Lightweight Periocular Recognition through Low-bit Quantization

Jan Niklas Kolf[1,2], Fadi Boutros[1,2], Florian Kirchbuchner[1,2], Naser Damer[1,2]
[1]Fraunhofer Institute for Computer Graphics Research IGD, Darmstadt, Germany
[2]Department of Computer Science, TU Darmstadt, Darmstadt, Germany
Email: jan.niklas.kolf@igd.fraunhofer.de

Abstract

Deep learning-based systems for periocular recognition make use of the high recognition performance of neural networks, which, however, is accompanied by high computational costs and memory footprints. This can lead to deployability problems, especially in mobile devices and embedded systems. Few previous works strived towards building lighter models, however, while still depending on floating-point numbers associated with higher computational cost and memory footprint. In this paper, we propose to adapt model quantization for periocular recognition. This, within the proposed scheme, leads to reducing the memory footprint of periocular recognition network by up to five folds while maintaining high recognition performance. We present a comprehensive analysis over three backbones and diverse experimental protocols to stress the consistency of our conclusions, along with a comparison with a wide set of baselines that prove the optimal trade-off between performance and model size achieved by our proposed solution. The code and pre-trained models have been made available at https://github.com/jankolf/ijcb-periocular-quantization.

1. Introduction

Biometrics deals with the recognition of people based on their physical or behavioural characteristics. The most relevant physical characteristics include the face, the iris or even fingerprints [39]. Especially for mobile devices, recognition based on biometric features became well-established [67, 69]. Modern smartphones are often equipped with a fingerprint sensor and a front camera that makes it possible to unlock the device using facial characteristics [69]. However, a front camera not only enables the use of the entire face but also various parts such as the iris or the periocular region around the eye. This region includes not only the eye itself but also the eyelids, the eyebrow, parts of the cheekbone, with rich textural and color information [42, 67]. In order to be able to use the iris as a feature, a good image with good exposure and no occlusion is required [67], which in reality, especially in the unconstrained scenario in everyday life, is not feasible for a mobile device [80]. The periocular region recognition does not rely on detailed image information as in iris recognition and it still provides enough discrepancy and uniqueness to allow recognition [80]. High-performing models for face recognition (FR), a well-established biometric modality, use over-parameterized deep neural networks (NN) which, however, are associated with high memory footprint and great computational complexity [11, 20, 25]. This severely restricts the use of these NN on embedded or mobile devices with limited computing possibilities, with power supply from a battery and with shared resources between several different tasks [19, 26, 50, 51]. To get lighter and faster NN and better recognition rates for mobile devices where the full capture of the face is not always feasible, periocular recognition (PR) can be used instead [67]. This is also the case for emerging deployments such as the verification of head-mounted display users in VR/AR applications [15, 16, 17]. This motivation has been additionally stressed lately with the COVID-19 pandemic and the limitations of FR systems and their components while wearing a mask [23, 29], which maintains the visibility of the periocular region [12, 28]. There are various approaches to developing efficient NNs and to overcome the limitations of mobile devices [24].
On the one hand, more efficient and lightweight network architectures are being developed, commonly for FR, whose computation is more efficient but can offer similar performance [18, 44, 45, 83] such as ShuffleFaceNet [49], Mixfacenets [9] or Mobilefacenets [21]. Another approach is knowledge distillation, where a complex teacher network is used to train a compact student network [10, 46]. While this can produce efficient and compact models, these techniques also rely on specific hardware, as the computation of such NNs is based on floating-point operations.

To target this limitation, model quantization approaches can be used to avoid this problem by reducing the parameter space to integer operations. Model quantization has been recently applied to regulate the computational cost of deep

FR [13]. However, it has not been yet proposed and investigated for PR in previous works [2, 43, 57, 61, 64, 67, 84]. Model quantization replaces floating-point parameters and operations with integer operations that have a shorter bit length [13, 27, 30, 47, 77]. The reduced bit length not only saves memory footprint but also speeds up the inference time of an NN, as integer operations can be performed faster depending on the architecture and hardware used. Deep learning program libraries such as PyTorch can also run a quantized model more efficiently, reducing computation time by a factor of between 2 and 4 [41, 55]. Due to the smaller memory size of quantized models, less data has to be transferred via the memory bus, which also reduces the memory bandwidth. Typical bit lengths for integer for the quantization approach are 8 bits [32, 38, 40, 41], a standard size for integer, and 4 bits [6, 79, 85].

In this paper, we propose to build lightweight and accurate PR models by adapting model quantization within a framework that ensures, to a large scale, the maintenance of the PR recognition performance of the full precision models. We demonstrate successfully that quantization can be used to reduce the computational complexity of NNs for PR while maintaining similar recognition rates. We investigate this proposal at different quantization levels and applied it on three diverse NN backbone architectures to insure generalized conclusions. Our models adapts the margin-based loss function ArcFace [25] in their training process on the recent large-scale UFPR-Periocular database [80]. As a supplementary contribution, our work investigates the optimal penalty margin in the training loss specifically for the PR problem. We additionally analysed the effect of encoding the periocular identity information on different embedding sizes on the performance of PR. Both identification and verification scenarios are evaluated [39], pointing out the success of our proposed solution and its ability to achieve an optimal trade-off between PR model size and performance in comparison to a wide set of baselines.

Section 3 presents our proposed framework, including the used angular margin penalty-based loss function with the evaluated margin values and embedding sizes, as well as the adapted quantization paradigm. In Section 4, the experimental setup including the dataset, evaluation metrics, and the quantization experimental details are presented. The results of the experiments are discussed in Section 5 and the effects of different quantization levels are analysed. The final conclusion is set out in Section 6.

2. Related Work

After Park et al. [53] used the periocular region with Local Binary Patterns (LBP) for recognition in a feasibility study, other studies followed up with similar approaches, e.g. Bharadwaj et al. [8] and Xu et al. [78]. Besides LBP, Scale Invariant Feature Transformation (SIFT) has also been applied, e.g. by Ross et al. [65], Alonso-Fernandez et al. [4] or by Ahuja et. al [1] and Raja et al. [59], who have also developed it explicitly for mobile devices. Tan et al. [71] used Leung-Mallik filters to acquire multiple features of the periocular region. In addition to various feature extraction methods, deep learning mechanisms were also used to perform PR on mobile devices. Nie et al. [52] applied a convolutional version of Restricted Boltzman Machines to learn features in an unsupervised approach which were then fused into a metric learning system. Zhang et al. [84] fused features of the iris and periocular region extracted from a convolutional neural network (CNN) for recognition. Raja et al. [58] also used deep learning to create features from the periocular region and used them for recognition. To do this, the authors used deep sparse filtering and evaluated their methodology on several smartphones. Reddy et al. [63] considered feature learning with a convolutional autoencoder trained in an unsupervised learning approach. With their method, the authors were able to beat a supervised trained ResNet50 in their evaluation. Proença et al. [56] have developed a data augmentation pipeline that modifies regions of the periocular area to implicitly provide prior information about the relevance of regions for biometric recognition to the CNN used. They have concluded that without the information and details of the ocular globe, recognition performance increases. Zanlorensi et al. [81] used a generative model to reduce within-class variability in unconstrained periocular recognition through normalization of the input. With this method, the authors increased the recognition performance of systems tested. Alonso-Fernandez et al. [5] used several feature extractors to combine them into one. They also did cross-spectrum experiments between visible and near-infrared images and achieved very good recognition rates. Rattani et. al [60] evaluated in a competition various submitted algorithms that showed very good recognition rates on different mobile devices and conditions. To reduce the computational effort for PR on mobile devices, Reddy et al. [62] proposed OcularNet, which uses multiple small CNNs to extract individual features from six overlapping patches from the periocular region, which were then combined into a feature vector. They were able to beat a ResNet50 model even though their model was 15.6x smaller. Almadan et al. [3] use pruning techniques to create a small and efficient NN for ocular recognition on mobile devices. In a similar direction Boutros et al. [10, 14] used knowledge distillation to transfer the knowledge learned by relatively larger networks to smaller ones, with the transfer performed in a traditional manner [10] or on the template level [14]. While many papers have considered periocular recognition for mobile devices, only a few put emphasis on the compactness of the used models. Additionally, none have proposed to use quan-

