# SIC-Gen: A Synthetic Iris-Code Generator

Pawel Drozdowski[1,2], Christian Rathgeb[2], Christoph Busch[2]

**Abstract:** Nowadays large-scale identity management systems enrol more than one billion data subjects. In order to limit transaction times, biometric indexing is a suitable method to reduce the search space in biometric identifications. Effective testing of such biometric identification systems and biometric indexing approaches requires large datasets of biometric data. Currently, the size of the publicly available iris datasets is insufficient, especially for system scalability assessments. Synthetic data generation offers a potential solution to this issue; however, it is challenging to generate data that is both statistically sound and visually realistic - for the iris, the currently available approaches prove unsatisfactory.

In this paper, we present a method for generation of synthetic binary iris-based templates, i.e. Iris-Codes, which are the *de facto* standard used throughout major biometric deployments around the world. We validate the statistical properties of the synthetic templates and show that they closely resemble ones produced from real ocular images. With the proposed approach, large databases of synthetic Iris-Codes with flexibly adjustable properties can be generated.

**Keywords:** Biometrics, Iris Recognition, Iris-Code, Synthetisation

## 1 Introduction

The iris is one of the most widely applied biometric modalities. In recent years, several large-scale deployments have been created, most notably the Indian National ID program [Un10], which has, at the time of this writing, enrolled over one billion subjects with biometric data including the irides. Despite using efficient comparators (e.g. Hamming distance for the iris) and parallelism, the computational load faced by such deployments in the identification scenario is extremely high. With biometric workload reduction as a motivation, many approaches for indexing of iris data have been developed [PN17]. However, evaluation of such approaches and their scalability is often questionable due to lack of large test datasets. While various publicly available iris databases with near-infrared (NIR) data exist, they are relatively small. At the time of this writing, some of the largest publicly available datasets, CASIA-IrisV4-Thousand and ND-CrossSensor-Iris-2013, contain merely 20.000 images from 1000 subjects and 146.550 images from 676 subjects, respectively. This is several orders of magnitude smaller than some of the large-scale deployments nowadays.

Synthetic data generation is one possible way of dealing with the issue of testing efficient indexing methods. Most of the existing approaches for synthetic iris generation attempt to synthesise an entire iris image or texture [Le03, Cu04, MR05, WSG05, ZS05, SR06,

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

[2] da/sec – Biometrics and Internet Security Research Group, Hochschule Darmstadt, Germany
{pawel.drozdowski,christian.rathgeb,christoph.busch}@h-da.de

ZSC07, WTS08, PN13]. The main issues with such approaches include the computational costs and the difficulty in guaranteeing the statistical properties of the real data. The vast majority of operational iris biometric systems are based on the Iris-Code [Da04], making it a *de facto* standard. Generating Iris-Codes (feature vectors) directly is therefore also viable and may offer better control over the statistical properties of the synthetic data. Recently, two such approaches have been proposed. Proença and Neves [PN13] provide a method of Iris-Code synthesis based on bit correlations; the method is shown to attain some of the desired statistical properties (the shapes of the genuine and impostor distributions). It is also somewhat flexible with adjustable parameters; however, it does not allow to generate a set of templates following a desired score distribution. Furthermore, the filter response resulting from the typical feature extraction process is not modelled (in other words, the produced synthetic Iris-Codes scantily resemble the ones produced from real iris images through the commonly used iris processing pipeline). Lastly, typical error patterns between two mated templates are not modelled. Daugman [Da16] proposed to use a simple hidden Markov model to generate a stream of bits and showed that it can be adjusted, so that the produced templates mimic the impostor distribution of real iris templates. However, the produced streams are 1-dimensional (i.e. do not model the correlation between the Iris-Code rows); furthermore, the method does not offer a way to generate more than one template per subject (i.e. it is not possible to use it for simulating genuine comparisons). As such, it might only be useful for stress-testing of iris identification systems.

