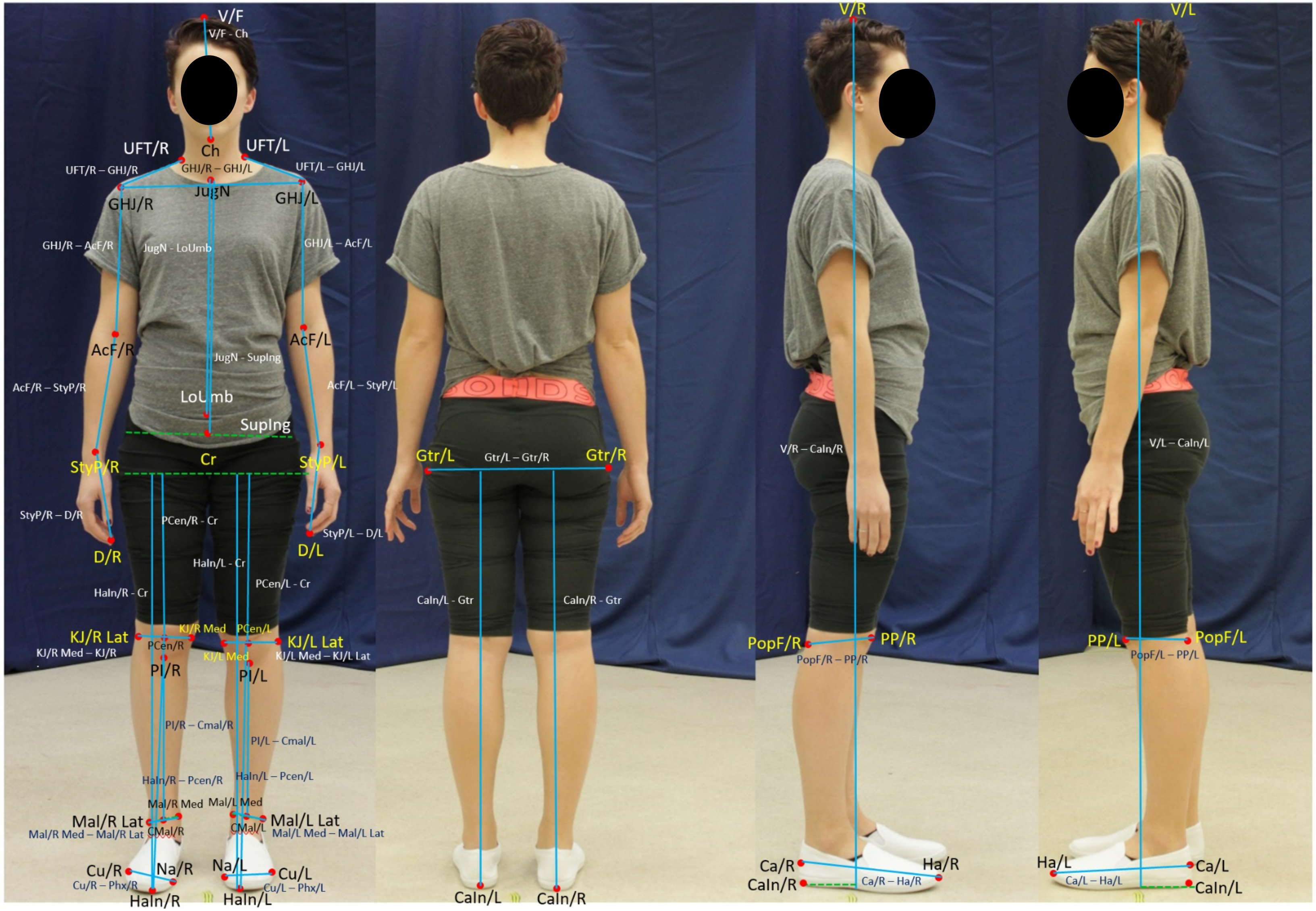

Figure 2 Unrefined Anthropometric Measurements Taken from all Views

The relaxed, natural position of the body during rest and walk, is referred to as 'normal' stance and gait respectively. Therefore, to maintain a realistic approach of comparing 'trace' and 'reference' materials, subjects assumed a 'normal' stance and gait. This stemmed from the unlikelihood of a person of interest presenting on CCTV footage and providing ideal evidence to allow precise anthropometric measurements. Standing was also assessed within this study, because although generally persons from CCTV footage are observed in motion, there are instances where they are seen to be standing, for example if they are waiting for the right moment to undertake any activity.

