A. Brömme, C. Busch, A. Dantcheva, K. Raja, C. Rathgeb and A. Uhl (Eds.): BIOSIG 2020,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2020 1

Compact Models for Periocular Verification Through Knowledge Distillation

Fadi Boutros[12], Naser Damer[12], Meiling Fang[12], Kiran Raja[3], Florian Kirchbuchner[1], Arjan Kuijper[12]

Abstract: Despite the wide use of deep neural network for periocular verification, achieving smaller deep learning models with high performance that can be deployed on low computational powered devices remains a challenge. In term of computation cost, we present in this paper a lightweight deep learning model with only 1.1m of trainable parameters, DenseNet-20, based on DenseNet architecture. Further, we present an approach to enhance the verification performance of DenseNet-20 via knowledge distillation. With the experiments on VISPI dataset captured with two different smartphones, iPhone and Nokia, we show that introducing knowledge distillation to DenseNet-20 training phase outperforms the same model trained without knowledge distillation where the Equal Error Rate (EER) reduces from 8.36% to 4.56% EER on iPhone data, from 5.33% to 4.64% EER on Nokia data, and from 20.98% to 15.54% EER on cross-smartphone data.

Keywords: Periocular recognition, Smartphone biometric verification, Knowledge distillation.

1 Introduction

The rapid growth of smartphone users (3.2 billion in 2019 [St20]) has also increased the interest in secure authentication application using smartphones. Biometric modalities like fingerprint, voice, periocular and face are widely employed on smartphones to achieve secure, convenient, and reliable authentication.

Of the many other modalities, periocular region provides a distinct trade-off between using iris or entire face for identity verification by considering a small area around the eye including eyelids, lashes, and eyebrows as biometric trait [PRJ09]. Given the performance under relaxed settings, periocular biometrics is recently well preferred for various use cases such as mobile platform [Al19] and embedded device [Bo19, Bo20a, Bo20b]. Motivated by such new applications, we focus on periocular modality for smartphone based biometric identity verification in this work.

Although the integration of biometrics in smartphone devices has enabled several advantages, deploying such a solution to a smartphone device faces several challenges. One of these challenges is the high variability between probe and gallery images produced when the images are acquired using different devices, different cameras, or under different environmental conditions, requiring a highly generalized solution. This challenge is well addressed in the literature as reported in the previous works [Al19, Ah17]. Yet another major challenge is related to the limited computational resources available in smartphone

[1] Fraunhofer Institute for Computer Graphics Research IGD, Darmstadt, Germany
[2] Mathematical and Applied Visual Computing, TU Darmstadt, Darmstadt, Germany
[3] The Norwegian Colour and Visual Computing Laboratory, NTNU, Gjovik, Norway

devices, especially when considering a solution based on a deep neural network with extremely high number of parameters. Recent works [Ga18, Ah17] have addressed this issue with the smartphone based periocular recognition using deep learning, albeit with less focus on the limited computation resource available on the smartphone devices where both models [Ga18, Ah17] contain more than 12 million of trainable parameters. Despite the use of deep learning, the challenge of customizing the solutions to smartphone devices with limited computational resources is not well addressed.

We therefore focus this work on reducing the number of parameters in the deep models to make them easily adaptable to mobile devices with limited computation resources by utilizing knowledge distillation noted as KD [HVD15] for periocular verification. To truly establish the applicability of the proposed approach for periocular verification, we provide the baseline performance of three DenseNet architectures [Hu17]: DenseNet-201, DenseNet-169, and DenseNet-121. Further, we propose a compact model which we refer to as DenseNet-20 based on the dense block containing 1.1 million of trainable parameters. The experimental results on VISPI dataset [KRB20] of 152 unique pericoular instances with 6682 images captured with 2 different smartphones (iPhone 5S and Nokia Lumia 1020) shows that the DenseNet-20 model achieves a comparable verification performance using a shallow architecture. With the obtained performance, we argue that deploying such a model to a low computational resource device is more realistic than other deeper models. Motivated by this, we also focus on enhancing the accuracy and generalizability of the shallow model for periocular recognition by successfully introducing the KD method to the training process. Although introducing knowledge distillation to the training process does not change the model capacity, the gradient descent induced by distillation loss function allows this model to find a very favorable minimum of the training objective [PL19]. Thus, our proposed approach improves the verification performance of the distilled model, in comparison to the same model trained without knowledge distillation, the Equal Error Rate (EER) is reduced from 8.36% to 4.56% on iPhone data, from 5.33% to 4.64% EER on Nokia smartphone data, and from 20.98% to 15.54% EER on cross-smartphone data.

