deWristified: Handwriting Inference Using Wrist-Based Motion Sensors Revisited

Raveen Wijewickrama
University of Texas at San Antonio
raveen.wijewickrama@utsa.edu

Anindya Maiti
University of Texas at San Antonio
a.maiti@ieee.org

Murtuza Jadliwala
University of Texas at San Antonio
murtuza.jadliwala@utsa.edu

ABSTRACT

Several recent research efforts have shown that privacy of handwritten information is vulnerable to inference threats that employ zero-permission motion sensors commonly found on wrist-wearables (e.g., smart watches and fitness bands) as information side-channels. While the adversary model in these earlier efforts have been reasonable and the proposed inference (or threat) frameworks themselves are practical and have technical merit, the related empirical evaluations suffer from several significant shortcomings, such as, use of specialized sensor hardware and highly constrained or restrictive experimental procedures, to name a few. As a result, it is hard to estimate the practical feasibility of these threats from existing research results in the literature, and thus, the extent to which end-users must be concerned about the possibility of such attacks in real-life. To answer the above question, this paper replicates some of the well-known wrist motion-based handwriting inference frameworks in the literature in order to (re)evaluate their success or accuracy in natural, unrestricted handwriting scenarios and settings by employing commercially available wrist-wearables. The results of these extensive replication and (re)evaluation studies highlight several characteristics in motion data corresponding to natural handwriting scenarios, which were either not observed or ignored by earlier efforts, and contribute to poor inference accuracy of the corresponding frameworks. In summary, accurate and practical handwriting inference using motion data (side-channeled) from consumer-grade wrist-wearables is difficult primarily due to unique and/or inconsistent handwriting behavior observed in natural writing.

Artifacts: `https://sprite.utsa.edu/art/dewristified`

ACM Reference Format:
Raveen Wijewickrama, Anindya Maiti, and Murtuza Jadliwala. 2019. deWristified: Handwriting Inference Using Wrist-Based Motion Sensors Revisited. In *12th ACM Conference on Security and Privacy in Wireless and Mobile Networks (WiSec '19), May 15–17, 2019, Miami, FL, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3317549.3319722

1 INTRODUCTION

Inference of sensitive user data by employing on-board sensors as information side-channels has been a significant privacy concern ever since the inception of commercial, consumer-grade wearable devices such as smart watches and fitness bands. Several proposals in the research literature have already demonstrated how data from zero-permission wrist-wearable sensors can be abused to infer keystrokes, user-activities, and behavior [11, 13–16, 19, 23, 25, 26, 28, 29]. In the same vein, multiple research efforts have also demonstrated the feasibility of inferring handwritten text using motion sensors (such as accelerometers and gyroscopes) present onboard these wrist-wearables. Some of the initial efforts in this direction showed the feasibility of inferring larger handwriting gestures, such as, writing on a whiteboard [6] or using hand/finger movements to write in the air [4, 5, 31]. More recent efforts have focused on inferring smaller and more natural handwriting gestures, such as, writing on a paper with pen/pencil [30]. Some of these works were presented with an adversary in mind, whereas others were presented merely as a mobile/wearable application or service. In this work, we focus on the problem of inferring handwritten text primarily from an adversarial point of view.

While these earlier research efforts concluded that their inference/classification frameworks were able to infer handwritten English letters and words from wrist-wearable motion data in an accurate and feasible manner, we observed that several of the assumptions made by them (implicitly or explicitly) were either not realistic or impossible to include in an adversarial setting. For example, Amma et al. [5] used specialized motion sensors and custom wrist-wearable hardware in their inference framework, which could sample at more than $800Hz$. However in practice, most common commercially-available smartwatch and fitness-band motion sensors have maximum (peak) sustainable sampling rates of around $200Hz$. The availability of such specialized sensors and hardware, and the extremely fine-grained motion data generated by it, may result in an accurate inference of handwriting, but would be difficult to assume in an adversarial setting where the target user is probably just wearing a consumer-grade wrist-wearable with sensors that have limited capabilities. Other limitations of some of the previous works include testing primarily in a personalized setting (training and testing data collected from the same participant), vague definition of segmentation techniques used to separate individual sentences and words, and disregard for varying writing styles of the same target user. The absence of these factors in their evaluation also gave us the impression that their data may have been collected in a tightly controlled fashion, which is not reflective of participants' natural handwriting and/or writing in a natural setting.

Motivated by these shortcomings of existing research efforts, in this paper we attempt to validate if the current empirical results on handwritten text inference using wrist-wearable motion sensors are generalizable and applicable under more practical adversarial

WiSec '19, May 15–17, 2019, Miami, FL, USA

ACM ISBN 978-1-4503-6726-4/19/05…$15.00
https://doi.org/10.1145/3317549.3319722

settings and handwriting scenarios. Broadly, our goal in this work is to evaluate if existing handwriting inference frameworks are a genuine and realistic threat to a wrist-wearable device user's privacy and security. In order to accomplish this goal in a structured fashion, we first closely replicate the four most notable inference frameworks in this direction, specifically, the ones defined by Xu et al. [31], Arduser et al. [6], Amma et al. [5], and Xia et al. [30]. Next, by means of contemporary, consumer-grade wrist-wearables, we collect natural handwriting related motion data from a large number of human subject participants in an unconstrained and non-restrictive setting for a variety of different writing scenarios. In order to showcase why it is much more difficult to infer natural handwriting, we then perform a detailed comparative analysis of our results with those obtained by previous efforts, across different writing scenarios. Our final contribution is an in-depth discussion on the factors that affect the success of handwriting inference attacks in real-life, supported by data and results obtained from our experimentation.

2 ADVERSARY MODEL AND BACKGROUND

Before outlining details of our replication and validation experiments, let us first provide an unambiguous description of the adversarial setting and capabilities assumed in this work, followed by a brief technical background of handwritten text inference. We also provide a detailed review of the four notable wrist motion-based handwriting inference schemes in the literature and summarize the key research gaps that we attempt to fill in this work.

2.1 Adversary Model

The use of hands (and gestures) to compose lingual texts continues to be one of the most prevalent methods of communication and many of these handwritten or hand-gestured text may contain sensitive information, such as personal identifiers and financial credentials. Our adversary in this work is as a malicious actor or entity whose goal is to infer such private handwritten text by employing some form of an information side-channel. While there exists several types of information side-channels, our focus in this paper is on zero-permission motion sensors (such as accelerometers and gyroscopes) found on most modern wrist-wearable devices such as fitness-trackers and smartwatches. Our adversary can gain access to this motion sensor information by tricking the victim user into installing on its wearable device a Trojan or malicious application that is masquerading as some useful application (for example, a game or a fitness application). Once this Trojan application is installed on the victim's wearable, it can sample and exfiltrate the motion sensor data back to the adversary using the device's network connection. The fact that popular wearable operating systems like Android Wear and watchOS allow third-party applications unrestricted access to back-end sensors such as gyroscope and accelerometer (thus, the term zero-permission), enables the malicious Trojan application to sense and exfiltrate this data without raising any flags or violating any system security policies. By masquerading as an application that would normally require access to these (motion) sensors for its operation, the malicious Trojan application can achieve further stealth.