2.2.2 Morphological Features

The combination of both anthropometric measurements and morphological classifications allows a global approach to determine variability among subjects. As stance within this study is regarded as permanent and gait is transient, differing variables were developed for each. All phases and events within the gait cycle were assessed within the morphological features examination.

Together with the features observed previously (Seckiner, 2014), on high quality footage and images, 44 features were produced (72 for both limbs) with 142 subclassifications for stance, then further refined to 35 features for both limbs. From this, a total of 14 stance variables regarding the limbs were adopted, but further refined from Bradshaw, (2007), and Wright, (2012), to provide a full body analysis. For gait, 39 classifications (63 for both limbs) and 118 subclassifications were produced then further refined to 51 features for both limbs (see Tables 6 and 7 in Seckiner *et al.*, (2022)). All features and their simple definitions are listed in Table 1 for stance and gait respectively; features in blue were features that were used for the final examination. Further, in future papers, the features that are extracted from this pool are further refined as the 'questioned material' is uncontrolled.

2.3 Data Processing of Images and Footage

Preceding the analysis, footage was reviewed, cut into stills and then compiled and standardised as follows: Images were resized and compiled into separate templates, with a grid overlaid. Following this, variables were measured and classified by referring to the developed templates.

Subjects were photographed, images were proportionally resized via the 'transform to scale' tool on Photoshop. Then each image was then overlaid with a removable 1cm x 1cm grid, this maintained consistency between images, particularly relevant for gait footage, as individuals were walking to and from the camera, producing inconsistent sizes. Although this may have led to a slight precision loss of the resolution from the images due to resizing, the main focus was to facilitate maximum consistency within the examination. Photoshop was used for anthropometric measurements on the images (4 images for body, 8 for gait), whereas the raw cut still and footage was used for morphological analysis (see Figure 6 in Seckiner *et al.*, (2022)).

For anthropometric assessments, the measuring tool in Photoshop was applied and values obtained, transferred onto Excel spreadsheets and converted into indices[5] and feature-to-height ratios, to eliminate the issue of scale between measurements. The usage of indices disregards image size, therefore allowing comparison of proportions. Categorical values obtained from morphological features (ordinal data) was recorded on separate datasheets for stance and gait (i.e., 1,2,3), then later converted into dichotomous (nominal) data to view the variants of each feature (i.e., 0,1,0 or 1,0,0 etc.), in turn determining the frequency of the feature variant. For example, if an individual would be a category '2' out of three possible feature variations, their nominal data would be '0,1,0'.

[5] By dividing the anthropometric measurement by the total sum of all measurements, indices are attained, subsequently proportions will be compared instead of sizes: Indices = $\frac{AnthropometricMeasurement}{Total \sum of alltheMeasurements}$

2.4 Repeatability Studies

To determine single observer repeatability dand to obtain data for the intra-variability of the measurement, an intra-observer error study was performed using the Technical Error of Measurement (TEM%) for anthropometry (Arroyo *et al.*, 2010; Goto and Mascie-Taylor, 2007) and Cohen's Kappa for morphology (Viera and Garrett, 2005), thus permitting interpretation of features over a period of time for a single observer. Repeatability studies were conducted through the assessment of five male and five female subjects (total of 10) randomly selected from the database.

