The Value of Multiple Viewpoints in Gesture-Based User Authentication

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

Abstract

Although traditionally used as a gesture recognition device, the Kinect has been recently leveraged for user entry control. In this context, a user admission decision is typically based on biometrics such as face, speech, gait and gestures. Despite being a relatively new biometric, gestures have been shown to be a promising authentication modality. These results have been achieved using a single Kinect camera. This paper aims to investigate the potential performance and robustness gains in gesture-based user authentication using multiple Kinects. We study the impact of multiple viewpoints on a dataset of 40 users that contains notable degradations from user memory and personal effects (multiple types of bags and outerwear). We found that two additional viewpoints can provide as much as 26–43% average relative improvement in the Equal Error Rate (EER) for user authentication, and as much as 16–68% average relative improvement in the Correct Classification Error (CCE) compared to using a single centered Kinect camera.

1. Introduction

The traditional model of biometrics in user authentication is plagued by the use of inherently nonrenewable information. For example, a person's face, iris, voice, or fingerprint is a natural human characteristic that cannot be easily changed. However, faces can be easily photographed in public, fingerprints are susceptible to being left on surfaces (e.g., Iphone 5S TouchID hack), while voices can be recorded. Once compromised, a biometric is no longer reliable for authentication, but changing it is a major inconvenience. Due to this, a renewable form of biometric would be invaluable. If such a biometric were to be (partially) compromised, one could (partially) change this biometric like a password.

In this paper, we consider a user gesture as a biometric of interest; although some gesture characteristics inherently depend on body build and thus are not renewable, the voluntary dynamics involved in performing a gesture can be altered – simply pick a new gesture. However, the appeal of gesture renewability alone is not sufficient for its wider adoption; an acceptable level of gesture-based authentication performance is necessary.

We focus on user gestures that have been captured with a Kinect camera due to its low-cost and well-supported video, depth, and audio sensors. In our study, we use skeletal joint coordinates, obtained from the Kinect SDK, as input data.

Since its introduction the Kinect has been used in a wide range of applications, from autonomous navigation (as a depth sensor for Roomba vacs and quadcopters) to human-computer interaction (as a skeleton sensor). In the latter context, the Kinect has been extensively used for gesture recognition [5, 7, 9, 12, 14, 15]. However, little work has been reported to date on gesture-based authentication using the Kinect. Lai *et al.* [8] have proposed using empirical log-covariance matrices across a sequence of body silhouettes, extracted from Kinect depth maps, for user authentication. Wu *et al.* [13] 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, both used a single Kinect camera to capture either depth maps or skeletal joints of a user facing the camera during enrollment and testing. A consistent user orientation captured from a single viewpoint is not easily enforceable in practice. Would an acquisition from multiple viewpoints help?

Our focus in this paper is not on mining for the best features and classifiers to maximize authentication performance. Instead, our goal is to systematically investigate the potential benefits of using data from multiple viewpoints *versus* a single viewpoint in gesture-based authentication. Without attempting to "re-invent the wheel", we simply adopt the current state-of-the-art features and classifiers that have proved successful in the context of recent work on gesture recognition and gesture-based authentication, and focus on the question of the utility of multiple viewpoints.

This work was supported by the National Science Foundation under award CNS-1228869.

Specifically, we use a popular covariance-based descriptor applied to skeletal joints captured by up to 4 Kinect cameras.

To the best of our knowledge this is the first work to explore gesture-based authentication using multiple viewpoints. We offer an insight into performance gains attributable to the use of multiple viewpoints as well as improved robustness in the presence of several real-world degradations (user memory and personal-effects such as backpacks/bags and outerwear) resulting from rigorous tests on a dataset of 40 users.

2. Feature extraction

In this section, we briefly review the process of extracting skeletal features using the Kinect.

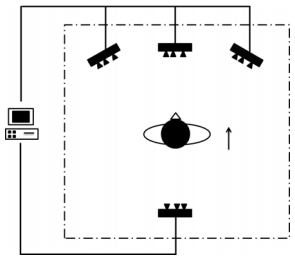


Figure 1. Experimental setup with four Kinect cameras. Three Kinects (left, center, and right) were placed in front of the user, at offset angles, and one was placed behind the user (back). Skeletal estimation was performed *independently* from each viewpoint.

2.1. Skeletons

Skeletal features are *independently* extracted from each of the Kinect depth cameras at 30 fps through the use of the Kinect SDK [1]. Frames captured by each Kinect are time-aligned through a host computer to be in sync with the central Kinect (Fig. 1). Technical details of the data acquisition setup are discussed in Section 5.