Once the adversary is able to remotely archive this exfiltrated sensor data, say, on its computation server, he can analyze it in an offline fashion to infer the actual written information from the data with as high accuracy as attainable. Such an adversary model is practical and has also been commonly employed in the literature for studying similar privacy threats due to zero permission mobile and wearable device sensors [13–15, 17, 23, 27, 28].

In this paper, we limit ourselves to the problem of inferring handwritten text in the English language only. This enables us to have a comprehensive and equitable comparison with other research efforts that have also inferred only English language written text. Additionally, we also assume that our adversary only employs the victim's hand movement data while writing, as perceptible on the victim's wrist-wearable motion sensors, for the inference attack. The adversary does not employ additional information, such as contextual dictionaries on the topic of writing and victim's language abilities, in order to improve the accuracy of the inference attacks. This is done to keep the adversary model practical and to achieve an equitable comparison with the inference frameworks being evaluated in this work. That being said, the adversary is free to use a generic language dictionary and well-known spelling correction techniques for improving handwriting inference.

2.2 Inferring Handwritten Text from Wrist Movements

Any framework for inferring handwritten text from wrist movements (or wrist motion sensor data) would ideally comprise of the following two key phases: (i) identifying the writing and non-writing parts in the sensor data stream in order to segment strokes, alphabets, words, and sentences, and (ii) using the segmented sensor data to perform character, word, or sentence recognition/classification. However, before we outline a concrete technical framework for handwriting inference from motion sensor data, let us first characterize the different writing styles and writing elements that may vary from person to person, and sometimes also for the same person. Writing in English language can be broadly categorized into the following two styles: lower versus upper case writing, and cursive versus non-cursive writing [7]. Irrespective of the style, writing in English (or any other) language comprises of drawing a series of alphabets, where each alphabet is nothing but a series of one or more strokes made with a writing apparatus such as a pen. A *stroke* can be defined as a continuous line drawn in one go, starting with a "pen-down" action and ending with a "pen-up" action. The same concept applies to all forms of writing, except that the pen can be replaced with another writing apparatus. Each handwritten alphabet can be uniquely described by the **number**, **direction** and **order** of strokes. It is possible to write the same alphabet in different cases with different number, direction and order of strokes. Even in the same case, the same alphabet can be written using a different number, direction and order of strokes. As a matter of fact, a written alphabet comprising of n strokes can have $n!$ different stroke order and directions depending on the writer. Thus, it is easy to imagine that handwriting related wrist or hand movements could differ significantly even for the same alphabet. For example, written in the same case, but with different number, direction, and/or order of strokes. These differences in movement are reflected on the motion sensors located on writer's wrist, resulting in different sensor data streams for the same alphabet. This is a significant challenge that must be overcome by any motion-based handwriting inference framework. It is worth pointing out that wrist movement-based handwriting inference is significantly different, and more challenging, than traditional image or pixel-based handwriting recognition [20, 21], primarily because image or pixel data does not include/consider information about the number, order and direction of strokes.

2.3 Previous Work and Motivation

As noted earlier, we are not the first to investigate handwriting inference threats using wrist-wearable motion sensors as an attack vector. Multiple previous research papers have demonstrated the feasibility of inferring different forms of handwritten information from motion data collected by means of wrist-wearables. We are particularly interested in the following four forms of handwriting scenarios that were evaluated earlier, primarily because they are the most commonly observed in real-life situations: (i) pen(cil) and paper writing [30], (ii) whiteboard writing [6], (iii) finger writing [31], and (iv) airwriting [3–5]. In this section, we describe the main strengths and shortcomings of these earlier research efforts, and outline our primary motivation for revisiting the problem of handwriting inference using wrist-wearables.

Pen(cil) and paper writing: Xia et al. [30] proposed an eavesdropping attack on the classical pen and paper based writing scenario using motion data recorded from a smartwatch worn on the writing hand. The threat was accomplished using wrist motion data that was sampled from the smartwatch's accelerometer and gyroscope at $200Hz$. Most modern smartwatches and fitness tracker motion sensors do support this sampling rate. The authors' employed a thresholding based word-wise segmentation on the continuous accelerometer data followed by an alphabet-wise segmentation using the gyroscope signal before classifying individual alphabets. The authors' did study a generalized setting for their classification algorithm, where training and testing data from different participants was used, making it realistic because it is generally difficult (or impossible) for an adversary to collect labeled training data from the victim or target. Given the above strengths, this work also has several shortcomings. First, the proposed inference framework is limited to non-cursive handwriting, and only lowercase alphabets were evaluated with the argument that, it can be easily extended for uppercase alphabets. Moreover, the framework is also difficult to replicate and generalize to other writing scenarios and settings due to the use of fixed thresholds (during segmentation) and specific features (during alphabet inference).

Whiteboard writing: Arduser et. al [6] proposed an inference framework that employs accelerometer and gyroscope data from a target victim's smartwatch to infer text written on a whiteboard. Besides the standard learning-based alphabet classification routine, the inference framework comprised of a pre-processing routine that first converted the motion data from device coordinates to whiteboard coordinates, which eliminated the effect of watch orientation when writing on different (top or bottom) areas of the whiteboard. Moreover, the motion data was sampled from standard consumer-grade smartwatches, which shows that such threats can be executed in the wild without using any specialized hardware. This research effort also suffers from several significant shortcomings. First, the proposed inference framework was used to evaluate only uppercase alphabets. Moreover, it is unclear whether the proposed framework can be easily replicated and generalized to other writing scenarios and settings as critical parameters (and description) related to the coordinate conversion routine and employed sensor sampling rates are unavailable. Lastly, it is unclear whether the provided empirical results are for a personalized (training and testing data from the same participant) or a generalized (training and testing data from different participants) classification setting.

Finger writing: Xu et al. [31] investigated the problem of alphabet recognition (with each character approximately $2.5'' \times 2.5''$ in size) when writing using the index finger on a surface by means of a Shimmer [2] device worn on the wrist of the writing hand. One of the most significant outcome of this research effort was that the authors were able to obtain very high (more than 90%) inference accuracy for their proposed inference framework. At the same time, one of the biggest drawback of their work was the use of sophisticated Shimmer devices in the inference framework, which are not as ubiquitous and popular as commercially available smartwatches and fitness-bands. In addition to this, the proposed inference framework was used to evaluate only writing of uppercase alphabet letters. This raises serious concerns about the broad applicability and generalizability of the proposed framework to other wrist wearable hardware and writing scenarios/settings. Also, the paper neither mentions the total number of unique participants for which the proposed inference framework was evaluated, making it hard to understand the statistical significance of the obtained results, nor does it provide details on whether a personalized or a generalized setting was used for the classifier evaluation.

