image will pass again to the CNN to achieve the factor 4 and so on.

In this work we take advantage of a common strategy used in image restoration, which is the extraction of patches and their representations as a series of pre-trained bases (such as PCA, DCT, Haar among other). Such filters are convolved with the image and in the case of this work will be optimized so that the mapping is the best possible. This can be done in one, two, or more layers and in the case of this work are followed by a reconstruction step which the predicted overlapping high-resolution patches are averaged to produce the final image. This strategy is used both in the SAEs and CNNs that will be explained in the next subsections.

A. Convolutional Neural Networks

Generally, the input of a CNN is a $(m \times m \times d)$ image where $(m \times m)$ is the dimension of the patch and d the number of channels (depth) of the image [19]. In this work, for the CNN training, patches are extracted from the HR images where $m = 33$ and $d = 1$, then the patches are downscaled (depending on the factor chosen for the method) and re-upscaled to the original size both using bicubic interpolation as it can be seen in the Figure 1.

In this work, the implemented CNN has three convolutional layers, where: the first layer consists of 64 filters of size 9x9x1 with stride 1 and padding 0, the second layer with 32 filters of size 1x1x64 with stride 1 and padding 0, and the last layer with 1 filter of size 5x5x32 with stride 1 and padding 0. With all paddings set to zero, the feature maps will decrease in size resulting in a patch of size 21x21. When the training is done, overlapping patches will be extracted from the LR images (upscaled using bicubic interpolation) with stride 1 and only the central pixel of the resulting feature map will be used which means that the smaller size of the resulted feature map will not influence the final image result.

After each convolutional layer a non-linearity (or activation) function is applied to the feature maps mainly to accelerate the convergence of the stochastic gradient algorithm called ReLU rectifier function: $f(x) = \max(0, x)$, where x is the neuron input.

For the training with the high-resolution patches with their correspondent low-resolution patches we use the Mean Squared Error (MSE) as the loss function trying to achieved the best PSNR as possible when the CNN is completely trained and the loss minimization is done using stochastic gradient descent with the standard backpropagation method.

In this work we tested three different approaches for the CNN training:

- From scratch (CNN FS): When the CNN weights are initialized randomly and trained according to the target image database (in the case of this work: the CASIA Interval V3 Iris Database) for the kernels domain adaptation, that is, to find the best way to map the data in order to perform the super-resolution.
- Transfer Learning (CNN TL): When an **off-the-shelf CNN** is chosen, which means that the CNN is pre-trained with a different database (in the case of this work: the ImageNet Database [20]) then it is used to perform the super-resolution in the target image database.
- Fine Tuning (CNN FT): The pre-trained network (off-the-shelf CNN) training is continued with new entries (with the target image database) for the weights to adjust properly to the new scenario reinforcing the more generic features with a lower probability of overfitting.

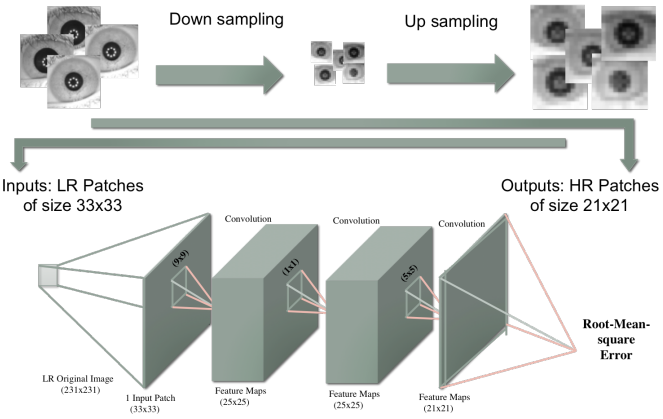

Fig. 1: An illustration of the Convolutional Neural Network architecture for Iris Super-Resolution.

