Machine Vision and Applications (2020) 31:67
https://doi.org/10.1007/s00138-020-01123-z

ORIGINAL PAPER

Benchmarking deep network architectures for ethnicity recognition using a new large face dataset

Antonio Greco[1] · Gennaro Percannella[1] · Mario Vento[1] · Vincenzo Vigilante[1]

Received: 25 November 2019 / Revised: 10 June 2020 / Accepted: 1 September 2020 / Published online: 14 September 2020

Abstract
Although in recent years we have witnessed an explosion of the scientific research in the recognition of facial soft biometrics such as gender, age and expression with deep neural networks, the recognition of ethnicity has not received the same attention from the scientific community. The growth of this field is hindered by two related factors: on the one hand, the absence of a dataset sufficiently large and representative does not allow an effective training of convolutional neural networks for the recognition of ethnicity; on the other hand, the collection of new ethnicity datasets is far from simple and must be carried out manually by humans trained to recognize the basic ethnicity groups using the somatic facial features. To fill this gap in the facial soft biometrics analysis, we propose the VGGFace2 Mivia Ethnicity Recognition (VMER) dataset, composed by more than 3,000,000 face images annotated with 4 ethnicity categories, namely African American, East Asian, Caucasian Latin and Asian Indian. The final annotations are obtained with a protocol which requires the opinion of three people belonging to different ethnicities, in order to avoid the bias introduced by the well-known other race effect. In addition, we carry out a comprehensive performance analysis of popular deep network architectures, namely VGG-16, VGG-Face, ResNet-50 and MobileNet v2. Finally, we perform a cross-dataset evaluation to demonstrate that the deep network architectures trained with VMER generalize on different test sets better than the same models trained on the largest ethnicity dataset available so far. The ethnicity labels of the VMER dataset and the code used for the experiments are available upon request at https://mivia.unisa.it.

Keywords Ethnicity recognition · Face analysis · Soft biometrics · Dataset · Deep learning · Benchmark

1 Introduction

The face is the part of the human body that contains most of the semantic information about an individual; the so-called facial soft biometrics, namely identity, gender, age, ethnicity, expression, have attracted in recent years the attention of the pattern recognition community thanks to the huge amount of possible applications in retail and video surveillance and to the intrinsic difficulty of designing effective and reliable algorithms in the challenging real-world scenarios. This trend is confirmed by the large amount of papers [10] describing the use of modern convolutional neural networks (CNNs) for solving problems as face recognition and verification [12], gender recognition [3,7,36], age estimation [6,13] and expression recognition [26,34].

In spite of this, ethnicity recognition, namely the capability of a system to determine whether an individual belongs to one of the $E = e_1, \ldots, e_E$ ethnicity groups according to facial appearance observations like skin color, morphology and other explicit patterns, has not received the same attention from the scientific community. The interest for ethnicity recognition is surely growing, considering that new methods and datasets [2,20,24,48] have been recently proposed to improve the accuracy of real applications currently achieving a performance biased by the ethnicity (face detection and recognition, gender classification, age estimation) or to give a definitive push to applications in forensics (ethnicity-

✉ Antonio Greco
agreco@unisa.it

Gennaro Percannella
pergen@unisa.it

Mario Vento
mvento@unisa.it

Vincenzo Vigilante
vvigilante@unisa.it

[1] Department of Information and Electrical Engineering and Applied Mathematics, University of Salerno, Fisciano, Italy

based subject identification for public safety). Nevertheless, the authors of a recent comprehensive survey [16] state that the growth of this research topic is mainly hindered by the lack of ethnicity data; in fact, in the era of deep learning it is necessary to have a large amount of data available to effectively train convolutional neural networks. Currently, there are no datasets for ethnicity recognition with a size comparable to the largest datasets available for the other facial soft biometrics [41]. To seal this observation, it has been recently shown [24] that the CNNs trained for ethnicity recognition on the currently available datasets have a limited capability to generalize on different test sets.