Airwriting: Amma et. al. [3–5] proposed an input mechanism (named *airwriting*) which detects and classifies hand writing gestures in the air using the motion sensor data collected from a specialized sensor-integrated hand glove. By comprehensively evaluating their inference frameworks in both personalized and generalized settings, the authors show that it is possible to obtain a reasonably high inference accuracy in this writing scenario/setting. However, while the use of a custom-designed hand glove capable of sampling motion sensors at $819.2Hz$ is viable for enabling novel HCI applications, such kind specialized hardware is not very popular and ubiquitous. This significantly limits the applicability of the above inference framework and the related results in an adversarial setting. Moreover, the proposed inference framework was used to evaluate only uppercase words from a dictionary as it did not include appropriate segmentation algorithms for separating out the alphabets within each word. This reliance on a dictionary for executing the inference algorithm significantly limits the type of information that can be inferred and is not very practical or realistic in an adversarial setting.

As evident from the strengths and shortcomings of each of the above research efforts, it was difficult for us to reasonably estimate whether these threats to users' handwritten information from current consumer-grade wrist-wearable sensors is practically feasible or not. And if it is, how would such an attack perform across a diverse group of users with different and unique handwriting styles? And, is there a way to develop a unified inference framework that

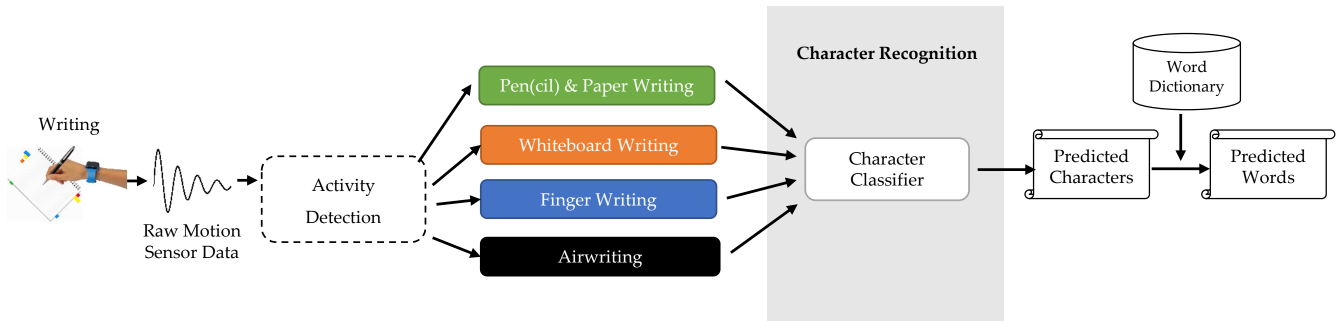


Figure 1: Generalized attack framework.

will not only work against diverse handwriting styles, but also different forms and cases of handwritten text? In order to fully understand the extent to which end-users must be concerned about the possibility of such attacks in real-life, it is paramount for us to answer these questions. As outlined earlier, we were unable to find these answers in the current research literature. In this paper, we attempt to seek these answers by closely replicating the implementation of the above four inference frameworks and re-evaluating them in realistic adversarial settings. In the next section we give details of the replicated experiments, and in Section 4 we present our evaluation results obtained using these replicated experiments.

3 EXPERIMENTAL SETUP

Due to the unavailability of publicly-available code, we replicated the inference frameworks of the four research efforts discussed above, i.e., pencil and paper writing [30], whiteboard writing [6], figure writing [31], and airwriting [4, 5], as closely as possible based on the information available in the corresponding papers. Below, we provide details of the experiments that we conducted using these replicated implementations of the inference frameworks.

3.1 Participants and Data Collection

To account for participant diversity in our experiments (with these frameworks), we recruited 28 participants aged between 18 and 30 ($\sigma = 4$) years, seven participants for each form of writing. 13 out of the 28 participants were male, and remaining 15 participants were female. All 28 participants were recruited using fliers posted around our University campus, and as a result they were from diverse demographic backgrounds. In order to test generalized inference models as an adversary would, and in-line with some of the previous works, only right-handed participants were used for the study (i.e., writing with the right hand). For the same reason, we also enforced non-cursive writing. In order to minimize bias in the collected motion sensor data, no other restrictions were imposed on the participants. Participants were not given any time limit to complete their writing tasks, and were encouraged to write using their normal or accustomed handwriting style. The entire experiment was also approved by our University's institutional review board (IRB).

3.2 Writing Scenarios

Participants in our experiments were asked to wear a smartwatch (Sony Smartwatch 3 or LG Watch Urbane, depending on the experimental scenario as detailed later) on their writing hand and perform the assigned writing tasks (as described below). Both accelerometer and gyroscope data were recorded at $200Hz$ from the smartwatch while the participants were undertaking their writing tasks. Depending on the writing scenario and setting, participants were provided with appropriate writing apparatus and environment. For example, in the pencil writing scenario participants were provided with a pencil, a chair to sit on, and positioned near a table with a sheet of paper on top of it (used as the writing surface). In the whiteboard writing scenario, participants were provided with a marker pen to write on a nearby whiteboard mounted on the wall. In the finger writing scenario, participants were provided with a chair to sit on, and positioned near a table with a touchscreen tablet computer on top of it (8″ Samsung GT-N5110 Android tablet, used as the writing surface). To match with [31], the writing area on the tablet screen was designed to be $2.5'' \times 2.5''$. Lastly, for the air-writing scenario, participants were provided with a chair to sit on, and ample free space around them to allow free movement of their arm. In all scenarios, the alphabets/words/sentence to be written were displayed on a nearby tablet screen, except in case of finger writing where the same tablet computer was used for both displaying the writing task and as the writing surface. The tablet was also used to record the ground truth alphabets/words, and additional ground truth spatial data in the finger writing scenario for in-depth empirical analysis of writing characteristics. Figure 2 show the setup of all the four writing scenarios described above.

3.3 Writing Tasks

In all of the four writing scenarios outlined above and depicted in Figure 2, participants performed the *same* set of writing tasks, where some of the subtasks were randomized in order to minimize bias in the writing activity. The design of our writing tasks was carefully undertaken so as to enable us to perform an equitable comparison of our results with the ones obtained earlier, while at the same time helping us gain more insight on the impact of different writing characteristics, settings, scenarios, etc. on the resulting inference accuracy. The writing subtasks were as follows:

(a) Pencil and Paper Writing (b) Whiteboard Writing (c) Finger Writing (d) Airwriting

Figure 2: Writing scenarios considered in our experiments.

- **Alphabets.** Participants wrote individual alphabets one at a time, covering all 26 alphabets in random order, in both upper and lower cases. Each alphabet was written 10 times, for a total of 260 alphabets written by each participant.
- **Words.** Participants wrote 4-8 alphabet words, one at a time, selected random from a vocabulary [9]. Each participant wrote 20 words, in both upper and lower cases.
- **Sentence.** Participants wrote a sentence covering all alphabets of the English language, "the five boxing wizards jump quickly", in both upper and lower cases.

In addition to the in-lab writing tasks, we also collected data from 2 participants to evaluate writing activity recognition among other daily activities. The participants wore a smartwatch for a day, and were asked to perform writing scenarios belonging to each of the above scenarios at random times during the day.

3.4 Inference Frameworks

We implemented the targeted inference frameworks [5, 6, 30, 31] primarily using Python 3.7, making use of the machine learning library *scikit-learn* wherever applicable. The specific implementation details for each of the writing scenarios are presented below.