B. Stacked Auto-Enconders

For the Layer-wise pre-training of Stacked Auto-Encoders we use the HR patches downscaled and upscaled again using bicubic interpolation in the same way as for the CNN. However, in this case, the matrix is turned to a vector in order to fit in the auto-encoder architecture. These vectors are used for the first auto-encoder as can be seen in Figure 2 that are trained until a threshold is reached. In the second auto-encoder, we use the vector that we got from the hidden layer of the previous trained auto-encoder as input, and proceed in the first auto-encoder. The same process is applied to the third layer and so on [21]. Then, we use the original images (HR patches) as the targets in the last layer of the output auto-encoder. These targets are used to update the parameter of the deep multi-layered neural network (Stacked Auto-Encoders) by means of a supervised error backpropagation algorithm. This process tries to reconstruct the image patch by generalizing the missing pixels with the auto-encoder weights learned from the all images of the training database.

When the training is completed, the auto-encoder is used to propagate all the LR patches upscaled using bicubic intepolation resulting in the reconstructed super-resolution patches in a magnification of 2 (when the training is done with this magnification). To achieve a magnification factor of 4, it is necessary to reinsert the reconstructed super-resolution images to the network in the same way as explained for the CNN approach.

For the experiments we trained four auto-encoders with the empirically chosen configuration: 1089-1000-1089 (where

1089 means the 33x33 input patches), 1000-2000-1000, 2000-2600-2000, 2600-2000-2600. Consequently, in the fine-tuning phase, the NN configuration for the Stacked Auto-Encoder experiment is: 1089-1000-2000-2600-2000-441. The size of the output (21x21 pixels) is because, in this case, a triangular architecture with more inputs than outputs can help the convergence in the fine-tuning phase.

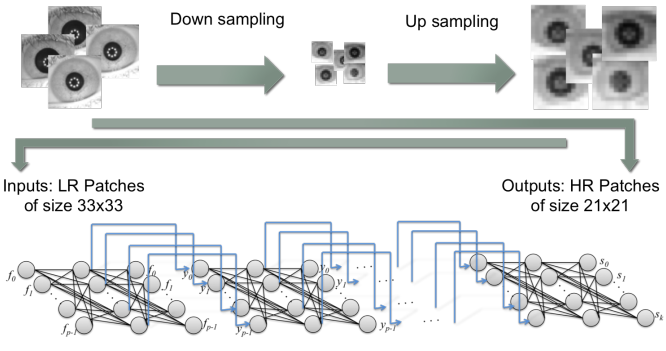

Fig. 2: An illustration of the Stacked Auto-Encoder architecture for Iris Super-Resolution.

III. Experimental Setup

For the experiments we use the CASIA Interval v3 iris database that contains a total of 2.655 NIR images of size 280x320 pixels, from 249 subjects captured with a self-developed close-up camera, resulting in 396 different eyes. Manual segmentation annotation of the database is available [3], which is used as input for our experiments. In the pre-processing step all images are resized via bicubic interpolation in order to have the same sclera radius and are aligned by extracting a square region of 231x231 around the pupil center. All images that do not fit in this requirement (for example when the eye is close to the image border) are discarded. After this, the 1.872 remaining images are used in the experiments. For the deep learning training and tests, the pre-processed dataset is divided into two separated sets: 925 images from the first 116 users for the training and 947 images from the remaining 133 users for the tests (we consider each eye as a different user). This set division by users is important to make sure that the same pattern (in the patches) will not be used both in training and testing steps.

To evaluate the performance of the methods by quality assessment algorithms we use the Peak Signal to Noise Ratio **(PSNR)**, that is the ratio between the peak signal and the power of corrupting noise that affects the fidelity of its representation, the Structural Similarity Index Measure **(SSIM)** that extracts three separate scores (visual influence, contrast and structural score) combining them to the final score, and the Visual Information Fidelity **(VIF)** that calculates the mutual information between input and the output of the HVS channel when no distortion is present and the mutual information between the input of the distortion channel and the output of the HVS channel for the test signal [22]. In these metrics, a high metric score reflects a high quality. For the quality tests, all images from the database were used in high resolution as reference images. We compare our method with bilinear and bicubic interpolation as well as to PCA hallucination of local patches used in [3].