2 Methodology

The goal of this work is to present a solution to improve the accuracy and generalizability of shallow CNN models for smartphone periocular verification. Particularly, we first evaluate deep representations extracted from periocular region using three different DenseNet [Hu17] architectures: DensNet-121, DensNet-169, and DenseNet-201. We further present our proposed compact CNN model, DenseNet-20, containing only 1.1 million trainable parameters. To further improve the generalizability and accuracy of the small CNN model, we introduce knowledge distillation (KD) to the DenseNet-20 model training process. This section presents the details of the employed DenseNet model along with the KD method.

2.1 Densely Connected Convolutional Networks

DenseNet [Hu17] is a convolutional neural network designed for image classification to achieve low classification error rates while having fewer parameters than ILSVRC 2015 winner, ResNet model [He16]. The architecture is based on connecting each convolutional layer to every other layer in a feed-forward fashion as shown in Figure 1. Thus, each layer

ℓ^{th} receives collective knowledge from all preceding layers $x_0, x_1, ..., x_{\ell-1}$ and passes on its knowledge to all subsequent layers. Given that each layer produces k feature maps, the input feature map for ℓ^{th} layer is $k_0 + k \times (\ell - 1)$ where k_0 is the number of channels in the input layer and k refers to the growth rate of the network. In this work, we evaluate three different DenseNet architectures as baselines: DenseNet-121, DenseNet-169, and DenseNet-201 where 121, 169, and 201 refer to the number of the convolutional layers in each model (network depth). The growth rate for all the networks is set to $k = 32$. The DenseNet-121, DenseNet-169, and DenseNet-201 models contain 7.1, 12.6 and 18.2m of trainable parameters, respectively.

We apply transfer learning on these models pretrained on ImageNet dataset [De09] by fine-tuning all the layers on images from our training dataset with Softmax classifier. In the test phase, the Softmax classifier is removed from all models and the feature f is extracted from the last layer which is of the dimension $7 \times 7 \times 1920$.

2.2 Proposed Compact DenseNet

We further propose a new model based on DenseNet architecture - DenseNet-20. Similar to the original DenseNet model, DenseNet-20 has 4 dense blocks with 1, 2, 8, and 6 layers in dense block 1, 2, 3, and 4, respectively. We train the compact DenseNet-20 model from scratch with Softmax classifier. The proposed DenseNet-20 contains 1.1m trainable parameters as compared to 18.2 million parameters with DenseNet-201. Similar to the original DenseNet models, the Softmax classifier is removed in the testing phase from the model to extract the feature f from the last layer with the dimension of $7 \times 7 \times 1920$.

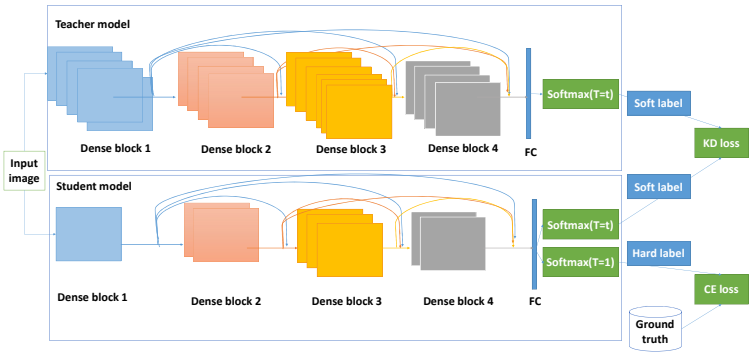

Fig. 1: An overview of the proposed KD approach for periocular verification based on DenseNet architecture.

2.3 Proposed Compact DenseNet-20 with Knowledge Distillation

We explore KD to improve the performance of DenseNet-20 model by employing a student-teacher relation where each of DenseNet-121, DenseNet-169, and DenseNet-201 models are used as a teacher to distill the knowledge to student model, DenseNet-20. We present the details of KD for the convenience of the readers.