Pencil writing: The raw sensor data was pre-processed by using Pauta Criterion [32] to eliminate outliers in the data, followed by a low pass filter with a range $1Hz$ to $25Hz$. As, our data collection already contains ground truth, no boundary detection was needed when training individual letter recognizer models. Then, for each letter sample in the preprocessed gyroscope data, entropy was calculated for each axis after a Fast Fourier Transform (FFT). For each axis, amplitudes for each frequency from 1 to $25Hz$ were computed. The remaining processing included finding peaks and valleys in the data and then computing features related to first peak and valley, last peak and the maximum peak. The resulting feature vector from this computation included a total of 115 features, which were then used to train a Random Forest classifier. Hyper-parameter tuning was done by using random search followed by a grid search to generate the best set of hyper-parameters. In the word detection phase, the top-5 letter predictions for each letter of a word were used to generate a list of letter sequences, which were then processed by a spell correcter module based on a comprehensive (100,000 words) dictionary, in order to obtain a list of corrected words. From the list of corrected words, the most frequently occurring word is selected as the final predicted word.

Whiteboard writing: The pre-processing done by Arduser et al. [6] involves converting the data from device coordinates to whiteboard coordinates. As sufficient information about this coordinate conversion was unavailable, we were unable to accurately replicate it. To compensate for this limitation and still have an equitable analysis, our whiteboard writing participants were asked to maintain a constant height when writing. The individual letters were then directly used in a Dynamic Time Warping (DTW) algorithm for character recognition. DTW has been traditionally used for time series alignment and calculation of a similarity distance between two time series [24]. A part of the collected wrist motion data (of users' handwriting) was used as templates for each letter in the alphabet, and these templates were then aligned with the query test sequences to compute the similarity distance score, where lower score would imply a better match. The results presented by Arduser et al. considered the presence of a written character in the top-3 predictions of the corresponding character as a successful inference. To compare against these results, we used the top-3 predictions for each character from DTW, i.e. the three lowest DTW scores. In the word detection phase, the audio collected alongside the motion data was used to segment letters within a word. A vector was constructed by summing the absolute amplitudes between 5 to $10KHz$ in the time series. Then, consecutive values greater than a threshold in this vector are combined, where the first value above the threshold is marked as the starting point and the possible end points are marked within a 2.5 *second* window from this start point. The corresponding accelerometer data between start and end point sequences is then used in the letter recognition, and used as constituents for word prediction.

Finger writing: Although no pre-processing step for raw data is mentioned in [31], certain individual features (from the data) were required to be computed after passing the data through a low-pass and band-pass filter. A set of features relating to motion energy, shape, posture were computed for both accelerometer and gyroscope data over all the axes. A detailed description of these features can be found in [18]. This resulted in a feature vector comprising of 46 features for each character window. These feature vectors were then used to train Naive Bayes, Logistic Regression and Decision Tree based classifier models for the character recognition tasks.

Airwriting: In this writing scenario, the raw sensor data was first normalized and then used to extract the average amplitude for each axis (from both the accelerometer and gyroscope data stream) resulting in a feature vector comprising of six features.

These features were then used to build and train a separate Hidden Markov Model (HMM) classifier [22] for each letter in the alphabet, resulting in a total of 26 models. These HMMs were implemented with left to right topology and 30 states. A Gaussian Mixture Model with six components was used to obtain observation probabilities for each state of the HMM.

4 INFERENCE ACCURACY RESULTS

In this section, we present evaluation results from our evaluation experiments with the replicated inference frameworks discussed above. First, we present inference accuracy results that were obtained in a generalized setting, followed by results obtained in a personalized setting. A summary of the inference accuracies obtained in our experiments compared with those obtained in the original papers of the four handwriting inference schemes can be found in Table 1.

4.1 Generalized Inference Accuracy

The generalized models were tested by using a Leave-One-Out-Cross-Validation (LOOCV) mechanism, where data from a single participant is used as the test set while the remaining participants' data is used as the training set for the corresponding classification model.

Pen(cil) writing: The extracted feature sets for each individual alphabets, as described in the previous section, were used in training a Random Forest classifier. Classification using the trained Random Forest classifier in the generalized setting yielded an average accuracy for lowercase alphabets of about 6% (σ = 1%) with alphabet *"l"* having the highest accuracy at 19% and all other alphabets had accuracies below 15%. The uppercase alphabets also resulted in similarly poor average accuracy of only 5% (σ = 2.0%) with alphabets *"E"* and *"L"* having the highest accuracy at 17% and 16%, respectively, and all other alphabets had accuracies below 10%. In comparison, authors of the original work [30] were able to obtain a mean accuracy of 50% (σ = 17%) for the lowercase writing scenario. Our poor alphabet-level accuracies were reflected in word prediction as well (< 1%). In comparison, authors in [30] obtained a word accuracy of about 33%.

Finger writing: The finger writing inference is done using three classifiers: Decision Tree, Naive Bayes and Logistic Regression. The classification accuracies of all the three classifiers turned out to be poor, with only 5% average accuracy (σ = 1%) for lowercase alphabets. Similar accuracies were also observed for the uppercase alphabets with 7% average accuracy (σ = 3%) for all the three classifiers. Word prediction was mostly unsuccessful (< 1% accuracy) due to the low alphabet-level inference accuracy.

Whiteboard writing: When each participant's dataset was tested against the templates taken from all other participants, an average accuracy of 20% (σ = 4%) for lowercase alphabets was observed. Alphabets *"l"* and *"z"* showed the highest accuracies at 55% and 47%, respectively, while all other alphabet accuracies were below 30%. We were able to obtain an average accuracy of around 27% (σ = 8%) for uppercase alphabets, where alphabets *A*, *H*, *L*, *M*, *N*, *W*, *Z* showed over 40% accuracy. We obtained an average accuracy of 39% (σ = 6%) for lowercase alphabets when top-3 predictions were considered, with alphabet *"o"* having an accuracy of 87%, followed by *"l"* having 75% accuracy. For uppercase alphabets, an accuracy of 44% (σ = 8%) was obtained with *"L"* having the highest accuracy at 72% followed by *M*, *N*, *H*, *W*, *V* having accuracies above 60%. It is unclear whether the results presented by Arduser et. al [6] were obtained in a generalized setting or a personalized one, but they were able to report a very high average inference accuracy of around 94%.

Airwriting: In the generalized setting, our average accuracy for the uppercase alphabets was only 9% (σ = 1%). The mean lowercase alphabet accuracy was a mere 5% (σ = 3%). The trained HMM-based character models were then concatenated and tested to infer words. The poor individual character inference accuracies were again reflected in word-level inferences, with less than 1% word inference accuracy.

4.2 Personalized Inference Accuracy

In the personalized setting, the classification models were evaluated by splitting the (motion) dataset of a participant into a training set and a testing set, and cross-validated wherever applicable.

Pen(cil) writing: The dataset of each participant was split into training and testing data using a 60:40 ratio. The alphabet inferences were poor even in the personalized setting, with only an 10% average accuracy for uppercase and 11% accuracy for lowercase alphabets. The authors in [30] do not provide any results for a personalized scenario mainly because their scheme was proposed as an attack. The poor alphabet inference accuracy was insufficient for a word inference, even with the help of a dictionary to recognize words.

