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

Soft-Biometrics Estimation In the Era of Facial Masks

Fernando Alonso-Fernandez[1], Kevin Hernandez Diaz[2], Silvia Ramis[3],
Francisco J. Perales[4], Josef Bigun[5]

Abstract: We analyze the use of images from face parts to estimate soft-biometrics indicators. Partial face occlusion is common in unconstrained scenarios, and it has become mainstream during the COVID-19 pandemic due to the use of masks. Here, we apply existing pre-trained CNN architectures, proposed in the context of the ImageNet Large Scale Visual Recognition Challenge, to the tasks of gender, age, and ethnicity estimation. Experiments are done with 12007 images from the Labeled Faces in the Wild (LFW) database. We show that such off-the-shelf features can effectively estimate soft-biometrics indicators using only the ocular region. For completeness, we also evaluate images showing only the mouth region. In overall terms, the network providing the best accuracy only suffers accuracy drops of 2-4% when using the ocular region, in comparison to using the entire face. Our approach is also shown to outperform in several tasks two commercial off-the-shelf systems (COTS) that employ the whole face, even if we only use the eye or mouth regions.

Keywords: Soft-Biometrics, Periocular, Gender, Age, Ethnicity.

1 Introduction

Recent research has explored the use of ancillary information, known as soft biometrics, which includes attributes like gender, age, ethnicity, etc. [DER16]. While they may not be sufficiently distinctive to allow accurate recognition, they can be used in a fusion framework to complement the primary system [Go18]. Automated soft-biometrics extraction has other applications as well, such as reducing the search space of subjects in large databases, locating specific individuals based on such semantic attributes, providing age-dependant access control, or customizing advertisements or customer recommendations [DER16].

Face is a natural way to recognize many soft-biometrics indicators. However, in unconstrained conditions, it may be partially occluded, accidentally or intentionally, as for example by the use of masks. Accordingly, we address the challenge of estimating soft-biometrics indicators when only images of face parts are available. This has been suggested in several studies with traditional features such as Local Binary Patterns or Histograms of Oriented Gradients [AFB16]. Here, we leverage the power of Convolutional Neural Networks (CNNs) pre-trained in the context of the ImageNet challenge with more than a million images to classify images into 1000 object categories. Based on [Ng18], the

[1] School of Information Technology, Halmstad University, Sweden, feralo@hh.se
[2] School of Information Technology, Halmstad University, Sweden, kevin.hernandez-diaz@hh.se
[3] Computer Graphics and Vision and AI Group, University of Balearic Islands, Spain, silvia.ramis@uib.es
[4] Computer Graphics and Vision and AI Group, University of Balearic Islands, Spain, paco.perales@uib.es
[5] School of Information Technology, Halmstad University, Sweden, josef.bigun@hh.se

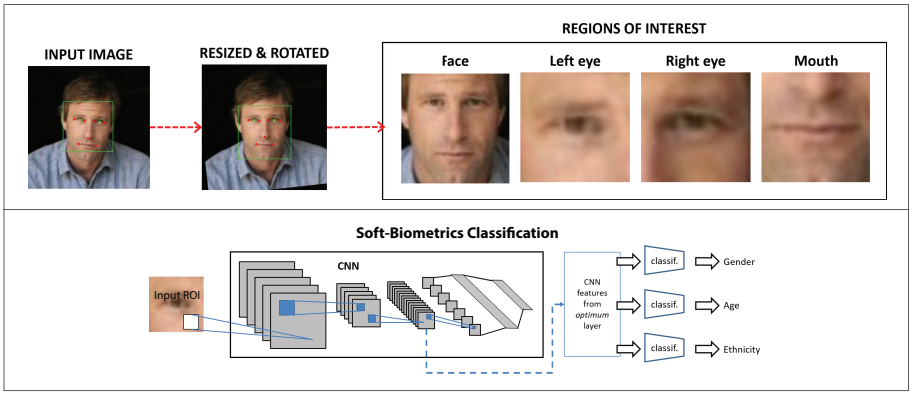

Fig. 1: Top: Extraction of the regions of interest. Bottom: Soft-biometrics classification framework.

authors in [HDAFB18, Al19] investigated the use of these off-the-shelf CNNs for periocular recognition, eliminating the necessity of designing and training new networks. Here, we further investigate their behaviour in soft-biometrics classification. Our experiments show that these off-the-shelf features are capable of measuring soft-biometrics using only the ocular or mouth regions, with negligible accuracy drops or even better performance in comparison to using the whole face. The proposed approach also compares favourably with two commercial off-the-shelf systems (COTS), outperforming them in several tasks.

