FaceGenderID: Exploiting Gender Information in DCNNs Face Recognition Systems

Ruben Vera-Rodriguez, Marta Blazquez, Aythami Morales
Biometrics and Data Pattern Analytics - BiDA Lab
Universidad Autonoma de Madrid, Madrid, Spain
{ruben.vera, aythami.morales}@uam.es, marta.blazquez.cortes@gmail.com

Ester Gonzalez-Sosa
Nokia Bell Labs, Madrid, Spain
ester.gonzalez@nokia-bell-labs.com

João C. Neves, Hugo Proença
Instituto de Telecomunicações,
University of Beira Interior, Portugal
{hugomcp, jcneves}@di.ubi.pt

Abstract

This paper addresses the effect of gender as a covariate in face verification systems. Even though pre-trained models based on Deep Convolutional Neural Networks (DCNNs), such as VGG-Face or ResNet-50, achieve very high performance, they are trained on very large datasets comprising millions of images, which have biases regarding demographic aspects like the gender and the ethnicity among others. In this work, we first analyse the separate performance of these state-of-the-art models for males and females. We observe a gap between face verification performances obtained by both gender classes. These results suggest that features obtained by biased models are affected by the gender covariate. We propose a gender-dependent training approach to improve the feature representation for both genders, and develop both: i) gender specific DCNNs models, and ii) a gender balanced DCNNs model. Our results show significant and consistent improvements in face verification performance for both genders, individually and in general with our proposed approach. Finally, we announce the availability (at GitHub[1]) of the FaceGenderID DCNNs models proposed in this work, which can support further experiments on this topic.

1. Introduction

Over the last years, face recognition is by far the biometric technology that has attracted the most attention from research and industry. Face recognition technology is nowadays used in many applications, from smart phone access to covert analysis of crowds using CCTV cameras [1]. Since 2014, breakthroughs in face recognition technology have been improving the recognition performance, now (at least) in line with humans thanks to the usage of deep convolutional neural networks (DCNNs) [2, 3, 4]. Also, the public availability of competitive DCNNs models, such as, VGG-Face [5] or ResNet-50 [6, 7], trained on databases comprising millions of face images [5, 7, 8, 9, 10] has boosted the research in this area.

Even though these DCNNs models already achieve really impressive high face recognition performance, some weaknesses remain. This is the case of the bias of face recognition performance, regarding covariates such as the gender, the ethnicity, or the age [11, 12]. As an example, a recent study [13] showed how commercial face recognition systems achieve better performance for lighter individuals and males and worse for darker females. Authors argue that the reason for these bias in face recognition technology origins from the datasets used for their training, which even if they are very large in number of images, they have a higher number of face images from Caucasian ethnicity and males. This is for example the case of VGGFace2 database [7], with 3.31 million images from which 74% are from Caucasians and 60% are males. Also, several published studies have confirmed the differences in recognition performance for different population demographic groups [12, 14, 15, 16].

As a response to this problem, new face datasets are appearing in the research community. Two examples are the DiveFace database[2] [17], which is an annotated subset from Megaface and contains balanced sets of face images regarding gender and three ethnic groups. Also, the DiF database [18] contains one million faces with a large

