Mitigating Bias in Gender, Age and Ethnicity Classification: a Multi-Task Convolution Neural Network Approach

Abhijit Das, Antitza Dantcheva and Francois Bremond

Inria, Sophia Antipolis, France
{abhijit.das, antitza.dantcheva, francois.bremond}@inria.fr

Abstract. This work explores joint classification of gender, age and race. Specifically, we here propose a Multi-Task Convolution Neural Network (MTCNN) employing joint dynamic loss weight adjustment towards classification of named soft biometrics, as well as towards mitigation of soft biometrics related bias. The proposed algorithm achieves promising results on the UTKFace and the Bias Estimation in Face Analytics (BEFA) datasets and was ranked first in the the BEFA Challenge of the European Conference of Computer Vision (ECCV) 2018.

Keywords: Bias, Facial analysis, Age, Gender and Race, Soft Biometrics, Facial Attributes

1 Introduction

The prevalent commercial deployment of automated face analysis systems (i.e., face recognition as a robust authentication method) has fueled increasingly the scientific attention. Current machine learning algorithms allow for a relative reliably detection, recognition, as well as categorization of face images w.r.t. age, race and gender. We note though, that training and evaluation data for such algorithms is often biased concerning factors such as age, gender, ethnicity, pose and resolution. Very recently Buolamwini and Gebru [1] reported that algorithms trained with such biased data are inherently bound to produce skewed results. This leads to a significant drop in performance of state of the art models, when applied to images of particular gender and / or ethnicity groups.

Motivated by the above, we here propose a Multi-Task Convolution Neural Network (MTCNN) employing joint dynamic loss weight adjustment targeted to jointly classify gender, age and race, as well as to minimize identified soft biometrics related bias.

The rest of the paper is organized as following: Section 2 reviews related work, Section 3 describes proposed methodology, in Section 4 the experimental results are presented and discussed. Finally Section 5 concludes the work.

2 Related work

2.1 Face analysis

Automated FA is instrumental in a wide range of applications including face detection [2], soft biometrics / face attribute classification [3] and face recognition [4]. Prominently, such applications have been integrated into the current generation of smartphones. Simultaneously, companies such as Google, IBM, Microsoft and Face++ have released commercial software on automated FA. Moreover nowadays, FA is not restricted to face recognition, but has transgressed onto emotion analysis [5], gender classification [6], as well as other facial characteristics [7–9].

Despite of the prevalence of face analysis, accuracies of face recognition systems are lower for particular subcategories of population, such as "female, black". Particularly individuals between the age of 18 and 30 years exhibit low accuracy for face recognition systems used in US-based law enforcement, as reported by Klare et al. [10]. With regards to gender, age and race classification, it is worth mentioning that these algorithms are biased as well, depending on related attribute-subcategories. The latest report on gender classification released by National Institute for Standards and Technology (NIST) reflects on the fact that algorithms performed worse for females than males [11].

Farinella and Dugelay [12] alleged that ethnicity has no effect on gender classification, adopting a binary ethnic categorization scheme for the experimentation: Caucasian and non-Caucasian. Dwork et al. [13] demonstrated the importance of understanding sensitive characteristics such as gender and race, in order to build demographically inclusive models. Proposals for fairness have included parity, such as demographic parity, and equality measures, which require equal false negative rates and false positive rates across subgroups. In a very recent work Buolamwini and Gebru [1] investigated gender and skin type bias in two facial analysis benchmarks, IJB-A and Adience, evaluating three commercial gender classification systems and concluded that darker-skinned females are the most misclassified group. They further recommended that such substantial disparities in the accuracy of classification of gender classification systems require urgent attention, in order to ensure that *FA-systems* are built genuinely *fair, transparent and accountable*. In addition, Garvie et al. [14] observed that African-Americans were more likely to be terminated by law enforcement and subjected to have a biased face recognition searches than individuals of other ethnicity. Therefore, monitoring phenotypic and demographic accuracy of these systems is necessitated.

One further biasing factor concerning gender classification constitutes age. Cheng et al. [15] performed an initial investigation, suggesting that the discriminable features for gender classification for children and adults were significantly different. This was affirmed by Dantcheva et al. [16] and Bilinski et al. [17].

