LEVERAGING SHAPE AND DEPTH IN USER AUTHENTICATION FROM IN-AIR HAND GESTURES

Jonathan Wu, James Christianson, Janusz Konrad, Prakash Ishwar

Department of Electrical and Computer Engineering, Boston University
8 Saint Mary’s Street, Boston, MA, 02215
[jonwu, jamesrc, jkonrad, pi]@bu.edu

ABSTRACT

Depth-sensors, such as the Kinect, have predominately been used as a gesture recognition device. Recent works, however, have proposed using these sensors for user authentication using biometric modalities such as: face, speech, gait and gesture. The last of these modalities – gestures, used in the context of full-body and hand-based gestures, is relatively new but has shown promising authentication performance. In this paper, we focus on hand-based gestures that are performed in-air. We present a novel approach to user authentication from such gestures by leveraging a temporal hierarchy of depth-aware silhouette covariances. Further, we investigate the usefulness of shape and depth information in this modality, as well as the importance of hand movement when performing a gesture. By exploiting both shape and depth information our method attains an average 1.92% Equal Error Rate (EER) on a dataset of 21 users across 4 predefined hand-gestures. Our method consistently outperforms related methods on this dataset.

Index Terms— User authentication, hand gestures, silhouette tunnels, depth sensors, biometrics

1. INTRODUCTION

Biometrics are a convenient alternative to traditional forms of access control such as passwords and pass-cards since they rely solely on user-specific traits. Unlike alphanumeric passwords, biometrics cannot be given or told to another person, and unlike pass-cards, are always “on-hand”. Biometrics encompass personal properties such as face, speech, iris, gait and gesture. In this paper, the last of these biometrics is investigated – gestures, specifically, using an intentional in-air hand motion “password” for authentication.

Gestures are an interesting biometric modality to investigate because they contain partially-renewable behavioral information unlike physiological biometrics such as face and fingerprint. Compared to other biometric modalities, a gesture biometric is easiest to change when compromised. This is because gestures have two types of information: hard-to-change hand shape and easily changed (“renewable”) hand motion. In comparison, changing one’s face or fingerprint would be very inconvenient. Changing a compromised gesture is simple – one simply selects a new motion much like a new password.

This paper focuses on hand-gestures that have been captured with a depth-sensor, in particular the Kinect v2 [1] due to its predecessor’s ubiquity. In our experiments, we focus on extracting gesture information purely from depth data – we do not use RGB information.

Although there has been extensive work in hand-gesture recognition with depth-sensors such as the Kinect [2, 3, 4, 5, 6], there has been little work in authentication. Perhaps the work most closely related is that of Aumi and Kratz [7] who propose using dynamic time warping (DTW) across six 3-D fingertip and palm coordinates (coarse hand pose-estimation) from a depth-image for hand-gesture authentication. Similar “signature”-type gestures have also been proposed using an accelerometer, gyroscope and touchscreen on a mobile phone.

Compass gesture (flat translation)

Piano gesture (subtle finger movements)

Push gesture (change in distance to sensor)

Flipping Fist gesture (occlusions in fingers)

Fig. 1. Kinect v2 depth images of the 4 gestures used for authentication. For visualization, images have been cropped, and only show the lower 4-bits of the 16-bit depth image.

This work is supported by the NSF under award CNS-1228869 and by Co-op Agreement No. EEC-0812056 from the ERCP, and by BU UROP.
Many of these methods only use a single-coordinate in space and propose elastic matching algorithms (much like DTW) for authentication. As a result, physiological shape information of the hand is lost in all these approaches. Although not the focus of this paper, authentication with American Sign Language (ASL) using RGB frames (one frame per signed letter) has been investigated using features such as color histogram, DCT, and entropy [12, 13]. Full-body gestures have also been investigated using body-silhouette and skeletal features [14, 15]. In such a scenario, however, the fine finger resolution of the hand is lost; the biometric information comes from the entire shape, build, and dynamics of the user’s body [16].

This paper proposes a new approach to in-air hand gesture authentication using a temporal hierarchy of depth-aware silhouette covariances. Unlike the other aforementioned approaches, we incorporate hand shape (lost in “signature”-type approaches), and explicitly use depth-information (additionally useful when hand pose-estimates are not reliable or available). We use the silhouette tunnel descriptors proposed for action recognition in [17], and adapt and extend it for hand-gesture authentication. We validate our results on a hand-gesture dataset collected across 21 users and 4 predefined hand gestures.