[1] https://github.com/BiDAlab/FaceGenderID

[2] Available at: https://github.com/BiDAlab/DiveFace

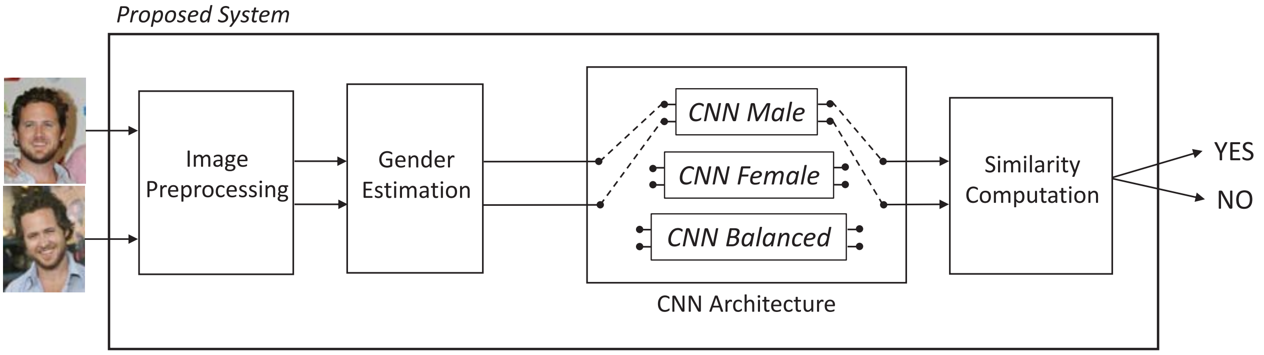


Figure 1. Architecture of our proposed approach for gender specific DCNNs models for face verification.

diversity and available labels for age and gender. Other face databases, with annotations for facial attributes can be found in [19, 20, 21].

On the other way, some works have reported that soft biometrics (such as the gender, ethnicity or age) can improve the performance of face recognition systems, such as [15, 22, 23]; and more recently in [20], fusing at the score level matching scores from a DCNNs face recognition system and scores from a system based on a bag of soft biometrics. Similarly, several other works explore facial soft biometrics [24, 25, 26, 27, 28] and body based soft biometrics to improve biometric systems [29, 30, 31]. In an original way, a recent work [17] proposed a method to remove the gender and ethnicity information from DCNN-based feature embbeddings, without dropping the face recognition performance. This method is claimed to avoid algorithmic discrimination and to comply with data protection regulations.

In this work, we propose to perform a gender-dependent training approach to improve the face verification performance of two very popular DCNNs models: VGG-Face and ResNet-50. In particular, we propose to use triplet loss learning algorithm [3, 5] to train both: *i*) gender specific DCNNs models, and *ii)* a gender balanced CNN model in order to develop a global face recognition system, which can make use of these three DCNNs models depending on the gender estimation label from the input images. Fig. 1 shows a diagram of the system architecture.

The remainder of the paper is organized as follows. Section 2 describes the proposed method. Section 3 reports the statistics of the database used in the our experiments. Section 4 provides the experiments carried out to validate the proposed approach. Finally, Section 5 draws the final conclusions and points out some lines for future work.

2. Proposed Method

The method proposed in this paper is focused on training a gender-dependent face representation to improve the performance of face verification systems. It is based on fine-tuning and in learning additional feature embeddings, using triplet loss from two well known pre-trained face recognition DCNNs models: VGG-Face and ResNet-50.

Fig. 1 provides a comprehensive perspective of the architecture of our system, to exploit gender information for improved face verification. At first, faces are detected from the input image, using the model provided by [32]. The face bounding box is then enlarged by a factor of 0.3 to include the whole head. Next, we use a gender estimator module to automatically distinguish males from females allowing to decide the network for analyzing the verification pair. If both face images have the same gender label, they both enter the gender specific DCNNs model. Otherwise, if the gender label is different (one male and one female), they enter the gender balanced DCNNs model. Finally, the Euclidean distance between the two feature embeddings is obtained to decide whether the pair sample regards the same person or not.

Apart from the strategy just described, it is also important to mention that the use of an alternative strategy has also been considered. This alternative strategy consists in just using the balanced DCNNs model, avoiding thus the use of a gender estimation module whose eventual errors compromise the further processing steps and also simplifying the overall architecture.

2.1. Pre-trained Face DCNNs Models

We use two popular DCNNs pre-trained models, which have recently achieved some of the best state-of-the-art performance in face recognition tasks: VGG-Face, proposed in [5] and ResNet-50, proposed in [6].

VGG-Face is a DCNNs with a VGG-16 architecture trained from scratch with a dataset that contains more than 2.6 million images of 2622 celebrities. The architecture comprises 8 blocks of convolutional layers followed by activation layers like ReLU or maxpooling, and 3 blocks of fully connected layers with ReLU activations. VGG-Face has an overall of 145,002,878 parameters split in 16 trainable layers (convolutional and fully connected layers).

