The Applicability of Ambient Sensors as Proximity Evidence for NFC Transactions

Carlton Shepherd, Iakovos Gurulian, Eibe Frank*, Konstantinos Markantonakis, Raja N. Akram, Emmanouil Panaousis†, Keith Mayes

Information Security Group, Royal Holloway, University of London, United Kingdom
* Dept. of Computer Science, University of Waikato, New Zealand
† University of Brighton, United Kingdom

IEEE Mobile Security Technologies '17

Contactless and Near-Field Communication (NFC)

- Contactless cards
  - First introduced by UK banks in 2007
  - Technicalities governed by ISO 14443
  - RFID induction at 13.56MHz (range: ~5cm)
  - 1 in 8 card payments are contactless in UK (UK Cards Association, 2016)

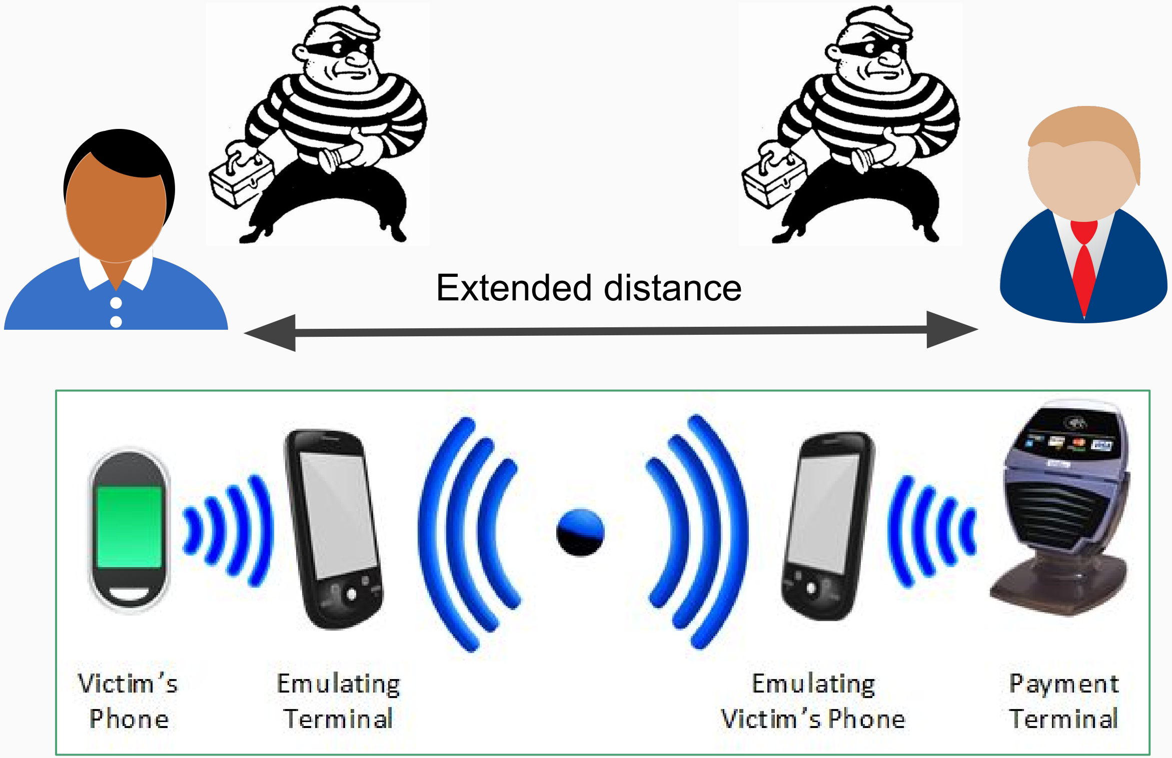
- NFC
  - Developed in 2002 by Sony and NXP
  - Contactless functionality on mobile platforms
  - NFC-enabled mobile devices can emulate a contactless card or reader

Relay Attacks

Passive man-in-the-middle attack in which an attacker extends the distance between the transaction terminal and payment instrument

Lack of proximity detection mechanism within NFC allows this. (“Is the device *really* <5cm away from the terminal?”)

