2016 23rd International Conference on Pattern Recognition (ICPR)
Cancún Center, Cancún, México, December 4-8, 2016

A Preliminary Study of CNNs for Iris and Periocular Verification in the Visible Spectrum

Karan Ahuja, Rahul Islam, Ferdous A. Barbhuiya
Indian Institute of Information Technology Guwahati
Assam, India, 781001
Email: {karan.ahuja, rahul.islam, ferdous}@iiitg.ac.in

Kuntal Dey
IBM Research India,
New Delhi, India
Email: kuntadey@in.ibm.com

Abstract—**Ocular biometrics in the visible spectrum has emerged as an area of significant research activity. In this paper, we propose two convolution-based models for verifying a pair of periocular images containing the iris, and compare the two approaches amongst each other as well as with a baseline model. In the first approach, we perform deep learning in an unsupervised manner using a stacked convolutional architecture, using external models learned *a-priori* on external facial and periocular data, on top of the baseline model applied on the provided data, and apply different score fusion models. In the second approach, we again use a stacked convolution architecture; but here, we learn the feature vector in a supervised manner. We obtain an AUROC of 0.946 and 0.981, and EER of 0.092 and 0.066, for the two models respectively. We further combine the two models, and observe the combined model to deliver the best performance in case the both the images arise from the same device type, but not necessarily so otherwise, obtaining a AUROC of 0.985 and EER of 0.057. Given the significant performance our methodology yields, our system can be used in real-life applications with minimal error.**

I. Introduction

Identifying individuals as genuine versus impostors, using facial features such as matching of the iris, periocular region and face, have emerged as areas of research interest. In this paper, following the definition given by [1], the term periocular means the area surrounding the eye as well as the eye - i.e. containing the iris and sclera also. The qualitative problem at hand is of verifying whether a pair of periocular images taken in the visible spectrum belong to the same person or not.

As the computer vision and image processing techniques have matured with time, several novel approaches towards this problem (and closely related problems) have been recently introduced by different schools of research. Early works by Park *et al.*, such as [2] and [1], have established the feasibility of using periocular images for identification. Prior works, such as [3] evaluated the utility of the periocular region appearance cues for biometric identification. Recently [4] reviewed the research progress in the area and discussed existing algorithms and the limitations of each of the biometric traits and information fusion approaches. In terms of iris segmentation, [5] presents an unsupervised iris defects detection method based on the underlying multispectral spatial probabilistic iris textural model. For iris recognition in the visible spectrum, [6] describes an integrated scheme for noisy iris recognition in adverse conditions. For mobile devices in particular, [7] proposes a system that combines the recognition of user's iris and user's devices for authentication of users and [8] proposes a fusion of face and iris features for recognition. In the recent ICIP periocular identification challenge [9], a database with the title VISOB was provided, and a number of approaches of identifying individuals were presented, where the images were collected under different lighting conditions, namely daylight, dim light and office lighting, and different devices, namely Oppo mobile phones, Samsung mobile phones and Apple iPhones. Deeply coupled auto-encoders [10] and deep sparse filter [11] based approaches were observed to have outperformed the remaining approaches, such as the 2-phase approach by Ahuja *et al.* [12] that uses a multinomial Bayesian Learning followed by Dense SIFT.

In this paper, we propose two models to solve the problem at hand. Both the models are based on deep convolutional neural networks (CNN), and both the models use a stacked layer architecture. Further, we treat the well-known Root SIFT method [13] as a baseline approach to solve the problem. Root sift calculates the image descriptor of the iris images given as part of the MICHE dataset, and subsequently matches an image with the other using a k-nearest neighbor approach.

In the first approach, we combine the baseline Root SIFT method with two external (pragmatic) knowledge sources. The first one is a 128-sized feature vector obtained from OpenFace [14], which is a general-purpose library for face recognition. The other pragmatic knowledge source is the VISOB dataset [9], which provides images of the periocular region. We use a 1024-sized feature vector of the periocular region, obtained by training on this dataset. We combine these three, namely the Root SIFT, feature vector of VISOBNet and that of the OpenFace, and combine the scores assigned by each of these three subsystems, to calculate a dissimilarity score, using simple averaging as well as linear regression based techniques. Thus, the first model is an unsupervised one, with respect to the provided MICHE-II database.