Towards eliminating bias Ranjan and Chellappa [18], as well as Ryu et al. [19] proposed the use of multi-task learning (MTL) networks, and fine-tuned a model trained for face recognition. In another work decoupled classifiers were

proposed to handle bias [20], where the learning of sensitive attributes can be separated from a downstream task in order to maximize both, fairness and accuracy. In particular they employed transfer learning to tackle bias of sub-type based domain adaptation.

2.2 Available datasets

In the context of bias estimation in FA, we note that generally publicly available datasets contain a significant demographic bias. For instance, Labeled Faces in the Wild (LFW), used widely as a benchmark, contains 77.5% male and 83.5% Caucasians [21]. While many works have reported high performance on LFW, the fine grained analysis of the performance considering race, age and gender sub-group has rarely been considered.

To mitigate these limitations, Intelligence Advanced Research Projects Activity (IARPA) introduced an initiative to release the IJB-A dataset as the most diverse set [22]. Further the Adience gender and age classification benchmark was released by Levi and Hassner [23]. As of 2017, NIST started a challenge to spur improvement in face gender classification by expanding on the former 2014-2015 study. Further, to increase the exposure of race diversity Escalera et al. [24] collected the Faces of the World (FotW) dataset, with the aim to achieve a uniform distribution across two genders and four ethnic groups.

While aforementioned datasets include annotation with gender as well as different skin color subgroups, the datasets lack the annotation for age. Consequently these datasets cannot be employed to study factors such as age, gender and race that can bias FA systems. Therefore, the UTKFace dataset was developed recently by Zhang et al. [25], consisting of all required labeling.

3 Proposed approach

3.1 Proposed MTCNN

We propose to use MTCNN with joint dynamic weight loss to classify gender, age and race and further mitigate related bias. The proposed method utilizes disjoint features of the fully connected layers of a Deep CNN using a separated fully connected layers for fulfilling multi-task learning that operates to aim better face attribute analysis. It exploits the synergy and the disjoint features among the tasks, boosting up performances. We exploit the fact that information contained in CNN features is hierarchically distributed throughout the network. Lower layers consist of feature such as edges and corners, and therefore contain better localization features. Hence they are more suitable for learning localization and pose estimation tasks. Whereas, on the other hand, deeper layers, e.g., higher top layers are class-specific and suitable for learning complex tasks such as face recognition and the fully connected layers involve for the classification task i.e., where the end to end system can learn and attempt to discriminate the salient features for different inherent tasks in a MTCNN scenario. Given the aforementioned MTCNN-characteristic, as well as the aim of the work to enhance face

attribute analysis, we propose to customize Facenet [26] for face recognition with ResNet V1 inception (as it is one of the prominent face architectures). A block digram of the proposed MTCNN is illustrated in Fig. 1.

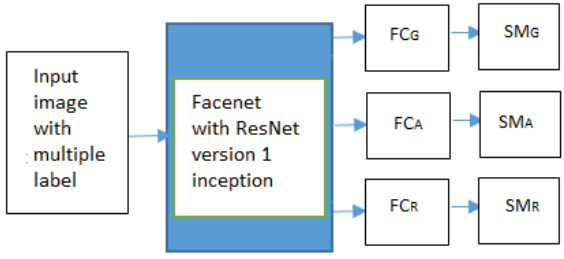

Fig. 1. Block diagram of proposed MTCNN for face attribute analysis (FC_R= fully connected layer of race classification task, FC_A= fully connected layer of age classification task, FC_G= fully connected layer of gender classification task, SM_R= Softmax layer of race classification task, SM_G= Softmax layer of gender classification task and SM_A= Softmax layer of age classification task).

The Facenet network consists of a batch of input layer and a Deep CNN architecture (ResNet V1 in our scenario) followed by L2 normalization, which results in face embedding. This is followed by the triplet loss during training. The architecture consists of a stream of convolution layers, normalization layer and pooling layers followed by 3 inception blocks and their reduction. The latter followed by dropout and a fully connected layer. At this point, we split the network into three separate branches corresponding to the different classification tasks (i.e., gender, race and age). We add three fully connected layers, one for race classification, second for gender classification and third for age classification. Finally, a Softmax layer is added to each of the branches to predict the individual task labels feature with L2 normalization and respective dropout layer. After each convolution a Rectified Linear Unit (ReLU) is deployed as activation function. The Facenet model turns an image of a face into a vector of 128 floating point numbers. These 128 embedding can be used as features for classification. While using Facenet we fine-tuned the fully connected layer.

