Creating Synthetic *IrisCodes* to Feed Biometrics Experiments

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

***Abstract*—The collection of iris data suitable to be used in experiments is difficult, mainly due to two factors: 1) the time spent by volunteers in the acquisition process; and 2) security / privacy concerns of volunteers. Even though there are methods to create images of artificial irises, there is no method exclusively focused in the synthesis of the iris biometric signatures (*IrisCodes*). In experiments related with some phases of the biometric recognition process (e.g., indexing / retrieval), a large number of signatures is required for proper evaluation, which, in case of real data, is extremely hard to obtain. Hence, this paper describes a stochastic method to synthesize *IrisCodes*, based on the notion of data correlation. These artificial signatures can be used to feed experiments on iris recognition, namely on the iris matching, indexing and retrieval phases. We experimentally confirmed that both the genuine and impostor distributions obtained on the artificial data closely resemble the values obtained in data sets of real irises. Finally, another interesting feature is that the method is easily parametrized to mimic *IrisCodes* extracted from data of varying levels of quality, i.e., ranging from data acquired in high controlled to unconstrained environments.**

I. 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 in a contactless way; 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.

The nationwide deployment of iris recognition systems is considered a success. In the last information update about the UAE system [5], over 2 million identities were included on its watch-list, and more than 350,000 deportees were prevented from entering the Emirates. The Unique Identification Authority of India [16] is deploying the system at the largest scale, with more than 300 million persons enrolled and adding about one million new identities per day, performing 6e^{14} daily cross-comparisons to search for duplicate identities.

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, at this time, these sets contain less than 10^4 identities, making 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 these 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 these values should be significantly *higher* than in the case of impostors.

The remainder of this paper is organized as follows: Section II summarizes the most relevant methods to synthesize iris data. Section III provides a description of the proposed method. Section IV presents and discusses the experiments. Finally, the conclusions are given in Section V.

II. 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, which is specially concerning in case that large data sets (e.g., for over 10^9 subjects) are required. This section summarizes the most relevant methods published in this scope.

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

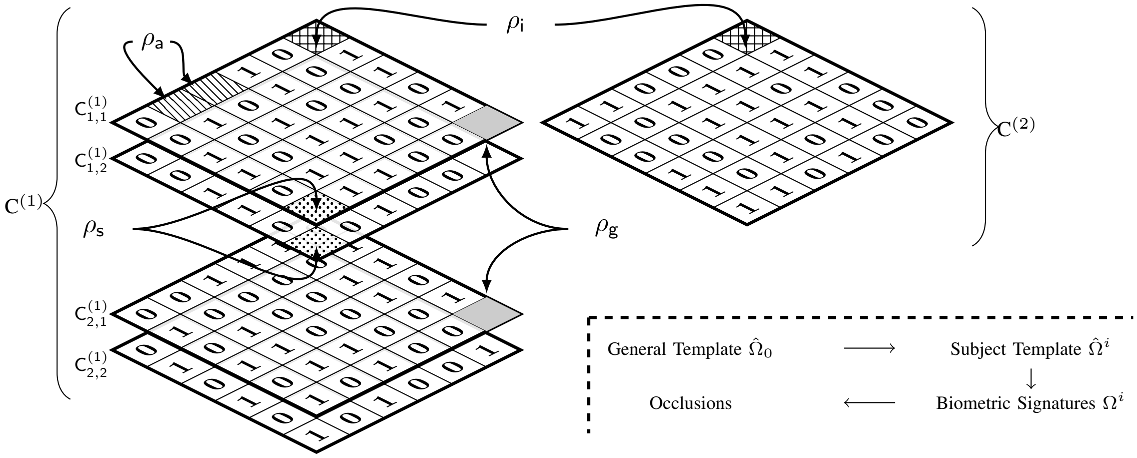

ρ_a
ρ_i
$C^{(1)}_{1,1}$
$C^{(1)}_{1,2}$
$C^{(1)}$
ρ_s
ρ_g
$C^{(1)}_{2,1}$
$C^{(1)}_{2,2}$
$C^{(2)}$
General Template $\hat{\Omega}_0$ ⟶ Subject Template $\hat{\Omega}^i$
↓
Occlusions ⟵ Biometric Signatures Ω^i

Fig. 1. Cohesive perspective 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 $C^{(1)}$ (each one with components extracted at two scales) and the right column illustrates an excerpt of an *irisCode* of subject $C^{(2)}$ ($C^{(i)}_{j,k}$ denotes a the jth code at the kth scale, from the ith person).

