

Privacy-Preserving Indexing of Iris-Codes with Cancelable Bloom Filter-based Search Structures

P. Drozdowski^{*†}, S. Garg[‡], C. Rathgeb^{*}, M. Gomez-Barrero^{*}, D. Chang[‡] and C. Busch^{*}

^{*} da/sec - Biometrics and Internet Security Research Group, Hochschule Darmstadt, Germany

[†] Norwegian Biometrics Laboratory, NTNU, Gjøvik, Norway

[‡] Indraprastha Institute of Information Technology, New Delhi, India

{pawel.drozdowski, christian.rathgeb, marta.gomez-barrero, christoph.busch}@h-da.de
{surabhigh, donghoon}@iiitd.ac.in

Abstract—Protecting the privacy of the enrolled subjects is an important requirement expected from biometric systems. In recent years, numerous template protection schemes have been proposed, but so far none of them have been shown to be suitable for indexing (workload reduction) in the computationally expensive identification mode. This paper presents a, best to the authors’ knowledge, first method in the scientific literature for indexing protected iris templates. It is based on applying random permutations to Iris-Code rows, and subsequent indexing using Bloom filters and binary search trees. In a security evaluation, the unlinkability, irreversibility and renewability of the method are demonstrated quantitatively. The biometric performance and workload reduction are assessed in an open-set identification scenario on the IITD and CASIA-Iris-Thousand datasets. The method exhibits high biometric performance and reduces the required computational workload to less than 5% of the baseline Iris-Code system.

I. INTRODUCTION

In recent years, interest in biometric systems have spiked with many large-scale deployments (*e.g.* national databases and border crossing control systems) appearing. Currently, the largest such system is the Indian National ID system, into which, at the time of this writing, 1.2 billion Indian residents have been enrolled [1] with multi-biometric data and unique identifier numbers. In the United Arab Emirates, the border control agency employs an iris-based blacklist system, which aims to prevent undesirable travellers (*e.g.* visa violators and criminals) from re-entering the country [2].

Those and similar deployments have to operate in the identification or duplicate-check modes. Due to the sheer size of such systems, they are faced with strenuous requirements in terms of biometric performance and computational workload. The naïve algorithm for such scenarios requires an exhaustive (1: N) database search, *i.e.* comparing the probe against all the references stored in the database. Notwithstanding the use of efficient hardware and parallelism, with the growing database sizes, the cost of executing such searches becomes computationally prohibitive. Simultaneously, the probability of false positives quickly becomes unacceptable. In [3], Daugman shows the probability of at least one false positive (P_N) occurring in a identification scenario to be: $P_N = 1 - (1 - P_1)^N$, where N is the number of enrolled subjects and P_1 the false positive probability of a one-to-one template comparison.

For this reason, research has been conducted into biometric *workload reduction*, whereby the exhaustive search is replaced with more advanced techniques. Those techniques often take advantage of the underlying biometric template data representation, thus facilitating efficient search strategies; for example through indexing or serial combination of algorithms. The aim thereof is to vastly reduce the necessary number of template comparisons per lookup, while maintaining or only insignificantly reducing the biometric performance achieved by the baseline, exhaustive algorithm. A biometric system in an open-set identification mode (*i.e.* without an identity claim) can be generalised to the classic nearest-neighbour search (NNS) problem. However, additional non-trivial challenges arise due to high dimensionality, as well as intra-class variation of the biometric data, which means that the biometric templates extracted from the reference and probe samples belonging to the same subject may be very similar, but (almost) never identical. Consequently, typical workload reduction approaches such as *indexing* need to be adapted to account for the challenging properties of the biometric data (see *e.g.* [4], [5], [6], [7], and [8] for a more comprehensive survey). Other approaches used in (iris) biometric systems include: *cascading algorithms*, whereby a computationally efficient (albeit less accurate) method first computes a shortlist of candidate identities, which is then searched exhaustively by a slower and more accurate comparator (see *e.g.* [9], [10], [11]); and *classification*, whereby the database is split into buckets containing certain template classes (*e.g.* based on gender, eye colour, some statistical properties etc.), with the exhaustive search only being performed inside the bucket corresponding to the probe (see *e.g.* [12], [13], [14]).