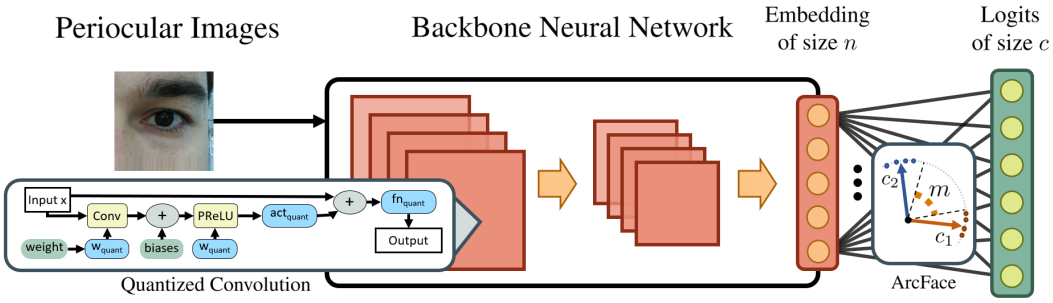

Figure 1. The figure illustrates the pipeline used, consisting of the input images, the backbone used including the quantized convolution module and the margin penalty-based loss function ArcFace [25]. UFPR-Periocular [80] is used as the image database and the images are scaled to 224×224. The backbone network creates an embedding feature vector sized n. From the embedding, the angular margin penalty-based loss function ArcFace with margin m calculates the logits for class activations. For c identities in the dataset there is one class activation, resulting in a c sized output vector. With quantization-aware training, individual simulated quantization operations are incorporated into the neural network to simulate the quantization behaviour during inference. Weights and activation functions as well as their results are quantized.

tization techniques to create a model without floating-point operations, enabling lighter deployment of the technology, as we do in this paper.

3. Methodology

To train a PR model, a network architecture and a loss function are needed. In this work, we incorporate the use of an angular margin penalty-based softmax losses, specifically the ArcFace loss [25], a well-performing loss for identity representation learning through classification, with which very good results for FR have been obtained [25, 75]. In order to create efficient and small NNs for PR, the weights and activation functions of a floating-point pre-trained model are reduced to b bit by a uniform quantization-aware training (QAT) procedure [31], which is explained in further detail in Section 3.2.

3.1. Angular Margin Penalty-based Loss Function

To achieve good recognition rates for PR, good inter-class discrepancies and intra-class compactness are needed [11]. A classic cross-entropy loss via softmax activations cannot achieve this distinction optimally [25], especially for verification approaches, since ambiguities occur especially at the decision boundaries [25]. In order to achieve a better separation between individual classes, angular margin penalty-based loss functions such as ArcFace [25] or ElasticFace [11] were introduced. The ArcFace loss used in this work is defined as [25]:

$$L = -\frac{1}{N}\sum_{i \in N} log \frac{e^{s(cos(\theta_{y_i}+m))}}{e^{s(cos(\theta_{y_i}+m))} + \sum_{j=1, j\neq y_i}^{c} e^{s\ cos(\theta_j)}}. \quad (1)$$

The loss function uses a feature vector of size n created by an NN to calculate the cosine similarity to the other class centers. To increase the separation to other classes, margins are used. By using different margins, individual samples are moved closer to their respective class centers. ArcFace loss for face recognition achieved top performances typically when setting $m = 0.5$, as investigated in [25]. However, due to the different nature of the identity information in the periocular region, in comparison to face, the optimal ArcFace margin value might be different. This assumption was not previously investigated and is a core part of this work. Additionally, and on the same level of importance, the optimal embedding size to represent the identity of the periocular region was also not previously investigated and is a contribution of this work. As a reference point, typical embedding size for face representations is 128 to 512 [11, 25, 74]. Therefore, as a further contribution, we have conducted an investigation that evaluates the influence of the embedding size n and the margin m in Section 5.1. The individual sizes and where they are located in the pipeline can also be seen in figure 1.

3.2. Model Quantization

The main aim of model quantization is to reduce the memory footprint of an NN that uses floating-point operations. Additionally, depending on the hardware used, the calculation of integer values is faster than the calculation of floating-point operations [41, 55, 76], which also can reflect in the inference time. The quantization approach is to replace the floating-point parameters of the NN with integer values with bit size b. Both the weights of the network and the activation functions themselves can be quantized. These can also be quantized to integers with different bit sizes. [33, 38]. However, the integer values must have the same influence as the floating-point parameters so that the network has the same behaviour. In practice, this is not always the case with direct quantization, which is why the network can be adapted to the new parameter values in an extra training step after quantization. This pro-

cedure is called quantization-aware training (QAT) [33]. A more detailed description of QAT follows in this chapter as it is adopted as the basis of this work. A signed integer r_Q with bit size b has 2^b different values and a range of values of $r_Q \in [-2^{b-1}, 2^{b-1}-1]$. Since the number of different digits of a b bit integer is smaller than that of 32 bit floating-point numbers (FP32) - in order to save computational effort - it is important to map the parameters to the entire value range of the integer. The theoretical value range of a FP32 number r according to the IEEE standard [86] is $r \in [-3.4 \times 10^{38}, 3.4 \times 10^{38}]$ [73]. For NN, in reality, a weight parameter and the results of arithmetic operations and activation functions do not lie in the entire range of values of a floating-point number. This is because some activation functions have a fixed range of values, there are normalisation and regularisation operations and for normalised inputs after training there are defined ranges of values for all parameters. For a FP32 parameter r within an NN let $r \in [\alpha, \beta]$ [31]. To represent FP32 value r with r_Q, the value range of r is mapped to that of r_Q. This can be done uniformly, where the values are mapped equally distributed to $[\alpha, \beta]$, or non-uniformly, where the distances between individual values of r in the new value range are different [31, 33]. Uniform mapping is used in this paper and is defined as [31, 38]

$$Q(r) = round(\frac{r}{S}) - Z, \tag{2}$$

where the operation $round(\cdot)$ rounds the floating-point number to the nearest integer. The constant Z is a b bit signed integer and is called zero-point. It specifies the zero-point shift of the mapping from the range of values of r and that of r_Q. The constant S is the scaling factor and divides the value range into equally sized partitions. Since the range of values from the r can be asymmetric ($-\alpha \neq \beta$) S and Z are defined as [38]

$$S = \frac{\beta - \alpha}{2^b - 1}, \tag{3}$$

$$Z = round(\beta \cdot \frac{2^b - 1}{\beta - \alpha} - 2^b - 1). \tag{4}$$

To reconstruct the original FP32 value from a quantized value, there is the dequantization operation [31, 38]:

$$D(r_Q) = S \cdot (r_Q + Z). \tag{5}$$

The range of r, limited by α and β, can be calculated either statically or dynamically. If it is calculated dynamically, the minimum and maximum values for individual parameters are stored during re-training and after a specified number of calibration steps this calculation is frozen and the values are set as α and β [31]. In this work, dynamic calibration is used to obtain a better parameter mapping.

