FAST PHYSICAL OBJECT IDENTIFICATION BASED ON UNCLONABLE FEATURES AND SOFT FINGERPRINTING

Taras Holotyak, Sviatoslav Voloshynovskiy, Oleksiy Koval, and Fokko Beekhof

University of Geneva
Department of Computer Science
7, route de Drize, CH-1227, Switzerland

ABSTRACT

In this paper we advocate a new technique for the fast identification of physical objects based on their physical unclonable features (surface microstructures). The proposed identification method is based on soft fingerprinting and consists of two stages: at the first stage the list of possible candidates is estimated based on the most reliable bits of a soft fingerprint and the traditional maximum likelihood decoding is applied to the obtained list to find a single best match at the second stage. The soft fingerprint is computed based on random projections with a sign-magnitude decomposition of projected coefficients. The estimate of a bit reliability is deduced directly from the observed coefficients. We investigate different decoding strategies to estimate the list of candidates, which minimize the probability of miss of the right index on the list. The obtained results show the flexibility of the proposed identification method to provide the performance-complexity trade-off.

Index Terms— Identification, computational complexity, fingerprint, Physical Unclonable Function (PUF).

1. INTRODUCTION

The drastic evolution of the modern digital world raises a lot of challenging and emerging issues that in most cases concern a person/object identification. In contemporary environment customers require a high performance and quick feedback (low complexity) of identification protocols that operate in large, dynamically changeable and unstructured databases. The relevant need of distributed system architecture and remote computing formulates the problem of data integrity preservation as well. The aspects of system secrecy and data privacy have also to be controlled during the design of identification system. All these conflicting requirements allow the formulation of the identification as a complex constraint optimization problem [1, 2]. In the scope of such a problem, this paper represents an attempt to design an identification method that will provide an flexible performance-complexity trade-off.

2. FAST IDENTIFICATION METHOD

To resolve the performance-complexity trade-off in the identification problem, a subspace extracting approach based on random projections and the concept of bit reliability was proposed [3]. The main idea behind the random projection application consists in the removal of ambiguity about the data $prior$ statistics, while information about bit reliabilities is exploited to reduce the identification complexity in an optimal way. The schematic diagram of the proposed approach is shown in Fig. 1. Under such a formulation, the identification

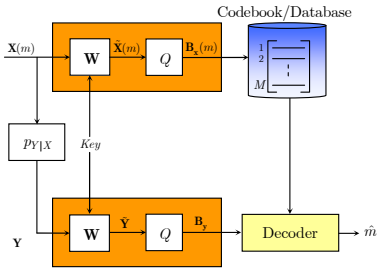


Fig. 1. Block-diagram of identification system.

system can be analyzed within digital communication framework. The presented identification system functions in two operating modes: enrollment and identification. During the enrollment, the PUFs, that are denoted as $\mathbf{x}(m) \in \mathbb{R}^N$, $m = 1, ..., M$, and considered as unique non-reproducible characteristics of objects, are acquired and transformed into the fingerprints using the following two stage procedure. First, $\mathbf{x}(m)$ of original dimensionality N are projected onto a J-dimensional $(J \leq N)$ space via:

$$\widetilde{\mathbf{x}}(m) = \mathbf{W}\mathbf{x}(m), \tag{1}$$

where $\mathbf{W} \in \mathbb{R}^{J \times N}$, $\mathbf{W} = (\mathbf{w}_1, ..., \mathbf{w}_J)^T$, and $W_{i,j} \sim \mathcal{N}(0, \frac{1}{N})$. The reason for such a projection matrix design is to ensure a certain invariance of the system to the input statistics of $\mathbf{X}(m)$. Indeed, it is not difficult to demonstrate that for any i.i.d. generated $\mathbf{x}(m)$, $\widetilde{\mathbf{x}}(m)$ will have Gaussian statistics with approximately preserved diagonal covariance

