DYNAMIC TIME WARPING FOR GESTURE-BASED USER IDENTIFICATION AND AUTHENTICATION WITH KINECT

Jonathan Wu, Janusz Konrad, and Prakash Ishwar

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

ABSTRACT

The Kinect has primarily been used as a gesture-driven device for motion-based controls. To date, Kinect-based research has predominantly focused on improving tracking and gesture recognition across a wide base of users. In this paper, we propose to use the Kinect for biometrics; rather than accommodating a wide range of users we exploit each user's uniqueness in terms of gestures. Unlike pure biometrics, such as iris scanners, face detectors, and fingerprint recognition which depend on irrevocable biometric data, the Kinect can provide additional revocable gesture information. We propose a dynamic time-warping (DTW) based framework applied to the Kinect's skeletal information for user access control. Our approach is validated in two scenarios: user identification, and user authentication on a dataset of 20 individuals performing 8 unique gestures. We obtain an overall 4.14%, and 1.89% Equal Error Rate (EER) in user identification, and user authentication, respectively, for a gesture and consistently outperform related work on this dataset. Given the natural noise present in the real-time depth sensor this yields promising results.

Index Terms— Dynamic Time Warping, Kinect, Revocable biometrics, Skeletal tracking, User Identification and Authorization

1. INTRODUCTION

The inherent nature of pure biometrics [1] is that it relies on irrevocable information. This information comes from natural characteristics: the face, iris, retina, and one's fingerprint. Unfortunately, some of these characteristics can be easily intercepted: the face is public information, and fingerprints can be unintentionally left behind on surfaces. As a result, once this information is stolen, it becomes difficult to reliably prove one's identity. This makes revocable alternatives desirable. This is achievable by adding additional information through body gestures, which can be altered and changed if necessary.

This work supported in part by the National Science Foundation under awards CCF-0905541 and CNS-1228869.

The Kinect's growing popularity has led to a wide assortment of depth-based applications ranging from gesture-based controls [2] to full body gait analysis [3]. A natural extension of these applications would be into the domain of user access controls. By performing a gesture in front of the Kinect and extracting its depth information, unique user information can be obtained.

This paper provides two key contributions to the area: an adapted dynamic time warping framework that uses the Kinect skeletal model with revocable gestures in the context of user access controls, and an in-depth analysis of relevant evaluation methods (authorized/unauthorized user group splits) in the scenarios of authorization and identification.

1.1. Overview and Related Work

There have been few related works in the context of Kinect-based user authentication and identification. Sae-Bae *et al.* [4] explore multi-touch gesture authentication using dynamic time warping with finger features. Ball *et al.* [5] propose using k-means clustering to identify users based on skeletal walking gait but report very limited experimental results. Lai *et al.* [6] use Kinect-based body silhouettes to identify users with an empirical log-covariance framework.

Our methodology expands on the approach of Lai *et al.* [6] by adopting the Kinect skeleton model instead of the body silhouette. Since the Kinect SDK returns labeled skeletal joints across time, we can treat the joint coordinates as separate time-series, and evaluate pairwise distances between joints using dynamic time-warping. Furthermore, we extend the method described in [6] to perform user authentication (the original method was applied to identification only[1]) and use it as a baseline against our newly proposed framework.

2. METHODOLOGY

2.1. Skeleton Model Features

The Kinect SDK [7] provides real-time 20-joint skeletal tracking by using depth information. These skeleton models

[1] Lai *et al.*[6] refer to identification as authentication

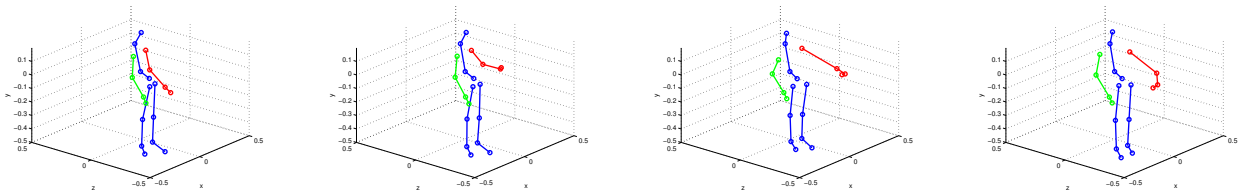

Fig. 1. Skeletal time-snapshots of a user swinging his left arm. Red and green indicate the left and right arms, respectively, and blue indicates the center of the body: head to spine, and both legs. All gestures are arm based.

track the following center-body joints: head, neck, spine, center hip, as well as the left- and right-side joints: hand, wrist, elbow, shoulder, hip, knee, ankle and foot. These joints are shown in Fig. 1.