In the MTCNN the network is split at the fully connected layer followed by individual Softmax layer of each task. Therefore an input layer tuple of the MTCNN, for a given training set T with N images contains T = Ii, Yi, where i=1:N, where Ii is the image and Yi is a vector consisting of the labels. In MTCNNs it is challenging to define the loss weight for each task. In previous works, this was dealt either by treating all tasks equally Dong et al,. [27], dynamic MTCNN Fang et al., [28], obtaining weights via brute-force search [28] or by dynamically assigning disjoint weights for the side task [29]. However neither of this strategies work in our setting. Unlike pose and illumination, gender, race

and age classification are closely related facial features. Moreover, they possess varying degrees of relevance for both intra-class and inter-class variation. Such relevance depends on their intensity exhibited per samples. Hence, we seek to optimize the effect of multi-task facial attribute classification (i.e., gender, age and race) by learning them jointly and dynamically, depending on the degree of relevance of the feature present for each classification task. Specifically, the MTCNN should directly learn classification task relations from data instead of subjective task grouping. Thereby deciding weight of the task sharing. Hence, we propose a joint dynamic weighting scheme to automatically assign the loss weights for the each task during training.

First, we find the summed weight for the each classification task by brute-force search on the validation set. Further by adding a fully connected layer and a Softmax layer to each task the model gets proficient to shared features from the last common layer, which is aimed at learning the dynamic weights for each iteration depending degree of relevance of the task. Therefore we obtain the dynamic weight percentages for each task from the fully connected layer. Further, the function of the soft-max layer converts the dynamic weights to positive values that sum to 1. Consequently, the most relevant task is to contribute predominantly to the final loss and the additional task is to contribute to the relevant task, in order to reduce the loss of the most relevant task. Thereby, the MTCNN should assign a higher weight for a non-relevant task with a lower loss, in order to reduce the overall loss. A mini-batch Stochastic Gradient Descent (SGD) was employed to solve the above optimization problem of loss weight. Further, the weights were averaged for each batch.

3.2 Implementation details

The Python library of Facenet[1] is used to calculate facial embedding of face images and developing the proposed MTCNN. The Facenet library was implemented in TensorFlow. It includes pre-trained Facenet models for face recognition. The models have been validated on the LFW database [30] and were trained on a subset of the MSCeleb-1M database [31]. The models architecture follows the Inception-ResNet-v1 network [32].

The Facenet library includes an implementation of detection, alignment and landmark estimation, as proposed by Zhang et al, [33] which we use for pre-processing for our images. The output of our proposed MTCNN is a 128 dimension floating point embedding, similar to Facenet. The Scikit-learns SVM version with RBF kernel is used with Tensorflow for classification of this embedding for face recognition. As the pre-trained model was trained on a much larger face dataset but with less similarity in respect to face attributes, we employ transfer learning. Specifically we freeze the initial layers of the pre-trained model and train the last top layers. Therefore, the top layers (which are known to contain the face attribute information) are customized to our setting of interest. During transfer learning we ensure that the final layers are not restored from the

[1] https://github.com/davidsa-ndberg/facenet

pre-trained model and we also to ensure that gradients are gated for all other parameters during training. While fine tuning the weight decay is set to 0.0005. All models are trained for 10 epochs with a batch size of 20. The learning rate starts at 0.001 and reduces at 5th, 7th, and 9th epochs with a factor of 0.1.

4 Experiment results and discussions

We proceed to present employed datasets, experimental protocol, implementation details, obtained results obtained, as well as to discuss these. The results achieved in the ECCV 2018 BEFA challenge employing the proposed MTCNN are also explained at the end of this section.

4.1 Datasets

We report experimental results on two datasets: UTKFace and BEFA challenge dataset, which we proceed to describe.

1. **UTKFace** dataset is a large-scale face dataset (over 20,000 face images) with a population coverage of long age span ranged from 0 to 116 years. Specifically, annotation for age includes following classes: baby: 0-3 years, child: 4-12 years, teenagers: 13-19 years, young: 20-30 years, adult: 31-45 years, middle aged: 46-60 years and senior: 61 and above years. The dataset additionally contains the labeling for gender (male and female), as well as five races (White, Black, Asian, Indian and other race). UTKFace includes large variations of pose, facial expression, illumination, occlusion, and resolution.

