The Value of Posture, Build and Dynamics in Gesture-Based User Authentication

Jonathan Wu, Prakash Ishwar, Janusz Konrad
Department of Electrical and Computer Engineering, Boston University
8 Saint Mary's Street, Boston, MA, 02215
[jonwu, pi, jkonrad]@bu.edu *

Abstract

User authentication based on biometrics such as fingerprint, iris, face, speech or gait has been around for many years. Recently, intentional user gestures have been shown to be a promising modality for user authentication. However, it is unclear how much of the performance can be attributed to pure biometric information that a user has no control over, such as individual limb lengths, and how much to the gesture dynamics, that a user can fully control. A related question is: How easy is it to copy these dynamics? In this paper, we propose a framework to decompose a gesture into three components: initial posture, limb proportions, and gesture dynamics. We then study the impact of each component and various component combinations on the performance of gesture-based user authentication using a dataset of 36 users performing 3 gestures of varying complexity. We also study spoof attacks using the same dataset and show, somewhat surprisingly, that amateurs are unable to copy gestures with sufficient accuracy so as to significantly degrade the overall authentication performance even when they are trained on users that they are closest to. While training certainly improves an attacker's ability to copy gesture dynamics, it seems that the unique limb proportions (which cannot be altered) and the initial posture (which amateurs attackers fail to pay attention to), more than make up for the loss due to compromised dynamics (which can always be renewed).

1. Introduction

Face, iris, voice and fingerprints contain natural human characteristics that are extremely inconvenient to change. For all practical purposes they are intrinsically non-renewable.[1] Clearly, a renewable form of biometric would be invaluable. If such a biometric were to be (partially) compromised, one could (partially) change it like a password.

In this paper, we consider *intentional* user gestures[2] as a form of biometric. Although user gestures inherently depend on body build which is not renewable, the voluntary dynamics involved in performing a gesture can be altered – simply pick a new gesture.

A gesture can be captured in many different ways, e.g., by means of accelerometers attached to user limbs, a video camera or an RGBD camera, such as the Kinect. The use of a Kinect camera is more convenient than using accelerometers (no need to attach sensors to the limbs) and more reliable than a video camera due to the extraction of depth information (using structured light or time of flight approach). Furthermore, the Kinect SDK produces skeletal joint coordinates that can be used directly as input data for an authentication algorithm.

In the context of human-computer interaction, the Kinect has seen extensive use in gesture recognition [6, 10, 14, 15, 18, 22, 23]. However, little work has been reported on gesture-based authentication using the Kinect [11, 20, 21]. Lai *et al.* [11] have proposed using empirical log-covariance matrices of features extracted from a sequence of body silhouettes computed from Kinect depth maps for user authentication. Wu *et al.* [20] proposed an alternative approach using dynamic time-warping (DTW) across the skeletal joint estimates obtained from the Kinect SDK. These two works have demonstrated the potential for a future use of gestures in authentication.

However, neither of these prior works nor, to the best of our knowledge, other works related to gait-based authentication [7, 12, 13, 19], have made an attempt to study and quantify which components of a gesture are the most informative for authentication and which are most resilient to spoofing. Is it the non-renewable information from a user's body build or is it the voluntary dynamics performed by a user? The aim of this study is to answer these questions by systematically decomposing a gesture into three funda-

*This work was supported by the National Science Foundation under award CNS-1228869. Additionally, we would like to acknowledge Luke Sorenson and Lucas Liang for their support in data gathering and tabulation.

[1] Unless, of course, they are combined with multiple factors or a key as in secure-biometrics.

[2] As opposed to unintentional user motion such as gait.

mental components: initial posture, limb proportions, and gesture dynamics. We also study the related issue of vulnerability to spoofing. If an attacker were able to capture a video recording of a user's gesture, would he/she be able to easily break-in by training to imitate the user's gesture? To what extent can the dynamics be copied? To what extent does the non-renewable component offer protection against such an attack?

2. Gesture Decomposition

We posit that a gesture contains three types of user information: **initial posture**, **limb proportions** (build), and **dynamics**. The first two types of information have no user *intent*. A posture is related to user habits, whereas limb proportions (body build) are inherent characteristics that cannot be easily changed. Contrary to this, dynamics have an intent – depending on the gesture, the dynamics change.

As we explain in the sequel, one or more of these three types of information can be individually isolated and suppressed using a spherical coordinate representation of limb vectors. The effect of the initial posture can be suppressed by initializing the limb orientations to a *standard initial posture* obtained by averaging the initial posture across all users. The effect of build can be suppressed by setting all limb proportions to *standard limb proportions* obtained by averaging the limb proportions across all users. Finally, the effect of dynamics, can be removed by only considering the first frame of a gesture. In the following sections, we detail these information suppression processes, and show how we manipulate this information to study the impact of components, in isolation or in combination, on user authentication.