2.5 Frequency Values within the Given Population

Following the repeatability studies for 10 random subjects, the features that were repeatable were added to the feature pool, and data analysis was completed for all subjects recruited. Once the feature extraction was complete, the rarity of the features developed and analysed, were first vetted through heat maps. The use of heat maps allowed any visual discrepancy to be revealed as well as highlighting very rare features. Once features were investigated thoroughly, relative frequency values for each feature and its subsequent rarity within the given population was surveyed. To do this, the categorical data was converted to dichotomous, which facilitated the relative frequency values to be determined and tabulated.

3. RESULTS

3.1 Morphology Repeatability Assessment

For morphology, measuring the inter-observer presence of true agreement and comparing to the amount of agreement based on chance is known as Cohen's Kappa statistic (Table 2) (Viera and Garrett, 2005). A Kappa value of 1 shows a perfect agreement, whereas a Kappa value of 0 indicates an agreement dependent on chance (Viera and Garrett, 2005). Furthermore, a value below 0 demonstrates a less than chance agreement (*ibid*). If Kappa values are above 0.5, they are considered reliable, whereas variables that fall below the 0.5 threshold are considered unreliable as a result of lacking reproducibility and reliability. If such results are obtained where they are considered unreliable, further refinement of such variables are recommended. To assess the agreement of values, Minitab Statistical software was used.

Table 2 The Interpretation of the Kappa Values (Kurande et al., 2013). The above figure indicates the Kappa values and the corresponding level of agreement of the results that are produced from the error study.

Kappa Value	**Strength of Reliability**
<0.0	Poor
0.01- 0.20	Slight
0.21 – 0.40	Fair
0.41 – 0.60	Moderate
0.61 – 0.80	Substantial
0.81 – 1.00	Almost Perfect

Repeatability for both stance and gait were completed for Cohen's Kappa statistics separately, and the following results indicate that levels for both were acceptable; values that were under 0.5 were considered to have too much error (see Figures 7 - 8 in Seckiner *et al.*, (2022)). An unacceptably poor repeatability performance was observed in four features in gait and one in stance, which contained lower levels of repeatability performance compared to the remainder of features. For gait these comprised of, 'backward arm swing: rotation of left hand', 'forward arm swing: level of elbow flexion of left arm', 'midstance: placement of right foot' and 'swing: lateral placement of lower right leg', and for stance, 'antero-posterior placement of right hand'. The features that performed consistently reliably between both gait and stance were observed to be placement of the feet, finger flexion, and hand rotation. Performance varied between left and right sides of the body, for example, the lateral placement of the arm upon backward swing had 0.68 for the right arm and 0.85 for the left. The performance was significantly reduced for the rotation of the hand upon backwards swing of the arm at 0.6 for the right hand and 0.2 for the left. This was not consistent, however, as some features had a perfect score of 1 for both right and left sides, such as the level of elbow flexion and lateral placement of the upper arm.

3.2 Anthropometry Repeatability Assessment

For anthropometric features, the Technical Error of Measurement (TEM%) was used to determine the standard deviation amongst repeated measures, thus concluding the precision of the observer (Arroyo *et al.*, 2010; Goto and Mascie-Taylor, 2007). The TEM% calculation is an index to measure the repeatability error of the observer (Perini *et al.*, 2005). The International Society for Advancement of Kinanthropometry (ISAK) determined that a repeatability error value above 1.5% for intra-observer was too excessive, and further training to minimise this variability is required (Perini *et al.*, 2005). However, this is not applicable for forensic image capture, as the ISAK error levels are based on constrained and controlled conditions and therefore new guidelines are required for forensic scenarios.

Anthropometric features were developed for the purpose of accommodating static, dynamic, and angle measurements (see Figure 9 in Seckiner *et al.*, (2022)). All measurements for stance, gait static and gait dynamic repeatability studies (aside from 'gait: right foot width') fell beneath the threshold, indicating that there was minimal error, and these variables were carried forward to the extraction and examination phase. The feature that performed the best with the single observer repeatability was height (0.25% for stance, 0.2% gait). Between stance and gait, the next most consistent feature that was highly repeatable was the leg length measurements (0.31% and 0.32% for stance, 0.42% and 0.48% for gait). For the angle measurements, however, significant repeatability error was observed within most angle measurements, with the highest level of repeatability error measured at 34.4%. These measurements were eliminated from analysis as they were unfit for assessment due to the process being repeated (original features redefined) with unsuccessful results.

