Iris Biometrics: A Method to Create Synthetic *IrisCodes*

Hugo Proença and João C. Neves
Department of Computer Science
IT - Instituto de Telecomunicações
University of Beira Interior, Portugal
{hugomcp,jcneves}@di.ubi.pt

Abstract

*The collection of iris data suitable to be used in experiments is difficult due to two factors: 1) the time spent by each volunteer in the acquisition process; and 2) security / privacy concerns of volunteers. Even though there are methods to synthesize images of artificial irises, there is no one exclusively focused in the synthesis of the iris biometric signatures (*IrisCodes*). This paper describes a stochastic method to synthesize* IrisCodes*, to feed experiments on iris matching, indexing and retrieval phases. We experimentally confirmed that both the genuine and impostor distributions obtained on the artificial data closely resemble values obtained in data sets of real irises. Also, the method is easily parameterized to mimic data of varying levels of quality.*

1. Introduction

Among multiple traits, the iris has made rapid strides in popularity due to the remarkable effectiveness of the deployed recognition systems [2] and to other interesting features: 1) its texture has a randotypic chaotic appearance possible to acquire contactless; 2) it has a simple shape, making easier its detection and segmentation; 3) it is roughly planar, enabling to compensate for deformations caused by camera-subject misalignments; and 4) most of its discriminating information lies in the lowest and middle-low frequency components of the signal, which are the most robust to noise. Accordingly, the nationwide deployment of iris recognition systems has already begun [3]. The Unique Identification Authority of India [16] is responsible for planning the largest-scale recognition system in the world (over 1 200 million persons) and the United Kingdom ID card initiative [7] intends to provide one biometric identity for each citizen.

To support research efforts, various iris image data sets are freely available (e.g., the CASIA [8], ICE [12], WVU [14], BATH [17], MMU [11], Olomuc [4] and UBIRIS [13]). However, up to the moment, these sets contain less than 10^4 identities, turning it hard to objectively assess the effectiveness of algorithms on large-scale scenarios. As a response, several attempts to create artificial iris images were done, which images acceptably resemble the appearance of real data.

In this paper we are particularly interested in providing data for the signatures matching and indexing / retrieval phases. We describe a stochastic method to obtain a large number of synthetic binary *IrisCodes*. The requirement of such type of method is evident, as generating a large number of artificial images is computationally expensive and unfeasible for practical scenarios. Also, the generation of binary signatures that closely resemble the extracted from real data is not straightforward, being important to account for the following factors:

- Impostors dissimilarity. The bit-by-bit comparison of signatures from different subjects should produce a *large* dissimilarity. The variability of this kind of scores should be relatively *small*.
- Genuine dissimilarity. The bit-by-bit comparison of signatures from the same subject should produce a *smaller* dissimilarity than for the impostors. Also, the variability of this kind of values should be significantly *higher* than in the case of impostors.

The remainder of this paper is organized as follows: Section 2 summarizes the most relevant published strategies to synthesize iris data. Section 3 provides a description of the proposed method. Section 4 presents and discusses the experiments. Finally, the conclusions are given in Section 5.

2. Related Work

As above stated, several methods were published to create artificial images of the iris that can be used for algorithm evaluation. However, the issue is their computational cost, turning hard to generate and transmit large data sets (e.g., for over 10^9 subjects). Even though, this section summarizes the most relevant methods published in this scope.

