SensitiveNets: Unlearning Undesired Information for Generating Agnostic Representations with Application to Face Recognition

Aythami Morales, Julian Fierrez, Ruben Vera-Rodriguez

Abstract—This work proposes a new neural network feature representation that help to leave out sensitive information in the decision-making process of pattern recognition and machine learning algorithms. The aim of this work is to develop a learning method capable to remove certain information from the feature space without drop of performance in a recognition task based on that feature space. Our work is in part motivated by the new international regulation for personal data protection, which forces data controllers to avoid discriminative hazards while managing sensitive data of users. Our method is based on a triplet loss learning generalization that introduces a sensitive information removal process. The method is evaluated on face recognition technologies using state-of-the-art algorithms and publicly available benchmarks. In addition, we present a new annotation dataset with balanced distribution between genders and ethnic origins. The dataset includes more than 120K images from 24K identities with variety of poses, image quality, facial expressions, and illumination. The experiments demonstrate that it is possible to reduce sensitive information such as gender or ethnicity in the feature representation while retaining competitive performance in a face recognition task.

Index Terms—face recognition, learning representation, agnostic representation, algorithmic discrimination.

1 Introduction

In May 2019, the Board of Supervisors of San Francisco banned the use of facial recognition software by the police and other agencies [1]. Face recognition algorithms are good examples of recent advances in Artificial Intelligence (AI). During the last ten years, the accuracy of face recognition systems has increased up to 1000x (it is probably the biometric technology with the greatest investment nowadays). The face recognition algorithms are dominated by Deep Neural Network architectures. These algorithms are trained with huge amounts of data with little control on what is happening during training (training typically focused on performance improvement). In other words: algorithms with excellent performance but quite opaque. This trend in AI (excellent performance + low transparency) can be observed not only in face biometrics, but also in many other AI applications [2]. At this point, and despite the advances in recognition performance, other factors such as the lack of transparency, discrimination, privacy, etc. are limiting many practical applications.

During the last decade, the accuracy has been the key concern for researchers developing automatic decision-making algorithms, e.g., in biometric systems for person recognition [3]. Recent progress under that umbrella has made possible and practical automatic decision-making in quite challenging problems, e.g., biometrics in the wild [4]. However, the recognition accuracy is not the only aspect to attend when designing learning algorithms. Algorithms have an increasingly important role in decision-making in several processes involving humans. These decisions have therefore increasing effects in our lives, and there is an increasing need for developing machine learning methods that guarantee fairness in such decision-making [5-11]. These methods must consider sociotechnical factors that involve a comprehensive framework including citizens, data, algorithms, laws, companies, governmental agencies, and therefore, the complete chain.

Discrimination can be defined in this context as the unfair treatment of an individual because of his or her membership in a particular group, e.g. ethnic, gender, etc. The right to non-discrimination is deeply embedded in the normative framework that underlies various national and international regulations, and can be found for example in Article 7 of the Universal Declaration of Human Rights, and Article 14 of the European Convention on Human Rights, among others. As a prove of these concerns, in April 2018 the European Parliament adopted a set of laws aimed to regularize the collection, storage and use of personal information, the General Data Protection Regulation (GDPR) [12]. According to paragraph 71 of GDPR, data controllers who process sensitive data have to "implement appropriate technical and organizational measures …" that "… prevent, inter alia, discriminatory effects".

• *Authors are with the Biometric and Data Pattern Analytics Lab – BiDA, Escuela Politecnica Superior, C/ Francisco Tomás y Valiente, 11 Universidad Autónoma de Madrid, Campus de Cantoblanco, 28049 Madrid. Email: {aythami.morales, julian.fierrez, ruben.vera}@uam.es*

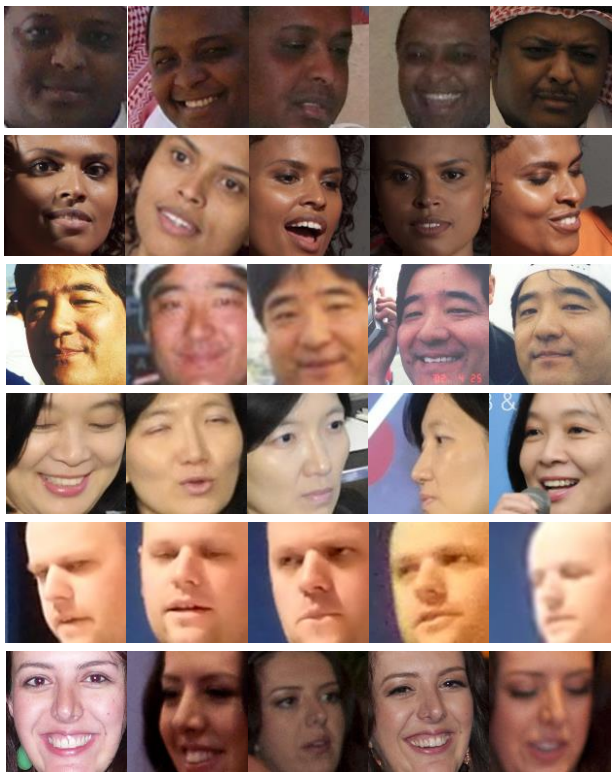


Fig. 1. Examples of the six classes available in DiveFace. Faces have been cropped using the algorithm proposed in [14].

The aim of this work is to develop a new agnostic feature representation capable to remove certain sensitive information. The resulting networks trained with the proposed representation, called SensitiveNets, can be trained for specific tasks (e.g. image classification, face recognition, speech recognition, ...), while minimizing the contribution of selected covariates, both for the task at hand and in the information embedded in the trained network. Those covariates will typically be the source of discrimination that we want to prevent (e.g. gender, ethnicity, age, and context).