KD is a technique to improve the performance and generalization ability of smaller models by transferring the knowledge learned by a cumbersome model (teacher) to a single

small model (student). The key idea is to guide the student model to learn the relationship between different classes discovered by the teacher model that contains more information beyond the ground truth labels [HVD15]. Suppose we have teacher model T, student model S, and training dataset $X, Y \in D$, where X is the training images and Y is their class labels. The output of the teacher model for any input $x_i \in X$ is a vector of class probabilities P^T computed for each class using softmax function by converting the logits, z^T into probabilities that sum to one $P^T(x) = softmax(z^T)$. Specifically, the probability p_i of class i is computed by comparing z_i with other logits as given: $p_i = \frac{exp(z_i)}{\sum_{j=1}^{N} exp(z_j)}$. This probability distribution will have a high probability value of p_i for the correct class $y_i \in Y$ with all other class probabilities close to zeros. Thus, it does not provide more valuable information than ground truth labels. Therefore, Hinton et al. [HVD15] proposed to scale the logits using a temperature parameter $t > 1$ before applying the softmax function. Thus, the teacher model can produce a softer distribution of the class probabilities, which provides more valuable information about classes similar to the predicted class. In this case, the output of the teacher model is $P_s^T(x) = Softmax(z^T/t)$ and the probability p_i of class i is given as: $p_i = \frac{exp(z_i/t)}{\sum_{j=1}^{N} exp(z_j/t)}$. Similarly, student S can produce a soft class probability distribution using the temperature parameter t, $P_s^S(x) = Softmax(z^S/t)$. The final loss for the student model is a weighted sum of two loss functions, cross-entropy loss L_{ce} and Kullback Leibler Divergence loss L_{kld}, as follows:

$$L_{KD} = \lambda * L_{ce}(Y, P^S(x)) + (1-\lambda) * t^2 * L_{kld}(P_s^S(x), P_s^T(x)),$$

where Y is the ground truth label, $P^S(x)$ standard softmax output produced by student, $P_s^S(x)$ parameterized softmax output produced by student, $P_s^T(x)$ parameterized Softmax output produced by teacher and $\lambda \in [0,1]$ is the weight parameter. Since the gradients of the L_{ce} loss is smaller than gradients of the L_{kld} where the logits used for L_{kld} is divided by t, the L_{kld} is multiplied by t^2 as suggested by Hinton et al. [HVD15].

We thus use the student-teacher based KD for all three DenseNet models - DenseNet-201, DenseNet-169 and DenseNet-121 by setting each of them as teacher and our proposed DenseNet-20 as the student model as shown in the Figure 1.

3 Experimental setup

To demonstrate the applicability of our proposed approach, we evaluate it on a public dataset of periocular images - VISPI database [KRB20]. We employ the subset of database containing 152 unique periocular instances captured from 76 unique subjects using two different smartphones - iPhone 5S and Nokia Lumia 1020. The 152 periocular instances are captured from both left and right eyes-76 instances are captured from the left eye and 76 instances are captured from the right eye. Each unique periocular image has multiple samples captured in different instances. The total distribution of the images in the database used for the evaluation in this work is presented in the Table 1.

Details	Smartphone	
	iPhone 5S	Nokia Lumia 1020
Capture Scenario	Mixed Illumination	Mixed Illumination
Resolution	12 Mp	41 Mp
Number of subjects	76	76
Unique periocular instances	152	152
Total images	3341	3341

Tab. 1: Distribution of periocular database employed in this work.

The ocular images are captured in a mixed illumination environment using the rear camera of the smartphones in a semi-cooperative manner. The images in the database also present everyday appearance variations that include the make-up and non-uniform illumination. Beside, the images in the VISPI database present various forms of degradation due to motion blur and eye blinking. Further, the influence of both the external sunlight illumination and the artificial room light illumination along with other degrading factors make the cross-sensor/cross-smartphone comparison challenging. The sample images from the periocular database as depicted in Figure 2 illustrate a set of variation and degradation in terms of appearance under different smartphones both across the phones and the subjects.