2.1. Skeletal Representation of Gestures

We can describe a gesture as a sequence of a user's skeletal joints in rectangular coordinates across time. The advantages of using skeletons over silhouettes, depth-maps, or images are two-fold: (i) skeletal data is sparse yet informative and (ii) skeletal data is relatively insensitive to changes in clothing, personal effects, and lighting conditions. Conveniently, the Kinect SDK [1, 17] provides rectangular coordinates of 20 skeletal joints on the human body for each frame at 30 frames per second. These coordinates are extracted from each depth frame and correspond to the following locations: head, neck, spine, center hip, and left and right versions of the hand, wrist, elbow, shoulder, hip, knee, ankle and foot (Fig. 1). A skeleton's evolution in time can be represented as a sequence of features $\mathbf{f}^t$ as follows:

$$\mathbf{f}^t := [\mathbf{s}_1^t, \mathbf{s}_2^t, ..., \mathbf{s}_{20}^t] \quad t = 1, ..., T, \tag{1}$$

where $\mathbf{s}_i^t = (x_i^t, y_i^t, z_i^t) \in \mathbb{R}^3$ is a 1×3 row vector which denotes the $x - y - z$ coordinates of the i-th skeletal joint in frame number t and T is the total number of frames.

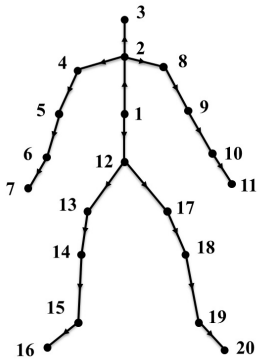


Figure 1. Example of a skeleton produced by the Kinect SDK.

2.2. Gesture Component Extraction

A skeleton can be represented in many equivalent ways. The simplest and most direct representation is as a tuple of 20 *joint* vectors in rectangular coordinates as in (1). An alternative representation which is more convenient for isolating the individual effects of initial posture, limb proportions, and dynamics, is as a tuple of 19 *limb* vectors together with one reference joint vector (see Fig. 1). In this context, it is useful to view the skeleton as a rooted tree with, for concreteness, the spine joint (joint number 1) as the root (or the reference joint) of the tree and the outgoing connected joints as the children. By knowing the coordinates of the root $\mathbf{s}_{spine}^t := \mathbf{s}_1^t$, and the outgoing edge vectors from the root, the entire skeleton can be reconstructed. If we denote the limb connecting joints i and j at time t by the limb vector $\mathbf{v}_{i,j}^t := \mathbf{s}_j^t - \mathbf{s}_i^t$, then

$$\mathbf{f}^t \equiv \{\mathbf{s}_{spine}^t, \mathbf{v}_{i,j}^t, i, j = 1, \ldots, 20 : i < j, (i,j) = \text{limb}\}. \tag{2}$$

To ensure that the initial position of the reference joint remains the same across different repetitions of a gesture, we subtract $\mathbf{s}_{spine}^1$ from all the joint vectors across all the frames, or equivalently, subtract $\mathbf{s}_{spine}^1$ from only the reference joint vector across all frames in the limb vector representation (2). By doing this we ensure that the spine joint in the first frame is always at the origin of the coordinate system.

Let $r_{i,j}^t, \theta_{i,j}^t$, and $\phi_{i,j}^t$ denote, respectively, the radius, azimuth angle, and elevation angle of the limb vector $\mathbf{v}_{i,j}^t$, i.e., the spherical coordinates of $\mathbf{v}_{i,j}^t$. Rectangular and spherical coordinates are information-equivalent representations of a vector and one can readily convert from one set of coordinates to another, i.e.,

$$\mathbf{v}_{i,j}^t = (x_i^t - x_j^t, y_i^t - y_j^t, z_i^t - z_j^t) \leftrightarrow (r_{i,j}^t, \theta_{i,j}^t, \phi_{i,j}^t).$$

The **initial posture** for each gesture can be thought of as the unintentional, habitual orientation of one's body parts.

This orientation can be described via the azimuth and elevation angles of all skeletal edges in the first frame, $\mathbf{f}^1$:

$$\{(\theta^1_{i,j}, \phi^1_{i,j}), i, j = 1, \ldots, 20 : i < j, (i,j) = \text{limb}\}.$$