The parameter ranges can also be calculated for different groups of parameters, this is called quantization granularity. In CNN, filters and layers can have different value ranges, which can therefore also be quantized differently. There are three different methodologies [31]: Layerwise, groupwise and channelwise quantization. In layerwise quantization, all filters of a convolution or the entire layer are considered in order to determine α and β. In groupwise quantization, groups are formed for which the range of values is determined. In channelwise quantization, the range of values is determined for individual filters of a convolutional layer. Due to its high popularity and good performance, channelwise quantization is used in this work [37, 82].

3.3. Quantization-Aware Training

The presented mapping involves a loss of representativity of the NN, as the learned parameters of the NN are not trained to deal with the lower expressiveness. This results in a loss of performance due to quantization. To counteract this, the NN is trained again in order to adjust the quantized parameters [31]. This process is called quantization-aware training and involves introducing simulated quantization operations. However, the respective forward and backward pass of the NN is performed in floating-point precision. After each gradient update, the parameters are quantized. To optimize the quantization parameter, the Straight-Through Estimator (STE) [7] is used, as the gradient of the quantized parameters through the flat step function is zero [31]. STE sets the derivative to 1, provided the input is within the learned value range $[\alpha, \beta]$, else it is set to zero.

4. Experimental Setup

In this Section, the model architectures as well as the dataset and training parameters used in this work are presented. The hardware used for the experiments is a GPU server with 2×16 core Intel Xeon Gold 6130s, 256GB RAM, and 4 Nvidia GeForce RTX 2080 Ti 11GB GPUs.

4.1. Dataset and evaluation protocols

The UFPR-Periocular [80] dataset consists of 33,660 images of periocular regions, left and right side of the face, from 1,122 subjects taken in 3 sessions with 196 different mobile devices. The images were not taken in a controlled scenario and include blur, occlusion and lighting variations. This dataset is chosen as it represent realistic use-case scenario in comparison to some of the available datasets mentioned in [80]. Additionally, this dataset is of a large scale in terms of different subjects, sessions and sensors [80], which makes it suitable for training NN models and of significant size to present realistic evaluation results. From the individual images, the periocular region was detected and extracted based on the eyes, rotated to a normalised eye position, and then the two periocular regions were cropped into

512×512 images each. All these preprocessing steps are performed by the data creators and are described in detail in [80]. The dataset contains three different training and evaluation protocols. One protocol is the closed world protocol and it represents an identification scenario. There, all identities contained in the training set are also contained in the validation and test set. The goal is to find the identity for an image from all other images of the test set from a 1:N comparison. The other two protocols are for the verification scenario, which is to check whether the identity the subject claims is correct. This is a 1:1 comparison between a sample, the given image, and a reference that the subject claims to be. In both verification protocols, the test set does not contain images of identities that are included in the training set. These protocols are noted as the open world scenario. The difference between the two protocols is that in one protocol, the identities from the training set are included in the validation set, which is called open world/closed validation. In the other verification protocol, there are different identities in the training, validation, and test set (i.e. identity-disjoint). This protocol is called open world/open validation. Each protocol is divided into three different folds, for each of which there is a training, validation and test set. For a more detailed description of the dataset, please refer to the work of Zanlorensi et al. [80]. We follow the procedure of Uzair et. al [72] and flip images containing the left eye horizontally so that they have the same orientation as images with right eyes. Images from the left and right eyes are considered to have the same identity in our experimental setup. All images are scaled to size 224×224 and each is normalized with mean and std. of 0.5. We follow the process described in [80] and train and evaluate the models and quantization for both the closed world identification protocol and the open world/closed validation verification protocol. The benchmark in each protocol is run and evaluated for all three folds. As described in [80], the rank 1 and rank 5 metrics, calculated from the pairwise cosine similarity between the embedding vectors, as well as the area under curve (AUC) and the equal error rate (EER) are calculated for the identification protocol. The metrics are explained in more detail in the following Section. For the evaluation of the verification, the AUC and the EER are calculated.

4.2. Evaluation metrics

The rank metric used for identification scenarios indicates in what percentage of the cases examined the identity of a person in the database is the 1st place (rank 1) or is within the first five rankings (rank 5). The ranking is created by the cosine similarity between the input and the database, the test set and is sorted in descending order according to the similarity score. The Equal Error Rate (EER) is defined as the operation point where the false match rate (FMR) is equal the false non-match rate (FNMR), based on the ISO/IEC 19795-1 [48] standard. The receiver operating characteristics (ROC) curve gives the individual $1 - \text{FNMR}$ and FMR for more operation points [39]. The area under curve (AUC) metric specifies how good the verification performance is. A 1.0 metric means that the system can perfectly distinguish comparisons between the identical identity and foreign identities.

4.3. Backbone architectures

All backbone full-precision models utilizing FP32 numbers are trained with the ArcFace [25] loss (as defined in Section 3.1) on the UFPR-Periocular [80] dataset. We are following [25] and set the scaling parameter s for the ArcFace loss to 64. The architectures used are ResNet18 [34] and ResNet50 [34] as well as MobileFaceNet [21]. These were chosen as they have shown good performances in different biometric applications [75]. All models are trained with a learning rate of 0.1 for 20 epochs, taking the model with the best metrics on the validation set. We follow a similar approach as [11] and reduce the learning rate by a factor of 10 after the 8th and 15th epoch. The optimiser used for all models is Stochastic Gradient Descent with a momentum of 0.9 and weight decay of $5e-4$. PyTorch [55] is used as the framework for the implementation.

4.4. Quantization Procedure

The weights and activations of all models, both for verification and identification, are quantized in three different bit sizes, each with 8 (W8A8), 6 (W6A6), and 4 (W4A4) bits. To adapt the quantized models, 20 epochs were again performed on the UFPR periocular dataset. The learning rate for the 8 and 6 bit quantization is 0.1, for the 4 bit quantization the learning rate 0.01 is used, because a learning rate of 0.1 leads to unstable optimisation. A batch size of 8 was used for ResNet18 and ResNet50, and a batch size of 16 for MobileFaceNet. Again, after the 8th and 15th epoch, the learning rate is reduced by a factor of 10. The first 8 epochs were used as calibration steps, after these epochs the values were taken as respective α and β values.