Of the 152 unique periocular instances, the first 100 instances (from 50 subjects, i.e., 50 instances captured from the left eye and 50 instances captured from the right eye) are used for the training and the other 52 instances (from 26 independent subjects, i.e., 26 instances captured from the left eye and 26 instances captured from the right eye) are used for testing. Further, a random subset of 200 images (two images per instance) is selected from the training split to validate the model during the training phase.

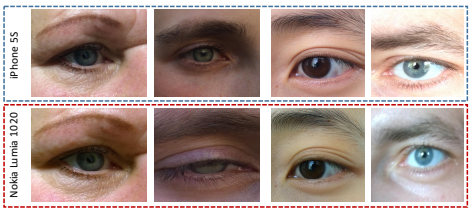

Fig. 2: Sample images from VISPI database

All the images are uniformly resized to a size of 224×224 pixels to match the input layer size of DenseNet model. The training data is augmented by applying horizontal and vertical random shifting by up to 20% of the image width and/or height, and random horizontal flipping. All models are trained with a batch size of 16 and SGD optimizer with Nesterov momentum 0.9. The initial learning rate is set to $\gamma = 0.001$ and $\gamma = 0.1$ for teacher models and student model, respectively and it is dropped by a factor of 0.1 when the accuracy on the validation dataset does not improve by a value of 0.1 for 5 consequent epochs. The initial number of epochs is set to 100 and early-stopping patience parameter is set to 10 causing DenseNet-20, DenseNet-121, DenseNet-169 and DenseNet-201 to stop after 29, 11, 11, 11 epochs, respectively. The training of the student model, DenseNet-20, trained KD loss stopped after 29, 34, and 28 epoch using teacher model DenseNet-121, DenseNet-169 and DenseNet-201, respectively. In practice, the training is performed offline once and the trained model is deployed on mobile devices, which makes the size of the model the most critical deployment factor. We followed the common choice for the KD hyperparameters [HVD15, CH19, Fu18] with Temperature $t >= 4$ and $\lambda = 0.9$.

The verification performance is reported using the cosine similarity measure for comparing the features extracted from the learnt models. The result is reported first for the DenseNet-20, DenseNet-121, DenseNet-169, and DenseNet-201 models without applying the KD. In addition, we report the result of the KD on the student model DenseNet-20 with DenseNet-121, DenseNet-169 or DenseNet-201 as a teacher which we note as as DenseNet-20-KD121, DenseNet-20-KD169 and DenseNet-20-KD201 respectively.

For each of the settings, we investigate the verification performance under three different evaluation scenarios defined as following:

• iPhone verification scenario: The reference and the probe images are acquired using the camera of iPhone smartphone.

• Nokia verification scenario: Similar to the previous scenario, the reference and the probe images are acquired using Nokia smartphone.

• Cross-smartphone verification scenario: the reference images are captured using iPhone camera and the probe images are captured using Nokia camera.

The verification performance is reported using Receiver Operating Characteristic (ROC) curves, Area under the curve (AUC), False Match Rate (FMR) at fixed False Non-Match Rate (FNMR) (FMR10, the lowest FNMR for FMR$\leq$10%), and Equal Error Rate (EER). The verification performances of the different experimental settings are presented in Figure 3 along with the EER and FMR10 values in Table 2. Each of the Figures 3.a-c shows the achieved ROC of iPhone, Nokia, and cross-smartphone verification scenario.

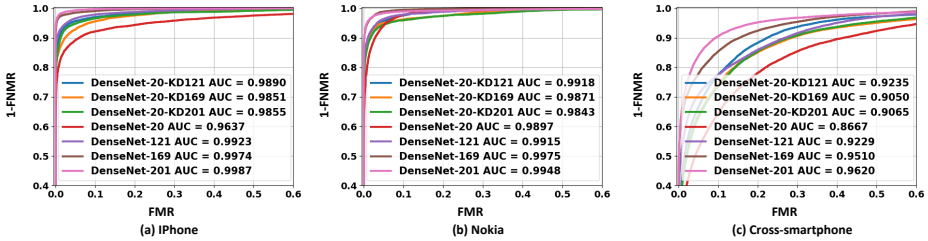

Fig. 3: The achieved ROC for different experimental settings. One can be clearly noticed the improvement in the DenseNet-20 verification performance using KD method.