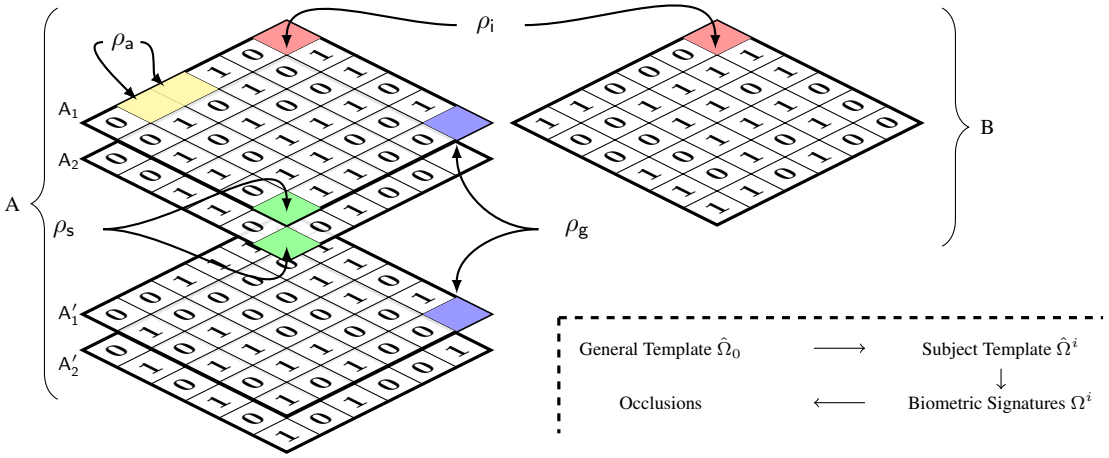


Figure 1. Cohesive overview of the parameters evolved in the synthesis of iris signatures. The different ρ values signal the correlation parameters . The left column represents two *IrisCodes* from subject A (each one with components extracted at two scales) and the right column illustrates an excerpt of an *irisCode* of subject B.

Lefohn *et* al. [9] proposed a method to create and render realistic looking irises by adding one layer at a time to the model and rendering an intermediate result, allowing incremental definition of the iris texture, using single layers taken from their standard library of textures. This method is useful in applications ranging from entertainment to ocular prosthetics. Cui *et* al. [1] proposed an iris synthesis method based on the analysis of principal components (PCA). They used an iris recognition algorithm based on PCA that operates on real images and allows to extract global feature vectors. These vectors were further used in image reconstruction. Iris samples that belong to the same class are constructed through letting the coefficients lie in the same sphere centered at a sample iris image in a high-dimensional space. To simulate different classes, they searched in a limited high-dimensional space. Also, authors concluded that super-resolution methods enhance the quality of the resulting images. Theoretical analysis and experimental results showed that the synthetic data mimics the traditional within-class and inter-class distances of real iris data. Shah *et* al. [15] proposed a technique to create digital versions of iris images used to evaluate the performance of iris recognition algorithms. Their scheme was divided into two phases: 1) at first, a Markov Random Field model generated a background texture that represents the global iris appearance; 2) next, a variety of iris features, radial and concentric furrows, collarette and crypts, were embedded in the texture field. Experiments with iris recognition algorithms validated the potential of this scheme. Zuo *et* al. [19] proposed a model and anatomy-based method for synthesizing iris images, having as purpose provide to the academia and industry a large data set to test iris recognition algorithms. This work also concerned about the bias that might be introduced by using synthetic data, having performed a comparison between the results observed for real and synthetic iris images. The comparison was quantified at three different levels: 1) global layout, 2) features of fine iris textures, and 3) recognition performance, including performance extrapolation capabilities. In most cases, the results confirm their expectation of a strong similarity between real and synthetic iris data generated using their model-based approach. Wei *et* al. [18] proposed an iris synthesis method and claimed to establish an effective paradigm to synthesize large iris databases, with the purpose to overcome the problems of data collection. Patch-based sampling was firstly employed to create prototypes, from where a number of intra-class samples were derived from each prototype. Experiments showed that the synthetic irises preserve the major properties of real ones and bear controllable statistics, turning them suitable for algorithm evaluation.

3. Proposed Method

Figure 1 illustrates the key parameters evolved in the synthesis of *IrisCodes*. The left column gives two *IrisCodes*, extracted at two scales of a given subject A. The right column gives an excerpt of the code of another subject (B). In real data and standard scenario, each code has $n = 2048$ bits extracted from the normalized images, with dimensions $r \times c$ at different scales s. Hence, four correlation parameters were used in the synthesis process: ρ_a (denoted by the yellow squares) dictates the strength of the linear correlation between spatially adjacent bits in the biometric signature. ρ_s (denoted by the green squares) corresponds to the strength of the linear correlation between

bits extracted from the same position of the iris at different scales. ρ_g (denoted by blue squares) is the strength of the linear correlation between corresponding bits of different signatures for each subject. Finally, ρ_i (represented by red squares) corresponds to the strength of the linear correlation between bits that correspond to the same region and scale of signatures extracted from different subjects.

The process is divided into three main phases: 1) at first, a generic template is created for the complete dataset, which will be used in the definition of the subjects templates. This template depends of the ρ_a parameter; 2) next, a template is created for each virtual subject. In this case, ρ_i dictates the dissimilarity between the templates generated for each subject; 3) then, a set of sample *IrisCodes* is created for each subject, considering the ρ_g parameter to control how much different will be these samples per subject; and 4) finally, occlusions in the irises are simulated, which correspond to regions of the *IrisCodes* where bits are purely random.