2 Soft-Biometrics Classification Approach

We extract features from different regions (Figure 1): face, left/right periocular, or mouth. For feature extraction, the following networks are used: AlexNet [KSH12], ResNet50 and ResNet101 [He16], DenseNet201 [Hu17], VGG-Face [PVZ15], and MobileNetv2 [Sa18]. These networks have gained in sophistication and depth, starting from AlexNet (with only 5 convolutional layers), to ResNet (50/101 layers) and DenseNet (201 layers). The latter were made possible thanks to concepts like residual connections [He16] and densely connected architectures [Hu17], with allowed the training of deeper networks. We also employ VGG-Face. Based on the generic VGG16, it is trained to recognize faces, so we believe that it can provide effective recognition in our tasks with data from facial regions. Finally, we use the network MobileNetv2, designed to have a smaller size while keeping accuracy. With these choices, we aim at comparing networks of different depths, and a network trained with faces as well. In using them, images are fed into each CNN. But instead of using the vector from the last layer, we employ as descriptor the intermediate layer identified as giving the best performance in periocular recognition [HDAFB18, Al19]. Since we will employ a similar type of data, we speculate that these layers will be useful for soft-biometrics as well. In particular, we use the layers: 14 (AlexNet), 73 (ResNet50), 165 (ResNet101), 223 (DenseNet201), 25 (VGG-Face) and 121 (MobileNetv2). Classification with each network is then done by training a linear Support Vector Machine (SVM) with the extracted feature vectors [Va95]. The complete procedure is shown in Figure 1 (bottom), whereas Table 1 indicates the size of the feature vector for each network.

Network	Layer	Size
AlexNet	14	43264
ResNet50	73	100352
ResNet101	165	50176
DenseNet201	223	6272
VGG-Face	25	100352

Network	Layer	Region	Size
MobileNetv2	121	-	7840
MobileNetv2	121	face	4763
+ PCA		left eye	4332
		right eye	4327
		mouth	4396

Tab. 1: Size of the feature vector per classification network.

Attribute					
Gender	Male (77.6%)	Female (22.4%)			
Age	Baby (<1%)	Child (<1%)	Youth (12.9%)	Adult (62.9%)	Senior (23.6%)
Ethnicity	White (81.6%)	Black (3.8%)	Asian (5.5%)	Indian (2.4%)	Other (6.7%)

Tab. 2: Statistics of soft-biometrics attributes of the LFW database.

3 Database and Protocol

We use the Labeled Faces in the Wild (LFW) database [Hu07]. It contains images of celebrities from the web with a large range of variations in pose, lightning, expression, etc. In particular, we use 12007 images, for which annotation of face landmarks is available. All images are rotated w.r.t. the axis crossing the eyes, and resized to an eye-to-eye distance of 42 pixels (average of the database). Then, a face image of 109×109 is extracted, together with the two periocular regions (43×43 each), and the mouth (49×49). Images are further resized to the input size of the networks. An example of this procedure is given in Figure 1. To train and evaluate our classification approach, we employ the ground-truth of [Go18]. Table 2 indicates the attributes employed and the statistics of the database. When there are more than two classes, a one-vs-one multi-class approach is used. For every feature and N classes, $N(N-1)/2$ binary SVMs are used. Classification is made based on which class has most number of binary classifications towards it (voting scheme). Evaluation is done with k-fold cross-validation (k=5), with k sets containing the same number of (non-overlapped) people. On each iteration, a set is retained for validation, and the remaining $k-1$ sets are used to train the SVMs. The average accuracy of the k iterations are then reported. The software employed was Matlab r2019a, which contains pre-trained models of all the CNNS, except VGG-Face which is from the Caffe Model Zoo.

4 Results

The performance of our soft-biometrics classification approach is reported in Figures 2-4 for gender, age, and ethnicity respectively. Accuracy is reported for each class (images of the class classified correctly), and for the whole database (images of the database classified correctly). We provide results using as input: *i*) the whole face, *ii*) the left/right eye separately, *iii*) both eyes together (by concatenating feature vectors), and *iv*) the mouth region.