Subsequent postures in the sequence pertain to the gesture's **dynamics**. The **limb proportions** describe the shape of a user's body (user build) regardless of the gesture that he/she performs. Ideally, limb lengths should not change across frames. However, the estimates of joint coordinates produced by the Kinect SDK are not perfect. We address this issue by computing the *average* length of each limb in a gesture across all frames and dividing it by the average length of the spine limb as follows:

$$\bar{r}_{i,j} = \frac{\sum_{t=1}^{T} r^t_{i,j}}{\sum_{t=1}^{T} r^t_{spine}} \tag{3}$$

where $r^t_{spine} = r^t_{1,2}$ is the spine limb length (Fig. 1) and $\bar{r}_{i,j}$ is the limb proportion for limb (i,j). Thus, for a given gesture sequence with 20 skeletal joints, there will be 19 limb proportions,

$$\bar{\mathbf{r}} := \{\bar{r}_{i,j}, i, j = 1, \ldots, 20 : i < j, (i,j) = \text{limb}\},$$

where $\bar{r}_{1,2} = 1$. From the above discussion it follows that we can represent any gesture as the combination of three sets of values: initial posture, $\{(\theta^1_{i,j}, \phi^1_{i,j}), \forall i, j : (i,j) = \text{limb}\}$, limb proportions $\bar{\mathbf{r}}$, and dynamics $\{(\theta^t_{i,j}, \phi^t_{i,j}), \forall i, j : (i,j) = \text{limb}, t = 2, \ldots T\}$.

2.3. Gesture Component Suppression

Our approach to study the individual and combined effects of initial posture, limb proportions and dynamics on user authentication performance is to first transform a given set of gestures to new ones (in rectangular coordinates) in which one or more gesture components (initial posture, limb proportions, dynamics) are either retained or suppressed and then evaluate the authentication performance on the transformed set of gestures.

The advantage of this approach, is that it allows us to use a single classifier and a single common feature space, namely the rectangular skeletal coordinates, for all the component combinations. If separate classifiers were developed for each combination of components (which live in different feature spaces), it would be unclear whether any performance differences are due to the components or/and the specific classifiers.

2.3.1 Suppressing Initial Posture

To remove the effects of user-specific initial posture, we introduce a limb-specific angular offset, $(\Delta\theta^{offset}_{i,j}, \Delta\phi^{offset}_{i,j})$, to every single frame. The goal of this is to orient the initial posture (1st frame), to a *standard* initial posture. As a result, this also re-orients subsequent frames in a sequence. The standard initial posture can be found by averaging the initial posture angles across all samples of all the users to yield $(\theta^{1,standard}_{i,j}, \phi^{1,standard}_{i,j})$. The angular offsets are then the angular differences between the standard initial posture and the user's initial posture:

$$(\Delta\theta^{offset}_{i,j}, \Delta\phi^{offset}_{i,j}) = (\theta^{1,standard}_{i,j} - \theta^1_{i,j}, \phi^{1,standard}_{i,j} - \phi^1_{i,j}) \tag{4}$$

The transformed gesture (in rectangular coordinates) with the initial posture suppressed (i.e., standardized) is then given by adding the angular offsets to the spherical coordinates of all frames and converting the result back to rectangular coordinates:

$$(\bar{r}_{i,j}, \theta^t_{i,j} + \Delta\theta^{offset}_{i,j}, \phi^t_{i,j} + \Delta\phi^{offset}_{i,j}) \rightarrow (\mathbf{v}^{t,noposture}_{i,j}). \tag{5}$$

2.3.2 Suppressing Limb Proportions (Build)

To remove the effects of limb proportions, we replace the radial distances (limb length proportions) with a set of *standard* limb proportions. Standard limb proportions are found by averaging the limb proportions across all samples of all users to obtain $\bar{\mathbf{r}}^{standard}$. The transformed gesture (in rectangular coordinates) with the limb proportions suppressed (i.e., standardized) is then given by replacing the radial distances (limb length proportions) with the standardized limb length proportions in all frames and converting the result back to rectangular coordinates:

$$(\bar{r}^{standard}_{i,j}, \theta^t_{i,j}, \phi^t_{i,j}) \rightarrow (\mathbf{v}^{t,nobuild}_{i,j}) \tag{6}$$

2.3.3 Suppressing Dynamics

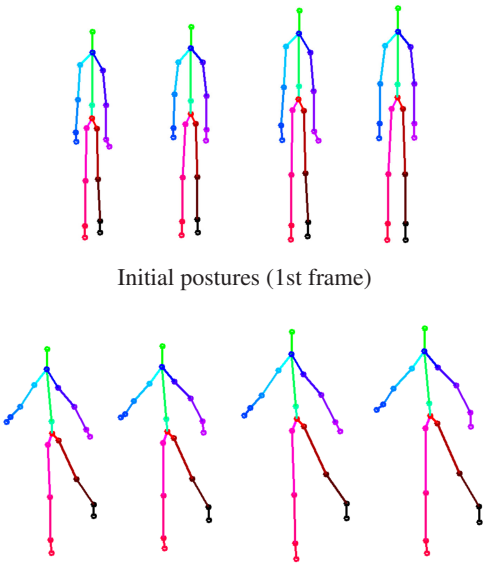
The suppression of dynamics is quite straightforward: just keep the first frame and discard the others.

