Gait Recognition Using Encodings With Flexible Similarity Measures

Michael B. Crouse Kevin Chen H. T. Kung
School of Engineering and Applied Sciences
Harvard University

Abstract

Gait signals detectable by sensors on ubiquitous personal devices such as smartphones can reveal characteristics unique to each individual, and thereby offer a new approach to recognizing users. Conventional pattern matching approaches use inner-product based distance measures which are not robust to common variations in time-series analysis (e.g., shifts and stretching). This is unfortunate given that it is well understood that capturing such variations is paramount for model performance. This work shows how machine learning methods which encode gait signals into a feature space based on a dictionary can use convolution and Dynamic Time Warping (DTW) similarity measures to improve classification accuracy in a variety of situations common to gait recognition. We also show that data augmentation is crucial in gait recognition, as diverse training data in practical applications is very limited. We validate the effectiveness of these methods empirically, and demonstrate the identification of user gait patterns where shift and stretch variations in measurements are substantial. We present a new gait dataset that contains a complete representation of the variations that can be expected in real-world recognition scenarios. We compare our techniques against the current state of the art gait period detection and normalization schemes on our dataset and show improved classification accuracy under all experimental scenarios.

1 Introduction

The Internet of Things (IoT) has created an explosion of sensor data due to the increased number of devices with embedded sensors, ranging from smart watches and phones to healthcare wearables and head-mounted devices. These devices, combined with new environments such as connected consumer smart homes, have opened a new set of applications and scenarios directly enabled by sensor data. Novel techniques that utilize and extract meaningful information from this vast stream of data are the key to success in this new area.

Much of this new sensor data is measured over time. The analysis of time-series data is a well studied subject explored in the signal processing community for applications such as networking, computer vision and speech recognition. More recently, there has been substantial interest in individual motion tracking and personalized gesture recognition for improved human and machine interaction with applications in user interface design and gaming. These efforts leverage the recent growth in the number and type of devices containing sensors for motion, audio, and video, yet there remains much work to be done in how to make best use of such a diverse set of sensors.

Gait recognition is one form of motion tracking that has been studied for a variety of reasons including fall detection, locomotion for robotics, and health/fitness monitoring. Human gait is the result of the cyclic motion of a set of human limbs [2]. Gait has also been used to uniquely identify a human using visual sensors (video cameras) as well as motion sensors. Using biometrics such as gait or fingerprints for identifying users is an important approach that will become increasingly useful in IoT. Gait is an ideal target as it can be obtained passively, monitored using a variety of sensors, and provides a fairly robust indicator of identity. Gait is also promising as a user identification mechanism that could improve the usability and security provided by current authentication schemes.

Machine learning techniques have achieved great successes in the fields of computer vision and speech recognition. In this paper, we adapt the same framework and propose a representation encoding method tailored to gait recognition, and report improved performance over the best published results.

2 Related Work

User identification using gait patterns with motion-based sensors has been the subject of much study over the past decade. Most of the approaches regarding gait recognition utilize accelerometers attached to subject for gathering data. In many cases, the accuracy of these systems are fairly high, but require multiple commercial sensors with several fixed points and extensive configuration [5, 10]. Recently, due to the integration of accelerometers and gyroscopes into smart phones, several new approaches have been proposed to reduce the number of sensors and relax the constrains on sensor placement [7, 4, 8, 9]. The most common technique for dealing with time-series data is through gait period detection, which involves locating the strike points of a subjects' gait signal. A strike point corresponds to a subject's heel striking the ground.

The work by Juefei-Xu et. al. is on recognizing a user's gait pattern recorded from off the shelf Android devices with users walking down a hallway [7]. Their technique relies heavily on normalization around the strike points. Their two normalization methods are: 1) centering measurements around a strike-point and 2) measurements between consecutive strike-points, interpolated to get uniform segment lengths. This approach attempts to address shift and stretch variance that naturally occurs in a human's gait between steps. Frank et. al. proposes the use of nonlinear dynamic systems to form a geometric time delay embedding per subject [4]. After training, they classify each new sample by selecting the nearest-neighbor using Euclidean distance. Their model has high accuracy but requires large segments of data to be able to perform classification accurately.