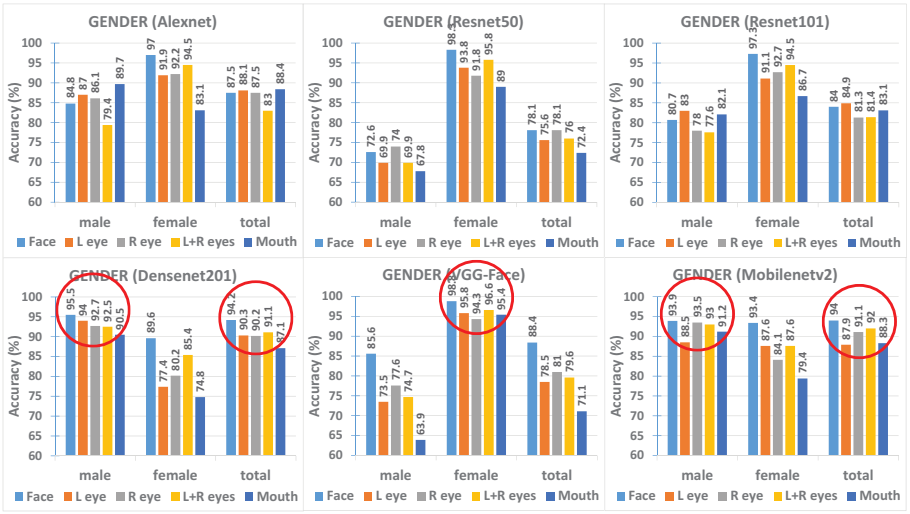

Fig. 2: Accuracy of gender estimation using different facial regions.

The size of some age groups ("baby", "child") is very small, see Table 2, so these groups been merged with the class "youth" into a single class that we call "minors".

The red circles indicate the top results for each class. A quick look reveals that three networks concentrate the top results (with few exceptions): DenseNet201, VGG-Face and MobileNetv2. The best networks overall ('total' accuracy) are DenseNet201 and MobileNetv2. This is interesting, since DenseNet201 is the deepest network employed, while MobileNetv2 is a lighter network designed to have much less depth and parameters. With DenseNet201, gender is estimated with an accuracy of 87.1-94.2% (depending on the image region), while age is estimated with 57.6-62%, and ethnicity with 76.8-81.6%. With MobileNetv2, gender is estimated with an accuracy of 87.9-94%, age with 55.5-63.8%, and ethnicity with 70.3-80.5%. It is also relevant that VGG-face does not systematically outperform the other networks, even if it is trained with facial data. DenseNet201 and MobileNetv2 are also the best network with the classes having more samples (Table 2): gender-male, age-adult, and ethnicity-white classes. On the other hand, VGG-Face wins with the classes that are less represented; a downside though is that its performance with the biggest classes is poor. The latter is also seen in the ResNet variants.

Interestingly, the feature vectors of DenseNet201 and MobileNetv2 are the smallest among those employed (Table 1). Therefore, a bigger feature vector does not correlate with a better performance, but the opposite. Also, MobileNetv2 stands out as a very balanced network, with top results with the biggest classes, and also relatively good performance with the others (with very few exceptions like ethnicity-indian or ethnicity-other classes, whose performance is very poor with any network). Given that the networks employed have not been specifically trained for soft-biometrics, and to eliminate feature redundancy, we carry out dimensionality reduction by Principal Component Analysis (PCA) [Jo02]. We retain the elements with 99% of the variance, with the PCA basis learnt using images from