In addition to the aforementioned need for workload reduction, potential of data exposure is a large concern in biometric system deployments, where the stored data is, in most cases, secured using traditional encryption algorithms [15]. Once compromised, this can lead to serious problems such as identity theft, cross-matching without consent and severely limited renewability. Furthermore, centralised storage of sensitive personal and biometric data has been increasingly receiving attention from the general public and various non-governmental organisations, thus leading to widened legisla-

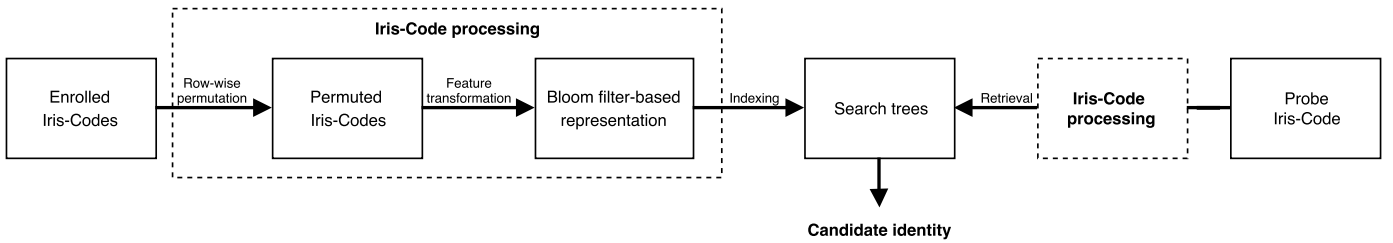


Fig. 1. An overview of the proposed system.

tion against privacy violations (e.g. GDPR in Europe [16]). Those matters have led to research into biometric *template protection* (see e.g. [17] and [18] for comprehensive surveys), with the aim of developing protection schemes especially dedicated for biometric data. Such systems must guarantee the properties stipulated by ISO/IEC IS 24745:2011 [19]:

Unlinkability It should be infeasible to determine whether or not two or more protected templates were derived from the same instance. This property prevents cross-matching across different databases.

Irreversibility Given a protected template and its corresponding secret, it should be infeasible to reconstruct the original biometric data. This property increases the security of the system against presentation and replay attacks.

Renewability It should be possible to issue new and revoke old protected templates from the same biometric instance and/or sample. This property ensures that in case of the biometric database being compromised, the data can be revoked and reissued, thereby preventing misuse.

Performance preservation The biometric performance is not significantly degraded by the template protection scheme.

With the aforementioned issues as motivation, this paper presents a, best to the authors' knowledge, first method in the scientific literature for indexing of protected iris templates. The method is based on Bloom filters and search trees (see [20] and [21]), which were previously shown to exhibit high workload reduction at an insignificant degradation to biometric performance, as well as scalability for an arbitrary number of enrollees. In this paper, said approach is extended by adapting ideas from [22] to accommodate cancelable iris templates which fulfil the aforesaid properties and are suitable for indexing.

The remainder of this paper is organised as follows: in section II, a method for privacy-preserving indexing of iris data is proposed. Section III presents the experiments and results, while section IV contains a summary and concluding remarks.

II. PRIVACY-PRESERVING INDEXING OF IRIS-CODES

In this section, the key components of the proposed system are presented. Subsection II-A describes a row-based permutation of Iris-Codes, while their transformation to a Bloom filter-based representation, as well as indexing and retrieval are outlined in subsection II-B. Figure 1 shows a schematic overview of the proposed system.

A. Row-based Permutation

To dissipate the statistical composition of the Iris-Code, a two-step feature rearrangement adapted from [22] is applied:

- 1) The Iris-Code is split into a small number of parts (IC_{parts}). The aim is to minimise the potential negative impact of the template protection on the biometric performance by preserving more spatial information. Several alternatives have been explored, namely: a) 2 parts – the real and imaginary response of the feature extractor; b) 4 or 8 parts – a further subdivision of each response into 2 or 4 parts, respectively.
- 2) A different row-based permutation is applied to each of the parts, which, as will be shown later (section III), makes inversion attacks infeasible (even under the *full-disclosure attacker model*, where the attacker is in possession of the permutation key). Potential loss of discriminative power due to the permutation is (mostly) avoided, since the horizontal neighbourhoods within rows persist. Note, that a column-wise permutation would not have had the desirable effect, due to the nature of Bloom filter-based Iris-Code representation explained in subsection II-B.

B. Indexing and Retrieval

Following the permutation of the Iris-Codes, the enrolled templates are organised into tree-based search structures following the methods of [20] and [21] described below.