3.3 Frequency Data

The frequency data provides valuable information on whether features are rarer within the population or more common (see Tables 8 - 11 in Seckiner *et al.*, (2022)). Within this study, the stance anthropometry features, including a shorter forearm length (relative frequency 2.87%) and leg length (relative frequency 4.17%) relative to their proportions were observed to be rare. For gait anthropometry distance between the toes upon heel strike had a relative frequency of 8.20%, thus observed to be rarer within the feature pool of the given population. For stance morphology and gait morphology, features including medial placement of the feet (stance 1.04% and gait 4.10%), moderate bow leggedness (stance 3.65% and gait 4.10%), or moderate knock kneed (stance 4.96% and gait 2.23%) were observed to be rare within the given population. Common features observed were an increased finger flexion in both stance (relative frequency 66.8%) and gait (relative frequency 83.20%).

4. DISCUSSION

The availability of surveillance footage of a person of interest varies on a case-by-case basis, and therefore an extensive variable feature set is desirable. Hence, the development of a framework for future operational applicability is essential to allow the assignment of the strength of evidence to the examined trace evidence. This study presents the first steps to developing an evaluation framework for improving the scientific approaches applicable to forensic gait analysis. To assign LRs, there are components that need to be fulfilled after the features are extracted, the two most important being the repeatability values from the single observer conducting the examination, followed by the relative frequency of the feature variants within the given population.

The first step involved the collection and assessment of data, followed by the production of repeatability scores from a single observer. A fixed set of features were used, as they constitute the set available for most subjects within the given population recruited within this study. However, its potential use upon examination of CCTV footage, an extensive variable set of features can be extracted to provide the strength of evidence.

Further to subjects recorded in normal stance, a variety of reasons (ranging from visibility of features/variables, to attire, to poor repeatability) resulted in the exclusion of features. The features that performed consistently in terms of repeatability between both gait and stance were placement of the feet, finger flexion, and hand rotation, which may be indicative of the examiner's capabilities, in that it may be attributable to their understanding of extracting those features, or alternatively, it may have been suggestive of features that were the most simple to extract. The features that contained a perfect score for morphological repeatability studies were observed for both right and left sides of the body, including the level of elbow flexion and lateral placement of the upper arm. These may have performed so well due to the ease of extracting the feature itself, or potentially the step-by-step instructions were defined in an accurate and adequate manner to successfully repeat the examination. The use of such instructions may be beneficial for the examination process to contribute to the accurate assessment of individuals, although, further research on the repeatability and reproducibility aspects needs to be completed before recommendations can be made.

A poorer repeatability score was observed in four features in gait and one in stance. For gait, 'backward arm swing: rotation of left hand', 'forward arm swing: level of elbow flexion of left arm', 'midstance: placement of right foot', 'swing: lateral placement of lower right leg' and for stance, 'antero-posterior placement of right hand' contained lower levels of repeatability performance. This was an expected outcome, and it is hypothesised that it may be attributed to the minor variances of the appendicular anatomy upon swing of the arms and legs, during normal gait – leading to varying result, and subsequently altering the performance.

Further, variability in the repeatability was observed for anthropometry, such as the comparison of the left and right side of the feet in the stance anthropometry results (1.2% left foot width and 2.43% right foot width). This may have occurred due to the varying positions of the foot (in toeing or out toeing) when captured in 2-Dimension (2D), combined with potential incorrect anatomical landmark placed during the repeatability examination process by the single observer. One way this can be prevented in future might be to redefine those anatomical landmarks further to improve the precision of placing those markers digitally, or possibly, the examination of individuals in 3-Dimension (3D) of the reference materials. It must be noted that current surveillance technology only provides 2D materials, and therefore the comparison between 2D and 3D needs to be explored in future studies. The limitations reside more in the limited information that is captured within the CCTV footage, which will be explored further in future papers of this series. Inter-observer repeatability tests using the manual (Seckiner, 2021) should be undertaken to determine whether similar results are attained, as well as validity studies on both observational and motion capture techniques (tracking such as silhouette, contour, skeletal and so on). Applying these techniques to covert scenarios as well as actual case footage will allow both scenario evaluations to be completed and pave way for operational evaluation in future.