In the second approach, we avoid using external *a-priori* knowledge, and solely rely upon the provided MICHE-II dataset to perform CNN-based deep neural network learning, using a supervised approach. We pass each training image through a 4-convolution CNN network, and subsequently create a 512-sized feature vector for each training image. For each test image, we construct its 512-feature vector, and compare this vector with each of the training vectors using cosine similarity, to find the best-match image. Note that, in both the models, we generate data using known augmentation techniques, to further improve the performance of our system. On the provided MICHE-II test dataset, we obtain an AUROC

978-1-5090-4847-2/16/$31.00 ©2016 IEEE

[7] of 0.946 and 0.981, and EER [9] of 0.092 and 0.066, for the two models respectively, when testing under the same device constraint. We also combine the two models, and observe that, the combined hybrid model outperforms all the remaining models, to deliver the most optimal performance under the same device constraint, achieving an EER of 0.057 and AUROC of 0.985. The high performances that our models yield, pose these models as reasonable candidates for deploying in real-life applications.

Thus, the contributions of our work are as follows.

- We propose two novel convolution-based stacked deep neural network models in order to compare a given periocular image containing the iris, with a set of periocular images.
- We create a first CNN-based unsupervised model, fusing the scores of two external feature vectors and a baseline Root SIFT model.
- We create a second CNN-based supervised model, that uses only the images from the MICHE-II dataset, and uses a cosine similarity metric on the derived feature vector for measuring similarity between image pairs.
- We provide an empirical comparison of the two approaches amongst each other as well as with a baseline Root SIFT model. We observe the second model to outperform the first one, the combined model to outperform the two independent models where the training and test images stemmed from the same device type, and further observe the baseline to be significantly outperformed by both the models we propose.

The rest of the paper is as follows. The details of our methodology, including the design principles and the models, are presented in Section II. Section III explores the outcome of applying our methodology on the given dataset. Finally, we provide a brief discussion in Section IV and conclude in Section V.

II. Methodology

We propose a baseline Root SIFT method and two further models to achieve our objective of verifying individuals. In the first model, we aim at learning an unsupervised metric [15], that generalizes well across several data-sets. In this setting, no training whatsoever is performed on the MICHE-II database [16]. In the second model, we employ a supervised learning paradigm that learns feature representation for comparison and verification on the MICHE-II data-set.

A. Baseline Model

Inspired by SIFT based models for ocular biometrics in the visible spectrum such as [9], [12] and [17], we make use of Dense SIFT keypoints for matching irises. First, the iris is extracted out of the image using the segmentation algorithm described in [5]. The algorithm provides us with the segmented and normalized iris image along with a defects mask. We first overlay the segmented iris image with the binary mask to get the iris image rid of any occlusions. We then compute Dense color Root SIFT [13] descriptors which gives us keypoints with identical size and orientation. The advantage of Root SIFT over traditional SIFT [18] is that it employs a Hellinger kernel instead of the standard Euclidean distance to measure the similarity between SIFT descriptors. Matching between descriptors is performed by comparing each local extrema using a nearest neighbor matcher [19]. The dissimilarity score d is given by:

$$d = \left(1 - \frac{|Matches|}{min(|KeyPts\_img1|, |KeyPts\_img2|)}\right) \quad (1)$$

B. Model 1

Figure 1 illustrates our proposed framework for Model 1. The model consists of three integral parts for the purpose of verification, namely OpenFace, Visobnet and RootSIFT.

1) OpenFace: OpenFace [14] is a general purpose face recognition library. Given a facial image, it outputs a 128-dimensional feature vector of that image. Although, it is crafted for face verification, we find it to perform well for the task of verification from partial face images as well. Therefore, we input the whole MICHE-II test image to the deep neural network without any preprocessing. OpenFace subsequently outputs the predicted similarity score of two images by computing the squared L2 distance between their representations. Since the representations are on the unit hypersphere, the scores range from 0 (the same picture) to 4.0. We then convert the score to a range from 0 to 1 to get the dissimilarity score. The use of OpenFace also helps us to compare as to what extent can existing state-of-the art methods for face verification be employed for the task of Ocular Bio-metrics.