We also conduct recognition experiments using reconstructed images to evaluate the iris recognition performance. In this procedure, first the iris is unwrapped to a normalized rectangle of 20x240 pixels using the Daugman's rubber sheet model [23], then a 1D Log-Gabor (LG) wavelet is applied with a phase binary quantization to 4 levels [24]. The comparison between the binary vectors is done by the normalized Hamming Distance [23] where the rotation is accounted for by shifting the grid of the query image in counter- and clock-wise directions, and selecting the lowest distance that corresponds to the best match. We also implemented a SIFT comparator in which SIFT feature points in scale space are extracted from the iris region (without unwrapping) and the comparison is performed based on the texture information around the feature points using the SIFT operator [25].

IV. Results

The results of the quality assessment for the test images and for the normalized iris region (20x240) are shown in Table I and Table II. It can be seen in Table I that the use of the Convolutional Neural Networks outperforms the traditional methods of interpolation (bicubic and bilinear) as well as the eigen-patch hallucination (PCA) method, mainly for small downscaling factors. It also can be noticed that the use of the Fine Tuning strategy improves the results by merging the use of natural and iris images during the CNN training. Also, when the CNN is trained with the same downscaling factor as the tests, the results are also becoming more resilient for lower resolutions. It can also be noticed that, for low resolutions, the quality assessment algorithms present different best results which can make the results interpretation difficult.

In iris recognition verification we consider two scenarios: 1) enrollment samples taken from original HR input images, and query samples taken from reconstructed super-resolution results (Table III) simulating a controlled enrollment scenario (for example, when the user is registered using a HR sensor and make use of the system using a cellphone camera with certain distance); and 2) both enrollment and query samples taken from the reconstructed super-resolution results (Table IV) simulating a totally uncontrolled scenario (for example, when the user is registered using a cellphone and make use of the system also using a cellphone camera with certain distance).

It can be observed that the performance of CNNs are the best for small downscaling factors in both scenarios in general, despite of the diversity of good results among the training approaches. Using the Log-gabor comparator the CNN using Fine Tuning and Transfer Learning approach beats the other methods except for the lowest resolution that PCA does best. For the SIFT comparator the CNNs are better. However, there is no particular winning training approach, in this case, using the downscaling factor of 2 the SAE method present

LR Size (scaling)		Bilinear	Bicubic	PCA	SAE	CNN FS Factor 2	CNN FS Factor 4	CNN TL Factor 2	CNN TL Factor 4	CNN FT Factor 2	CNN FT Factor 4	CNN FS Factor 8	CNN FT Factor 8	CNN FS Factor 16	CNN FT Factor 16
115x115 (1/2)	psnr	32.17	34.04	34.63	32.56	35.51	-	35.63	-	**35.93**	-	-	-	-	-
	ssim	0.892	0.926	0.934	0.897	0.945	-	0.946	-	**0.948**	-	-	-	-	-
	vif	0.813	0.819	0.771	0.724	0.821	-	0.823	-	**0.833**	-	-	-	-	-
57x57 (1/4)	psnr	27.64	29.17	29.89	28.06	30.43	30.81	30.65	30.44	30.69	**30.89**	-	-	-	-
	ssim	0.773	0.805	0.809	0.773	0.828	0.834	0.833	0.831	0.834	**0.837**	-	-	-	-
	vif	0.543	0.536	0.443	0.467	0.535	0.534	0.534	0.519	**0.546**	0.534	-	-	-	-
29x29 (1/8)	psnr	24.38	25.32	26.72	24.58	25.83	26.17	26.22	26.20	26.08	26.34	25.56	**28.31**	-	-
	ssim	0.682	0.700	0.709	0.680	0.710	0.720	0.723	0.721	0.718	0.727	0.707	**0.741**	-	-
	vif	**0.382**	0.376	0.254	0.333	0.340	0.330	0.327	0.322	0.340	0.320	0.299	0.326	-	-
15x15 (1/16)	psnr	21.94	22.85	**24.31**	22.07	23.26	20.98	23.63	23.66	23.36	20.98	-	-	22.01	23.16
	ssim	0.626	0.640	0.655	0.628	0.646	0.619	0.657	0.655	0.649	0.619	-	-	0.648	**0.670**
	vif	0.299	**0.304**	0.170	0.208	0.268	0.190	0.251	0.231	0.259	0.180	-	-	0.218	0.260

TABLE I: Results with different downscaling factors and two different factors (average values on the test dataset).

