Handwriting Watcher: A Mechanism for Smartwatch-Driven Handwriting Authentication

Conference Paper · October 2017

DOI: 10.1109/BTAS.2017.8272701

Isaac Griswold-Steiner
Texas Tech University

Richard Matovu
Texas Tech University

To appear in IEEE International Joint Conference on Biometrics (IJCB 2017), Denver, Colorado

Handwriting Watcher: A Mechanism for Smartwatch-Driven Handwriting Authentication

Isaac Griswold-Steiner, Richard Matovu, Abdul Serwadda
Texas Tech University, Lubbock, TX 79409
{isaac.griswold-steiner, richard.matovu, abdul.serwadda}@ttu.edu

Abstract

Despite decades of research on automated handwriting authentication, there is yet to emerge an automated handwriting authentication application that breaks into the mainstream. In this paper, we argue that the burgeoning wearables market holds the key to a practical handwriting authentication app. With potential applications in online education, standardized testing and mobile banking, we present Handwriting Watcher, a mechanism which leverages a wrist-worn sensor-enabled device to authenticate a user's free handwriting. Through experiments capturing a wide range of writing scenarios, we show Handwriting Watcher attains mean error rates as low as 6.56% across the population. Our work represents a promising step towards a market-ready, generalized handwriting authentication system.

1. Introduction

Handwriting authentication — the verification of a claimed identity based on handwriting — is largely accomplished manually,[1] despite the wide range of automated handwriting authentication methods proposed by researchers over the years (e.g., see [31, 19, 25]). For example, the standard protocol through which instructors verify authorship of an assignment is by visually comparing the texts in question. In today's digital age, such a task could have been automated long ago, in much the same way plagiarism checking is built into every course management platform.

Because existing automated handwriting authentication methods come with significant usability challenges, handwriting authentication remains a largely manual task. For example, some methods are based on special pens that are unavailable or not easily accessible on the open market (e.g., see [28]); others are entirely built for touch sensitive devices and thus cannot be used for traditional pen and paper settings (e.g., see [30]); while another sizable chunk relies on intricate image processing techniques, calling for a cumbersome document scanning step to digitize the text before any processing can be done (e.g., see [18]). Imagine a high school teacher faced with the challenge of scanning the daily homework of 30+ students for the sole purpose of verification.

In this paper, we tap into the burgeoning trend of fitness trackers, watches and other sensor-enabled wrist-worn devices to front *Handwriting Watcher*. This is an exciting new mechanism through which the motion of a wrist-worn sensor-enabled device, such as a smart watch, is used as a basis to authenticate (or to "*watch*") a user's handwriting. Relative to existing handwriting authentication approaches, our method offers a number of advantages that include:

(1) *Non intrusiveness* — Just like the bulk of apps on the mobile markets, our method only requires an end user wearing the device to download and run the corresponding app; imposing little or no extra operational responsibilities on the writer or the verifier. This attribute sets our method apart from those requiring document scanning and is of particular benefit in pen and paper scenarios where the handwriting authentication needs to be done on a large scale (see examples in Section 6),

(2) *Ubiquity and multi-purpose nature of wrist-worn devices* — Wrist worn devices have already taken up the lion's share of the wearables market, prompting some analysts to declare "the predicted wearables boom to be all about the wrist [8]". The fact that our method is built upon a fast growing technology increases the chances of breaking into the mainstream market, a feat that no other automated handwriting authentication method has achieved. Further, the fact that wrist-worn devices are being used for a wide range of applications (e.g., traditional time measurement, fitness tracking, sleep monitoring, etc), increases the utility that users can derive from these devices. The multipurpose nature of these accessories means that handwriting authentication would become one of many uses for a device already selling for as little as $30 [6].

[1] We refer to manual handwriting authentication in the context of free text and not signatures.

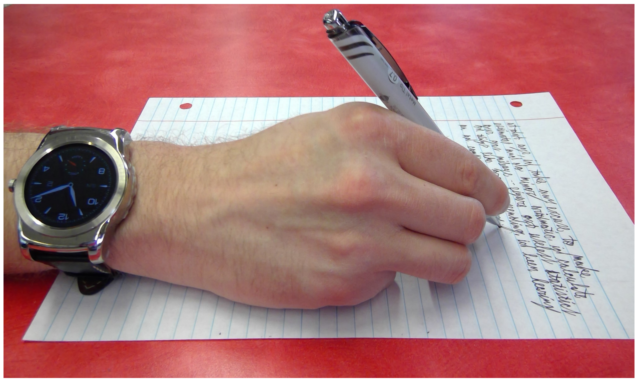


Figure 1: A user in the middle of writing a paragraph. As the fingers and pen meticulously trace out the strokes forming each character, the watch, largely constrained by the table-top, undergoes limited motion relative to the pen tip. We investigate the question of whether these subdued movements could capture the user-specific traits exhibited at the much more mobile pen tip.

(3) *Versatility* — Our method is not tailored to a particular writing surface. Although our research was specific to writing on a piece of paper, as long as one wears the device it could be trained for a different writing surface, such as a touch screen.