2) VisobNet: Deep learning systems have achieved state of the art accuracies in face recognition tasks [20]. However, they require large a large training database to learn their models. Alternatively, the use of transfer learning [21] is often used to solve this problem. Here the feature representation is learned on an external dataset. Motivated by the success of such approaches, we employ a similar approach in which we train our model on the VISOB Database [9]. The model automatically learns appearance-based features by using a deep convolutional Neural network. We train our CNN on a multi-class face recognition task, namely to classify the identity of the periocular image.

The overall architecture is depicted in Table I. First, the periocular region is extracted from the given image by creating a rough bounding box around the eye, the dimension of which are given as a function of the iris center and radius returned by Haindl *et al.* [5]. This RGB (3 channel) periocular image is re-sized to 32 pixels $\times$ 48 pixels and given as input to the convolution layer 1. We use a convolution kernel of size 3×3 in all the convolution layers and the activation function is ReLU [22] except in the last layer where we use a softmax classifier [23]. The dense layers of the network represent the fully connected layers. We train the CNN using Stochastic Gradient Descent (SGD) [24] with standard back-propagation and Momentum (set to 0.9) [25]. We train the model with a learning rate of 0.01 for all layers and a batch size of 256 for 1500 epochs. We also employ data augmentation [26] to increase the samples for training. We use the Keras library [27] for training our model. We take the output of the Fully Connected layer 1 to get the $1,024$-dimensional feature vector

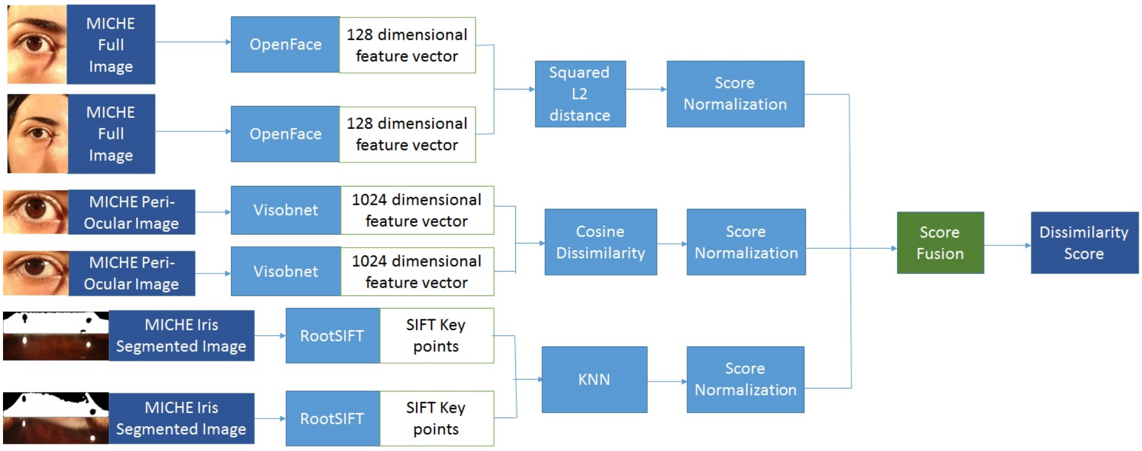

Fig. 1: Schematic representation of proposed model 1.

of the periocular image. We then compute the cosine similarity [28] between the two feature vectors of the periocular images.

3) RootSIFT: This is the Root SIFT based baseline model, described in II-A.

4) Score Fusion: We first normalize the scores [7] to bring them within the fixed numerical range of [0,1]. We then employ two approaches for score fusion. In the first approach we simply take an average of all the scores. Hence, the model and dissimilarity metric remains completely unsupervised. In the second approach we train a linear regressor on 5% of the total image pairs of the MICHE-II database [16]. Note that, this is a supervised approach, carried out to compare its results with the unsupervised verification metric.