The Kinect SDK [1] performs skeletal pose estimation by using a variant of an algorithm developed by Shotton *et al.* [10]. Twenty $x-y-z$ joint coordinates are estimated from each Kinect camera (viewpoint) and each depth frame: head, neck, spine, center hip, and left-right versions of the hand, wrist, elbow, shoulder, hip, knee, ankle, and foot. An entire gesture for a single viewpoint can be represented as a sequence of skeletal-pose feature vectors:

$$\mathbf{f}_n^v := [x_{1,n}^v, y_{1,n}^v, z_{1,n}^v, ..., x_{20,n}^v, y_{20,n}^v, z_{20,n}^v]^T, \quad (1)$$

where $x_{i,n}^v, y_{i,n}^v, z_{i,n}^v, i = 1, ..., 20$ are the coordinates of the individual joints for viewpoint number v ($v = 1, ..., V$) and frame number n ($n = 1, ..., N$), with V being the total number of viewpoints (4 in this paper) and N being the total number of frames in the sequence. Two examples of joint positions are shown in Fig. 2. Let $\mathbf{F}_v = [\mathbf{f}_1^v, \mathbf{f}_2^v, \cdots, \mathbf{f}_N^v]$ denote a $d \times N$ feature matrix ($d = 60$) from the v-th Kinect.

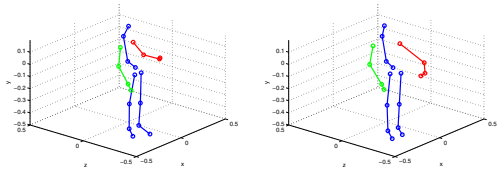


Figure 2. Skeletal joints of a user swinging his left arm. Red and green indicate the left and right arms, respectively, and blue indicates the center of the body and two legs.

2.2. Feature normalization

Due to the nature of gesture dynamics and body build, individual feature elements (e.g., coordinates $x_{i,n}^v$ and $y_{j,n}^v$) may have significantly different dynamic ranges. A feature element with a large amplitude would then influence an overall error metric more than a feature with a small amplitude. In order to avoid developing a complicated metric with unequal weights for individual elements, we adopt the approach of Hussein *et al.* [5] and normalize the matrix $\mathbf{F}_v$ along rows (time-wise) as follows:

$$\mathbf{F}_v^{norm}[i,j] = \frac{\mathbf{F}_v[i,j] - \min_k \mathbf{F}_v[i,k]}{\max_k \mathbf{F}_v[i,k] - \min_k \mathbf{F}_v[i,k]} \quad (2)$$

where $\mathbf{F}_v[i,j]$ denotes the value in the i-th row and j-th column of $\mathbf{F}_v$. The above normalization ensures that the values of all feature elements are contained within the dynamic range $[0, 1]$.

3. Authentication method

Similar to previous work on gesture recognition [7] and gesture-based authentication [8, 13], we base authentication decisions on thresholding the nearest-neighbor distance of a test sample to the (training) set of authorized samples. The distance metric used in [7, 8] is based on covariance descriptors of bags of features extracted from skeleton or silhouette sequences. Hussein *et al.* [5] also use a special covariance descriptor, Cov3DJ, constructed across a 3-level temporal hierarchy, for gesture recognition. However, unlike [7, 8] where the covariance descriptors are combined with nearest-neighbor classifiers equipped with a carefully constructed metric, Hussein *et al.* use their Cov3DJ descriptors as inputs to an SVM. Since the method in [5] achieves state-of-the-art performance in gesture recognition, we adopt Cov3DJ but combine it with nearest-neighbor classification with a simple linear distance metric as described below.

3.1. Covariance descriptor

A skeletal sequence can be also viewed as a "bag of features", where each frame (skeleton) is associated with

a $d \times 1$ feature vector. A $d \times d$ empirical covariance matrix C of a collection of feature vectors (normalized according to (2)) provides a low-dimensional, second-order representation of the entire feature-vector collection:

$$C := \frac{1}{N} \sum_{1}^{N} (\mathbf{f}_n^{norm} - \boldsymbol{\mu})(\mathbf{f}_n^{norm} - \boldsymbol{\mu})^T, \qquad (3)$$

where $\boldsymbol{\mu}$ is the empirical mean of normalized feature vectors $\mathbf{f}_n^{norm}$. For skeletons, $d = 60$ and N is the number of frames in the skeletal sequence. Since C is a symmetric matrix, its upper-triangular part of size $(d^2 + d)/2$ can be used as an independent gesture descriptor.