2.3.4 Suppressing Component Combinations

In the last few sections, we described how to remove each of the three types of information. To remove more than one type of information at a time, we only need to combine the procedures of the information we want to remove. Table 1 describes various combinations of information that we evaluate. We do not evaluate the case where all components are suppressed, as all gesture samples would be identical. Using this methodology we show in Fig. 2 a few samples of so-constructed gestures including the standard posture and build sample.

Information Suppressed	Initial Posture	Limb Proportions	Dynamics
Nothing	✓	✓	✓
Dynamics	✓	✓	
Build	✓		✓
Posture		✓	✓
Dynamics+Build	✓		
Dynamics+Posture		✓	
Posture+Build			✓

Table 1. Various combinations of components we consider when reconstructing gesture sequences.



Initial postures (1st frame)

Intermediate posture.

Figure 2. Skeletons with various components suppressed. Top row suppressions (left to right): Dynamics, Dynamics+Posture, Dynamics+Build, Dynamics+Posture+Build (standard initial posture and standard build). Bottom row suppressions (left to right): Nothing, Posture, Build, Posture+Build.

3. DTW-based Skeletal Distance

We aim to understand the impact of different gesture components on authentication performance. In this work we do not aim to develop the best conceivable authentication algorithm for our problem by optimizing features, learning algorithms, and tuning parameters. Keeping this in mind, we focus on a simple yet time-tested authentication algorithm that is based on thresholding the nearest distance between a query and the enrolled gesture samples as explained in Section 6. We use dynamic time warping (DTW) to measure the distance between two gesture sequences of possibly different durations. DTW is a non-linear alignment algorithm that is relatively popular and has been extensively studied in the literature [3, 9, 16]. A modified version of this algorithm, as suited to our problem, is detailed below.

Let $\mathbf{F}_{g_1} = [\mathbf{f}_{g_1}^1, \cdots, \mathbf{f}_{g_1}^{T_1}]$ and $\mathbf{F}_{g_2} = [\mathbf{f}_{g_2}^1, \cdots, \mathbf{f}_{g_2}^{T_2}]$ be two feature matrices of skeletal features corresponding to gestures g_1 (T_1 frames long) and g_2 (T_2 frames long). A distance based on the cost of *aligning* $\mathbf{F}_{g_1}$ and $\mathbf{F}_{g_2}$ can be computed from a $T_1 \times T_2$ cost matrix. Let the cost matrix's i,j-th entry be the cost of aligning the skeletal feature in frame-i of gesture g_1 with the skeletal feature in frame-j of gesture g_2:

$$\text{cost}(\mathbf{f}_{g_1}^i, \mathbf{f}_{g_2}^j) = \sum_{p=1}^{20} ||\mathbf{s}_{p,g_1}^i - \mathbf{s}_{p,g_2}^j||_2.$$

An admissible alignment scheme is a path $\mathbf{P}$ through the cost matrix defined as follows

$$\mathbf{P} = \{(i_k, j_k), k = 1, \ldots, K : i_1 = j_1 = 1, i_K = T_1, \\ j_K = T_2, \forall k, i_{k+1} - i_k, j_{k+1} - j_k \in \{0, 1\}\}$$

where $\max(T_1, T_2) \leq K \leq T_1 + T_2$ is the path-length. The cost of a path is defined as follows:

$$\text{pathcost}(\mathbf{P}, \mathbf{F}_{g_1}, \mathbf{F}_{g_2}) = \sum_{(i_k, j_k) \in \mathbf{P}} \text{cost}(\mathbf{f}_{g_1}^{i_k}, \mathbf{f}_{g_2}^{j_k})$$

The path of interest is the one with the least cumulative cost. This path can be solved recursively using dynamic programming in quadratic time. The final cost is defined as follows:

$$d_{DTW}(\mathbf{F}_{g_1}, \mathbf{F}_{g_2}) = \min_{\mathbf{P}} \text{pathcost}(\mathbf{P}, \mathbf{F}_{g_1}, \mathbf{F}_{g_2})$$

4. Dataset

Datasets for gesture recognition and gesture authentication share some commonalities. The goal in recognition is to identify the gesture performed irrespective of the user, whereas the goal in authentication is to identify the user irrespective of the gesture. It might seem that a given dataset can be used interchangeably for both problems, e.g., analyzing user-authentication performance using a gesture recognition dataset. In reality, however, this is not the case since gesture datasets are typically ***gesture-centric*** meaning that they have a large gestures-per-user ratio (many gestures to classify, few users performing them) whereas studying authentication requires the opposite, namely a ***user-centric*** dataset which has a high users-per-gesture ratio, since one of the goals of authentication is to to ensure resilience against intentional and unintentional copying of gestures. With this guideline, our acquired dataset maintains a high users-per-gesture ratio of 12 (36 users, 3 gestures).

