Dempster-Shafer for score fusion with uncertainty in multi-modal continuous authentication schemes on mobile devices

Max Smith-Creasey, Muttukrishnan Rajarajan
City, University of London, Northampton Square, Clerkenwell, London, EC1V 0HB, UK
max.smith-creasey@city.ac.uk, r.muttukrishnan@city.ac.uk

Abstract

Interest in continuous mobile authentication schemes has increased in recent years. These schemes use sensors on mobile devices to collect biometric data about a user. The use of multiple sensors in a multi-modal scheme has been shown to improve accuracy but sensor scores are often combined naïvely using averaging techniques. The effects of uncertainty in score fusion has not been explored. We present a novel Dempster-Shafer based score fusion approach for continuous authentication schemes. Our approach combines sensor scores factoring in the uncertainty of the sensor. We propose and evaluate five techniques for computing uncertainty. Our proof-of-concept system is tested on three state-of-the-art datasets and compared with common fusion techniques. We find our approach yields the highest accuracies and achieves equal error rates as low as 8.05%.

1. Introduction

Mobile devices have become a ubiquitous part of modern life. This creates a need for protecting private data stored on devices. Traditional authentication techniques are knowledge-based where a user provides input such as a PIN, password or pattern. However, these have been shown to be vulnerable to smudge attacks and shoulder surfing [1].

The inclusion of sensors in mobile devices has seen biometrics used to authenticate users. Common biometrics used to authenticate include fingerprints and facial recognition. However, these continue to suffer from spoof attacks [2]. Furthermore, all discussed security mechanisms so far are *one-shot* authentication approaches; the device does not re-authenticate after it is unlocked.

Continuous authentication techniques have been proposed to alleviate issues with current authentication techniques [17]. Such schemes create a biometric profile using sensor data such that future data can then be compared to the user profile to authenticate. Most schemes combine several modalities due to enhanced accuracy and robustness [24].

One popular way to continuously authenticate is to sample device sensors periodically to build a profile of passively collected behavioural biometrics [12]. Such schemes combine multiple modalities (e.g.: movement sensors, location and wi-fi hotspots) and rely on the fact that humans are creatures of habit to authenticate. When attackers take a device from a user the behaviour will deviate from that of the genuine user and the device can subsequently lock.

Whilst there is research on including a variety of different modalities, there is little research on the effective fusion of scores obtained from sensor readings. Most continuous authentication approaches combine scores using techniques such as averaging. This does not factor uncertainty into the fusion and assumes absolute levels of belief in scores.

In this paper we propose a new continuous authentication scheme that uses Dempster-Shafer theory to incorporate uncertainty into sensor scores. We evaluate and discuss different ways of computing uncertainty. We show how our approach yields better accuracy against fusion techniques used in other multi-modal continuous authentication systems. The primary contributions of this paper are:

i) A new scheme for continuous authentication using Dempster-Shafer theory for sensor score uncertainty.

ii) We propose and evaluate multiple mechanisms for computing the uncertainty of sensor scores.

iii) We propose a *generalised impostor* for characterisation so the system does not require all impostor data.

The remainder of this paper is organized as follows. Section 2 explores the related work. Section 3 presents the underlying theory our scheme implements. In Section 4 our general idea is described. Section 5 discusses the experiments and discusses results. Section 6 concludes the research and discusses future work.

2. Related Work

In this section we discuss studies about biometric fusion, multi-modal authentication and Dempster-Shafer theory.

2.1. Biometric Fusion

In [18] the authors provide an overview of fusion in biometric systems. There are three levels at which data can be fused: the extraction-level, the score-level and the decision-level. Extraction-level fusion will often concatenate features into a single vector before classification. Score level combines scores from classifiers. Decision level uses binary decisions from classifiers to form a decision.

The most common method for biometric fusion is score-level fusion [7]. This level offers an ease of accessing and combining scores and does not suffer from the rigidity of decision-level fusion or potential feature incompatibilities of extraction level-fusion [9]. There are several techniques for combining classifier scores [11]. Commonly used techniques, due to simplicity, are averages, minimum score, maximum score, sum score and the product of all scores.

2.2. Multi-modal Continuous Authentication

In [22] a scheme is built on two weeks of phone, SMS, browser and location data from over 50 subjects. Probability density functions model the behaviour at different times of different days. Scores for different modalities are fused using weighted sum and product techniques. In [12] the authors authenticate with app usage. They train rule-based and neural-network techniques on user behaviour. The scheme achieves an equal error rate (EER) of 9.8% and can be adapted to a false acceptance rate (FAR) of 4.17% and a false rejection rate (FRR) of 11.45% through parameter adjustment. The scheme is, however, limited to authenticating with app usage. This is a visible behaviour and may vulnerable to shoulder-surfing attacks. Furthermore, once an app is authenticated there is no re-authentication during usage.