In the dataset used, a single gesture consists of the joint information (x,y,z coordinates) across all 20 skeletal joints over the span of 30 frames (1 second). Eight different gestures were performed by 20 different users, with each user repeating the same gesture 5 times. The performed gestures include the following: right-arm swing, right-arm push, right-arm back, left-arm swing, left-arm push, left-arm back, zoom-in (arms moving outwards), zoom-out (arms closing inwards). In total, this dataset consists of 8×20×5 = 800 short gesture sequences across all users. This dataset was also used in [6, 8].

2.1.1. Normalization of joint coordinates

In order to remove natural biases in the data, robust normalization is necessary. In this scenario, the bias stems from relative translational and depth positioning, as well as from natural size variations among subjects. To rectify this, spine-based centering and length normalization were applied as follows:

$$\begin{aligned} \mathbf{x}^g_{i,t} &= (x^g_{i,t}, y^g_{i,t}, z^g_{i,t}) \\ \mathbf{x}^g_{center,i,t} &= (x^g_{i,t} - x^g_{spine,t}, y^g_{i,t} - y^g_{spine,t}, z^g_{i,t} - z^g_{spine,t}) \\ \mathbf{x}^g_{norm,i,t} &= \frac{\mathbf{x}^g_{center,i,t}}{||\mathbf{x}^g_{center,neck,t} - \mathbf{x}^g_{center,spine,t}||}, \end{aligned}$$

where $x^g_{i,t}, y^g_{i,t}, z^g_{i,t}$ are the 3D coordinates of joint number i at time t for gesture g. To center a gesture at the spine, the spine joint coordinates were subtracted from all joints at every time instant. To normalize the subject's size, all the coordinates were scaled by the distance between the neck and spine joints.

2.2. Dynamic Time Warping Review

Dynamic time warping (DTW) is a well-known sequence alignment algorithm that finds a non-linear warping path between two time-varying sequences. In our case, the time sequences are of the same length. Regardless, this warping path finds the minimum cumulative cost to align the sequences.

This algorithm has been thoroughly investigated and its properties explored in several publications [9, 10, 11, 12]. We review this algorithm briefly below, and then adapt it to our application. Consider two time sequences $\mathbf{X}^{g_1}$ and $\mathbf{X}^{g_2}$ resulting from gestures g_1 and g_2. We define a time sequence for gesture g as a collection of skeletal joints using the following notation:

$$\begin{aligned} \mathbf{X}^g &= (\mathbf{X}^g_1, \mathbf{X}^g_2, \ldots \mathbf{X}^g_n), \\ \mathbf{X}^g_t &= (\mathbf{x}^g_{norm,1,t}, \mathbf{x}^g_{norm,2,t}, \ldots \mathbf{x}^g_{norm,d,t}), \end{aligned}$$

where n is the length of the sequence and $\mathbf{X}^g_t$ denotes the collection of d joints for gesture g at time t. For our case, we use the entire set of body joints, where $d = 20$.

To align two sequences $\mathbf{X}^{g_1}$, and $\mathbf{X}^{g_2}$, we define $cost$ to be the $n \times n$ cost matrix where the cost associated with time instants (i, j) is given by:

$$cost(\mathbf{X}^{g_1}_i, \mathbf{X}^{g_2}_j) = \sum_{p=1}^{d} ||\mathbf{x}^{g_1}_{norm,p,i} - \mathbf{x}^{g_2}_{norm,p,j}||$$

where $||\cdot||$ is the Euclidean distance between d pairs of corresponding joints in the skeletal models.

Let $\mathbf{P}$ be defined as a possible path along the cost matrix $cost$ as follows:

$$\begin{aligned} \mathbf{P} = \{(i_k, j_k), k = 1, \ldots, K : i_1 = j_1 = 1, i_K = j_K = n, \\ i_k \leq i_{k+1}, j_k \leq j_{k+1}\} \end{aligned}$$

where K is the path-length which can range from n to $2n$, inclusive. The cost of this path is defined as follows:

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

For our final DTW cost, we are interested in the path with least cost among all paths that begin at $(1, 1)$ and end at (n, n). The least cost path can be found via dynamic programming [9]. The final cost between gesture g_1 and gesture g_2 is given by:

$$DTW(\mathbf{X}^{g_1}, \mathbf{X}^{g_2}) = \min_{\mathbf{P}} pathcost(\mathbf{P}, \mathbf{X}^{g_1}, \mathbf{X}^{g_2}) \quad (1)$$

The run-time complexity of baseline DTW is $O(n^2)$. Since Euclidean distance computation has $O(d)$ complexity which is linear in the number of joints, the overall complexity is $O(dn^2)$. However, it has been shown that the amortized cost for large datasets is typically of order $O(n)$ [11]. Thus with our cost function, the complexity is typically of order $O(dn)$.