In total, about 1.8 hours of data were recorded, with each user averaging 3 minutes of data (each sample about 3 seconds long). Users were all college-affiliated (25 males, 11

females) mostly in the age range of 18-33 years. In order to strive for realistic intra-class variability and reduce pose bias, users were instructed to leave (for approximately one minute) and re-enter the recording area between gesture samples. This dataset is available online [2].

The 3 gestures, designed to be of increasing complexity, involved movement in both the upper and the lower body (Fig. 3):

- ***Left-right gesture***: user reaches right shoulder with left hand, and then reaches left shoulder with right hand,
- ***Double-handed arch gesture***: user draws an arch from left to right with both hands,
- ***Balancing gesture***: user first raises right arm forward while pulling left arm back, then balances by forward sweeping left leg while simultaneously tucking left arm in and bringing right arm to rest.

Gesture samples were collected in two sessions that were separated by one week. In the first session, each user was instructed how to perform each gesture through a text and video prompt (a multi-modal instruction scheme). In the literature, a multi-modal instruction scheme is known to improve gesture reproducibility over a single-modal instruction scheme (e.g. text or video only) [4]. After instruction, users performed each gesture 10 times.

In order to facilitate our gesture spoofing study (Section 5), each of the 3 gestures of each user was matched to an *attack target* after the first session. Attack targets were found by comparing the "centroid" samples of each user. If $\mathcal{A} := \{\mathbf{S}_1, \ldots, \mathbf{S}_m\}$ denotes a user's first-session samples (feature matrices) of a given gesture, then we define the user's "centroid" sample for that gesture as follows:

$$\mathbf{S}_{centroid} = \arg\min_{\mathbf{S}\in\mathcal{A}} \sum_{i=1}^{m} d_{DTW}(\mathbf{S}, \mathbf{S}_i). \qquad (7)$$

A user's attack target (in the second session), for a given gesture, is the owner of the closest centroid sample (nearest-neighbor) for that gesture. The aim of matching attackers to their "easiest victims" is threefold: 1) all participants can serve as attackers in the study 2) no participant is asked to attack more than one user which balances the burden across all participants, and 3) the odds of users succeeding as attackers are improved, which somewhat compensates for the lack of experience and the limited practice-time available for an attack. Under this matching scheme, vulnerable users would get attacked more often than others (very distinct users never get attacked). Furthermore, attackers may end up attacking up-to three distinct users (one for each gesture) Most users had one attacker. The maximum number of attackers that a user had was seven. We describe these attacks in the following section.

We would like to note that both the data-capture and the subsequent processing for authentication are in real-time with a delay on the order of half a second on a modern laptop or PC. Although our dataset is somewhat small (only 36 users and 3 gestures), in total there are 20 samples per gesture (10 own and 10 attack) for each user. The accuracy margin of error-rate-calculations is therefore on the order of 1/360. It is our hope that our dataset, which is publicly available online, will serve as a small starting point for future research in gesture-based authentication.

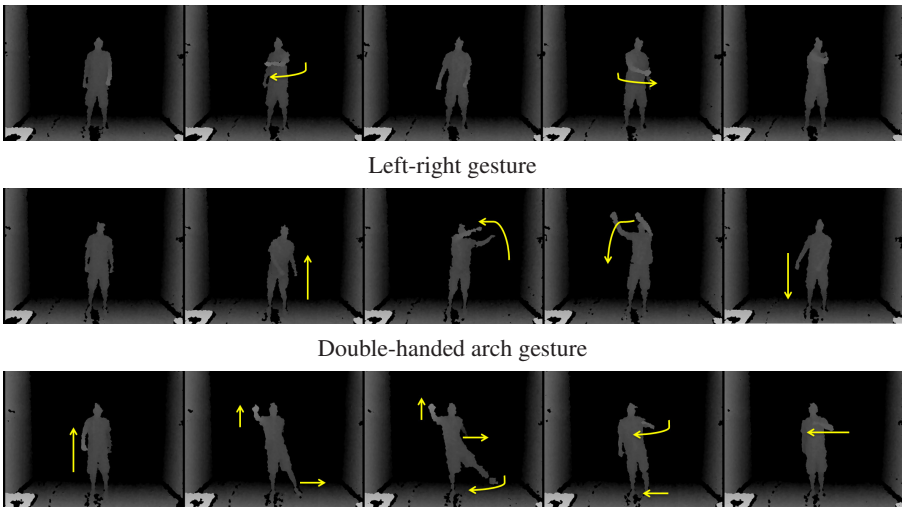
5. Spoofing Study

A natural question for any form of authentication is how easy is it to spoof? In the context of our study this question acquires an additional dimension. How much resilience to spoofing does each gesture component offer? We aimed to answer this question for gesture authentication by having amateur attackers attempt to mimic authorized users that they are closest to with some additional training.