To the best of our knowledge ours is the first paper to investigate how a wrist-worn device could be leveraged to authenticate a user's free handwriting. The only other work that used data from a similar device to authenticate users' writing patterns was the paper by Nassi *et al*. [25]. However, the written text studied were *signatures*, which by their static nature pose a significantly different pattern precognition problem from that seen with the freely written text studied in this paper (see detailed distinction between the two papers as well as other lines of related work in Section 3).

The contributions of our paper are summarized below:

1. We designed *Handwriting Watcher*, a mechanism through which an individual's free handwriting can be authenticated based on motion sensor data collected from a wrist-worn device. In our best performing configuration, we find the mechanism to have a mean Equal Error Rate (EER) of 6.56%, with certain users having EERs as low as 0%. These numbers point to the potential of our method to serve as part of a multi-modal writer authentication scheme (e.g., the method could be used as a trigger for other more intrusive schemes).

2. We rigorously evaluate *Handwriting Watcher* under a number of settings, including those where: (1) users freely compose text in response to a range of questions (i.e., pauses due to cognitive load impact the writing process), (2) users copy text from different kinds of transcripts, and, (3) training and testing is done based on data collected over several weeks. These wide ranging investigations provide insights into the potential of our method for different kinds of application scenarios.

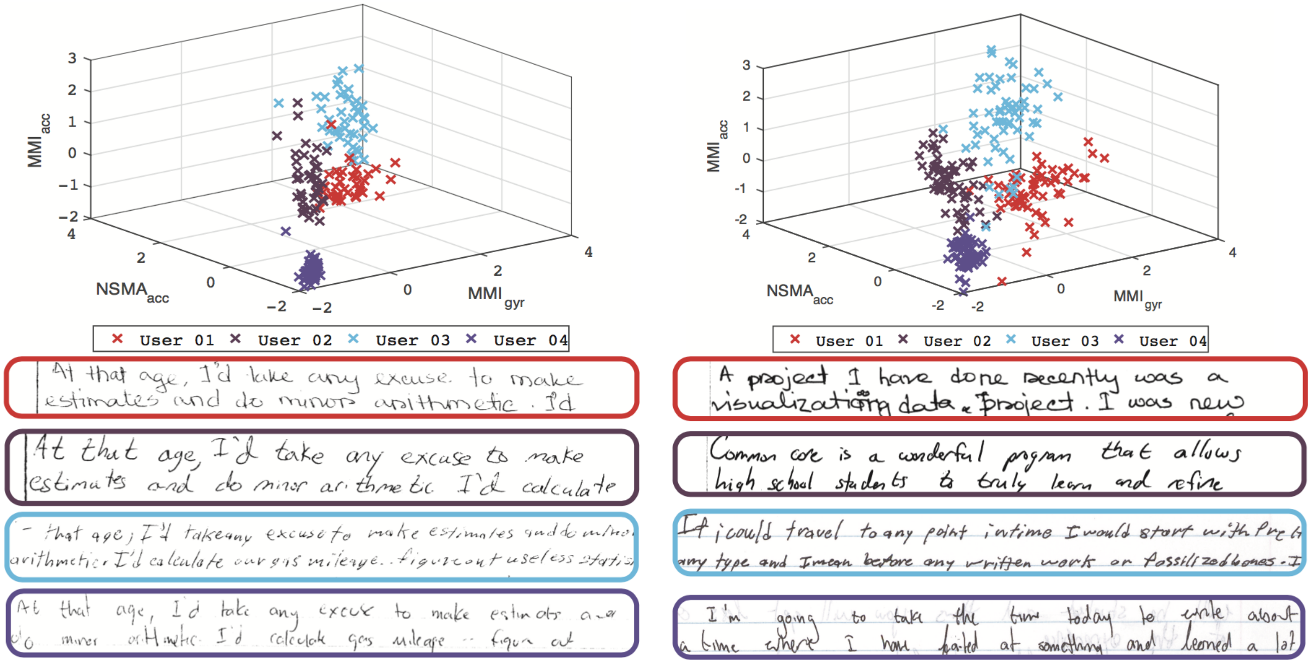
3. To drive our method, we extract a large body of 364 features and evaluate their performance before selecting a compact set of highly performing features. Findings on the performance of the various features will guide future research in the realm of wearables-driven handwriting authentication.

Road-map: The rest of the paper is organized as follows. In Section 2, we present the basic intuition behind the potential discriminative ability of wrist movements. We present related work in Section 3, describe our data collection experiments and machine learning framework in Section 4 and present the authentication results in Section 5. We finally discuss a potential application of our method in Section 6.

2. Building the intuition — Can wrist movement patterns capture the discriminative information in an individual's handwriting?

Relative to the pen (or pen tip), the wrist sees a much smaller amount of movement during writing (see Figure 1 for visual perspective of how pen location (and by extension, motion) relates to watch location and motion). This difference in motion dynamics between the pen tip and wrist raises questions on whether wrist movement patterns could capture the discriminative information embedded in the writer's handwriting. For example, let us consider an individual who exhibits highly discriminative behavior during the execution of the strokes and curves making up different characters; a sensor built into the pen would (at least theoretically) capture the vast majority of this information, while a wrist-worn device, due to the limited amount of movement transmitted to it, would capture only a small fraction of this information. A question then arises of whether the amount of information transmitted to this device is adequate to capture the general attributes unique to the individual's handwriting.