We evaluate the new proposed representation removing gender and ethnicity information from the decision-making of state-of-the-art face recognition systems. The proposed representation is evaluated on face recognition technology because of: i) the high level of sensitive information present in face imaging (e.g. gender, age, ethnicity, health); ii) it is a very active research topic with interest from both academia and industry; and iii) it is a challenging pattern recognition problem with multiple sources of variations (e.g. pose, illumination, image quality). Facial attributes revealing the gender, age or ethnicity have the potential to discriminate citizens based on the group to which that person belongs [13].

The contributions of this work are two-fold: i) a new feature representation aimed at eliminating sensitive information from the decision-making of recognition algorithms based on deep neural network embeddings. This new representation is able to work with pre-trained models, therefore it is compatible with existing networks (e.g., ResNet trained on VGGFace2 used in our experiments). To the best of our knowledge, this is the first work that has addressed this challenge for face recognition algorithms. And ii) a new annotation dataset (DiveFace) with uniform distribution between genders and ethnic origins, used in our experiments. The dataset includes more than 120K images from 24K identities with variety of poses, image quality, facial expressions and illumination (see Fig. 1). This database is a step forward with respect to the resources available in this area (e.g., DiveFace is ten times bigger than the dataset recently used in [10]).

The rest of the paper is structured as follows: end of section 1 summarizes related works on this area. Section 2 introduces the proposed agnostic representation learning process. Section 3 describes the database generated for this study, while Section 4 presents experiments and results. Finally, Section 5 will summarize the main conclusions.

1.1 Related Works

The evaluation of the fairness of an algorithm/technology requires a deep analysis of the sociotechnical context of its usage [8]. We cannot separate technology from the context of its usage. The study of discrimination-aware information technology is not new and includes efforts from different research communities. In [14-16] researchers analyzed several techniques to improve fairness through discrimination-aware data mining. Those three approaches proposed to act on the decision rules in order to compensate discriminatory effects. Similarly, a modified Bayes classifier focused on reducing discriminatory effects was proposed in [19]. Researchers proposed to modify the probability distributions of the classifiers as well as balanced models to guarantee fair decision-making [18]. How to detect discriminatory effects on web-platforms was discussed in [20] with approaches focused on auditing algorithms and its decision rules. All those approaches [14-19] proposed methods to act on the decisions rather than the learning processes.

The study of new learned representations to improve fairness of the learning processes has attracted other researchers [21][22]. In [21] researchers proposed projection methods to preserve individual information while obfuscating membership to specific groups. The main drawback of the proposed techniques was that discrimination was modelled as statistical imparity, which is not applicable when the classification goal is not directly correlated with membership in a specific group.

The concept of agnostic neural network as it is used in the present work was introduced in [22]. That work proposed a learning method to recognize images minimizing the usage of context information (e.g. recognize a whale without using information from sea regions in the image). The main problem of that approach was the large performance drop when not using the context.

Bias and discrimination are related to each other but

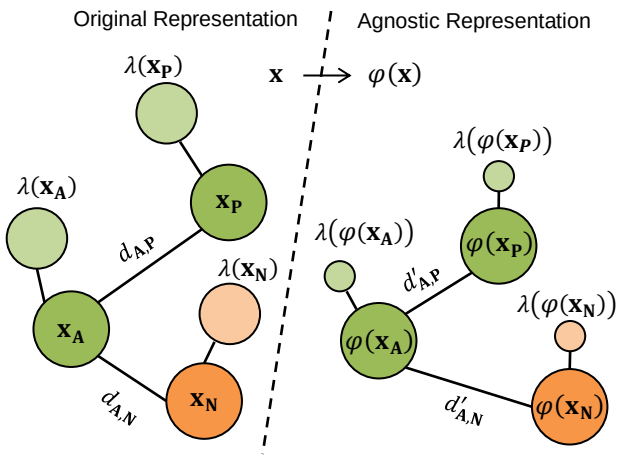


Fig. 2. Notation and data involved to generate the new feature representation $\varphi(\mathbf{x})$. As a result of the learning process, the distance between embeddings from the same class in the new representation and the sensitive information are reduced ($d'_{\mathbf{A,P}} < d_{\mathbf{A,P}}$ and $\lambda(\varphi(\mathbf{x})) < \lambda(\mathbf{x})$) while the distance between different classes is augmented ($d'_{\mathbf{A,N}} > d_{\mathbf{A,N}}$).

they are not necessarily the same thing. Bias is traditionally associated with unequal representation of classes in a dataset [23]. Dataset bias can produce unwanted results in the decision-making of algorithms, e.g., different face recognition accuracy depending of your ethnicity [24][25]. Researchers have explored new representations capable to compensate this dataset bias [10][26]. The proposal in [10] is based on a joint learning and unlearning algorithm inspired in domain and task adaptation methods [27]. Similarly, the method in [26] proposed an approach based on Multitask Convolutional Neural Networks to compensate the dataset bias in face attribute classification. These works reported encouraging results showing that it is possible to remove sensitive information (named as spurious variations in [10]) for age, gender, ancestral origin and pose in face processing. However, those works did not address the problem on face verification tasks.

De-identification is another approach proposed to improve the privacy in face processing. In [28][29] researchers proposed de-identification techniques that obfuscate gender attributes while preserving face verification accuracy. The main drawback of these techniques is that gender information is not eliminated but distorted.