5. Results

In this Section we discuss our achieved results under three main investigations. First, we analyse the benefits of adapting a margin-based penalty loss for PR, including a comprehensive analyses of the suitable margin value for PR, along with investigating the most suitable embedding size for PR, both previously unstudied issues. Second we presents and discuss the performance of our different quantized models under various evaluation protocols and experimental settings. Third, we put the performance and size of our quantized models in perspective by comparing it to a large set of baselines.

Table 1. The results for the AUC and EER metric for different margin values m and embedding sizes n tested with a ResNet18 FP32 on the open world/open validation protocol of the UFPR-Periocular [80] dataset. The overall best value combination is in bold. It can be seen that nearly for all margin values the embedding size $n = 512$ is best. Also, $m = 0.5$ and $n = 512$, the values ArcFace [25] uses for FR, is not the best combination. The best results for PR for the verification task are achieved with $m = 1.0$ and $n = 512$.

m	n	**AUC (%)**	**EER (%)**
0.0	128	96.964 ± 0.329	9.042 ± 0.497
	256	96.984 ± 0.267	8.825 ± 0.449
	512	96.906 ± 0.267	8.931 ± 0.406
0.1	128	97.188 ± 0.247	8.505 ± 0.377
	256	97.314 ± 0.290	8.242 ± 0.493
	512	97.311 ± 0.265	8.174 ± 0.455
0.2	128	97.278 ± 0.169	8.368 ± 0.239
	256	97.464 ± 0.257	7.862 ± 0.435
	512	97.539 ± 0.254	7.706 ± 0.430
0.3	128	97.244 ± 0.201	8.442 ± 0.427
	256	97.598 ± 0.291	7.634 ± 0.459
	512	97.724 ± 0.231	7.338 ± 0.421
0.4	128	97.376 ± 0.269	8.162 ± 0.482
	256	97.752 ± 0.245	7.341 ± 0.472
	512	97.934 ± 0.226	6.888 ± 0.399
0.5	128	97.517 ± 0.208	7.890 ± 0.407
	256	97.862 ± 0.267	7.149 ± 0.566
	512	98.040 ± 0.205	6.656 ± 0.437
0.6	128	97.527 ± 0.187	7.806 ± 0.357
	256	97.928 ± 0.217	6.943 ± 0.448
	512	98.071 ± 0.193	6.509 ± 0.454
0.7	128	97.657 ± 0.230	7.624 ± 0.472
	256	98.002 ± 0.165	6.772 ± 0.314
	512	98.178 ± 0.188	6.313 ± 0.429
0.8	128	97.726 ± 0.176	7.431 ± 0.355
	256	98.044 ± 0.121	6.742 ± 0.327
	512	98.130 ± 0.199	6.412 ± 0.477
0.9	128	97.736 ± 0.227	7.419 ± 0.445
	256	98.044 ± 0.127	6.708 ± 0.349
	512	98.200 ± 0.181	6.272 ± 0.378
1.0	128	97.774 ± 0.184	7.320 ± 0.327
	256	98.053 ± 0.148	6.686 ± 0.403
	512	$\mathbf{98.200 \pm 0.152}$	$\mathbf{6.240 \pm 0.412}$
1.1	128	85.140 ± 12.87	20.443 ± 12.883
	256	98.054 ± 0.176	6.647 ± 0.432
	512	98.144 ± 0.187	6.371 ± 0.422
1.2	128	70.511 ± 1.922	35.275 ± 1.723
	256	64.947 ± 1.474	39.440 ± 1.031
	512	98.042 ± 0.123	6.550 ± 0.289

5.1. Investigating margin-based penalty loss and embedding size

To investigate the impact of margin values and embedding sizes on PR performance, we train an FP32 ResNet18 model on the UFPR open world/open validation protocol. ArcFace is used as a loss function. The parameter m which specifies the margin in ArcFace loss and the embedding size n are varied. Values from 0.0 to 1.2 are tested for m with increments of 0.1. For $m \geq 1.3$, the training is unstable, and therefore no greater margin value is evaluated. This conclusion sets the optimal m for PR far from that for face, which is specified in [25] to be 0.5. The embedding sizes n evaluated were 128, 256 and 512. The results across all test folds including mean and standard deviation for EER and AUC in percent are presented in table 1. As shown in the table, the embedding size $n = 512$ is the best for nearly each margin value. This indicates that the smaller embedding sizes do not have sufficient dimensionality to encode the identities efficiently, under the investigated setup. The best margin value is $m = 1.0$. With increasing margin values, the performance decreases again.

5.2. The effect of model quantization

The results for the full-precision (FP32) and quantized models (W8A8, W6A6, W4A4) on the open world/closed validation protocol for the verification tasks of the UFPR dataset are shown in table 2. For each of the three models used, ResNet18, ResNet50, and MobileFaceNet, the results are listed first for the FP32 and then for the quantized models. The number of parameters and the required memory size are also given. quantization not only reduces the required memory size but also the required memory bandwidth. This is especially important for mobile devices in order to ensure fast execution and verification or identification. Therefore, a trade-off between the required memory space and the recognition performance is important. Following the same format, table 3 shows the results for the identification tasks according to the closed world protocol. The rank 1 and rank 5 metrics show the identification performance, while the AUC and EER show the verification performance on the closed world set. Based on the two tables, the following observations can be made:

a) Influence of 8 bit quantization: The quantization to 8 bit results in a model size of one quarter of the originally required memory size. In the verification scenario in the open world protocol, all three models only slightly lose performance. ResNet18 performance drops from 5.76% EER to 5.99%, a similar drop is observed for ResNet50 (5.88% EER to 5.99%) and MobileFaceNet (3.86% to 4.22%). In the identification scenario, the drop in performance due to the closed world protocol is lower. There, the quantized ResNet18 and ResNet50 models achieve approximately the same EER and AUC values. Only in the rank 1 and rank 5 metrics is a slight drop noticeable. MobileFaceNet and ResNet50 in quantized form beat the full-precision model in the EER metric and are also very close to the full-precision model in the rank 1 and rank 5 metrics, although it only

Table 2. The results for all three trained backbones with their respective quantization levels. Training and testing were carried out on the open world/closed validation protocol in the verification scenario. It can be seen that with 8 or 6 bit quantization the models have a similar recognition performance and only slightly lose recognition rates. Although the models are 25% smaller, the AUC and EER metrics only deteriorate slightly.