4 Results and Discussion

One can clearly notice from Table 2 that the verification performances are consistently better based on the model size when same training procedure is followed. The first four rows in Table 2 present the verification performances of the DenseNet-201, DenseNet-169, DenseNet-121, and DenseNet-20 trained without the KD method. The highest verification performance among all evaluated models is achieved by the DenseNet-201 model, where the achieved EER was 9.71% for cross-smartphone verification scenario, 1.60% for iPhone verification scenario, and 2.57% for Nokia verification scenario. Also, it can be observed from the Table 2 that the DenseNet-20 model aims at maintaining (to a large degree) the verification performance of deeper model where the achieved EERs were 8.36%, 5.33% and 20.98% for iPhone, Nokia and cross-smartphone verification scenarios, respectively.

It can be further noticed that the verification performances degrade for all models when the references and probes images are captured from different smartphones in comparison to the case where the probe and the reference images are captured from the same smartphone as shown in the Table 2. However, this degradation in the performance is a common problem for cross-smartphone verification scenario as reported in the previous works [Al19].

4.1 Impact of Knowledge Distillation

The results of the proposed approach based on KD are presented in the Table 2 and Figure 3. We make the following observations from the obtained results:

Model	Inference time	Num. of. parameters	Teacher	iPhone		Nokia		Cross-smartphone	
				EER	**FMR10**	**EER**	**FMR10**	**EER**	**FMR10**
DenseNet-201	5.4ms	18.2	-	0.0160	0.0026	0.0257	0.0105	0.0971	0.0949
DenseNet-169	4.7ms	12.6	-	0.0220	0.0093	0.0256	0.0052	0.1224	0.1459
DenseNet-121	3.8ms	7.1	-	0.0396	0.0212	0.0417	0.0213	0.1666	0.2257
DenseNet-20	2.1ms	1.1m	-	0.0836	0.0782	0.0533	0.0227	0.2098	0.3556
DenseNet-20-KD201	2.1ms	1.1m	DenseNet-201	0.0515	0.0340	0.0538	0.0404	0.1709	0.2640
DenseNet-20-KD2169	2.1ms	1.1m	DenseNet-169	0.0617	0.0440	0.0496	0.0376	0.1711	0.2582
DenseNet-20-KD121	2.1ms	1.1m	DenseNet-121	**0.0456**	**0.0298**	**0.0464**	**0.0240**	**0.1554**	**0.2251**

Tab. 2: Performance obtained for different experimental settings along with inference time (in millisecond) and the number of trainable parameters (in million) for each of the evaluated models. The first four rows of the table present the achieved result for the three teacher models and for the student models (without using KD). The last three rows of the table present the achieved verification performance by including KD in the training process.

• It is noticed that introducing the KD to the DenseNet-20 model training significantly improved the verification performance and outperforms teacher model in some cases. For example, in the cross-smartphone verification scenario, the student outperformed its teacher DenseNet-121 where the achieved EER by the student was 15.54% and by its teacher was 16.66%. Similar observations is also reported in in previous work [Fu18].

• The best verification performance is achieved using DenseNet-121 model as teacher, where the achieved EERs in this case were 5.56% 4.64% and 15.54% for iPhone, Nokia and cross-smartphone verification scenarios.

• Using a larger and more accurate teacher model did not serve as better supervision to the student model as seen in Table 2. This can be explained by the fact that as the teacher model becomes more accurate using a deeper architecture, the soft probabilities produced by the teacher will contain more complex information about the class distributions and the small student model will not be able to learn all this complex information considering the small student capacity. Similar conclusion is also reported in the previous work [CH19].

5 Conclusion

We presented in this work a new approach for periocular verification exploiting the idea of Knowledge distillation (KD). The proposed models have resulted in significantly lower model size but with comparable performance to larger deep models. Through the experiments on public periocular dataset consisting of 152 unique periocular instances captured with two different smartphones, we showed that applying KD on DenseNet-20 training process achieves an EER of 4.5% on iPhone data, 4.6% on Nokia data, and 15.54% on cross-smartphone data, in comparison to EER of 8.36% on iPhone data, 5.33% on Nokia data, and 20.98% on cross-smartphone data when the same model trained without KD. In the future works in this direction, we intend to investigate the proposed method on larger datasets captured in multiple sessions to gain insights on generalizability aspects.