C. Model 2

In this model we employ a supervised CNN to learn discriminative feature representations from the MICHE-II dataset. In the domain of face verification and other recognition tasks, supervised methods tend to show a clear advantage over unsupervised ones [15]. We train our model as a recognition system on 80% of the MICHE-II dataset and use the remaining 20% for validation. The details of our CNN Model are captured in Table II. The advantage of employing appearance based Convolution Neural Network is that it is able to visualize the iris from periocular region on the fly without prior need for segmentation. As MICHE-II dataset contains only a few over $3,000$ images, we resort to data augmentation to increase the robustness and generality of our model. For data augmentation, we rotate the image between 0 to 30 degrees, randomly shift the images horizontally and vertically by 0.1 of their total width and height respectively. We also flip the images horizontally and also zoom it in between 0.7 to 1.3 times it's original size. The CNN model details are similar to Section II-B2 in terms of learning algorithm, rates and kernel sizes, with the only difference being that we train this model for a $1,000$ epochs because it is shallower in comparison and hence converges faster. The input of the model is a resized RGB image from the MICHE-II data-set having dimensions of 64×96, and it's output is a 512 dimensional feature vector. Similar to Section II-B2 we employ a cosine similarity to get the similarity between two feature vectors.

D. Hybrid Model

This model is a amalgamation of our unsupervised and supervised model. Here the score of Model 2 is used in conjunction with the score of Model 1 to compute a fused dissimilarity score.

III. Evaluation

We evaluate the performance of the two proposed CNN-based models, as well as the Root SIFT baseline model, on the provided MICHE-II test dataset. For experiments, we use a hardware configuration of Intel Pentium CPU 2020M @ 2.40 GHz and 4 GB RAM. Our methodology for Model 1 achieves an execution time of approximately 1.7 seconds for inference and 1,300 seconds to run the externally provided segmentation method for a given image pair. As our Model 2 does not require any prior segmentation, it achieves a smaller execution time of 0.6 seconds for verifying a given image pair.

A. Data Description

The training dataset, as well as the test dataset, are drawn from the MICHE-II dataset, which in turn are taken from the same paradigm as MICHE-I with respect to environment, mode of capture *etc.* [16]. Two device types have been used to capture the data in the MICHE-II test data-set, namely Samsung Galaxy S4 and Apple iPhone 5. The training dataset comprises of over $3,000$ images, across all environments, devices and eyes (left/right), and has 75 distinct labels (unique subjects). The provided MICHE-II test dataset comprises of 120 images of the left and the right eyes combined. While some of the subjects present in it are part of the MICHE-II training database, most of it's subjects are new and mutually exclusive from it.

B. Model Evaluation

A test verification process is carried out, by comparing each test dataset image with one another, in all possible combinations, under each of the above settings. We perform empirical evaluation of our models under the following paradigms.

Same-Eye versus Cross-Eye: Under the *same-eye paradigm*, we hypothesize that the left and right iris of a given person are different from each other. Hence, we compare the Left Eye Images with Left Eye Images and Right Eye Images with Right Eye Images. Under the *cross-eye paradigm*, we ignore the possibility that left and right eyes could produce different features, and merge all the eye images for the comparison. The rationale behind making this apparently counter-intuitive assumption are to exploit the following. (a) Data augmentation with horizontal flip: In the data augmentation process during the deep CNN training, we also perform horizontal flip of the images, thereby the left and right eyes also getting ”interchanged” in the learning process. (b) Feature similarity: In the given image dataset features, only minor dissimilarities exist between left and right eye images of most of the given persons. We observe similar performances in these two paradigms.