LR Size (scaling)		Bilinear	Bicubic	PCA	SAE	CNN FS Factor 2	CNN FS Factor 4	CNN TL Factor 2	CNN TL Factor 4	CNN FT Factor 2	CNN FT Factor 4	CNN FS Factor 8	CNN FT Factor 8	CNN FS Factor 16	CNN FT Factor 16
115x115 (1/2)	psnr	34.27	36.22	36.83	34.69	37.69	-	37.80	-	**38.08**	-	-	-	-	-
	ssim	0.930	0.951	0.955	0.923	0.963	-	0.963	-	**0.964**	-	-	-	-	-
	vif	0.812	0.848	0.824	0.766	0.859	-	0.864	-	**0.872**	-	-	-	-	-
57x57 (1/4)	psnr	29.27	31.14	32.13	29.94	32.76	33.34	33.02	32.73	33.03	**33.40**	-	-	-	-
	ssim	0.853	0.873	0.874	0.852	0.887	0.891	0.890	0.889	0.891	**0.893**	-	-	-	-
	vif	0.583	0.601	0.550	0.540	0.614	0.626	0.625	0.617	0.630	**0.632**	-	-	-	-
29x29 (1/8)	psnr	25.59	26.67	**28.61**	25.86	27.25	27.56	27.74	27.73	27.56	27.91	25.56	28.31	-	-
	ssim	0.791	0.803	0.811	0.788	0.810	0.818	0.820	0.819	0.816	0.823	0.806	**0.837**	-	-
	vif	0.456	**0.459**	0.399	0.429	0.443	0.449	0.451	0.444	0.449	0.450	0.425	0.440	-	-
15x15 (1/16)	psnr	22.96	23.97	**25.82**	23.08	24.42	24.55	24.94	24.96	24.55	22.15	-	-	23.31	24.67
	ssim	0.748	0.760	0.774	0.749	0.763	0.743	**0.774**	0.772	0.766	0.743	-	-	0.761	**0.785**
	vif	0.417	0.414	0.335	0.417	0.393	0.350	0.386	0.374	0.386	0.342	-	-	0.407	**0.419**

TABLE II: Results with different downscaling factors and two different factors for the unwrapped iris region (average values on the test dataset).

the best result for the scenario 2. It also can be seen that for the SIFT comparator the performances of the Bicubic and Bilinear methods degrade rapidly when the resolution decreases, whereas the CNN methods show high resiliency.

It is interesting to notice in scenario 1 (Table III) that CNN methods, Bicubic and Bilinear interpolations perform better in factor 2 and 4 than using the original images without downscaling which means that it, in terms of recognition, it is better to downscale the original image (i.e. apply a blur filter) and apply the deep-learning methods from the sensor before comparison. This can be explained by the fact that the image downscaling and subsequent upscaling performs a form of denoising process that can help the recognition system.

In the recognition experiments we also perform a significance test to calculate a boundary on the significance between the best results presented and the results from the original database using the Chi-squared distribution according to [26]. With $\tilde{\chi}^2 = 15.977$ the values that are significantly better than the original results are underlined in Table III and IV.

V. Conclusion

In this work we investigated deep learning single-image super-resolution methods using Stacked Auto-Encoders and Convolutional Neural Networks to increase the resolution of iris images. To address the problem we tested if the end-to-end mapping between low and high resolution images can be successful applied using different strategies as transfer learning and fine-tuning to improve the results.

Evaluation performed on a database of near-infrared iris images with different upscaling factors both in the training process and in the tests show the superiority of the tested methods over the compared methods in terms of quality assessment, with the CNN using Fine Tuning approach presenting the best results on average. When we evaluate the recognition rate by iris comparison experiments, the CNNs in general presented better results, but there was no particular CNN approach being the best in all scenarios. We also showed that an uncontrolled scenario (scenario 2 in the EER verification results) is feasible since the deep learning approach in scenario 2 presented better accuracy results than the scenario 1. With this, it can be concluded that in practical tests, when the verification images are in low-resolution and the enrollment images are in high-resolution it is better to downscale the enrollment images and perform the super-resolution in both databases to achieve better recognition results.

Also, it is important to notice that recognition performed is not considerably degraded until image is downscaled by 1/8 or higher factors, allowing to use both query and test images of reduced size which can be an advantage for systems under low storage or data transmission capabilities.