Before venturing into building machine learning models of users' handwriting, we undertook preliminary analysis to gain insight into the above question. Specifically, to gain visual understanding of how users relate to each other in these low dimensional (2D and 3D) feature spaces, we made plots of features in pairs and in triplets. If certain features can be found to discriminate between users in these low dimensional feature spaces, then it is plausible that a larger set of



At that age, I'd take any excuse to make estimates and do minor arithmetic. I'd

At that age, I'd take any excuse to make estimates and do minor arithmetic I'd calculate

- that age, I'd takeany excuse to make estimates and do minor arithmetic. I'd calculate our gas mileage. . figure out useless statis

At that age, I'd take any excuse to make estimates and do minor arithmetic. I'd calculate gas mileage -- figure out

(a) All four users copied the same text

A project I have done recently was a visualizationg data project. I was new

Common core is a wonderful program that allows high school students to truly learn and refine

If i could travel to any point in time I would start with Pre h any type and I mean before any written work or fossilized bones. I

I'm going to take the time today to write about a time where I have failed at something and learned a lot

(b) Each of the four users wrote their own unique text

Figure 2: Preliminary exploration of whether smartwatch movement patterns could contain enough information to separate writers. Based on just three features from our larger feature-set, the plots reveal a clear delineation between a group of four users regardless of whether the users copied from a common transcript or wrote independent texts in response to a series of questions. In Figure (2a), each user had fewer samples than in Figure (2b) due to variations in the amount of text copied. Regardless of sample size variations however, the separability between the users is clear.

carefully selected features could enable handwriting recognition with reasonable accuracy.

Figure 2 shows some of our findings from this preliminary analysis. The three features represented in the figure are the Normalized Signal Magnitude Area (NSMA) for the accelerometer and the Mean Movement Intensity (MMI) for the accelerometer and gyroscope (A complete description of our full feature-set can be found in Section 4). In Figure (2a), these features were extracted from data generated by a set of users who wrote exactly the same text (see snippets of the text below the figure). Figure (2b) on the other hand was drawn based on data collected while each user wrote unique text in response to an open ended question (see snippets of text below the figure). Both figures are based on a fixed subset of 4 users. For each user, the corresponding text snippet below the figure is bounded with a color matching the user's data in the 3D plots.

The figures reveal a clear separation between the four users. In particular, Figure (2a) reveals that even when the individuals write exactly the same text, the distinction persists, implying the existence of a pattern independent of the text in question. Similar traits were observed with different features and other user groups (not shown here due to space limitations). With these promising preliminary results, we were motivated to extend our work to the classification process on the full dataset and a large feature-set (Complete classification results presented in Section 5).

3. Related Work

Handwriting authentication using wrist-worn devices: The only other study on authenticating users based on writing patterns, using data from a *wrist-worn* sensor-enabled device, was the recent work by Nassi *et al.* [25]. In that paper, a Microsoft Band [7] gathered accelerometer and gyroscope data from users as they wrote their *signatures*. Using Dynamic Time Warping (DTW), the authors extracted features from the signature samples and evaluated the performance of a range of classification techniques that included logistic regression, nave Bayes, random forests, and neural networks (as implemented in Weka with default parameters). The lowest EER of about 0.05 was obtained with the logistic regression classifier. Nassi *et al.*'s work provides solid evidence of data from the sensors embedded in wrist-worn devices being able to capture a user's signature with high accuracy. *A critical question that is not tackled by their work however is that of whether a user's*

Accelerometer	ReliefF Score	Gyroscope	ReliefF Score
absolute difference of the roll	0.1627	entropy of x	0.0785
mean of the roll	0.1295	mean magnitude (mag)	0.077
entropy of y	0.085	mean movement intensity	0.077
max of power spectral density of the mag	0.0808	normalized signal magnitude area	0.0743
std dev of power spectral density of the mag	0.0797	area under the curve of the mag	0.0732
std dev of power of the mag	0.0796	mean of the second eigenvector	0.0688
std dev of FFT of the mag	0.0796	entropy of the mag	0.0687
max of power of the mag	0.0793	std dev of x	0.0681
area under the curve of the mag	0.0774	std deviation of gradient of x	0.0664
normalized signal magnitude area	0.0755	std dev of power spectral density of the mag	0.0654
absolute difference of x	0.0732	variance of the third eigenvector	0.0649
mean movement intensity	0.0732	max power spectral density of the mag	0.0644
mean of the mag	0.0732	std dev of power of the mag	0.0643
entropy of z	0.0699	std dev of FFT of the mag	0.0643
area under the curve of x	0.0692	variance of the second eigenvector	0.0639

Table 1: The top performing features extracted from the accelerometer and gyroscope data. The features for each data source are ranked according to their ReliefF score.

free handwriting could produce a similarly stable pattern based on data from the sensors in a wrist-worn device.

A signature, by virtue of being centered on a fixed set of meticulously shaped and highly practiced characters, is quite stable and easily recognized through modeling and mapping aspects of the character shapes (e.g., through DTW in [25]). Free handwriting on the other hand is highly unstable relative to a signature, and can exhibit significant variations even when two samples of the same text written by a user are compared. Where handwriting patterns from two different texts are to be compared, the distinction from a signature is even more pronounced since the motion sensor data collected from the two instances maps to two completely different character sequences. In light of these differences, the signature recognition problem in the study by Nassi *et al.* [25] provides little or no information on whether a user's free handwriting, as encoded in terms of motion sensor data generated by a wrist-worn device, could be used to identify a user.