Formally, the process is based in the notion of linear correlation. Let u be a random value drawn from an uniform distribution $U \sim \mathbb{U}(0,1)$. u is quantized into binary values, and similar probabilities for each value are maintained:

$$u_q = \begin{cases} 1 & \text{, if } u \leq 0.5 \\ 0 & \text{, if } u > 0.5 \end{cases} \quad (1)$$

Let $\rho_.$ be a correlation value, (either ρ_a, ρ_i, ρ_g and ρ_s). Every bit of code c at position (x, y) is generated in top-left to bottom-right order in the following manner:

$$c(x,y) = 1 - \Big(H(t_0^r - \frac{r^2}{2}) \otimes H(\frac{(1 + erf(|t_0^r - 0.5|)\ \rho_.)}{2} - u_q)\Big) \quad (2)$$

being t_0^r is the total number of '0' bits in a neighborhood of radius r, *erf* is the sigmoid error function, $\otimes$ the exclusive OR logical operation and H the Heaviside function, defined as follows:

$$H(x) = \begin{cases} 0 & \text{, if } x \leq 0 \\ 1 & \text{, if } x > 0 \end{cases} \quad (3)$$

The top-left bit of the generic template of the data set is draw in a purely random way. Then, all the template is generated according to (2), using ρ_a as correlation parameter and $r = 1$. Next, the first scale of the templates for each subject is generated, using the ρ_i value and obtaining t_0^r from the generic template. For all subsequent scales, ρ_s controls the correlation and t_0^r is taken from the anterior scale. In a third step, the samples per subject are created, according to the ρ_g value and taking t_0^r from the subject template at the corresponding scale. In order to simulate different quality acquisition environments, a quality parameter *env* $\in [0, 1]$ weights the values of ρ_g, i.e., $\rho_{g'} = env \cdot \rho_g$. Table3 summarizes the parameters evolved in the above described synthesis process.

Parameter	Range	Description
ρ_s	[0,1]	Scale correlation. Controls the probability that bits extracted from the same positions of the iris at different scales have similar value.
ρ_a	[0,1]	Spatial correlation. Controls the probability that bits extracted from adjacent positions of the iris have similar values.
ρ_g	[0,1]	Genuine correlation. Controls the probability that bits extracted from images of a given subject have similar values.
ρ_i	[0,1]	Impostors correlation. Controls the probability that bits extracted from images of different subjects have similar values.
env	[0,1]	Corresponds directly to the *quality* data generated. "0" corresponds to data of poorest quality and "1" simulates signatures extracted from high quality data.

Table 1. Summary of the parameters evolved in the proposed method for the synthesis of *IrisCodes*.

Examples of the *IrisCodes* generated are shown in figure 2, illustrating the effect of the ρ_a parameter. Here, large values increase the correlation between adjacent bits (upper rows), whereas small values decrease this dependency and turn (for $\rho_a = 0$) the values of each bit independent of its neighborhood. The upper row of Figure 3 illustrates the ρ_s parameter. Here, two-scale signatures from subjects A and B are shown. The bottom row gives the effect of ρ_g by showing two additional samples B_1 and B_2 from subject B. The bottommost table gives the pair wise distances between *IrisCodes*, confirming that all requirements about codes dissimilarity were faithfully modeled.

4. Experiments

Figure 4 show histograms of the genuine and impostor comparisons obtained, regarding the env parameter that simulates the quality of the data from where the signatures would have been extracted. Previous studies shown that the conditions in the acquisition environment have a strong effect in the genuine comparisons, which was also confirmed in our observations. The top-left figure gives the distributions for an environment of relatively good quality (Env. A). Then, for remaining environments, quality decreases and, in the case of Env. D, there is a significant overlap between both distributions, as happens in real-world scenarios when the iris is not properly acquired.