2. SILHOUETTE REPRESENTATION

A compact silhouette representation can be extracted from any gesture sequence consisting of depth frames.

First, a sequence of binary silhouettes of a hand-gesture is extracted by thresholding the difference of each depth-frame from a known depth-background. Afterwards, the largest two-dimensional, 8-connected component of each frame is taken to be the user’s hand silhouette tunnel.

An approach similar to the one proposed by Guo et al. [17] and Lai et al. [18] is used to extract silhouette tunnel features. Let \( n = 1, \ldots, N \) index all \( N \) pixels within the silhouette tunnel and let \( (x, y, t) \) denote the space-time coordinates of pixel number \( n \). The following 14-dimensional feature vector \( f^n \) is computed at each silhouette pixel which captures the shape, depth, and dynamics of a gesture:

\[
 f^n = f(x, y, t) := [x, y, t, z, d_E, d_W, d_N, d_S, \\
 d_{NE}, d_{SW}, d_{SE}, d_{NW}, d_{T+}, d_{T-}]^T
\]

where \( z \) is the depth value at \( (x, y, t) \), and \( d_{dir} \) denotes the Chebyshev distance between a pixel \( n \) and its nearest silhouette boundary pixel in direction \( dir \). The first 8 directions are in the \( x, y, \) spatial plane (4 cardinal directions and 4 inter-cardinal directions), and the last 2 are in the temporal direction (forward and backward in time). Further, let \( F = [f^1, f^2, \ldots, f^N] \) denote a \( 14 \times N \) matrix that is computed from any silhouette tunnel. We rescale our features by normalizing each row of \( F \) to range from 0 to 1.

3. COVARIANCE DESCRIPTOR

The aforementioned silhouette representation can be seen as a “bag of features” since each silhouette pixel has an associated \( 14 \times 1 \) feature vector. A \( 14 \times 14 \) empirical covariance matrix \( C \) of the collection of feature vectors can be used to provide a low-dimensional, second-order descriptor:

\[
 C := \frac{1}{N} \sum_{n=1}^{N} (f^n - \mu)(f^n - \mu)^T,
\]

where \( \mu \) is the empirical mean of feature vectors \( f^n \). Since \( C \) is a symmetric matrix, the upper-triangular portion of size \((14^2 + 14)/2 = 105\) can be used as an equivalent gesture descriptor. Further, a simple way to compare the distance between two descriptors is to compute the Euclidean distance (Frobenius norm) between them. We denote this by \( d_{U.tri.Eucl} \).

4. ADDING A TEMPORAL HIERARCHY

One key property of the covariance descriptor is that pixel ordering does not matter. Even if the order of pixels were scrambled, the covariance matrix would be unaffected.

In order to incorporate temporal ordering information in a gesture, Hussein et al. [19] suggested using a hierarchy of covariance descriptors across temporal partitions at various scales. In this paper, we apply this idea with 3 temporal levels. At level \( i \), \( 2^{i-1} \) equal-length, non-overlapping temporal partitions are computed across the entire sequence. For example, at the 3rd level in the hierarchy there would be 4 equal-length partitions each of length \( N/4 \). (temporal ranges: \( 1 \) to \( \lfloor N/4 \rfloor \), \( \lfloor N/4 \rfloor + 1 \) to \( \lfloor N/2 \rfloor \), \( \lfloor N/2 \rfloor + 1 \) to \( \lfloor 3N/4 \rfloor \), and \( \lfloor 3N/4 \rfloor + 1 \) to \( N \)). The covariance matrix corresponding to a specific temporal range is computed using the feature vectors of only those pixels whose time index \( t \) falls within the specified temporal range. The computation of these covariance matrices can be sped up using the method of integral signals [19, 20].

The upper triangular portion of each covariance matrix computed from this hierarchy is concatenated together to yield one long gesture descriptor. With 3 levels, this yields 7 covariance matrices, giving a final descriptor length of \( 105 \times 7 = 735 \). We use the Frobenius norm between two descriptors of this type as the distance metric.