2 Agnostic Representation: Our Method

Triplet loss was proposed to improve the performance of face descriptors in verification algorithms [30][31]. In this work we propose a generalization of triplet loss to train a new representation that maintains a competitive recognition performance without using sensitive information during the decision-making. Assume that each image is represented by an embedding descriptor $\mathbf{x} \in \mathbb{R}^d$ 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 images from the same class (e.g. an identity in face recognition), and Negative (**N**) is an image from a different class. We form a list of triplets **T** that violate the recognition performance constraint:

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

where i is the index of the triplet, $\|\cdot\|$ is the Euclidean Distance and α is a real numbered threshold. In words, we select difficult triplets where the inter-class distance is smaller than the intra-class distance.

In order to eliminate the sensitive information from the pre-trained model without a significant drop of recognition performance, we propose a modified loss function:

$$loss = \sum_{i \in \mathbf{T}} \left[\left\|\varphi(\mathbf{x}_\mathbf{A}^i) - \varphi(\mathbf{x}_\mathbf{P}^i)\right\|^2 - \left\|\varphi(\mathbf{x}_\mathbf{A}^i) - \varphi(\mathbf{x}_\mathbf{N}^i)\right\|^2 + \Lambda^i\right], \tag{2}$$

where $\varphi(\mathbf{x})$ is the projection function of the new feature representation, **T** is a list of triplets, and Λ^i is the amount of sensitive information present in each triplet. Λ^i is calculated as:

$$\Lambda^i = \lambda\left(\varphi(\mathbf{x}_\mathbf{A}^i)\right) + \lambda\left(\varphi(\mathbf{x}_\mathbf{P}^i)\right) + \lambda\left(\varphi(\mathbf{x}_\mathbf{N}^i)\right) + \alpha, \tag{3}$$

where $\lambda(\varphi(\mathbf{x}))$ is obtained from the projected feature representation according to the equation:

$$\lambda_\mathbf{A}^i = \lambda\left(\varphi(\mathbf{x}_\mathbf{A}^i)\right) = \log\left(1 + \sum_{s=1}^{C}\left|\frac{1}{C} - p_s(\varphi(\mathbf{x}_\mathbf{A}^i))\right|\right), \tag{4}$$

being $p_s(\varphi(\mathbf{x}_\mathbf{A}^i))$ the sensitivity detection output (softmax function) corresponding to the probability $P\left(s\middle|\varphi(\mathbf{x}_\mathbf{A}^i)\right)$, C is the number of sensitive classes (e.g. two for gender), and $|\cdot|$ is the absolute value. $\lambda_\mathbf{A}^i$ will tend to zero when $p_s(\varphi(\mathbf{x}_\mathbf{A}^i))$ approximates the random chance ($1 / C$) for all classes s. $\lambda_\mathbf{N}^i$ and $\lambda_\mathbf{P}^i$ are calculated similarly. The notation and data involved are summarized in Fig. 2. The proposed loss function minimizes the sensitive information λ, while maintains distances between positive and negative embeddings.

The method proposed to generate $\varphi(\mathbf{x})$ is data driven and potentially applicable to any feature representation. In this paper we show its application in face biometrics to remove information such as gender and ethnicity. Without loss of generality, our training algorithm for obtaining $\varphi(\mathbf{x})$ works as follows when applied to this problem:

a. We train a sensitivity detector $p(\mathbf{x})$ using the face representation generated by the pre-trained model **x** and the sensitive feature to remove (e.g. gender). The detector is a dense classification layer (softmax function) with number of units equal to the number of classes C of that feature. Given a face descriptor $\mathbf{x} \in \mathbb{R}^d$, the output of the detector $p(\mathbf{x}) \in \mathbb{R}^C$ will be the probability to belong to the different classes of the feature to remove.

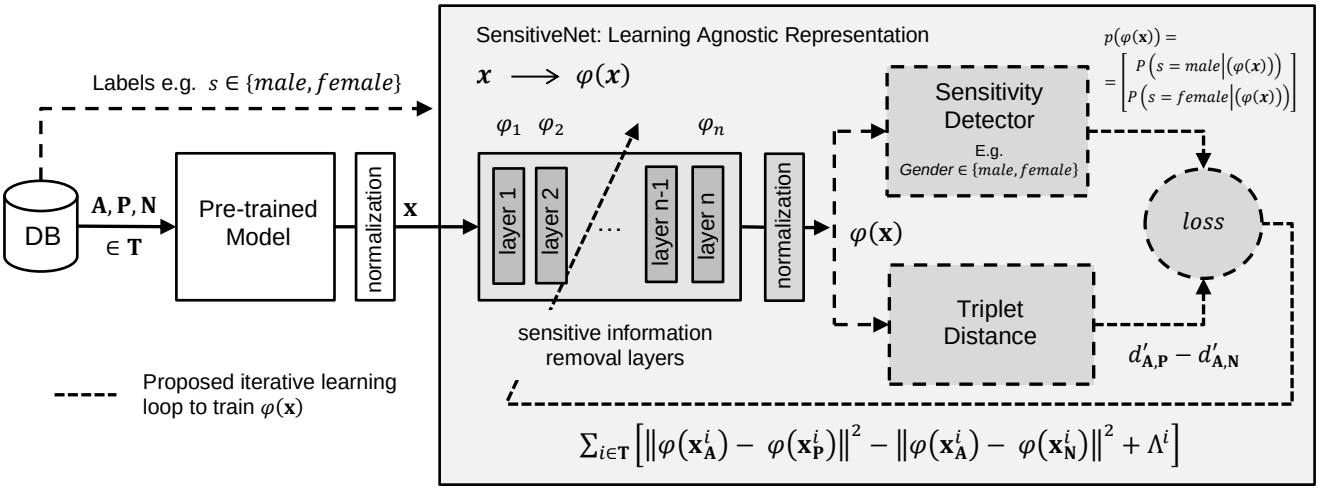


Fig. 3. Training process of SensitiveNets to remove sensitive information from the pre-trained embedding representation **x**. The normalization is a l_2-norm and the Sensitivity Detector is trained using a softmax classification layer. The resulting feature representation is $\varphi(\mathbf{x})$.