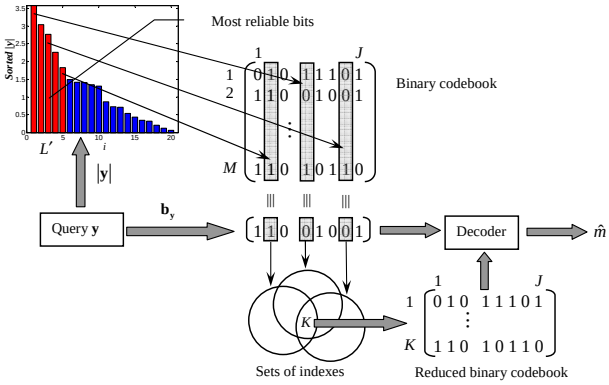


Fig. 2. Bit-reliability based identification decoder.

matrix, i.e., $\mathbf{K}_{\widetilde{x}} \approx \sigma_X^2 \mathbb{I}_J$, where $\mathbb{I}_J$ is an identity matrix in J-dimensional space, since $\mathbf{W}\mathbf{W}^{\mathbf{T}} \approx \mathbb{I}_J$. Secondly, the projection output is converted into the binary form as follows:

$$\mathbf{b}_{\mathbf{x}}(m) = sign\left(\mathbf{W}\mathbf{x}(m)\right). \tag{2}$$

The main purpose of the binarization stage is to amplify privacy of identification, as well as tackle data storage and computational complexity aspects. It is well known [1] that privacy protection is of primary importance in biometric applications. However, recent technological progress allowing microstructure modifications at nano level [4], extends the importance of privacy issues to a broader scope of problems, where the PUF based object identification is required.

In the identification mode, the query, which is distorted by discrete memoryless channel (DMC), $p\left(\mathbf{y}|\mathbf{x}\right)$, version of one of the enrolled objects, is converted farther to the binary format according to:

$$\widetilde{\mathbf{y}} = \mathbf{W}\mathbf{y} = \mathbf{W}(\mathbf{x}+\mathbf{z}),\ \mathbf{b}_{\mathbf{y}} = sign\left(\mathbf{W}\mathbf{y}\right). \tag{3}$$

It is important to note that, similarly to $\mathbf{x}$, any additive i.i.d. $\mathbf{z}$ will be converted into the additive Gaussian one with $\mathbf{K}_{\widetilde{z}} \approx \sigma_Z^2 \mathbb{I}_J$. Finally, the decoder that observes $\mathbf{y}$ and has access to the enrolled database should decide, which one out of M alternatives is present at the system input. In most identification system designs, the maximum likelihood (ML) decoder is used. In order to find a match between the channel output and the given codebook, this method performs an exhaustive search over the entire codebook according to the rule:

$$\hat{m} = \arg \max_{1 \le m \le M} (p(\mathbf{b}_{\mathbf{y}}|\mathbf{b}_{\mathbf{x}}(m))). \tag{4}$$

For the identification setup this solution is optimal in terms of performance, but characterized by complexity $\mathcal{O}(MJ)$. A possible alternative that leads to a reduction of the decoding complexity was proposed in [3]. In this case the match is performed within a reduced set of codewords only that is defined using bit reliability. For the appropriate selection of the set of reliable bits one can conjecture the presence of the sought codeword in this subset with a high probability. The structure of the proposed decoder is shown in Fig. 2. First, the given query is decomposed into magnitude and sign parts. The sign part is preserved for further matching, while magnitude components are used to evaluate bit reliability. Then, for each of L' most reliable bits of the query, the set of codewords $\mathcal{K}_i$, for which reliable bit matches exactly with the corresponding codeword bit, is composed. The final list of the candidates can be obtained by the exclusive combination of the above mentioned sets $\mathcal{K} = \cap_{i=1}^{L'} \mathcal{K}_i$. Finally, the ML decoding in this extracted subspace finalizes the identification procedure.