Whiteboard writing: The whiteboard writing scenario was analyzed in a personalized setting by using 50% of the alphabet samples per participant as a training set (or set of templates for the DTW algorithm), while testing was performed using the remaining alphabet samples. This resulted in a 51% mean alphabet accuracy (σ = 0.13) for the lowercase alphabets, with alphabets *c*, *p*, *v*, and *z* having accuracies over 60%. We were able to obtain a 56% mean accuracy for uppercase alphabets (σ = 12%), with alphabets *B*, *I*, *M*, *N*, *S*, *Z* having accuracies over 70%. Figure 3 shows that alphabets *"n"* was often misclassified with *"h"* (and vice-versa). Similarly, letters *"i"* and *"j"* were also often misclassified with each other due to the high similarity between their strokes. Figure 4 shows that in the uppercase alphabets, *"P"* and *"D"* were often misclassified with each other along with *"U"* and *"V"*. We also tested the prediction accuracy by considering the top-3 guesses. This resulted in 73% accuracy (σ = 10%) for uppercase alphabets. The alphabets *C*, *M*, *S*, *Z* had over 80% accuracy. For lowercase alphabets, 69% accuracy (σ = 10%) was obtained, and only the alphabets *c*, *f*, *k*, *p*, *z* had accuracies over 75%. In comparison, [6] presented results only for uppercase alphabets and had a 99% accuracy within 3 guesses.

Finger writing: For this scenario, all the three classifiers were evaluated with a 60:40 train:test ratio. We were able to observe an average accuracy of around 8% for all the three classifiers (σ = 2%) for inferring lowercase alphabets. Inference of uppercase alphabets produced a slightly higher average accuracy of around 17% (σ = 10%). In comparison, Xu et. al [31] had over 85% accuracy for all the three classifiers for uppercase alphabets. Word prediction was

	Personalized				Generalized			
	Lowercase		Uppercase		Lowercase		Uppercase	
	Original Work	Our Replication	Original Work	Our Replication	Original Work	Our Replication	Original Work	Our Replication
Pencil writing (Xia et al. [30])	–	11%	–	10%	50%	6%	–	5%
Finger writing (Xu et al. [31])	–	8%	91%	17%	–	5%	–	7%
Whiteboard writing (Arduser et al. [6])	–	51%	94%	56%	–	20%	–	27%
Airwriting (Amma et al. [4, 5])	–	14%	95%	14%	–	5%	82%	9%

Table 1: Comparison of alphabet inference accuracies. Empty fields imply that the original work did not test those setting.

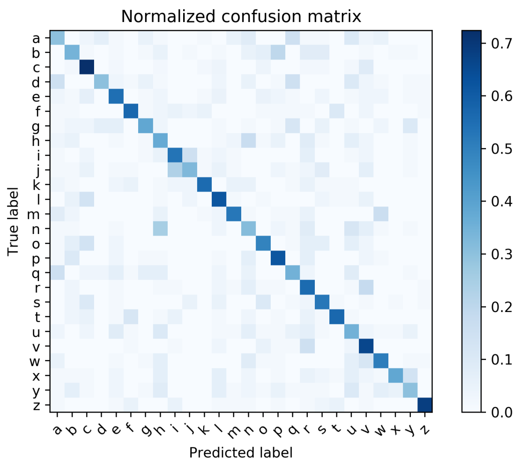


Figure 3: Confusion matrix for lowercase alphabets in whiteboard writing.

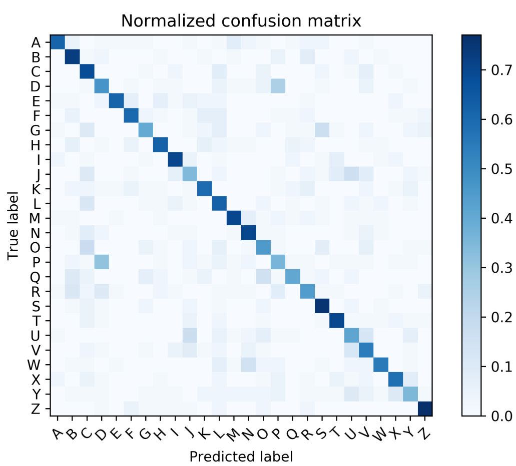


Figure 4: Confusion matrix for uppercase alphabets in whiteboard writing.

mostly unsuccessful (< 1% accuracy) in our test due to the low alphabet-level inference accuracies.

Airwriting: For this scenario, the data set for each participant was split using a 65:35 training:testing ratio. We observed an uppercase alphabet inference accuracy of around 14% (σ = 3%). Inference of lowercase alphabets also resulted in an average accuracy of around 14% (σ = 5%). The individual character HMM models were then concatenated to build word models for every word in a vocabulary of approximately 1000 words [9]. The word-level data collected per participant was then tested against these word models to predict words. The low individual character accuracies were reflected in the word-level inferences and we obtained an average word-level inference accuracy of less than < 1%.

4.3 Writing Activity Detection

Among the four handwriting schemes considered in this work, only [5] (airwriting) and [31] (finger writing) evaluated the problem of detecting writing related gestures/activities among other activities. In real-life deployment, this step is equally essential for both HCI applications and for an adversary trying to infer private handwritten data. Xu et al.[31] considered gestures relating to the index finger, the hand and the arm, and classified them using the same set of features used for alphabet inference. They obtained true positive rates of over 90%. Amma et al. [5] used a binary SVM classifier to identify the airwriting motion in a continuous motion stream and achieved a recall of 99% and a precision of 26%. As both of these prior works used specialized devices, while we only considered wrist motion data available from the smart watches (used in our experiments), it was challenging to perform an equitable comparative evaluation. We replicated the handwriting activity detection model used by Xu et al. [31], but tweaked it so that it can be used to identify any of the four writing scenarios (i.e., pencil writing, whiteboard writing, finger writing and airwriting). In a personalized setting with a user's labeled data included in the training, our activity detection model achieved around 56% recall (and 57% precision) for air and finger writing scenarios while pencil writing achieved 39% recall (47% precision). Whiteboard writing resulted in the lowest recall value at only 23% (34% precision). When considering each writing scenario against all the other writing and non-writing activities, whiteboard and finger writing resulted in over 90% recall with under 40% precision, and airwriting and pencil writing resulted in recall of 78%. A generalized testing of our activity detection model achieved around 35-40% recall for airwriting, whiteboard writing and pencil writing, whereas finger writing resulted in the lowest recall at just 8%. The whiteboard writing achieved highest precision at 65%, while the other three writing scenarios had a lower precision in the range of 20-40%.

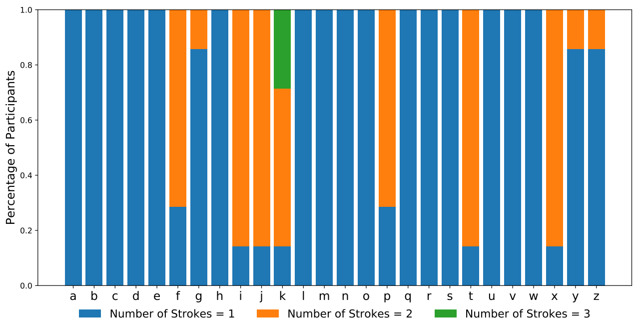


Figure 5: Number of strokes for the same letter for different participants (lowercase).