b. We add a dense connected layer with l units (linear activation function). The feature projection $\varphi_1(\mathbf{x})$ is trained to minimize the loss function presented in equation (2) according to a list **T** of triplets that satisfy the equation (1) and the detector trained in step *a*.

c. We re-train $p(\mathbf{x})$ using the projected representation $\varphi_1(\mathbf{x})$ instead of the original **x**. This sensitivity detector serves to reveal new covariates of the projected feature representation $\varphi_1(\mathbf{x})$.

d. We add a new dense connected layer with l units (linear activation function). The embedding projection $\varphi_2(\varphi_1(\mathbf{x}))$ is trained to minimize the loss function presented in equation (2) according to a list **T** of triplets that violate the equation (1) and the detector trained in step *c*.

e. We repeat steps *c* and *d* using projections trained in each step. The process ends when the accuracy of the sensitive detector is below a threshold. Fig. 3 illustrates the proposed training method.

The final result of the learning process is a projection function $\varphi(\mathbf{x})$ trained to remove sensitive information and maintain recognition performances.

3 DiveFace Database: An annotation dataset for face recognition trained on diversity

The database presented in this work is generated using the Megaface MF2 training dataset [32]. We decided to leverage MF2 after exploring other public datasets including CelebA [33], VGGFace2 [34], and LFW [35]. MF2 is part of the publicly available Megaface dataset with 4.7 million faces from 672K identities and it includes their respective bounding boxes. All images were obtained from Flickr Yahoo's dataset [36].

DiveFace[1] contains annotations equally distributed among six classes related to gender and three ethnic groups. Gender and ethnicity have been annotated following a semi-automatic process. There are 24K identities (4K for each class). The average number of images per identity is 5.5 with a minimum number of 3 for a total number of images greater than 120K. Users are grouped according to their gender (male or female) and three categories related with ethnic physical characteristics:

- East Asian: people with ancestral origin in Japan, China, Korea and other countries in that region.
- Sub-Saharan and South Indian: people with ancestral origins in Sub-Saharan Africa, India, Bangladesh, Bhutan, among others.
- Caucasian: people with ancestral origins from Europe, North-America and Latin-America (with European origin).

We are aware about the limitations of grouping all human ethnic origins into only 3 categories. According to studies, there are more than 5K ethnic groups in the world. We categorized according to only three groups in order to maximize differences among classes. As we will show in the experimental section, automatic classification algorithms based on these three categories show performances up to 98% accuracy.

DiveFace is employed to train the method proposed in Section 2. In order to demonstrate the generalization capacity of the method, we evaluate the verification results over another two popular face datasets. Labeled Faces in the Wild (LFW) [35] is a database for research on unconstrained face recognition [37]. The database contains more than 13K images of faces collected from the web. We employ the aligned images from the test set provided with view 1 and its associated evaluation protocol. CelebA [33] is a large-scale face attributes dataset with more than 200K celebrity images. While the gender attributes are provided together with the CelebA dataset, ethnicity

[1] Dataset available here: https://github.com/BiDAlab/DiveFace

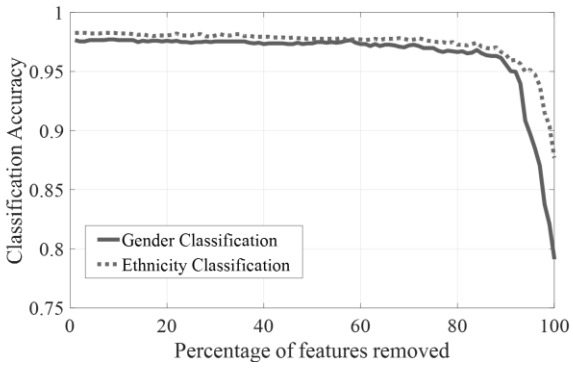


Fig. 4. Classification accuracy for gender (continuous line) and ethnicity (dotted line) vs percentage of features removed from the feature space before training. Accuracy obtained with Resnet-50 embeddings (2048 features) and a softmax classification layer trained after feature removal.

was labeled according to a commercial COTS ethnicity detection system. These three databases are composed by images acquired in the wild, with large pose variations, and varying face expressions, image quality, illuminations, and background clutter, among other variations [4].

4 Experiments

4.1 Pre-trained model

The performance of face recognition technology has been boosted by deep convolutional neuronal networks that have drastically reduced the error rates in the last decade [38]. On the other hand, a face image reveals information not only about who we are but also about what we are (e.g. gender, ethnicity, age). Researchers have proposed to exploit such auxiliary data of the users to improve face recognition [24][39][40]. These auxiliary data are also known as soft biometrics, which refer to those biometrics that can distinguish different groups of people but do not provide enough information to uniquely identify a person [40]. Those attributes can be extracted with high accuracy using just one face picture [37][39].

In our experiments we employ the popular face recognition pre-trained model ResNet-50. This model has been tested on competitive evaluations and public benchmarks [34]. ResNet-50 is a convolutional neural network with 50 layers and 41M parameters initially proposed for general purpose image recognition tasks [41]. The main difference with traditional convolutional neural networks is the inclusion of residual connections to allow information skip layers and improve gradient flow.