The second step was to evaluate the rarity of the features through heat maps (Seckiner, 2021) to determine if there were rare people and/or features respectively. The relative frequency values for each feature within the given population were then assessed to observe whether features were more rare or common. Those that were determined to be rare were the features that occurred at a low frequency within the subject pool. The highlighted features were those values under 50, equating to 13.05% of the given population for stance and 18.6% of the given population for gait. It appears that the features that were seen less frequently within the given population were those that were more marked, such as the moderate knock knees and moderate bow leggedness observed for stance morphology. Within gait, lateral rotation of the hand during backwards or forward arm swing were very rare, and extended fingers were only observed in one person during forward arm swing, indicating a high rarity, which may be attributed to increased speed. This is reinforced by Birch *et al.*, (2013), who highlighted that the primary feature that aided in the analysis of the study was the arm swing. However, studies such as Veres *et al.*, (2004) and Zhao *et al.*, (2006), favour the analysis of the lower body due to the high variability of the upper limbs, as a result of measurements being deemed unreliable, following unsuccessful tracking of the upper limbs. The next step for creating an LR model is defining a set of independent features that will be used to assign the LRs.

This study is not without limitations, the main one being, is that the quarter (or 'oblique') view of an individual requires research with the morphometric techniques applied and the feature sets examined. As the trace is rarely exactly parallel or perpendicular to the camera these variations and its evaluation are lacking within the forensic literature. As the trace can walk diagonal to the camera, the quarter views and three-quarter oblique views need to be assessed and any observed features extracted for analysis. If this were to occur, all angles of the body (quarter views - midpoint between a frontal and profile view and posterior and profile view) can be observed, features extracted, and analyses conducted, this will allow further robustness to the technique, as all views of the body can be assessed with the relevant features developed and extracted for analysis.

Combined with the quarter view assessment, improving the standardisation of the photographic conditions through a variety of mobile phone and various types of surveillance cameras and environments is also required. For this study, only one camera was available, however, different types of cameras will further highlight the limitations and requirements of the examination of gait and its forensic evaluation. The evaluation of the trace from varying cameras and their associated qualities (ranging from good to poor) is necessary to further the research, to approximate the redundancy of the footage for gait analysis.

As the ground truth of the measurements were not established (since you cannot obtain *in situ* measurements from a trace recorded on CCTV footage), it is possible that measurements of features did not fully correspond to that of the 'ground truth' of the participant's measurements. However, it is important to note that all measurements were completed by a single observer, thus allowing consistent measurements across all subjects, and allowing precision of the measurements (reinforced by the repeatability studies) taken by the single observer and potentially reducing the repeatability error.

The above-mentioned components are required constituents that will allow the assignment of the weight of the evidence for body/gait measurements. This in turn is a step towards overcoming the current challenge, which is the lack of a logical evaluation framework within the forensic gait analysis discipline. This paper forms the foundation for implementing such a logical framework within gait analysis, paving the way for further advancement.

5. CONCLUSIONS

The LR approach has been proposed and implemented within various forensic disciplines, including forensic biometrics, for example speaker recognition (Champod and Meuwly, 2000).

New and emerging technologies and validation studies have in turn improved the validity of the examination of the trace.

For this study, a total of 17 anthropometric features for body/stance, and 25 for gait were extracted and observed as consistent across the data pool. For morphology, 35 for stance and 51 for gait was extracted from the volunteer database. For a logical evaluation of the evidence, repeatability and frequency studies are required. The repeatability studies were from a single observer for stance (morphology and anthropometry) and gait (morphology and anthropometry [static, dynamic, angle]). The angle measurements in this study contained too many discrepancies and were therefore removed from the study. Further, performance varied between left and right sides of the body based on position, where for instance the left foot width for gait performed at a TEM% of 1.2%, whereas the right foot performed at a TEM% of 2.43%. The feature which had the lowest TEM% (and therefore the most repeatable scores) were height and leg measurements for anthropometry. For morphology, the highest Kappa scores (the most repeatable) for both stance and gait included the placement of the feet, finger flexion, and hand rotation.