Model	Params	Bits	Size (MB)	Verification (1:1)	
				AUC (%)	EER (%)
ResNet18	62.560M	FP32	250.24	98.51 ± 0.15	5.76 ± 0.38
		W8A8	62.56	98.41 ± 0.16	5.99 ± 0.39
		W6A6	46.92	98.35 ± 0.15	6.17 ± 0.40
		W4A4	31.28	75.69 ± 3.17	29.18 ± 2.70
ResNet50	82.125M	FP32	328.50	98.47 ± 0.17	5.88 ± 0.38
		W8A8	82.13	98.39 ± 0.18	5.99 ± 0.41
		W6A6	61.59	98.29 ± 0.18	6.28 ± 0.41
		W4A4	41.06	66.02 ± 1.07	37.78 ± 0.94
MobileFaceNet	1.276M	FP32	5.10	99.23 ± 0.05	3.86 ± 0.21
		W8A8	1.28	99.11 ± 0.05	4.22 ± 0.09
		W6A6	0.96	99.18 ± 0.06	4.02 ± 0.19
		W4A4	0.64	55.01 ± 3.24	50.49 ± 8.39

Table 3. The results for all three trained backbones with their respective quantization levels. Training and testing were carried out on the closed world protocol for the identification and verification scenario. It can be seen that with 8 or 6 bit quantization the models have a similar recognition performance and only slightly lose recognition rates. All models can keep or beat the previous recognition rates for a quantization level of 8 bits. The results for bit size 6 are also very close to the original values, depending on the model. Bit size 4 is not sufficient to ensure good recognition.

Model	Params	Bits	Size (MB)	Identification (1:N)		Verification (1:1)	
				Rank 1 (%)	Rank 5 (%)	AUC (%)	EER (%)
ResNet18	62.560M	FP32	250.24	99.61 ± 0.08	99.88 ± 0.03	99.77 ± 0.01	1.75 ± 0.05
		W8A8	62.56	99.54 ± 0.11	99.85 ± 0.05	99.76 ± 0.01	1.75 ± 0.02
		W6A6	46.92	99.48 ± 0.11	99.79 ± 0.05	99.73 ± 0.01	1.92 ± 0.04
		W4A4	31.28	50.96 ± 12.54	62.01 ± 11.25	86.18 ± 3.96	19.17 ± 4.12
ResNet50	82.125M	FP32	328.50	99.54 ± 0.04	99.81 ± 0.04	99.76 ± 0.02	1.76 ± 0.04
		W8A8	82.13	99.47 ± 0.02	99.82 ± 0.04	99.75 ± 0.01	1.74 ± 0.02
		W6A6	61.59	99.34 ± 0.07	99.71 ± 0.08	99.73 ± 0.01	1.95 ± 0.09
		W4A4	41.06	33.24 ± 3.82	46.29 ± 4.09	80.55 ± 2.03	26.75 ± 1.98
MobileFaceNet	1.276M	FP32	5.10	99.87 ± 0.06	99.92 ± 0.03	99.86 ± 0.01	1.26 ± 0.09
		W8A8	1.28	99.79 ± 0.05	99.91 ± 0.03	99.86 ± 0.01	1.18 ± 0.08
		W6A6	0.96	99.80 ± 0.06	99.91 ± 0.03	99.86 ± 0.01	1.22 ± 0.13
		W4A4	0.64	5.53 ± 1.13	11.19 ± 2.46	62.28 ± 3.30	42.50 ± 3.99

requires a quarter of the memory space.

b) Influence of 6 bit quantization: Both ResNet models slightly lose performance when quantized to 6 bits in comparison to 8 bit, while the drop in the open world/closed validation protocol is a bit larger. It is noteworthy that although the models are only 18% of the original size, the performance drop of the EER in the closed world protocol is only from 1.75% to 1.92% for ResNet18. MobileFaceNet beats the EER of the FP32 model in the closed world protocol and achieves similar results in the open world/closed validation protocol as well. Despite the drastic reduction in memory requirements, the networks can demonstrate very good PR performance.

b) Influence of 4 bit quantization: All models lose immense performance when quantized to 4 bit, which is 12.5% of the original bit size. MobileFaceNet in particular jumps from 99.87% in the full-precision model to 5.53% in the rank 1 metric. In this experimental setup, the 4 bits used do not have sufficient significance to learn encoding the identities sufficiently well and to produce a good separation between them.

An overall observation from the results indicates that our proposed quantization of various periocular recognition models is extremely successful in maintaining the accuracy of the full-precision models with very minor drop in the recognition performance. The model size, however, could be reduced immensely to 25% or 18.75% of the original memory footprint.

5.3. Comparison with State-Of-The-Art

The goal of this work is to enhance periocular model compactness through model quantization and is not concerned with specifically achieving the highest recognition accuracy. However, in table 4 and table 5 we report the performances of the quantized models proposed in this work and a set of previously reported periocular recognition solutions (in [80]) along with the model size of the utilized architectures. Considering both the performance on the individual evaluation protocols and the model footprint, the models presented in this work outperform the baselines with comparable footprint in most experimental settings and offer an optimal trade-off between performance and model size.

Table 4. Comparison of our quantized models and a set of baselines trained using the same protocol. All the baselines are reported as in [80]. The comparison is given as the verification performance in the open world/closed validation protocol and the model size in MB. When considering the model size, our quantized models outperform many of the much larger models.

Model	Params	Size (MB)	Verification (1:1)	
			AUC (%)	EER (%)
VGG16 [68, 80]	135.89M	1088	97.38 ± 0.53	8.52 ± 0.92
VGG16-Face [54, 80]	135.89M	1088	97.70 ± 0.42	7.78 ± 0.75
InceptionResNet [70, 80]	55.25M	445	99.10 ± 0.24	4.61 ± 0.65
ResNet50V2 [35, 80]	49.79M	400	98.73 ± 0.28	5.69 ± 0.64
ResNet50 [34, 80]	24.61M	198	98.60 ± 0.28	5.98 ± 0.67
ResNet50-Face [20, 80]	24.61M	198	99.18 ± 0.16	4.38 ± 0.47
Xception [22, 80]	21.91M	176	98.93 ± 0.16	5.23 ± 0.42
DenseNet121 [36, 80]	7.79M	64	99.51 ± 0.12	3.39 ± 0.46
Multi-task [80]	4.49M	37	99.67 ± 0.08	2.81 ± 0.39
MobileNetV2 [66, 80]	3.13M	26	99.56 ± 0.08	3.17 ± 0.33
Siamese [80]	2.55M	21	97.27 ± 0.64	8.10 ± 1.01
Pairwise [80]	2.35M	20	98.62 ± 0.72	5.77 ± 1.57
ResNet18 W8A8	62.56M	62.56	98.41 ± 0.16	5.99 ± 0.39
ResNet18 W6A6	62.56M	46.92	98.35 ± 0.15	6.17 ± 0.40
ResNet50 W8A8	82.13M	82.13	98.39 ± 0.18	5.99 ± 0.41
ResNet50 W6A6	82.13M	61.59	98.29 ± 0.18	6.28 ± 0.41
MobileFaceNet W8A8	1.28M	1.28	99.11 ± 0.05	4.22 ± 0.09
MobileFaceNet W6A6	1.28M	0.96	99.18 ± 0.06	4.02 ± 0.19

Table 5. Comparison of our quantized models and a set of baselines trained using the same protocol. All the baselines are reported as in [80]. The comparison is given as the identification and verification performance in the closed world protocol and the model size in MB. When considering the model size, our quantized models outperform many of the much larger models.