3.2. Temporal hierarchies

A key problem with the covariance descriptor is that the *ordering* of frames does not matter. If this order were to be scrambled, the covariance matrix would remain unchanged. In order to emphasize the importance of frame ordering in a gesture, Hussein *et al.* [5] suggested using a hierarchical computation of covariance descriptors across temporal windows at various scales. In this way, given any scrambling of the frames, the covariance matrices across smaller time windows would be different. In our work, we follow this idea for 3 temporal levels. At level i, 2^{i-1} equal-length, non-overlapping windows are computed across the entire sequence. For example, at the 3rd level in hierarchy there would be 4 equal-length windows each of length $N/4$. (temporal ranges: 1 to $\lfloor \frac{N}{4} \rfloor$, $\lfloor \frac{N}{4} \rfloor + 1$ to $\lfloor \frac{N}{2} \rfloor$, $\lfloor \frac{N}{2} \rfloor + 1$ to $\lfloor \frac{3N}{4} \rfloor$, and $\lfloor \frac{3N}{4} \rfloor + 1$ to N). All these covariance matrices can be computed quickly through the use of *integral signals* [5, 11].

For each covariance matrix that is computed from the temporal hierarchy, the upper triangular portion serves as our gesture descriptor, and all these descriptors are concatenated into one long gesture descriptor vector. For the case of 3 layers, there are 7 covariance matrices. Each upper triangular matrix is of length $(60^2 + 60)/2 = 1,830$, which concatenated together, yields a total length of $7 \times 1,830 = 12,810$. Thus, for a single gesture sequence from one Kinect camera using a 3-layer temporal hierarchy with Cov3DJs will yield a descriptor of length 12,810. We will refer to this final descriptor for view number v as $\mathbf{b}_v$.

3.3. Euclidean distance scores

A simplistic way to compare any two gesture descriptors, $\mathbf{b}_{1,v}$ and $\mathbf{b}_{2,v}$, is to compute their Euclidean distance. For each viewpoint v, this will yield a score. For each v the score is given by

$$d_v := ||\mathbf{b}_{1,v} - \mathbf{b}_{2,v}||_2. \qquad (4)$$

We can generate a set of V scores $S = \{d_1, d_2, ...d_V\}$ between any two gesture sequences.

4. Multiview score and feature fusion

Given multi-view data, we consider two simple fusion schemes: score fusion and feature fusion.

In score fusion, we consider each Kinect viewpoint to be an *independent* system. Each system computes a score for a given query gesture against a template from the enrollment database, and an aggregate score across all systems is used to determine an acceptance or rejection. In Section 3.3, we described how the individual viewpoint scores give rise to the score set S. To get a fused score, we apply one of the following operations on the set S: *min*, *mean*, *median*.

In feature fusion, we consider *concatenation*: combining features *before* a covariance matrix is computed (3). We simply concatenate the feature vectors $\mathbf{f}_n^{norm}$ (1) across all viewpoints. Whereas a single viewpoint would yield a feature vector of size 60×1, across V viewpoints this vector is now $60V \times 1$ and the final gesture descriptor is of length $7(60^2V^2 + 60V)/2$. Naturally, this yields only one score, so no subsequent score fusion step is necessary.

The final score obtained through aggregation of individual scores or through feature concatenation is used to evaluate authentication performance. Although more sophisticated fusion techniques could be applied, we believe the key insights into the benefits of using multiview data would remain unchanged.

5. Multi-view dataset

The problems of gesture recognition and gesture-based authentication are similar is some ways. Both require a group of users performing a set of gestures. However, in the former case the goal is to recognize the gesture regardless of the user, whereas in the latter case it is to recognize the user regardless of the gesture. Although it might seem that a given dataset can be used interchangeably for both tasks, e.g., analyzing user authentication performance using a gesture recognition dataset, this is not the case. Datasets for gesture recognition 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. With this guideline, our acquired dataset maintains a high users-per-gesture ratio of 20 (40 users, 2 gestures), under a variety of real-world recording conditions. This dataset has been made available at [2].

5.1. Setup for data acquisition

Each user gesture was captured from 4 Kinects. 3 Kinects were placed facing the user and 1 Kinect was directly behind (Fig. 1). Of the forward-facing Kinects, 2 were offset by about 35 degrees to the side, with 1 device directly in front. All devices were set up approximately 2

meters away from the user. Users were primarily facing the center camera for the duration of the performed gesture. All the Kinects were connected to a single PC to assure time synchronization. Captured frames were synced to the frame-rate of the center Kinect.