5. ADDING HAND MORPHOLOGY

Additional biometric information may be found by leveraging hand morphology (such as the thenar eminence in Fig. 2). We investigated whether silhouette parts (perhaps pertaining to the hand’s morphology) could be used to improve authentication performance. To achieve this goal, we applied a crude segmentation algorithm, based on depth, to partition a hand silhouette into multiple (“sub”-silhouette) regions. Across a
sequence of frames, these mutually-exclusive regions form additional “sub”-silhouette tunnels. To generate these tunnels, we find multiple frame-dependent depth thresholds \( r \), as follows.

Consider the frame \( t \), where we desire to find \( K \) segmented silhouettes (we set \( K = 3 \)). Let \( (r_{t,1}, r_{t,2}, \ldots, r_{t,K+1}) \), be ordered depth-thresholds with the property that the range between any consecutive threshold pair \((r_{t,i}, r_{t,i+1})\) contains a fraction \( \frac{r}{2} \) of all the depth values in the given frame. Using these thresholds, the segmentation at frame \( t \) associated with the \( i \)-th threshold pair yields frame \( t \) in the \( i \)-th “sub”-silhouette tunnel. As these segmentations can be noisy, we remove connected components with less than 20 pixels.

We compute a covariance matrix for each of these “sub”-silhouette tunnels. For fusion, we average the Frobenius norms (of covariance matrix differences between the query and enrolled samples) across 4 tunnels: the 1 full-silhouette and the 3 i-corresponding “sub”-silhouette covariances. Further, each of these matrices can use a temporal hierarchy as described in the last section.

6. GESTURE DATASET

At first glance, datasets that are suitable for hand-gesture recognition might appear to be suitable for hand-gesture authentication. Unfortunately, this is not the case – recognition datasets tend to be gesture-centric, focusing on many gesture types collected across few users (or few samples per user e.g., MSRGesture3D [5] with 2-3 samples per user). Gesture authentication seeks an opposite type of dataset, a user-centric one, that focuses on many users instead of many gesture types.

Aiming to create such a user-centric dataset (datasets from related work were not publicly available), we collected hand gestures from 21 college-affiliated users consisting of 15 males, and 6 females. A Kinect v2, a time-of-flight depth-sensor, was used to acquire a 512x424 depth image of each gesture sample at 30 fps. Each gesture sample was recorded in near proximity to the sensor (approximately 50 cm) so as to maximize hand resolution. Each user was asked to perform 4 unique types of hand-gestures, each type designed to be a few seconds in duration. Ten samples of each gesture type were recorded, with users instructed to leave and re-enter the recording area to reduce arm- and hand-pose biases between samples. Further, users were instructed using text and video since it is known that using both improves gesture reproducibility compared to the instructions with only one (e.g., either text or video) [21]. All 4 gestures were performed with the right hand starting in a “rest” position: the hand extended downwards on top of a ceiling-facing Kinect sensor, with fingers spread comfortably apart. The orientation of our sensor was designed to mimic an authentication terminal, where typically only a single user is visible. These gestures are:

**Compass:** users trace the compass directions of North, East, South and West with an open hand with the restriction that after each compass direction has been reached, the user must return to the center position before tracing a subsequent direction; this gesture evaluates planar translation,

**Piano:** users use their fingers to “press” the keys of an imaginary keyboard – fingers are pressed one-by-one starting from the thumb and ending with the pinky; this gesture evaluates subtle fingertip movements,

**Push:** users pull the arm back, and push towards the sensor; this gesture evaluates depth translation,

**Flipping Fist:** users first flip the hand over and close it into a fist, and then flip the fist over and open it back to the starting hand pose in front of the sensor; this gesture evaluates the effect of occlusions and more sophisticated fingertip motion.

7. PERFORMANCE EVALUATION

In this paper, we focus on entry control performance in the context of user authentication [22]. In authentication, a user provides his/her claimed identity and a query sample (a biometric gesture). If the query sample is a close match to an enrolled sample of the claimed identity, the user is considered a match and allowed entry. Otherwise, the user is rejected. Authentication can have two types of errors: false acceptances and false rejections. The false acceptance rate (FAR) reflects the rate at which unauthorized users are allowed entry and is a measure of security. The false rejection rate (FRR) reflects the rate at which authorized users are denied entry and is a measure of convenience. Usually, these two rates will have performance trade-offs which can be found by applying various acceptance thresholds across the system. A frequently used metric is the equal error rate (EER), which is the value at which FAR and FRR are equal. EER is computed for each authorized user separately (with 21 user-specific EERs, each with their own acceptance threshold). We report a final resulting average EER and its standard deviation.