- 1) The Iris-Codes are evenly split into j equally sized blocks of adjustable height and width ($H \times W$). Subsequently, a simple transformation function is applied to the blocks column-wise, whereby each column (a binary string), is mapped to its corresponding decimal value.
- 2) For each block, an empty (*i.e.* all bits set to 0) Bloom filter (\mathbf{b}) of length 2^H is created and the indices corresponding to the decimal column values are set to 1.
- 3) Hence, the resulting template (\mathbf{B}) is a sequence of j such Bloom filters - $[\mathbf{b}_1, \dots, \mathbf{b}_j]$.
- 4) The dissimilarity (DS) between two Bloom filter-based templates (denoted \mathbf{B} and \mathbf{B}') can be efficiently computed (utilising intrinsic CPU operations and trivially parallelisable), as shown in the equation below, where $|\cdot|$ represents the population count, *i.e.* Hamming weight.

$$DS(\mathbf{B}, \mathbf{B}') = \frac{1}{j} \sum_{i=1}^j \frac{|\mathbf{b}_i \oplus \mathbf{b}'_i|}{|\mathbf{b}_i| + |\mathbf{b}'_i|}$$

The Bloom filter-based templates are, to a certain degree, rotation-invariant, which means that contrary to the Iris-Codes, no alignment compensation is needed during the template comparison stage. Furthermore, the data representation is sparse, which is a crucial property for the indexing step described below. The representation sparseness is guaranteed, since for each Bloom filter of length 2^H , at most W (in practice fewer – due to the bit correlations in the Iris-Codes) indices are activated, and for the considered system configurations $W \ll 2^H$.

- 1) The list of N enrolled templates is (approximately evenly) split and assigned to T trees. This step is needed (for any sizeable N values) to maintain the sparseness of the data representation.
- 2) Each node of a tree (containing $M = \frac{N}{T}$ templates) is constructed through a union of templates, which corresponds to the binary OR applied to the individual Bloom filters in the sequence. The tree root is constructed from all templates assigned to the respective trees (*i.e.* $\bigcup_{m=1}^M \mathbf{B}_m$), while the children at subsequent levels are created each from half of the templates from their parent node (*e.g.* at first level – the children of the root node – $\bigcup_{m=1}^{\frac{M}{2}} \mathbf{B}_m$ and $\bigcup_{m=\frac{M}{2}+1}^M \mathbf{B}_m$).
- 3) The templates ($\mathbf{B}_1, \dots, \mathbf{B}_M$) are inserted as tree leaves.

After constructing the trees, the retrieval can be performed as shown below.

- 1) A small number of the most promising trees (t) out of T constructed trees can be pre-selected (denoted $\frac{1}{t}$) based on comparison scores between the probe and root nodes.
- 2) The chosen trees are successively checked until the first candidate identity is found or all the pre-selected trees have been visited. Note, that for the genuine transactions, thanks to the pre-selection step, the trees most likely to contain the sought identity are visited first.

A tree is traversed by, at each level, computing the comparison score between its nodes and the probe, and choosing the path with the best score. Once a leaf is reached, a final comparison takes place. The idea is based on the representation sparseness: as long as, at each level, the relation $DS_{genuine} \ll DS_{impostor}$ generally holds true, the genuine probes will be able to traverse the tree using the correct path to reach a matching leaf template. Note, that the row-based permutation (subsection II-A) does not, in any way, impair the representation sparseness, since the average number of activated indices remains identical for the Bloom filters produced from permuted and unpermuted Iris-Codes.

The complexity of a single lookup is $O(T + t * (2 * \log M))$. As it is sufficient to pre-select only a small fraction of the constructed trees, *i.e.* $t \ll T$, the lookup workload remains low, while arbitrarily many enrollees can be accommodated. For reference, figure 2 shows the indexing and retrieval in a single tree. If multiple trees are constructed, the search is trivially parallelisable by simultaneously traversing many trees at once.