ResNet-50 is another DCNNs based on a residual neuronal network architecture. The key of this network is to insert shortcut connections among blocks, which turn the network into a residual network version. ResNet-50 has a total of 41,192,951 parameters split in 34 residual layers for training. ResNet-50 was adapted to face recognition in [7], where this network was trained from scratch with VGGFace2 dataset. This is the model used in this study.

2.2. Gender Estimation

The gender estimator module used in this work (see Fig. 1) is built by fine-tuning the pre-trained model ResNet-50, since it attains higher recognition rates when compared to other similar models (e.g., VGG-Face). For this reason, the weights from the pre-trained model are frozen except for the last fully connected layer, which is then retrained with gender balanced data from VGGFace2 dataset as described in Sect. 4.1. Finally, a feed-forward layer with a softmax activation is added as final layer, in order to provide a binary output score (male and female).

2.3. Gender Specific Models for Face Recognition using Triplet Loss

The gender specific models for face verification (see CNN architecture in Fig. 1) are developed using feature embeddings from the last fully connected layer of each DCNNs pre-trained models (VGG-Face and ResNet-50). Then, a triplet loss learning algorithm [3, 5] is used to infer a new feature representation with better person discrimination within subjects of the same gender. Triplet-loss can be used as a domain adaptation method. Our hypothesis is that gender-dependent domains can outperform a gender agnostic domain. Therefore, we train a specific model for males, another for females, and a final one using gender balanced data as specified in Sect. 4.1. In general, we assume that each image is represented by an embedding descriptor $\mathbf{x} \in \mathrm{R}^d$ obtained by a pre-trained model. A triplet is composed by three different images from two different classes: Anchor (**A**) and Positive (**P**) are different face images from the same subject, and Negative (**N**) is an image from a different subject. We form a list of triplets **T** that satisfies the following condition:

$$\left\|\mathbf{x}_{\mathbf{A}}^{i} - \mathbf{x}_{\mathbf{P}}^{i}\right\|^2 - \left\|\mathbf{x}_{\mathbf{A}}^{i} - \mathbf{x}_{\mathbf{N}}^{i}\right\|^2 > \alpha \qquad (1)$$

where i is the index of the triplet, $\|\cdot\|$ is the Euclidean distance and α is a threshold value. In words, we select difficult triplets where the inter-class distance is smaller than the intra-class distance. As Fig. 2 shows, we pass the original feature embeddings through the triplet loss model, yielding a feature embedding for which we have set its dimension to be of M = 1,024 (smaller than the input, which is of 4,096 for the case of VGG-Face and of 2,048 for ResNet-50).

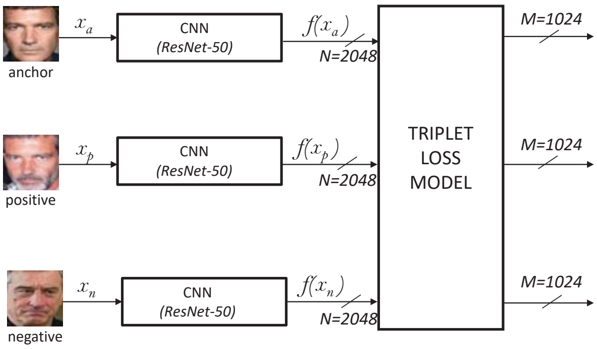


Figure 2. Triplet loss architecture for the case of using ResNet-50 DCNN pre-trained model. A similar architecture is used for the case of VGG-Face model, where the input dimension to the triplet loss model is N = 4,096, instead.

With this process, we would have trained the three DCNNs models of the proposed system, as shown in Fig. 1. One DCNNs model would be used to compare male images (CNN Male), another one to compare female images (CNN Female), and the last model that would be trained using all triplets generated for the previous two models, and therefore having gender balanced data (CNN Balanced). This last model would be used to compare images with different gender label (male/female, and therefore would be very likely to observe only impostor comparisons, with exception of failures in the gender estimation module). The system proposed could be tuned to find a specific threshold to minimize the verification error for each of the three possible cases (male, female and mixed gender comparisons). Otherwise, the matching scores from the three systems could be used from the training data to find a global threshold. Another option we consider is that all face image comparisons would go through the gender balanced DCNNs model, avoiding the need for a gender estimation module. All these possibilities are discussed in the experimental work (Sect. 4).