Relay attacks allow attackers to use victims’ credentials for their benefit. **Use cases: access control, transportation, purchasing goods…**

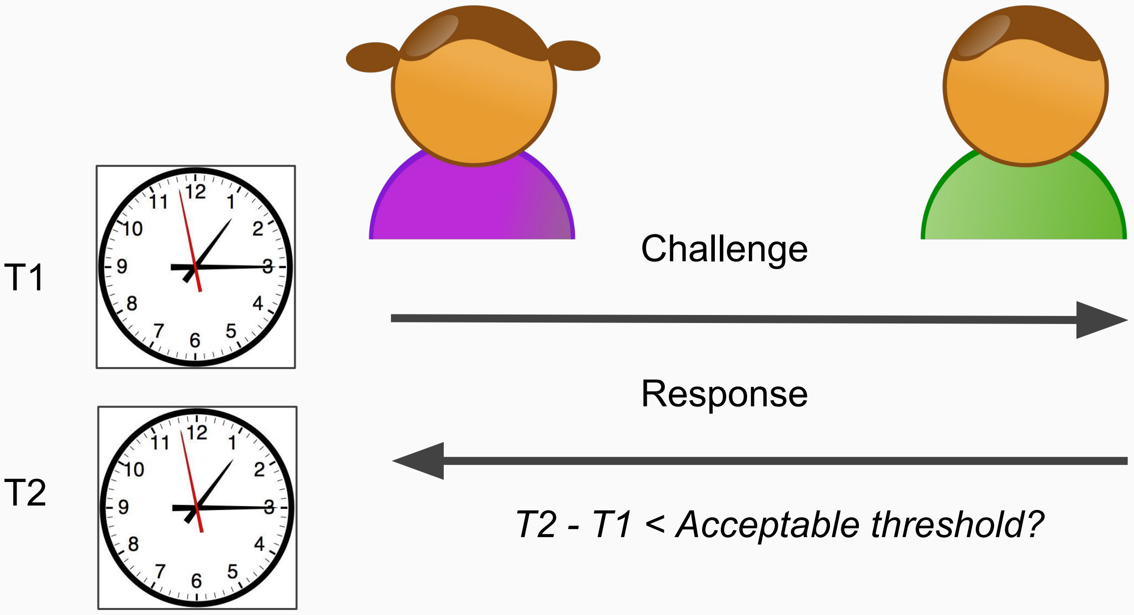


Proximity Detection

The proximity problem is well-known with conventional contactless cards; solved by **distance-bounding protocols**

Same attack applies with mobile devices; **distance-bounding very difficult due to hardware/software variations between devices**

Distance Bounding by Time



Proximity Detection via Sensing

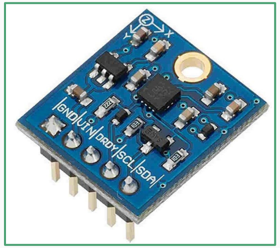
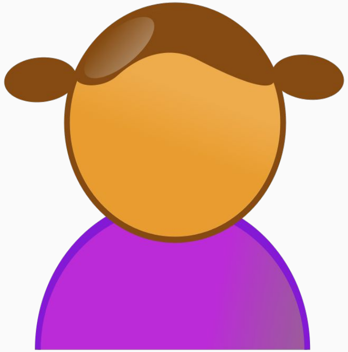
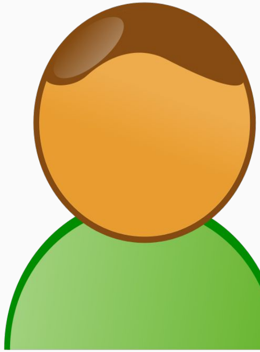
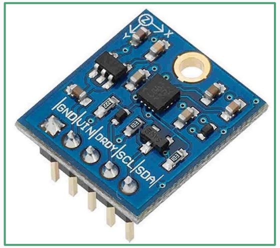
- **Ambient sensing** proposed in countless papers to address the proximity detection problem with mobile devices, e.g. Varshavsky et al. [1]

- **Assumption:** environmental conditions of the transaction terminal and mobile device are uniquely similar, e.g. sound of a loud cafeteria

- ...but how well does this assumption hold in practice? This is the aim of our investigation

1. Varshavsky et al., “Amigo: proximity-based authentication of mobile devices.”, UbiComp (2007), Springer, 253-27

Distance Bounding by Sensing

S1 = {*measurements*}

S2 = {*measurements*}

Send S2

“Are S1 and S2 similar enough?”