Additionally, *IrisCodes* were validated in terms of the performance attained by three state-of-the-art indexing / retrieval strategies, when compared to the originally described by authors in their experiments on real iris data. The selected methods are due to Gadde *et* al. [5], which analyzed

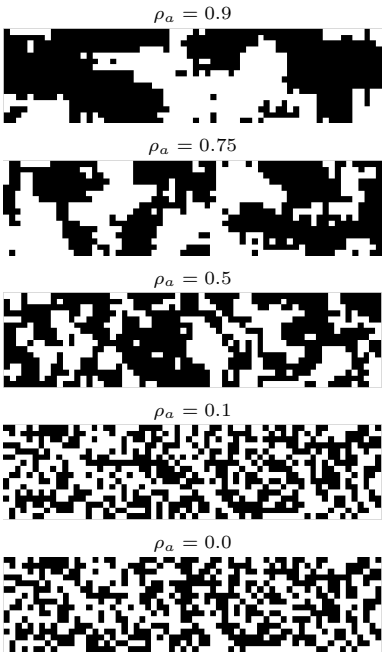
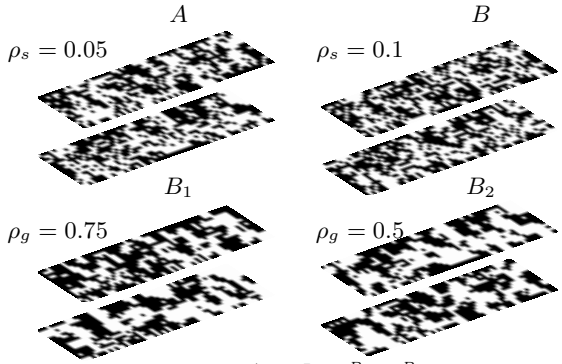


Figure 2. Effect of the spacial correlation parameter (ρ_a). Larger values augment the probability that neighbor codes have similar values, whereas the zero value turns the spatial location of each bit independent of its value.



	A	B	B_1	B_2
A	0	0.49	0.49	0.49
B	-	0	0.22	0.29
B_1	-	-	0	0.36
B_2	-	-	-	0

Figure 3. Images at the top row illustrate the effect of the ρ_s value. Images at the bottom illustrate the effect of ρ_g. A and B are signatures from different subjects. (B_1 and B_2 are samples from subject B). The bottommost table gives the pair wise Hamming distances between A and B's.

the distribution of intensities and selected patterns with low coefficients of variation (CVs) as indexing pivots. For each probe represented in the polar domain, a radial division of

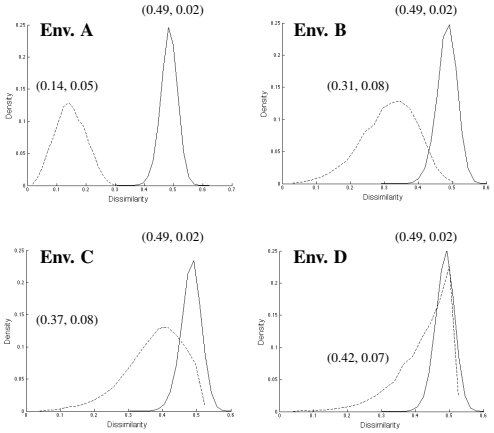


Figure 4. Illustration of the separation between genuine (dashed lines) and impostor (continuous lines) comparisons, for different levels of *quality*. At the top-left, histograms corresponding to data acquired in heavily controlled scenarios are shown (A). Data separability decreases in the bottom-right direction, and the poorest separable data at far right (D) is only suitable for soft biometric recognition purposes.

n-bands was performed and indexed using the radial band of the highest density of CV patterns. Also, Hao *et* al. [6] used the spatial spread of the most reliable bits, they propose an indexing technique based on the notion of multi-collisions. In the retrieval process, a minimum of k collisions between the probe and gallery samples is required to identify a potential match. Finally, Mukherjee and Ross [10] approached the problem from two different perspectives, by analyzing the iris texture and the *IrisCode*. The best results in the latter case were attained when each code was split into fixed-size blocks. First-order statistics for each block were used as the primary indexing value. A k-means strategy was used to divide the feature space into different classes. For comprehensibility, a single numeric score was used to assess levels of performance, in terms of the relation between hit and penetration rates, as suggested by Mukherjee and Ross [10] :

$$\tau = \sqrt{h(1-p)} \tag{4}$$

,

being h and p the hit and penetration rates. Table 4 compares the results announced by authors in their experiments with real irises data sets and the results obtained for synthetic *IrisCodes*, according to the method proposed in this paper. For contextualization, four different environments are shown (columns A to D), corresponding to the histograms of Figure 4. For both the methods of Gadde *et* al and Mukherjee and Ross, the results observed for synthetic data were poorer than those reported by authors, enabling