In future work we intend to focus on the Convolutional Neural Network approach testing new methods as the use of recursive layers and investigate the use of other loss functions as perceptual loss functions as well as explore other datasets with different semantic knowledge to perform the fine tuning approach.

Acknowledgment

This research was partially supported by CNPq-Brazil for Eduardo Ribeiro under grant No. 00736/2014-0.

LR Size (scaling)		Bilinear	Bicubic	PCA	SAE	CNN FS Factor 2	CNN FS Factor 4	CNN TL Factor 2	CNN TL Factor 4	CNN FT Factor 2	CNN FT Factor 4	CNN FS Factor 8	CNN FT Factor 8	CNN FS Factor 16	CNN FT Factor 16
115x115	LG	**0.69**	**0.69**	0.73	3.00	0.72	-	0.76	-	**0.69**	-	-	-	-	-
(1/2)	SIFT	4.05	**3.51**	3.81	4.21	4.01	-	4.21	-	4.01	-	-	-	-	-
57x57	LG	0.69	0.68	0.73	1.34	0.69	0.68	0.72	0.72	0.68	**0.67**	-	-	-	-
(1/4)	SIFT	10.42	7.41	5.20	10.13	4.95	**4.34**	4.47	4.67	4.41	**4.34**	-	-	-	-
29x29	LG	1.61	1.42	1.11	2.33	1.18	1.18	1.07	1.10	1.09	**1.02**	1.53	1.37	-	-
(1/8)	SIFT	28.23	24.99	15.86	35.31	17.50	**14.26**	16.31	17.34	17.96	15.87	20.65	17.65	-	-
15x15	LG	10.39	9.59	**7.29**	14.29	9.07	18.72	8.96	9.67	9.43	17.84	-	-	19.53	15.74
(1/16)	SIFT	50.52	47.33	36.51	48.02	41.76	42.06	38.23	**36.36**	40.99	39.08	-	-	42.60	45.35

TABLE III: Verification results (EER) of the scenario 1 (original vs. downscaled) considered for different downscaling factors. The results for the original database with no scaling for the LG and SIFT are respectively 0.76 and 4.19.

LR Size (scaling)		Bilinear	Bicubic	PCA	SAE	CNN FS Factor 2	CNN FS Factor 4	CNN TL Factor 2	CNN TL Factor 4	CNN FT Factor 2	CNN FT Factor 4	CNN FS Factor 8	CNN FT Factor 8	CNN FS Factor 16	CNN FT Factor 16
115x115	LG	**0.61**	0.73	0.72	0.66	0.72	-	0.72	-	0.72	-	-	-	-	-
(1/2)	SIFT	3.01	3.13	3.71	**2.54**	3.82	-	3.80	-	3.82	-	-	-	-	-
57x57	LG	0.76	0.65	0.68	0.72	0.68	0.65	**0.60**	0.66	0.62	0.68	-	-	-	-
(1/4)	SIFT	4.26	3.08	3.37	3.45	2.09	2.23	2.41	2.29	**1.94**	2.50	-	-	-	-
29x29	LG	2.38	1.88	1.18	2.14	1.30	1.95	**0.98**	1.24	1.14	1.26	1.71	1.41	-	-
(1/8)	SIFT	14.82	11.6	7.54	15.82	6.50	**6.26**	7.33	8.14	7.30	7.26	8.65	7.45	-	-
15x15	LG	11.03	11.25	**4.79**	8.58	9.10	14.31	6.26	8.18	7.88	11.64	-	-	12.43	11.46
(1/16)	SIFT	41.66	36.37	19.50	36.35	22.64	20.12	22.28	**17.26**	22.78	19.08	-	-	19.59	26.40

TABLE IV: Verification results (EER) of the scenario 2 (downscaled vs. downscaled) considered for different downscaling factors. The results for the original database with no scaling for the LG and SIFT are respectively 0.76 and 4.19.

REFERENCES

[1] K. Nguyen, C. Fookes, S. Sridharan, and S.n Denman, "Feature-domain super-resolution for iris recognition," *Computer Vision and Image Understanding*, vol. 117, no. 10, 2013.