Sensing for Proximity Detection

- Most modern mobile devices contain an array of sensors
  - Motion: **accelerometer, gyroscope, gravity…**
  - Environmental: **light, temperature, humidity, sound (via microphone)…**
  - Position: **GPS location, rotation vector, proximity…**

- Plenty of proposals on using these for payments, access control etc. [1-3].

- **Problem:** long sampling durations (up to 30 seconds). Impractical for impromptu payments: EMV mandates max transaction time of 500ms.

1. Halevi et al., "Secure Proximity Detection for NFC Devices Based on Ambient Sensor Data", ESORICS 2012
2. Mehrnezhad et al., "Tap-Tap and Pay: Preventing MITM Attacks in NFC Payments using Mobile Sensors", SSR 2015
3. Truong et al., "Comparing and Fusing Different Sensing Modalities for Relay Attack Resistance in ZIA", PerCom 2014

Outline

- **How well does ambient sensing fare under EMV restrictions?**

- We evaluate **17 sensors** available through the Android platform.

- Each sensor, where feasible (more later), was used to record **1,000 contactless transactions at four locations**, with a test base of **252 users**

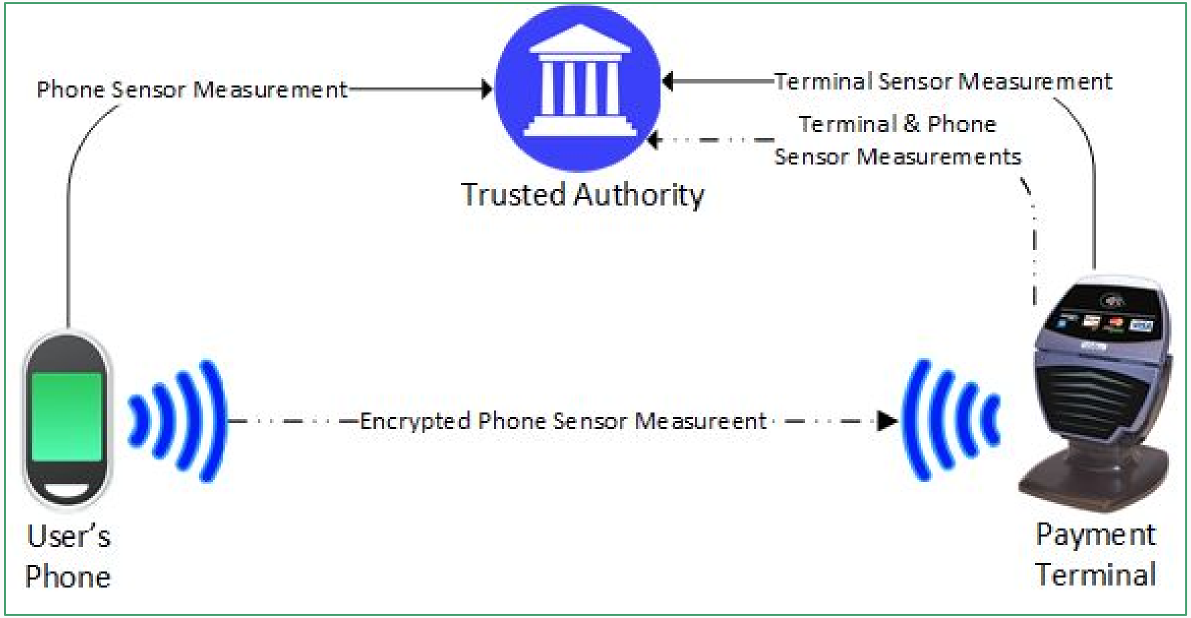
- Collected data was subjected to two evaluations:
    - **Threshold-based**: classic methodology for binary classification used in some work
    - **Machine learning:** evaluate several classifiers, e.g. SVM, Random Forest, Logistic Regression