Our approach uses a general machine learning model coupled with traditional signal processing similarity measures. Specifically, we utilize a dictionary/encoding framework and use convolution and Dynamic Time Warping (DTW) measures for encoding to provide more robust representations for classification in a scenario with large intra-class variation. We contribute a unique dataset that was captured by a smart phone under a wide variety of conditions including different paces, phone orientations, and over different days. We report the highest accuracy compared with the traditional gait period detection and normalization methods.

3 Problem Definition

Utilizing accelerometer and gyroscope measurements of users' gait for recognition is fundamentally a time-series analysis problem. As previously mentioned, time-series analysis is a well-studied area explored mostly by the signal processing community, one of the key applications being voice recognition and speech translation. While gait recognition involves similar problems, it has several unique properties and challenges that need to be addressed. Unlike audio data captured from speech, gait pattern data captured by motion sensors is periodic in nature. Periodicity in the data provides both benefits in terms of recognition and challenges in determining the appropriate unit to perform recognition (a single step or an ensemble of steps). Additionally, measured gait patterns for a single user can have large variations between days due to factors such as phone placement on a user's body and terrain differences. Another source of variation stems from different paces, as humans rarely maintain a single pace while walking and between walks due to an assortment of variables including mood, environment and destination.

Past research has shown the feasibility of gait pattern recognition using data captured from accelerometers and gyroscopes. However, it remains a challenging problem due to several reasons: (1) lack of public datasets, (2) the recognition scheme needs to work with a very small amount of training data, (3) high degree of signal variations (pace/phone placement, etc).

We have identified several imperative criteria for any system that attempts gait recognition:

1. robust under different phone placement
2. robust under different walking speed
3. robust between different days
4. requiring only a small amount of training data

Our approach utilizes conventional machine learning techniques coupled with traditional time-series similarity measures to address each of these necessary conditions.

4 Methodologies

With sufficient samples, we can separate classes well using our feature representations, as demonstrated by the 99% accuracy for the complete training case in Table 2. We have observed the same phenomena in another public dataset [3], of which more than 98% of the samples are linearly separable if trained on the entire dataset.[1] In the following section we describe our approach of addressing the situation where only limited amount of samples are available. Specifically, we need to handle large variations in gait signals of the same user over different days and at different paces using the very few samples available during the training phase for better handling of variations such as shifts and stretches.

[1] This is generally not true in most other applications.

We propose a pipeline that includes (1) preprocessing, (2) feature encoding, (3) linear Support Vector Machine (SVM) classification, and (4) data augmentation. Preprocessing is a fixed process that makes our method insensitive to sensor orientation. Feature encoding deals with more variations, and requires (unlabeled) data for training. We use a linear Support Vector Machine (SVM) to classify samples in their feature representations. Finally, data augmentation is used to increase sample diversity.

4.1 Preprocessing

Gyroscopes and accelerometers both report readings for 3 directions (or axes). Let a segment x_{raw} be a $3 \times T$ matrix, where T is the length of the signal. We compute the 3×1 principal eigenvector v from x_{raw}, and then compute $x = v^T x_{raw}$. This makes x insensitive to sensor orientation.

4.2 Feature Encoding

We use a dictionary/coding framework for representation encoding. In particular, we consider the simplest setting: random patch dictionary and distance encoding. We construct the random patch dictionary D by selecting random samples from the training set. In distance encoding, a feature vector f for a given sample x is computed by $f_i = \mathsf{dist}(x, d_i)$ with components f_i where $\{d_i\}$ are the entries in the dictionary, and dist is some distance measure. By encoding we mean the process of computing the feature vector f for a given sample x. As we will see, samples corresponding to gait signals of different users are readily separable with linear SVM when samples are expressed in their feature representations.

- **Encoding with Convolutional Distance Measure** The most obvious variation in gait signal is perhaps the shift. A shift is a change translation in time of a gait signal. When a gait signal is captured from a user, the alignment for comparison with other signals is unknown and must be accounted for. Therefore, we want the feature encoding to be shift invariant, so that shifted versions of the same pattern would be transformed to the same feature vector. To this end, we use the distance measure defined as follows:

  $$\mathsf{dist}_{\mathsf{conv}}(a,b) = \mathsf{max}(|\mathsf{conv}(a_r, b)|)$$

  where a_r is a in the reverse order. This finds the offset that gives maximum correlation. We will use $\mathsf{Enc}_{\mathsf{conv}}$ to denote encodings with $\mathsf{dist}_{\mathsf{conv}}$.