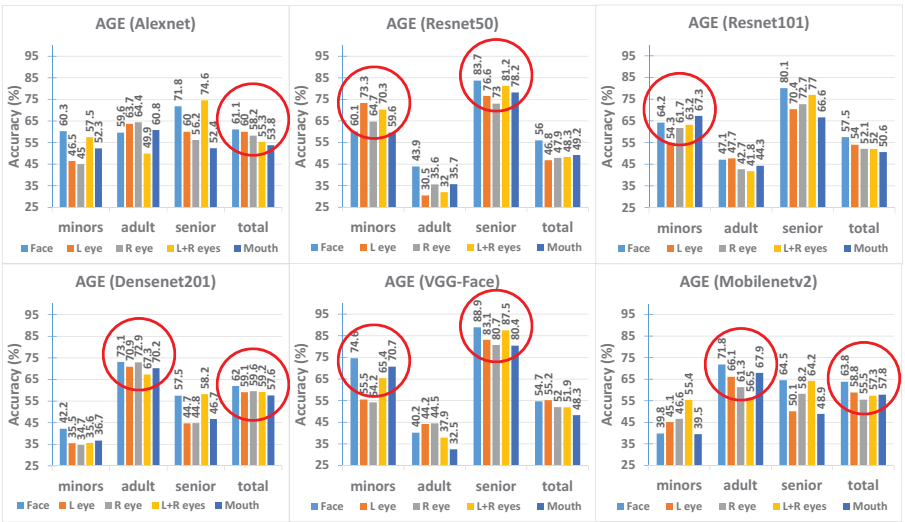

Fig. 3: Accuracy of age estimation using different facial regions.

the training set on each validation iteration. In our experiments, we have observed that PCA provides further performance improvement with DenseNet201 and MobileNetv2 in the majority of classes. On the other hand, results with the other networks are not consistent, showing improvement with some classes, while decreasing substantially in others. Due to space, we only show results of MobileNetv2 (Table 3). Also, Table 1 (right) gives the average number of retained coefficients for the different regions.

As it can be observed in Table 3, in overall terms ('total' columns), PCA provides an extra improvement. The performance of the biggest classes (gender-male, age-adult, and ethnicity-white) is better, and improvements happen as well with several less-represented classes. It happens though that some small classes worsen after PCA, e.g. age-senior, ethnicity-black, or ethnicity-other. Regarding the use of different facial regions, it can be observed that using only the periocular or mouth regions is not necessarily worse than using the whole face. This is not only seen with MobileNetv2 (Table 3), but with other networks as well (Figures 2-4). When estimating gender with MobileNetv2, the best accuracy is obtained with the whole face (95.8%). With a combination of both eyes, accuracy is just 2.4% below (93.4%), and with only one eye, it drops a further 0.8% only (92.6%). Accuracy with only the mouth region is also comparably good (90.5%), although its accuracy with the gender-female class is much worse than the other facial regions. In a similar vein, the whole face provides the best overall performance in age (64.5%) and ethnicity (83.3%) estimation, and the use of facial parts results in a small accuracy drop only. Age with only the mouth is estimated with an accuracy of 59.6%, which goes up to 60% when both eyes are used, and even better with the left eye only (60.2%). Similarly, ethnicity with both eyes or the mouth is estimated with an accuracy of 81.3%/81.5%, and even better with the right eye only (82.9%). It is worth noting as well that combining both eyes does not necessarily

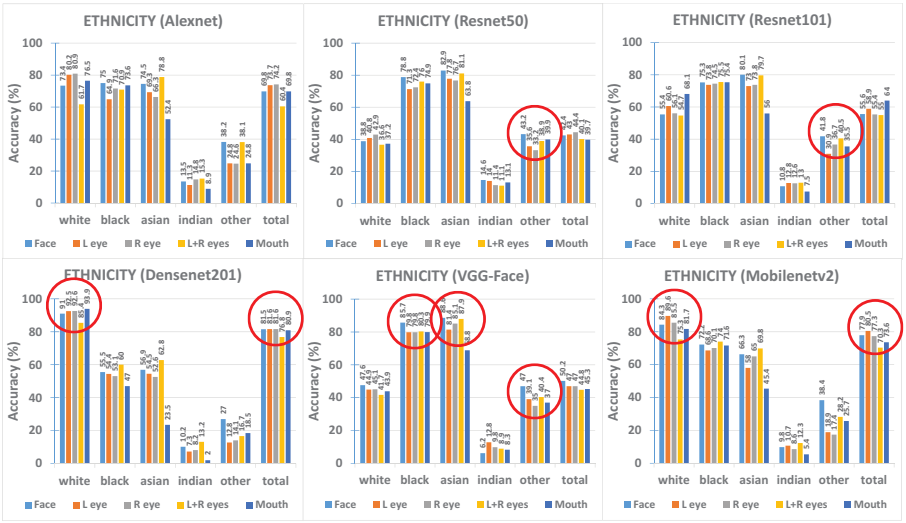

Fig. 4: Accuracy of ethnicity estimation using different facial regions.

produces better accuracy, in comparison to using one eye only. In addition, uncorrelated areas such as the eye or mouth provides a relatively similar performance.

