Big Wave Shoulder Surfing

Neil Patil
neilpatil@utexas.edu
University of Texas at Austin

Brian Cui*
briancui@utexas.edu
University of Texas at Austin

Hovav Shacham
hovav@cs.utexas.edu
University of Texas at Austin

ABSTRACT

Camera technology continues to improve year over year with advancements in both hardware sensor capabilities and computer vision algorithms. The ever increasing presence of cameras has opened the door to a new class of attacks: the use of computer vision techniques to infer non-digital secrets obscured from view. We prototype one such attack by presenting a system which can recover handwritten digit sequences from recorded video of pen motion. We demonstrate how our prototype, which uses off-the-shelf computer vision algorithms and simple classification strategies, can predict a set of digits that dramatically outperforms guessing, challenging the belief that shielding information in analog form is sufficient to maintain privacy in the presence of camera surveillance. We conclude that addressing these new threats requires a new method of thinking that acknowledges vision-based side-channel attacks against physical, analog mediums.

CCS CONCEPTS

• **Security and privacy → Human and societal aspects of security and privacy**; **Security in hardware**; • **Computing methodologies → Computer vision**.

KEYWORDS

video camera, computer vision, pen and paper, side-channel attack, image analysis, mobile devices, motion tracking, handwriting

ACM Reference Format:
Neil Patil, Brian Cui, and Hovav Shacham. 2019. Big Wave Shoulder Surfing. In . ACM, New York, NY, USA, 10 pages.

1 INTRODUCTION

Cameras are widely understood as capturing a limited 2D projection of a 3D world. However, improvements in camera hardware and image processing are enabling cameras to infer aspects of the surrounding 3D space. Advancements in motion tracking, object recognition, and scene reconstruction from the Computer Vision community, combined with higher megapixel counts and video framerates, are enabling advanced image processing features on consumer devices such as "Night Sight" (for low light photography) and "Portrait Mode" (for simulated depth-of-field).

*The first two authors contributed equally to this work.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

, ,
© 2019 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00

Figure 1: An example scenario. Even though the writing is not visible, the pen's motion can be observed and the writing reconstructed by a digital camera. Our prototype generalizes this attack to less optimal camera angles, distances, and image quality.

Meanwhile, analog mediums continue to be trusted platforms for sensitive information. Writing on paper, physical ID cards, and face-to-face meetings are considered more trustworthy and less accessible to foreign adversaries when compared to digital communication. The airgapped nature of analog, "real-life" activities is assumed to avoid many of the risks stemming from the digital realm.

Given ongoing trends regarding the proliferation of cameras, improvements to sensor hardware, and new computer vision techniques, we argue that the safety of analog data - and in particular, the process of its creation - is overestimated in today's environment. The increasing availability and effectiveness of cameras, present in both pockets and on property, mean that common areas where physical data is frequently produced cannot be trusted as confidential, and core societal processes dependent on analog platforms have not yet adapted to the increasing capabilities of cameras.

While cameras add permanency to what can be seen directly in their field of view, they can also *indirectly observe* hidden information in and outside the frame. As a result, the number of opportunities for computer-vision based side channel attacks is increasing. Advanced image processing techniques can recover visual artifacts from what was previously thought as noise [17]. Even in designated secure environments, such as a bank vault or government embassy, remote recording devices intended for maintaining surveillance and security can be used to breach the analog sphere. As an example, we believe the following attacks will become feasible within the coming years:

- An attacker with access to a surveillance camera observes the hand and pen motion of someone writing on a paper

form at a bank, inferring what was written without a direct view of the paper. In this paper, we prototype a working version of this attack.
- The camera present in an airplane infotainment tablet observes the laptop vibrations and arm motions of a passenger, inferring what he or she is typing. Attacks making use of hardware vibrations from user input have already been demonstrated, inferring keypresses from the vibration of a laptop [16] and touchscreen input from the vibrations of a tablet [38].
- An overhead camera records a customer using an ATM and uses the victim's muscle movements to decipher their PIN, even if they have shielded the keypad with their hand. Prior work has already shown that hand motions can be used to infer a user's smartphone PIN upon entry [36].

We explore the first attack in depth: recovering handwritten text through distant observation of pen motion alone. Such an attack would be effective even if the victim deliberately attempts to obscure their writing from view by covering the tip of their pen or the writing itself. To test the attack, we prototype a system that demonstrates promising accuracy, while only making use of a single fixed camera, off-the-shelf 2D motion tracking, and a relatively simple classification strategy.

In summary, the contributions of this paper are:

- We outline a new class of attacks that allow digital cameras to steal information through indirect observation of *analog* input devices and storage mediums.
- We demonstrate a prototype of an example attack that allows an attacker to steal hidden written text from camera footage of pen motion alone.
- We discuss ongoing social trends as well as hardware and software capabilities that will make this new class of attacks a serious concern in the future.

2 BACKGROUND

Side channel attacks leveraging mobile sensors have grown in popularity over recent years. Attacks using microphones to infer keyboard input [2] and accelerometers to infer smartphone input [3] have been demonstrated to be effective. However, the progress of these attacks has been slowed due to diminishing returns in hardware development and sensor fidelity - in part because of a lack of consumer interest (microphone quality is not usually a distinguishing hardware feature) and physical constraints (such as the inverse square law).

On the other hand, camera technology continues to improve year-over-year. Image and video quality improvements open the door to new vision based side channels that may have previously been considered infeasible. In this section, we identify hardware, software, and societal factors responsible for these changes.