- **Encoding with DTW Distance Measure** Another classical measure for evaluating similarity between time series is dynamic time warping (DTW) [11]. DTW finds an optimal "path" that can morph one signal to another. For gait signal, this means DTW will likely consider x and x' similar if they are compressing/stretching variations of one another. Figure 2 shows a cartoon that demonstrates the flexibility of this measure. The DTW distance measure is defined as follows:

  $$\mathsf{dist}_{\mathsf{dtw}}(a,b) = \mathsf{DTW}_{\mathsf{cost}}(a,b)$$

  We will use $\mathsf{Enc}_{\mathsf{dtw}}$ to denote encodings with $\mathsf{dist}_{\mathsf{dtw}}$. Note that in the DTW algorithm, one may specify the largest matching range, which in our case can be conveniently set to half of the largest step size.

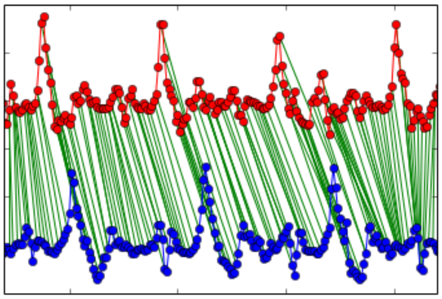


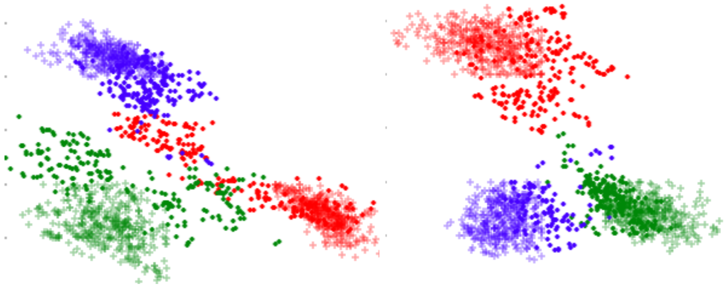
Figure 2: DTW alignment between two segments of the same subject. Blue is sampled from a normal pace session, and red is from a fast pace session. Note that $\mathsf{dist}_{\mathsf{dtw}}$ for this pair of signals would be small because DTW found an optimal way to align them.

In Figure 1(a), we illustrate how our encoding methods can make data more separable. This visualization is made by projecting encoded samples from 3 randomly selected classes onto 2D. One can see that the training data is very separable and form tight clusters for $\mathsf{Enc}_{\mathsf{conv}}$. However, the test examples do not necessarily fall into the correct clusters and may be on the wrong side of the decision boundary. We alleviate this problem with data augmentation.

4.3 Data Augmentation

Besides designing variation-tolerant encodings, data augmentation is another way to deal with variations in samples. For example, applying a shift-invariant encoding corresponds to augmenting the data with shifted-variations of observed data in terms of improving the match between the training and testing distributions.

We identify three major natural variations in gait patterns: shifts, stretching, and compression. Stretching and compression corresponds to scaling the signal in the



(a) Enc_{conv}

(b) Enc_{conv} with data augmentation

Figure 1: Samples projected onto decision planes trained on encoded representations in the feature space with and without augmentation. Transparent plus signs (+) represent samples from the training set, and opaque dots are samples from testing set. Note that data augmentation reduces the amount of overlap between classes.

time axis. We can acquire these variations in the training set by sampling a session in the following way: (1) select a random starting point, (2) select random window size and re-sample signal within the window into a fixed length. Samples generated this way have all three types of variations. In Figure 1(b), we show the advantage of data augmentation using convolution as the similarity measure.

5 Dataset Generation

In this section, we will describe how our dataset is generated and collected. We also provide a brief overview of accelerometers and gyroscopes in terms of what they measure.