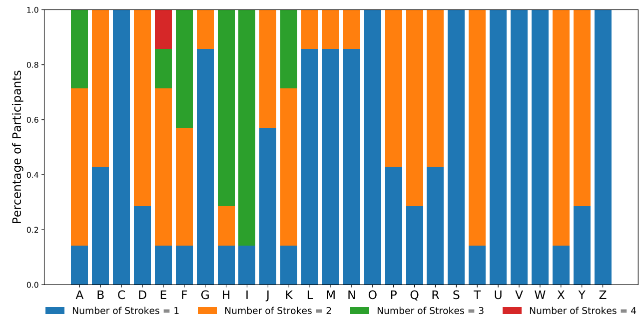


Figure 6: Number of strokes for the same letter for different participants (uppercase).

5 FACTORS AFFECTING INFERENCE ACCURACY

As evident from Section 4, our replicated experiments did not perform as well as some of the previous works. In this section, we analyze the factors that we believe were the main causes of the poor inference accuracies, which indirectly determines the practicality of such handwriting inference attacks in real-life settings.

5.1 Number of Strokes

One of the main factors that influences a person's handwriting (and thus, its inference by the described frameworks) is the number of strokes the person uses to write each alphabet of a language. During our experiments, we observed varying writing styles among different participants which effectively resulted in varying number of strokes (across participants) for writing the same alphabet. Figure 5 and 6 shows that the same alphabet is written using different number of strokes by different participants. For lowercase alphabets, we observed that *a*, *b*, *d*, *e*, *f*, *g*, *h*, *k*, *p*, *q*, *w*, *x*, *y* and *z* have varying number of strokes. For alphabet *k*, we observed some participants used just one stroke and some other participants used up to three strokes. Uppercase writing shows more variation in number of strokes compared to lowercase writing, where except for alphabets *C*, *H*, *O*, *S*, *U* and *W*, all other alphabets show variations. Notably, the alphabet *E* was written using number of strokes ranging from three to four.

We also observed that even the same participant sometimes use varying number of strokes for the same alphabet. Figure 7 shows the mean of variances for the number of strokes calculated per participants. It is observed that except for alphabets *c*, *o*, *q* and *s*, all other lowercase alphabets show some variance in the number of strokes used. In uppercase alphabets, only *C*, *L*, *O*, *S*, *U*, *V*, *W* and *X* were consistent when it comes to the number of strokes used. The high variance of *k* is possibly due to the multiple ways that alphabet can be written, in which number of strokes ranging from one to three can be used to write it. Similarly, for alphabet *E*, possible methods of writing includes using number of strokes ranging from two to four (considering a stroke to be the writing segment from one pen-down to the next pen-up).

5.2 Order of Strokes

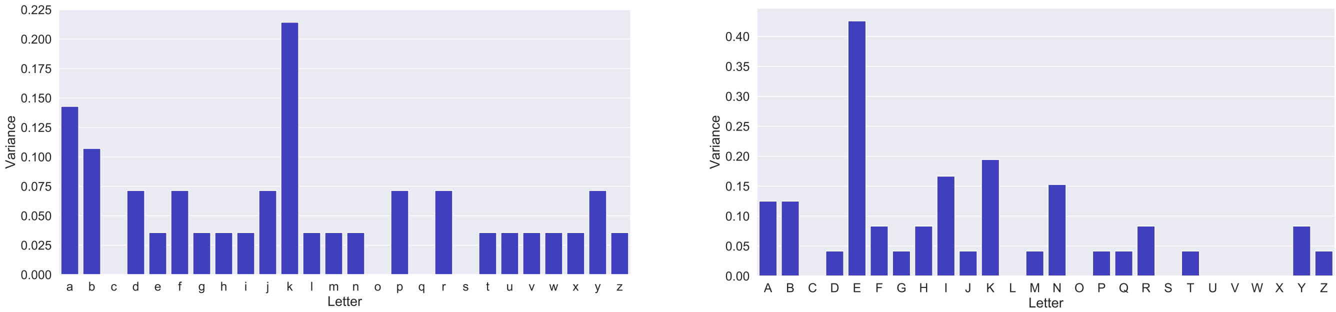
Additionally, we also observed that alphabet letters written using more than one stroke introduces another element of confusion, which is the order in which the strokes are written. A simple example would be writing uppercase alphabet *T*, in which we predominantly observe two strokes. These two strokes can be either written as a horizontal stroke followed by a vertical stroke, or vice versa. Such variations in order of strokes is likely to cause high degree of misclassification during inference.

5.3 Direction of Strokes

Another important factor, especially considering that we are using motion signals to infer handwriting, is the direction of the strokes used to write an alphabet. One common way of writing *f* is starting from the curved top and writing the vertical stroke, followed by the horizontal stroke. But, depending on a person's writing habit, one could also write *f* with a vertical stroke from the bottom and curve it on the top. Figure 8 shows how two different participants have written the alphabet *N*. One participant started from the bottom of the first vertical stroke and continued the same stroke to end of alphabet. The other participant started the first vertical stroke from the top and then went up through the same stroke to complete the alphabet. This is a clear depiction of how even the direction of strokes differ among various writers. Further, the overall shape of the alphabets also could differ across participants. Figure 9 shows two different styles of writing lowercase *y* with a curved style strokes and non-curved strokes. Such variations in direction of strokes is likely to cause high degree of misclassification during inference.

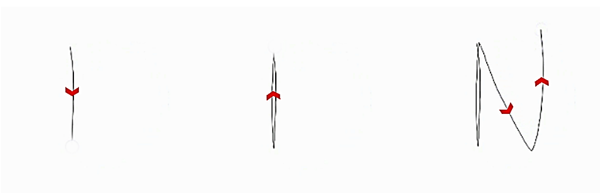
5.4 Training Data Relevance

Even the same person can write differently based on the time, location, or some other context. The variations of number of strokes can be assumed as a possible indicator of such context-depended writing characteristics. For example, while in haste a writer may choose to use lesser number of strokes than usual for writing an alphabet letter or write certain alphabets differently than usual, whereas in more leisurely settings, the same writer may be more careful and consistent in his/her writing. Also, the previous alphabet could affect how the next alphabet is started when writing in a

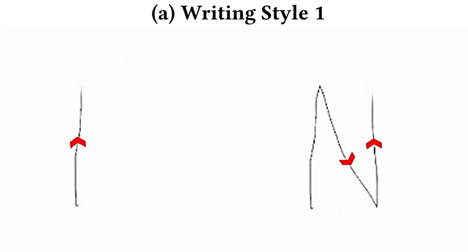


(a) Lowercase **(b) Uppercase**

Figure 7: Variance in number of strokes per alphabet per participant, averaged for all participants.