2.1 Improvements in Hardware

Spurred by consumer interest, improvements in high-precision mass manufacturing have resulted in smartphone camera megapixel counts improving year-over-year [33] [13]. These improvements have been accompanied by increases in video capture framerates up to 240 fps, enabling slow-motion video recording previously only available in specialized high speed cameras [18]. Minute visual details useful for side-channel attacks - like the vibrations of a bag of chips caused by noise [11] - will begin to become more easily captured by general purpose consumer-grade camera hardware.

Furthermore, manufacturers are beginning to exceed single-camera configurations in favor of housing multiple cameras in their devices for the capture of "3D pictures" and depth perception effects. The recently released Nokia 9 PureView features five rear-facing cameras [28], and the next iPhone is expected to house three [20]. As sensor costs decrease over the coming years, these multi-camera setups are predicted to make their its way to lower-end phones and other consumer hardware [23].

2.2 Improvements in Software

Improvements in photo processing have made their way to the consumer market, as innovations in computer vision pave the way for intelligent analysis and interpretation of noisy raw camera data [17]. "Night Sight", present in the Google Pixel line of phones, can capture details in near-total darkness using relatively short shutter speeds. "Portrait Mode", available in smartphones from several manufacturers including the iPhone, uses computer vision techniques to estimate depth from one or more cameras. Google Clips, a pocket-sized camera released in 2018, uses machine learning to control the shutter and automatically take "interesting" photos [14].

Recent advancements in computer vision have also enabled computers to fill in hidden visual content. Scene reconstruction of 3D environments from a set of disjoint 2D images has been studied extensively over past years [12]. Adobe's "content aware fill" feature, now present in Adobe After Effects, is capable of seamlessly inferring and filling in missing sections of an image [7]. These techniques could be applied towards ascertaining aspects about the environment that are not immediately shown in the camera frame.

Machine learning has also driven improvements in discerning patterns from noisy signal data [17]. This has already made possible several side channel attacks against the physical space: mobile keyloggers have been built from smartphone microphones [2] and gyroscopes [8]. We believe similar techniques can be carried over to analyze camera images and enable new computer vision side channel attacks.

2.3 Societal Trends

We believe ongoing societal trends will multiply the quantity of cameras and their frequency of use in the coming years, amplifying the plausibility of visual side-channel attacks.

The norms around disclosure of surveillance are being relaxed. The release of Google Glass in 2013 was followed by significant criticism due to public privacy concerns (early adopters were commonly referred to as "glassholes"), which contributed to the project's eventual shutdown [37]. However, five years after Google Glass, similar "always-on, always-connected" devices like Snap Spectacles [10] and Google Clips [14] have been released that have not been met with the same skepticism.

Moreover, the public is becoming more comfortable with - or often unaware of - camera observation, in part due to the growth of social media making camera recording commonplace. Additionally,

public attitudes may begin to shift towards accepting a loss in privacy in exchange for digital conveniences, despite civil liberty concerns imposed by government surveillance programs [31]. Recent polling of the Chinese public has indicated citizens are becoming more supportive of surveillance because it improves perceptions of safety [43].

Finally, the decreasing cost of general purpose hardware, such as Android tablets used in car and airplane infotainment systems, can result in cameras being present in unexpected places. For example, cameras have been found inadvertently embedded in the tablets that make up airplane seatback entertainment displays, garnering privacy concerns [27]. As many of these devices are networked with the outside world, they present new vectors for attackers to gain access to camera footage.

3 CAMERAS STEALING DIGITAL INFORMATION

Prior work has already taken advantage of the new wave of high quality camera hardware, developing new side-channel attacks that extract information from the use of digital devices.

Many papers have gone beyond traditional "shoulder surfing" by investigating using *reflections* off of common surfaces to capture information from computer screens. Reflective surfaces like sunglasses [30], teapots, and even the human eye can be read at a distance to compromise displays [5] [4].

There is also prior work on inferring digital information through solely body motion. Chen et al. showed that eye movements could be used as a reliable source of touchscreen keystroke inference [9]. Shukla et al. showed how cameras can already use observed hand motions to determine the PIN entry process on common smartphones [36].

Yue et al. showed that cameras can infer keyboard inputs a user is typing on a tablet, even without a direct view of the screen itself [42]. Subsequent research showed that this can even be performed at great distances via use of a drone flying overhead [41].

This line of work on vision-based side channel attacks has focused on stealing information through interaction with *digital devices*. As a result, countermeasures often make use of software solutions - such as randomizing number input locations on a PIN entry screen [1] - which can be quickly deployed by a security software update.

Our focus for this paper is investigating the feasbility of *analog* information theft. We believe camera attacks on the analog medium present a greater threat than similar attacks on digital platforms, as common sense understanding of pen-and-paper's affordances has remained unchanged for millenia. Norms surrounding the safety and security of covert writing are far more engrained in human behavior, and will be more difficult to change.

4 ATTACKS ON HANDWRITING

To demonstrate one of the possible analog side-channel attacks that are now possible thanks to camera quality and availability improvements, we prototyped a novel *attack on handwriting* that demonstrates how digital sensors can be used to breach the analog space. Our attack applies simple computer vision techniques to exploit *pen motion* as a source of information. This is based off of the observation that as a person writes, their pen follows a trajectory dependent on the individual characters being written - for example, an "O" might project a more circular motion than a "W".

Figure 2: A frame capture representative of a test video that fits our security model. Note how the writing is deliberately obscured from view. Image has been magnified and brightened for viewing clarity.

We hypothesized that analyzing this motion could allow an attacker to reverse-engineer the original text written on the page. Our paper presents an end-to-end classification program which accepts recorded video of a victim writing numbers and produces a ranking of output predictions for each digit.