5.1 Devices

IoT devices, specifically wearables, are becoming prevalent and most include some form of motion sensing chip. For example, wearables such as the Pebble Watch and Nike+ FuelBand contain 3-axis accelerometers [1]. These sensors are prevalent in smart phones including Apple's iPhone and most Android devices. Our experimental device is an Android HTC Droid DNA placed in the subject's front left pocket. The Android smart phone captures both accelerometer and gyroscope data and records it locally before it is uploaded as a batch for processing. We were able to sample at 50 Hz for both the accelerometer and gyroscope, which should be sufficient for capturing the necessary characteristics unique to a subject's gait.

5.2 Accelerometer

An accelerometer measures the change of position of a test mass. Accelerometers react to a large number of external forces including linear motion, gravity, centripetal force, and other motions [6]. The measurement taken from an accelerometer is the sum of all these forces in terms of acceleration. A 3-axis accelerometer provides measurements along 3 orthogonal axes, x, y, z.

day 1	day 2
P=1, O=1, Pace=1	P=1, O=1, Pace=1
P=2, O=1, Pace=1	P=1, O=1, Pace=2
-	P=1, O=2, Pace=1

Table 1: Experimental configurations recorded for each subject. Each cell is a single session corresponding to 50 seconds sensor data. For day 1 we collect 2 sessions for each setting. P=path (1,2), O=orientation (1,2), Pace (1=normal, 2=fast)

5.3 Gyroscope

Gyroscopes are sensors that measure the angular velocity of an object. They measure the rate of rotation around a single axis. Smart phones and wearables often contain 3-axis gyroscopes capable of extremely accurate, low-latency measurements. These sensors provide a good balance to the accelerometer as the measurements are not biased by gravity or magnetic forces and are less noisy [6].

5.4 Data Collection and Description

Gait measurements were collected under a variety of device orientations, paces, over different paths and days. We define a session of gait measurements as a single walk around our pre-defined course by a subject. Each session is between 40 and 50 seconds in duration where a smart device is placed in the subjects front left pocket

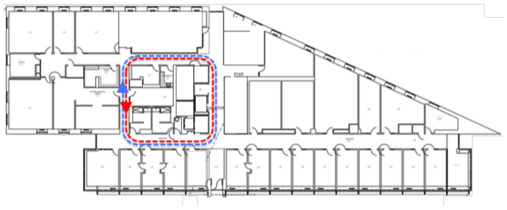


Figure 3: Subjects walk on two paths: counterclockwise and clockwise around the hallway Each subject walks each path twice for a total of 4 recorded walks.

and the watch placed on their left wrist. The orientation of the device is always vertical, with the device's z-axis pointed upward, and we alter whether the device screen faces outward, either away from or facing the subject. We have two different paces, one normal corresponding to 3 to 5 feet per second and one brisk at 4 to 6 feet per second. The two paths can be seen in Figure 3. The paths correspond to a loop around an office hallway, one clockwise and the other counter-clockwise. We recorded all of the subjects on two separate days. The break-down of session configurations is listed in Table 1. We recorded 31 different subjects from our office building, an assortment of students and staff. We have full data for 31 subjects under the basic walk experimental configuration. Our dataset is robust in the sense that we have 9 subjects with multiple days and variations including phone orientation and pace.

6 Experiments

We describe 5 different experiments using our gait dataset. The complete training set serves as a baseline for evaluating the capability of our model and confirming our intuition described in Section 4. The other 4 experiments are designed to observe increasing levels of variance between training and testing data. We consider the following settings:

1. **Complete Training Set**: We take all sessions from the dataset, break them down into segments, and split the set of segments into training and testing sets. The training set is complete because it contains samples from every session. Note that the testing set is still using disjoint set of samples.

2. **Different Sessions**: We take sessions from the same day under the same orientation and pace setting, and split them into training and testing sessions.

3. **Different Orientation**: Sessions are taken from the same day under same pace, and then split by orientation into training and testing sessions.

4. **Different Days**: Sessions are taken under same pace and same orientation, and then split by days into training and testing sessions.

5. **Different pace**: Sessions are taken from the same day under same orientation, and then split by pace into training and testing sessions.