In [8] a scheme is shown combining orientation, accelerometer and touch-gesture data. The study evaluates min, max, product and sum fusion techniques with sum yielding the lowest EER of 0.31%. Similarly, in [15] the authors combine touch-gestures with accelerometer, orientation and power consumption data. They achieve EERs ranging from 6.1% to 6.9% using a majority voting fusion approach. Touch-gestures are also employed in [25] where passively collected multi-modal sensor scores adjust the threshold of touch-gesture authentication. The approach to the multi-modal component uses an averaging approach for score fusion. This unrealistically implies the same and constant certainty applies to all sensor scores.

In [20] the authors build a multi-modal scheme employing linguistic analysis, keystroke dynamics and behavioural profiling to authenticate. They fuse these modalities using sum and weighted average techniques, achieving an EER of 8%. The authors expand this work in [19]. The study achieves EERs as low as 3.3%. The authors implement sum fusion and matcher weighting [26] to combine scores but does not consider flexibility based on certainty. The scheme in [4] uses face, voice, keystroke and touch-gesture modalities on mobile devices. The system provides a framework but does not provide an implementation or results so the practicality is unknown. The scheme selects weighted sum score fusion, noting the ease and accuracy of the approach, but does not consider potential uncertainty.

Calls, SMS, browser and wifi data are used in [29] to authenticate users with an accuracy of 98.36%. However, these are limited modalities because they can be infrequent. This is expanded on in [30] with an adaptive neuro-fuzzy inference system. The scheme achieves competitive accuracies of 94.95%. The study uses only 5 users and therefore results are limited in their robustness. Sum fusion is used to combine some of the scores in the papers. In [5] stylometry, app usage, browsing and location data are used achieving an EER of 5%. However, the scheme uses binary classifiers that unrealistically use data from all impostors. Also, their applied data fusion centre requires a priori probabilities and does not consider uncertainty which limits flexibility.

A multi-modal scheme is proposed in [10] using probability density models spatial and temporal contexts. It shows impostor detection rates ranging from 53-99% depending on different attacker types. Scores are fused via a simple average. The study does not consider the influence of uncertainty in scores. In [13] the authors expand on this study to produce a more adaptive sensor sampling scheme due to the computational cost of sampling. However, the averaging is still used to fuse scores. In [6] a scheme is presented that collects wifi, bluetooth and location every 5 minutes to infer the environment security. Modality scores are fused via averaging techniques. The precision and recall the system achieves is 85% and 91%, respectively.

2.3. Dempster-Shafer for Sensor Systems

Dempster-Shafer theory has commonly been applied to sensor-based systems employing multiple sensors that may vary in their certainty or trustworthiness [27, 3]. The use of Dempster-Shafer theory in biometric systems is limited. In [23] the authors use the predictive ability of classifiers based on previous performance to form a belief in a hypothesis. In [16] Dempster-Shafer theory is applied to a biometric system incorporating face, fingerprint and iris. The authors construct uncertainty values as a function of the EERs achieved by the classifiers. They show that Dempster-Shafer improves fusion accuracy. However, the study is limited to three modalities and is not continuous.

3. Dempster-Shafer Theory

The scheme we propose uses Dempster-Shafer (D-S) theory [21] as the theoretical basis. D-S theory is a generalisation of probability theory and provides an evidence based framework that incorporates uncertainty. The theory is based on the ideas of obtaining degrees of belief for a

question and combining degrees of belief such they provide combined nuanced beliefs in hypotheses. The motivation for employing D-S theory into our framework is the ability to factor in the reliability in the sensor scores as uncertainty.

D-S theory requires a finite set of mutually exclusive and wholly exhaustive possibilities (e.g.: system states). This is the *frame of discernment*, represented by Ω such that $\Omega = \{P_1, P_2, ..., P_n\}$ where P_k represents a possibility. Any hypothesis, A, refers to a subset of Ω. The *power set*, 2^Ω, contains all possible hypotheses (known as *focal elements*) and can be derived from Ω. The power set 2^Ω contains all subsets of Ω, including itself and the null set $\emptyset$.

All subsets in 2^Ω are assigned a probability (degree of belief) known as a *basic belief mass* in a process called *basic belief assignment*. Belief in a hypothesis is computed by the sum of all sets in which it appears, contrary to traditional probability theory in which a single probability represents one atomic hypothesis. The mass function *m* maps each hypothesis A in 2^Ω to a value between 0 and 1 such that:

$$m(\emptyset) = 0, \text{ and } \sum_{A \in 2^\Omega} m(A) = 1 \tag{1}$$

The next assignment is given by a *belief function* that attributes a value $bel(A)$ between 0 and 1 to hypothesis A such that:

$$\text{bel}(A) = \sum_{B|B \subseteq A} m(B) \tag{2}$$

The final assignment to the hypotheses in 2^Ω is given by a *plausibility function* that assigns a value $pls(A)$ between 0 and 1 to hypothesis A such that:

$$\text{pls}(A) = \sum_{B|B \cap A \neq \emptyset} m(B) \tag{3}$$

The D-S framework carries the benefit that it is possible to derive the results of any two assignments provided that one assignment is available. Belief and plausibility are related by the following:

$$\text{pls}(A) = 1 - \text{bel}(\overline{A}) \tag{4}$$

Combining evidence from multiple observations for a set of focal elements is provided by Dempster's rule of combination. The *joint mass* in a hypothesis A provided by two observers *1* and *2* is given by $m(A) = m_1(A) \oplus m_2(A)$. This is calculated as the following orthogonal sum:

$$m_1(A) \oplus m_2(A) = \frac{1}{1-K} \sum_{B \cap C = A \neq \emptyset} m_1(B) m_2(C) \tag{5}$$

where:

$$K = \sum_{B \cap C = \emptyset} m_1(B) m_2(C) \tag{6}$$

4. Proposed Approach

In this section we describe our continuous authentication scheme as well as the components, architecture and how we incorporate Dempster-Shafer theory.

4.1. General Idea

This study introduces a novel and flexible continuous multi-modal behavioural authentication scheme for mobile devices. We hypothesise that introducing uncertainty into sensor scores can improve the accuracy and robustness of the scheme. Furthermore, we aim to show how uncertainty components can be manipulated due to temporal context or a growing lack of trust in sensor scores.

Our scheme continuously samples readings from sensors S_n on the mobile device. After a training phase of *m* days, collected sensor data is used to create behavioural models for the user. Because different behaviour can be expected at different times of day we use the concept of *anchors* [10, 25] where each anchor represents a behavioural model for a sensor for a period of the day. The probabilistic models are trained on the collected sensor data resulting in a model for each sensor for each anchor (assuming sensor data was available for that anchor). During this training phase the collected sensor data is also used to establish the uncertainty in the scores produced by each sensor (see Section 4.5).

Upon training the probabilistic behavioural models the scheme switches to an authentication mode. Sensors are continuously sampled during the this phase such that authentication is truly continuous. After each time period *p*, a collected window *w* of sensor readings is classified by the appropriate classifiers. If more than one sample from an individual sensor is present the scores are averaged.

Once likelihood scores are given by the behavioural models for each sensor for the period *p* they are fused in a score-level fusion strategy. Our novel approach applies D-S theory for score-level fusion (Section 4.4 expands on this). We use the obtained scores and the computed uncertainty *U* in those scores to model the belief in the related hypotheses. Dempster's rule of combination [25] is then used to fuse the belief that the user is genuine with consideration to the uncertainty represented in the scores. The final score is compared to a threshold. If the score surpasses the threshold access is maintained, otherwise an explicit authentication method (such as a PIN) is triggered.

4.2. User Profile Creation

As discussed in Section 4.1, our scheme uses probabilistic models to maintain the behavioural profiles of a user. These models are realised through the use of kernel density estimators (KDEs) (for sensors that output continuous values) and histograms (for sensors that output discrete values). These techniques are used due to their simplicity and popularity in similar studies [10, 25, 13].

The anchors are constructed to model temporal contexts and represented a time period of the day (e.g.: each hour). Sensor samples scored against these models result in a likelihood score that the sample is that of the genuine user. If there is no probabilistic model for a sensor at an anchor then it is omitted from the score fusion.

4.3. Generalised Impostor

Some uncertainty strategies in our scheme require knowledge of impostor scores to identify score distributions or sensor accuracies. In practice, we cannot know the scores that would result from all real-world impostors because we would not have access to their data. This unrealism is present in schemes such as [5] where the classifier is trained on the genuine user and then on all impostors (including the impostors the scheme is tested on). Our scheme therefore models a *generalised impostor* from a set of impostors not currently involved in the experiments.

The approach of a generalised impostor has been used with success in [28] for touch-gestures. The generalised impostor data can be scored on the trained scheme to generate generalised impostor scores. Therefore our scheme does not require data of real-world impostors but may use an initial set of impostors to calibrate the scheme. This approach is predicated on the assumption and likelihood that real-world impostors will share more in common with the generalised impostor than the genuine user.

4.4. Application of Dempster-Shafer Score Fusion

For a sensor S_n sample with a score of s and a current uncertainty U, the masses attributed to the hypotheses to satisfy the D-S formulation are given as:

$$m(\emptyset) = 0 \tag{7}$$

$$m(G) = (1 - U) \times s \tag{8}$$

$$m(I) = (1 - U) \times (1 - s) \tag{9}$$

$$m(E) = U \tag{10}$$

As defined for D-S in Section 3, $m(\emptyset) + m(G) + m(I) + m(E) = 1$. The uncertainty U is computed as defined in Section 4.5. When multiple sensors provide mass for a hypothesis, the combined belief can be computed using Dempster's rule of combination, discussed in Section 3.