4.1 Security Model

We assume that an attacker has access to camera footage of a victim writing on a piece of paper, and wants to determine what the victim has written. However, the camera footage cannot view the paper or pen tip directly - instead, it can only view the back end of the victim's pen. For example, the subject could be writing on the page and deliberately covering the text with their hand, but leaving the rear of the pen visible. An example frame capture demonstrating this view is shown in Figure 2.

We assume the footage comes from a high resolution (1080p and above) zoomable camera capable of recording video of at least 30 FPS - hardware capabilities easily surpassed by modern devices. For this prototype, we restrict the class of written symbols the attacker is attempting to decipher to the digits 0-9. Numbers keep our search space tractable, but also regularly contain sensitive information of known lengths (such as credit cards or social security numbers).

We test our model on a suite of videos recorded from varying camera angles and distances. In our test footage, the victim writes with a font size equal to that used when entering social security numbers on the IRS W-9 form [15], emulating a real-world scenario.

4.2 Prior Work

Prior work on pen observation influenced our system implementation. Using only a top-down camera view, Munich and Perona [26] constructed a stylus-like human-computer interface using a combination of edge detection and Kalman filtering to track the pen

tip. Seok et al. extended this work to track pens against non-blank paper by using a color and shape matching strategy [35].

Wu et al. [39] augmented a pen by attaching a 3D printed dodecahedron with tracking marks printed on each face, enabling real time pose estimation with submilimeter accuracy. While our security model only emcompasses “unaugmented" commodity pens, the effectiveness of the work by Wu et al. suggested to us that knowledge of the rear end of the pen could nonetheless be used to infer motion of the front.

Yasuda et al. [40] proposed a signature verification system which compares the motion of a pen’s tip against a pre-recorded signature path, making use of temporal sequence similarity metrics to determine if two pen paths “match". Our model takes advantage of a similar metric, Time Warp Edit Distance (TWED) [25], in order to compare temporal sequences.

4.3 System Design

In order to use the program, the attacker must supply 1) video footage of the user writing, 2) the estimated locations of the four corners of the paper in the video frame, 3) a bounding box around the rear end of the pen to motion track, and 4) the quantity of digits to predict, n.

Once these details are provided, the program executes the following steps in sequence:

(1) The program estimates the pose of the camera relative to the paper. More formally, the program determines the position of a 3D paper from its 2D projection on the camera view, outputting a rotation matrix R and translation vector t. We make use of the OpenCV implementation of the Perspective-n-Point algorithm by Li et al. [22] with the four corners of an 8.5x11 inch page (estimated by the attacker) as the reference points.

(2) The set of 3D ground truth digit sequences are transformed to match the camera’s perspective. Our model applies a linear transformation by multiplying the the generated rotation matrix R by the positions stored by the time series ground truth data (discussed in the following section). In effect, the ground truth motion sequences are transformed to what they would look like as if they were viewed in 2D at the given camera angle.

(3) The path of the rear end of the pen is determined using motion tracking. The attacker must initially draw a bounding box around the rear end of the pen, which is then tracked through successive frames and used to construct a time series that estimates the pen’s path in the 2D projection. Our model uses the Discriminative Correlation Filter with Channel and Spatial Reliability (CSRT) [24] [19] algorithm, a state-of-the-art motion tracking algorithm based on optical flow. We use the implementation of CSRT supplied by OpenCV 3.4.3. An example of a motion tracked writing path of the digit “3" is shown in Figure 5.

(4) The multi-digit path is partitioned into individual digit paths using a discretization heuristic. This helps distinguish “pen down" motion (when ink is flowing) from “pen up" motion (spaces and transitions). In order to do so, we observed that spaces are often indicated by a short burst of rapid motion as the victim shifts their pen quickly in preparation for writing the next digit. Accordingly, our model greedily finds points in the path that have the greatest speed, flagging the nearby surrounding region as a space until $n - 1$ spaces have been found. Figure 4 displays the normalized speed of a path “1-2-3" from our ground truth set, with circles indicating the beginning of spaces according to our heuristic.

(5) Each digit chunk is compared against the set of pre-recorded and transformed ground truth paths (from step 2) and ranked. In order to compare these different temporal sequences, we make use of Time Warp Edit Distance (TWED) [25], a variant of the Fast Dynamic Time Warping algorithm [34]. The score of a sequence is then computed by taking the sum of squared distance (TWED score) for the individual X and Y axis temporal sequences. The system then finds the digit class with a sequence that minimizes this score.

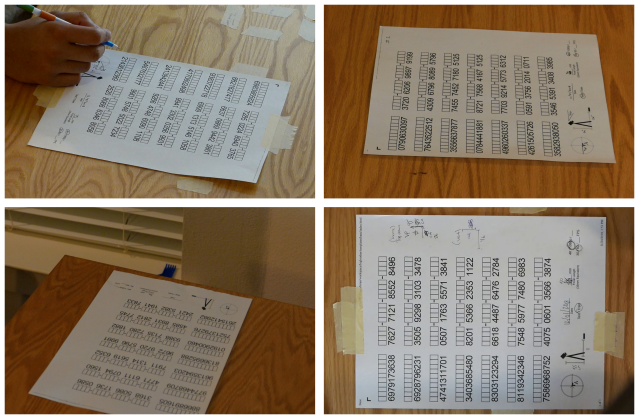

Figure 3: Frame captures of the (0, 55, 100), (0, -45, -90), (0, -30, -130) test cases and ground truth collection case, respectively. The images here have been cropped, brightened and magnified significantly for viewing clarity.

The final output of the model is a ten class ranking for every digit in the sequence.

Our model was developed using Python 3.6.5 and the OpenCV 3.4.3 computer vision library [29].