Our paper provides the first results on the performance of smart-watch driven, handwriting based, user authentication for a wide range of writing scenarios. This study potentially paves the way for a host of new applications (See example application in Section 6).

Other past research relating to our work includes: (1) handwriting recognition research in both sensor independent settings (i.e., traditional pen and paper settings) and sensor-augmented settings (i.e., settings involving touch-sensitive screens and sensor-augmented pens), and (2) general research on the use of wearable devices for authenticating users based on various behavioral patterns. Next, we briefly discuss how these lines of past research relate to our work.

Sensor-independent and sensor-oriented handwriting recognition: Sensor-independent handwriting research (e.g., see [14, 26, 29]) involves the application of image processing techniques to segment and match characters or shapes or both. Rather than relying on image processing, our method uses accelerometer and gyroscope readings as the only inputs. This difference in the nature of inputs places our work apart from other methods; particularly when considering the operational dynamics of the authentication system and the application scenarios of the methods.

The cluster of sensor-oriented handwriting recognition research possesses more similarities with our technique because they also use sensor data instead of static images of the written text. One subset of studies in this line of work used sensor-oriented pens that capture attributes such as the x,y components of the force, and the pressure at the tip of the pen at different time instants during writing (e.g., see Crane *et al.* [15] and Gruber *et al.* [17]). The other subset uses touch-sensitive devices (e.g., tablets) to capture similar information except with x,y coordinates instead of force vectors and all the sensing built into the writing surface as opposed to the pen (e.g., see Lee *et al.* [22], Muramastu *et al.* [24] and Yi *et al.* [30]). A key distinction between our research and these two families of research is that our method relies on *the movement dynamics of the writer's wrist*; a location that is not only a significant distance from the tip of the pen, but also undergoes very subtle movement that might not manifest as significant and consistent variations as the user writes out different characters. This is in stark contrast to previous sensor-oriented handwriting recognition research where the pen or touch screen provided precise information on the spatiotemporal attributes of each character and associated curves, enabling the use of meth-

Classifier	E1		E2		E3	
	Inter	Intra	Inter	Intra	Inter	Intra
SVM	9.25 (5.60)	9.46 (6.84)	11.30 (7.02)	7.12 (4.23)	16.54 (10.59)	9.44 (7.51)
MLP	8.40 (5.85)	8.85 (6.55)	11.71 (7.29)	6.62 (3.58)	17.68 (11.52)	10.04 (7.15)

Table 2: A comparison of mean and standard deviation EER values by experiment and classifier. EERs are displayed as percentages on the [0,100] interval. The format is mean (standard deviation).

ods which rely on matching the shapes of the characters. Due to the absence of this precise information, our method does not involve any shape matching and instead authenticates users' handwriting based on the general movement pattern exhibited by the user's wrist.

General research on the use of wearables for user authentication: The last lines of work that relate to our research are the studies which have used data collected from various body-worn devices to authenticate users. For example, Shrestha *et al.* [27] and Kumar *et al.* [21] used accelerometer and gyroscope data collected from a wrist-worn smart-watch to authenticate users based on their hand movement patterns while they walked. These two works are a variant of a larger body of gait recognition research which has used sensors located at different parts of the body to authenticate users based on their walking patterns (e.g., see Mantyjarvi *et al.* [23], Alvarez-Alvarez *et al.* [13], and Derawi *et al.* [16]). Although this body of work uses the same kinds of sensors we made use of, none have tackled the handwriting recognition problem.

4. Data Gathering and Processing

Following IRB approval; we conducted three experiments (E1, E2, E3) to test how effective writing is for user authentication in a variety of scenarios. All experiments involved two sessions on separate days. For E1, all subjects copied the same excerpt from a Jeff Bezos graduation speech. E2 was a slight variation on this, requiring all participants to copy a unique and randomly assigned page of text. E3 imitated real world conditions by requiring written responses to unique (across sessions) questions. Before collecting data in E3, subjects were allowed to see the questions and take some notes on what they wanted to write about.

E1 comprised 20 subjects, while E2 and E3 had 21. Subjects were all students and staff at our university, most were undergraduate and graduate students, and 80% were male. They were informed that they would be copying text and providing written answers to questions during the experiments. After agreeing to participate, each subject was given a unique and anonymous ID.

4.1. Recording Devices

Many wearable devices have a limited API, so it is difficult to retrieve raw values for the accelerometer or gyroscope sensor measurements. This makes it hard to extract features capable of identifying a particular user. An Android Wear device was chosen because it provided unrestricted access to sensor data at a 100 Hz rate. The Android Wear transmitted linear acceleration[2], acceleration, gyroscope, and magnetometer measurements across three axis (x, y, and z). For this paper we used the linear acceleration and gyroscopic measurements. We chose linear acceleration over acceleration measurements which included gravity because we wanted to isolate the motion dynamics of the wrist.