3. EVALUATION

We define a gesture sample, or simply a sample, as one execution of an arbitrary gesture. Multiple samples are associated with each user, all of them resulting from a variety of gestures. In our dataset, there are 20 users with 40 samples per user (each of 8 gestures executed 5 times). We evaluate user-access control in two possible scenarios:

- **User Authentication**: User provides sample and identity information (analogous to ID and password in computer login) for admission eligibility.
- **User Identification**: User provides sample but no identity information. This is a harder task than authentication as the system must evaluate admission eligibility against the entire set of authorized users.

In order to exhaustively test the proposed method, K users are split into L authorized users, and $(K-L)$ unauthorized users, called a $L/(K-L)$ split. For every split, there are "K choose L" combinations of authorized versus unauthorized users. Samples of every such combination are evaluated for a given distance threshold θ to obtain the false acceptance and rejection rates (FAR and FRR).

In particular, in one test, each sample of each unauthorized user (query sample) is compared against samples from authorized users. If the distance between the query sample and the closest authorized user's sample is below a threshold θ, a false acceptance is declared. In another test for the same threshold θ, leave-one-out cross-validation (LOOCV) is performed on the samples of authorized users, i.e., one (query) sample of an authorized user is removed and evaluated against the remaining authorized samples. If the distance between the query sample and the closest of the remaining samples is above threshold θ, a false rejection is declared. FAR is defined as the fraction of unauthorized user samples that were falsely accepted. FRR is the fraction of authorized user samples that were falsely rejected. These values are computed for a range of θ values. For a suitable choice of θ, the FAR and FRR values become equal. This common value is called the equal error rate (ERR).

We now describe our evaluation methodology in detail. For a given $L/(K-L)$ split, let each of the "K choose L" combinations yield the following: the set $\mathcal{A}$ containing all samples s belonging to L authorized users, and its complement of samples belonging to $K-L$ unauthorized users, $\mathcal{U}$:

$$\mathcal{A} = \{s_1, \ldots s_m\}, \quad \mathcal{U} = \mathcal{A}^C, \quad \mathcal{S} = \mathcal{A} \cup \mathcal{U}, \quad \mathcal{A} \cap \mathcal{U} = \emptyset,$$

where m is the number of authorized samples, and $\mathcal{S}$ contains all n samples (authorized and unauthorized), i.e., $\mathcal{S} = \{s_1, \ldots s_m, \ldots s_n\}$.

Let $d(s_i, s_j)$ denote a distance function between two samples, here, this is the dynamic time warping cost (1), and the log-covariance metric in the work of Lai *et al*. [6]. The distance of sample s_i from authorized samples in $\mathcal{A}$ can then be defined as follows:

$$d(s_i, \mathcal{A}) = \min_{s_j \in \mathcal{A}} d(s_i, s_j)$$

For a given threshold θ, and set of authorized samples $\mathcal{A}$ from a single split we can obtain the false acceptance and false rejection counts (FAC and FRC) for identification (ID) and authentication (AU) as follows:

$$\begin{aligned}
FAC^{ID}(\mathcal{A}, \theta) &= \sum_{s \in \mathcal{U}} 1(d(s, \mathcal{A}) < \theta) \\
FRC^{ID}(\mathcal{A}, \theta) &= \sum_{s \in \mathcal{A}} 1(d(s, \mathcal{A} \backslash \{s\}) \geq \theta) \\
FAC^{AU}(\mathcal{A}_u, \theta) &= \sum_{s \in \{\mathcal{S} \backslash \mathcal{A}_u\}} 1(d(s, \mathcal{A}_u) < \theta) \\
FRC^{AU}(\mathcal{A}_u, \theta) &= \sum_{s \in \mathcal{A}_u} 1(d(s, \mathcal{A}_u \backslash \{s\}) \geq \theta)
\end{aligned}$$

where $\mathcal{A}_u \subset \mathcal{A}$ is the set of samples within $\mathcal{A}$ that belong to a specific user u, and $1(\mathbf{B})$ is the indicator function which equals 1 when $\mathbf{B}$ is true and 0 otherwise. In both counts, the threshold θ compares the minimum distance a sample is from all authorized samples (1-nearest-neighbor) through the indicator function. For identification, the authorized sample is any sample within the authorized group. For authentication, the authorized sample is any sample that belongs to the authorized user whose identity is being claimed.