2. **BEFA** challenge dataset[2] is the official dataset of the related challenge. It contains 13431 test images. It has been annotated for age (baby, child, teenagers, young, adults, middle age and senior), gender (male and female), and ethnicity (white, black, Asian and Indian). As per challenge goals, these soft biometric traits and trait instances are represented in a balanced manner.

4.2 Experimental protocol

We classify gender, age and race, respectively, analyzing the trait instances of the remaining facial attributes. For instance for age classification we report the age true classification rates for males, females, as well as for each category separately.

1. **UTKFace**: The dataset consists of three parts. Part I, II and III consist of 10437, 10719 and 3252 images, respectively. We use part I for training (three forth for fine tuning and rest for training the classifier), part II for testing and part III for validation.

[2] https://sites.google.com/site/eccvbefa2018/

2. **BEFA**: The BEFA dataset was used only for testing. Training and fine tuning were performed as above. As per the challenge protocol, we classify the soft biometric traits, analyzing all possible categories. Hence all 56 are considered (i.e., one possible category being "male, young, Indian").

4.3 Results and analysis

1. **UTKFace**: The overall mean classification accuracy for race, gender and age classification is summarized in Table 1. We observe that the proposed approach significantly improves accuracies achieved by Facenet and its fine-tuned model, consistently for all attributes.

Table 1. Mean classification accuracy of the race, gender and age classification [%].

	Race	Gender	Age
Facenet	85.1	91.2	56,9
Finetuned Facanet (FFNet)	86.1	96.1	64
Proposed MTCNN	90.1	98.23	70.1

Table 2. The classification accuracy of the **race classification** considering gender and age instances [%].

	FFNet	Proposed MTCNN
Male	87	90.9
Female	84.3	89.1
Baby	100	100
Child	84.3	88.8
Teenager	85.8	89.1
Young	84.9	88.9
Adult	88.1	91.5
Middle	87.7	90.7
Senior	81.9	87.7

The results in Table 2 suggest the presence of a bias in **race classification** w.r.t. gender and age for the FFNet approach. While babies are categorized to 100% ethnically, the accuracy decreases down to 81.9% for seniors. Less profound, but similarly while females are categorized to 84.3% ethnically, males reach 87% true classification accuracy. We note, that such bias has been mitigated to a large extent by the proposed MTCNN. The remaining

Table 3. The classification accuracy of the **age classification** considering race and gender instances [%].

	FFNet	Proposed MTCNN
Male	61,5	69.1
Female	66,8	70.9
White	61,8	69.6
Black	59,2	68.9
Asian	78,7	80.1
Indian	63,6	69.5
Others	64.5	66.7

low level of bias still persists among the senior age category, possibly due to lower race exhibits in older subjects.

Table 3 summarizes the accuracies of **age classification**. Again, FFNet's performance indicates for a significant bias, with classification rates ranging from 61.5% for males to 78.7% for Asians. We note a classification accuracy higher for females than for males. Such bias in gender estimation has been nearly mitigated by the proposed MTCNN. A certain low level of bias can be still found among race instances.

The accuracies in Table 4 indicate that for gender classification, bias w.r.t. race and age can be mainly observed for the instances baby, child and teenagers employing FFNet. Again, this bias has been nearly mitigated by the proposed MTCNN. Remaining bias is accredited to low sexual dimorphism in babies.

Table 4. The classification accuracy of the **gender classification** considering gender and race instances [%].

	FFNet	Proposed MTCNN
Baby	70	80.5
Child	79,6	96.7
Teenager	92	95.
Young	96,8	97
Adult	97,7	98.3
Middle	96,6	97.7
Senior	95,1	96.5
White	97	98.7
Black	95	98.6
Asian	97,5	99.3
Indian	97,8	99.1
Others	97.5	99

2. **BEFA**: The overall mean classification accuracies of the BEFA challenge dataset employing the proposed MTCNN are summarized in Table 5.

Table 5. The overall mean classification accuracy of race, gender, age and overall attribute classification on BEFA challenge dataset [%].