The lack of ethnicity data is mainly due to the intrinsic difficulties related to their collection and annotation procedure. In fact, the concept of ethnicity is rather controversial: unlike other kinds of biometrics, for example gender, it can be defined qualitatively and not quantitatively, being complex the identification of universal distinguishing features. It has been demonstrated [31] that the "ethnicity", as intended by humans, has no biological validity, since there are no genetic characteristics that allow individuals to be grouped according to the commonly defined "ethnicities". Therefore, the categorization is done according to visual differences in the somatic facial features universally recognized by humans. It implies that an automatic annotation procedure using, for example, the place of birth cannot be defined; the groundtruths of the ethnicity groups must be manually done by human annotators, and the reliability of the ethnicity labels strongly depends on the ability of the annotator.

We describe the existing datasets and methods in the following subsections, discussing the issues related to the ethnicity annotation and the limited generalization capabilities achievable with the currently available datasets.

1.1 Existing datasets

The publicly available datasets for ethnicity recognition, summarized in Table 1, have three main drawbacks.

First, the ethnicity groups are completely different among the datasets. This is often a direct consequence of the absence of a standard categorization, while sometimes the difference is due to the specificity of the application context: for example many datasets contain faces belonging to a single macro-ethnicity (e.g., Chinese, Brazilian, Japanese, Iranian, Saudi Arabia); of course, they cannot be used for a general ethnicity recognition system, but are eventually integrated into larger datasets. In other cases, namely CAFE, FERET, PubFig, EGA, LFWA+, UTKFace and FairFace, the annotations are more heterogeneous but, anyway, inconsistent with each other, so making the same experiments not always reproducible on different benchmarks.

Second, the datasets are composed by few thousands of images and, consequently, are not large enough (or not enough for all the ethnicities) to train deep networks. FERET [35] is surely the most popular dataset for training and testing ethnicity classifiers, but it does not provide an official division in training and test set; therefore, the experiments performed by different authors are not easily reproducible. MORPH-II only contains enough examples for two classes (European and Asian), while the remaining classes only include the 4% of the total images. EGA [38], collected fusing 6 already existing datasets (including FERET), is among the most widely used benchmark for ethnicity recognition since it includes more than 70,000 samples; however, the reduced number of subjects is a significant limitation for training modern deep networks. LFWA+ and UTKFace have been widely adopted in recent years for training and testing modern neural networks, since they are annotated at least with the most common labels (White, Black, Asian); nevertheless, they consist of a few thousands of images and cross-dataset experiments demonstrated that the deep networks trained with their images are not able to generalize on different test sets [24]. FairFace is probably the most complete dataset currently available for ethnicity recognition, since it is composed by 108,192 images, already divided in training and test sets. It is balanced, in the sense that the samples are almost equally distributed among the 7 different ethnicity groups. The tests performed in [24] show that a network trained with FairFace generalizes better than the same network trained with LFWA+ and UTKFace.

Third, the annotation procedures typically do not take into account the Other Race Effect (ORE). It has been scientifically established that in face analysis tasks the humans perform significantly better when dealing with faces of people of their own ethnicity than with individuals belonging to other ethnicities; for this reason, this effect is also called own race bias [16]. Therefore, involving people of different ethnicities in the annotation procedure is the only way to achieve independence from the race bias. To the best of our knowledge, this aspect is not taken into account in the existing datasets.

The review of the literature points out the need to collect a sufficiently large and heterogeneous dataset for ethnicity recognition, defining and applying a clear and reproducible standard, which takes into account the ORE, for defining and annotating the ethnicity groups.

1.2 State-of-the-art approaches

In this section, we focus our attention on the automatic ethnicity recognition approaches and their results. The common solutions focus on two, three or four ethnicity groups; typical choices are among Caucasian (or White), African American (or Black), East Asian (or Asian), Asian Indian (or Indian) and Latin (or Hispanic) categories.

Table 1 Publicly available datasets of face images with ethnicity groups

Dataset	Images (subjects)	Ethnicity groups
FERET [35]	14,126 (1199)	Caucasian, Asian, Oriental African
JAFFE [30]	2130 (10)	Japanese
IFDB [4]	3600 (616)	Iranian
CASPEAL [18]	30,900 (1040)	Chinese
MORPH-II [37]	55,134 (13,618)	African, European, Asian, Hispanic, Others
FEI [43]	2800 (200)	Brazilian
PubFig [25]	58,797 (200)	Asian, Caucasian, African American, Indian
CUN [17]	112,000 (1120)	Chinese
HUDA [47]	N/A	Saudi Arabia
EGA [38]	72,266 (469)	African American, Asian, Caucasian, Indian, Latin
CAFE [29]	1192 (154)	Caucasian, East Asians, Pacific Region
LFWA+ [28]	13,233 (5749)	White, Black, Asian
UTKFace [48]	20,000 (N/A)	White, Black, Asian, Indian, Others
FairFace [24]	108,192 (N/A)	White, Black, East Asian, Southeast Asian, Indian, Middle Eastern, Latin