Same-Device versus Cross-Device: Under the *same-device paradigm*, we compare images taken from the same device type with each other. That is, we compare images taken from Samsung Galaxy S4 only with other images taken from Samsung Galaxy S4, and images taken from Apple iPhone 5 only with other images taken from Apple iPhone 5. Under the *cross-device paradigm*, we compare between the images agnostic of the device type from which any of the images were taken from. Note that, we experiment with both the *same-eye (SE)* and *cross-eye (CE)* with the *same-device (SD)* and *cross-device (CD)* paradigms, and observe similar performance outcomes between same-eye and cross-eye testing, whereas there is a stark improvement in results when migrating from cross-device to same-device paradigm. This can be seen in Tables III and IV which correspond to EER and AUROC respectively for the various methods. Here, Model 1 LR refers to the Linear Regression based supervised score fusion technique, as opposed to EQ which refers to the unsupervised average based score fusion. Figures 2, 3, 4, 5, 6, 7 and 8 showcase the ROC curves of the various methods discussed in Section II. In these figures, the label *Default* corresponds to the CD_SE paradigm, and *Same Device* corresponds to the SD_SE paradigm. For our hybrid model we achieve an EER of 0.352 and 0.057, and AUROC of 0.736 and 0.985 in the CD_SE and SD_SE paradigms respectively. The FAR-FRR Curve for this can be found in Figure 9.

IV. Discussion

As shown in Section III, our system delivers a stark improvement over the baseline approach. This can be attributed to the use of deep learning neural network models. While supervised models clearly outperform the unsupervised ones, it is interesting to note that unsupervised models learnt on different (external) datasets also provide reasonable accuracy, when applied on the current dataset. One interesting observation is that OpenFace, a model created for facial recognition, performs reasonably well on the MICHE-II test database, where only partial faces are visible. The success of such models, opens further avenues such as using of pre-trained

Layer	Output Shape	Params
convolution2d_1	(32, 32, 48)	896
activation_1	(32, 32, 48)	0
convolution2d_2	(32, 30, 46)	9248
activation_2	(32, 30, 46)	0
maxpooling2d_1	(32, 15, 23)	0
dropout_1	(32, 15, 23)	0
convolution2d_3	(64, 15, 23)	18496
activation_3	(64, 15, 23)	0
convolution2d_4	(64, 13, 21)	36928
activation_4	(64, 13, 21)	0
maxpooling2d_2	(64, 6, 10)	0
dropout_2	(64, 6, 10)	0
convolution2d_5	(128, 6, 10)	73856
activation_5	(128, 6, 10)	0
convolution2d_6	(128, 4, 8)	147584
activation_6	(128, 4, 8)	0
maxpooling2d_3	(128, 2, 4)	0
dropout_3	(128, 2, 4)	0
flatten_1	(1024)	0
dense_1	(1024)	1049600
activation_7	(1024)	0
dropout_4	(1024)	0
dense_2	(586)	600650
activation_8	(586)	0
	Total params	1937258

TABLE I: Our architecture for Visobnet features. The output size is given by filters×rows×cols.

Layer	Output Shape	Params
convolution2d_1	(32, 64, 96)	896
activation_1	(32, 64, 96)	0
convolution2d_2	(32, 62, 94)	9248
activation_2	(32, 62, 94)	0
maxpooling2d_1	(32, 31, 47)	0
dropout_1	(32, 31, 47)	0
convolution2d_3	(64, 31, 47)	18496
activation_3	(64, 31, 47)	0
convolution2d_4	(64, 29, 45)	36928
activation_4	(64, 29, 45)	0
maxpooling2d_2	(64, 14, 22)	0
dropout_2	(64, 14, 22)	0
flatten_1	(19712)	0
dense_1	(512)	10093056
activation_5	(512)	0
dropout_3	(512)	0
dense_2	(75)	38475
activation_6	(75)	0
	Total params	10197099

TABLE II: Our architecture for Model 2-based CNN. The output size is given by filters×rows×cols.

METHOD	CD_CE	CD_SE	SD_CE	SD_SE
Root SIFT	0.508	0.517	0.518	0.554
Visobnet	0.421	0.435	0.116	0.120
OpenFace	0.360	0.354	0.147	0.148
Model 1 LR	0.368	0.369	0.139	0.139
Model 1 EQ	0.409	0.417	0.106	0.092
Model 2	0.271	0.278	0.067	0.066

TABLE III: Equal Error Rate

METHOD	CD_CE	SD_SE	SD_CE	CD_SE
Root SIFT	0.500	0.453	0.486	0.486
Visobnet	0.637	0.924	0.928	0.623
OpenFace	0.619	0.924	0.922	0.694
Model 1 LR	0.691	0.956	0.956	0.688
Model 1 EQ	0.664	0.946	0.948	0.653
Model 2	0.827	0.981	0.984	0.815