4.5. Uncertainty Computation

In this section we describe the novel mechanisms we have devised for deriving the mass attributed to the uncertainty of each sensor score.

Accuracy: In this approach the uncertainty of a sensor score is based on the accuracy (in terms of EER) of the classifier, similar to [16]. Thus, uncertainty is proportional to the ability of a classifier to distinguish between genuine and impostor sensor data. The motivation for using accuracy-based uncertainty is that scores which are highly distinct will reduce uncertainty and increase trust. The uncertainty U_{S_n} for a score from a sensor S_n is given in Equation 11.

$$U_{S_n} = max(0,\ 1 - (2 * EER_{S_n})) \tag{11}$$

Quality: Here the uncertainty of a score is based on its statistical quality and stability. The aim is to increase uncertainty for scores that are sporadic and unstable. Therefore, uncertainty U_{S_n} is computed on prior scores from a sensor S_n for a user by a function of the mean μ (establishing how well the classifier recognises the scores) and the range r (identifying score consistency). It is given by the below:

$$U_{S_n} = ((1 - \mu_{s_n}) \ +\ r_{s_n})/2 \tag{12}$$

Temporally-aware Accuracy This method builds on our accuracy approach and introduces a temporal awareness at an anchor a_t to the uncertainty applied to each sensor score. The approach is based on the idea that different sensors will have different amounts of uncertainty present in the scores at different times of day. It is computed by:

$$U_{S_n,a_t} = max(0,\ 1 - (2 * EER_{S_n,a_t})) \tag{13}$$

Temporally-aware Quality This approach expands on quality by incorporating temporal awareness for anchor a_t to the uncertainty U_{S_n,a_t} applied to sensor scores. The approach is derived from the assumption that different sensors will have different amounts of uncertainty present in the scores at different times. It is given by:

$$U_{S_n,a_t} = ((1 - \mu_{s_n,a_t}) + r_{s_n,a_t})/2 \tag{14}$$

Temporally-aware Quality & Accuracy This approach combines the above temporal quality and accuracy uncertainty U_{S_n,a_t} mechanisms by computing their product. This mechanism is based on the assumption that the two separate mechanisms may compliment each other for improved accuracy. The equation is given as:

$$\begin{aligned} U_{S_n,a_t} = (((1 - \mu_{s_n,a_t}) + r_{s_n,a_t})/2) * \\ max(0,\ 1 - (2 * EER_{S_n,a_t})) \end{aligned} \tag{15}$$

5. Experimental Results and Discussion

In this section, we evaluate our framework. We first discuss the datasets and evaluation metrics. We then perform and discuss the experiments we perform with on our system to assess the accuracy and robustness.

Table 1. This table shows the modalities selected from each dataset to be used in our experimentation. For tri-axis motion sensor data the magnitude of the axes is used to represent a reading.

Modality	**Sherlock** [14]	**MSC** [25]	**GCU** [10]
Wi-fi	✓	✓	✓
Accelerometer	✓	✓	✓
Bluetooth	✓	✓	×
Location	✓	✓	×
Cell Towers	✓	✓	✓
Device Volume	✓	×	×
Gravity	×	✓	×
Gyroscope	✓	✓	✓
Ambient Light	✓	✓	✓
Call Info	✓	×	×
SMS Messages	✓	×	×
Magnetometer	✓	✓	✓
Activity	×	✓	×
Ambient Noise	✓	✓	✓

5.1. Datasets

In order to robustly validate our approach we use three state-of-the-art datasets. A summary of the modalities in these datasets is given in Table 1.

SHR (Sherlock) Dataset This is a long-term data collection project [14] employing 50 users on Samsung Galaxy S5 devices. Data collected is extensive with approximately 670 million records per month from sensors detailing hardware, software, telephony, network, motion, application and environment. Data is separated into quarterly partitions. We use the first quarter of 2016 and include 39 users that have consistent and daily samples from all sensors.

MSC (Mobile Sensor Collection) Dataset This dataset [25] was collected in 2017 and contains data from six volunteers using Nexus 4 mobile devices for a minimum of two-weeks each. Data is sampled continuously from motion, network, location and environmental sensors. We use all six participants in our study.

GCU (Glasgow Caledonian University) Dataset The GCU dataset [10] contains sensor data collected from the university staff and students in 2013. The dataset comprises of sensor readings from networks, applications, motion and environment. The data is collected from each user for a minimum of 14 days. The publicly available dataset that is used in this study contains data from four users.

5.2. Evaluation Metrics

We test the accuracy of our system using the following evaluation metrics that are common to biometric systems:

False Acceptance Rate (FAR): This is the rate that an impostor is wrongly classified as the genuine user.

False Rejection Rate (FRR): This is the rate that the genuine user is wrongly classified as an impostor.