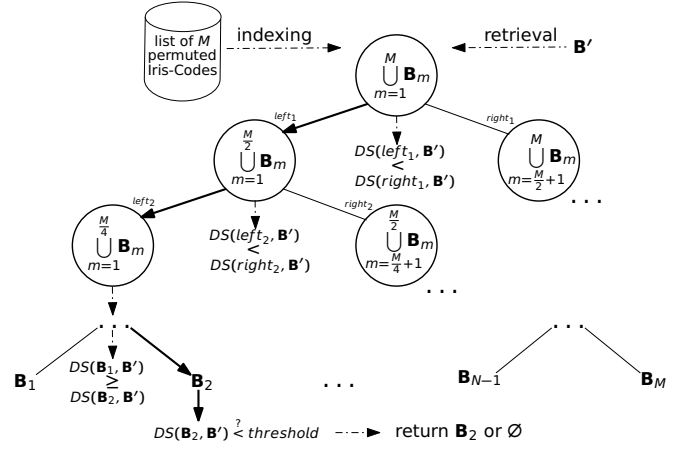


Fig. 2. Indexing and retrieval in the Bloom filter-based system. In this case, the retrieval follows the bold arrow path down to a leaf, where the final decision is made.

III. EXPERIMENTS

This section presents experiments performed to assess the proposed system. The experimental setup is outlined in subsection III-A, while the performance and privacy evaluations are presented in subsection III-B.

A. Experimental Setup

Two publicly available datasets of near-infrared iris images were chosen for the experiments: IITDv1 [23] and CASIA-IrisV4-Thousand [24] (henceforth referred to as "IITD" and "CASIA", respectively). They contain 1120 and 20000 images from 224 and 1000 subjects, respectively. Example images from the datasets are shown in figure 3.

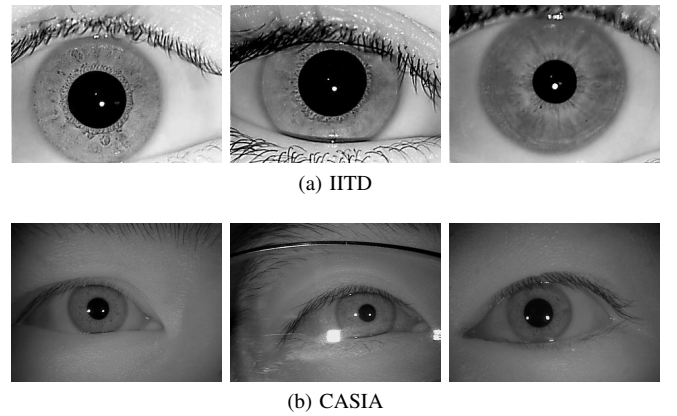


Fig. 3. Example images from the chosen datasets.

The raw images were processed with the commonly used methods using open-source libraries: OSIRIS [25] and USIT [26]. After segmentation, where the iris and pupil boundaries are located, the iris textures were normalised according to the rubbersheet model [27] and subsequently enhanced by applying Contrast Limited Adaptive Histogram Equalization (CLAHE). Features were extracted with the Daugman-like 1D-LogGabor algorithm (LG), generating 512x20 bits Iris-Codes.

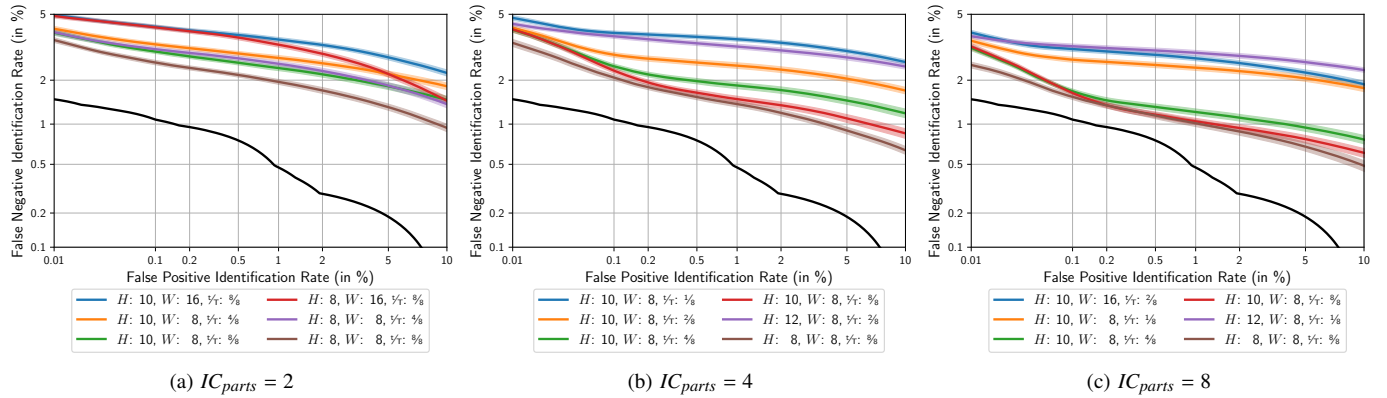


Fig. 4. DET curves for the proposed system. The faint colours around the curves represent the 95% confidence interval, while the black line represents the baseline (with EER of 0.66) – an Iris-Code system performing an exhaustive search and using ± 4 bit-shifts for sample alignment compensation.