Each training and testing segment contains 300 samples, which roughly correspond to 6 to 7 seconds.

6.1 Evaluation of Data Augmentation and Feature Encoding

We explore the intra-class variations under different settings, and show how different schemes performs under these variations. The reported performance is the multi-class classification accuracy in each of the described settings. Classification accuracy is the number of correctly labeled predictions over the number of total examples in the test set. We discuss the impact of data augmentation and the empirical results of two similarity measures, convolution and DTW.

6.1.1 Data Augmentation

We extract segments of randomized length from training sessions, and then re-sample them with bi-cubic interpolation into segments of a fixed length. In these experiments, the window size is sampled from $\mathsf{Gaussian}(300, 30)$, where the standard deviation (30) is selected empirically to give the best result. The resulting segments are of length 300.

The performance of the convolution encoding with and without data augmentation is shown in the first 2 columns of Table 2. For the complete training set, classification is near perfect because there is little variation within the same session (as shown earlier in Section 4). The convolution measure captures most of the variations between different sessions with the larger number of subjects. It is also able to capture most of the variations under changes in orientations with 88% accuracy. It is important to note that this is with a lower number of subjects.

For sessions over different days we see a larger accuracy drop, implying a more significant change in gait pattern over days. The lower accuracy for different pace is expected, because the convolve-transform is not intended to be scale invariant. With data augmentation, the accuracy is comparable in the first three setting because the variation introduced by our synthetic data is not necessarily important for the testing set. On the other hand, we get significantly better results by having synthetic data for predicting sessions of different pace. This suggests

setting (# subjects)	$\mathsf{Enc}_{\mathsf{conv}}$	$\mathsf{Enc}_{\mathsf{conv}} + \mathsf{aug}$	$\mathsf{Enc}_{\mathsf{dtw}}$	α	β
complete training set (31)	.99	.99	.99	.94	.90
different sessions (31)	.88	.88	.99	.90	.88
different orientation (9)	.88	.92	.73	n/a	n/a
different days (9)	.69	.77	.76	.48	.53
different pace (9)	.48	.70	.79	.75	.76

Table 2: Comparison across different methods. It shows a significant improvement of $\mathsf{Enc}_{\mathsf{conv}}$ by data augmentation. Data augmentation does not change the outcome of $\mathsf{Enc}_{\mathsf{dtw}}$ much and is therefore not listed on the table. α and β are normalization-based methods proposed by Juefei-Xu et. al. Note the first two experiments consist of 31 subjects while the others are only 9.

that our data augmentation captures some variation between different pace and over days.

6.1.2 Encoding Gains with $\mathsf{Enc}_{\mathsf{conv}}$ and $\mathsf{Enc}_{\mathsf{dtw}}$

As described in Section 4, it is important for the transformation to capture legitimate variances related to pace. In Section 4 we hypothesize that the purpose for encoding in our case is to generalize training data. We demonstrate the performance of $\mathsf{Enc}_{\mathsf{dtw}}$ and $\mathsf{Enc}_{\mathsf{conv}}$ in Table 2 (column 2 and 4) in order to empirically validate these claims.

Our method outperforms the state-of-art for general, individual sessions, and inter-day sessions. Many existing approaches rely on normalization techniques to reduce intra-class variations. While they can achieve reasonable accuracy with normalized data, the normalization process has two drawbacks. The process requires a a long sequence to capture the periodicity necessary for determining gait cycles. Secondly, the normalization process is based on heuristics that require parametric tuning per sensor.

Juefei-Xu et. al. proposes two normalization schemes, α and β, that consist of gait period detection via using peak finding heuristics [7]. The normalization method α places the strike point in the center of each segment. The normalization method β uses sampled data between two strike points, using interpolation and re-sampling to yield equal length sequences. Their β normalization provides stretch invariance due to re-sampling the measurements between pairs of steps to yield segments of equal length.

We implemented both of their normalization methods and performed parameter screening to yield the best possible performance on our dataset using their approach (their dataset was not publicly available at the time of this writing).