The frequency studies consisted of observing the frequency of the feature variants. It was seen that the rarest feature variants were the lateral rotation of the hand during stance, which was only observed in one participant (relative frequency 0.26%), and the in-toeing of the feet, seen in four subjects (relative frequency 1.04%). Full extension of the fingers during gait were also seen to be rare, which was a feature extracted from one subject (relative frequency 0.37%).

To evaluate the observations and measurements from CCTV footage, the logical framework should be used to evaluate the strength of the evidence for a trace recorded on surveillance materials. This paper serves to provide the data necessary for applying such a logical framework for forensic body and gait analysis, which will be discussed and explored in future papers of this series. Within this paper, recruitment of participants for analysis was completed and a demographic obtained that will be useful for future studies, where the logical framework will not only be applied for an Australian population, but for other populations as well.

It is imperative for the development and implementation of a logical evaluation framework within the forensic disciplines to assign strength to the evidence for court processes. This paper serves as a preliminary step to contribute to the probabilistic evaluation for body and gait materials.

Ethics

Under ethics approval (UTS HREC Ref No. 2015000451), subjects were recruited and photographed/filmed. All subject data was anonymised, and all personal information was stored separate to number coded images (in a PIN access room, on a password protected computer and in a locked safe). All CCTV footage obtained was also stored under the same conditions.

Acknowledgements

The authors would like to acknowledge all the subjects who volunteered in this study and were photographed/filmed under ethics approval at the University of Technology Sydney. Without their generous contribution, this study would not be possible.

6. REFERENCES

Arroyo, M., Freire, M., Ansotegui, L., Rocandio, A.M. Intraobserver error associated with anthropometric measurements made by dietitians. Journal of Nutricion Hospitalaria, 2010. 25: p. 1053-1056.

Birch, I., Birch, M., Lall, J. The accuracy and validity of the Sheffield Features of Gait Tool. Journal of Science and Justice, 2021. 61(1), pp 72-78.

Birch, I., Nirenberg, M., Vernon OBE, W., Birch, M. 2020. Forensic Gait Analysis: Principles and Practice. CRC Press: Taylor and Francis Group CLC.

Birch, I., Birch, M., Rutler, L., Brown, S., Burgos, L.R., Otten, B., Wiedemeijer, M. The repeatability and reproducibility of the Sheffield Features of Gait Tool. Journal of Science and Justice, 2019. 59, 544-551.

Birch, I., Raymond, L., Christou, A., Fernando, M.A., Harrison, N., Paul, F. The Identification of Individuals by Observational Gait Analysis using Closed Circuit Television Footage. Science and Justice, 2013. 53: p. 339-342.

Cardoso, J.R., Pereira, L.M., Iversen, M.D., Ramos, A.L. What is the fold standard and what is ground truth? Evidence-based Orthodontics, 2014. 19(5): p. 27 – 30.

Champod, C., and Meuwly, D. The inference of identity in forensic speaker recognition. The Journal of Speech Communication, 2000. 31: p. 193 – 203.

Christensen, A.M., Crowder, C.M., Ousley, S.D., Houck, M.M. Error and its meaning in forensic science. Journal of Forensic Science, 2013. 59:1, p. 123-126.

(ENFSI) European Network of Forensic Science Institutes. Guideline for evaluative reporting in forensic science. Strengthening the evaluation of forensic results across Europe (STEOFRAE), 2015. Version 3.0. p. 98.

Goto, R., Mascie-Taylor, C.G.N. Precision of Measurement as a Component of Human Variation. Journal of Physiological Anthropology, 2007. 26: p. 253-256.