Model	Params	Size (MB)	Identification (1:N)		Verification (1:1)	
			Rank 1 (%)	Rank 5 (%)	AUC (%)	EER (%)
VGG16 [68, 80]	135.89M	1088	50.56 ± 3.30	68.73 ± 3.01	99.41 ± 0.11	3.59 ± 0.32
VGG16-Face [54, 80]	135.89M	1088	56.29 ± 1.62	73.84 ± 1.48	99.43 ± 0.08	3.44 ± 0.28
InceptionResNet [70, 80]	55.25M	445	65.16 ± 2.45	81.53 ± 1.99	99.78 ± 0.15	1.85 ± 0.40
ResNet50V2 [35, 80]	49.79M	400	63.18 ± 2.14	77.79 ± 1.81	99.74 ± 0.04	2.24 ± 0.18
ResNet50 [34, 80]	24.61M	198	71.06 ± 1.14	85.22 ± 0.82	99.89 ± 0.02	1.41 ± 0.10
ResNet50-Face [20, 80]	24.61M	198	73.76 ± 1.43	86.86 ± 1.02	99.83 ± 0.03	1.74 ± 0.12
Xception[22, 80]	21.91M	176	57.43 ± 1.43	75.88 ± 1.52	99.77 ± 0.04	2.19 ± 0.18
DenseNet121[36, 80]	7.79M	64	75.54 ± 1.36	88.53 ± 0.97	99.93 ± 0.02	1.11 ± 0.09
Multi-task [80]	4.49M	37	84.32 ± 0.71	94.55 ± 0.58	99.96 ± 0.01	0.81 ± 0.06
MobileNetV2 [66, 80]	3.13M	26	77.98 ± 1.08	90.19 ± 0.79	99.93 ± 0.01	1.13 ± 0.07
Siamese [80]	2.55M	21	–	–	98.94 ± 0.22	4.86 ± 0.44
Pairwise [80]	2.35M	20	–	–	99.44 ± 0.66	3.06 ± 1.84
ResNet18 W8A8	62.56M	62.56	99.54 ± 0.11	99.85 ± 0.05	99.76 ± 0.01	1.75 ± 0.02
ResNet18 W6A6	62.56M	46.92	99.48 ± 0.11	99.79 ± 0.05	99.73 ± 0.01	1.92 ± 0.04
ResNet50 W8A8	82.13M	82.13	99.47 ± 0.02	99.82 ± 0.04	99.75 ± 0.01	1.74 ± 0.02
ResNet50 W6A6	82.13M	61.59	99.34 ± 0.07	99.71 ± 0.08	99.73 ± 0.01	1.95 ± 0.09
MobileFaceNet W8A8	1.28M	1.28	99.79 ± 0.05	99.91 ± 0.03	99.86 ± 0.01	1.18 ± 0.08
MobileFaceNet W6A6	1.28M	0.96	99.80 ± 0.06	99.91 ± 0.03	99.86 ± 0.01	1.22 ± 0.13

6. Conclusion

This work is the first to propose the adaption of model quantization for periocular recognition. Three different popular neural networks were quantized to different bit sizes to insure the generalizability of the conclusions. The evaluation of the models shows that similar recognition performance can be achieved with quantization to 8 and 6 bit, while reducing the model size to 25% and 18.75% of its original full precision size, respectively. This contribution was accompanied by detailed analyses on the optimal loss penalty margin for PR and the optimal embedding size to represent identity information in the periocular region. Model quantization, within the presented framework, is therefore proved, in a comprehensive comparison to a set of baselines, to be a good approach for creating lightweight periocular recognition systems.

Acknowledgement This research work has been funded by the German Federal Ministry of Education and Research and the Hessian Ministry of Higher Education, Research, Science and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE.

References

[1] K. Ahuja, A. Bose, S. Nagar, K. Dey, and F. Barbhuiya. Isure: User authentication in mobile devices using ocular biometrics in visible spectrum. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 335–339. IEEE, 2016. 2

[2] K. Ahuja, R. Islam, F. A. Barbhuiya, and K. Dey. Convolutional neural networks for ocular smartphone-based biometrics. *Pattern Recognition Letters*, 91:17–26, 2017. 2

[3] A. Almadan and A. Rattani. Compact cnn models for on-device ocular-based user recognition in mobile devices. In *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1–7. IEEE, 2021. 2

[4] F. Alonso-Fernandez and J. Bigun. Best regions for periocular recognition with nir and visible images. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 4987–4991. IEEE, 2014. 2

[5] F. Alonso-Fernandez, K. B. Raja, R. Raghavendra, C. Busch, J. Bigun, R. Vera-Rodriguez, and J. Fierrez. Cross-sensor periocular biometrics in a global pandemic: Comparative benchmark and novel multialgorithmic approach. *Information Fusion*, 83:110–130, 2022. 2

[6] R. Banner, Y. Nahshan, and D. Soudry. Post training 4-bit quantization of convolutional networks for rapid-deployment. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[7] Y. Bengio, N. Léonard, and A. Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013. 4

[8] S. Bharadwaj, H. S. Bhatt, M. Vatsa, and R. Singh. Periocular biometrics: When iris recognition fails. In *2010 fourth IEEE international conference on Biometrics: Theory, Applications and Systems (BTAS)*, pages 1–6. IEEE, 2010. 2

[9] F. Boutros, N. Damer, M. Fang, F. Kirchbuchner, and A. Kuijper. Mixfacenets: Extremely efficient face recognition networks. In *2021 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–8. IEEE, 2021. 1

[10] F. Boutros, N. Damer, M. Fang, K. Raja, F. Kirchbuchner, and A. Kuijper. Compact models for periocular verification through knowledge distillation. In *2020 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–5. IEEE, 2020. 1, 2

[11] F. Boutros, N. Damer, F. Kirchbuchner, and A. Kuijper. Elasticface: Elastic margin loss for deep face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2022*. Computer Vision Foundation / IEEE, 2022. 1, 3, 5

[12] F. Boutros, N. Damer, J. N. Kolf, K. B. Raja, F. Kirchbuchner, R. Ramachandra, A. Kuijper, P. Fang, C. Zhang, F. Wang, D. Montero, N. Aginako, B. Sierra, M. Nieto, M. E. Erakin, U. Demir, H. K. Ekenel, A. Kataoka, K. Ichikawa, S. Kubo, J. Zhang, M. He, D. Han, S. Shan, K. Grm, V. Struc, S. Seneviratne, N. Kasthuriarachchi, S. Rasnayaka, P. C. Neto, A. F. Sequeira, J. R. Pinto, M. Saffari, and J. S. Cardoso. MFR 2021: Masked face recognition competition. In *International IEEE Joint Conference on Biometrics, IJCB 2021, Shenzhen, China, August 4-7, 2021*, pages 1–10. IEEE, 2021. 1

[13] F. Boutros, N. Damer, and A. Kuijper. Quantface: Towards lightweight face recognition by synthetic data low-bit quantization. In *26th International Conference on Pattern Recognition, ICPR 2022, Montreal,Quebec ,August 21-25, 2021*. IEEE, 2022. 2

[14] F. Boutros, N. Damer, K. B. Raja, F. Kirchbuchner, and A. Kuijper. Template-driven knowledge distillation for compact and accurate periocular biometrics deep-learning models. *Sensors*, 22(5):1921, 2022. 2