In this paper, we present a synthetic Iris-Code generator, which both reflects the statistical properties of the real Iris-Codes and resembles the real templates visually. An important feature of the proposed approach is its flexibility, in that it allows to generate Iris-Codes with an arbitrary resolution and an arbitrary score distribution of mated templates, unlike any of the approaches currently in the literature. To facilitate reproducible research, the software written in Python3 programming language, is released to the scientific community under a permissive license.

The remainder of this paper is organised as follows: section 2 describes the proposed method of synthetic Iris-Code generation. In section 3 the properties of the generated templates are validated, while section 4 contains concluding remarks.

## 2 Proposed Method

When generating synthetic Iris-Codes, several matters have to be taken into account:

- Dataset

  **Score distributions** The distributions of Hamming distance scores must closely resemble the ones produced by real data.

  **Degrees of freedom** Based on a large number of comparison scores from non-mated templates, the effective number of independent bits (degrees of freedom) can be calculated. Degrees of freedom can be seen as discrimination entropy as a measure of information content in iris images and has to be close to that of the real data.

- Individual templates

  **Bit correlation** The bits in an Iris-Code are far from independent. There exist correlations between both rows and columns, which result in long sequences of identical consecutive bits. The reason for this is partially the anatomy of iris patterns, as well as the nature of the commonly used feature extractors [Da16]. Those correlations have to be reflected in the synthetic data.

  **Error patterns** The majority of bit mismatches between two mated Iris-Codes occurs for bits resulting from wavelet response close to 0 (i.e. where the response phase changes). Those occur mostly on the edges of the bit sequences, and are called the ”fragile” bits [HBF09]. They have to be present in the synthetic data. Additional noise sources, such as the occlusions resulting from the eyelids, have to be modelled as well.

  **Rotation** In the real data, rotations of the eye, which are mainly caused by head tilts (i.e. roll pose), potentially result in misalignment between two mated samples. In Iris-Codes, this is represented by circular horizontal shifts of the matrix columns, which have to be modelled in the synthetic data.

The proposed generator synthesises Iris-Codes as pairs of mated templates, referred to as Iris-Codes *IC1* and *IC2* in the algorithm description and figure 1 below. The bold-filled arrows denote the changes to the template throughout the process, while the thin arrows denote the system parameters.

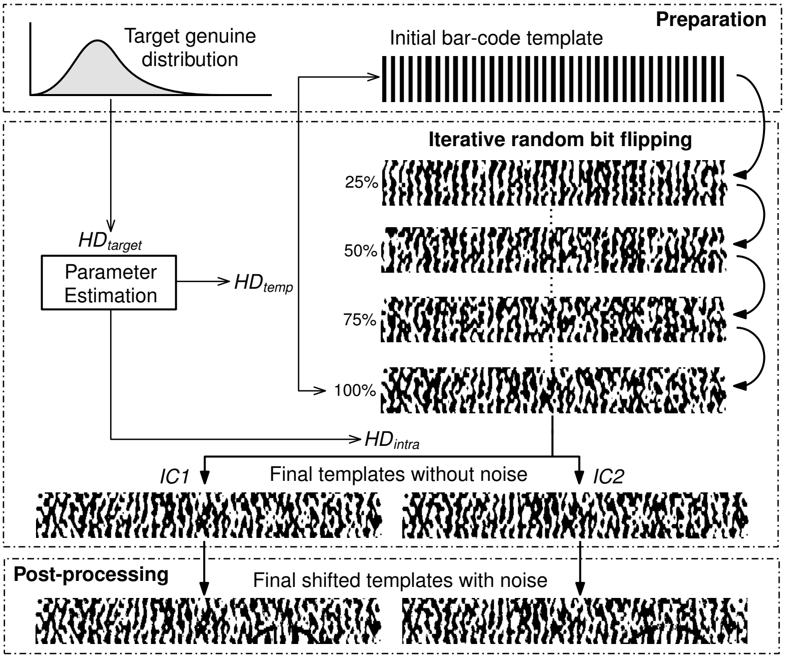

Preparation
Target genuine distribution
Initial bar-code template
Iterative random bit flipping
25%
50%
75%
100%
$HD_{target}$
Parameter Estimation
$HD_{temp}$
$HD_{intra}$
IC1
Final templates without noise
IC2
Post-processing
Final shifted templates with noise