TABLE IV: AUROC

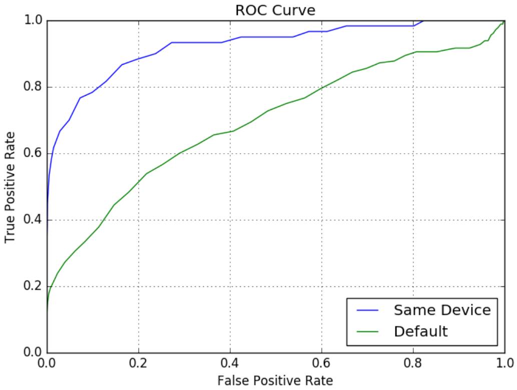

Fig. 2: OpenFace ROC

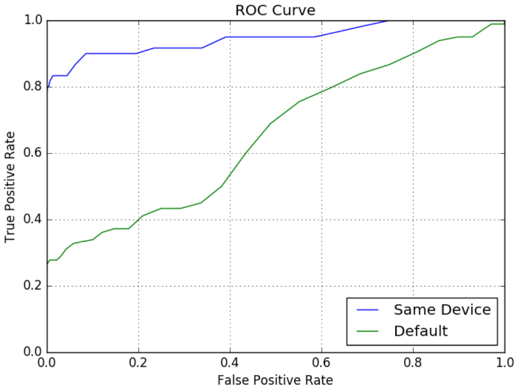

Fig. 3: Model 1 EQ ROC

models - trained on a larger (external) dataset, albeit for a slightly different task such as face or periocular recognition, and fine-tuning its last layers for addressing the complexities of the target dataset; thus maintaining the generality and robustness of the system, and at the same time fitting the model better for the target dataset. It will also be of interest to explore feature embeddings that directly correspond to image similarity, such as the Weighted X^2 distance.

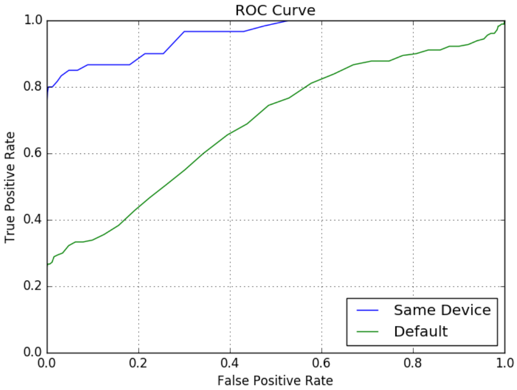

Fig. 4: Model 1 LR ROC

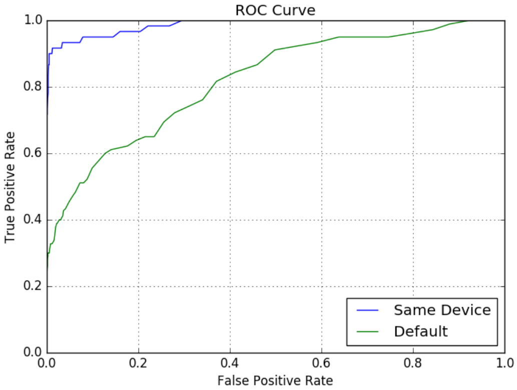

Fig. 5: Model 2 ROC

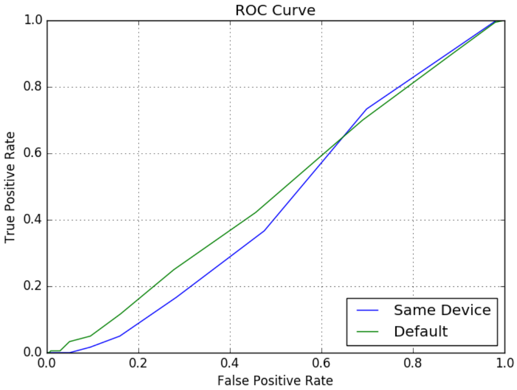

Fig. 6: Root SIFT ROC

V. Conclusion

In this paper, we proposed a baseline model, namely Root SIFT, and two stacked convolution-based deep learning learning models, for identifying an individual from a periocular image. This was obtained by training the CNNs on a given set of periocular images as part of the learning phase, and verifying a pair of images during the testing phase.