Acknowledgment: this research work has been funded by the German Federal Ministry of Education and Research and the Hessen State Ministry for Higher Education, Research and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE.

References

[Ah17] Ahuja, Karan; Islam, Rahul; Barbhuiya, Ferdous A; Dey, Kuntal: Convolutional neural networks for ocular smartphone-based biometrics. Pattern Recognition Letters, 2017.

[Al19] Alonso-Fernandez, Fernando; Raja, Kiran B.; Raghavendra, Ramachandra; Busch, Christoph; Bigün, Josef; Vera-Rodríguez, Rubén; Fiérrez, Julian: Cross-Sensor Periocular Biometrics: A Comparative Benchmark including Smartphone Authentication. CoRR, abs/1902.08123, 2019.

[Bo19] Boutros, Fadi; Damer, Naser; Kirchbuchner, Florian; Kuijper, Arjan: Eye-MMS: Miniature Multi-Scale Segmentation Network of Key Eye-Regions in Embedded Applications. In: Proceedings of the IEEE/CVF ICCVW. Oct 2019.

[Bo20a] Boutros, F.; Damer, N.; Raja, K.; Ramachandra, R.; Kirchbuchner, F.; Kuijper, A.: Periocular Biometrics in Head-Mounted Displays: A Sample Selection Approach for Better Recognition. In: 2020 8th IWBF. pp. 1–6, 2020.

[Bo20b] Boutros, Fadi; Damer, Naser; Raja, Kiran; Ramachandra, Raghavendra; Kirchbuchner, Florian; Kuijper, Arjan: Iris and Periocular Biometrics within Head Mounted Displays: Segmentation, Recognition, and Synthetic Generation. Image Vis. Comput., 2020.

[CH19] Cho, Jang Hyun; Hariharan, Bharath: On the efficacy of knowledge distillation. In: Proceedings of the IEEE ICCV. pp. 4794–4802, 2019.

[De09] Deng, Jia; Dong, Wei; Socher, Richard; Li, Li-Jia; Li, Kai; Li, Fei-Fei: ImageNet: A large-scale hierarchical image database. In: CVPR 2009, 20-25 June 2009, Miami, Florida, USA. IEEE Computer Society, pp. 248–255, 2009.

[Fu18] Furlanello, Tommaso; Lipton, Zachary Chase; Tschannen, Michael; Itti, Laurent; Anandkumar, Anima: Born-Again Neural Networks. In (Dy, Jennifer G.; Krause, Andreas, eds): Proceedings of the 35th ICML 2018, Sweden. volume 80 of PMLR, pp. 1602–1611, 2018.

[Ga18] Garg, Rishabh; Baweja, Yashasvi; Ghosh, Soumyadeep; Singh, Richa; Vatsa, Mayank; Ratha, Nalini: Heterogeneity aware deep embedding for mobile periocular recognition. In: 9th BTAS. IEEE, 2018.

[He16] He, Kaiming; Zhang, Xiangyu; Ren, Shaoqing; Sun, Jian: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778, 2016.

[Hu17] Huang, Gao; Liu, Zhuang; Van Der Maaten, Laurens; Weinberger, Kilian Q: Densely connected convolutional networks. In: Proceedings of the IEEE CVPR. pp. 4700–4708, 2017.

[HVD15] Hinton, Geoffrey E.; Vinyals, Oriol; Dean, Jeffrey: Distilling the Knowledge in a Neural Network. CoRR, abs/1503.02531, 2015.

[KRB20] Kiran Raja; Ramachandra, Raghavendra; Busch, Christoph: Collaborative representation of blur invariant deep sparse features for periocular recognition from smartphones. Image and Vision Computing, p. 103979, 2020.

[PL19] Phuong, Mary; Lampert, Christoph: Towards understanding knowledge distillation. In: International Conference on Machine Learning. pp. 5142–5151, 2019.

[PRJ09] Park, Unsang; Ross, Arun; Jain, Anil K: Periocular biometrics in the visible spectrum: A feasibility study. In: 2009 IEEE 3rd BTAS. IEEE, pp. 1–6, 2009.

[St20] Number of smartphone users worldwide from 2016 to 2021. "https://www.statista.com/statistics/330695/number-of-smartphone-users-worldwide/".