4.2. Feature Extraction and Selection

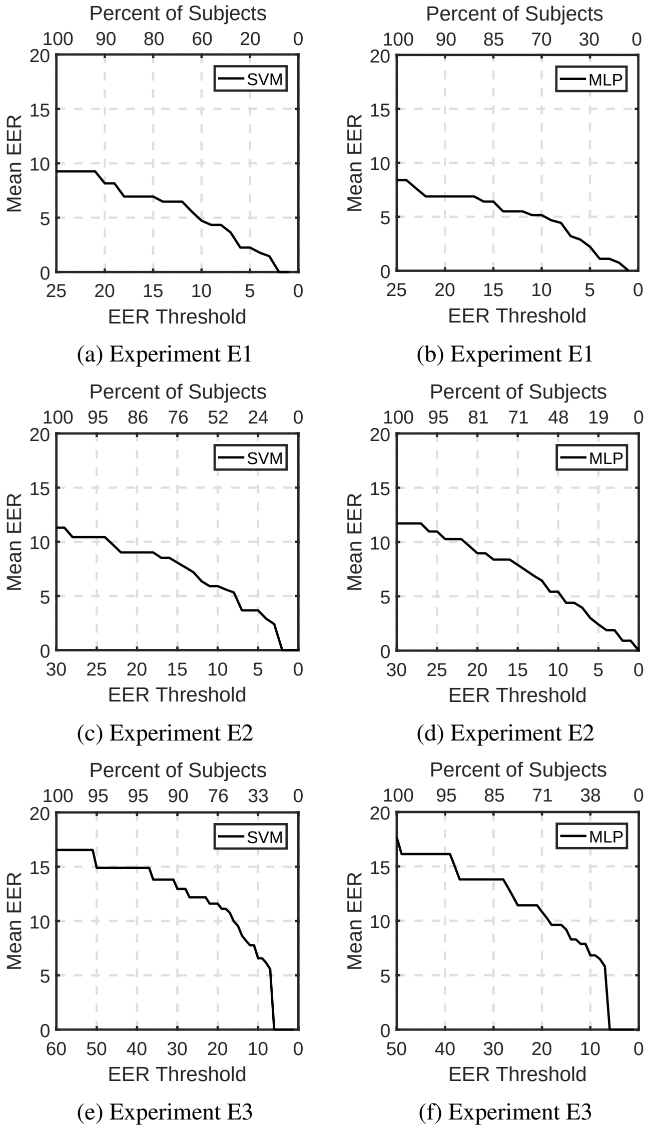
Given raw data from the accelerometer and gyroscope sensors, we used a lowpass filter to smooth the data and reduce the influence of outliers or inconsistencies. We then selected 30 second sliding windows with a 50% overlap to select chunks of raw data for feature extraction. The window size was chosen based on experimentation, as is discussed in Section 5.3.

We extracted 364 features from the accelerometer and gyroscope data produced by the smart-watch. Based on the x, y, and z components of the given sensor we calculate the magnitude. This gives rise to four vectors over a sliding window. We refer to these four vectors as Type A feature sources.

We also create three feature vectors known as Type B feature sources, two are the Fast Fourier Transform (FFT) and Power Spectral Density (PSD). The third is the squared magnitude of each FFT coefficient, which we call power. All Type B feature sources are calculated for each Type A feature source. For example, we compute the PSD (Type B) of the x component (Type A).

Type C feature sources are RFFT, DCT, DST, gradient, and eigenvectors; all generated for each of the x, y, and z sensor data components in a window. Eigenvectors were computed as follows: first, we v-stacked the x, y, and z components of the sensor data in a window to form a matrix

[2] According to the Android API specifications, linear acceleration is the accelerometer measurements excluding gravity.



(a) Experiment E1

(b) Experiment E1

(c) Experiment E2

(d) Experiment E2

(e) Experiment E3

(f) Experiment E3

Figure 3: User level performance and how it impacts population performance across all experiments and classifiers. This helps see a more granular representation of how effective the system will be for the average user.

M, second, we took the covariance of M and calculated its eigenvectors.

Type D feature sources were pitch and roll. Note that Type A, B, C, and D feature sources are all vectors that we use to extract features. Next we describe the features we extracted from these feature sources.

We calculated the mean, variance, and standard deviation of all feature source types. Min and max were calculated for power, PSD, gradient, and the eigenvectors. Mean and variance of movement intensity (calculated with x, y, and z) provided two features. The absolute difference, zero crossing rate, and root mean square were calculated for Type A feature sources, pitch, and roll. From Type A feature sources we calculated the skew, kurtosis, area under the curve (AUC), and information entropy. The combination of Type A feature sources was used to calculate normalized signal magnitude area. We also selected the top five FFT coefficients for the x, y, and z sensor data components. Finally, the correlation between the (x, y), (x, z), and (y, z) component pairs were also computed. This diverse set of 182 features was calculated for both the accelerometer and gyroscope.

After extracting features we used the Scikit-Learn StandardScaler with default settings for feature normalization. To cut down this large set of features to a compact and highly discriminative subset, applied the ReliefF feature ranking method and retained the top 60 features [20]. Table 1 shows 30 of the best features (15 for each data source) and their ReliefF scores. We only show 30 features due to space limitations.