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, making them suitable for algorithm evaluation.

III. Proposed Method

According to the most acknowledged iris recognition algorithm [3], an *IrisCode* results from the convolution between the normalised iris image and a bank of Gabor filters at different scales. Then, the signal of the resulting complex coefficients determines the binary components of the signature. In agreement to that algorithm, the method proposed in this paper generates a set of binary values grouped in different scales, using the notion of *correlation*. For comprehensibility, let $C^{(i)}_{j,k}$ denote the jth code at the kth scale, from the ith person and $c(x, y, s)$ denote a bit of an *IrisCode* at position (x, y) and scale s.

Figure 1 gives an overall perspective of the parameters evolved in the synthesis of *IrisCodes*. The left column shows two *IrisCodes* of subject $C^{(1)}$, extracted at two scales. The right column gives an excerpt of the code of another subject ($C^{(2)}$). As in real signatures, each code has $n = 2048$ bits, with dimensions $n_r \times n_c$ at different scales n_s. Hence, four correlation parameters are used in the synthesis process: ρ_a dictates the strength of the linear correlation between bits that are spatially adjacent in the biometric signature. ρ_s corresponds to the strength of the linear correlation between bits extracted from the same position of the iris at different scales. ρ_g is the strength of the linear correlation between the corresponding bits of different signatures of subject $C^{(1)}$. Finally, ρ_i corresponds to the strength of the linear correlation between bits of the same position and scale of signatures extracted from different subjects.

The process is divided into three main phases: 1) a general template is created, which determines the subjects' templates. This general 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 of subjects; 3) 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) occlusions in the irises are simulated, which correspond to regions of the *IrisCodes* where bits are purely random.

Formally, let u be a random value drew from a uniform distribution $U \sim \mathbb{U}(0,1)$. u is quantized into binary value, maintaining similar probabilities for 0's and 1's:

$$u_q = \left\{ \begin{array}{ll} 1 & \text{, if } u \leq 0.5 \\ 0 & \text{, if } u > 0.5 \end{array} \right. \tag{1}$$

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

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

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

$$H(x) = \left\{ \begin{array}{ll} 0 & \text{, if } x \leq 0 \\ 1 & \text{, if } x > 0 \end{array} \right. \tag{3}$$

The top-left bit of the general template of the data set is purely random. Then, all the bits in the subjects' template are 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 $\xi \in [0,1]$ weights the values of ρ_g, i.e., $\rho_{g'} = \xi \rho_g$.

The final step simulates the regions of the iris that are occluded (e.g., due to eyelids or eyelashes). Two semi-circles of radii r_1 and r_2 are draw and placed in the bottom part of the *IrisCode*, which are known to be the regions that are most frequently occluded in the normalized images. Next, considering that bits extracted from eyelids or eyelashes do not possess any discriminating ability, bits inside these circles are generated in a purely random way (1), disregarding all the correlation ρ values. Figure 2 illustrates two occluded regions in *IrisCodes*, where the image at the top represents an heavily occluded image, in opposition to the bottom image that is almost noise-free.

Table III summarises the parameters evolved in the above described synthesis process, giving the range of values allowed for each one. Additionally, the bottom rows of that Table give examples of the parameters used to simulate environments of ideal conditions, and heavily unconstrained environments. These values were used to generate the data sets of Env. A and Env. D and are given for guidance of readers.

Examples of the *IrisCodes* generated are shown in Figure 3, 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 4 illustrates the ρ_s parameter. Here, two-scale signatures from subjects $C^{(1)}$ and $C^{(2)}$ are shown. The bottom row gives the effect of ρ_g by showing two additional signatures of subject $C^{(2)}$. The bottommost table gives the pairwise distances between *IrisCodes*, confirming that all requirements about codes dissimilarity were faithfully modeled.