Fig. 1: The process of generating an Iris-Code pair with SIC-Gen

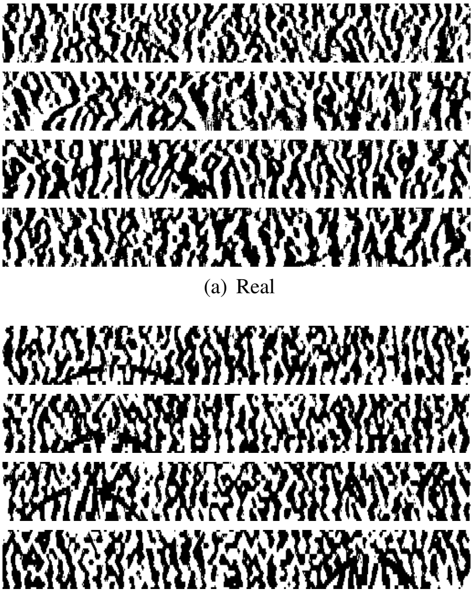
1. **Preparation**, during which a base Iris-Code matrix is created as follows:
   - The first row is created by generating alternating sequences of 0's and 1's with lengths drawn from a normal distribution. The distribution parameters can be estimated empirically, by measuring the sequence lengths in real Iris-Codes.
   - By duplicating that row, a simple bar-code pattern is generated.
2. **Parameter Estimation**, during which system configuration variables are calculated based on the user input.
   - A target Hamming distance ($HD_{target}$) between *IC1* and *IC2* is drawn from a random distribution.
   - $HD_{temp}$ and $HD_{intra}$ (see figure 1 and next step of the process description), are estimated based on $HD_{target}$. Following relations are satisfied: $HD_{temp} + HD_{intra} = C$ and $HD_{target} = 2HD_{intra} - O$, where $O$ is the expected overlap of bit mismatches introduced by the process described in the next step; $C$ remains constant for a batch of generated templates, and affects the effective number of independent bits (degrees of freedom) in the synthetic data.
3. **Iterative bit flipping**, during which a pair of mated Iris-Code templates is created from the base Iris-Code.
   - The bits at the edges of consecutive bit sequences (i.e. where sequences of 1's turn to 0's and vice versa) are randomly flipped. After $HD_{temp}$ from the original bar-code template is reached, the template is split into *IC1* and *IC2*. Subsequently, bit flipping occurs until $HD_{intra}$ between them is reached.
   - Additionally, majority voting and median filtering are applied to make the patterns visually smoother. Furthermore, the chances of bit flips are adjusted on per-row basis to simulate the collarette and furrow structures in real irides.
   - This step can be accelerated by applying an initial shifting pattern to the bar-code template produced in step 1.
4. **Post-processing**, during which additional noise factors are accounted for. Those include:
   - Adding the characteristic pattern resulting from an eyelid, as well as the noise beneath it.
   - Adding additional noise in the row near the pupil and simulating occlusions.
   - Storing the noise masks.
   - Applying circular shifts to the Iris-Code to simulate sample roll pose.

The process generates Iris-Codes of a default size; smaller sizes, if desired, are sampled from this size. The default dimension is motivated by the ISO/IEC international standard on Biometric sample quality [IS15]. There, the minimum iris radius is recommended to be at least 80 pixels (for the smallest reported human iris), which corresponds to a texture width of $80 * 2\pi \approx 502$ pixels when unrolled. The recommended optimal iris-pupil ratio is

0.2, which corresponds to a pupil of $80*0.2 = 16$ pixels, and thus an iris texture of 64 rows. Thus, the default size of the generated Iris-Codes is $512\times64$ bits. There are numerous adjustable parameters, which allow to mimic different properties of the Iris-Code (e.g. the correlations between rows and columns, noise). Notably, it is also possible to *guarantee* an arbitrary distribution of genuine scores and thereby simulate sample quality. For the data generated in this paper, the HDs are drawn from a Weibull distribution, due to its close resemblance to real data; another candidate could be the Gamma distribution. Yet another approach could be to empirically estimate a distribution from real data and use it instead.