Generic Architecture

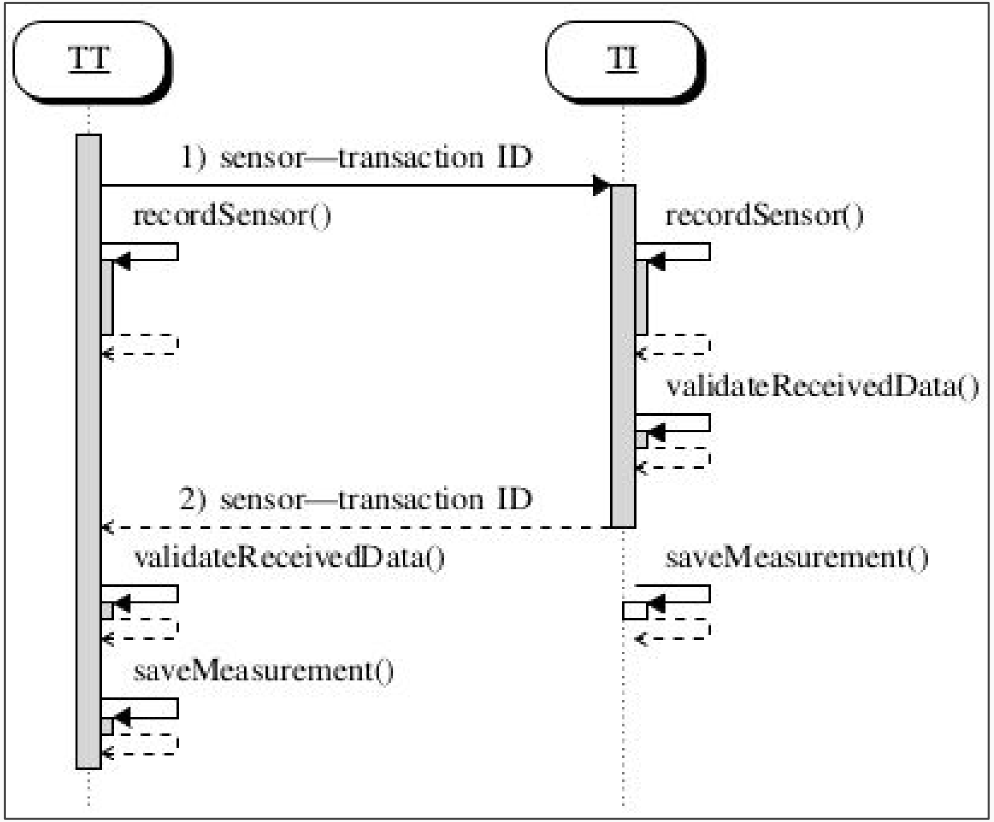
During the transaction, both the payment instrument (phone) and terminal collect measurements for a given sensor over 500ms

Sensor measurements are judged to be acceptable by some authority: on the terminal itself (locally), or transmitted to a remote authority

Transaction is rejected if sensor measurements are not 'similar' enough, implying a relay attack



Test-bed Overview



Sensor Selection

Problem 1: no single device includes all possible sensors

Four devices used to capture the widest range modalities: Nexus 9, Nexus 5, Samsung Galaxy S4 and SGS5 Mini

Problem 2: some sensors simply returned no values (or extremely few) within the 500ms limit, e.g. GPS and nearby WiFi access points.

For this paper, we removed these sensors from further analysis; 500ms limit was maintained throughout

TABLE 2: Sensor Availability

Sensors	Nexus 9 (1)	Nexus 9 (2)	Nexus 5	SGS5 mini
PI-PT Pair: Nexus 9 (1) → Nexus 9 (2)				
Accelerometer	✓	✓	✓	✓
Bluetooth	*	*	*	*
GRV†	✓	✓	*	✓
GPS	*	*	*	*
Gyroscope	✓	✓	✓	✓
Magnetic Field	✓	✓	✓	✓
Network Location	✓	✓	✓	✓
Pressure	✓	✓	✓	✗
Rotation Vector	*	*	*	*
Sound	✓	✓	✓	*
WiFi	*	*	*	*
PI-PT Pair: SGS5 mini → Nexus 5				
Gravity	○	○	✓	✓
Light	*	*	✓	✓
Linear Acceleration	○	○	✓	✓
Proximity	✗	✗	✓	✓
Unsupported				
Relative Humidity	‡	‡	‡	‡
Ambient Temperature	‡	‡	‡	‡

✓: Working properly. ✗: Not present on device. *: Technical limitations.
‡: Evaluated using Samsung Galaxy S4. ○: Returned only zero-values.
† Geomagnetic Rotation Vector.