Following the work of Gafurov *et al.*[5] we consider *minimal effort impersonation attacks* where our attackers have basic knowledge of the system (a gesture used for authentication), a limited time to study a target, and a set number of attacks. We limited the time to study a target to 1 minute and permitted 10 trials (10 recorded samples) by each attacker on a single target per gesture. For practice, attackers were allowed to view a looping video recording of their target's "centroid" sample from the first session. Attackers were given a chance to "mirror" the gesture by being shown streaming video of their practice. Once they were comfortable or a minute had elapsed, the spoof attempts were recorded.

6. Performance Evaluation

We consider entry control performance in the context of ***authentication*** [8]. In *authentication*, a user provides two pieces of information: his/her claimed identity and a biometric. If the biometric closely matches an enrolled sample of the given identity, the user is allowed entry. Otherwise, he/she is rejected. Two kinds of errors are considered in this case: false acceptance and false rejection. The false acceptance rate (FAR) is the rate at which *unauthorized* users are allowed entry. The false rejection rate (FRR) is the rate at which *authorized* users are denied entry. In any practical system, FAR and FRR will have trade-offs. One can find these trade-offs by applying various acceptance thresholds across the system. A common metric of performance is the equal error rate (EER) which occurs when FAR and FRR are equal. We compute EER scores for each authorized user separately (user-specific EER), and report the resulting average EER and standard deviation. We briefly recap this process below.



Left-right gesture

Double-handed arch gesture

Balancing gesture

Figure 3. Snapshots of the gestures each user performed in our dataset (Kinect depth shown).

Let $\mathcal{A}_i = \{\mathbf{S}_1, \ldots \mathbf{S}_m\}$ be a set containing m gesture samples from a single authorized user i. Let $\mathcal{U}_i$ be a set of gesture samples that do not come from authorized user i. The FRR is found by comparing samples in $\mathcal{A}_i$ amongst themselves (each sample in $\mathcal{A}_i$ is treated as a query sample), and the FAR is found by comparing samples in $\mathcal{U}_i$ to samples in $\mathcal{A}_i$ (each sample in $\mathcal{U}_i$ is treated as a query sample). We use a nearest-neighbor criterion $d_{NN}(\cdot,\cdot)$ to compare a single query sample $\mathbf{Q}$ to the authorized set $\mathcal{A}_i$.

$$d_{NN}(\mathbf{Q}, \mathcal{A}_i) \quad = \min_{\mathbf{S}\in\mathcal{A}_i} d_{DTW}(\mathbf{Q}, \mathbf{S}).$$

For a given threshold value θ, the FAR and FRR are calculated by:

$$FRR(\mathcal{A}_i, \theta) \;=\; \frac{\sum_{\mathbf{Q}\in\mathcal{A}_i} \mathbf{1}(d_{NN}(\mathbf{Q}, \mathcal{A}_i \backslash \{\mathbf{Q}\}) \geq \theta)}{|\mathcal{A}_i|}$$

$$FAR(\mathcal{A}_i, \mathcal{U}_i, \theta) \;=\; \frac{\sum_{\mathbf{Q}\in\mathcal{U}_i} \mathbf{1}(d_{NN}(\mathbf{Q}, \mathcal{A}_i) < \theta)}{|\mathcal{U}_i|}$$

where the indicator function $\mathbf{1}(condition)$ equals 1 if the '$condition$' is true and equals 0 otherwise. Note that for the FRR, we use leave-one-out cross validation such that each sample in the authorized set $\mathcal{A}_i$ is compared to the set $\mathcal{A}_i \backslash \{\mathbf{Q}\}$ with itself removed.

The EER for $(\mathcal{A}_i, \mathcal{U}_i)$ can be found by first computing these FAR and FRR values for different values of θ. Afterwards, the EER is determined by finding the boundary of the convex hull of these FAR-FRR pairs, and locating the point on the boundary of the convex hull where FAR equals FRR. This process is repeated for each authorized user who each has his/her own unique set $(\mathcal{A}_i, \mathcal{U}_i)$. All user EERs are then aggregated to yield the values shown in our results (mean).

Information Suppressed	Left-right	Double-handed arch	Balancing
Nothing	**1.97%**	**0.25%**	**0.68%**
Dynamics	3.83%	3.01%	2.12%
Build	2.09%	0.38%	1.20%
Posture	3.75%	0.61%	1.30%
Dynamics +Build	4.29%	4.88%	3.72%
Dynamics +Posture	8.22%	4.76%	4.39%
Posture +Build	6.91%	0.91%	3.22%

Table 2. User authentication EER (average of user-specific EERs) with zero effort attacks when various components are suppressed (please see Table 1 for component combinations). The best-performing EERs for each gesture are in bold-face text.