Table 2 provides a comparison of our encoding schemes versus their gait period detection and normalization methods. We are able to provide improved accuracy on our dataset in all settings for $\mathsf{Enc}_{\mathsf{dtw}}$ and all but pace under $\mathsf{Enc}_{\mathsf{conv}}$ with data augmentation. The pace scenario shows that the normalization methods α and β do provide some stretch invariance whereas $\mathsf{Enc}_{\mathsf{conv}}$ does not. However, we have shown in the previous section that the gap can be bridged by data augmentation. We also show that $\mathsf{Enc}_{\mathsf{dtw}}$ outperforms the normalization-based technique by being more tolerant to variation.[2]

We have observed that insufficient training data is a challenge for gait recognition — segments within a session are practically identical, so most gait datasets have effectively less than 5 samples from each class. However, if given enough samples, gait signals seem linearly separable without involving complex non-linear transformations typical of many machine learning techniques (see first row in Table 2). Unlike most machine learning problems, the task in gait signal recognition is not about finding a non-linear transformation such that classes become linearly separable, but should instead be focused on generalizing the available training data.

7 Conclusion

We have shown that gait signals are readily separable using our encoding which requires almost no data preprocessing. This is observed in our dataset as well as other public gait datasets. The implication is that unlike many other classical classification problems (e.g., computer vision), there is no need to learn a complicated nonlinear transform to make the data easily separable.

We identify that the main challenge in gait pattern classification is the disparity (e.g., different pace) between training and testing data, due to a small number of effective samples. We discuss the characteristics of gait signals, and show how feature encoding and data augmentation alleviates the this problem. Our encoding based method outperforms the best published result in terms of short-segment gait signal classification.

[2] We attempted to introduce our orientation transformations to their normalization schemes but it significantly degraded their performance in all scenarios.

8 Acknowledgements

This material is based on research sponsored in part by the Intel Corporation. The authors would like to thank Marcus Comiter for useful discussion and suggestions. We would also like to acknowledge all the participants that contributed walks to our dataset.

References

[1] Pebble teardown. http://www.ifixit.com/. Accessed: 2014-03-04.

[2] Boyd, J. E., and Little, J. J. Biometric gait recognition. In *Advanced Studies in Biometrics*. Springer, 2005, pp. 19–42.

[3] Frank, J., Mannor, S., and Precup, D. Data sets: Mobile phone gait recognition data, 2010.

[4] Frank, J., Mannor, S., and Precup, D. A novel similarity measure for time series data with applications to gait and activity recognition. In *Proceedings of the 12th ACM international conference adjunct papers on Ubiquitous computing-Adjunct* (2010), ACM, pp. 407–408.

[5] Gafurov, D., Snekkenes, E., and Bours, P. Gait authentication and identification using wearable accelerometer sensor. In *Automatic Identification Advanced Technologies, 2007 IEEE Workshop on* (2007), IEEE, pp. 220–225.

[6] Goehl, D., and Sachs, D. Motion sensors gaining inertia with popular consumer electronics. *White Paper, IvenSense Inc* (2007).

[7] Juefei-Xu, F., Bhagavatula, C., Jaech, A., Prasad, U., and Savvides, M. Gait-id on the move: pace independent human identification using cell phone accelerometer dynamics. In *Biometrics: Theory, Applications and Systems (BTAS), 2012 IEEE Fifth International Conference on* (2012), IEEE, pp. 8–15.

[8] Kobayashi, T., Hasida, K., and Otsu, N. Rotation invariant feature extraction from 3-d acceleration signals. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on* (2011), IEEE, pp. 3684–3687.

[9] Ngo, T. T., Makihara, Y., Nagahara, H., Mukaigawa, Y., and Yagi, Y. The largest inertial sensor-based gait database and performance evaluation of gait-based personal authentication. *Pattern Recognition 47*, 1 (2014), 228–237.

[10] Rong, L., Jianzhong, Z., Ming, L., and Xiangfeng, H. A wearable acceleration sensor system for gait recognition. In *Industrial Electronics and Applications, 2007. ICIEA 2007. 2nd IEEE Conference on* (2007), IEEE, pp. 2654–2659.

[11] Sakoe, H., and Chiba, S. A dynamic programming approach to continuous speech recognition. In *Proceedings of the seventh international congress on acoustics* (1971), vol. 3, pp. 65–69.