4.3. Training and Testing Details

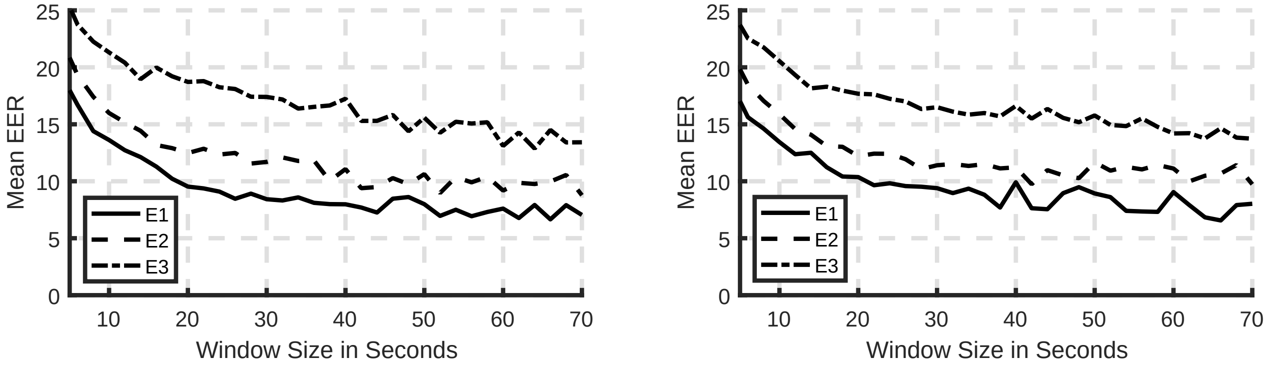
We sampled a number of classifiers in the process of reaching our results. However, only the two top performing classifiers are utilized in this paper due to space constraints. One of the two was a Support Vector Machine (SVM) with a linear kernel, C value of 0.03, and a balanced class weight. Grid-search was used to find the optimal values for C and the other hyper-parameters. The second classifier was a Multilayer Perceptron (MLP) with 60 neurons in the first layer and 30 in the second. Layer size and the number of layers was determined through experimentation. Implementations of each classifier came from scikit-learn [10], an accessible industry standard tool.

The training data for each model was a subset of the total training data. Specifically, we used a 10 to 1 ratio of imposter to authentic user samples. That ratio of imposters was chosen after running experiments to determine what provided the best results. For the testing set we randomly selected an equal number of imposter and authentic users from the testing data for verification. The probabilistic measures of certainty were used to calculate the EER for each user and a overall EER was calculated as an average of all users.

5. Experimental Results

5.1. Mean Performance Across the Experiments

Table 2 summarizes the intra-session and inter-session results for each classifier and experiment. The SVM and MLP classifiers had similar EERs between all sessions, with the largest difference being for E3 where the mean EER is 1.14 higher for the MLP. The MLP achieved the best results for E1 with a mean EER of 8.40%, while the SVM had the



(a) MLP Classifier

(b) SVM Classifier

Figure 4: Graphs showing the change in mean EER according to window size. Overall there is a steady decrease in the mean EER as window size increases.

best result for E2 and E3 with mean EERs of 11.30% and 16.54%. Upon further examination, the significant difference in results between E1 or E2 and E3 is due to a small number of users with high EERs. Section 5.2 delves into the user level results further and their impact on the findings reported in Table 2.

We considered a number of reasons for the differences in results between different experiments and the most plausible are as follows. Less variation in features is seen between training and testing sessions when the content is the same, rather than unique as in E2 and E3. We believe this explains most of the difference between E1 and E2 results. The large gap between E1 or E2 and E3 can be attributed to additional inconsistent pauses that occurred while the subjects were thinking about what to write. While some students wrote without hardly ever pausing, others would pause for 10 or more seconds at a time. Subjects also occasionally got stuck on a question during one of the two sessions but not the other. This would cause even greater variation in the extracted feature vector for the session in which the long pause took place. These trends prompt us to hypothesize that augmenting our methods with an algorithm that detects and filters out pauses might significantly improve performance.

5.2. User-Level Performance

It's useful to see the distribution of average EERs per user and how they impacted the overall mean. This analysis, shown in Figure 3, is based on the results derived with 30 second sliding windows.

In Figure 3, the bottom x-axis represents the EER threshold, below which a certain percentage of the subject population (given by top x-axis) falls. For example in Figure 3b, 70% of subjects have an EER of below 10%. This subpopulation has a mean EER of nearly 5%. Figure 3 shows that for all three experiments and both classifiers, a *small* subset of users with bad results significantly increased the average EER.

Further examination of Figure 3 reveals that E3 was the most significantly impacted by a few subjects with high EERs. While the worst 30% of subjects increase the mean EER for E1 and E2 by 2-4%, for E3 a similar number of subjects with the worst EERs contributes 5% to the mean population EER. As discussed in Section 5.1, we note that an algorithm detecting or managing the dynamics of pauses might improve authentication for these under-performing users. We believe this will most heavily impact E3, where users were most likely to pause for extended periods of time.

5.3. Impact of Window Size

In Figure 4 we demonstrate the impact of different W_s lengths on the mean EER. Both graphs stop at 70 seconds because we began to run low on training samples per user. What can be seen is that the mean EER decreases as the window size (W_s) increases.