3. PROPOSED METHOD ENHANCEMENT

Previous results show that the reduction of complexity in the fast identification method leads to the performance loss [3]. The ML identification system is equivalent to a Bayesian M-ary hypothesis testing, where checking of the match is performed for all codewords (in the codebook of cardinality M), while the fast identification method can be represented by two stage hypothesis testing, where the match is sought within only the preselected subset of cardinality $K = |\mathcal{K}|$. For such a consideration, the average probability of error, P_e, of the ML decoder due to the symmetry of the codebook construction is defined as [5]:

$$P_{e\,ML}^{M} = \sum_{m=1}^{M} \Pr[\hat{m} \neq m|m] \Pr[m] \tag{5}$$

with $Pr[m] = 1/M$.

For the proposed fast method the probability of identification error consists of two terms and equals:

$$\begin{aligned} P_{e\,fast} &= P_{le} + (1 - P_{le}) P_{e\,ML}^{K} = \Pr[m \notin \mathcal{K}] \\ &+ (1 - \Pr[m \notin \mathcal{K}]) \sum_{m=1}^{K} \Pr[\hat{m} \neq m|m] \Pr[m], \end{aligned} \tag{6}$$

where P_{le} defines the list error and means that the correct codeword is not included into the reduced codebook $\mathcal{K}$ at the first stage of identification, $P_{e\,ML}^{K}$ represents the probability of identification error in the reduced codebook. Keeping in mind that $\mathcal{K} \subset \mathcal{M}$ ($\mathcal{M} = \{1, 2, \ldots, M\}$), the accuracy of identification in the reduced and entire codebooks satisfy $P_{e\,ML}^{K} \le P_{e\,ML}^{M}$, therefore, (6) can be simplified as follow:

$$P_{e\,fast} \le P_{le}(1 - P_{e\,ML}^{M}) + P_{e\,ML}^{M}. \tag{7}$$

Therefore, the performance degradation of the fast method vs. the ML decoder is mainly due to the $P_{le}(1 - P_{e\,ML}^{M})$, where the list error P_{le} plays a crucial role. Thus, further improvement of the fast identification method will consist in optimization of this parameter.

3.1. List decoding based on threshold combining of decision sets

The procedure of splitting the codeword bits into subsets of L' reliable and $(J - L')$ non-reliable ones was proposed to generate the reduced codebook based on reliable data only. The analysis in [3] shows that the subset of reliable bits is characterized by a significantly lower average probability of bit error $\bar{P}_b$ in comparison to the case of the entire codeword. However, since the probability of the list error is defined as follows:

$$P_{le} = 1 - \prod_{i=1}^{L'} (1 - P_{b\,i}), \tag{8}$$

where $P_{b\,i}$ is the probability of error in the i-th bit of the fingerprint, a failure in a single reliable bit identification can lead to a miss of the correct codeword while constructing the reduced codebook. $P_{b\,i} > 0$ highlights the disadvantage of the exclusive strategy for candidate indexes combining and requires a decisions fusion rules, which are resilient to errors.

This paper considers the threshold combining principle as one of the possible solutions. Each codeword is scored by the number of times the corresponding reliable bits of a codeword and the query coincide. The reduced binary codebook will be designed with the codewords, whose scores overcome a certain threshold Th. Obviously, this threshold belongs to the interval $[0, L']$, where $(Th = L')$ corresponds to the case of the exclusive combination [3]. The robustness of the threshold combining method is justified by a selection of the threshold that corresponds to the number of reliable bits reduced by the number of errors in the subset of reliable bits, i.e.,

$$Th = L' - T_{err}, \tag{9}$$

where $T_{err} = F^{-1}\left(1 - \epsilon, L', \bar{P}_b^{rel}\right)$ defines the number of errors that can appear in the subset of reliable bits with confidence $(1 - \epsilon)$, $\bar{P}_b^{rel}$ denotes average probability of bit error in the subset of reliable bits, $F^{-1}(.)$ is the inverse cumulative distribution function of the Binomial distribution.