7. Results

7.1. Effects of Posture, Build, and Dynamics

We computed authentication EER for all 36 users from first session samples for each of the 3 gestures. This is equivalent to considering all 36 users as performing zero-effort attacks against one another in the worst case scenario when they all select the same gesture. The 7 combinations of gesture components that we described in Section 2.2 were applied to each of the 3 gestures, as shown in Table 2. If each user has a different gesture, the EER performance would only be better (lower) than the values shown here.

In terms of gestures, the "double-handed arch" performs

Gesture	Information Suppressed	Matched Zero-Effort EER	Matched Spoof EER	$\mathrm{EER}_{\mathrm{Spoof}} - \mathrm{EER}_{\mathrm{Zero-Effort}}$
Left-right	Nothing	2.78%	2.35%	-0.43
	Posture+Build	7.33%	10.28%	+2.95
Double-handed arch	Nothing	1.24%	1.13%	-0.11
	Posture+Build	3.78 %	4.22 %	+0.44
Balancing	Nothing	2.66%	2.06%	-0.60
	Posture+Build	5.60%	6.36%	+0.76

Table 3. EER shown for matched zero-effort attacks, and matched spoofing attacks. We show results for when no information is suppressed (Nothing), and when user-unique initial posture and build information are removed.

best, followed by "balancing", and then the "left-right" gesture. The "left-right" gesture should be expected to perform the worst as it is the least sophisticated (complex) of the three gestures. We originally expected the "balancing" gesture to perform the best due to its high complexity (it requires hand-leg coordination and body balancing). Surprisingly, it was only second-best. This can be explained, in retrospect, by the difficulty of reliably reproducing a complex gesture which has the effect of increasing the FRR and thereby the EER. So while complex gestures may be psychologically appealing as having higher discriminative power, they may actually be counterproductive because they can be difficult to reproduce.

In terms of gesture components, the suppression of dynamics has the single largest impact on the EER for every gesture followed by, somewhat surprisingly, the initial posture, and finally build. For example, for the "double-handed arch" gesture, the EER increases by 2.76% (from 0.25%) when the dynamics are suppressed, by 0.36% when the posture is suppressed, and by 0.13% when the build is suppressed (Table 2). Clearly dynamics play an important role. However, the role of posture and build is not insignificant. For instance, for the "left-right" gesture, the EER with posture and build retained but with dynamics suppressed is 3.83% which is lower than 6.91% when only dynamics are preserved. When all components are used, the EER is 1.97%. Similarly for the "balancing" gesture the EER with only posture and build (no dynamics) is 2.12% which is smaller than 3.22% when only dynamics are preserved. When all components are used, the EER is 0.68%. Thus, while dynamics is the most significant component of the three, the combination of all components results in a significant improvement.

7.2. Effects of Amateur Spoofing Attacks

In order to evaluate spoofing attacks, we considered EER in two contexts: matched zero-effort EER and matched spoofing EER. In order to compute the matched zero-effort EER, we only use samples from the first session and only consider the pool of authorized users who will be attacked in the second session (approximately 16 users attacked for each gesture). For each authorized user, we only consider unauthorized samples from users who will attack them in the second session. As we only use first-session samples, all these unauthorized samples are "matched" zero-effort attacks.

Following this train of thought, the matched spoofing EER is computed across the same authorized users with the only difference being unauthorized samples that are now second-session spoof attacks instead of first-session ones. These results are shown in Table 3.

Intuitively, one would expect the EER to increase after a matched spoofing attack relative to a matched zero-effort attack. Surprisingly, for our dataset, the EER performance actually slightly improves for all 3 gestures. This suggests that it is non-trivial for lay persons to effectively copy a user's gesture even when they are explicitly asked to attack their most vulnerable target and they have the opportunity to practice using a video-recording of their target performing his/her gesture.

Despite the unexpected decrease in EER of the matched spoofing attack relative to the matched zero-effort attack, interestingly, the EER based on dynamics alone, i.e., with posture and build suppressed actually increases consistently across all three gestures (see the last column of Table 3). This suggests that training does improve the ability of a lay user to copy the *dynamics*. Thus, body build and initial posture offer a limited but non-negligible level of protection against spoofing attacks.

8. Conclusions

The authentication power of a gesture does not lie solely in posture, body build, or dynamics. Each gesture component plays a non-trivial role. For zero-effort attacks with a single component suppressed, the suppression of dynamics causes the most significant degradation followed by posture and then build (Table 2). In the context of matched spoofing, although attackers may be able to improve their ability to replicate dynamics through training, posture and limb

proportions serve to make the gesture more secure. This effectively comes full circle – gestures do *indeed* serve as a renewable biometric. Should a breach occur and dynamics get compromised, the biometric can be partially changed by performing a different dynamic.