Our experiments include a ResNet-50 model trained from scratch using VGGFace2 dataset [34]. Before applying the proposed method, we cropped the face images using the algorithm proposed in [14]. The pre-trained model is used as embedding extractor. Those embeddings are then l_2-normalised to generate our input representation $\mathbf{x}$. The similarity between two face descriptors $(\mathbf{x}^i, \mathbf{x}^j)$ is calculated as the Euclidean distance $\|\mathbf{x}^i - \mathbf{x}^j\|$. The verification accuracy is obtained comparing the distances between positive matches ($\mathbf{x}^i$ and $\mathbf{x}^j$ belong to the same identity) with negative matches ($\mathbf{x}^i$ and $\mathbf{x}^j$ belong to different identities). Two face descriptors are assigned to the same identity if their distance is smaller than a threshold τ. The pre-trained model used in this work achieved a verification accuracy (test set from view 1 experimental protocol) of 98.4% on the LFW benchmark [35].

4.2 Experiment 1: Sensitive information in deep learning face descriptors

The first experiment aims to demonstrate the high level of sensitive information that form part in decision-making of state-of-the-art face recognition algorithms. Using the pre-trained model described in Section 4.1, we trained a classification layer (softmax function) composed by two or three neurons (for gender or ethnicity respectively). We used 9000 and 1800 images from DiveFace dataset for training and testing respectively (images and identities not employed during training). We kept frozen the parameters of the pre-trained models to train only the parameters of the classification layer. Implementation details: 150 epochs, Adam optimizer (learning rate=0.001, $\beta_1 = 0.9$, and $\beta_2 = 0.999$), and batch size of 128 samples. In order to demonstrate the high presence of sensitive information in the embedding generated by the pre-trained model, we report the classification accuracy for an increasing number of features removed before the classification.

Fig. 4 shows results for gender and ethnicity classification. Results show how it is possible to accurately classify both gender and ethnicity even with only 10% of the features from the pre-trained model. It is important to highlight that Resnet-50 was trained for face verification, not gender or ethnicity classification. Although this model was trained for person recognition, sensitive information is deeply embedded in its feature representation. According to these results, we can argue that gender and ethnicity features are present in the decision-making of a system based on this popular face recognition model.

4.3 Experiment 2: Removing sensitive information from face descriptors

The learning method proposed in Section 2 for obtaining the mapping function $\varphi(\mathbf{x})$ is trained using two different subsets of DiveFace. The sensitivity detector is trained with 3K different identities (3 images per identity) balanced between genders and ethnic groups. The list of triplets $\mathbf{T}$ are generated with the remaining 21K identities (all images available per identity) according to the equation (1) with $\alpha = 0.2$ and $l = 1024$.

The aim of the proposed method is to maintain the face recognition performance while removing the sensitive information considered (gender and ethnicity). To analyze the effectiveness of the proposed method, we conducted two experiments including two datasets not used during

TABLE I. CLASSIFICATION ACCURACIES FOR EACH TASK BEFORE AND AFTER APPLYING THE PROJECTION INTO THE NEW FEATURE REPRESENTATION. THE REDUCTION COLUMN SHOWS THE RELATIVE DIFFERENCE IN ACCURACY BEFORE AND AFTER PROJECTION WITH RESPECT TO RANDOM CHANCE. IDENTITY_G AND IDENTITY_E REPRESENT IDENTITY VERIFICATION ACCURACIES WHEN GENDER AND ETHNICITY ARE REMOVED RESPECTIVELY.

Task	Before	After	Reduction*	Random
Identity_G	98.4%	96.8%	3.3%	50%
Identity_E		96.2%	4.5%	
Neural Network				
Gender	97.7%	58.8%	81.5%	50%
Ethnicity	98.8%	55.1%	66.4%	33%
Support Vector Machine				
Gender	96.2%	56.3%	86.4%	50%
Ethnicity	98.2%	54.1%	67.6%	33%
Random Forest				
Gender	95.1%	54.6%	89.8%	50%
Ethnicity	97.3%	53.5%	68.1%	33%

**Reduction = (Before-After)/(Before-Random)*

the training phase of the agnostic features:

a) Verification accuracy: we calculated the face verification accuracy using either original embeddings $\mathbf{x}$ or their projections $\varphi(\mathbf{x})$ according to the evaluation protocol of the popular benchmark of LFW [35]. Table I and Fig. 5 (top) show the accuracies of embeddings generated by the pre-trained model before and after the projection. The results show a very small drop of performance when the projection is applied, which demonstrates the success of our method in preserving the accuracy in the main task here, i.e., face verification.

b) Removing sensitive information: we train different gender and ethnicity classification algorithms (Neural Networks, Support Vector Machines and Random Forests) either on original embeddings $\mathbf{x}$ or on their projections $\varphi(\mathbf{x})$. The algorithms were trained and tested with 9000 and 1800 images respectively (same set employed in Section 4.3). Fig. 5 (middle and bottom) and Table I show the accuracies obtained by each classification algorithm before and after the projections. Results show a quite significant drop of performance in both gender and ethnicity classification when the proposed representation is applied, which demonstrates the success of our proposed approach in removing the considered sensitive information (gender and ethnicity in this case) from the embeddings.

4.4 Discussion

The sensitive information is extracted using an iterative process with different layers because of the high prevalence of this information in the original embeddings. For the problem experimentally addressed here (i.e., face recognition using a gender- and ethnicity-blind representation based on state-of-the-art deep networks and datasets), we have observed that it is necessary at least n=4 layers to make the considered models agnostic.