ISO/IEC. Information technology – Biometric performance testing and reporting. Part 1: Principles and framework, 2021. Reference number: ISO/IEC 19795-1:2021(E).

Kendall, F.P., McCreary, E.K., Provance, P.G., Rodgers, M.M., Romani, W.A. Muscles-Testing and Function with Posture and Pain, 2005. 5th Edition, Lippincott: Williams and Wilkins.

Langenburg, G.M. A Critical Analysis and Study of the ACE-V Process. University of Lausanne, 2012.

Macoveciuc, I., Rando C.J., Borrion, H. 2019. Forensic Gait Analysis and Recognition: Standards of Evidence Admissibility. Journal of Forensic Sciences. 64:5, pp 1294 – 1303.

Meuwly, D. and Veldhuis, R. Forensic biometrics: From two communities to one discipline, in 2012 BIOSIG - Proceedings of the International Conference of Biometrics Special Interest Group (BIOSIG). 2012: Darmstadt, Germany.

Meuwly, D., Ramos, D., Haraksim, R. A guideline for the validation of likelihood ratio methods used for forensic evidence evaluation. Forensic Science International, 2017. 276, pp 142-153.

Montani, I., Marquis, R., Anthonioz, N.E., Champod, C. Resolving differing expert opinions. Science and Justice, 2019. 59, pp 1-8.

Morrison, G.S., Thiruvaran, T., Epps, J. Estimating the Precision of the Likelihood-Ratio Output of a Forensic Voice Comparison System. The speaker and Language Recognition Workshop, 28 June – 1 July, 2010, Brno, Czech Republic.

Neumann, C., Evett, I.W., Skerrett, J. Quantifying the weight of evidence form a forensic fingerprint comparison: a new paradigm. Journal of the Royal Statistical Society, Series A, 2021. 175(2), pp. 371-415.

NIFS National Institute of Forensic Science Australia New Zealand. An introductory guide to Evaluative Reporting, 2017.

Perini, T.A., Oliveira, G.L., Ornellas, J.S., Oliveira, F.P. Technical Error of Measurement in Anthropometry. Brazilian Journal of Sports Medicine, 2005. 11: p. 86-90.

Roux, C., Bucht, R., Crispino, F., Forest, P.D., Lennard, C., Margot, P., Miranda, M., NicDaeid, N., Ribaux, O., Ross, A., Willis, S. The Sydney Declaration – Revisiting the essense of forensic science through its fundamental principles. Forensic Science International, 2022. 332: p. 1 – 10.

Seckiner, D., Meuwly, D., Roux, C., Gittelson, S., Maynard, P., Mallett, X. Dataset of Distinctive Extracted Features for Forensic Evaluation of Body Stance and Gait. Data in Brief (submitted for publication).

Seckiner, D. The development and testing of a forensic interpretation framework for use on anthropometric and morphological data collected during stance and gait. University of Technology Sydney, 2021.

Seckiner, D., Mallett, X., Meuwly, D., Maynard, P., Roux, C. Forensic Gait Analysis – Morphometric Assessment from Surveillance Footage. Forensic Science International, 2019. 296: p. 57-66.

Seckiner, D., Mallett, X., Roux, C., Meuwly, D., Maynard, P., Forensic Image Analysis – CCTV distortion and artefacts. Forensic Science International, 2018. 285: p. 77-85.

Seckiner, D. Forensic Body Mapping: Morphometric Gait Analysis and Quantification of Associated CCTV Image Distortions. Honours Thesis. University of Technology Sydney, 2014

Veres, G.V., Gordon, L., Carter, J.N., Nixon, M.S. What image information is important in silhouette-based gait recognition? Journal of Computer Vision and Pattern Recognition, 2004: p. 776-782.

Viera, A.J., Garrett, J.M. Understanding Interobserver Agreement: The Kappa Statistic. Journal of Family Medicine, 2005. 37: p. 360-363.

Zhao, G., Liu, G., Li, H., Pietikainen, M. 3D Gait Recognition Using Multiple Cameras. Automatic Face and Gesture Recognition, 2006: p. 1-6.