For the experiments, 256 references (from the IITD dataset – left and right eye instances are mutually independent and thus treated as separate subjects) were enrolled. The entire CASIA dataset together with the remainder of the IITD data are used to supply an ample number of impostor comparison trials. To make the evaluation more robust, 50 random permutations are generated and used throughout the experiments. In other words, the performance evaluation for each system configuration is repeated 50 times with the different permutations of the Iris-Code templates.

Following metrics were used for evaluation of the various aspects of the proposed system:

Biometric performance: ISO/IEC IS 19795-1:2006 [28] metrics are used.

- The false positive and false negative identification rates plotted as detection error trade-off (DET) curves.
- The equal-error-rate (EER).

Workload: metrics from ISO/IEC IS 19795-1:2006 [28], and proposed in [21] is used.

- The penetration rate (p).
- The required number of bit comparisons per identification transaction expressed as a fraction (F) of the number of required Iris-Code baseline bit comparisons.

Template protection: metrics introduced in [29] are adopted.

- Unlinkability: the overall measure of the linkability of a given biometric template protection system ($D_{\leftrightarrow}^{\text{sys}}$). It is computed in terms of the probabilities of having a mated or non-mated comparison for each possible linkage score between templates enrolled in different applications. It yields values in the closed range $[0, 1]$, and reports a decreasing degree of unlinkability (*i.e.* increasing degree of linkability).
- Irreversibility: the success probability (P_{guess}) of guessing an original biometric template given a protected template under *full-disclosure attacker model* (*i.e.* the used permutation sequence is known to the attacker).
- Renewability: the number of possible permutation sequences, $|K|$ (*i.e.* the size of the key space).

B. Performance Evaluation and Protection Analysis

Figure 4 shows DET curves (with axes using a standard deviate scale [30]) for some of the best performing system configurations. Plots for each $IC_{\text{parts}} \in \{2, 4, 8\}$ show the three best configurations in terms of biometric performance and three best configurations in terms of workload reduction. The proposed protected indexing system exhibits high biometric performance, albeit naturally suffering a relatively small loss from the baseline Iris-Code based system. It can also be observed, that splitting the Iris-Code into more groups than just the real and imaginary parts prior to applying the permutation, is beneficial for the biometric performance. This is due to the fact that by splitting the Iris-Code into more parts, the potential for information loss due to permutation is decreased by preserving more spatial information. The plotted confidence intervals show that the biometric performance of the proposed system is stable across different permutations (in other words, changing the applied random permutation does not adversely affect the biometric performance of the system).

In table I, the workload and security parameters of the proposed system (for the configurations plotted in figure 4) are listed. A significant workload reduction is noticeable – the proposed system only requires between 1% and 10% of the workload incurred by the baseline system. This is achieved partly by decreasing the penetration rate as can be seen in the table, and partly by reducing the size of the biometric templates in terms of number of bits. Table I also shows the unlinkability, irreversibility and renewability of the proposed system¹. It can be readily seen, that the keyspace ($|K|$) for the proposed system is huge, thereby ensuring renewability and contributing to the infeasibility of reversing the protected templates (P_{guess}), which is further enhanced by the nature

¹In calculations, the average number of activated bits in the Bloom filters must be rounded to the nearest integer, thus in some cases the resulting P_{guess} may be equal for different H values (particularly when $H = 8$ or $H = 10$). Furthermore, since the *full-disclosure attacker model* is used, the further effort of reversing the row-wise permutation (which would have been different depending on H values) is not included in P_{guess} , since the attacker is assumed to be in possession of the used permutation sequences.

of the Bloom filter based representation (some loss of local information). Lastly, the measure of global unlinkability ($D_{\leftrightarrow}^{sys}$) for the tree leaves puts the proposed system (depending on the configuration) in close to fully unlinkable and semi-unlinkable region (as defined in [29]). Thus, for appropriate configuration selection, the security goals of a cancelable template protection scheme are accomplished.