3. Database

The database used in the experimental work of this paper is VGGFace2 [7]. This is a large-scale face dataset, which contains 3.31 million images of 9,131 subjects, with an average of 362.6 images per subject (varying between 80 and 843 images for each identity). There is a specific set of data from 500 subjects to be used for evaluation purposes. Images were downloaded from Google Image Search and have large variations in pose, age, illumination and ethnicity. The dataset is not balanced regarding the gender, having 59.3% of male subjects and 40.3% of female subjects. Regarding the ethnicity the differences are much greater, having the great majority of 72.2% of Caucasian subjects, 15.8% of African, 6.0% of Asian and 4.0% of Indian.

4. Experiments

4.1. Experimental Protocol

The experimental protocol used in this paper was designed to: *i*) train an automatic face gender estimator using a gender balanced dataset (Experiment 1); *ii*) analyse the effect of using pre-trained DCNNs models originally trained with gender unbalanced datasets, over gender specific datasets (Experiment 2); and *iii*) train gender specific systems and a gender balanced system for face verification (Experiment 3), in order to try to improve the overall system performance by minimizing the gender covariate.

The first experiment is based on training an automatic face gender estimator. For this, the ResNet-50 pre-trained model is retrained to give two outputs: male or female, for an input face image. In this case VGGFace2 dataset is divided into Training set with 200K images (100K images per gender), Validation set with 30K images (15K per gender) and Evaluation set with 132K images (62K per gender).

The second experiment analyses the face verification performance per gender for both VGG-Face and ResNet-50 pre-trained models as baseline. As this is just an evaluation experiment, 10K face image pairs are randomly selected per gender (5K genuine pairs and 5K impostor pairs) from the original evaluation set of VGGFace2 dataset experimental protocol, comprised of images from 500 subjects.

The third Experiment is based on training: *a*) specific DCNNs models per gender, and *b*) a DCNNs model with a gender balanced set of images. For the first case (Exp. 3(a)), i.e., training a specific DCNNs model for each gender, VGGFace2 set is divided into a Training set comprised 240 images for 2500 subjects of each gender (600K images per gender) and a Validation set comprised of 60 images for the same 2500 subjects of each gender (150K images per gender). For training the triplet loss learning algorithm, triplets of images (anchor, positive and negative) are formed from the data of the Training set. As the amount of data for training was really large, we limited each subject not to be included in more than 30 triplets. The triplets generated were ranked regarding the α threshold in Equation (1). The 1000 triplets with highest α value were discarded as there were cases of mistakes in the labels or were really hard cases. Then, the following 75K triplets in terms of α were selected for each gender separately. Finally, the Validation data used was comprised of 10K image pairs per gender (5K genuine pairs and 5K impostor pairs), and the evaluation set was exactly the same one used in Experiment 2, i.e., a set of 10K image pairs per gender from the Evaluation set. For the second case (Exp. 3(b)), i.e., training a DCNNs model with a gender balanced set of images, again the triplet loss learning algorithm was trained in this case using 150K triplets (formed by using both sets of 75K triplets for each gender from the previous Exp. 3(a)). The Validation and Evaluation data used for Exp. 3(b) where the same as per Exp. 3(a).

Finally, a final evaluation of the different configurations, which are: the baseline systems, the balanced gender models, and the gender specific models both using manual and automatic annotations of the gender were evaluated using a total of 25K pairs of images from the Evaluation set, comprising 10K genuine pairs (5K genuine pairs for each gender), 10K impostor pairs (5K impostor pairs for each gender), the same one used in Exp. 2 and 3, and an additional set of 5K mixed gender impostor pairs (that would go through the gender balanced model). This evaluation set is a subset extracted from the original evaluation set of VGGFace2 dataset experimental protocol, comprised of images from 500 subjects.