The project source code, including all training and testing data, is available online at `https://github.com/patil215/bigbrother`.

4.4 Training Data Collection

We recorded a corpus of 3D “ground truth" data for our model to compare inputs against for classification. Our ground truth digit paths were collected using a two camera setup to capture the motion of the pen in 3D, using a professional 12MP Panasonic Lumix DMC-LX100 camera and a 12MP GoPro Hero 5 aligned only the X-Y and X-Z plane respectively. Both cameras were set to record at 1080p and 60 frames per second. The footage was then manually segmented and discretized to create a corpus of paths for each digit. A sample 3D path is shown in Figure 5.

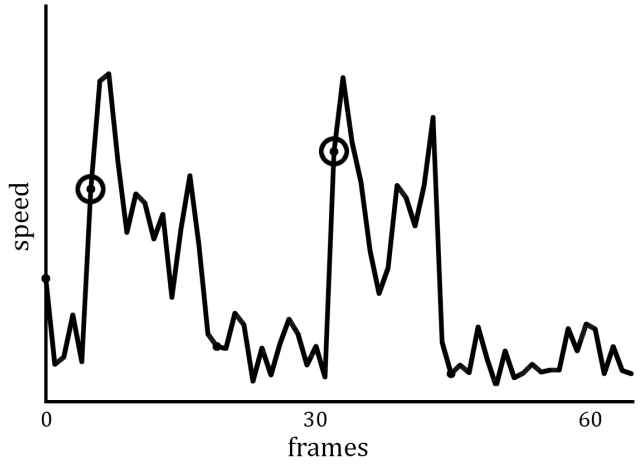

Figure 4: Chart showing normalized speed of the pen end for each frame in the sequence "1-2-3". The circles indicate the frame right after the beginning of spaces, as determined by our greedy heuristic.

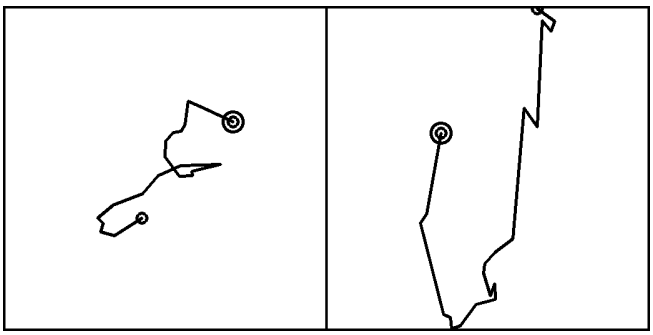

Figure 5: An example ground truth path for the digit "3", captured from motion tracking the rear end of the pen. The left image is from the top down XY perspective; the right image is from the side YZ perspective. The start and end of the path is designated by the larger and smaller circles, respectively.

Our ground truth data set of 253 single-digit samples consisted of about 25 examples per digit class, recorded from a single test subject (Patil).

4.5 Results

4.5.1 Test Data. To evaluate the effectiveness of our model, we recorded a suite of test videos comprising multiple camera angles, distances, and digit sequence lengths of our test subject (Patil) writing random numbers on the same test bed template shown in Figure 6. All videos were recorded with the Panasonic Lumix camera at 4K resolution and 30 FPS. Test paths were captured using the motion tracking after manually drawing a bounding box around the rear end of the pen. Frame captures from each angle are shown in Figure 3. All in all, our test video data comprises 1213 paths spanning 12.4 gigabytes of compressed video footage.

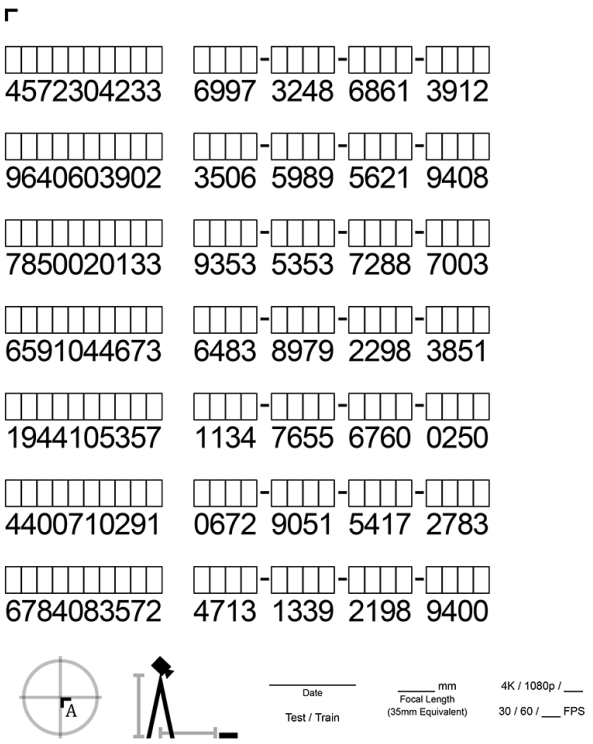

Figure 6: Example test bed template sheet, printed out on standard 8.5 in by 11 in paper. The digit entry boxes are .5 cm wide and .85 cm tall to emulate the digit entry fields on the IRS W-9 form [15]. The numbers are randomly generated with each test sheet.

The following table provides a snapshot of the quantity of test cases, organized by sequence length. We focus on test paths of "interesting" lengths: credit card numbers are grouped by four digit sequences, phone numbers are seven digits, and social security numbers are nine digits. Results from input sequences of other lengths are present in the project source code.

Euler Angles	Distance (m)	1	4	7	9
0, 55, 100	2.4	80	56	32	16
0, -30, -130	1.4	69	49	28	14
0, -45, -90	1.7	70	49	28	14

4.5.2 Evaluation. As our model offers a ranking of all ten possible digits for each of the n digits in the sequence, we evaluate our model's performance by investigating the likelihood of our model guessing the correct sequence within r guesses for each digit.