3.2. Complexity issues of fast identification

Identification of items in unstructured databases is known to be an NP-hard problem and characterized by complexity that grows exponentially with the codeword length, i.e., $M \leq 2^{JR_{id}}$, where R_{id} denotes the identification rate [6]. Estimation of the method's complexity usually assumes evaluation of the worst case scenario of its behavior (complexity upper bound or big $\mathcal{O}$ notation) and neglecting the insufficient terms and coefficients. However, being NP-hard, both exhaustive search (ML decoder) and all existing methods with reduced complexity differ by only a coefficient in the order of the exponent. Therefore, the complexity analysis will be performed without such a simplification. Moreover, the complexity of the proposed method is parameterized by the threshold selection requiring the evaluation of not only the upper, but also the lower bound on the method complexity (big Ω notation) [7].

Assuming ideal conditions of the data access (data transfer between memory and processor has no computational costs), the identification complexity will be evaluated in terms of the number of summation/multiplication operations. For the simplicity of the complexity analysis the computational costs of one summation are considered to be equivalent to the costs of one multiplication. The complexity of the ML decoder can be expressed according to the following formula:

$$\mathcal{O}(M \cdot (2J - 1) + M) = \mathcal{O}(2MJ), \tag{10}$$

where term $\mathcal{O}(M \cdot (2J - 1))$ defines the number of the summation / multiplication operations for pair-wise similarity measures and $\mathcal{O}(M)$ is the complexity of the best candidate selection. The complexity of the fast method consists of the next terms: the first stage of identification (a search of the match with L' reliable bits $\mathcal{O}(ML')$, the codeword scores generation $\mathcal{O}(M(L' - 1))$, and the design of the reduced binary codebook $\mathcal{O}(M)$; the second stage of identification (a pair-wise similarity measurement in the reduced binary codebook $\mathcal{O}(K \cdot (2J - 1))$, and the resulting candidate selection $\mathcal{O}(K)$. Depending of the required performance-complexity trade-off, the complexity of the proposed method is limited by

$$\Omega\left(M \cdot \left(\frac{2J}{2^{L'}} + 2L'\right)\right) \text{ for } (Th = L') \tag{11}$$

and

$$\mathcal{O}\left(M \cdot (2J + 2L')\right) \text{ for } (Th = 1). \tag{12}$$

For the considered later experimental setup ($M = 16384, J = 512$) the above relations are visualized in Fig. 3. As it can be noted, the proposed method of fast identification operates in a wide range of complexities due to its parametrization by the threshold selection $(Th = 1, \ldots, L')$. In practice, operating at the first stage of the method with a small amount of reliable bits (region of small L'/J ratios) makes it possible to provide the average computational complexity of the identification method close to its lower bound with an acceptable loss in performance.

4. EXPERIMENTAL VALIDATION

This Section contains the experimental results that justify the properties of the described above fast identification method. The fingerprint database was simulated by the artificially generated set of texture patterns [8] with the cardinality of $M = 16384$. All acquisition and data processing distortions were modeled by the additive white Gaussian noise. The variance of the noise was selected to satisfy signal-to-noise-ratio (SNR) range -15...10 dB. Each PUF of the original size of 64×64 pixels $(N = 4096)$ was further projected onto $(J = 512)$-dimensional space. To demonstrate the relationship between different identification methods we

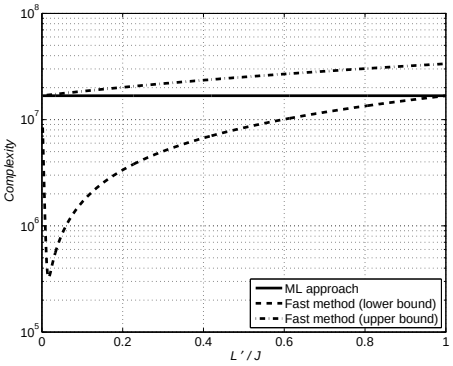


Fig. 3. Complexity of identification techniques.