Equal Error Rate (EER): When FAR is equal to FRR. FAR and FRR sets are obtained as an acceptance threshold is adjusted and are correlated such that if one increases the other decreases. In our experiments, EER=(FAR+FRR)/2 for the FAR and FRR with the smallest difference.

Receiver Operating Characteristic (ROC): This plots a curve that is used to assess the performance of a binary classifier system. ROC uses the axis of true positive rate and false positive rate for acceptance thresholds.

Table 2. This table shows EERs for the different uncertainty mechanisms for all datasets. Note that A, Q, TA, TQ and TQA are Accuracy, Quality, Temporally-aware Accuracy, Temporally-aware Quality and Temporally-aware Quality & Accuracy, respectively.

	A	**Q**	**TA**	**TQ**	**TQA**
SHR	15.60%	22.48%	15.12%	22.05%	16.91%
MSC	10.52%	39.68%	9.94%	36.42%	19.48%
GCU	14.10%	17.65%	13.83%	18.71%	15.77%

5.3. Uncertainty Experimentation

In this experiment we evaluate the uncertainty computation techniques (described in Section 4.5) that produce the lowest EER during D-S score fusion. We further demonstrate the uncertainty allocation to some sensors in the scheme. This experiment is performed on all datasets.

This experiment uses a training duration of 10 days because similar behavioural studies indicate this provides time for the profile to stabilise [10, 25]. We use the subsequent 7 days of data for testing (if available for the user). A data collection period of 1 minute is used such that authentication occurs every 1 minute on the data collected during that period. This ensures fast and responsive authentication. A period of 1 hour for the temporal anchors windows is used.

Characterisation in some uncertainty computation strategies requires positive and negative scores. We use 5-fold cross validation on the 10 days of the genuine user's training data to produce positive scores. To produce negative scores a generalised impostor is created by generating scores for the training data of all users not involved in the current test during the 5-fold cross validation. The scores are randomly selected such that the total for the positive and negative scores are equal (such that the system is kept lightweight).

Results for this experiment are shown in Table 2. We see the uncertainty that yields the lowest EER for all datasets is the temporal accuracy uncertainty, which is modelled on the ability of the sensor classifier to correctly distinguish a genuine user from an impostor during a time period. In Figure 1 we show uncertainties for sensors at different times of day to illustrate temporal differences in uncertainty. We find some sensors have consistently low uncertainty (such as wi-fi and Bluetooth) and others have frequently high uncertainty (such as light and noise sensors).

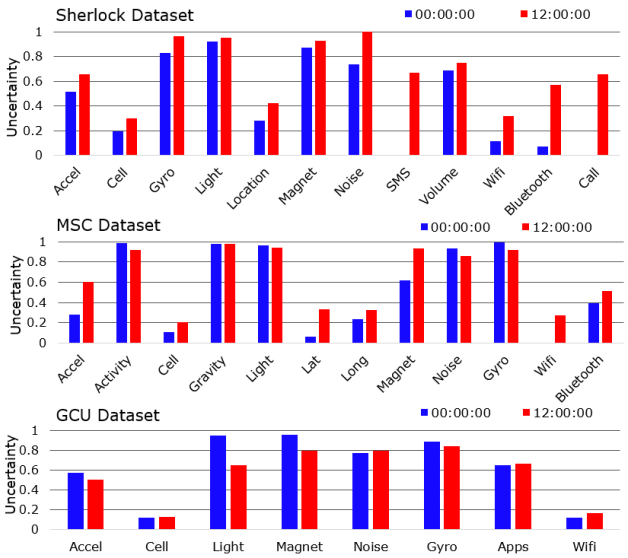

Figure 1. This figure shows the average uncertainties in the different sensor scores for users in all datasets at two different hours of day (00:00-01:00 and 12:00-13:00).

5.4. Comparison with other Fusion Techniques

In this experiment we compare our Dempster-Shafer based scheme with other commonly used biometric fusion techniques identified in the literature in Section 2. We find the most common approaches to score fusion in multimodal continuous authentication schemes are average and weighted average so these are used for comparison.

We adopt the approach used in the previous experiment with the uncertainty mechanism that provided the lowest EER. This time scores are also fused using average and weighted average fusion techniques. The weighting applied to scores in the weighted average technique is computed as $1 - U$. This experiment is performed on all datasets.

The results for this experiment are shown as ROC curves in Figure 2. We see that for all datasets the D-S fusion approach yields lower EERs than average and weighted average approaches. The EERs for each individual user in each dataset are shown in Figure 3. The D-S fusion approach shows lowered EER score groups when compared to the other techniques. These results indicate that D-S can yield lower EERs and achieve higher levels of accuracy.