We define the *guess magnitude* as the size of the set of possible digit sequences given r guesses for each digit. Given an n digit sequence and $1 \leq r \leq 10$ guesses for each of the digits in the sequence, the total guess magnitude is then r^n. (In the worst case, $r = 10$ and thus the guess magnitude is 10^n, equivalent to brute force search.)

A given model's *accuracy score* is computed as the likelihood that the correct digit sequence appears among all guesses. We compare our model's accuracy to a baseline model where each digit is randomly picked out of a size of set r. When picking randomly among each digit given 10^n total n-digit sequences, the baseline random guess accuracy score is then $\frac{r^n}{10^n} = (\frac{r}{10})^n$.

When classifying single digits, our model is able to achieve 61.64% accuracy with only a single guess, compared to a 10% probability for the baseline. Given a set of five guesses, our model manages to contain the correct digit within the set 95.89% of the time, compared to a 50% probability for the baseline.

Our model's accuracy scores significantly beat the baseline in all multiple digit sequence cases we tested as well. For example, for $n = 9$, the set of guesses generated by taking the top seven ($r = 7$) guesses for each digit from our model will contain the actual nine digit sequence with 51.22% probability, whereas the baseline only offers a 4.04% probability of guessing the nine digit sequence correctly given seven random guesses for each digit. This also yields a signficant reduction in search space: a brute force search over nine digits would give a guess magnitude of 10^9, which is $\frac{10^9}{7^9} = 729$ times larger than the selection offered by our model for $r = 7$. Moreover, our model's suggestions are still useful when an exhaustive search is needed, placing more likely sequences candidates first.

Table 1 includes a breakdown of how well our model classifies digit sequences of various lengths. Our model performs better than the baseline by a significant margin when given the same number of guesses. Graphs comparing our model's accuracy to the baseline for $n = 1, 4, 7$ are shown in Figure 7.

The model performs better on some camera vantage points than others - for the single-digit case, the average rank (average value of r needed to correctly guess the sequence) at the best angle (0, -45, -90) is 1.5, while at the worst angle (0, 55, 100), the average max rank is 2.0. Table 2 provides complete statistics of average max rank by sequence length and camera angle.

Our model's performance validates our hypothesis by demonstrating pen motion, however subtle, can be analyzed to recover information about the text being written. Additionally, the success of the single digit case from chained application of relatively unsophisticated techniques for data collection and classification demonstrates this information may be surprisingly accessible.

4.6 Discussion

The outcome of our prototype demonstrates that cameras can be used to indirectly capture information that has been hidden from an attacker. We stress that our model does not take advantage of more sophisticated techniques that are already available today that would make it more effective:

(1) Our model uses a relatively small amount of information available from the camera, tracking the rear end of the pen as a single point. This ignores the full 3D geometry of the pen, which can be used to estimate the pose more accurately. Research has demonstrated this is possible given a predefined pen reference model for pose estimation [21]. Additionally, inputs from other side channels can be considered simultaneously with camera vision, such as the use of the sound for digit segmentation.

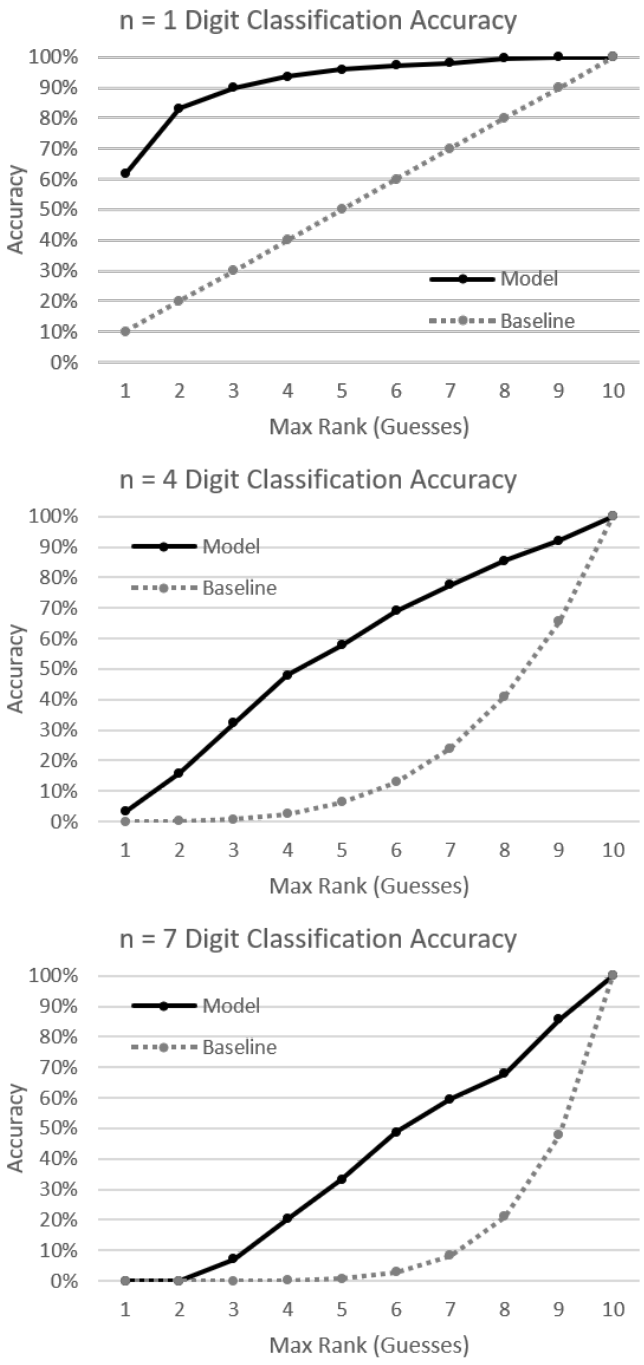

Figure 7: Accuracy scores for sequence lengths of $n = 1, 4, 7$.