Every experiment improved by 2-4% at a higher W_s as compared to a 30 second interval. For E1 this meant going from a mean EER of around 9% at a 30 second window to 7% at a 70 second window. A larger impact was demonstrated for E3, dropping the mean EER from between 16.5% and 17.5% to slightly more than 13. The impact of a larger W_s was greater for the MLP classifier than for the SVM, however this is only significant for E1 and E2. In fact, for the E2 results, the mean EER stops declining around the 42s mark. Overall, we believe that larger W_s lengths were more effective because larger collections of data enabled the computation of features that more accurately represent the unique characteristics of the writing style rather than the content.

The downside of a larger W_s is that data must be consistent and uninterrupted the entire window and it reduces the number of ”attempts” that the user has over a limited period of time. So although we chose W_s of size 30 seconds for our final results in Table 2, one might choose a different W_s depending on the application in question. For example, in applications where authentication is not real-time it might be more appropriate to use a larger W_s that takes advantage of more predictive features. While real-time authentication scenarios might require shorter window lengths to account for faster authentication time requirements.

6. Example Application of our Method

Our method can be leveraged to add a layer of trust to exam proctoring tools in scenarios where such mechanisms are susceptible to abuse or students have high motivation to cheat. The standardized tests (such as GRE [2], TOEFL [12], ACT [1], and SAT [9]) used in the admission process for graduate and undergraduate study in the USA are examples of exams fitting this classification[3]. To administer these exams, the US Education Testing Service (ETS) contracts proctors in almost all countries of the world. The integrity of the exams depends on the rigor with which these third party proctors apply the user verification mechanisms specified by the ETS. Currently, the primary mechanism of user verification in these exams is the picture ID presented by the student. A video recording of the exam process may be added; this provides the option of later verification by the ETS.

As evidenced from the plethora of cases where students have successfully used fake IDs to take these exams (e.g., see [3, 5, 4]), user ID-based verification is far from foolproof. If proctors collude with the students (e.g., see [11]), even video based verification could easily be manipulated through still images presented to the web cam or surveillance cameras aligned to record an entity other than the student legitimately taking the exam.

Combining smartwatch-based handwriting recognition with either or both of the above described authentication methods creates a multi-modal authentication mechanism that is not easily circumvented. Handwriting verification is a secondary modality that can be investigated in real time or used for later verification by ETS. This would provide additional confidence that all segments of the writing sample belong to the genuine user. In cases where a video recording (via web cam) of the user’s face is used as an additional layer of security, handwriting mismatches can be a part of the suite of real-time tests that can trigger closer evaluation of video segments for potential violations.

To combat the threat of proctors who might collude with the students, the data to be used for training the handwriting recognition algorithms could be collected outside of the final exams. For example, such data could be collected at an alternate proctoring location, or from high schools or undergraduate institutions attended by the students. This data collection could be gathered over an extended period of time, taking place weeks or months before the standardized tests are given. In general, because students generate handwriting samples routinely during classroom activities, the ten to fifteen minutes of data needed for training would allow for many different unobtrusive avenues of collection.

The ideas discussed above also apply to online education — e.g., with the aid of our method, online education platforms could provide the professor with some measure of authentication of handwritten assignments. Further interventions would then be based on this measure (e.g., more intrusive modalities that give higher recognition accuracies could come into play when a certain threshold is not met by a student’s submission).

7. Conclusion

Our objective with this research has been to contribute a method of text-independent user authentication that relies on readily available hardware. *Handwriting Watcher* is a demonstration that this is possible using sensor data from a wrist-worn device. For applications ranging from in-class exams to mobile banking, we believe that this form of writing authentication is easier to implement than many other previously proposed methods.

Our experiments were designed to differentiate between different types of writing scenarios and show how the viability of the system was impacted by users responding to prompts rather than copying text. To model the *motion dynamics* of writing we selected 60 features with the highest ReliefF score. Training a SVM and MLP on these 60 features resulted in mean EERs of as low as 6.56%. As mentioned in Section 5.1, we will be exploring ways of increasing performance by reducing the impact of pauses while a user is writing. We will also investigate other classification and feature engineering techniques that might be able to better utilize the time-series nature of our data.

8. Acknowledgment

This research was supported by National Science Foundation Award Number: 1527795.

References

[1] The ACT test for students and parents. `http://www.act.org/content/act/en/products-and-services/the-act.html`. Last accessed in April, 2017.

[2] The GRE tests. `https://www.ets.org/gre/`. Last accessed in March, 2017.

[3] The SAT and ACT essay portions are entirely paper-based while the GRE and TOEFL exams have paper and computer-based versions.

[3] How sophisticated test scams from china are making their way into the u.s. https://www.theatlantic.com/education/archive/2016/03/how-sophisticated-test-scams-from-china-are-making-their-way-into-the-us/474474/. Last accessed in March, 2017.

[4] Indicted for cheating. https://www.insidehighered.com/news/2015/05/29/chinese-nationals-indicted-elaborate-cheating-scheme-standardized-admissions-tests. Last accessed in March, 2017.

[5] Just how common is SAT cheating? http://schoolsofthought.blogs.cnn.com/2011/12/16/just-how-common-is-sat-cheating/. Last accessed in March, 2017.

[6] KOOGOGO u8 bluetooth smart watch. https://www.amazon.com/gp/offer-listing/B06XVSGXZF/ref=dp_olp_0?ie=UTF8&condition=all&qid=1492477985&sr=1-1. Last accessed in April, 2017.