To get the FAR and FRR for a given θ across all given splits for L authorized users, we consider all "K choose L" authorized/unauthorized user combinations for identification as follows:

$$\begin{aligned}
FAR^{ID}(L, \theta) &= \frac{\sum_{\mathcal{A} \subset \mathcal{S}} FAC^{ID}(\mathcal{A}, \theta)}{\binom{K}{L} |\mathcal{U}|}, \\
FRR^{ID}(L, \theta) &= \frac{\sum_{\mathcal{A} \subset \mathcal{S}} FRC^{ID}(\mathcal{A}, \theta)}{\binom{K}{L} |\mathcal{A}|},
\end{aligned}$$

where we compute all possible L user subsets $\mathcal{A}$ within $\mathcal{S}$, and for authentication:

$$\begin{aligned}
FAR^{AU}(L, \theta) &= \frac{\sum_{\mathcal{A}_u \in \mathcal{A} \subset \mathcal{S}} FAC^{AU}(\mathcal{A}_u, \theta)}{\binom{K}{L} (|\mathcal{S}| - |\mathcal{A}|/L)(L)} \\
FRR^{AU}(L, \theta) &= \frac{\sum_{\mathcal{A}_u \in \mathcal{A} \subset \mathcal{S}} FRC^{AU}(\mathcal{A}_u, \theta)}{\binom{K}{L} (|\mathcal{A}|/L)(L)}
\end{aligned}$$

Gesture	Identification						Authentication					
	Proposed method			Lai et. al.[6]			Proposed method			Lai et. al.[6]		
Group Split	19/ 1	15/ 5	10/ 10	19/ 1	15/ 5	10/ 10	19/ 1	15/ 5	10/ 10	19/ 1	15/ 5	10/ 10
Right Swing	**6.02%**	**6.02%**	**5.28%**	7.07%	6.95%	5.74%	**3.98%**	**3.98%**	**3.98%**	4.04%	4.04%	4.01%
Right Push	**3.99%**	**3.22%**	**2.91%**	8.11%	8.31%	8.70%	**2.03%**	**2.03%**	**1.98%**	3.74%	3.73%	3.73%
Right Back	1.01%	1.01%	1.00%	**0.00%**	**0.00%**	**0.00%**	1.01%	1.00%	1.03%	**0.00%**	**0.00%**	**0.00%**
Left Swing	4.08%	**4.02%**	**2.99%**	**4.03%**	4.03%	3.99%	**1.12%**	**1.11%**	**1.11%**	2.01%	2.01%	2.01%
Left Push	9.05%	8.58%	7.61%	**5.04%**	**4.99%**	**4.04%**	2.02%	**2.01%**	**1.96%**	**2.01%**	**2.01%**	2.01%
Left Back	1.01%	0.99%	1.01%	**0.00%**	**0.00%**	**0.00%**	**0.00%**	**0.00%**	**0.00%**	**0.00%**	**0.00%**	**0.00%**
Zoom-in	**5.02%**	**4.94%**	**4.10%**	9.57%	9.05%	7.99%	**1.02%**	**1.02%**	**0.97%**	2.45%	2.45%	2.45%
Zoom-out	**7.97%**	**6.31%**	**5.71%**	10.95%	8.95%	7.65%	**2.59%**	**2.59%**	**2.59%**	7.97%	8.02%	7.83%
All Gestures	**4.14%**	**4.12%**	**3.51%**	6.92%	6.49%	6.16%	**1.89%**	**1.89%**	**1.89%**	2.79%	2.73%	2.73%

Table 1. Equal Error Rate (EER) of various gestures for identification and authentication compared to the baseline [6]. Group splits denote authorized/unauthorized users, i.e. 19/1 denotes the EER for 19 authorized/1 unauthorized user splits.

where we compute all possible L user subsets $\mathcal{A}$ within $\mathcal{S}$, across every subset $\mathcal{A}_u$ within $\mathcal{A}$. We average this across the total possible number of false rejections and acceptances.

Authorized group splitting is useful because it evaluates a methodology for different amounts of authorized users. For example, a large authorized group split (typically harder to solve), depicts a scenario where user access is shared amongst many users (such as a door), and a smaller authorized group split, depicts a scenario where user access is more personal (such as a computer). We evaluate the aforementioned methods against previous work [6] on this dataset for the splits (various L values) of 19 authorized users, 15 authorized users, and 10 authorized users.