We also provide (Table 4) the results of two COTS systems, Face++[3] and Microsoft Cognitive Toolkit[4], given in [Go18]. These systems estimate soft-biometrics attributes based on deep learning architectures. The results in Table 4 have been obtained using images of the whole face. Note that not all the classes employed in this paper are provided. Ethnicity is only given by Face++, giving only the classes white (caucasian), black and asian. Regarding age, the results in [Go18] are separated by the five age groups of Table 2. By comparing Tables 3 and 4, we observe that the performance of our suggested framework using MobileNetv2 outperforms the gender estimation of these COTS systems. Regarding age estimation, the COTS systems are better for age classes involving minors (which represent only about 13% of the data), but they show poorer performance with age-adult or age-senior groups. Regarding ethnicity, our approach outperforms the COTS systems for white and black classes. It is also worth noting that in the classes where our approach outperforms the COTS systems, the superiority is observed as well if we only employ the eye or mouth regions.

5 Conclusions

We suggests the use of off-the-shelf CNN architectures, pre-trained in the context of the ImageNet Large Scale Visual Recognition Challenge, for the task of soft-biometrics classification with facial images. More importantly, giving the current context where face en-

[3] https://www.faceplusplus.com/
[4] https://www.microsoft.com/cognitive-services/

	GENDER			AGE			
	male	female	total	minors	adult	senior	total
face	93.9	**93.4**	94	39.8	71.8	**64.5**	63.8
face + PCA	**97.6**	90.1	**95.8**	45	**75.6**	53.1	**64.5**
left eye	88.5	87.6	87.9	45.1	66.1	50.1	58.8
left eye + PCA	95.8	80.8	92.5	46.1	69.4	46.7	60.2
right eye	93.5	84.1	91.1	46.6	61.3	58.2	55.5
right eye + PCA	94.6	85.3	92.6	39.9	70.5	46	57.8
both eyes	93	87.6	92	**55.4**	56.5	64.2	57.3
both eyes + PCA	94.6	89.7	93.4	45.9	72.8	45.9	60
mouth	91.2	79.4	88.3	39.5	67.9	48.9	57.8
mouth + PCA	95.2	74.6	90.5	41.9	71.9	44	59.6

	ETHNICITY					
	white	black	asian	indian	other	total
face	84.3	72.2	66.3	9.8	**38.4**	77.9
face + PCA	91.1	**78.1**	66.8	7.7	32	**83.3**
left eye	89.6	68.6	58	10.7	18.9	80.5
left eye + PCA	90.2	68.5	60.4	**13.4**	23.5	81.4
right eye	85.5	70.1	65	8.6	17.4	77.3
right eye + PCA	**93.2**	63.8	55.3	8.2	17	82.9
both eyes	75.3	74	69.8	12.3	28.2	70.3
both eyes + PCA	88.8	70.8	**76.8**	11.1	25.3	81.3
mouth	81.7	71.6	45.4	5.4	25.7	73.6
mouth + PCA	92	69.5	38.8	4.3	21.7	81.5

Tab. 3: MobileNetv2 network: Accuracy of soft-biometrics estimation with and without PCA reduction using different facial regions. For each region, the best accuracy (between using/not using PCA) is highlighted with a grey background. The best overall accuracy of each class is marked in bold.

gines are forced to work with images of people wearing masks, we evaluate the feasibility of using partial images containing only the ocular or mouth regions (Figure 1). In this paper, we test popular generic architectures, with features extracted from intermediate layers identified in previous studies as providing good person recognition with ocular images. Prediction is then done with SVM classifiers. They are evaluated with 12007 annotated images of the LFW database [Hu07, Go18]. Our results indicate the possibility of performing soft-biometrics classification using images containing only the ocular or mouth regions, without a significant drop in performance in comparison to using the entire face. An overall accuracy of 95.8/64.5/83% in gender/age/ethnicity estimation is obtained with images of the entire face using the MobileNetv2 architecture. Using only images of one eye, the best accuracy in these tasks is 92.6/60.2/82.9% respectively, and using images of the mouth area, we obtain an accuracy of 90.5/59.6/81.5%. The proposed approach also compares well with two COTS systems by Face++ and Microsoft, outperforming them in the gender estimation task, and in several classes of the age and ethnicity tasks.