[7] Microsoft band. https://www.microsoft.com/microsoft-band/en-us. Last accessed in March, 2017.

[8] The predicted wearables boom is all about the wrist. https://www.statista.com/chart/3370/wearable-device-forecast/. Last accessed in April, 2017.

[9] SAT suite of assessments. https://collegereadiness.collegeboard.org/sat?navId=www-sat. Last accessed in March, 2017.

[10] scikit-learn. http://scikit-learn.org/stable/. Last accessed in April, 2017.

[11] TOEFL fraud could be tip of the iceberg. http://timesofindia.indiatimes.com/city/hyderabad/Toefl-fraud-could-be-tip-of-the-iceberg/articleshow/46114710.cms. Last accessed in April, 2017.

[12] The TOEFL test. https://www.ets.org/toefl. Last accessed in March, 2017.

[13] A. Alvarez-Alvarez, G. Trivino, and O. Cordon. Human gait modeling using a genetic fuzzy finite state machine. *IEEE Trans. on Fuzzy Sys.*, 20(2):205–223, 2012.

[14] M. Bulacu and L. Schomaker. Text-independent writer identification and verification using textural and allographic features. *IEEE transactions on pattern analysis and machine intelligence*, 29(4), 2007.

[15] H. D. Crane and J. S. Ostrem. Automatic signature verification using a three-axis force-sensitive pen. *IEEE Trans. on Sys., Man, and Cyber.*, (3):329–337, 1983.

[16] M. O. Derawi, C. Nickel, P. Bours, and C. Busch. Unobtrusive user-authentication on mobile phones using biometric gait recognition. In *Intell. Inf. Hiding and Multim. Sig. Proc. (IIH-MSP), 2010 Sixth Int. Conf.*, pages 306–311. IEEE, 2010.

[17] C. Gruber, T. Gruber, S. Krinninger, and B. Sick. Online signature verification with support vector machines based on lcss kernel functions. *IEEE Trans. on Sys., Man, and Cyber., Part B (Cyber.)*, 40(4):1088–1100, 2010.

[18] A. Hamadene and Y. Chibani. One-class writer-independent offline signature verification using feature dissimilarity thresholding. *IEEE Transactions on Information Forensics and Security*, 11(6):1226–1238, 2016.

[19] A. Humm, J. Hennebert, and R. Ingold. Combined handwriting and speech modalities for user authentication. *IEEE Trans. on Sys., Man, and Cyber.-Part A: Sys. and Hum.*, 39(1):25–35, 2009.

[20] I. Kononenko, E. Šimec, and M. Robnik-Šikonja. Overcoming the myopia of inductive learning algorithms with RELIEFF. *Applied Intelligence*, 7(1):39–55, 1997.

[21] R. Kumar, V. V. Phoha, and R. Raina. Authenticating users through their arm movement patterns. *arXiv preprint arXiv:1603.02211*, 2016.

[22] L. L. Lee, T. Berger, and E. Aviczer. Reliable online human signature verification systems. *IEEE Trans. on Pat. Analys. and Mach. Intell.*, 18(6):643–647, 1996.

[23] J. Mantyjarvi, M. Lindholm, E. Vildjiounaite, S.-M. Makela, and H. Ailisto. Identifying users of portable devices from gait pattern with accelerometers. In *Acous., Speech, and Sig. Proc., 2005. Proceedings.(ICASSP'05). IEEE Int. Conf. on*, volume 2, pages ii–973. IEEE, 2005.

[24] D. Muramatsu, M. Kondo, M. Sasaki, S. Tachibana, and T. Matsumoto. A markov chain monte carlo algorithm for bayesian dynamic signature verification. *IEEE Trans. on Inf. Foren. and Sec.*, 1(1):22–34, 2006.

[25] B. Nassi, A. Levy, Y. Elovici, and E. Shmueli. Handwritten signature verification using hand-worn devices. *arXiv preprint arXiv:1612.06305*, 2016.

[26] L. Schomaker and M. Bulacu. Automatic writer identification using connected-component contours and edge-based features of uppercase western script. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(6):787–798, 2004.

[27] B. Shrestha, M. Mohamed, and N. Saxena. Walk-unlock: Zero-interaction authentication protected with multi-modal gait biometrics. *arXiv preprint arXiv:1605.00766*, 2016.

[28] B. L. Van, S. Garcia-Salicetti, and B. Dorizzi. On using the viterbi path along with hmm likelihood information for online signature verification. *IEEE Trans. on Sys., Man, and Cyber., Part B (Cyber.)*, 37(5):1237–1247, 2007.

[29] X. Wu, Y. Tang, and W. Bu. Offline text-independent writer identification based on scale invariant feature transform. *IEEE Transactions on Information Forensics and Security*, 9(3):526–536, 2014.

[30] J. Yi, C. Lee, and J. Kim. Online signature verification using temporal shift estimated by the phase of gabor filter. *IEEE Trans. on Sig. Proc.*, 53(2):776–783, 2005.

[31] X.-Y. Zhang, G.-S. Xie, C.-L. Liu, and Y. Bengio. End-to-end online writer identification with recurrent neural network. *IEEE Transactions on Human-Machine Systems*, 2016.