4.2. Experimental Results

4.2.1 Face Gender Estimator

The gender estimator was trained as specified in Sect. 4.1 using 70 epochs. Gender estimation accuracy was maximized on the Validation set at epoch 27. This model was then applied to the Evaluation data obtaining a final accuracy of 96.8%. These results could be further improved using a more advanced gender estimator system, but this was not the major purpose of this work. In the following sections, the results of using this gender estimator are compared to the case of having the gender labels through manual annotation.

4.2.2 Baseline Face Verification Systems

This experiment shows the results for the analysis carried out using both VGG-Face and ResNet-50 pre-trained DCNNs models for each gender independently. Fig. 3 shows the Receiver Operating Characteristic (ROC) curves obtained, giving the area under the curve (AUC) values for each pre-trained model and gender. As can be seen, in general ResNet-50 model achieves better results compared to VGG-Face, around 4.5% of higher AUC value in average. Regarding the gender, in both DCNNs models, male achieves higher performance compared to female, 0.6% AUC higher for the case of ResNet-50, and 0.8% AUC higher for the case of VGG-Face. These results confirm the findings of previous works, being able to see the bias of the face verification results regarding the gender. Some reasons for this can be the larger amounts of male images compared to female images used to train the DCNNs models, but also the higher intra-class variability present in females due to the higher usage of make up, different hair styles or accessories among other factors.

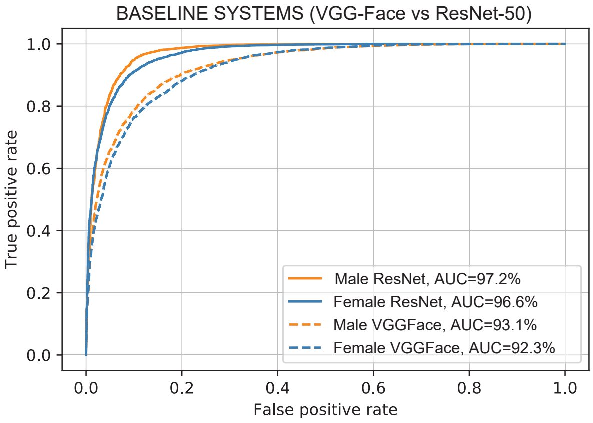


Figure 3. ROC curves observed for the baseline systems (VGG-Face and ResNet-50) for both male and female genders independently.

VGG-Face	**TAR @ FAR**		**AUC**
	0.01	**0.1**	
Baseline	0.402	0.815	94.0
Gender Balanced	0.393	0.828	94.4
Proposed Manual	0.480	0.860	95.2
Proposed Auto	0.439	0.848	94.9

Table 1. Performance comparison between the four DCNNs systems proposed in this paper, using the VGG-Face architecture. TAR values for FAR=0.01 and FAR=0.1, and AUC (in %).

ResNet-50	**TAR @ FAR**		**AUC**
	0.01	**0.1**	
Baseline	0.518	0.946	97.4
Gender Balanced	0.566	0.942	97.5
Proposed Manual	0.600	0.950	97.8
Proposed Auto	0.589	0.948	97.7

Table 2. Performance comparison between the four DCNNs systems proposed in this paper, using the ResNet-50 architecture. TAR values for FAR=0.01 and FAR=0.1, and AUC (in %).