[2] K. Bowyer, K. Hollingsworth, and P. Flynn, "Image understanding for iris biometrics: A survey," *Computer Vision and Image Understanding*, vol. 110, no. 2, 2008.

[3] F. Alonso-Fernandez, R. A. Farrugia, and J. Bigun, "Eigen-patch iris super-resolution for iris recognition improvement," in *2015 23rd European Signal Processing Conference (EUSIPCO)*, Aug 2015.

[4] J. B. Huang, A. Singh, and N. Ahuja, "Single image super-resolution from transformed self-exemplars," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 5197–5206.

[5] J. Li, Y. Qu, C. Li, Y. Xie, Y. Wu, and J. Fan, "Learning local gaussian process regression for image super-resolution," *Neurocomputing*, vol. 154, 2015.

[6] R. Timofte, V. DeSmet, and L. VanGool, "A+: Adjusted anchored neighborhood regression for fast super-resolution," in *12th Asian Conference on Computer Vision*, Cham, 2015, Springer International Publishing.

[7] J. Yang, Z. Wang, Z. Lin, S. Cohen, and T. Huang, "Coupled dictionary training for image super-resolution," *IEEE Transactions on Image Processing*, vol. 21, no. 8, Aug 2012.

[8] N. Wang, D. Tao, X. Gao, X. Li, and J. Li, "A comprehensive survey to face hallucination," *International Journal of Computer Vision*, vol. 106, no. 1, 2014.

[9] K. Nguyen, S. Sridharan, S. Denman, and C. Fookes, "Feature-domain super-resolution framework for gabor-based face and iris recognition," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, June 2012.

[10] H.C. Burger, C.J. Schuler, and S. Harmeling, "Image denoising: Can plain neural networks compete with bm3d?," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, June 2012.

[11] V. Jain and S. Seung, "Natural image denoising with convolutional networks," in *Advances in Neural Information Processing Systems 21*. Curran Associates, Inc., 2009.

[12] Zhen Cui, Hong Chang, Shiguang Shan, Bineng Zhong, and Xilin Chen, "Deep network cascade for image super-resolution," in *Computer Vision ECCV 2014*, vol. 8693 of *Lecture Notes in Computer Science*. Springer International Publishing, 2014.

[13] J. Kim, J. K. Lee, and K. M. Lee, "Deeply-recursive convolutional network for image super-resolution," *CoRR*, vol. abs/1511.04491, 2015.

[14] J. Johnson, A. Alahi, and Fei-Fei Li, "Perceptual losses for real-time style transfer and super-resolution," *CoRR*, vol. abs/1603.08155, 2016.

[15] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. P. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," *CoRR*, vol. abs/1609.04802, 2016.

[16] W. Shi, J.Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," *CoRR*, vol. abs/1609.05158, 2016.

[17] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, Feb 2016.

[18] CASIA Iris Image Database, "http://biometrics.idealtest.org/," .

[19] E. Ribeiro, A. Uhl, G. Wimmer, and M. Häfner, "Exploring deep learning and transfer learning for colonic polyp classification," *Computational and Mathematical Methods in Medicine*, pp. 1–16, 2016.

[20] Alex K., I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*. Curran Associates, Inc., 2012.

[21] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, Aug 2013.

[22] H. Hofbauer and A. Uhl, "Identifying deficits of visual security metrics for images," *Signal Processing: Image Communication*, vol. 46, 2016.

[23] J. Daugman, "How iris recognition works," *IEEE Trans. Cir. and Sys. for Video Technol.*, vol. 14, no. 1, Jan. 2004.

[24] L. Masek, "Recognition of human iris patterns for biometric identification," Tech. Rep., The University of Western Australia, 2003.

[25] F. Alonso-Fernandez, P. Tome-Gonzalez, V. Ruiz-Albacete, and J. Ortega-Garcia, "Iris recognition based on sift features," in *2009 First IEEE International Conference on Biometrics, Identity and Security (BIdS)*, Sept 2009.

[26] H. Hofbauer and A. Uhl, "Calculating a boundary for the significance from the equal-error rate," in *2016 International Conference on Biometrics (ICB)*, June 2016, pp. 1–4.