Each Kinect camera captured a 640x480 depth image and skeletal joint coordinates at 30 fps. All data was captured using the official Microsoft Kinect SDK [1]. Using

Figure 3. DC motors attached to each Kinect reduce structured light interference between multiple Kinects.

multiple Kinect cameras introduces structured light interference, which reduces the quality of the estimated depth maps. In order to reduce this interference, we applied an approach similar to the one developed by Butler *et al.* [3]; we attached Amico DC motors with different revolutions-per-minute (RPM) to each camera as shown in Fig. 3.

Figure 4. RGB and depth images from 4 Kinects of a user performing his own gesture. In this scenario, the user is wearing a bag (one of the degradations considered).

5.2. Acquisition methodology

Over a two week period, gesture samples from 40 different college-affiliated users (27 males, 13 females) primarily 18-33 years old were collected. Each subject was asked to perform 2 unique short gestures (approximately 3 seconds long), each with 20 samples. Following the work of Fothergill *et al.* [4] on how to best instruct users to perform gestures, instructions were given through text and video. To reduce pose bias between gestures, users were told to leave and re-enter the recording area between samples. Both gestures involved motion in the upper and lower body (Fig. 5):

- ***S gesture***: drawing an "S" shape with both hands,
- ***User-defined gesture***: user chooses his/her own gesture (no instruction).

5.3. Degradations

Four different types of gesture scenarios were considered: *no degradations*, *personal effects*, *user memory*, and *gesture reproducibility*. In the case of *personal effects*, users either wear or carry something during their gestures. Half of our users were told to wear heavier clothing, and the other half were told to carry some type of a bag. Users wore a variety of heavier clothing: sweatshirts, windbreakers, and jackets. They carried backpacks (either on a single shoulder or both), messenger bags, and purses.

The impact of *user memory* was tested by collecting samples a week after a user first performed a gesture. Users were first asked to perform the gesture without any video or text prompt. After a few samples were recorded, the user was shown a prompt and asked to perform the gesture again. These last samples measured *reproducibility*. Of the 20 samples recorded for each gesture, each of the described scenarios had 5 samples recorded. Table 1 shows a summary of the degradation scenarios that were used for each gesture.

6. Performance evaluation methodology

Entry control performance can be evaluated according to one of two entry-scenarios: ***authentication*** or ***identification*** [6]. In this section, we detail each of these scenarios and their associated performance metrics.

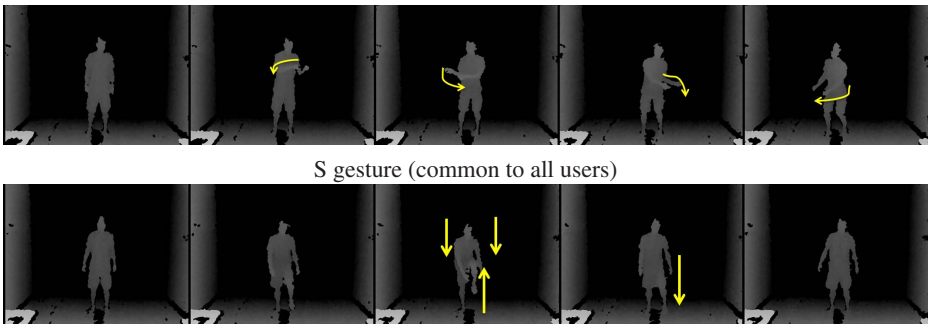
6.1. Authentication

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.

6.1.1 Computing authentication EER

Let $\mathcal{A} = \{s_1, \ldots s_m\}$ be a set containing m gesture samples (covariance/feature matrices) with known identities from the authorized users in the system, and let $\mathcal{Q}$ be a set of unidentified query gesture samples q. $\mathcal{A}$ and $\mathcal{Q}$ can be considered as sets of training and testing samples. Naturally, this assumes that $\mathcal{A} \cap \mathcal{Q} = \emptyset$. Let $\mathcal{Q}_{true}$ and $\mathcal{Q}_{false}$ denote the subsets of $\mathcal{Q}$ where the true identities of the samples are authorized and unauthorized, respectively. Thus, $\mathcal{Q} = \mathcal{Q}_{true} \cup \mathcal{Q}_{false}$. Let $d_{NN}(\cdot,\cdot)$ denote the nearest-neighbor score function between a query sample q and the set $\mathcal{A}$:

$$d_{NN}(q, \mathcal{A}) \quad = oper(\min_{s \in \mathcal{A}} d_{q,s,v}), \quad q \in \mathcal{Q}$$



S gesture (common to all users)

User-defined gesture knee-lift (will vary between users)

Figure 5. Snapshots of the gestures each user performed in our dataset (Kinect depth shown).