# 3 Validation

In this section, the properties of the synthetically generated data are validated with respect to the requirements outlined in section 2. The visual comparison between real and synthetic Iris-Codes can be seen in figure 2. The real Iris-Codes were produced by using the OSIRIS toolkit [ODGS16] to process the near-infrared images from the iris subset of the BioSecure [Or10] database. The toolkit provides the commonly used 2D-Gabor feature extraction algorithm to produce the Iris-Codes. The synthetic Iris-Codes bear an excellent resemblance to the real ones.

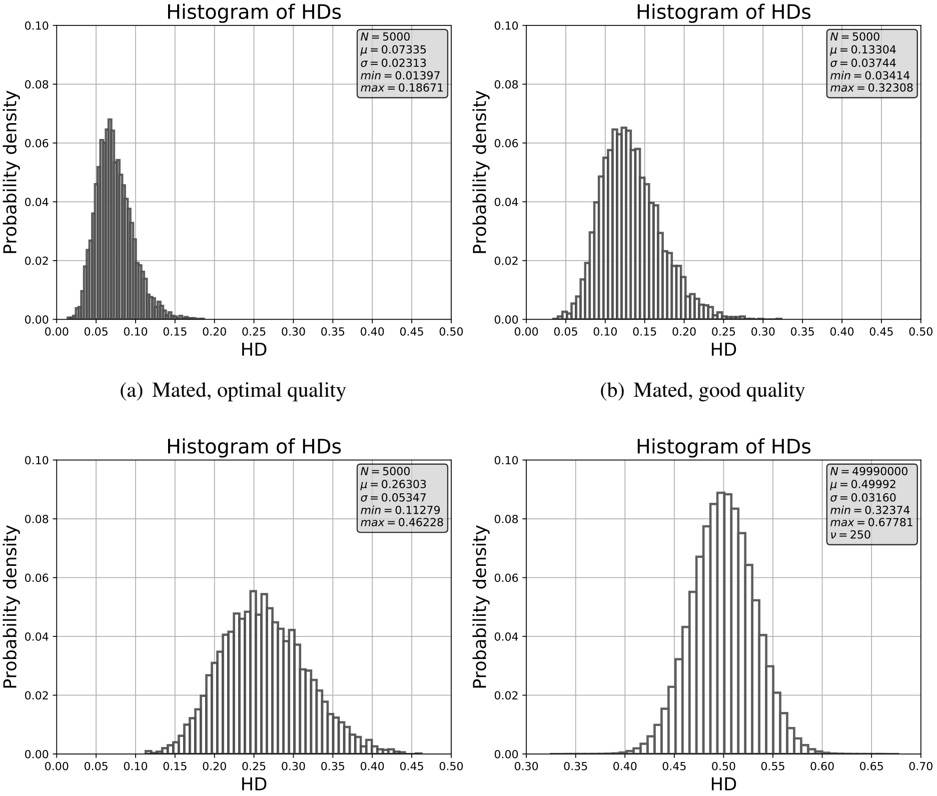


(a) Real

(b) Synthetic

Fig. 2: Example Iris-Codes produced from real eye images and generated by the proposed method

After confirming the visual appearance of the synthetic Iris-Codes to closely resemble that of the real data, their statistical properties are validated. Figure 3(d) shows the distribution of scores for non-mated templates for a large number of comparisons ($N$). The resulting distribution and its statistical properties (the yellow box in the image), including degrees of freedom ($\nu$), are identical to that exhibited by the real data, shown by Daugman in [Da04]. In figures 3(a), 3(b) and 3(c), example distributions of comparison scores for mated templates are shown, representing simulating of optimal, good and non-optimal quality data, respectively. As mentioned earlier, the mated distributions can be specified arbitrarily due to the nature of the template generation process (see section 2). The score distributions in figure 3 were produced using Iris-Codes of size 256×8 bits (same as used by Daugman in the paper cited above), sampled from the default size Iris-Codes generated by the process described in the previous section.