	All Attributes	Race	Gender	Age
MTCNN	56.37	84.29	93.72	71.83

The related confusion matrices of gender, age, race classification and attribute classification considering age, gender and race along intersection for all the attributes are illustrated in Fig. 2 to Fig. 5. It can be concluded that age, gender and race classification performance achieved by the proposed MTCNN on the BEFA dataset is rather encouraging, whilst there are some remaining scenarios with low performance. For instance, it can be observed from Fig 2 that the performance of gender classification was low among the baby subgroup, moreover it was worst for female babies with Asian race. In terms of age classification, we observe from Fig 3 that the classification performance was low for the adults for most instances of race and gender, specifically with lowest accuracies for black females. Further from Fig 4, we see that for race classification the performance is lowest for Asian and Indian seniors. While considering all attributes together (age, race and gender), the performance dropped, with lowest accuracies for teenagers and adult black females.

5 Conclusions

In this paper, we presented an approach for gender, age and race classification targeted to minimize inter-class bias. The proposed multi-task CNN approach utilized joint dynamic loss, providing promising results on the UTKFace and the Bias Estimation in Face Analytics (BEFA) challenge datasets. The proposed algorithm was ranked first in the related BEFA-challenge of the European Conference on Computer Vision (ECCV) 2018. In future work, we intend to extend the current study onto other facial attributes. In addition, we intend to explore the approach presented in this work in the context of mitigating biases in face recognition.

Acknowledgement

A. Das was supported by the research program FER4HM funded by Inria and CAS. A. Dantcheva was supported by the French Government (National Research Agency, ANR) under grant agreement ANR-17-CE39-0002.

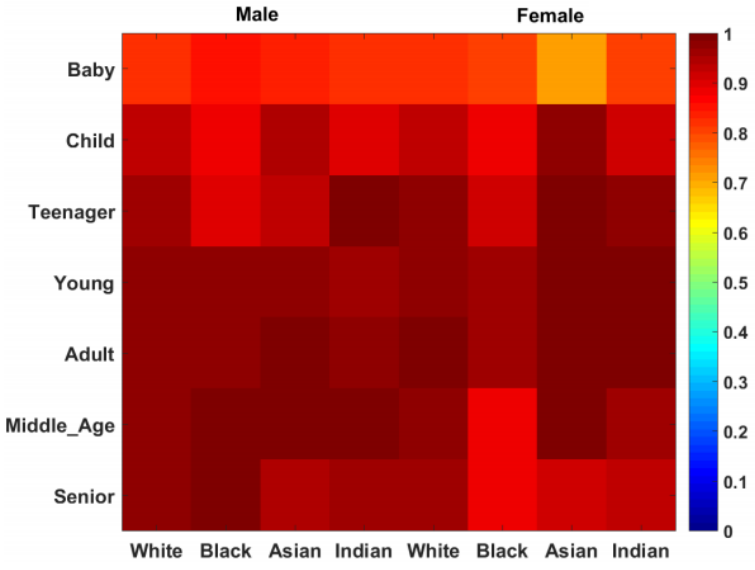

Fig. 2. Confusion matrix of gender classification along intersections of all attributes.

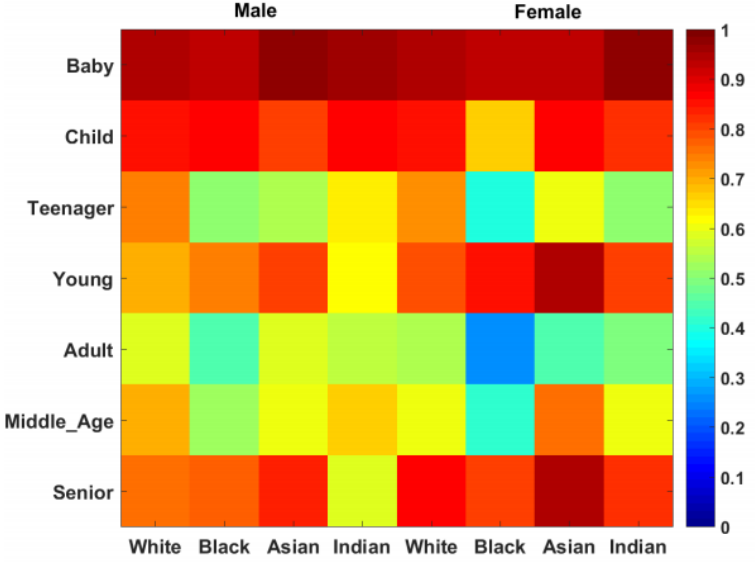

Fig. 3. Confusion matrix of age classification along intersections of all attributes.