Data Collection

Implemented a test-bed using the chosen sensors (using Android)

At four locations around our university: cafeteria, lab, dining hall and library

Location entered before deployment

User taps payment device on the terminal, NFC connection formed, both devices record measurements for 500ms for a given sensor

Users, recruited from nearby, were allowed to conduct as many transactions as they wanted (252 users in total)

Mock terminal (Nexus 5)

Mock payment device (Nexus 5)

Undergrad recruitment equipment (chocolate)

Sensor Reliability

Firstly, 100 test transactions were conducted to judge whether sensors could *collect anything within 500ms*

Suspected previously that collecting nearby WiFi APs and Bluetooth devices would struggle

Suspicions were also confirmed for GPS, temperature and humidity; these were discarded

Some sensors recorded values but the overall transaction failed, e.g. lost NFC connection. (Interestingly, highest rates were recorded with the SGS5 mini; device choice is a significant influence on transaction success)

TABLE 4: Usability and Reliability Analysis

Sensors	Total Transactions	Transaction Failures	Sensor Failures
Accelerometer	1025	13 (1.26%)	0 (0%)
Bluetooth	101	1 (0.99%)	99 (99.1%)
GRV	1019	8 (0.78%)	0 (0%)
GPS	101	1 (0.99%)	100 (99.10%)
Gyroscope	1022	11 (1.07%)	0 (0%)
Magnetic Field	1027	17 (1.65%)	0 (0%)
Network Location	1053	15 (1.42%)	960 (91.17%)
Pressure	1018	10 (0.98%)	0 (0%)
Rotation Vector	1023	14 (1.36%)	0 (0%)
Sound	1047	4 (0.38%)	0 (0%)
WiFi	100	0 (0%)	100 (100%)
Gravity	1165	143 (12.27%)	0 (0%)
Light	1057	37 (3.50%)	0 (0%)
Linear Acceleration	1175	159 (13.53%)	3 (0.3%)
Proximity	1071	58 (5.41%)	0 (0%)
Ambient Temperature	50	0 (0%)	47 (94%)
Relative Humidity	50	0 (0%)	47 (94%)

Evaluation Process

1. **Pre-analysis:** rule out any ineffective sensors under the EMV time limit

2. **Collection:** measurements for the remaining 11 sensors over approximately 1,000 individual transactions (ready for *off-line* analysis)

3. Two analyses
   - **Threshold-based:** can we find a simple threshold, *t*, which separates all il-/legitimate transactions? (Popular method in related work using the EER method)
   - **Machine learning:** accuracy of correctly identifying legitimate and legitimate transactions over a variety of algorithms (more powerful classification technique)

Evaluation Metrics (1)

- Chose Equal Error Rate (EER), popular metric for binary classification problems, e.g. fingerprint authentication
    - EER defined as the intersection of False Acceptance Rate (FAR) and False Rejection Rate (FRR)
        - A broad 'balancing' of usability (FRR) and security (FAR)

- Each transaction, T_i, has a corresponding transaction terminal (TT) and transaction instrument (TI) measurement set, i.e. $T_i = (TT_i, TI_i)$

- A transaction is legitimate if TT and TI are 'similar enough' (with respect to known legitimate and illegitimate transactions)

Evaluation Metrics (2)

- $T_i = (TT_i, TI_i)$ are considered to be legitimate transactions (1,000 per sensor)
- Illegitimate transaction set generated by pairing each TT_i with TI_j from other transactions ($i \neq j$)
    - Recall assumption that measurements are unique
        - Even those in the same location
    - Why? Relay attacks can occur in the same location
        - Imagine an attacker behind a victim in a store
- Huge dataset of ~1 million transactions

Threshold-based Analysis

- 'Similar enough' data implies the presence of a threshold, t, such that *similarity*(TT_i, TI_i) < t implies a legitimate T_i

- Calculate Equal Error Rate (EER) of each sensor over a range of observed thresholds from the collected data; compute FAR and FRR at each threshold, and find intersect

- Thresholds computed according to similarity measures:
    - Pearson's Correlation Coefficient [1]
    - Mean Absolute Error [2]
    - Many, many other similarity metrics possible, but we scope this paper to these