Gesture:	S gesture	User Defined
Session I:	1. Observe video and text description of gesture 2. No degradation: Perform gesture normally (5 times) 3. Personal effects: Wear a coat, or carry a bag 4. Perform gesture with personal effect (5 times)	1. Create custom gesture 2. No degradation: Perform gesture normally (5 times) 3. Personal effects: Wear a coat, or carry a bag 4. Perform gesture with personal effect (5 times)
Session II (a week after):	1. Memory: Perform gesture from memory (5 times) 2. Observe video and text description of gesture 3. Reproducibility: Perform gesture (5 times)	1. Memory: Perform gesture from memory (5 times) 2. Observe video of prior performance from Session I. 3. Reproducibility: Perform gesture (5 times)

Table 1. Details of recording procedure for the dataset. Users were instructed to "reset" initial position between gesture performances. Half of the users wore coats, and the other half carried bags.

where $d_{q,s,v}$ is the covariance Euclidean metric (4) between samples q and s for the view v. For single viewpoint, v is fixed to the viewpoint of interest (left, right, or center), and $oper(\cdot,\cdot)$ is the identity function. For fusion using feature-vector concatenation, we consider the "combined" viewpoint, where $d_{q,s,v}$ is instead drawn from concatenated feature vectors. Similar to single viewpoint, $oper(\cdot,\cdot)$ will also be the identity function. For multiple viewpoint score fusion, we consider all viewpoints $v = 1,...,V$, and set $oper(\cdot,\cdot)$ to be either the minimum, mean or median. Effectively, this is applying an operator across the nearest-neighbor individual viewpoint metrics as described in Sec. 4. For a given threshold value θ, the false acceptance and false rejection paired counts (FAC and FRC) can be calculated as follows:

$$FAC(\mathcal{A},\mathcal{Q},\theta) = \sum_{q\in\mathcal{Q}_{false}} \mathbf{1}(d_{NN}(q,\mathcal{A}) < \theta)$$

$$FRC(\mathcal{A},\mathcal{Q},\theta) = \sum_{q\in\mathcal{Q}_{true}} \mathbf{1}(d_{NN}(q,\mathcal{A}) \geq \theta)$$

where the indicator function $\mathbf{1}(condition)$ equals 1 if the $condition$ is true and equals 0 otherwise. After normalization of these counts by the number of samples in each set, we obtain the FAR and FRR:

$$FAR(\mathcal{A},\mathcal{Q},\theta) = \frac{FAC(\mathcal{A},\mathcal{Q},\theta)}{|\mathcal{Q}_{false}|}$$

$$FRR(\mathcal{A},\mathcal{Q},\theta) = \frac{FRC(\mathcal{A},\mathcal{Q},\theta)}{|\mathcal{Q}_{true}|}$$

In order to test the authentication performance of different methods using a single dataset, multiple pairs of training and testing sets $\mathcal{A}_i$ and $\mathcal{Q}_i$ can be created. Subsequently, the overall FAR and FRR across such pairs can be computed. If $\mathbf{A} := \cup_i \mathcal{A}_i$ and $\mathbf{Q} = \cup_i \mathcal{Q}_i$ then the overall FAR and FRR for the tuple $(\mathbf{A},\mathbf{Q})$ is defined as follows:

$$FAR(\mathbf{A},\mathbf{Q},\theta) = \frac{\sum_i FAC(\mathcal{A}_i,\mathcal{Q}_i,\theta)}{\sum_i |\mathcal{Q}_{i,false}|}$$

$$FRR(\mathbf{A},\mathbf{Q},\theta) = \frac{\sum_i FRC(\mathcal{A}_i,\mathcal{Q}_i,\theta)}{\sum_i |\mathcal{Q}_{i,true}|}$$

These overall FAR and FRR values are for a fixed threshold θ. The EER for $(\mathbf{A},\mathbf{Q})$ can be found by first computing these FAR and FRR values for different values of θ, then finding the boundary of the convex hull of these FAR-FRR pairs, and finally locating the point on the boundary of the convex hull where FAR equals FRR.

For authentication, each unique user defines multiple $\mathcal{A}_i$, $\mathcal{Q}_i$ pairs. For a given user, there exists a single $\mathbf{Q}_{i,false}$ which is the set of samples of all the other users. However, each user has multiple pairs of $(\mathbf{Q}_{i,true},\mathcal{A}_i)$'s. Each $\mathbf{Q}_{i,true}$ of a user is made up of a single sample of that user that is left out for testing with the remaining samples of that user forming the paired $\mathcal{A}_i$ set. Thus there are as many $\mathbf{Q}_{i,true}$'s (or equivalently $\mathcal{A}_i$'s) for a user as there are number of samples of that user. Described in another way, our authentication performance metrics are based on a leave-one-out cross validation (LOOCV) test where the user

sample that is left out is the authorized query sample which forms a $\mathbf{Q}_{i,true}$ set.