This study focused on the Kinect sensor, but the broad conclusions are not device-centric. Although the cost of the Kinect is not negligible today, ubiquitous depth sensor integration is expected in next-generation smartphones, PCs, and tablets. As a final point, this work focused on understanding the value of posture, build, and dynamics in gesture-based authentication and not so much on optimizing the system for yielding the best possible EERs. Yet, the EERs reported in this study are reasonably small. Still, we believe that there is much room for improvement, since all users in our study used the same gestures. In a realistic system, users would select different gestures which would only lead to lower EERs.

References

[1] KinectSDK. www.microsoft.com/en-us/kinectforwindows/, 2013. 2

[2] BodyLogin Dataset. http://vip.bu.edu/projects/hcis/body-login, 2014. 5

[3] H. Ding, G. Trajcevski, P. Scheuermann, X. Wang, and E. Keogh. Querying and mining of time series data: experimental comparison of representations and distance measures. *Proc. Very Large Data Bases Endow.*, 1(2):1542–1552, Aug. 2008. 4

[4] S. Fothergill, H. M. Mentis, P. Kohli, and S. Nowozin. Instructing people for training gestural interactive systems. In J. A. Konstan, E. H. Chi, and K. Höök, editors, *CHI*, pages 1737–1746. ACM, 2012. 5

[5] D. Gafurov, E. Snekkenes, and P. Bours. Spoof attacks on gait authentication system. *Information Forensics and Security, IEEE Transactions on*, 2(3):491–502, 2007. 5

[6] M. E. Hussein, M. Torki, M. A. Gowayyed, and M. El-Saban. Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, pages 2466–2472. AAAI Press, 2013. 1

[7] H. Iwama, M. Okumura, Y. Makihara, and Y. Yagi. The ou-isir gait database comprising the large population dataset and performance evaluation of gait recognition. *Information Forensics and Security, IEEE Transactions on*, 7(5):1511–1521, 2012. 1

[8] A. K. Jain, A. A. A. Ross, and K. Nandakumar. *Introduction to biometrics*. Springer, 2011. 5

[9] E. Keogh. Exact indexing of dynamic time warping. In *Proc. 28th International Conference on Very Large Data Bases*, pages 406–417, 2002. 4

[10] K. Lai, J. Konrad, and P. Ishwar. A gesture-driven computer interface using kinect. In *Image Analysis and Interpretation (SSIAI), 2012 IEEE Southwest Symposium on*, pages 185 – 188, April 2012. 1

[11] K. Lai, J. Konrad, and P. Ishwar. Towards gesture-based user authentication. In *Advanced Video and Signal-Based Surveillance (AVSS), 2012 IEEE Ninth International Conference on*, pages 282 –287, Sept. 2012. 1

[12] L. Lee and W. E. L. Grimson. Gait analysis for recognition and classification. In *Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on*, pages 148–155. IEEE, 2002. 1

[13] Y. Makihara, H. Mannami, A. Tsuji, M. A. Hossain, K. Sugiura, A. Mori, and Y. Yagi. The ou-isir gait database comprising the treadmill dataset. *IPSJ Trans. on Computer Vision and Applications*, 4(0):53–62, 2012. 1

[14] L. Miranda, T. Vieira, D. Martinez, T. Lewiner, A. W. Vieira, and M. F. Campos. Real-time gesture recognition from depth data through key poses learning and decision forests. In *Graphics, Patterns and Images (SIBGRAPI), 2012 25th SIBGRAPI Conference on*, pages 268–275. IEEE, 2012. 1

[15] M. Raptis, D. Kirovski, and H. Hoppe. Real-time classification of dance gestures from skeleton animation. In *Proc. 2011 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pages 147–156, 2011. 1

[16] C. Ratanamahatana and E. Keogh. Everything you know about dynamic time warping is wrong. In *Third Workshop on Mining Temporal and Sequential Data, in conjunction with the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2004. 4

[17] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1297–1304, 2011. 2

[18] J. Wang, Z. Liu, Y. Wu, and J. Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1290–1297, 2012. 1

[19] L. Wang, T. Tan, H. Ning, and W. Hu. Silhouette analysis-based gait recognition for human identification. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 25(12):1505–1518, 2003. 1

[20] J. Wu, J. Konrad, and P. Ishwar. Dynamic time warping for gesture-based user identification and authentication with kinect. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 2371–2375, 2013. 1

[21] J. Wu, J. Konrad, and P. Ishwar. The value of multiple viewpoints in gesture-based user authentication. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 90–97, 2014. 1

[22] L. Xia, C.-C. Chen, and J. Aggarwal. View invariant human action recognition using histograms of 3d joints. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, pages 20–27. IEEE, 2012. 1

[23] X. Yang and Y. Tian. Eigenjoints-based action recognition using naive-bayes-nearest-neighbor. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, pages 14–19, 2012. 1