One of the questions that motivated this work was to analyze how important are sensitive features such as gender

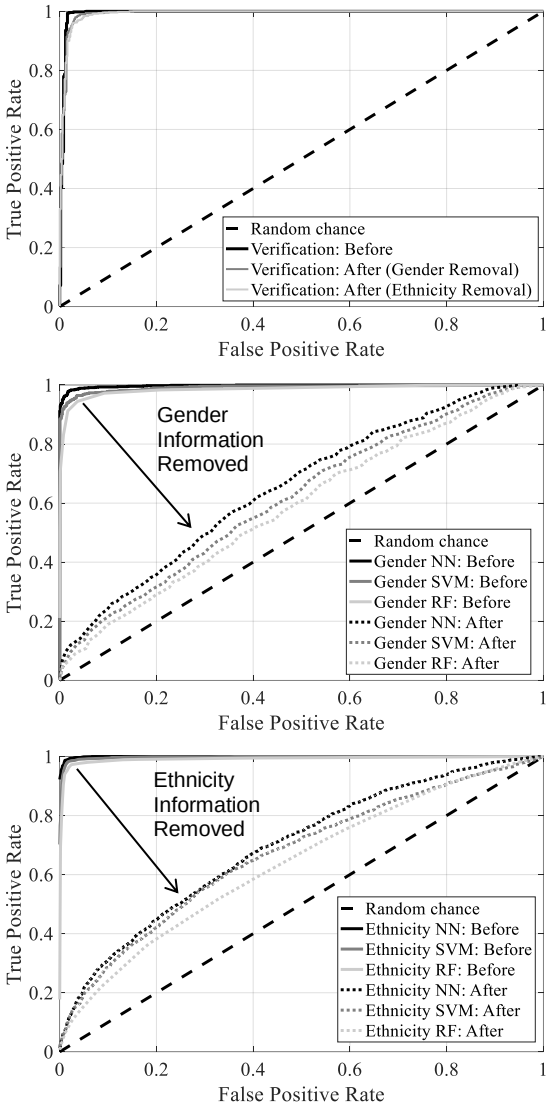


Fig. 5. ROC curves for different tasks before and after applying the proposed embedding projection aimed at reducing sensitive features (gender and ethnicity in our experiments). Verification accuracy (top), gender classification accuracy (middle) and ethnicity classification accuracy (bottom).

or ethnicity for person recognition based on face biometrics. Gender and ethnicity may help to identify a person, but our results suggest that this information is not critical for identification.

On a different front, domain adaptation methods are also capable to minimize the contribution of selected covariates [10][27]. However, existing works studying that line are limited to classification problems with few classes and a large number of samples per classes. The application of domain adaptation methods for generating agnostic representations in person recognition is a subject for future research. The large number of classes and reduced number of samples make this problem quite challenging for domain adaptation methods [10][27].

5 Conclusions

This work has proposed a new general representation trained to eliminate sensitive information from decision-making of recognition algorithms that employ deep neural networks embeddings. Our representation is applicable to any pattern recognition and machine learning problem, but we have studied it in the framework of face recognition. Sensitive information such as gender or ethnicity is highly embedded in the feature space used by most approaches in face recognition.

We propose an iterative learning method developed to maintain recognition performance while minimizing the contribution of selected covariates such as gender and ethnicity. The method is based on a triplet loss generalization.

Additionally, we make available a new annotation database very useful to train unbiased and discrimination-aware face recognition algorithms. The database comprises labels from more than 150K images and 30K identities. Our proposed method for generating agnostic representations is evaluated using two popular databases. The results show how it is possible to reduce the performance of gender and ethnicity detectors by 60-80% while the face verification performance is only reduced by 3-4%.

With more and more algorithms participating in the decision-making of human lives, if we want to guarantee trust of users it is important to promote a new discrimination-aware machine learning. This is especially important in pattern recognition fields such as biometrics, where better discrimination-aware and privacy-preserving methods [42] are very much needed [43].

Acknowledgment

This work was supported in part by the Project CogniMetrics, under Grant TEC2015-70627-R and Bio-Guard (Ayudas Fundación BBVA a Equipos de Investigación Científica 2017). Spanish Patent Application (P201831278).

References

[1] K. Conger, R. Fausset, S. F. Kovaleski, "San Francisco Bans Facial Recognition Technology", New York Times, May 2019.

[2] Iyad Rahwan, Manuel Cebrian, et al., "Machine behaviour", *Nature,* vol. 568, pp. 477–486. April 2019.

[3] A. K. Jain, K. Nandakumar, A. Ross, "50 years of Biometric Research: Accomplishments, Challenges, and Opportunities", *Pattern Recognition Letters*, vol. 79, pp. 80-105, 2016.

[4] H. Proenca, M. Nixon, M. Nappi, E. Ghaleb, G. Ozbulak, H. Gao, H. K. Ekenel, K. Grm, V. Struc, H. Shi, X. Zhu, S. Liao, Z. Lei, S. Z. Li, W. Gutfeter, A. Pacut, J. Brogan, W. J. Scheirer, E. Gonzalez-Sosa, R. Vera-Rodriguez, J. Fierrez, J. Ortega-Garcia, D. Riccio and L. De Maio, "Trends and Controversies", *IEEE Intelligent Systems*, vol. 33, no. 3, pp. 41-67, 2018.

[5] B. Goodman and F. Flaxman, "European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation"", *AI Magazine*, vol. 38, no. 3, 2016.

[6] J. Kleinberg, J. Ludwig, S. Mullainathan, C. Sunstein, "Discrimination in the Age of Algorithms". *Journal of Legal Analysis*, 2019. Also appears as NBER Working Paper Number 25548, February 2019.

[7] J. Kleinberg, S. Mullainathan, "Simplicity Creates Inequity: Implications for Fairness, Stereotypes, and Interpretability". *Proc. 20th ACM Conference on Economics and Computation* (EC), 2019.

[8] A. D. Selbst, et al. "Fairness and abstraction in sociotechnical systems", *Proc. of the Conference on Fairness, Accountability, and Transparency*. ACM, 2019.