1. Mehrnezhad et al., "Tap-Tap and Pay: Preventing MITM Attacks in NFC Payments using Mobile Sensors", SSR 2015
2. Halevi et al., "Secure Proximity Detection for NFC Devices Based on Ambient Sensor Data", ESORICS 2012

Threshold Results

Findings: for both metrics, EERs are substantially above acceptable levels

Best performing sensor: Pressure with MAE (circled): 27% EER

This still implies accepting ~27% of illegitimate transactions incorrectly and rejecting the same number of legitimate ones

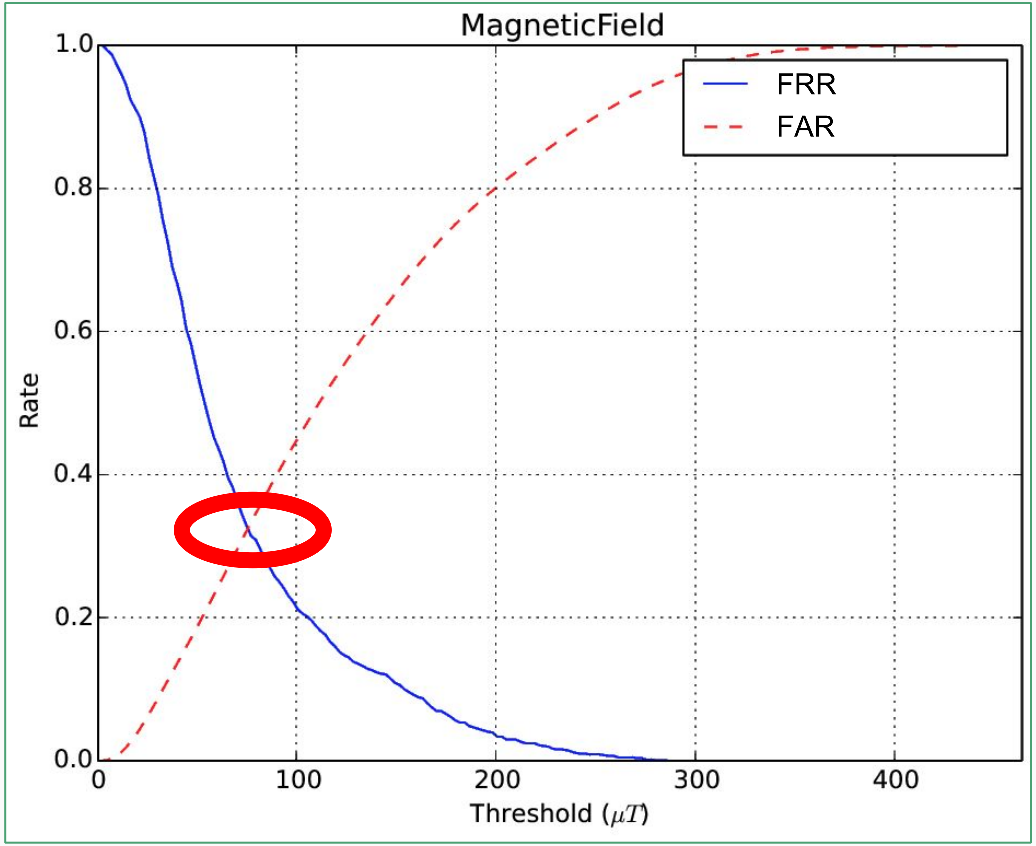
Most other sensors perform higher, e.g. 30-49% EER, indicating that observed sensor data isn't sufficiently discriminatory for these metrics (little difference between sensor pairs)

TABLE 3: Optimum Thresholds and Associated EERs

Sensors	**Optimum Threshold**$_{MAE}$	EER_{MAE}	**Optimum Threshold**$_{corr}$	EER_{corr}
Accelerometer	0.784	0.434	0.596	0.458
Ambient Temperature	–	–	–	–
Bluetooth	–	–	–	–
GRV	0.499	0.384	0.556	0.486
GPS	–	–	–	–
Gyroscope	0.614	0.443	0.636	0.441
Magnetic Field	76.12	0.323	0.495	0.384
Network Location	8.532	[illegible]	N/A*	N/A
Pressure	2.787	0.270	0.329	0.492
Rotation Vector	1.281	[illegible]	0.011	0.466
Relative Humidity	–	–	–	–
Sound	8.22	0.417	-0.022	0.488
WiFi	–	–	–	–
Gravity	9.93e-3	0.429	0.596	0.424
Light	182.1	0.488	0.020	0.496
Linear Acceleration	1.361	0.496	-0.020	0.426
Proximity	N/A†	N/A	N/A	N/A

*Insufficient data to calculate correlation
†All transactions contained the same value for both devices.