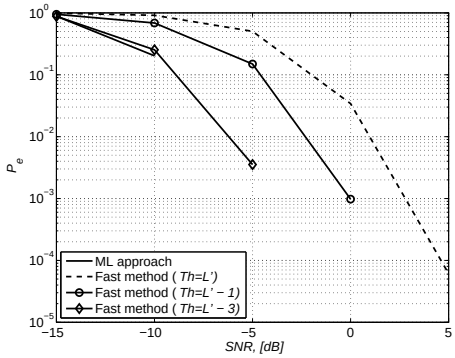
have compared: 1) the exhaustive search method (the ML approach); 2) the fast identification method with exclusive decisions fusion [3]; 3) the proposed identification technique with robust fusion of decisions. In two-stage identification (2) and (3), the reduced binary codebook was generated using $L' = 12$ the most reliable bits. The results of the experimental validation are presented in Fig. 4. They allow to conclude that the proposed fast identification technique based on the threshold combining can successfully and flexibly solve the identification performance-complexity trade-off. Various solutions of this trade-off were obtained for different threshold values. In particular, using the proposed strategy, about 10 time reduction in complexity is attained for an asymptotically small loss in performance for the SNR range of -15...-10 dB.

5. CONCLUSIONS

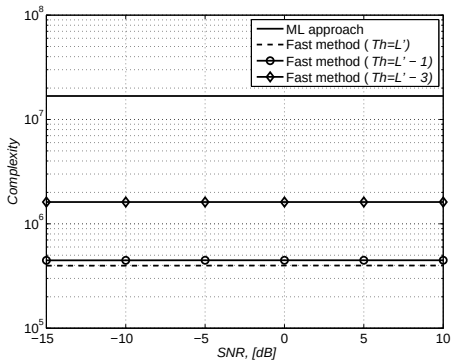
This paper considers the problem of performance-complexity trade-off in PUF identification with emphasis on the particularities of the fast search in unstructured databases. Based on the accurate analysis of the existing methods we have extended the proposed earlier fast identification technique that utilizes the concept of bit reliability. The robustness of the proposed method to different acquisition and data processing distortions is provided by application of the threshold combining approach to the reliable bit based decisions. Being parametrized by the threshold selection, the proposed identification method combines a nearly optimal performance with a reasonable computational complexity. The numerical experiments confirm the predicted characteristics of the proposed identification method.

6. REFERENCES

[1] B. Skoric P. Tuyls and T. Kevenaar, *Security with Noisy Data: Private Biometrics, Secure Key Storage and Anti-Counterfeiting*, Springer-Verlag, 2007.



(a) Identification performance



(b) Identification complexity

Fig. 4. Experimental comparison of identification methods.

[2] R. Pappu, *Physical One-Way Functions*, Ph.D. thesis, MIT, 2001.

[3] T. Holotyak, S. Voloshynovskiy, F. Beekhof, and O. Koval, "Fast identification of highly distorted images," in *Proc. of SPIE Photonics West, Electronic Imaging 2010 / Media Forensics and Security XII*, San Jose, USA, Jan. 21–24 2010.

[4] A. Margomenos, *Three dimensional integration and packaging using silicon micromachining*, Ph.D. thesis, University of Michigan, 2003.

[5] T. Cover and J. Thomas, *Elements of information theory*, Wiley, 1991.

[6] F. Beekhof, S. Voloshynovskiy, O. Koval, and T. Holotyak, "Fast identification algorithms for forensic applications," in *Proc. of IEEE Int. Workshop on Inf. Forensics and Security*, Dec. 6–9 2009.

[7] T. Cormen, C. Leiserson, R. Rivest, and C. Stein, *Introduction to Algorithms, 2nd Edition*, McGraw-Hill, 2001.

[8] J. Portilla and E. Simoncelli, "A parametric texture model based on joint statistics of complex wavelet coefficients," *Int. J. Comput. Vision*, vol. 40, no. 1, pp. 49–70, 2000.