In more detail, let \( \mathcal{A}_i = \{S_1, \ldots, S_m\} \) be a set containing \( m \) gesture samples from a single authorized user \( i \). Let \( \mathcal{U}_i \) be a set of all gesture samples that do not come from authorized user \( i \) (i.e., the 20 other users). 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 \( Q \)). This is done using leave-one-out cross validation (LOOCV) such that each sample is compared to the set \( \mathcal{A}_i \setminus \{Q\} \), i.e., with the query itself removed. The FAR is found by comparing samples in \( \mathcal{U}_i \) to samples...
### Table 1

<table>
<thead>
<tr>
<th>Method</th>
<th>Compass</th>
<th>Piano</th>
<th>Gesture Used</th>
<th>Flipping Fist</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Baseline</td>
<td>2.97%±0.69%</td>
<td>4.32%±0.99%</td>
<td>5.94%±1.31%</td>
<td>3.85%±0.91%</td>
<td>4.27%±0.51%</td>
</tr>
<tr>
<td>2. Temporal Hierarchy</td>
<td>1.98%±0.57%</td>
<td>2.90%±0.82%</td>
<td>5.28%±1.14%</td>
<td>2.40%±0.73%</td>
<td>3.14%±0.43%</td>
</tr>
<tr>
<td>3. Additional Tunnels + Temporal Hierarchy</td>
<td>0.44%±0.16%</td>
<td>1.29%±0.46%</td>
<td>4.89%±0.95%</td>
<td>1.05%±0.54%</td>
<td>1.92%±0.35%</td>
</tr>
<tr>
<td>4. First Frame of Silhouette Tunnel</td>
<td>5.33%±1.04%</td>
<td>7.00%±1.30%</td>
<td>7.94%±1.37%</td>
<td>6.44%±1.27%</td>
<td>6.68%±0.62%</td>
</tr>
</tbody>
</table>

Table 1. Average of user-specific EER and the confidence interval for various methods and gestures. The best-performing EER for each gesture is in bold-face. Please refer to Section 8 for additional information about these methods.

Incorporating additional silhouette tunnels from depth-information and the temporal hierarchy representation yields the best result on average (method 3) with an 1.92% EER. Comparing this to our baseline with a 4.27% average EER, there is a 2.35% EER reduction.

The gestures in order of performance from best to worst are: Compass, Flipping Fist, Piano, then Push. The push gesture always performs the worst. We believe this is due to poorer hand segmentations that are prevalent in this gesture. We noticed a few users would pull the arm back too close to the body, resulting in silhouettes that sporadically include portions of the chest. Since only depth information is used in our approach, it is difficult to differentiate between the body parts. Using RGB information may be useful in this case since skin pigment should be easy enough to differentiate from clothing.

As expected, method 4, using only the first frame performs the worst out of all the methods. This enforces the notion that there exists a unique behavioral movement in each user’s hand gesture.

It is important to point out that the Flipping Fist gesture performs quite well (2nd best). This result indicates that occluded shapes of the hand (such as a fist where all five fingers are hidden) still contain useful biometric information. This suggests that hand gestures need not be limited to cases where all fingers are visible, and that in these cases, leveraging features based on shape and depth (as features based on fingertip locations are ill-defined) is useful for authentication.

### 8. RESULTS AND DISCUSSION

Table 1 shows the authentication EER and confidence interval for 4 methods. The first method is a baseline, which we use to highlight the improvements that can be gained from incorporating depth-information and leveraging a temporal hierarchy. In this method, we use a single silhouette tunnel and a 13 × 13 covariance matrix that does not use the depth value z from the feature vector in equation (1). Subsequent methods use the full 14 × 1 feature vector (with z). The second method, incorporates the temporal hierarchy representation as described in Section 4. The third method, incorporates the temporal hierarchy representation and additional silhouette tunnels (a total of 4) from Sections 4 and 5. The last method (fourth) shows the value of motion in a gesture. Since, all gestures start with the right-hand in a neutral position, using only the first frame will have no dynamics. Further, there is no time information in the first frame, as the features associated with \((t, d_{t-}, d_{t+})\) in the covariance matrix become irrelevant.
10. REFERENCES