For $n = 1$, out of a set of 219 test samples, our model correctly classifies 61.64% with only a single guess, with the correct class appearing within the top 5 guesses 95.89% of the time. The baseline is given by $\frac{rank}{10}$.

For $n = 4$, out of a set of 152 test samples, 57.89% of sequences were included in the set of sequences given by five guesses per digit. The baseline is given by $(\frac{rank}{10})^4$.

For $n = 7$, out of a set of 84 test samples, our model considered 33.33% sequences correctly within the top 5 guesses for each digit. The baseline is given by $(\frac{rank}{10})^7$.

Seq. Length	1		4		7		9	
Max Rank (r)	Model (%)	Base (%)	Model (%)	Base (%)	Model (%)	Base (%)	Model (%)	Base (%)
1	61.64	10.00	3.29	0.01	0.00	0.00	0.00	0.00
2	83.11	20.00	15.79	0.16	0.00	0.00	0.00	0.00
3	89.95	30.00	32.24	0.81	7.14	0.02	2.44	0.00
4	93.61	40.00	48.03	2.56	20.24	0.16	9.76	0.03
5	95.89	50.00	57.89	6.25	33.33	0.78	21.95	0.20
6	97.26	60.00	69.08	12.96	48.81	2.80	36.59	1.01
7	98.17	70.00	77.63	24.01	59.52	8.24	51.22	4.04
8	99.54	80.00	85.53	40.96	67.86	20.97	60.98	13.42
9	100.00	90.00	92.11	65.61	85.71	47.83	80.49	38.74
10	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00

Table 1: Percent accuracy results for sequence lengths of 1, 4, 7, and 9 compared to baseline results. All percentages are rounded to two decimal places. Our model outperforms the baseline even as sequence length increases.

Sequence Length (n)	Average Max Rank (0, -45, -90)	Average Max Rank (0, -30, -130)	Average Max Rank (0, 55, 100)	Average (all angles)	Std. Dev. (all angles)
1	1.50	1.83	2.06	1.95	0.28
4	3.82	5.98	5.71	5.85	1.18
7	5.43	7.42	7.47	7.45	1.16
9	6.14	7.82	8.13	7.98	1.07

Table 2: Average Max Rank (AMR) for a few representative angles for sequence lengths of 1, 4, 7, and 9. AMR helps estimate our model's performance on a given test dataset; this score can vary depending on camera angle and distance.

(2) The algorithms used for in tracking and classification are general purpose and non-domain-specific. Our current TWED and nearest neighbor classification strategy is relatively unsophisticated compared to more complex classification models such as neural networks [32]. A more sophisticated scoring kernel such as multidimensional TWED [6] could better distinguish handwritten characters from each other. Additionally, a more powerful motion tracking strategy such as a color and shape matching approach could be able to track the pen with greater accuracy [35].
(3) Our model makes use of relatively unsophisticated hardware compared to what is entering the market. High speed video recording capabilities (240 FPS) are becoming standard smartphone features, which can lead to improvements in object motion tracking accuracy [18]. Stereo camera hardware offering depth estimation is also becoming available [20], with some applications like virtual reality require more robust depth tracking using dedicated infrared sensor arrangements. Making use of these sensors would allow for 3D depth estimation, making the motion tracking more powerful.

We believe these improvements represent relatively low-hanging fruit and are thus not out of reach of a government agency, computer vision research team, or specialized corporation.

5 COUNTERMEASURES

The most direct mitigation against the presence of camera surveillance is straightforward: conduct business in an area free of it. However, society today has many scenarios where people cannot or do not think to take this mitigation, as common sense behavior surrounding pen-and-paper is entrenched within our society. Businesses and government offices contain employees and customers working on forms and other confidential data among many attack venues - overhead security cameras, videoconferencing hardware, and personal devices. Polling places have shielded voting booths, but these still usually leave the eye, body, and arm motions of voters open to indirect observation. Consumers are actively inviting IOT devices and camera-based security systems into their homes. Public spaces like banks or stores use cameras to monitor customers, ironically as a security measure.

Ensuring that all of these spaces are free of indirect camera surveillance is a daunting task, but without greater exploration of the full capabilities of cameras, these attacks will go unnoticed and these activities unadjusted.

Fundamentally, these attacks stem from a mis-estimation of privilege - not only of the details of what cameras can infer, but the abilities of nearby objects - vibrating chip bags [11], reflective surfaces [5], moving hands and pens [36] - to indirectly signify the activities around them. In order to identify these opportunities in the future, computer security experts will need to work with physical security, sensor hardware, and computer vision experts.

6 CONCLUSION

The growing availability of cameras is blurring the line between the analog and digital space. Innovations in both camera hardware and computer vision software accompanied by changing social norms have opened the door to a new type of side channel attack: inference of information believed to be hidden from view.

In this paper, we present a novel information stealing attack against pen-and-paper which demonstrates the ability of cameras to infer aspects of their environment that are deliberately occluded. We show that pen motion is a salient source of information about written output, even when the writing itself is not directly visible. We present a prototype that uses off-the-shelf computer vision algorithms and simple classification strategies to predict written digit outputs with promising accuracy. Given that our prototype does not make use of more specialized techniques that have already proven to be effective, we have no doubt that improvements are possible that would allow for prediction of any handwritten text from auxiliary motion alone.