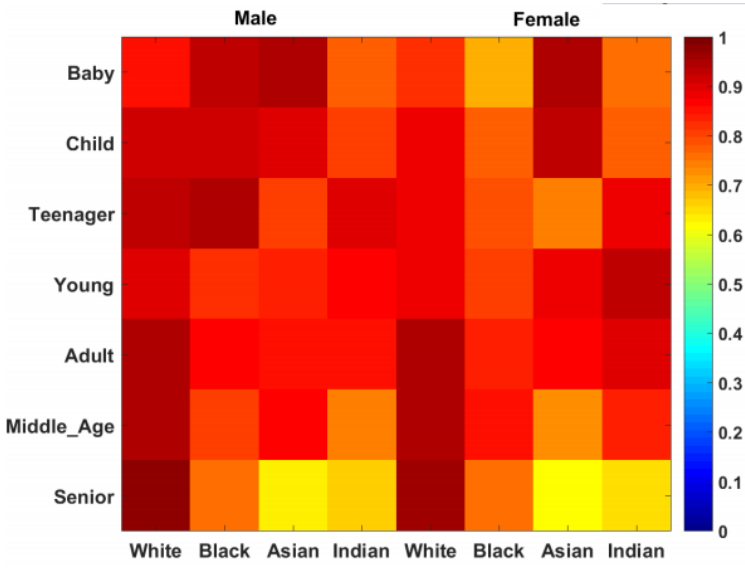

Fig. 4. Confusion matrix of race classification along intersections of all attributes.

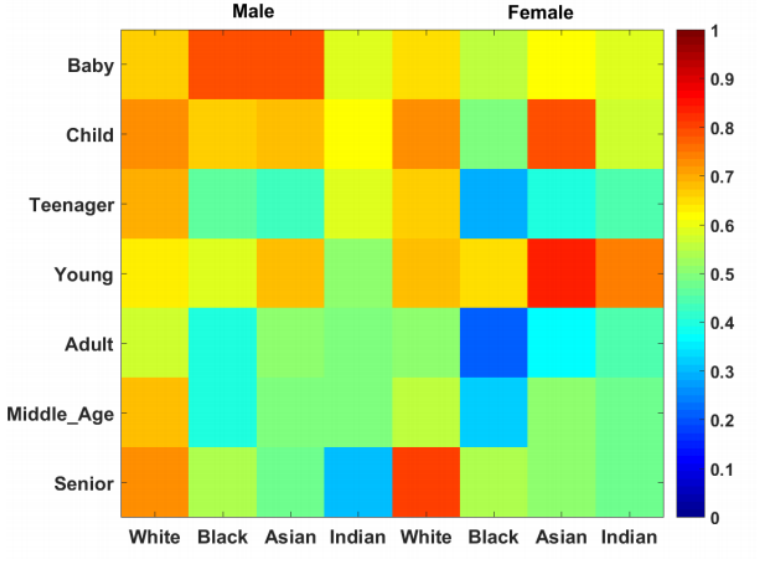

Fig. 5. Confusion matrix of attributes classification considering all attributes together (gender, age and race) along intersections of all attributes.