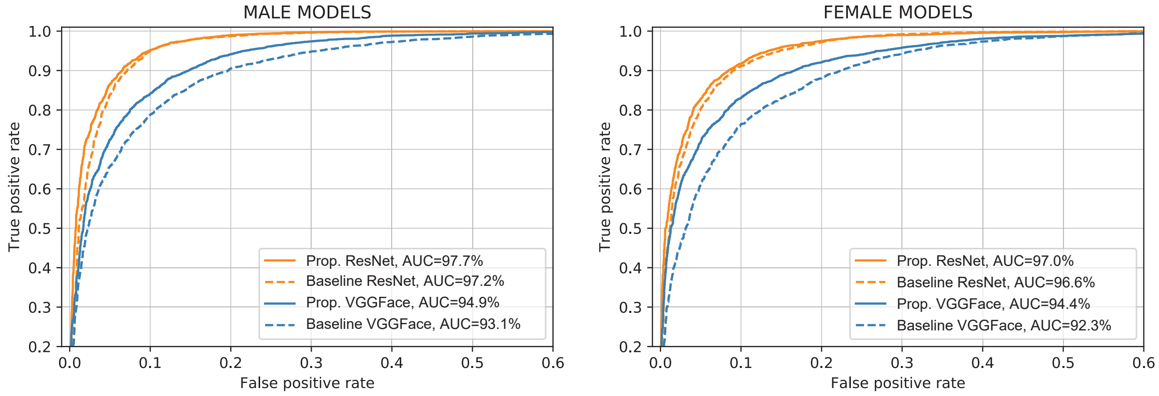
4.2.3 Proposed Face Verification Systems

This section shows the results of our proposed face verification systems, which train a gender-dependent face representation to improve the performance for each gender separately, and therefore the overall system performance.

Fig. 4 shows the ROC curves obtained for both baseline systems and our proposed gender specific systems (only the male and female specific DCNNs models, not the gender balanced one). Fig. 4(a) shows the results for the male gender, obtaining improvements of AUC for both DCNNs models used (VGG-Face and ResNet-50). For ResNet-50, our proposed “CNN Male” system achieves an AUC of 97.7%, 0.5% AUC higher compared to the baseline system. For VGG-Face, our proposed “CNN Male” system achieves an AUC of 94.9%, 1.8% AUC higher compared to the baseline system, a very significant improvement of performance. Fig. 4(b) shows the results for the female gender, obtaining also improvements of AUC for both DCNNs models used. For ResNet-50, our proposed “CNN Female” system achieves an AUC of 97.0%, 0.4% AUC higher compared to the baseline system. For VGG-Face, our proposed “CNN Female” system achieves an AUC of 94.4%, 2.1% AUC higher compared to the baseline system, also a very significant improvement in performance. Even though the system performance for each gender independently improves with our proposed DCNNs gender specific models, a better performance is still observed for males. Accordingly, further work has to be carried out to try to reduce this bias, for which - probably - the best solution would be to train from scratch gender specific DCNNs models.

Our next experiment describes the performance of the final global system, with a clearly defined task: given a pair of input images, output a matching score and a genuine/impostor decision. First, thresholds that maximize the TAR at FAR=0.1% were analyzed for both male and female DCNNs models, for the results shown in Fig. 4, achieving very similar threshold values for both cases. Therefore, the system proposed is based on obtaining a matching score coming from computing the Euclidean distance of a pair of feature embbeddings from one of the three DCNNs options (CNN Male, CNN Female and CNN Balanced shown in Fig. 1. Then, based on these matching scores a global evaluation was performed as described in Sect. 4.1. Tables 1 and 2 show the TAR results at two FAR values (0.01 and 0.1) and also the AUC value for the system based on VGG-Face and ResNet-50 DCNNs architectures respectively. Each table shows the results for four different systems considered in order to carry out a fair comparative analysis: *i)* using the baseline DCNNs pre-trained model, *ii)* using the gender balanced DCNNs model for all face image comparisons, and *iii)* using the proposed system based on three possible DCNNs models (male, female and gender balanced) depending on the gender labels of the faces to be compared. For this last case, we provide results using the original gender labels manually annotated and also using the output from our gender estimator, which would be the case to be deployed in a real application.

Let us analyze first the results for the VGG-Face architecture, shown in Table 1. Based on both the TAR values and the AUC, the best system performance is achieved by our proposed gender specific face verification system using manual gender labels (95.2% AUC compared to the 94.0% AUC for the baseline system), followed by the same system but using automatic gender labels (94.9% AUC). For ResNet-50 architecture results are shown in Table 2, giving similar trends. Best performance is achieved by our proposed gender specific face verification system using manual



(a) Male Gender

(b) Female Gender

Figure 4. Comparison between the ROC curves of the VGG-Face and ResNet-50 (baseline) and the proposed systems for (a) male and (b) female genders.

gender labels (97.8% AUC compared to the 97.4% AUC for the baseline system), followed by the same system but using automatic gender labels (97.7% AUC). Overall performance improvement is smaller for the case of ResNet-50 architecture, but this baseline system already achieves a very competitive performance.