TABLE I
RESULTS

IC_{parts}	H	W	$\frac{1}{4}$	EER	p	F	$D_{\leftrightarrow}^{sys}$	P_{guess}	$ K $	
2	8	8	%	1.96	0.31	0.063	0.32	2^{-960}	2^{10097}	
			%	2.15	0.19	0.038	0.29			
	10	8	%	3.01	0.31	0.063	0.45	2^{-1472}	2^{4414}	
			%	2.11	0.31	0.063	0.09	2^{-960}	2^{10097}	
4	8	8	%	2.71	0.19	0.038	0.10	2^{-1536}	2^{4414}	
			%	1.46	0.31	0.063	0.31			
			%	1.55	0.31	0.063	0.10	2^{-960}	2^{20195}	
			%	1.92	0.19	0.038	0.10			
	10	8	%	2.04	0.11	0.022	0.09			
			%	2.97	0.07	0.014	0.09			
	8	12	8	%	2.87	0.11	0.022	0.07	2^{-1080}	
				%	1.12	0.31	0.063	0.31		
10		8	%	0.92	0.31	0.063	0.11	2^{-960}	2^{40390}	
			%	1.15	0.19	0.038	0.09			
12	8	%	2.51	0.07	0.014	0.09				
		%	2.36	0.11	0.022	0.16	2^{-1536}	2^{17655}		
							2^{-1080}	2^{40390}		

IV. SUMMARY

In this paper, an approach for indexing cancelable iris templates has been proposed. The approach is based on a row-wise permutation of the Iris-Code rows and indexing them in Bloom filter-based tree structures. The experiments show that the proposed system fulfils the pre-requisites stipulated by ISO/IEC IS 24745:2011 for biometric template protection schemes (unlinkability, irreversibility, renewability and biometric performance), and additionally vastly reduces the workload associated with identification scenario – to less than 5% of the baseline system.

ACKNOWLEDGEMENTS

This work was partially supported by the German Federal Ministry of Education and Research (BMBF) as well as by the Hessen State Ministry for Higher Education, Research and the Arts (HMWK) within Center for Research in Security and Privacy (CRISP). One of the authors is supported by the TCS PhD Fellowship at IIIT-Delhi.

REFERENCES

- [1] Unique Identification Authority of India, “AADHAAR dashboard,” https://uidai.gov.in/aadhaar_dashboard/, last accessed: 2018-02-27.
- [2] A. N. Al-Raisi and A. M. Al-Khoury, “Iris recognition and the challenge of homeland and border control security in UAE,” *Telematics and Informatics*, vol. 25, no. 2, pp. 117–132, May 2008.
- [3] J. Daugman, “Biometric decision landscapes,” University of Cambridge - Computer Laboratory, Tech. Rep. UCAM-CL-TR-482, January 2000.
- [4] R. Mukherjee and A. Ross, “Indexing iris images,” in *2008 19th Intl. Conf. on Pattern Recognition*, 2008, pp. 1–3.
- [5] F. Hao, J. Daugman, and P. Zielinski, “A fast search algorithm for a large fuzzy database,” *IEEE Trans. on Information Forensics and Security*, vol. 3, no. 2, pp. 203–212, 2008.