We also evaluate our method with various sets of $\mathcal{S}$. We consider 9 sets in total: 8 sets of users only performing single gestures, and the set of all the gestures. Effectively, in this evaluation we associate only a single gesture with a user, or multiple gestures with a user. For the single-gesture sets, 5 samples are associated with each user (depending on the gesture) and for the comprehensive set (all gestures), all 40 gesture-samples are associated with each user. As Table 1 shows, the proposed method provides a $32-40\%$ overall EER improvement for "All Gestures" for both identification and authentication over [6]. This improvement over [6] is also shown in Fig. 2 which shows FAR vs. FRR plots for the 19/1 split.

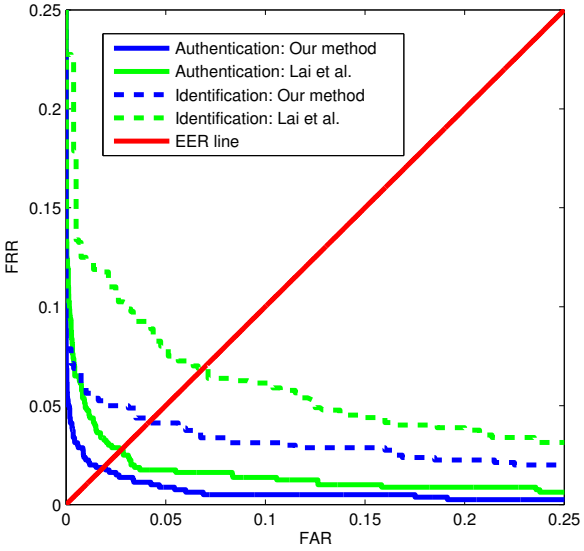

Fig. 2. FAR vs FRR performance in identification and authentication for the proposed metric evaluated against Lai *et. al* [6] for the 19/1 split.

4. CONCLUSIONS

We have proposed and shown the viability of an extension to Kinect skeletal tracking in the domain of user access control.

In the most ideal scenario, skeletal tracking is not biased by one's body shape (i.e., influences from additional layers of clothing) and only tracks one's underlying joints. This can give it more robustness than methods that purely rely on the body silhouette. In our method, we use skeletal tracking from the Kinect SDK. Although there exist alternative, potentially more robust methods [13] to calculate skeletal models from depth maps aside from the baseline Kinect SDK, our framework can easily be adapted to future improvements to skeletal tracking.

This work has shown that the skeletal model contains unique user information, and coupled with a user's simple gesture can be potentially used as a revocable biometric. Future work may look into more sophisticated gestures, as well as an extension of this work into continuous authentication, where users are recognized across a longer period of time.

5. REFERENCES

[1] A.K. Jain, A.A. Ross, and K. Nandakumar, *Introduction to Biometrics*, Springer, 2011.

[2] 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*, 2011, pp. 147–156.

[3] M. Gabel, E. Renshaw, A. Schuster, and R. Gilad-Bachrach, "Full body gait analysis with kinect," in *Proc. Engineering in Medicine and Biology Society*, Sept. 2012, pp. 1964–1967.

[4] N. Sae-Bae, K. Ahmed, K. Isbister, and N. Memon, "Biometric-rich gestures: a novel approach to authentication on multi-touch devices," in *Proc. 2012 ACM Annual Conference on Human Factors in Computing Systems*, 2012, pp. 977–986.

[5] A. Ball, D. Rye, F. Ramos, and M. Velonaki, "Unsupervised clustering of people from 'skeleton' data," in *Proc. Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction*, 2012, pp. 225–226.

[6] 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*, Sept. 2012, pp. 282 –287.

[7] 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*, June 2011, pp. 1297–1304.

[8] 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*, April 2012, pp. 185 –188.

[9] T.K. Vintsyuk, "Speech discrimination by dynamic programming," *Cybernetics*, vol. 4, pp. 52–57, 1968.

[10] C.S. Myers and L.F. Habiner, "A comparative study of several dynamic time-warping algorithms for connected-word," *Bell System Technical Journal*, vol. 60, Sept. 1981.

[11] C.A. Ratanamahatana and E. Keogh, "Everything you know about dynamic time warping is wrong," in *Proc. 3rd Workshop on Mining Temporal and Sequential Data, in conjunction with the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug. 2004, pp. 22–25.

[12] X. Wang, A. Mueen, H. Ding, G. Trajcevski, P. Scheuermann, and E. Keogh, "Experimental comparison of representation methods and distance measures for time series data," *Data Mining and Knowledge Discovery*, vol. 26, pp. 275–309, Mar. 2013.

[13] W. Shen, K. Deng, X. Bai, T. Leyvand, B. Guo, and Z. Tu, "Exemplar-based human action pose correction and tagging," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, June 2012, pp. 1784 –1791.