A limitation to overcome is the class imbalance of our database. Also, the CNN layers employed were optimized for periocular recognition, but it might be that the best layer for soft-biometrics or for the entire face or the mouth region is different. We are also looking

GENDER

Face++			Microsoft		
male	female	total	male	female	total
92.2	87.5	91.1	93.5	91.1	92.9

AGE

Face++					
baby	child	youth	adult	senior	total
100	53.2	81.4	32	33.4	38.8

ETHNICITY

Face++					
white	black	asian	indian	other	total
88.3	76.2	83.1	-	-	87.4

AGE

Microsoft					
baby	child	youth	adult	senior	total
100	45.2	92.2	52.5	59.6	59.3

Tab. 4: Performance of Face++ and Microsoft COTS [Go18].

into fine-tuning CNN architectures to do the classification directly, thanks to newer annotated repositories [MFV19]. We also foresee that improvements can be obtained by joint estimation of soft-biometrics indicators by sharing weights between different networks, since a single facial feature carry information about different soft-biometrics [DER16].

Acknowledgment

This work was partly done while F. A.-F. was a visiting researcher at the University of Balearic Islands (UIB), funded by the visiting lecturers program of the UIB. Authors F. A.-F., K. H.-D. and J. B. also would like to thank the Swedish Research Council for funding their research. Part of the computations were enabled by resources provided by the Swedish National Infrastructure for Computing (SNIC) at NSC Linköping. We also gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan V GPU used for this research.

References

[AFB16] Alonso-Fernandez, F.; Bigun, J.: A survey on periocular biometrics research. Pattern Recognition Letters, 82:92–105, 2016.

[Al19] Alonso-Fernandez, F.; Raja, K. B.; Raghavendra, R:; Busch, C.; Bigün, J.; Vera-Rodríguez, R.; Fiérrez, J.: Cross-Sensor Periocular Biometrics: A Comparative Benchmark including Smartphone Authentication. CoRR, abs/1902.08123, 2019.

[DER16] Dantcheva, A.; Elia, P.; Ross, A.: What Else Does Your Biometric Data Reveal? A Survey on Soft Biometrics. IEEE TIFS, 11(3):441–467, 2016.

[Go18] Gonzalez-Sosa, E.; Fierrez, J.; Vera-Rodriguez, R.; Alonso-Fernandez, F.: Facial Soft Biometrics for Recognition in the Wild: Recent Works, Annotation and COTS Evaluation. IEEE TIFS, 13(8):2001–2014, August 2018.

[HDAFB18] Hernandez-Diaz, K.; Alonso-Fernandez, F.; Bigun, J.: Periocular Recognition Using CNN Features Off-the-Shelf. In: Proc BIOSIG. pp. 1–5, Sep. 2018.

[He16] He, K.; Zhang, X.; Ren, S.; Sun, J.: Deep Residual Learning for Image Recognition. In: Proc CVPR. pp. 770–778, June 2016.

[Hu07] Huang, G. B. et al.: Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. TR 07-49, Univ of Massachusetts, Oct 2007.

[Hu17] Huang, G. et al.: Densely Connected Convolutional Networks. Proc CVPR, 2017.

[Jo02] Jolliffe, Ian: Principal component analysis. Springer Verlag, New York, 2002.

[KSH12] Krizhevsky, A. et al.: ImageNet Classification with Deep Convolutional Neural Networks. Advances in Neural Informat Proc Systems 25. Curran Associates, Inc., 2012.

[MFV19] Morales, A.; Fierrez, J.; Vera-Rodríguez, R.: SensitiveNets: Learning Agnostic Representations with Application to Face Recognition. CoRR, abs/1902.00334, 2019.

[Ng18] Nguyen, K.; Fookes, C.; Ross, A.; Sridharan, S.: Iris Recognition With Off-the-Shelf CNN Features: A Deep Learning Perspective. IEEE Access, 6:18848–18855, 2018.

[PVZ15] Parkhi, O. M.; Vedaldi, A.; Zisserman, A.: Deep Face Recognition. Proc BMVC, 2015.

[Sa18] Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.: MobileNetV2: Inverted Residuals and Linear Bottlenecks. In Proc CVPR. pp. 4510–4520, 2018.

[Va95] Vapnik, V. N.: The Nature of Statistical Learning Theory. Springer-Verlag, 1995.