[15] F. Boutros, N. Damer, K. B. Raja, R. Ramachandra, F. Kirchbuchner, and A. Kuijper. Fusing iris and periocular region for user verification in head mounted displays. In *IEEE 23rd International Conference on Information Fusion, FUSION 2020, Rustenburg, South Africa, July 6-9, 2020*, pages 1–8. IEEE, 2020. 1

[16] F. Boutros, N. Damer, K. B. Raja, R. Ramachandra, F. Kirchbuchner, and A. Kuijper. Iris and periocular biometrics for head mounted displays: Segmentation, recognition, and synthetic data generation. *Image Vis. Comput.*, 104:104007, 2020. 1

[17] F. Boutros, N. Damer, K. B. Raja, R. Ramachandra, F. Kirchbuchner, and A. Kuijper. Periocular biometrics in head-mounted displays: A sample selection approach for better recognition. In *8th International Workshop on Biometrics and Forensics, IWBF 2020, Porto, Portugal, April 29-30, 2020*, pages 1–6. IEEE, 2020. 1

[18] F. Boutros, P. Siebke, M. Klemt, N. Damer, F. Kirchbuchner, and A. Kuijper. Pocketnet: Extreme lightweight face recognition network using neural architecture search and multi-step knowledge distillation. *IEEE Access*, pages 1–1, 2022. 1

[19] Y. Cai, Z. Yao, Z. Dong, A. Gholami, M. W. Mahoney, and K. Keutzer. Zeroq: A novel zero shot quantization framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13169–13178, 2020. 1

[20] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE, 2018. 1, 8

[21] S. Chen, Y. Liu, X. Gao, and Z. Han. Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices. In *Chinese Conference on Biometric Recognition*, pages 428–438. Springer, 2018. 1, 5

[22] F. Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017. 8

[23] N. Damer, F. Boutros, M. Süßmilch, F. Kirchbuchner, and A. Kuijper. Extended evaluation of the effect of real and simulated masks on face recognition performance. *IET Biom.*, 10(5):548–561, 2021. 1

[24] A. Das, C. Galdi, H. Han, R. Ramachandra, J.-L. Dugelay, and A. Dantcheva. Recent advances in biometric technology

for mobile devices. In *2018 IEEE 9th international conference on biometrics theory, applications and systems (BTAS)*, pages 1–11. IEEE, 2018. 1

[25] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 1, 2, 3, 5, 6

[26] J. Deng, J. Guo, D. Zhang, Y. Deng, X. Lu, and S. Shi. Lightweight face recognition challenge. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 1

[27] R. Dong, Z. Tan, M. Wu, L. Zhang, and K. Ma. Finding the task-optimal low-bit sub-distribution in deep neural networks. *arXiv preprint arXiv:2112.15139*, 2021. 2

[28] M. Fang, F. Boutros, A. Kuijper, and N. Damer. Partial attack supervision and regional weighted inference for masked face presentation attack detection. In *16th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2021, Jodhpur, India, December 15-18, 2021*, pages 1–8. IEEE, 2021. 1

[29] M. Fang, N. Damer, F. Kirchbuchner, and A. Kuijper. Real masks and spoof faces: On the masked face presentation attack detection. *Pattern Recognit.*, 123:108398, 2022. 1

[30] W. Fei, W. Dai, C. Li, J. Zou, and H. Xiong. General bitwidth assignment for efficient deep convolutional neural network quantization. *IEEE Transactions on Neural Networks and Learning Systems*, 2021. 2

[31] A. Gholami, S. Kim, Z. Dong, Z. Yao, M. W. Mahoney, and K. Keutzer. A survey of quantization methods for efficient neural network inference. *arXiv preprint arXiv:2103.13630*, 2021. 3, 4

[32] J. Gong, H. Shen, G. Zhang, X. Liu, S. Li, G. Jin, N. Maheshwari, E. Fomenko, and E. Segal. Highly efficient 8-bit low precision inference of convolutional neural networks with intelcaffe. In *Proceedings of the 1st on Reproducible Quality-Efficient Systems Tournament on Co-designing Pareto-efficient Deep Learning*, page 1. 2018. 2

[33] Y. Guo. A survey on methods and theories of quantized neural networks. *arXiv preprint arXiv:1808.04752*, 2018. 3, 4

[34] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5, 8

[35] K. He, X. Zhang, S. Ren, and J. Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016. 8

[36] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 8

[37] Q. Huang, D. Wang, Z. Dong, Y. Gao, Y. Cai, T. Li, B. Wu, K. Keutzer, and J. Wawrzynek. Codenet: Efficient deployment of input-adaptive object detection on embedded fpgas. In *The 2021 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*, pages 206–216, 2021. 4

[38] B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. Howard, H. Adam, and D. Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2704–2713, 2018. 2, 3, 4

[39] A. K. Jain, A. Ross, and S. Prabhakar. An introduction to biometric recognition. *IEEE Transactions on circuits and systems for video technology*, 14(1):4–20, 2004. 1, 2, 5

[40] Q. Jin, J. Ren, R. Zhuang, S. Hanumante, Z. Li, Z. Chen, Y. Wang, K. Yang, and S. Tulyakov. F8net: Fixed-point 8-bit only multiplication for network quantization. *arXiv preprint arXiv:2202.05239*, 2022. 2

[41] R. Krishnamoorthi. Quantizing deep convolutional networks for efficient inference: A whitepaper. *arXiv preprint arXiv:1806.08342*, 2018. 2, 3

[42] P. Kumari and K. Seeja. Periocular biometrics: A survey. *Journal of King Saud University-Computer and Information Sciences*, 2019. 1

[43] P. Kumari and K. Seeja. A novel periocular biometrics solution for authentication during covid-19 pandemic situation. *Journal of Ambient Intelligence and Humanized Computing*, 12(11):10321–10337, 2021. 2

[44] H.-C. Li, Z.-Y. Deng, and H.-H. Chiang. Lightweight and resource-constrained learning network for face recognition with performance optimization. *Sensors*, 20(21):6114, 2020. 1

[45] X. Li, F. Wang, Q. Hu, and C. Leng. Airface: Lightweight and efficient model for face recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 1

[46] J. Liu, H. Qin, Y. Wu, J. Guo, D. Liang, and K. Xu. Coupleface: Relation matters for face recognition distillation. *arXiv preprint arXiv:2204.05502*, 2022. 1

[47] Y. Ma, T. Jin, X. Zheng, Y. Wang, H. Li, G. Jiang, W. Zhang, and R. Ji. Ompq: Orthogonal mixed precision quantization. *arXiv preprint arXiv:2109.07865*, 2021. 2

[48] A. Mansfield. Information technology–biometric performance testing and reporting–part 1: Principles and framework. *ISO/IEC*, pages 19795–1, 2006. 5

[49] Y. Martindez-Diaz, L. S. Luevano, H. Mendez-Vazquez, M. Nicolas-Diaz, L. Chang, and M. Gonzalez-Mendoza. Shufflefacenet: A lightweight face architecture for efficient and highly-accurate face recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 1

[50] Y. Martínez-Díaz, H. Méndez-Vázquez, L. S. Luevano, L. Chang, and M. Gonzalez-Mendoza. Lightweight low-resolution face recognition for surveillance applications. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 5421–5428. IEEE, 2021. 1