- [6] H. Mehrotra, B. Srinivas, and B. Majhi, “Indexing iris biometric database using energy histogram of DCT subbands,” *J. of Communications in Computer and Information Science*, vol. 40, pp. 194–204, 2009.
- [7] R. B. Gadde, D. Adjeroh, and A. Ross, “Indexing iris images using the burrows-wheeler transform,” in *2010 IEEE Intl. Workshop on Information Forensics and Security*, 2010, pp. 1–6.
- [8] H. Proença, “Iris biometrics: Indexing and retrieving heavily degraded data,” *IEEE Trans. on Information Forensics and Security*, vol. 8, no. 12, pp. 1975–1985, 2013.
- [9] J. E. Gentile, N. Ratha, and J. Connell, “An efficient, two-stage iris recognition system,” *Intl. Conf. on Biometrics: Theory, Applications and Systems*, pp. 211–215, 2009.
- [10] M. Konrad, H. Stögner, A. Uhl, and P. Wild, “Computationally efficient serial combination of rotation-invariant and rotation compensating iris recognition algorithms,” in *Intl. Conf. on Computer Vision Theory and Applications*, 2010, pp. 85–90.
- [11] C. Rathgeb, A. Uhl, and P. Wild, “Incremental iris recognition: A single-algorithm serial fusion strategy to optimize time complexity,” *Intl. Conf. on Biometrics: Theory, Applications and Systems*, 2010.
- [12] X. Qiu, Z. Sun, and T. Tan, “Global texture analysis of iris images for ethnic classification,” *Advances in Biometrics*, pp. 411–418, 2005.
- [13] A. Ross and M. S. Sunder, “Block based texture analysis for iris classification and matching,” *Conf. on Computer Vision and Pattern Recognition - Workshops*, pp. 30–37, 2010.
- [14] H. Zhang, Z. Sun, T. Tan, and J. Wang, “Iris image classification based on hierarchical visual codebook,” *Trans. on Pattern Analysis and Machine Intelligence*, vol. 36, no. 6, 2014.
- [15] K. Nandakumar and A. K. Jain, “Biometric template protection: Bridging the performance gap between theory and practice,” *IEEE Signal Processing Magazine*, vol. 32, no. 5, pp. 88–100, September 2015.
- [16] European Parliament, “Regulation (EU) 2016/679,” *Official Journal of the European Union*, vol. L119, pp. 1–88, April 2016.
- [17] C. Rathgeb and A. Uhl, “A survey on biometric cryptosystems and cancelable biometrics,” *EURASIP J. on Information Security*, 2011.
- [18] C. Rathgeb, J. Wagner, and C. Busch, “Iris biometric template protection,” in *Iris and Periocular Biometric Recognition*. IET, 2017.
- [19] ISO/IEC JTC 1/SC 27 IT Security techniques, *ISO/IEC 24745:2011. Information technology – Security techniques – Biometric information protection*, International Organization for Standardization and International Electrotechnical Committee, June 2011.
- [20] C. Rathgeb, F. Breiting, H. Baier, and C. Busch, “Towards bloom filter-based indexing of iris biometric data,” in *2015 Intl. Conf. on Biometrics (ICB)*, May 2015, pp. 422–429.
- [21] P. Drozdowski, C. Rathgeb, and C. Busch, “Bloom filter-based search structures for indexing and retrieving iris-codes,” *IET Biometrics*, 2017.
- [22] M. Gomez-Barrero, C. Rathgeb, G. Li, R. Ramachandra, J. Galbally, and C. Busch, “Multi-biometric template protection based on Bloom filters,” *Information Fusion*, vol. 42, pp. 37–50, July 2018.
- [23] A. Kumar and A. Passi, “Comparison and combination of iris matchers for reliable personal authentication,” *Pattern Recognition*, vol. 43, no. 3, pp. 1016–1026, March 2010.
- [24] Chinese Academy of Sciences’ Institute of Automation, “CASIA iris image database,” <http://biometrics.idealtest.org/>, December 2010, last accessed: 2018-02-27.
- [25] N. Othman, B. Dorizzi, and S. Garcia-Salicetti, “OSIRIS: An open source iris recognition software,” *Pattern Recognition Letters*, vol. 82, no. 2, pp. 124–131, September 2016.
- [26] C. Rathgeb, A. Uhl, P. Wild, and H. Hofbauer, “Design decisions for an iris recognition SDK,” in *Handbook of Iris Recognition*, 2nd ed., ser. Advances in Computer Vision and Pattern Recognition. Springer, 2016.
- [27] J. Daugman, “How iris recognition works,” *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 14, no. 1, pp. 21–30, January 2004.
- [28] ISO/IEC 19795-1:2006. *Information Technology – Biometric Performance Testing and Reporting – Part 1: Principles and Framework*, International Organization for Standardization and International Electrotechnical Committee, April 2006.
- [29] M. Gomez-Barrero, J. Galbally, C. Rathgeb, and C. Busch, “General framework to evaluate unlinkability in biometric template protection systems,” *IEEE Trans. on Information Forensics and Security*, vol. 3, no. 6, pp. 1406–1420, June 2018.
- [30] A. Nautsch, D. Meuwly, D. Ramos, J. Lindh, and C. Busch, “Making likelihood ratios digestible for cross-application performance assessment,” *IEEE Signal Processing Letters*, vol. 24, no. 10, pp. 1552–1556, October 2017.