References

1. Buolamwini, J., Gebru, T.: Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Conference on Fairness, Accountability and Transparency. (2018) 77–91
2. Zafeiriou, S., Zhang, C., Zhang, Z.: A survey on face detection in the wild: past, present and future. Computer Vision and Image Understanding **138** (2015) 1–24
3. Dantcheva, A., Elia, P., Ross, A.: What else does your biometric data reveal? a survey on soft biometrics. IEEE Transactions on Information Forensics and Security (2015) 1–26
4. Ranjan, R., Sankaranarayanan, S., Castillo, C.D., Chellappa, R.: An all-in-one convolutional neural network for face analysis. In: Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on, IEEE (2017) 17–24
5. Dehghan, A., Ortiz, E.G., Shu, G., Masood, S.Z.: Dager: Deep age, gender and emotion recognition using convolutional neural network. arXiv preprint arXiv:1702.04280 (2017)
6. Wang, Y., Kosinski, M.: Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. Journal of personality and social psychology **114**(2) (2018) 246
7. Wu, X., Zhang, X.: Automated inference on criminality using face images. arXiv preprint arXiv:1611.04135 (2016) 4038–4052
8. Dantcheva, A., Bremond, F., Bilinski, P.: Show me your face and i will tell you your height, weight and body mass index. In: International Coference on Pattern Recognition (ICPR). (2018)
9. Dantcheva, A., Dugelay, J.: Female facial aesthetics based on soft biometrics and photo-quality. In: IEEE International Conference on Multimedia and Expo (ICME). (2011)
10. Klare, B.F., Burge, M.J., Klontz, J.C., Bruegge, R.W.V., Jain, A.K.: Face recognition performance: Role of demographic information. IEEE Transactions on Information Forensics and Security **7**(6) (2012) 1789–1801
11. Ngan, M., Ngan, M., Grother, P.: Face recognition vendor test (FRVT) performance of automated gender classification algorithms. US Department of Commerce, National Institute of Standards and Technology (2015)
12. Farinella, G., Dugelay, J.L.: Demographic classification: Do gender and ethnicity affect each other? In: Informatics, Electronics & Vision (ICIEV), 2012 International Conference on, IEEE (2012) 383–390
13. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. In: Proceedings of the 3rd innovations in theoretical computer science conference, ACM (2012) 214–226
14. Clare Garvie, A.B., Frankle, J.: The perpetual line-up: Unregulated police face recognition in america. (2016)
15. Cheng, Y.D., O'Toole, A.J., Abdi, H.: Classifying adults' and children's faces by sex: Computational investigations of subcategorical feature encoding. Cognitive science **25**(5) (2001) 819–838
16. Dantcheva, A., Bremond, F.: Gender estimation based on smile-dynamics. IEEE Transactions on Information Forensics and Security **12**(3) (2017) 719–729
17. Bilinski, P., Dantcheva, A., Brémond, F.: Can a smile reveal your gender? In: Biometrics Special Interest Group (BIOSIG), 2016 International Conference of the, IEEE (2016) 1–6

18. Ranjan, R., Castillo, C.D., Chellappa, R.: L2-constrained softmax loss for discriminative face verification. arXiv preprint arXiv:1703.09507 (2017)
19. Ryu, H.J., Adam, H., Mitchell, M.: Inclusivefacenet: Improving face attribute detection with race and gender diversity. In: Workshop on Fairness, Accountability, and Transparency in Machine Learning (FAT/ML). (2018)
20. Dwork, C., Immorlica, N., Kalai, A.T., Leiserson, M.D.: Decoupled classifiers for group-fair and efficient machine learning. In: Conference on Fairness, Accountability and Transparency. (2018) 119–133
21. Han, H., Jain, A.K.: Age, gender and race estimation from unconstrained face images. Dept. Comput. Sci. Eng., Michigan State Univ., East Lansing, MI, USA, MSU Tech. Rep.(MSU-CSE-14-5) (2014)
22. Klare, B.F., Klein, B., Taborsky, E., Blanton, A., Cheney, J., Allen, K., Grother, P., Mah, A., Jain, A.K.: Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2015) 1931–1939
23. Levi, G., Hassner, T.: Age and gender classification using convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. (2015) 34–42
24. Escalera, S., Torres Torres, M., Martinez, B., Baró, X., Jair Escalante, H., Guyon, I., Tzimiropoulos, G., Corneou, C., Oliu, M., Ali Bagheri, M., et al.: Chalearn looking at people and faces of the world: Face analysis workshop and challenge 2016. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. (2016) 1–8
25. Zhang, Z., Song, Y., Qi, H.: Age progression / regression by conditional adversarial autoencoder. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE (2017)
26. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2015) 815–823
27. Yi, D., Lei, Z., Liao, S., Li, S.Z.: Learning face representation from scratch. arXiv preprint arXiv:1411.7923 (2014)
28. Fang, Y., Ma, Z., Zhang, Z., Zhang, X.Y., Bai, X.: Dynamic multi-task learning with convolutional neural network
29. Yin, X., Liu, X.: Multi-task convolutional neural network for pose-invariant face recognition. IEEE Transactions on Image Processing (2017)
30. Huang, G.B., Mattar, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In: Workshop on faces in'Real-Life'Images: detection, alignment, and recognition. (2008)
31. Guo, Y., Zhang, L., Hu, Y., He, X., Gao, J.: Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In: European Conference on Computer Vision, Springer (2016) 87–102
32. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-resnet and the impact of residual connections on learning. (2017)
33. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters **23**(10) (2016) 1499–1503