[51] Y. Martinez-Diaz, M. Nicolas-Diaz, H. Mendez-Vazquez, L. S. Luevano, L. Chang, M. Gonzalez-Mendoza, and L. E. Sucar. Benchmarking lightweight face architectures on specific face recognition scenarios. *Artificial Intelligence Review*, 54(8):6201–6244, 2021. 1

[52] L. Nie, A. Kumar, and S. Zhan. Periocular recognition using unsupervised convolutional rbm feature learning. In *2014 22nd International Conference on Pattern Recognition*, pages 399–404. IEEE, 2014. 2

[53] U. Park, A. Ross, and A. K. Jain. Periocular biometrics in the visible spectrum: A feasibility study. In *2009 IEEE 3rd international conference on biometrics: theory, applications, and systems*, pages 1–6. IEEE, 2009. 2

[54] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. 2015. 8

[55] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 2, 3, 5

[56] H. Proença and J. C. Neves. Deep-prwis: Periocular recognition without the iris and sclera using deep learning frameworks. *IEEE Transactions on Information Forensics and Security*, 13(4):888–896, 2017. 2

[57] K. B. Raja, N. Damer, R. Ramachandra, F. Boutros, and C. Busch. Cross-spectral periocular recognition by cascaded spectral image transformation. In *2019 IEEE International Conference on Imaging Systems and Techniques, IST 2019, Abu Dhabi, United Arab Emirates, December 9-10, 2019*, pages 1–7. IEEE, 2019. 2

[58] K. B. Raja, R. Raghavendra, and C. Busch. Collaborative representation of deep sparse filtered features for robust verification of smartphone periocular images. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 330–334. IEEE, 2016. 2

[59] K. B. Raja, R. Raghavendra, M. Stokkenes, and C. Busch. Smartphone authentication system using periocular biometrics. In *2014 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–8. IEEE, 2014. 2

[60] A. Rattani, R. Derakhshani, S. K. Saripalle, and V. Gottemukkula. Icip 2016 competition on mobile ocular biometric recognition. In *2016 IEEE international conference on image processing (ICIP)*, pages 320–324. IEEE, 2016. 2

[61] N. Reddy, A. Rattani, and R. Derakhshani. Comparison of deep learning models for biometric-based mobile user authentication. In *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–6. IEEE, 2018. 2

[62] N. Reddy, A. Rattani, and R. Derakhshani. Ocularnet: deep patch-based ocular biometric recognition. In *2018 IEEE international symposium on technologies for homeland security (HST)*, pages 1–6. IEEE, 2018. 2

[63] N. Reddy, A. Rattani, and R. Derakhshani. Robust subject-invariant feature learning for ocular biometrics in visible spectrum. In *2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–9. IEEE, 2019. 2

[64] N. Reddy, A. Rattani, and R. Derakhshani. Generalizable deep features for ocular biometrics. *Image and Vision Computing*, 103:103996, 2020. 2

[65] A. Ross, R. Jillela, J. M. Smereka, V. N. Boddeti, B. V. Kumar, R. Barnard, X. Hu, P. Pauca, and R. Plemmons. Matching highly non-ideal ocular images: An information fusion approach. In *2012 5th IAPR International Conference on Biometrics (ICB)*, pages 446–453. IEEE, 2012. 2

[66] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 8

[67] R. Sharma and A. Ross. Periocular biometrics and its relevance to partially masked faces: A survey. *arXiv preprint arXiv:2203.15203*, 2022. 1, 2

[68] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 8

[69] R. Spolaor, Q. Li, M. Monaro, M. Conti, L. Gamberini, and G. Sartori. Biometric authentication methods on smartphones: A survey. *PsychNology Journal*, 14(2), 2016. 1

[70] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*, 2017. 8

[71] C.-W. Tan and A. Kumar. Human identification from at-a-distance images by simultaneously exploiting iris and periocular features. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, pages 553–556. IEEE, 2012. 2

[72] M. Uzair, A. Mahmood, A. Mian, and C. McDonald. Periocular biometric recognition using image sets. In *2013 IEEE Workshop on Applications of Computer Vision (WACV)*, pages 246–251. IEEE, 2013. 5

[73] S. Van Der Walt, S. C. Colbert, and G. Varoquaux. The numpy array: a structure for efficient numerical computation. *Computing in science & engineering*, 13(2):22–30, 2011. 4

[74] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu. Cosface: Large margin cosine loss for deep face recognition. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 5265–5274. Computer Vision Foundation / IEEE Computer Society, 2018. 3

[75] M. Wang and W. Deng. Deep face recognition: A survey. *Neurocomputing*, 429:215–244, 2021. 3, 5

[76] O. Weng. Neural network quantization for efficient inference: A survey. *arXiv preprint arXiv:2112.06126*, 2021. 3

[77] J. Wu, C. Leng, Y. Wang, Q. Hu, and J. Cheng. Quantized convolutional neural networks for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4820–4828, 2016. 2

[78] J. Xu, M. Cha, J. L. Heyman, S. Venugopalan, R. Abiantun, and M. Savvides. Robust local binary pattern feature sets for periocular biometric identification. In *2010 Fourth IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, pages 1–8. IEEE, 2010. 2

[79] Z. Yao, Z. Dong, Z. Zheng, A. Gholami, J. Yu, E. Tan, L. Wang, Q. Huang, Y. Wang, M. Mahoney, et al. Hawq-v3: Dyadic neural network quantization. In *International Conference on Machine Learning*, pages 11875–11886. PMLR, 2021. 2

[80] L. A. Zanlorensi, R. Laroca, D. R. Lucio, L. R. Santos, A. S. Britto Jr, and D. Menotti. Ufpr-periocular: a periocular

dataset collected by mobile devices in unconstrained scenarios. *arXiv preprint arXiv:2011.12427*, 2020. 1, 2, 3, 4, 5, 6, 7, 8

[81] L. A. Zanlorensi, H. Proença, and D. Menotti. Unconstrained periocular recognition: Using generative deep learning frameworks for attribute normalization. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 1361–1365. IEEE, 2020. 2

[82] D. Zhang, J. Yang, D. Ye, and G. Hua. Lq-nets: Learned quantization for highly accurate and compact deep neural networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 365–382, 2018. 4

[83] P. Zhang, F. Zhao, P. Liu, and M. Li. Efficient lightweight attention network for face recognition. *IEEE Access*, 10:31740–31750, 2022. 1

[84] Q. Zhang, H. Li, Z. Sun, and T. Tan. Deep feature fusion for iris and periocular biometrics on mobile devices. *IEEE Transactions on Information Forensics and Security*, 13(11):2897–2912, 2018. 2

[85] B. Zhuang, C. Shen, M. Tan, L. Liu, and I. Reid. Towards effective low-bitwidth convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7920–7928, 2018. 2

[86] D. Zuras, M. Cowlishaw, A. Aiken, M. Applegate, D. Bailey, S. Bass, D. Bhandarkar, M. Bhat, D. Bindel, S. Boldo, et al. Ieee standard for floating-point arithmetic. *IEEE Std*, 754(2008):1–70, 2008. 4