As a conclusion, our proposed FaceGenderID system based on gender specific DCNNs models obtains improvements of performance compared to the baseline DCNNs architectures on an general evaluation. Also, this system achieves better performance than a gender balanced DCNNs model. These FaceGenderID DCNNs models are freely available for research purposes at GitHub[3].

5. Conclusions

The work described in this paper is rooted in an extensively reported observation: the remarkable performance of state-of-the-art face recognition systems is still conditioned by some data covariates, such as the subjects gender, ethnicity or age. In particular, we addressed the effect of gender in recognition accuracy, observing a consistently better performance for males than for females. According to this, we used a triplet loss learning algorithm to exploit the gender information by inferring: *i*) gender specific DCNNs, and *ii*) gender balanced DCNNs. This procedure was observed to reduce the effect of the gender as a recognition covariate.

As future work, we will focus on training the whole proposed DCNNs architecture from scratch, as probably just training the last triplet loss layer is not enough to fully remove the gender bias of the datasets used for training the original pre-trained DCNNs models used.

Acknowledgments

This work has been supported by projects: BIBECA (TEC2018 MINECO), Bio-Guard (Ayudas Fundación BBVA a Equipos de Investigación Científica 2017). We also acknowledge the support of NVIDIA Corporation, with the donation of one Titan X GPU to UBI.

References

[1] H. Proenca, M. Nixon, M. Nappi, E. Ghaleb, G. Özbulak, H. Gao, H. Ekenel, K. Grm, V. Struc, H. Shi, X. Zhu, S. Liao, Z. Lei, S. Li, W. Gutfeter, A. Pacut, J. Brogan, W. Scheirer, E. Gonzalez-Sosa, R. Vera-Rodriguez, J. Fierrez, J. Ortega-Garcia, D. Riccio, and L. Maio. Trends and controversies. *IEEE Intelligent Systems*, 33(3):41–67, 2018. 1

[2] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1701–1708, 2014. 1

[3] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, 2015. 1, 2, 3

[4] R. Ranjan, S. Sankaranarayanan, A. Bansal, N. Bodla, J. Chen, V. M. Patel, C. D. Castillo, and R. Chellappa. Deep learning for understanding faces: Machines may be just as good, or better, than humans. *IEEE Signal Processing Magazine*, 35(1):66–83, 2018. 1

[5] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *Proceedings of the British Machine Vision Conference*, 2015. 1, 2, 3

[6] 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 (CVPR)*, pages 770–778, 2016. 1, 2

[3] https://github.com/BiDAlab/FaceGenderID

[7] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *Proc. International Conf. on Automatic Face and Gesture Recognition (FG)*, 2018. 1, 3

[8] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard. The megaface benchmark: 1 million faces for recognition at scale. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4873–4882, 2016. 1

[9] D. Yi, Z. Le, S. Liao, and S. Z. Li. Learning face representation from scratch. *arXiv:1411.7923*, 2014. 1

[10] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. MS-Celeb-1M: A dataset and benchmark for large scale face recognition. In *Proc. of European Conference on Computer Vision*, 2016. 1

[11] A. Acien, A. Morales, R. Vera-Rodriguez, I. Bartolome, and J. Fierrez. Measuring the gender and ethnicity bias in deep models for face recognition. In *Proc. of the IAPR Iberoamerican Congress on Pattern Recognition (CIARP)*, 2018. 1

[12] B. Lu, J. Chen, C. D. Castillo, and R. Chellappa. An experimental evaluation of covariates effects on unconstrained face verification. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 1(1):42–55, 2019. 1

[13] J. Buolamwini and T. Gebru. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proc. of Conference on Fairness, Accountability and Transparency*, volume 81, pages 77–91, 2018. 1

[14] A. J. O'Toole, X. An, P. J. Phillips, and J. Dunlop. Demographic effects on estimates of automatic face recognition performance. In *Proc. of International Conf. on Automatic Face and Gesture Recognition (FG)*, pages 83–90, 2011. 1