Our first model, an unsupervised one, exploited *a-priori* knowledge to perform transfer learning, stemming from (a) a

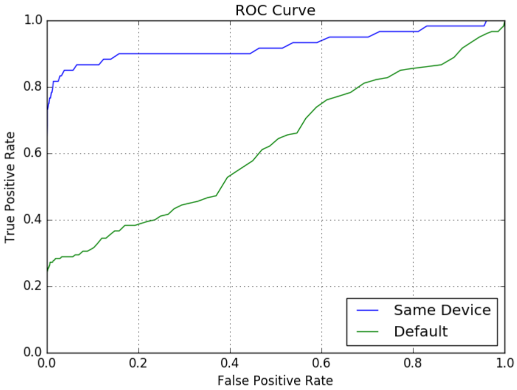

Fig. 7: Visobnet ROC

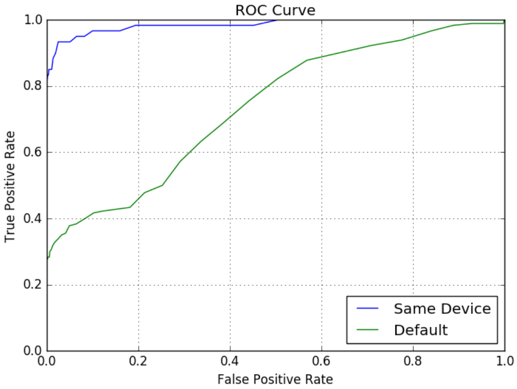

Fig. 8: Hybrid Model ROC

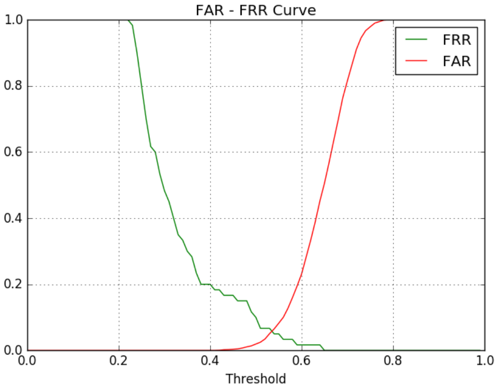

Fig. 9: Hybrid Model FAR-FRR in the SD_SE paradigm

128-dimensional facial feature vector exposed by OpenFace [14], and (b) a $1,024$ dimensional feature vector of the periocular region trained on VISOB database [9]. It obtained similarity scores for each source-target pair using each of the two methods, used Root SIFT on the provided (MICHE-II) test data to obtain a dissimilarity score, and finally applied an average-based and a linear regression based score fusion technique to identify the best-matching source-target pair. The second model, on the other hand, used a 4-layer stacked convolution network followed by a 512-dimensional feature vector for supervised learning, and used cosine similarity for testing purposes. The first model produces a best-case AUROC of 0.956 and EER of 0.092, and the second produces a best-case AUROC of 0.981 and EER of 0.066, respectively. Both significantly outperform the baseline Root SIFT method applied on the provided data, which yields a best-case performance of 0.453 and EER of 0.554. Further, a combination of the two models, is observed to deliver the best performance, under the constraint that the training and test data arise from the same device type, achieving an AUROC of 0.985. The encouraging performance delivered by both our models, signify the potential of these models as candidates for deployment in real-life applications.

References

[1] U. Park, R. R. Jillela, A. Ross, and A. K. Jain, "Periocular biometrics in the visible spectrum," *IEEE Transactions on Information Forensics and Security*, vol. 6, no. 1, pp. 96–106, 2011.

[2] U. Park, A. Ross, and A. K. Jain, "Periocular biometrics in the visible spectrum: A feasibility study," in *Biometrics: Theory, Applications, and Systems (BTAS'09).* IEEE, 2009, pp. 1–6.