The list does not include databases of 3D faces

Most available methods are mainly based on handcrafted features. Such methods specifically design a procedure that encapsulates in a face descriptor the properties that allow to distinguish ethnicity groups from one another. The commonly considered traits encoded in the descriptor include the color of the skin, the shape of the eyes, the facial landmarks and so on; the obtained feature vector is then fed to a classifier, which predicts the ethnicity. Between the newer methods, there are some approaches based on automatic representation learning, such as CNNs; those methods perform better when trained with large quantities of representative data and typically achieve the best accuracy and generalization capability when this condition is respected.

The skin color is the most popular feature to recognize the ethnicity; color values or color histograms are used as feature vectors to train a SVM classifier in [11] obtaining a result of 78.5% on the FERET dataset, including Black, White and Asian classes. The authors of [25] also use color as a base feature and develop attribute prediction for multiple attributes on their own PubFig dataset. In other cases, a feature selection step is implemented, using algorithms like KCFA [45] or Adaboost [39]. However, these types of features are not invariant to illumination, which can substantially affect the performance in real environments.

Other approaches are instead based on the use of texture or shape descriptors (or a combination of them) to detect the facial ethnicity differences that are unrelated to the skin color. For example, Wu et al. [44] use Haar features and Adaboost and experiment on their own private database with 3 classes, while Hosoi et al. [22] and Lin et al. [27] compute a face descriptor based on Gabor features, eventually selected with Adaboost [27], and perform the ethnicity classification with a SVM. The methods in [40] and [32] rely on the use of LBP histograms and a KNN classifier, but the former computes the face descriptor selecting the most discriminant LBP and Haar features with a PCA, while the latter combines a descriptor based on LBP with a Weber Local Descriptor (WLD) [8]. Many of these works collect their own datasets due to the drawbacks of the existing ones.

The best results, as expected, are mostly obtained with automatically learned features. In [1], the authors successfully train a CNN for different face-related tasks, including ethnicity recognition. They make large use of data augmentation for training their network and yield an accuracy of 93.9% on the FERET dataset with 3 classes (White, Asian, Other).

On the same dataset and classes, much more recently, an accuracy of 98.9% has been achieved [2]; the method is based on a fine-tuned VGG-Face used as a feature extractor on input faces normalized using canonical alignment; the feature vector is then fed to a linear SVM for performing the ethnicity classification.

Yi et al. [46] extract multiple patches from the aligned face image at 4 different scales and applies 23 different shallow multi-task CNNs to classify them, fusing the decision in the output layer, which provides both the ethnicity group and the age estimation. The resulting accuracy on the MORPH-II dataset is 99.11%. The authors only use Black and White as classes and ignore Asian, Hispanic and Other since those first two classes represent the 96% of the images.

On the same dataset, Hu et al. [20] obtain a result of 98.6%. The authors expand the annotation of the popular LFW+ dataset for different facial attributes, including eth-

nicity. The network architecture is a multi-task version of AlexNet.

Guo et al. [19] also realize a multi-task classifier: the faces are detected and aligned, cropped and resized to 60 × 60 and then used in grayscale to extract using "BIF" features, that are biologically inspired. Exploiting a feature selection approach, they can distinguish Black from White people with 99% accuracy.

Karkkainen et al. [24] evaluate the ethnicity recognition performance of a ResNet-34 model trained using different training sets. The experimental protocol involves evaluations on different test sets, in order to investigate the generalization capabilities achieved by the network trained with a specific set. The experiments show that the ResNet-34 trained with FairFace generalizes substantially better than the same model trained with UTKFace and LFWA+, so demonstrating the importance of the dataset adopted as training set.