6.2. Identification

In *identification*, a user presents his/her biometric sample to a system which retrieves an enrolled identity through a one-to-many match. This is called the closed-set identification problem, i.e., classification under the assumption that the query user's identity is enrolled. The correct classification rate (CCR), and its complement error, CCE = 1-CCR, can be used to express accuracy. This value is also computed with LOOCV, where each user is labeled with the identity of his/her nearest-neighbor match.

7. Results and discussion

Various results for authentication and identification are shown in Tables 2 and 3, respectively. Six types of training and testing scenarios are considered. For simplicity, we denote each scenario by its test-set: "No degradations", "Personal effects", "User memory", "Reproducibility", "All of the above", and "Everything". The "No degradations" evaluation scenario trains and tests off gesture samples that have no degradations, while "Everything" trains and tests off *all* samples with and without degradations. In between these two scenarios are "Personal effects", "User Memory", "Reproducibility", and "All of the above", where the training is on normal samples that are free of degradations, but the testing is on data with either one type of degradation ("Personal effects", "User Memory", "Reproducibility") or all of them ("All of the above"). These scenarios are important for the following reason: they highlight the very realistic case in a practical system, where you cannot always capture all types of degradations in enrollment. Think of the similar case in photo identification where it is not practical to take photos of people with every type of glasses, sunglasses, haircuts, facial hair, etc. Effectively, we apply the same reasoning to gestures.

Due to degraded skeletal pose estimates from the rear-facing Kinect, we only consider the frontal 3 cameras in our multi-viewpoint evaluations. Under our methodology, including the 4-th camera does not improve results.

At a high level, looking at the row "Everything" in Table 2, we observe that S gestures are outperformed by user-defined gestures. This should come as no surprise, as it should be harder to distinguish between users when they *all* perform the same gesture. In consequence, it is not surprising that the presence of degradations introduced into the test set causes a more significant performance drop for S gestures than for user-defined gestures.

The introduction of any degradation causes a performance drop, although more for some degradations than others. In particular, time-related degradations (samples after a week), as seen in "User Memory" and "Reproducibility" rows, produce a larger drop than "Personal effects" degradation.

There are a few peculiarities in our results – we attempt to reason as to why. In S gestures, there is a decrease in performance from user memory to reproducibility, while in the user-defined case there is an increase. Upon closer inspection of our data, we noticed that a few users performed a mirror-image of their gestures during the user memory degradation tests. Gestures that were normally left to right, were performed right to left. This tended to happen more frequently in user-defined gestures. When users were shown the gesture again, user performance improved – which is what we see for the case of user-defined gestures. The reverse is true in S gestures, which we attribute to a slight difference in protocol. Whereas in the user-defined case the user's original performance from a week earlier was shown, in the S-gesture case a generic action recording (from an individual who was not part of the dataset) was shown. From our empirical results, it would seem that users tend to perform differently when instructed to replicate themselves or another person.

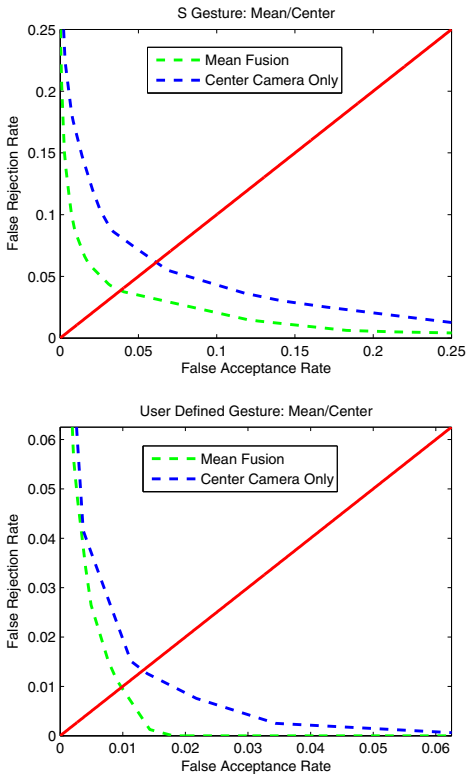


Figure 6. Convex hull of the ROC curves illustrating the EER improvement from using multiple views: mean fusion is compared to a single view (center). These results correspond to the EER values from the train-set/test-set "No degradations/All of the above", in Table 2 for both the S gesture and user-defined gesture.