Method	Real	A	B	C	D
Gadde *et* al. [5]	0.909	0.650	0.637	0.588	0.583
Hao *et* al. [6]	0.997	**0.999**	0.981	0.761	0.740
Mukherjee and Ross [10]	0.858	0.675	0.651	0.593	0.568

Table 2. Comparison between the results (expressed in terms of (4) obtained by three state-of-the-art iris indexing / retrieval methods on signatures extracted from real irises (*Real Irises* column) and using synthetic *IrisCodes* generated by the proposed method.

to conclude about an extremely high quality level of the images used in their experiments. I the case of the method of Hao *et* al, results obtained in the synthetic *IrisCodes* were quite close to the reported by authors, specially in the case of environment A (highlighted in bold) which genuine / impostor distributions closely resemble the corresponding data given by authors in their paper. This fact was positively regarded as a strong indicator of the quality of the synthetic codes.

5. Conclusions

This paper described a stochastic method to generate synthetic *IrisCodes*. When performing an *all-agains-all* comparison between the generated codes, we confirmed that the resulting genuine and impostor comparisons faithfully resemble the corresponding distributions observed for real iris data. Also, an additional empirical validation was carried out by comparing the results obtained by three state-of-the-art indexing / retrieval techniques on real and artificial *IrisCodes*. It should be highlighted the easy parameterization of the proposed method, so to resemble the conditions in acquisition environments of varying quality. The proposed method is able to fed experiments on signature matching and indexing / retrieval phases, which is particularly important due to the eminent deployment of nationwide iris recognition systems.

References

[1] J. Cui, Y. Wang, J. Huang, T. Tan, and Z. Sun. An iris image synthesis method based on pca and super-resolution. *Proceedings of the IEEE*, 94(11):1927–1935, 2006.

[2] J. Daugman. Probing the uniqueness and randomness of iriscodes: Results from 200 billion iris pair comparisons. *Proceedings of the 17th International Conference on Pattern Recognition*, 4:471–474, 2004.

[3] J. Daugman and I. Mallas. Iris recognition border-crossing system in the UAE. *International Airport Review*, 8(2):49–53, 2004.

[4] M. Dobes and L. Machala. [online], `http://phoenix.inf.upol.cz/iris/`.

[5] R. Gadde, D. Adjeroh, and A. Ross. Indexing iris images using the burrows-wheeler transform. *Proceedings of the iEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6, 2010.

[6] F. Hao, J. Daugman, and P. Zielinski. A fast search algorithm for a large fuzzy database. *IEEE Transactions on Information Forensics and Security*, 3(2):203–211, 2008.

[7] Identity and Passport Service. [online], `http://www.direct.gov.uk/en/travelandtransport`, accessed on june, 2012.

[8] Institute of Automation, Chinese Academy of Sciences. [online] `http://www.sinobiometrics.com`.

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

[10] R. Mukherjee and A. Ross. Indexing iris images. *Proceedings of the 19th International Conference on Pattern Recognition*, pages 1–4, 2008.

[11] Multimedia University. [online] `http://pesona.mmu.edu.my/ccteo`.

[12] National Institute of Standards and Technology. [online] `http://iris.nist.gov/ICE/`.

[13] H. Proença, S. Filipe, R. Santos, J. Oliveira, and L. A. Alexandre. The ubiris.v2: A database of visible wavelength iris images captured on-the-move and at-a-distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(8):1502–1516, 2010.

[14] A. Ross, S. Crihalmeanu, L. Hornak, and S. Schuckers. A centralized web-enabled multimodal biometric database. *Proceedings of the 2004 Biometric Consortium Conference (BCC)*, 2004.

[15] S. Shah and A. Ross. Generating synthetic irises by feature agglomeration. *Proceedings of the 2006 IEEE International Conference on Image Processing*, pages 317–320, 2006.

[16] Unique Identification Authority of India. [online], `http://uidai.gov.in/about-uidai.html`, accessed on june, 2012.

[17] University of Bath. [online], `http://www.bath.ac.uk/elec-eng/pages/sipg/`.

[18] Z. Wei, T. Tan, and Z. Sun. Synthesis of large realistic iris databases using patch-based sampling. *Proceedings of the 19th International Conference on Pattern Recognition*, pages 1–4, 2008.

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