5.5. Generalised Impostor Size

The experiments so far have used all users not involved in the test to model the generalised impostor. This experiment evaluates EERs whilst varying the number of impostors forming the generalised impostor. This experiment follows the previous approaches and uses the uncertainty mechanism that yielded the lowest EER (in Section 5.3). The experiment randomly selects an increasing number of users not involved in the each test to form the generalised impostor until EERs stabilise. This experiment uses only the Sherlock dataset because it provides the greatest number of users to vary the generalised impostor size.

The experiment is run and each time the impostors in the generalised impostor is increased. The results are shown in Figure 4. The figure shows a single impostor as the generalised impostor yields an EER of 18.85% but the EER stabilises at approximately 9 users and provides an EER of 15.13% (comparable to when the generalised user is made up of all 37 impostors in Table 2). We therefore posit that 9 users are sufficient for a generalised impostor.

5.6. Multiple Scores Window

Some continuous authentication systems use multiple authentication scores to form a final decision because it can be less volatile than a single result. This experiment therefore evaluates the EERs using a window of multiple D-S fused scores to form a decision. To do this we adopt the experimental approach used in the previous experiments that have yielded the lowest EERs so far.

Here, multiple scores are collected during a time window providing n D-S fused scores. We firstly use majority voting whereby the user is deemed genuine if a majority of scores are greater than the threshold. The second technique used is applying an average over the score window and comparing the average to the threshold.

The experiments show results for a window of 5 to 30 scores in Figure 5. We see the EER decreases as the window size increases. The change in EER appears small because score changes seem to occur for significant durations and therefore cannot be contained in the window sizes tested. The lowest EER is 8.05% for the MSC dataset with majority voting but requires a 30 minute window size. We note that larger window sizes improve EER but increases the window for attacks (due to larger time before re-authentication).

6. Conclusion

In this paper a novel continuous authentication scheme is produced for mobile devices using Dempster-Shafer theory for incorporating uncertainty in score fusion. The paper has evaluated multiple methods of computing uncertainty and compared the fusion approach to other commonly used used approaches. The scheme shows a generalised impostor provides characterisation without requiring data from the impostor being tested. Finally, it is shown that windows of scores are used to form a more accurate decision.

Future work will explore how the behavioural models can be adapted over time to include new behaviours such that concept drift is minimised. This will allow the scheme to mitigate the act of re-training the entire system every time behaviour changes. Furthermore, the future work will explore novel ways to decide when to collect data and authenticate. This is so the scheme can preserve battery life.

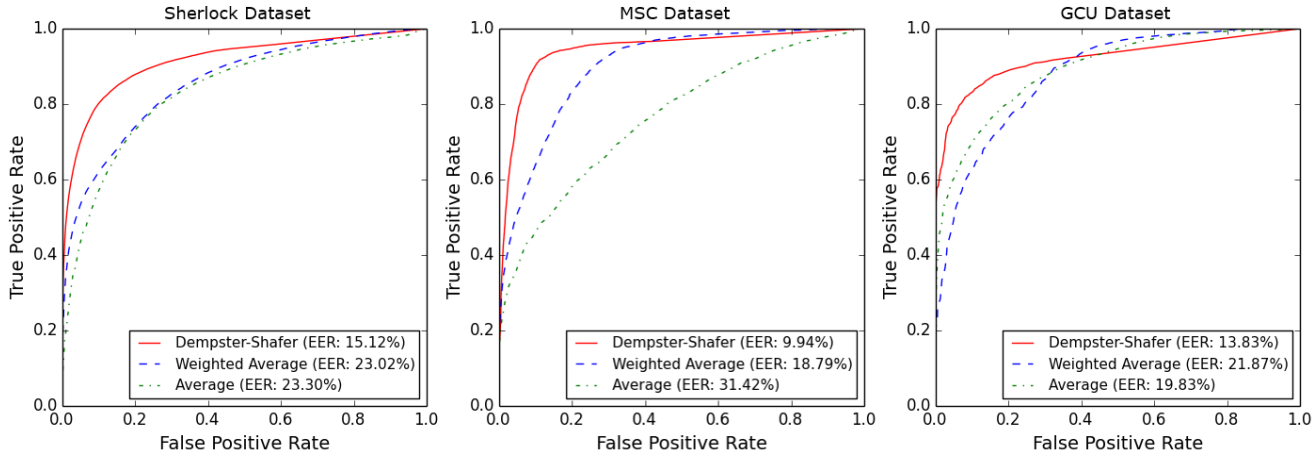

Figure 2. This figure shows the ROC curves for all three datasets when scores of the readings of different sensors are fused using average, weighted-average and Dempster-Shafer theory.