It turns out that for multiview results, the center cam-

User Authentication Equal Error Rate (EER)

Gesture	Train-set/Test-set	Single-Viewpoint			Multi-Viewpoint Fusion				Multi minus Single	
	Camera/Method:	**Left**	**Right**	**Center**	**Min**	**Mean**	**Median**	**Concat.**	**BestM-BestS**	**Mean-Center**
S Gesture	No degradations/No degradations	**4.0%**	5.5%	9.0%	**5.0%**	5.5%	**5.0%**	6.5%	1.0%	-3.5%
	No degradations/Personal effects	**7.5%**	**7.5%**	9.0%	7.5%	7.5%	**7.0%**	8.0%	-0.5%	-1.5%
	No degradations/User Memory	**11.5%**	**11.5%**	14.0%	11.5%	**10.5%**	11.0%	11.0%	-1.0%	-3.5%
	No degradations/Reproducibility	**11.5%**	12.5%	13.0%	**10.5%**	**10.5%**	11.0%	11.0%	-1.0%	-2.5%
	No degradations/All of the above	**9.5%**	**9.5%**	12.0%	**8.5%**	9.0%	9.0%	9.5%	-1.0%	-3.0%
	Everything*/Everything*	4.1%	**4.0%**	6.1%	3.9%	3.9%	**3.8%**	4.1%	-0.3%	-2.3%
	Column Averages:	8.0%	8.4%	10.5%	7.8%	7.8%	7.8%	8.4%	-0.2%	-2.7%
User Defined	No degradations/No degradations	1.5%	2.0%	**1.0%**	1.0%	**0.5%**	**0.5%**	1.0%	-0.5%	-0.5%
	No degradations/Personal effects	**1.5%**	2.0%	2.0%	1.5%	**1.0%**	1.5%	2.0%	-0.5%	-1.0%
	No degradations/User Memory	2.5%	**1.5%**	2.5%	2.0%	**1.5%**	**1.5%**	2.0%	0.0%	-1.0%
	No degradations/Reproducibility	3.5%	**2.0%**	3.5%	3.0%	**2.0%**	2.5%	3.5%	0.0%	-1.5%
	No degradations/All of the above	2.5%	**2.0%**	2.5%	2.0%	**1.5%**	2.0%	2.0%	-0.5%	-1.0%
	Everything*/Everything*	1.5%	**1.0%**	1.4%	1.3%	**0.9%**	**0.9%**	1.4%	-0.1%	-0.5%
	Column Averages:	2.2%	1.8%	2.1%	1.8%	1.2%	1.5%	2.0%	-0.5%	-0.9%

Table 2. Equal error rate (EER) for authentication shown to 1 digit of precision. Smaller is better. FRR, and thereby EER, is rounded off to the nearest accuracy margin which is one over the number of positive samples in each train-set (this margin is $1/200 = 0.5\%$ for all scenarios except the "Everything*" scenario for which it is $1/800 = 0.125\%$). The best results for single-viewpoint and multi-viewpoint are shown in boldface. **BestM-BestS** denotes the error difference between the *best* performing result from multi-viewpoint vs. single-viewpoint (best multi minus best single). **Mean-Center**, similarly denotes the error difference between the mean multi-viewpoint scheme and the center camera (mean scheme minus center). "Everything*" contains samples with and without degradations.

Closed-set Identification (CCE)

Gesture	Train-set/Test-set	Single-Viewpoint			Multi-Viewpoint Fusion				Multi minus Single	
	Camera/Method:	**Left**	**Right**	**Center**	**Min**	**Mean**	**Median**	**Concat.**	**BestM-BestS**	**Mean-Center**
S Gesture	No degradations/No degradations	**1.5%**	3.0%	3.5%	**2.0%**	2.5%	2.5%	2.5%	0.5%	-1.0%
	No degradations/Personal effects	6.5%	**6.0%**	11.1%	7.0%	6.5%	7.0%	**6.0%**	0.0%	-4.5%
	No degradations/User Memory	19.5%	20.0%	**17.5%**	18.0%	15.5%	16.5%	**14.5%**	-3.0%	-2.0%
	No degradations/Reproducibility	28.5%	26.0%	**22.5%**	26.5%	21.5%	22.0%	**20.5%**	-2.0%	-1.0%
	No degradations/All of the above	14.0%	13.8%	**13.6%**	13.4%	11.5%	12.0%	**10.9%**	-2.8%	-2.1%
	Everything*/Everything*	1.3%	**1.1%**	1.3%	**0.5%**	0.8%	1.3%	0.6%	-0.6%	-0.5%
	Column Averages:	11.9%	11.7%	11.6%	11.2%	9.7%	10.2%	9.2%	-2.4%	-1.9%
User Defined	No degradations/No degradations	**0.0%**	0.5%	0.5%	**0.0%**	**0.0%**	**0.0%**	**0.0%**	0.0%	-0.5%
	No degradations/Personal effects	0.5%	**0.0%**	1.0%	**0.0%**	0.5%	1.0%	0.5%	0.0%	-0.5%
	No degradations/User Memory	**1.5%**	2.5%	3.5%	3.0%	1.5%	2.0%	**1.0%**	-0.5%	-2.0%
	No degradations/Reproducibility	**1.0%**	**1.0%**	1.5%	1.0%	**0.0%**	1.0%	0.5%	-1.0%	-1.5%
	No degradations/All of the above	**0.8%**	1.0%	1.6%	1.0%	**0.5%**	1.0%	0.5%	-0.3%	-1.1%
	Everything*/Everything*	**0.1%**	0.3%	0.1%	**0.0%**	0.1%	0.1%	0.1%	-0.1%	0.0%
	Column Averages:	0.7%	0.9%	1.4%	0.8%	0.4%	0.9%	0.4%	-0.2%	-0.9%