[9] J. Buolamwini and T. Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification", *Proc. of the ACM Conf. on Fairness, Accountability, and Transparency*, New York, USA, pp. 81:1-15, 2018.

[10] M. Alvi, A. Zisserman, C. Nellaker, "Turning a Blind Eye: Explicit Removal of Biases and Variation from Deep Neural Network Embeddings", *Proc. of European Conf. on Computer Vision*, Munich, Germany, 2018.

[11] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, K.W. Chang, "Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints", *Proc. of the Conf. on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, pp. 2979-2989, 2017.

[12] EU 2016/679 (General Data Protection Regulation). Available online at: https://gdpr-info.eu/, last accessed 2018/11/13.

[13] A. Acien, A. Morales, R. Vera-Rodriguez, I. Bartolome, J. Fierrez, "Measuring the Gender and Ethnicity Bias in Deep Models for Face Recognition, *Proc. of IAPR Iberoamerican Congress on Pattern Recognition*, Madrid, Spain, 2018.

[14] K. Zhang, Z. Zhang, Z. Li, Y. Qiao, "Joint Face Detection and Alignment using Multi-task Cascaded Convolutional Networks", *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499-1503, 2016.

[15] B. Berendt and S. Preibusch, "Exploring Discrimination: A User-centric Evaluation of Discrimination Aware Data Mining", *Proc. of the IEEE Int. Conf. on Data Mining Workshops*, Brussels, Belgium, pp. 344-351, 2012.

[16] B. Berendt and S. Preibusch, "Better Decision Support through Exploratory Discrimination-Aware Data Mining: Foundations and Empirical Evidence", *Artificial Intelligence and Law*, vol. 22, no. 2, pp. 175-209, 2014.

[17] S. Hajian, J. Domingo-Ferrer, A. Martinez-Ballester, "Discrimination Prevention in Data Mining for Intrusion and Crime Detection", *Proc. of IEEE Symposium on Computational Intelligence in Cyber Security*, Paris, France, 2011.

[18] T. Kehrenberg, Z. Chen, N. Quadrianto, "Interpretable Fairness via Target Labels in Gaussian Process Models", *arXiv:1810.05598*, 2018.

[19] T. Calders, S. and Verwer "Three Naive Bayes Approaches for Discrimination-Free Classification", *Data Mining and Knowledge Discovery*, vol. 21, no. 2, pp. 277-292, 2010.

[20] C. Sandvig, K. Hamilton, K. Karahalios, C. Langbort, "Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms", *Proc. of the Annual Meeting of the International Communication Association*, Seattle, USA, 2014.

[21] E. Raff and J. Sylvester "Gradient Reversal against Discrimination", *Proc. of Workshop on Fairness, Accountability, and Transparency in Machine Learning*, Stockholm, Sweden, 2018.

[22] S. Jia, T. Lansdall-Welfare, N. Cristianini "Right for the Right Reason: Training Agnostic Networks", *Proc. of Int. Symposium on Intelligent Data Analysis*, Hertogenbosch, Netherlands, pp. 164-174, 2018.

[23] A. Torralba and A. A. Efros "Unbiased Look at Dataset Bias", *Proc. of IEEE Int. Conf. on Computer Vision and Pattern Recognition*, Colorado Springs, USA, pp. 1521-1528, 2011.

[24] B. F. Klare, M. J. Burge, J. C. Klontz, R. W. V. Bruegge, A. K. Jain "Face Recognition Performance: Role of Demographic Information", *IEEE Trans. on Information, Forensics and Security*, vol. 7, no. 6, pp. 1789-1801, 2012.

[25] M. Orcutt "Are Face Recognition Systems Accurate? Depends on Your Race", *MIT Technology Review*, 2016.

[26] A. Das, A. Dantcheva , F. Bremond "Mitigating Bias in Gender, Age, and Ethnicity Classification: a Multi-Task Convolution Neural Network Approach", *Proc. of European Conf. on Computer Vision Workshops*, Munich, Germany, 2018.

[27] E. Tzeng, J. Hoffman, T. Darrell, K. Saenko " Simultaneous Deep Transfer Across Domains and Tasks", *Proc. of IEEE Int. Conf. on Computer Vision*, Santiago, Chile, pp. 4068-4076, 2015.

[28] V. Mirjalili, S. Raschka, A. Ross, "Gender Privacy: An Ensemble of Semi Adversarial Networks for Confounding Arbitrary Gender Classifiers", *Proc. of IEEE 9th International Conference on Biometrics: Theory, Applications and Systems*, Los Angeles, USA, 2018.

[29] V. Mirjalili, S. Raschka, A. Namboodiri, A. Ross "Semi-Adversarial Networks: Convolutional Autoencoders for Imparting Privacy to Face Images", *Proc. of IAPR Int. Conf. on Biometrics*, Gold Coast, Australia, 2018.

[30] O. M. Parkhi, A. Vedaldi, A. Zisserman, "Deep Face Recognition", *Proc. of British Machine Vision Conf.*, Swansea, UK, 2015.

[31] F. Schroff, D. Kalenichenko, J. Philbin "FaceNet: A Unified Embedding for Face Recognition and Clustering", *Proc. of the IEEE Int. Conf. on Computer Vision and Pattern Recognition*, Boston, USA, pp. 815-823, 2015.

[32] I. Kemelmacher-Shlizerman, S. Seitz, D. Miller, E. Brossard "The MegaFace Benchmark: 1 Million Faces for Recognition at Scale", *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, Las Vegas, USA, pp. 4873-4882, 2016.

[33] S. Yang, P. Luo, C. C. Loy, X. Tang " From Facial Parts Responses to Face Detection: A Deep Learning Approach", *Proc. of IEEE Int. Conf. on Computer Vision*, Santiago, Chile, pp. 3676-3684, 2015.