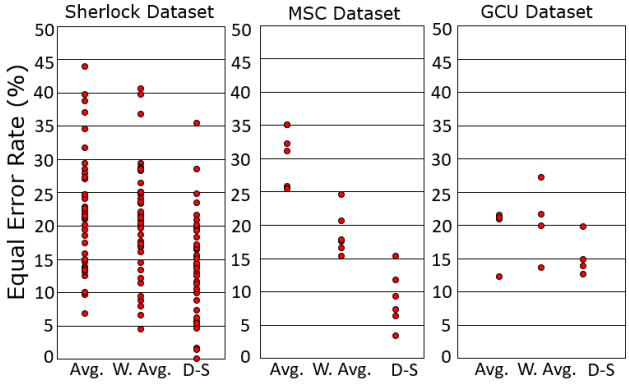

Figure 3. This figure shows the EERs obtained for each user (represented by red dots) in each dataset for the average, weighted average and Dempster-Shafer fusion techniques.

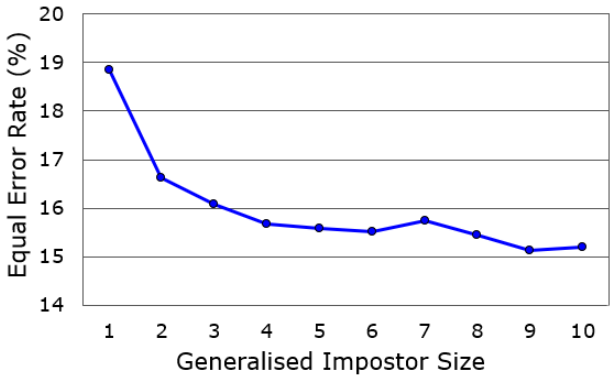

Figure 4. This figure shows the EERs obtained when the number of impostors forming the generalised impostor is varied.

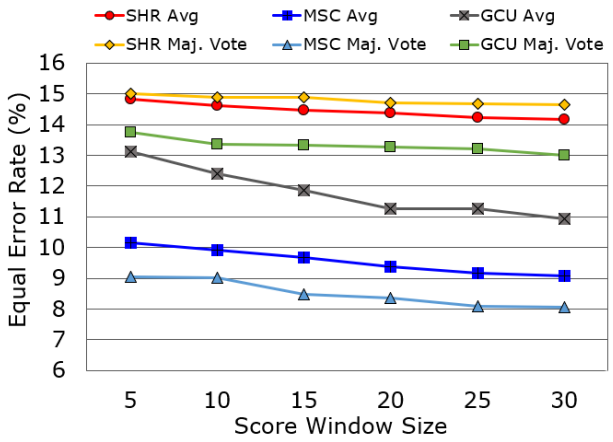

Figure 5. This figure shows EERs obtained when windows of multiple D-S fused scores are used to make a decision.

References

[1] F. Brudy, D. Ledo, S. Greenberg, and A. Butz. Is anyone looking? mitigating shoulder surfing on public displays through awareness and protection. In *Proceedings of The International Symposium on Pervasive Displays*, PerDis '14, pages 1:1–1:6, New York, NY, USA, 2014. ACM.

[2] K. Cao and A. K. Jain. Hacking mobile phones using 2d printed fingerprints. In *MSU Technical report*, 2016.

[3] T. M. Chen and V. Venkataramanan. Dempster-shafer theory for intrusion detection in ad hoc networks. *IEEE Internet Computing*, 9(6): pages 35–41, Nov 2005.

[4] G. Fenu, M. Marras, and L. Boratto. A multi-biometric system for continuous student authentication in e-learning platforms. *Pattern Recognition Letters*, 113: pages 83 – 92, 2018.

[5] L. Fridman, S. Weber, R. Greenstadt, and M. Kam. Active authentication on mobile devices via stylometry, application usage, web browsing, and gps location. *IEEE Systems Journal*, 11(2): pages 513–521, June 2017.

[6] A. Gupta, M. Miettinen, N. Asokan, and M. Nagy. Intuitive security policy configuration in mobile devices using context profiling. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, pages 471–480, Sept 2012.

[7] M. He, S.-J. Horng, P. Fan, R.-S. Run, R.-J. Chen, J.-L. Lai, M. K. Khan, and K. O. Sentosa. Performance evaluation of score level fusion in multimodal biometric systems. *Pattern Recognition*, 43(5): pages 1789 – 1800, 2010.

[8] A. Jain and V. Kanhangad. Exploring orientation and accelerometer sensor data for personal authentication in smartphones using touchscreen gestures. *Pattern Recogn. Lett.*, 68(P2): pages 351–360, Dec. 2015.

[9] A. Jain, K. Nandakumar, and A. Ross. Score normalization in multimodal biometric systems. *Pattern Recogn.*, 38(12): pages 2270–2285, Dec. 2005.

[10] H. G. Kayacik, M. Just, L. Baillie, D. Aspinall, and N. Micallef. Data driven authentication: On the effectiveness of user behaviour modelling with mobile device sensors. In *Proceedings of the Third Workshop on Mobile Security Technologies (MoST)*, 2014.

[11] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas. On combining classifiers. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(3): pages 226–239, Mar. 1998.