(a) Mated, optimal quality

(b) Mated, good quality

(c) Mated, non-optimal quality

(d) Non-mated

Fig. 3: Distributions of Hamming distances for a large number of comparisons between synthetic templates

Due to correlations between bits in an Iris-Code, its rows comprise of sequences of consecutive identical bits. It is of interest to verify, that the synthetic data follows that property. As real data reference, sequence lengths for all templates from the iris subset of the BioSecure database were computed. In figure 4, those distributions are shown, along with sequence lengths produced by Daugman's HMM from [Da16]. The distribution for the synthetic data generated by SIC-Gen closely follows the one exhibited by the real data.

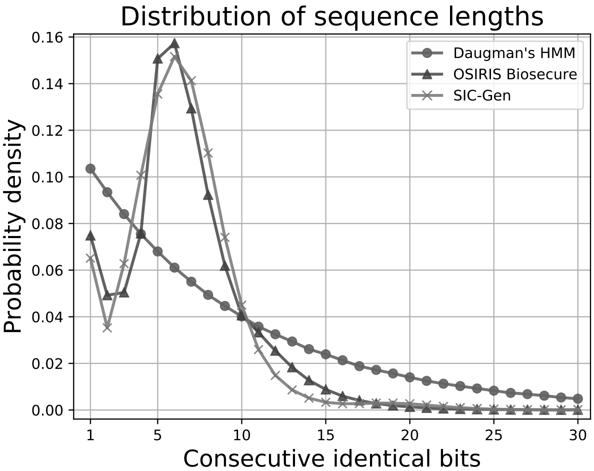
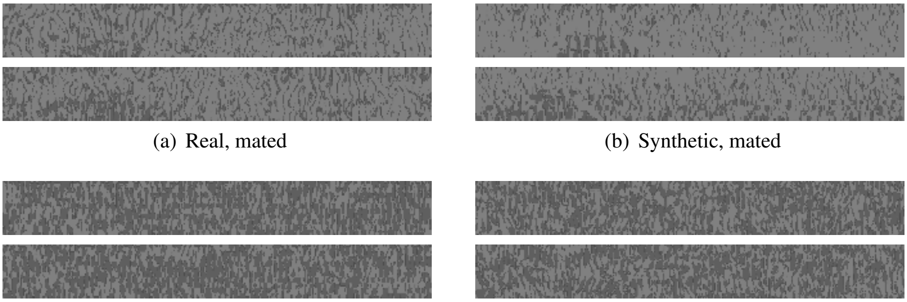


Fig. 4: Visualisation of lengths of sequences of consecutive bits in real data from BioSecure database, SIC-Gen synthetic templates and synthetic templates generated with Daugmann's HMM

Figure 5 shows example error patterns for comparisons between mated and non-mated templates. For the mated template pairs, the bit mismatches occur at the edges of sequences of consecutive identical bits, resulting in the pattern akin to that shown in real data by Hollingsworth *et al.* [HBF09].



(a) Real, mated

(b) Synthetic, mated

(c) Real, non-mated

(d) Synthetic, non-mated

Fig. 5: Example error patterns for comparisons between the real Iris-Codes from the BioSecure dataset and between the synthetic Iris-Codes

## 4 Conclusion and Future Work

In this paper, a method for generating synthetic Iris-Codes has been presented. The proposed method allows for a flexible specification of the score distribution between mated templates, to allow simulating different sample quality, acquisition environments etc.; the bit mismatches between two mated templates follow the so-called "fragile bits" patterns observed in real data. Simultaneously, the important statistical properties (e.g. degrees of freedom) of the distribution of non-mated comparison scores are maintained. Additionally, the synthetic Iris-Codes resemble the real ones visually. They reflect the correlations between Iris-Code bits resulting in long sequences of consecutive identical bits, as well as the typical noise sources, such as the eyelid pattern, circular shifts, wavelet noise and additional noise near the pupil. By accounting for all the aforementioned statistical and visual properties of real iris data, the proposed method represents a significant improvement over the current state-of-the-art and can be used in research cases where large iris datasets are needed, but unavailable. In future work, the authors intend to employ the synthetic Iris-Codes in large-scale testing of biometric indexing approaches, as well as to attempt to generate iris textures and/or images from the synthetic data using learning-based methods, e.g. Galbally *et al.* [Ga13].