We believe this example is representative of a larger class of attacks that challenge the belief that shielding analog information is sufficient to maintain privacy in the presence of cameras. Given improving camera hardware and vision algorithms, as well as changing social norms, such attacks could eventually disrupt important social processes that are particularly difficult to change, such as voting. We conclude by suggesting that computer scientists directly collaborate with signal processing engineers, physical security specialists, and computer vision and policy experts in order to address future threats.

REFERENCES

[1] MD Amruth and K Praveen. 2016. Android smudge attack prevention techniques. In *Intelligent Systems Technologies and Applications*. Springer, 23–31.

[2] D. Asonov and R. Agrawal. 2004. Keyboard acoustic emanations. In *IEEE Symposium on Security and Privacy, 2004. Proceedings. 2004.* 3–11. https://doi.org/10.1109/SECPRI.2004.1301311

[3] Adam J. Aviv, Benjamin Sapp, Matt Blaze, and Jonathan M. Smith. 2012. Practicality of Accelerometer Side Channels on Smartphones. In *Proceedings of the 28th Annual Computer Security Applications Conference (ACSAC '12)*. ACM, New York, NY, USA, 41–50. https://doi.org/10.1145/2420950.2420957

[4] M. Backes, T. Chen, M. Duermuth, H. P. A. Lensch, and M. Welk. 2009. Tempest in a Teapot: Compromising Reflections Revisited. In *2009 30th IEEE Symposium on Security and Privacy*. 315–327. https://doi.org/10.1109/SP.2009.20

[5] M. Backes, M. DÃijrmuth, and D. Unruh. 2008. Compromising Reflections-or-How to Read LCD Monitors around the Corner. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*. 158–169. https://doi.org/10.1109/SP.2008.25

[6] Muzaffar Bashir and Jürgen Kempf. 2008. Reduced dynamic time warping for handwriting recognition based on multidimensional time series of a novel pen device. *International Journal of Intelligent Systems and Technologies, WASET* 3, 4 (2008), 194.

[7] Ajay Bedi, Sourabh Gupta, and Saurabh Gupta. 2017. Content aware fill based on similar images. US Patent 9,697,595.

[8] Liang Cai and Hao Chen. 2011. TouchLogger: Inferring Keystrokes on Touch Screen from Smartphone Motion. *HotSec* 11 (2011), 9–9.

[9] Y. Chen, T. Li, R. Zhang, Y. Zhang, and T. Hedgpeth. 2018. EyeTell: Video-Assisted Touchscreen Keystroke Inference from Eye Movements. In *2018 IEEE Symposium on Security and Privacy (SP)*. 144–160. https://doi.org/10.1109/SP.2018.00010

[10] Josh Constine. 2018. Snapchat launches Spectacles V2, camera glasses you'll actually wear. https://techcrunch.com/2018/04/26/snapchat-spectacles-2/. https://techcrunch.com/2018/04/26/snapchat-spectacles-2/ Accessed: 04-11-2018.

[11] Abe Davis, Michael Rubinstein, Neal Wadhwa, Gautham J Mysore, Frédo Durand, and William T Freeman. 2014. The visual microphone: passive recovery of sound from video. (2014).

[12] Charles R Dyer. 2001. Volumetric scene reconstruction from multiple views. In *Foundations of image understanding*. Springer, 469–489.

[13] Tobias Fiebig, Jan Krissler, and Ronny Hänsch. 2014. Security impact of high resolution smartphone cameras. In *8th {USENIX} Workshop on Offensive Technologies ({WOOT} 14)*.

[14] Brian Heater and Brian Heater. 2018. Google Clips review. https://techcrunch.com/2018/02/27/google-clips-review/. https://techcrunch.com/2018/02/27/google-clips-review/

[15] IRS [n. d.]. Form W-9 (Rev. October 2018) - fw9.pdf. https://www.irs.gov/pub/irs-pdf/fw9.pdf. Accessed: 12-05-2018.

[16] K. Jin, S. Fang, C. Peng, Z. Teng, X. Mao, L. Zhang, and X. Li. 2017. ViViSnoop: Someone is snooping your typing without seeing it!. In *2017 IEEE Conference on Communications and Network Security (CNS)*. 1–9. https://doi.org/10.1109/CNS.2017.8228624

[17] Nima Khademi Kalantari, Steve Bako, and Pradeep Sen. 2015. A machine learning approach for filtering Monte Carlo noise. *ACM Trans. Graph.* 34, 4 (2015), 122–1.

[18] Hamed Kiani Galoogahi, Ashton Fagg, Chen Huang, Deva Ramanan, and Simon Lucey. 2017. Need for speed: A benchmark for higher frame rate object tracking. In *Proceedings of the IEEE International Conference on Computer Vision*. 1125–1134.

[19] Matej Kristan, Jiri Matas, Ales Leonardis, Michael Felsberg, Luka Cehovin, Gustavo Fernandez, Tomas Vojir, Gustav Hager, Georg Nebehay, and Roman Pflugfelder. 2015. The visual object tracking vot2015 challenge results. In *Proceedings of the IEEE international conference on computer vision workshops*. 1–23.

[20] Yoko Kubota and Takashi Mochizuki. 2019. Apple Plans Three New iPhones This Year, Plays Catch-Up on Cameras. =https://www.wsj.com/articles/apple-plans-new-lcd-iphone-this-year-despite-xrs-stumble-11547199263. https://www.wsj.com/articles/apple-plans-new-lcd-iphone-this-year-despite-xrs-stumble-11547199263

[21] Vincent Lepetit, Pascal Fua, et al. 2005. Monocular model-based 3d tracking of rigid objects: A survey. *Foundations and Trends® in Computer Graphics and Vision* 1, 1 (2005), 1–89.