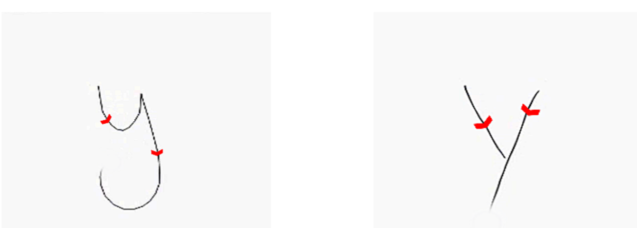


(a) Writing Style 1



(b) Writing Style 2

Figure 8: Different ways of writing alphabet uppercase N.



(a) Writing Style 1 **(b) Writing Style 2**

Figure 9: Different ways of writing alphabet lowercase y.

natural setting. This specific transition motion that occur between alphabets, could vary mainly based on how the previous alphabet was written, i.e. the number and/or the order of the strokes. And these transition motions intrinsically could affect alphabet classification, because the writing motion of an alphabet preceded or followed by the transition motion could easily be mistaken for a totally different alphabet. To this end, the authors of airwriting [5] specified that they used a separate HMM model for such transitions, but do not provide details on how the data to train this model was obtained. And as pointed out above, these factors vary based on the writer and also the writing conditions (time, location, writing surface, pen/pencil used, position, etc.). Therefore, training such a model would be a complicated task due to the highly irregular nature of these transitions.

5.5 Uppercase vs Lowercase

Most of handwriting inference or recognition related works in the literature consider only one of the alphabet cases, i.e., lowercase or uppercase, along with claims that their framework can be easily extended to perform inference in the other case. However, we observed that frameworks which rely on feature-based parametric classifiers do not extend well into the other case. This was observed for finger writing and airwriting in which the original frameworks only considered uppercase, and our results indicate slightly better accuracy for uppercase alphabets over lowercase alphabets. Authors of pencil writing [30] utilized highly specific features designed to assist in the classification of alphabets normally containing exactly two strokes, such as *f*, *i*, *j*, *p*, *t*, and *x*. As a result, it is unclear how such features would perform for uppercase alphabets which can have more (or less) than two strokes.

5.6 Wearables Not on the Writing Wrist

According to an ongoing online study [1] taken part by more than 5667 participants, only about 39.93% users prefer to wear their watch on their dominant hand. This indicates that a majority of users would wear their (smart)watch or fitness band on their non-dominant hand, or in our context the non-writing hand. Needless to say, the outlined handwriting inference attacks using wrist-wearable motion sensors will fail if the eavesdropping smartwatch or fitness-band is not worn on the writing hand.

5.7 Specialized Devices

As mentioned earlier, the original airwriting [5] and finger writing [31] works used specialized hardware, such as a glove with a much higher sampling rate (more than 800Hz) than that supported by consumer-grade smartwatches. Such specialized hardware is advantageous to an attacker since it allows capturing more sensitive and comprehensive information on hand movements. While the requirement of specialized hardware currently limits the scope of handwriting inference attacks, recently researchers have shown that consumer-grade smartwatches may have more potential than previously known. For example, Laput et. al. [12] show that consumer-grade smartwatch OS impose an artificial limit on the sensor sampling rate, which can be bypassed by modifying the OS kernel. With a modified kernel, Laput et. al. [12] were able to record accelerometer data at $4KHz$ in order to detect hand gestures and detect objects grasped by the hand. Therefore, it is possible that in future, sensor sampling rate becomes much higher in smartwatches, which would allow adversaries to capture more sensitive hand motion data capable of more accurate handwriting inference.

6 DISCUSSION

6.1 Limitations

Our main objective in this work was to investigate whether state-of-the-art wrist motion based handwriting inference techniques do actually work "as advertised" in realistic (uncontrolled and unconstrained) writing settings and scenarios. Our overarching goal was to determine if these schemes pose a significant privacy threat and can be deployed as a feasible adversarial tool to infer sensitive handwritten text. Although we demonstrate that existing wrist motion-based handwriting inference techniques do not perform well in realistic writing scenarios using modern consumer-grade wrist wearable devices and would not be very feasible adversarial tools, our work stops at that point. In this paper, we do not make any attempt to propose novel inference frameworks that outperform the existing ones considered in the earlier research efforts. However, the lessons learned from this research effort will definitely be useful in such endeavors in the future.

Despite our best efforts to collect participant handwriting data in a natural and unconstrained setting, we were obviously not able to capture all possible writing situations. Our data collection was still in a conventional writing setting and we did not include/evaluate non-conventional scenarios such as writing too quickly (due to one being in a haste) or writing too slowly. Moreover, our experiments only considered a set of standard and popularly used writing apparatus and surfaces, and we did not evaluate these existing inference mechanisms for a variety of other alternate writing tools (such as stylus, marker, chalk, etc.) and surfaces (such as curved or angled writing surfaces). It should also be noted that while participants were given complete freedom (and recommended) to write in a natural, unconstrained fashion (with the only limitation being non-cursive), we were unable to control environmental factors such as preferred ambient light and temperature which can also potentially impact a person's natural writing ability or style. Furthermore, all the wrist motion based handwriting inference schemes analyzed in this work consider only non-cursive writing scenarios primarily because of the inherent complexity of inferring inter-connected letters within a word in cursive writing. In addition to this, these schemes only collected data from right-handed writers for consistency reasons. In order to accomplish an equitable comparison of the obtained inference results, we also carried out our data collection experiments only for non-cursive handwriting and for right-handed writers. A more comprehensive analysis in this direction should also include data from left-handed and cursive writers.

6.2 Replicability

When trying to replicate the results of previous handwriting inference works, a significant amount of our efforts went in to re-implementing the inference frameworks and re-collecting data in realistic unconstrained writing settings. Our research would have been less demanding if authors of these earlier works would have made their research reproducible. Unfortunately, this is not surprising as replicability has been a significant issue in the security and computer systems community [8, 10]. To make our research effort more useful to the community, we have made all our data and source code publicly available. A web link to these artifacts can be found at the end of the abstract of this paper. Researchers working in the same domain will now be able to comparatively analyze their proposals to the existing ones in the literature.

7 CONCLUSION

In this paper, we investigated how frameworks on wrist-wearable motion sensor based handwriting inference attacks perform in realistic day-to-day writing situations. We carefully analyzed the major factors that bring complexity to wrist motion based handwriting recognition by highlighting specific ambiguities we observed in the order of the strokes, number of strokes, and direction of strokes when writing a character, followed by the overall shape of a character. In addition to these writing characteristics being different among different users, we also observed inconsistencies within the same user's handwriting. Our investigation depicts that due to highly varying nature of handwriting from person to person, wrist motion sensor based inference attacks are unlikely to pose a substantial threat to users of current consume-grade smartwatches and fitness bands.

ACKNOWLEDGMENT

Research reported in this publication was supported by the Division of Computer and Network Systems (CNS) of the National Science Foundation (NSF) under award number 1828071 (originally 1523960).