## 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).

## References

[Cu04] Cui, J.; Wang, Y.; Huang, J.; Tan, T.; Sun, Z.: An iris image synthesis method based on PCA and super-resolution. In: 17th Intl. Conf. on Pattern Recognition. volume 4, pp. 471–474, August 2004.

[Da04] Daugman, J.: How iris recognition works. IEEE Trans. on Circuits and Systems for Video Technology, 14(1):21–30, January 2004.

[Da16] Daugman, J.: Information theory and the IrisCode. IEEE Trans. on Information Forensics and Security, 11(2):400–409, February 2016.

[Ga13] Galbally, J.; Ross, A.; Gomez-Barrero, M.; Fierrez, J.; Ortega-Garcia, J.: Iris Image Reconstruction from Binary Templates: An Efficient Probabilistic Approach Based on Genetic Algorithms. Computer Vision and Image Understanding, 117(10):1512–1525, October 2013.

[HBF09] Hollingsworth, K. P.; Bowyer, K. W.; Flynn, P. J.: The best bits in an Iris Code. IEEE Trans. on Pattern Analysis and Machine Intelligence, 31(6):964–973, June 2009.

[IS15] ISO/IEC JTC1 SC37 Biometrics: . ISO/IEC 29794-6:2015. Information technology – Biometric sample quality – Part 6: Iris image data. International Organization for Standardization and International Electrotechnical Committee, July 2015.

[Le03] Lefohn, A.; Budge, B.; Shirley, P.; Caruso, R.; Reinhard, E.: An ocularist's approach to human iris synthesis. IEEE Computer Graphics and Applications, 23(6):70–75, November 2003.

[MR05] Makthal, S.; Ross, A.: Synthesis of iris images using Markov random fields. In: 13th European Signal Processing Conf. pp. 1–4, September 2005.

[ODGS16] Othman, N.; Dorizzi, B.; Garcia-Salicetti, S.: OSIRIS: An open source iris recognition software. Pattern Recognition Letters, 82(2):124–131, September 2016.

[Or10] Ortega-Garcia, J. et al.: The Multiscenario Multienvironment BioSecure Multimodal Database (BMDB). IEEE Trans. on Pattern Analysis and Machine Intelligence, 32(6):1097–1111, June 2010.

[PN13] Proença, H.; Neves, J. C.: Creating synthetic IrisCodes to feed biometrics experiments. In: IEEE Workshop on Biometric Measurements and Systems for Security and Medical Applications. pp. 8–12, September 2013.

[PN17] Proença, H.; Neves, J. C.: Iris biometric indexing. In (Rathgeb, C.; Busch, C., eds): Iris and periocular biometric recognition, p. 25. IET, 2017.

[SR06] Shah, S.; Ross, A.: Generating synthetic irises by feature agglomeration. In: Intl. Conf. on Image Processing. pp. 317–320, October 2006.

[Un10] Unique Identification Authority of India (UIDAI): , Aadhaar issued summary. `https://portal.uidai.gov.in/uidwebportal/dashboard.do`, 2010. Last accessed: 2017-07-31.

[WSG05] Wecker, L.; Samavati, F.; Gavrilova, M.: Iris synthesis: a reverse subdivision application. In: 3rd Intl. Conf. on computer graphics and interactive techniques in Australasia and South East Asia. pp. 121–125, November 2005.

[WTS08] Wei, Z.; Tan, T.; Sun, Z.: Synthesis of large realistic iris databases using patch-based sampling. In: 19th Intl. Conf. on Pattern Recognition. pp. 1–4, December 2008.

[ZS05] Zuo, J.; Schmid, N. A.: A model based, anatomy based method for synthesizing iris images. In: Intl. Conf. on Biometrics. pp. 428–435, January 2005.

[ZSC07] Zuo, J.; Schmid, N. A.; Chen, X.: On generation and analysis of synthetic iris images. IEEE Trans. on Information Forensics and Security, 2(1):77–90, March 2007.