Table 3. Correct classification error (CCE) for closed-set identification shown to 1 digit of precision. All query samples are assumed to have been enrolled into the system beforehand. See the caption of Table 2 for explanations.

era is not always the best performing one across all test-set scenarios; the side cameras (left and right) consistently outperform the center camera. For the S gesture, the training sample that is closest to the test sample belongs to the center camera for only about 22% of the test samples ($\sim$32% match to the left and $\sim$46% to the right). For the user-defined gesture, about 39% of the test samples find their best match among the center training samples, about 28% with the left, and about 33% with the right. This can be explained as follows. If a part of the body is occluded during

a gesture in one viewpoint, another camera may be able to see the gesture more clearly without this occlusion. Inherently, this shows the value of multiview acquisition during enrollment and testing.

If we had to pick one multi-viewpoint fusion method out of all the methods we applied, we would pick the mean fusion scheme. This is because it performs the best across both gesture test-sets (mean fusion scores are bolded the most). Thus, in comparison to the single-viewpoint **authentication** setup which, traditionally, will only consist of a single *centered* camera, we find an average EER decrease of 2.7% (∼26% relative improvement) and 0.9% (∼43% relative improvement), for the S and user-defined gestures, respectively, due to multi-viewpoint mean fusion. Similarly, in comparison to the single-viewpoint **identification** setup, we find an average CCE decrease of 1.9% (∼16% relative improvement) and 0.9% (∼68% relative improvement), for the S and user-defined gestures, respectively, due to multi-viewpoint mean fusion. In every testing scenario that we compare to the center-viewpoint, multiple viewpoints always outperform – they are always more informative.

Even if we were to pick the best performing single-viewpoint camera and compare its results against the best performing fusion scheme, separately for each of the six training/testing scenarios, we still find an overall improvement in performance by using multiple viewpoints. Specifically, the average EER decreases by 0.2% (∼3% relative improvement) and 0.2% (∼33% relative improvement), and average CCE decreases by 2.4% (∼23% relative improvement) and 0.2% (∼32% relative improvement), respectively for the S and user-defined gestures.

Finally, a finer perspective of the benefit of multiple viewpoints can be obtained by examining the ROC curves for single-view and multi-view authentication for "No degradations/All of the above" scenario shown in Fig. 6.

8. Conclusions

This paper presented a look into multi-viewpoint gesture-based authentication and the benefits it gives in comparison to a single-viewpoint authentication system. To the best of our knowledge, this is the first such study. Based on the empirical results presented here, multi-viewpoints undeniably offer clear and significant benefits in terms of both performance and robustness against degradations, over the traditional single-viewpoint setup.

References

[1] KinectSDK. www.microsoft.com/en-us/kinectforwindows/, 2013. 2, 4

[2] BodyLogin Dataset. http://vip.bu.edu/projects/hcis/body-login, 2014. 3

[3] D. A. Butler, S. Izadi, O. Hilliges, D. Molyneaux, S. Hodges, and D. Kim. Shake'n'sense: reducing interference for overlapping structured light depth cameras. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems*, pages 1933–1936. ACM, 2012. 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. 4

[5] 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, 2, 3

[6] A. K. Jain, A. A. A. Ross, and K. Nandakumar. *Introduction to biometrics*. Springer, 2011. 4

[7] 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, 2

[8] 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, 2

[9] 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

[10] 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

[11] O. Tuzel, F. Porikli, and P. Meer. Pedestrian detection via classification on riemannian manifolds. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(10):1713–1727, 2008. 3

[12] 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

[13] 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, 2

[14] 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

[15] 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