[22] Shiqi Li, Chi Xu, and Ming Xie. 2012. A Robust O(n) Solution to the Perspective-n-Point Problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34 (2012), 1444–1450.

[23] Research ltd and Markets. [n. d.]. Global Smartphone 3D Camera Market Size, Market Share, Application Analysis, Regional Outlook, Growth Trends, Key Players, Competitive Strategies and Forecasts, 2018 To 2026. https://www.researchandmarkets.com/reports/4749574/global-smartphone-3d-camera-market-size-market

[24] Alan Lukezic, Tomás Vojír, Luka Cehovin, Jiri Matas, and Matej Kristan. 2016. Discriminative Correlation Filter with Channel and Spatial Reliability. *CoRR* abs/1611.08461 (2016). arXiv:1611.08461 http://arxiv.org/abs/1611.08461

[25] Pierre-François Marteau. 2009. Time warp edit distance with stiffness adjustment for time series matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31, 2 (2009), 306–318.

[26] Mario E Munich and Pietro Perona. 1996. Visual input for pen-based computers. In *Image Processing, 1996. Proceedings., International Conference on*, Vol. 2. IEEE, 173–176.

[27] Christine Negroni. 2019. There Are Probably Cameras on Your Flight, but Relax, They're Not On (Yet). https://www.nytimes.com/2019/04/02/business/airlines-cameras-privacy.html

[28] Nokia [n. d.]. Nokia 9 PureView. https://www.nokia.com/phones/en_int/nokia-9-pureview/#camera. https://www.nokia.com/phones/en_int/nokia-9-pureview/#camera Accessed: 04-11-2019.

[29] OpenCV [n. d.]. OpenCV Library. https://opencv.org/. Accessed: 12-08-2018.

[30] R. Raguram, A. M. White, Y. Xu, J. Frahm, P. Georgel, and F. Monrose. 2013. On the Privacy Risks of Virtual Keyboards: Automatic Reconstruction of Typed Input from Compromising Reflections. *IEEE Transactions on Dependable and Secure Computing* 10, 3 (May 2013), 154–167. https://doi.org/10.1109/TDSC.2013.16

[31] Lee Rainie and Maeve Duggan. 2016. Americans' opinions on privacy and information sharing. https://www.pewinternet.org/2016/01/14/privacy-and-information-sharing/

[32] A Rajavelu, Mohamad T Musavi, and Mukul Vassant Shirvaikar. 1989. A neural network approach to character recognition. *Neural networks* 2, 5 (1989), 387–393.

[33] Kor Renkema. 2007. Moore's law and mold making: staying in the megapixel race. In *Optical Manufacturing and Testing VII*, Vol. 6671. International Society for Optics and Photonics, 66710K.

[34] Stan Salvador and Philip Chan. 2007. Toward Accurate Dynamic Time Warping in Linear Time and Space. *Intell. Data Anal.* 11, 5 (Oct. 2007), 561–580. http://dl.acm.org/citation.cfm?id=1367985.1367993

[35] Jae-Hyun Seok, Simon Levasseur, Kee-Eung Kim, and JinHyung Kim. 2008. Tracing handwriting on paper document under video camera. ICFHR.

[36] Diksha Shukla, Rajesh Kumar, Abdul Serwadda, and Vir V. Phoha. 2014. Beware, Your Hands Reveal Your Secrets!. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security (CCS '14)*. ACM, New York, NY, USA, 904–917. https://doi.org/10.1145/2660267.2660360

[37] David Streitfeld. 2013. Google Glass Picks Up Early Signal: Keep Out. https://www.nytimes.com/2013/05/07/technology/personaltech/google-glass-picks-up-early-signal-keep-out.html

[38] Jingchao Sun, Xiaocong Jin, Yimin Chen, Jinxue Zhang, Rui Zhang, and Yanchao Zhang. 2016. VISIBLE: Video-Assisted Keystroke Inference from Tablet Backside Motion. https://doi.org/10.14722/ndss.2016.23060

[39] Po-Chen Wu, Robert Wang, Kenrick Kin, Christopher Twigg, Shangchen Han, Ming-Hsuan Yang, and Shao-Yi Chien. 2017. DodecaPen: Accurate 6DoF Tracking of a Passive Stylus. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology*. ACM, 365–374.

[40] Kumiko Yasuda, Daigo Muramatsu, Satoshi Shirato, and Takashi Matsumoto. 2010. Visual-based online signature verification using features extracted from video. *Journal of Network and Computer Applications* 33, 3 (2010), 333–341.

[41] Q. Yue, Z. Li, C. Gao, W. Yu, X. Fu, and W. Zhao. 2018. The Peeping Eye in the Sky. In *2018 IEEE Global Communications Conference (GLOBECOM)*. 1–7.

https://doi.org/10.1109/GLOCOM.2018.8647787

[42] Qinggang Yue, Zhen Ling, Xinwen Fu, Benyuan Liu, Kui Ren, and Wei Zhao. 2014. Blind Recognition of Touched Keys on Mobile Devices. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security (CCS '14)*. ACM, New York, NY, USA, 1403–1414. https://doi.org/10.1145/2660267.2660288

[43] Hanqi Zhang, Jianfeng Guo, Chao Deng, Ying Fan, and Fu Gu. 2019. Can Video Surveillance Systems Promote the Perception of Safety? Evidence from Surveys on Residents in Beijing, China. *Sustainability* 11, 6 (2019). https://doi.org/10.3390/su11061595