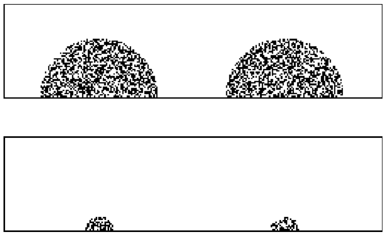


Fig. 2. Illustration of the bits in *IrisCodes* that are extracted from regions of the iris occluded by eyelids or eyelashes. These regions do not possess any discriminating ability between the genuine and impostors comparisons and - as such - the corresponding bits are draw in a purely random way (1).

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.
ξ	[0,1]	Corresponds directly to the *quality* of the data generated. "0" corresponds to data of poorest quality and "1" simulates signatures extracted from high quality data.
Optimal Environment		ρ_s=0.15, ρ_a=0.22, ρ_g=0.7, ρ_i=0.05, ξ=1
Unconst. Environment		ρ_s=0.15, ρ_a=0.22, ρ_g=0.1, ρ_i=0.07, ξ=0

TABLE I. SUMMARY OF THE PARAMETERS EVOLVED IN THE PROPOSED METHOD FOR THE SYNTHESIS OF *IrisCodes*.

IV. EXPERIMENTS

Figure 5 shows two histograms of the genuine (dashed lines) and impostor (continuous lines) matching scores, with respect to the ξ parameter. In all these plots, results regard 50,000 *IrisCodes* from 10,000 simulated different subjects. 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 topmost figure gives the distributions for an environment of relatively good quality (Env. A). Then, for the remaining environments, quality decreases and, in the case of Env. D, there is a significant overlap between both distributions, as it happens in uncontrolled scenarios.

Additionally, the synthetic *IrisCodes* were validated in terms of the performance attained by three state-of-the-art

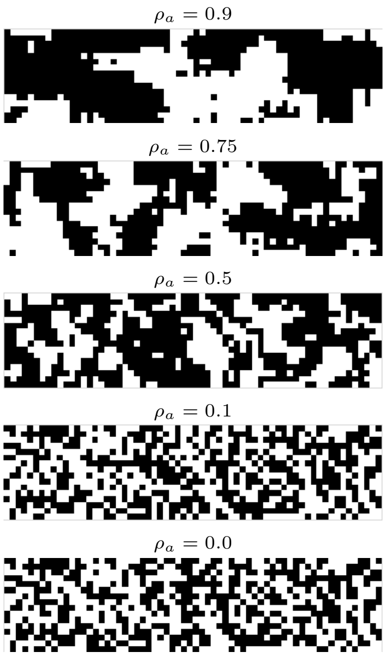
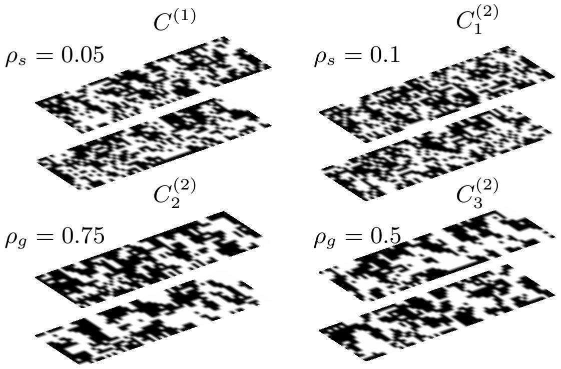


Fig. 3. Effect of the parameter ρ_a, that determines the spatial correlation between adjacent bits. Larger values augment the probability that neighbour codes have similar values, whereas the zero value turns the value of a bit independent of its spatial location.



	$C^{(1)}$	$C_1^{(2)}$	$C_2^{(2)}$	$C_3^{(2)}$
$C^{(1)}$	0	0.49	0.49	0.49
$C_1^{(2)}$	-	0	0.22	0.29
$C_2^{(2)}$	-	-	0	0.36
$C_3^{(2)}$	-	-	-	0

Fig. 4. Images at the top row illustrate the effect of the ρ_s value. Images at the bottom illustrate the effect of ρ_g. $C^{(1)}$ and $C^{(2)}$ are signatures from different subjects. The bottommost table gives the pairwise Hamming distances between $C^{(1)}$ and $C^{(2)}$.