[34] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, A. Zisserman, "VGGFace2: A Dataset for Recognising Face Across Pose and Age", *Proc. Int. Conf. on Automatic Face and Gesture Recognition*, Xian, China, pp. 67-74, 2018.

[35] E. Learned-Miller, G. B. Huang, A. RoyChowdhury, H. Li, G. Hua, "Labeled Faces in the Wild: A Survey", *Advances in Face Detection and Facial Image Analysis*, Michal Kawulok, M. Emre Celebi, and Bogdan Smolka eds., Springer, pp. 189-248, 2016.

[36] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, L.-J. Li, "The New Data and New Challenges in Multimedia Research", arXiv preprint arXiv:1503.01817, 2015.

[37] E. Gonzalez-Sosa, J. Fierrez, R. Vera-Rodriguez, F. Alonso-Fernandez, "Facial Soft Biometrics for Recognition in the Wild: Recent Works, Annotation and COTS Evaluation", *IEEE Trans. on Information Forensics and Security*, vol. 13, no. 7, pp. 2001-2014, 2018.

[38] R. Ranjan, S. Sankaranarayanan, A. Bansal, N . Bodla, J.-C. Chen, V.M. Patel, C.D. Castillo, R.Chellappa, "Deep Learning for Understanding Faces: Machines May Be Just as Good, or Better, than Humans", *IEEE Signal Processing Magazine*, vol. 35, pp. 66-83, 2018.

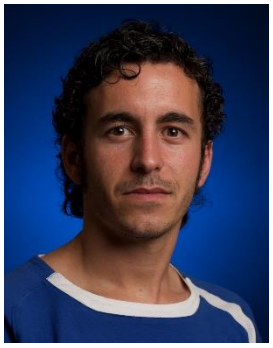
[39] Z. Liu, P. Luo, X. Wang, X. Tang, "Deep Learning Face Attributes in the Wild", *Proc. of Int. Conf. on Computer Vision*, Santiago, Chile, pp. 3730-3738, 2015.

[40] P. Tome, J. Fierrez, R. Vera-Rodriguez, M. Nixon "Soft Biometrics and their Application in Person Recognition at a Distance", *IEEE Trans. on Information Forensics and Security*, vol. 9, no. 3, pp. 464-475, 2014.

[41] K. He, X. Zhang, S. Ren, J. Sun, "Deep Residual Learning for Image Recognition", *Proc. of IEEE Int. Conf. on Computer Vision and Pattern Recognition*, Las Vegas, USA, pp. 770–778, 2016.

[42] M. Gomez-Barrero, J. Galbally, A. Morales, J. Fierrez, "Privacy-Preserving Comparison of Variable-Length Data with Application to Biometric Template Protection", *IEEE Access*, vol. 5, pp. 8606-8619, 2017.

[43] J. Fierrez, A. Morales, R. Vera-Rodriguez, D. Camacho, "Multiple Classifiers in Biometrics. Part 2: Trends and Challenges", *Information Fusion*, vol. 44, no. 103-112, 2018.



Aythami Morales Moreno All received the M.Sc. degree in telecommunication engineering from the Universidad de Las Palmas de Gran Canaria in 2006 and the Ph.D. degree from La Universidad de Las Palmas de Gran Canaria in 2011. Since 2017, he is Associate Professor with the Universidad Autonoma de Madrid. He has conducted research stays at the Biometric Research Laboratory, Michigan State University, the Biometric Research Center, Hong Kong Polytechnic University, the Biometric System Laboratory, University of Bologna, and the Schepens Eye Research Institute. He has authored over 70 scientific articles published in international journals and conferences. He has participated in national and EU projects in collaboration with other universities and private entities, such as UAM, UPM, EUPMt, Indra, Union Fenosa, Soluziona, or Accenture. His research interests are focused on pattern recognition, computer vision, machine learning, and biometrics signal processing. He has received awards from the ULPGC, La Caja de Canarias, SPEGC, and COIT.

Julian Fierrez received the M.Sc. and Ph.D. degrees in electrical engineering from the Universidad Politecnica de Madrid, Spain, in 2001 and 2006, respectively. Since 2002, he has been with the Biometrics and Data Pattern Analytics LabATVS, Universidad Autonoma de Madrid, where he has been an Associate Professor since 2010. From 2007 to 2009, he held a Postdoctoral position at Michigan State University under a Marie Curie Fellowship. He has been actively involved in the last 15 years in large EU projects focused on biometrics (such as BIOSECURE, TABULA RASA, and BEAT). He was a recipient of a number of distinctions, including the EAB European Biometric Industry Award 2006, the EURASIP Best Ph.D. Award 2012, and the IAPR Young Biometrics Investigator Award 2017. Since 2016, he has been an Associate Editor of the IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY and the IEEE Biometrics Council Newsletter.

Ruben Vera-Rodriguez received the M.Sc. degree in telecommunications engineering from Universidad de Sevilla, Spain, in 2006, and the PhD degree in electrical and electronic engineering from Swansea University, U.K., in 2010. Since 2010, he has been with the Biometric Recognition Group, Universidad Autonoma de Madrid, Spain, first as the recipient of a Juan de la Cierva Post-Doctoral Fellowship from the Spanish Ministry of Innovation and Sciences, and as an Assistant Professor since 2013. His research interests include signal and image processing, pattern recognition, and biometrics, with emphasis on signature, face and gait verification and forensic applications of biometrics. Dr Vera-Rodriguez is actively involved in several National and European projects focused on biometrics. He was the recipient of the best paper award at the 4th International Summer School on Biometrics, Alghero, Italy, in 2007.