[3] D. L. Woodard, S. J. Pundlik, J. R. Lyle, and P. E. Miller, "Periocular region appearance cues for biometric identification," in *CVPR Workshops*. IEEE, 2010, pp. 162–169.

[4] I. Nigam, M. Vatsa, and R. Singh, "Ocular biometrics: A survey of modalities and fusion approaches," *Information Fusion*, vol. 26, pp. 1–35, 2015.

[5] M. Haindl and M. Krupička, "Unsupervised detection of non-iris occlusions," *Pattern Recognition Letters*, vol. 57, pp. 60–65, 2015.

[6] M. De Marsico, M. Nappi, and D. Riccio, "Noisy iris recognition integrated scheme," *Pattern Recognition Letters*, vol. 33, no. 8, pp. 1006–1011, 2012.

[7] C. Galdi, M. Nappi, and J.-L. Dugelay, "Multimodal authentication on smartphones: Combining iris and sensor recognition for a double check of user identity," *Pattern Recognition Letters*, 2015.

[8] M. De Marsico, C. Galdi, M. Nappi, and D. Riccio, "Firme: face and iris recognition for mobile engagement," *Image and Vision Computing*, vol. 32, no. 12, pp. 1161–1172, 2014.

[9] A. Rattani, R. Derakhshani, S. K. Saripalle, and V. Gottemukkula, "Icip 2016 competition on mobile ocular biometric recognition," in *ICIP*. IEEE, 2016, pp. 320–324.

[10] R. Raghavendra and C. Busch, "Learning deeply coupled autoencoders for smartphone based robust periocular verification," in *ICIP*. IEEE, 2016, pp. 325–329.

[11] K. B. Raja, R. Raghavendra, and C. Busch, "Collaborative representation of deep sparse filtered features for robust verification of smartphone periocular images," in *ICIP*. IEEE, 2016, pp. 330–334.

[12] K. Akuja, A. Bose, S. Nagar, K. Dey, and F. Barbhuiya, "Isure: User authentication in mobile devices using ocular biometrics in visible spectrum," in *ICIP*. IEEE, 2016, pp. 335–339.

[13] R. Arandjelović and A. Zisserman, "Three things everyone should know to improve object retrieval," in *CVPR*. IEEE, 2012, pp. 2911–2918.

[14] B. Amos, B. Ludwiczuk, and M. Satyanarayanan, "Openface: A general-purpose face recognition library with mobile applications," CMU-CS-16-118, CMU School of Computer Science, Tech. Rep., 2016.

[15] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *CVPR*, 2014, pp. 1701–1708.

[16] M. De Marsico, M. Nappi, D. Riccio, and H. Wechsler, "Mobile iris challenge evaluation (miche)-i, biometric iris dataset and protocols," *Pattern Recognition Letters*, vol. 57, pp. 17–23, 2015.

[17] 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)*. IEEE, 2009, pp. 1–8.

[18] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[19] M. Muja and D. G. Lowe, "Fast approximate nearest neighbors with automatic algorithm configuration." *VISAPP (1)*, vol. 2, no. 331-340, p. 2, 2009.

[20] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *CVPR*, 2014, pp. 1891–1898.

[21] X. Cao, D. Wipf, F. Wen, G. Duan, and J. Sun, "A practical transfer learning algorithm for face verification," in *ICCV*, 2013, pp. 3208–3215.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *CVPR*, 2015, pp. 1026–1034.

[23] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *ECCV*. Springer, 2014, pp. 818–833.

[24] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Cognitive modeling*, vol. 5, no. 3, p. 1, 1988.

[25] I. Sutskever, J. Martens, G. E. Dahl, and G. E. Hinton, "On the importance of initialization and momentum in deep learning." *ICML (3)*, vol. 28, pp. 1139–1147, 2013.

[26] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," *arXiv preprint arXiv:1405.3531*, 2014.

[27] F. Chollet, "Keras," https://github.com/fchollet/keras, 2015.

[28] H. V. Nguyen and L. Bai, "Cosine similarity metric learning for face verification," in *Asian Conference on Computer Vision*. Springer, 2010, pp. 709–720.