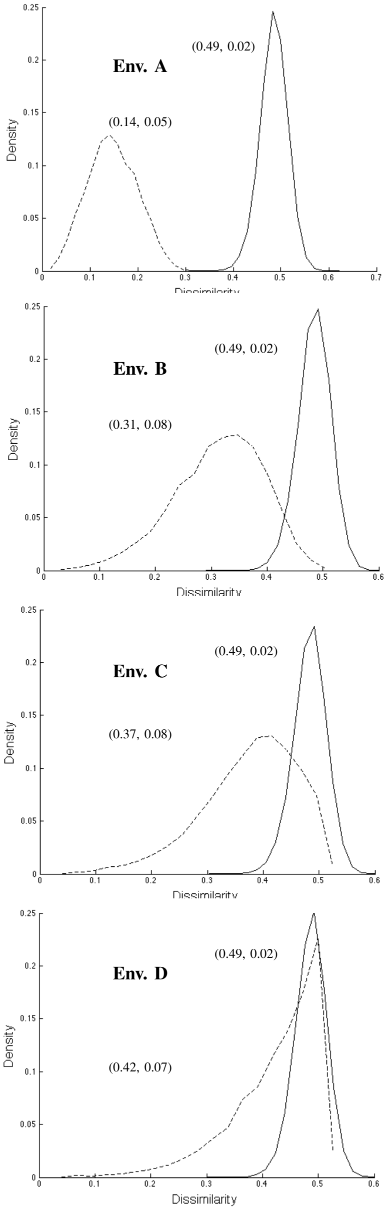


Fig. 5. Illustration of the separation between genuine (dashed lines) and impostor (continuous lines) comparisons, for different levels of *quality*.

indexing / retrieval strategies, comparing the results to the ones reported by authors in their experiments. The selected methods are due to: 1) Gadde *et* al. [6], which analyzed 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 n-bands was performed and indexed using the radial band of the highest density of CV patterns. 2) Hao *et* al. [7] 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, 3) 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

Method	**Real**	**Env. A**	**Env. B**	**Env. C**	**Env. D**
Gadde *et* al. [6]	0.909	0.650	0.637	0.588	0.583
Hao *et* al. [7]	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 II. RESULTS OBTAINED BY THREE STATE-OF-THE-ART IRIS INDEXING / RETRIEVAL METHODS ON SIGNATURES EXTRACTED FROM REAL IRISES (COLUMN *Real*) AND USING THE SYNTHETIC *IrisCodes* GENERATED BY THE PROPOSED METHOD (COLUMNS *Env. A-D*).

the 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 IV compares the results announced by authors in their experiments (Column *Real*) to the results obtained for synthetic *IrisCodes*, according to the method proposed in this paper. For contextualization, four different environments are shown (columns *Env. A* to *Env. D*), corresponding to the histograms of Figure 5. 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 to conclude about an extremely high quality level of the images used in their experiments. Also, we noted that both indexing methods are extremely sensitive to slight changes in the distributions of genuine / importer scores. Specifically, they significantly increase their effectiveness when the genuine distribution is positively skewed, which does not happens in the generated data sets. In the case of the method of Hao *et* al., results obtained in the synthetic *IrisCodes* were close to the reported by authors, specially in the case of environment A (highlighted in bold), in which the genuine / impostor distributions closely resemble the results given by authors. This fact was positively regarded as a strong indicator of the quality of the synthetic codes.

V. CONCLUSIONS

Aiming to resemble the *IrisCodes* that result from the most acknowledged iris recognition algorithm (Daugman's [3]), this paper described a stochastic method to generate synthetic *IrisCodes*, based on the notion of *linear correlation*. This method can be used to create an extremely large number of iris signatures, used to evaluate / validate different phases of the iris recognition process (e.g., iris matching, indexing / retrieval algorithms). When performing an *all-against-all* comparison between the generated codes, we confirmed that the resulting genuine and impostor matching scores 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*.The easy parameterization of the proposed method should be highlighted, to resemble the conditions in acquisition environments of varying quality.

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. How Iris Recognition Works. *IEEE Transactions on Circuits and Systems for Video Technology*, 14(1): 21–30, 2004.

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

[5] Emirates ID Head Office. Iris scan prevents entry of 350,000 deportees: Saif Bin Zayed. http://www.emiratesid.gov.ae/en/media-centre/news/, assessed on June 2013.

[6] 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.

[7] 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.

[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, 2013.

[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.