Example EER Curve: Magnetic Field with MAE



Machine Learning Analysis (1)

- *Can we do better than naive threshold-based measures?*
  Machine learning exists for such discrimination problems...

- Explored multitude of supervised learning classifiers: SVM, Naive Bayes, Decision Tree (C4.5), Random Forest, Logistic Regression and ML Perceptron

- Feature vector was the individual measurement differences between TT and TI
  - Rationale: simple similarity metrics across the measurement sets might not be a good starting point for providing discrimination between il-/legitimate transactions
  - Perhaps interactions between individual measurements can make this possible

Machine Learning Results

- Employed stratified 10-fold cross-validation per classifier (10 times)
  - Conducted using the WEKA toolkit
  - Six classification algorithms
- Best case: 9.2% EER for pressure sensor with Decision Tree

TABLE 5: Estimated EER for machine learning algorithms, obtained by repeating stratified 10-fold cross-validation 10 times

Dataset	Random Forest	Naive Bayes	Logistic Regression	Decision Tree	Support Vector Machine	Multilayer Perceptron
Accelerometer	62.6±2.4	50.9± 2.6	52.6± 2.3	50.0± 0.0	**49.8**± 2.5	55.1± 2.5
GeomagneticRotationVector	**43.5**±2.1	44.7± 2.4	47.4± 3.1	50.0± 0.0	48.9± 3.6	45.0± 2.6
Gravity	87.4±1.8	57.9± 2.0	57.9± 2.4	**50.0**± 0.0	**50.0**± 2.6	74.6±11.2
Gyroscope	68.3±2.7	**49.9**± 2.4	54.3± 2.4	50.0± 0.0	51.1± 2.5	51.4± 2.5
Light	57.6±2.6	51.5± 2.4	53.3± 2.5	**50.0**± 0.0	50.8± 2.4	51.3± 2.8
LinearAcceleration	60.3±2.5	50.7± 2.7	54.3± 2.3	**50.0**± 0.0	**50.0**± 2.1	55.4± 2.8
MagneticField	**29.2**±2.1	31.9± 2.0	32.2± 2.0	[illegible]	39.8± 4.6	32.9± 2.6
Pressure	10.3±1.0	10.7± 1.0	28.7± 1.3	**9.2**± 5.4	31.9± 4.5	11.4± 1.9
Proximity	49.9±3.1	53.7± 6.9	**47.6**±18.8	[illegible]	54.3± 25.4	50.8±19.7
RotationVector	**27.6**±4.6	56.3±24.3	59.6±23.3	50.0± 0.0	51.3± 24.3	48.8±24.5
Sound	**28.8**±1.9	31.4± 2.2	31.0± 2.1	34.7±13.6	41.1± 4.1	30.6± 2.0

Conclusion

- Evaluated a multitude of sensors using a variety of techniques
- Grounded ambient sensing under real-world constraints (EMV)
- Best result: 9.2% EER
  - Still too high as a suitable defence for sensitive scenarios, e.g. payments
  - What is acceptable?
    - Imagine ~1-in-10 transactions being denied at a crowded location, e.g. London Underground system (metro)
    - <1%, perhaps?

Future Research

- Generate data from a test-bed that reflects an actual relay attack, rather than synthetically generating illegitimate measurements
  - *We've already performed this; recently accepted at IEEE TrustCom '17*
  - *Sadly, results are still similar…*

- Use multiple sensors simultaneously
  - We used an in-depth but single sensor approach in this study
  - Multiple sensors to discriminate better, e.g. light **and** sound of a quiet, brightly-lit room
  - *Some challenges:*
    - Numerous sensor fusion techniques exist...
    - …and combinatorial explosion of potential sensors: which *n* sensors? *n*=3, 4,…,10?

Thanks for listening

Any questions?

Download our datasets and try yourself (link in the paper!)