REFERENCES

[1] 2019. Poll: What is Your Hand-Orientation & What Wrist Do You Wear Your Watch On? http://www.ablogtowatch.com/poll-your-hand-orientation-what-wrist-wear-your-watch/. Online; accessed 2019-01-25.
[2] 2019. Shimmer Sensing. http://www.shimmersensing.com/. Online; accessed 2019-01-25.
[3] Christoph Amma, Dirk Gehrig, and Tanja Schultz. 2010. Airwriting recognition using wearable motion sensors. *ACM International Conference Proceeding Series*.
[4] Christoph Amma, Marcus Georgi, and Tanja Schultz. 2012. Airwriting: Hands-free mobile text input by spotting and continuous recognition of 3D-space handwriting with inertial sensors. In *ACM International Symposium on Wearable Computers (ISWC)*.
[5] Christoph Amma, Marcus Georgi, and Tanja Schultz. 2014. Airwriting: A wearable handwriting recognition system. *Springer Personal and Ubiquitous Computing* 18, 1 (2014).
[6] Luca Ardüser, Pascal Bissig, Philipp Brandes, and Roger Wattenhofer. 2016. Recognizing text using motion data from a smartwatch. In *IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops)*. 1–6.
[7] Yoshua Bengio, Yann LeCun, Craig Nohl, and Chris Burges. 1995. LeRec: A NN/HMM Hybrid for On-Line Handwriting Recognition. *Neural Computation* 7, 6 (1995).

[8] Christian Collberg and Todd A Proebsting. 2016. Repeatability in computer systems research. *Commun. ACM* 59, 3 (2016).

[9] Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. In *Language Resources and Evaluation Conference*, Vol. 29.

[10] Peter Ivie and Douglas Thain. 2018. Reproducibility in Scientific Computing. *ACM Computing Surveys (CSUR)* 51, 3 (July 2018).

[11] Cagdas Karatas, Luyang Liu, Hongyu Li, Jian Liu, Yan Wang, Sheng Tan, Jie Yang, Yingying Chen, Marco Gruteser, and Richard Martin. 2016. Leveraging Wearables for Steering and Driver Tracking. In *IEEE International Conference on Computer Communications (INFOCOM)*.

[12] Gierad Laput, Robert Xiao, and Chris Harrison. 2016. Viband: High-fidelity bio-acoustic sensing using commodity smartwatch accelerometers. In *ACM Annual Symposium on User Interface Software and Technology*.

[13] Xiangyu Liu, Zhe Zhou, Wenrui Diao, Zhou Li, and Kehuan Zhang. 2015. When Good Becomes Evil: Keystroke Inference with Smartwatch. In *ACM Conference on Computer and Communications Security (CCS)*.

[14] Anindya Maiti, Oscar Armbruster, Murtuza Jadliwala, and Jibo He. 2016. Smartwatch-Based Keystroke Inference Attacks and Context-Aware Protection Mechanisms. In *ACM Asia Conference on Computer and Communications Security (AsiaCCS)*.

[15] Anindya Maiti, Ryan Heard, Mohd Sabra, and Murtuza Jadliwala. 2018. Towards Inferring Mechanical Lock Combinations Using Wrist-Wearables As a Side-Channel. In *ACM Conference on Security and Privacy in Wireless and Mobile Networks (WiSec)*.

[16] Anindya Maiti, Murtuza Jadliwala, Jibo He, and Igor Bilogrevic. 2015. (Smart)Watch Your Taps: Side-channel Keystroke Inference Attacks Using Smartwatches. In *ACM International Symposium on Wearable Computers (ISWC)*.

[17] Anindya Maiti, Murtuza Jadliwala, Jibo He, and Igor Bilogrevic. 2018. Side-Channel Inference Attacks on Mobile Keypads Using Smartwatches. *IEEE Transactions on Mobile Computing* 17, 9 (2018).

[18] Emmanuel Munguia Tapia. 2008. *Using machine learning for real-time activity recognition and estimation of energy expenditure*. Ph.D. Dissertation. Massachusetts Institute of Technology.

[19] Sashank Narain, Triet D Vo-Huu, Kenneth Block, and Guevara Noubir. 2016. Inferring User Routes and Locations using Zero-Permission Mobile Sensors. In *IEEE S&P*.

[20] Krishna S Nathan, Homayoon SM Beigi, Jayashree Subrahmonia, Gregory J Clary, and Hiroshi Maruyama. 1995. Real-time on-line unconstrained handwriting recognition using statistical methods. *International Conference on Acoustics, Speech, and Signal Processing* 4 (1995).

[21] Réjean Plamondon and Sargur N Srihari. 2000. Online and off-line handwriting recognition: a comprehensive survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22, 1 (2000).

[22] Lawrence R Rabiner. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* 77, 2 (1989).

[23] Mohd Sabra, Anindya Maiti, and Murtuza Jadliwala. 2018. Keystroke Inference Using Ambient Light Sensor on Wrist-Wearables: A Feasibility Study. In *ACM Workshop on Wearable Systems and Applications (WearSys)*.

[24] Skyler Seto, Wenyu Zhang, and Yichen Zhou. 2015. Multivariate time series classification using dynamic time warping template selection for human activity recognition. *arXiv:1512.06747* (2015).

[25] Muhammad Shoaib, Stephan Bosch, Ozlem Durmaz Incel, Hans Scholten, and Paul JM Havinga. 2016. Complex Human Activity Recognition Using Smartphone and Wrist-Worn Motion Sensors. *Sensors* 16, 4 (2016), 426.

[26] Chen Wang, Xiaonan Guo, Yan Wang, Yingying Chen, and Bo Liu. 2016. Friend or Foe?: Your Wearable Devices Reveal Your Personal Pin. In *ACM Asia Conference on Computer and Communications Security (AsiaCCS)*. 189–200.

[27] Chen Wang, Xiaonan Guo, Yan Wang, Yingying Chen, and Bo Liu. 2016. Friend or Foe?: Your Wearable Devices Reveal Your Personal PIN. In *ACM Asia Conference on Computer and Communications Security (AsiaCCS)*. 12.

[28] He Wang, Ted Tsung-Te Lai, and Romit Roy Choudhury. 2015. MoLe: Motion Leaks Through Smartwatch Sensors. In *ACM Annual International Conference on Mobile Computing and Networking (MobiCom)*.

[29] Hongyi Wen, Julian Ramos Rojas, and Anind K Dey. 2016. Serendipity: Finger gesture recognition using an off-the-shelf smartwatch. In *ACM CHI Conference on Human Factors in Computing Systems*.

[30] Qingxin Xia, Feng Hong, Yuan Feng, and Zhongwen Guo. 2018. MotionHacker: Motion sensor based eavesdropping on handwriting via smartwatch. In *IEEE Conference on Computer Communications Workshops (INFOCOM Workshops)*. 468–473.

[31] Chao Xu, Parth H Pathak, and Prasant Mohapatra. 2015. Finger-writing with smartwatch: A case for finger and hand gesture recognition using smartwatch. In *ACM Workshop on Mobile Computing Systems and Applications (HotMobile)*. 9–14.

[32] Liang Zhang, Yongyuan Qin, and Jinliang Zhang. 2011. Study of polynomial curve fitting algorithm for outlier elimination. In *IEEE International Conference on Computer Science and Service System (CSSS)*.