[15] B. F. Klare, M. J. Burge, J. C. Klontz, R. W. Vorder Bruegge, and A. K. Jain. Face recognition performance: Role of demographic information. *IEEE Transactions on Information Forensics and Security*, 7(6):1789–1801, 2012. 1, 2

[16] C. M. Cook, J. J. Howard, Y. B. Sirotin, J. L. Tipton, and A. R. Vemury. Demographic effects in facial recognition and their dependence on image acquisition: An evaluation of eleven commercial systems. *IEEE Trans. on Biometrics, Behavior, and Identity Science*, 1(1):32–41, 2019. 1

[17] A. Morales, J. Fierrez, and R. Vera-Rodriguez. SensitiveNets: Learning Agnostic Representations with Application to Face Recognition. *arXiv:1902.00334*, 2019. 1, 2

[18] Michele Merler, Nalini Ratha, Rogerio S. Feris, and John R. Smith. Diversity in Faces. *arXiv:1901.10436*, 2019. 1

[19] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proc. of IEEE International Conf. on Computer Vision (ICCV)*, pages 3730–3738, 2015. 2

[20] E. Gonzalez-Sosa, J. Fierrez, R. Vera-Rodriguez, and F. Alonso-Fernandez. Facial soft biometrics for recognition in the wild: Recent works, annotation and cots evaluation. *IEEE Transactions on Information Forensics and Security*, 13(7), 2018. 2

[21] J. Neves, J. Moreno, and H. Proenca. Quis-campi: an annotated multi-biometrics data feed from surveillance scenarios. *IET Biometrics*, 7(4):371–379, 2018. 2

[22] H. Zhang, J. Ross Beveridge, Bruce A. Draper, and P. Jonathon Phillips. On the effectiveness of soft biometrics for increasing face verification rates. *Computer Vision and Image Understanding*, 137(C):50–62, 2015. 2

[23] H. Proença, J. C. Neves, S. Barra, T. Marques, and J. C. Moreno. Joint head pose/soft label estimation for human recognitionin-the-wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(12):2444–2456, 2016. 2

[24] A. Dantcheva, P. Elia, and A. Ross. "What else does your biometric data reveal? a survey on soft biometrics". *IEEE Transactions on Information Forensics and Security*, 11(3):441–467, 2016. 2

[25] A. Dantcheva, C. Velardo, A. D'Angelo, and J.L. Dugelay. Bag of soft biometrics for person identification. *Multimedia Tools and Applications*, 51(2):739–777, 2011. 2

[26] P. Samangouei, V. M. Patel, and R. Chellappa. Facial attributes for active authentication on mobile devices. *Image and Vision Computing*, 58:181 – 192, 2017. 2

[27] H. Han, A. K. Jain, F. Wang, S. Shan, and X. Chen. Heterogeneous face attribute estimation: A deep multi-task learning approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(11):2597–2609, 2018. 2

[28] P. Tome, R. Vera-Rodriguez, J. Fierrez, and J. Ortega-Garcia. Facial soft biometric features for forensic face recognition. *Forensic Science International*, 257:171–284, 2015. 2

[29] P. Tome, J. Fierrez, R. Vera-Rodriguez, and M. S. Nixon. Soft biometrics and their application in person recognition at a distance. *IEEE Transactions on Information Forensics and Security*, 9(3):464–475, 2014. 2

[30] E. Gonzalez-Sosa, A. Dantcheva, R. Vera-Rodriguez, J.L. Dugelay, F. Bremond, and J. Fierrez. Image-based gender estimation from body and face across distances. In *Proc. of International Conf. on Pattern Recognition (ICPR)*, 2016. 2

[31] R. Vera-Rodriguez, P. Marin-Belinchon, E. Gonzalez-Sosa, P. Tome, and J. Ortega-Garcia. Exploring automatic extraction of body-based soft biometrics. In *Proc. of the International Carnahan Conf. on Security Technology (ICCST)*. 2

[32] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016. 2