[12] F. Li, N. Clarke, M. Papadaki, and P. Dowland. Active authentication for mobile devices utilising behaviour profiling. *International Journal of Information Security*, 13(3): pages 229–244, Jun 2014.

[13] N. Micallef, H. G. Kayack, M. Just, L. Baillie, and D. Aspinall. Sensor use and usefulness: Trade-offs for data-driven authentication on mobile devices. In *2015 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pages 189–197, March 2015.

[14] Y. Mirsky, A. Shabtai, L. Rokach, B. Shapira, and Y. Elovici. Sherlock vs moriarty: A smartphone dataset for cybersecurity research. In *Proceedings of the 2016 ACM Workshop on Artificial Intelligence and Security*, AISec '16, pages 1–12, New York, NY, USA, 2016. ACM.

[15] R. Murmuria, A. Stavrou, D. Barbara, and D. Fleck. Continuous authentication on mobile devices using power consumption, touch gestures and physical movement of users. In *Proceedings of the 18th International Symposium on Research in Attacks, Intrusions, and Defenses - Volume 9404*, RAID 2015, pages 405–424, New York, NY, USA, 2015. Springer-Verlag New York, Inc.

[16] K. Nguyen, S. Denman, S. Sridharan, and C. Fookes. Score-level multibiometric fusion based on dempstershafer theory incorporating uncertainty factors. *IEEE Transactions on Human-Machine Systems*, 45(1): pages 132–140, Feb 2015.

[17] V. M. Patel, R. Chellappa, D. Chandra, and B. Barbello. Continuous user authentication on mobile devices: Recent progress and remaining challenges. *IEEE Signal Processing Magazine*, 33(4): pages 49–61, July 2016.

[18] A. Ross and A. Jain. Information fusion in biometrics. *Pattern Recognition Letters*, 24(13): pages 2115 – 2125, 2003. Audio- and Video-based Biometric Person Authentication (AVBPA 2001).

[19] H. Saevanee, N. Clarke, S. Furnell, and V. Biscione. Continuous user authentication using multi-modal biometrics. *Computers & Security*, 53: pages 234 – 246, 2015.

[20] H. Saevanee, N. L. Clarke, and S. M. Furnell. Multi-modal behavioural biometric authentication for mobile devices. In D. Gritzalis, S. Furnell, and M. Theoharidou, editors, *Information Security and Privacy Research*, pages 465–474, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[21] G. Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, 1976.

[22] E. Shi, Y. Niu, M. Jakobsson, and R. Chow. Implicit authentication through learning user behavior. In *Proceedings of the 13th International Conference on Information Security*, ISC'10, pages 99–113, Berlin, Heidelberg, 2011. Springer-Verlag.

[23] R. Singh, M. Vatsa, A. Noore, and S. K. Singh. Dempster-shafer theory based classifier fusion for improved fingerprint verification performance. In P. K. Kalra and S. Peleg, editors, *Computer Vision, Graphics and Image Processing*, pages 941–949, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

[24] M. Smith-Creasey and M. Rajarajan. A continuous user authentication scheme for mobile devices. In *2016 14th Annual Conference on Privacy, Security and Trust (PST)*, pages 104–113, Dec 2016.

[25] M. Smith-Creasey and M. Rajarajan. Adaptive threshold scheme for touchscreen gesture continuous authentication using sensor trust. In *2017 IEEE Trustcom/BigDataSE/ICESS*, pages 554–561, Aug 2017.

[26] R. Snelick, U. Uludag, A. Mink, M. Indovina, and A. Jain. Large-scale evaluation of multimodal biometric authentication using state-of-the-art systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(3): pages 450–455, March 2005.

[27] H. Wu, M. Siegel, R. Stiefelhagen, and J. Yang. Sensor fusion using dempster-shafer theory [for context-aware hci]. In *IMTC/2002. Proceedings of the 19th IEEE Instrumentation and Measurement Technology Conference (IEEE Cat. No.00CH37276)*, volume 1, pages 7–12 vol.1, May 2002.

[28] H. Xu, Y. Zhou, and M. R. Lyu. Towards continuous and passive authentication via touch biometrics: An experimental study on smartphones. In *10th Symposium On Usable Privacy and Security (SOUPS 2014)*, pages 187–198, Menlo Park, CA, 2014. USENIX Association.

[29] F. Yao, S. Y. Yerima, B. Kang, and S. Sezer. Event-driven implicit authentication for mobile access control. In *2015 9th International Conference on Next Generation Mobile Applications, Services and Technologies*, pages 248–255, Sept 2015.

[30] F. Yao, S. Y. Yerima, B. Kang, and S. Sezer. Continuous implicit authentication for mobile devices based on adaptive neuro-fuzzy inference system. In *2017 International Conference on Cyber Security And Protection Of Digital Services (Cyber Security)*, pages 1–7, June 2017.