From the analysis of the state of the art, we notice three main things: (i) recent literature mostly considers the problem of ethnicity recognition in multi-task settings, as an addition to other soft biometric tasks, or to evaluate the impact of the ethnicity on other attributes such as gender recognition and age estimation; (ii) the application of many modern methods, including CNN architectures, has saturated the capabilities of datasets acquired in controlled laboratory conditions as FERET and MORPH-II; (iii) among the currently available datasets, only FairFace is able to provide the network models with generalization capabilities, even if a substantial performance variation is observed on different test sets.

For all these considerations, a public large and challenging dataset, reliably annotated with the most common ethnicities, can surely be a panacea for training and benchmarking new approaches. We believe that there is space for further improvement of existing methods for ethnicity recognition, increasing their accuracy in real-world situations ("in-the-wild").

1.3 Contributions

We propose in this paper a new set of labels for ethnicity recognition. In particular, we annotated more than 3 millions images of 9129 identities publicly available in the VGGFace2 dataset [5] with 4 ethnicity groups, namely African American (AA), East Asian (EA), Caucasian Latin (CL) and Asian Indian (AI) and we make the whole benchmark publicly available with the name VGGFace2 Mivia Ethnicity Recognition (VMER) dataset. To avoid the bias possibly introduced by the other race effect, we asked the opinions of three people belonging to different ethnicities, namely one African American, one Caucasian Latin and one Asian Indian, choosing the final ethnicity group through a majority voting. The opinion of a fourth annotator has been required in case of a tie.

The utility for the scientific research is at least threefold: (i) the high number of samples available in the dataset allows the training of CNN networks, including larger architectures, that are more prone to overfitting; (ii) the annotation protocol, designed for reducing the own race bias, ensures the accuracy of the ethnicity labels; (iii) the availability of other annotations, namely identity and gender, makes the dataset suited for future advanced analyses, such as multi-task learning or forensics applications.

As a matter of fact, we use this dataset for training modern deep network architectures, as MobileNet v2, ResNet-50, VGG-16 and VGG-Face, obtaining more than 94% of accuracy on the test set. In addition, following on the experiments carried out in [24], we perform a cross dataset evaluation demonstrating that neural networks (ResNet-34 and VGG-Face) trained with VMER are able to better generalize on a different test set (UTKFace) with respect to the same networks trained with FairFace, so confirming the effectiveness of the new set of labels. We also visualize the features learned by a CNN trained with VMER and demonstrate that they correspond to distinctive facial traits typically adopted also by humans.

We consider our results as a baseline of the performance achievable with the modern deep network architectures and assume that this contribution can pave the way for future experiments and applications in this research field.

1.4 Organization of the paper

The paper is organized as follows: in Sect. 2, we describe the dataset, giving details about the available face images, the characteristics of the considered ethnicity categories and the pre-processing of the face images; in Sect. 3, we report the results of our experimental analysis; finally, in Sect. 4 we draw the conclusions, defining the possible future directions of the research in this field.

2 VMER dataset

2.1 Description

The proposed VMER dataset is composed by images collected from the original VGGFace2 [5], which is so far the largest face dataset in the world including more than 3.3 millions face images, with an average of about 362 samples per subject (minimum 87 images per subject). It also includes gender labels and consists of 62% males and 38% females.

The images in the dataset have been acquired in different lighting and occlusion conditions, and the faces of the subjects are characterized by different pose, age, ethnicity and size. In particular, more than 75% of the faces have a resolu-

Table 2 Number of images and subjects for each ethnicity available in the VMER dataset, already divided in training and test set

Ethnicity	Training Images (Subjects)	Test Images (Subjects)	Training %	Test %
African American	242,783 (712)	10,373 (34)	7.7	6.1
East Asian	187,893 (533)	18,750 (62)	6.0	11.1
Caucasian Latin	2,507,837 (6854)	130,900 (380)	79.9	77.4
Asian Indian	202,205 (530)	9001 (24)	6.4	5.3
Total	3,140,718 (8629)	169,024 (500)	3,309,742 (9129)	

Fig. 1 Samples of African American, East Asian, Caucasian Latin and Asian Indian people available in the VMER dataset

tion between 50×50 and 200×200 pixels, which is less than the input size of most of the popular CNNs. It is important to take into account this aspect when dealing with these face images, as we will show in our experimental analysis.

2.2 Ethnicity annotation

The categorization of the ethnicity is a task anything but simple even for a human, as witnessed by the scientific literature in this field [16]; imagine how complex this classification can be for a computer vision algorithm, which can only make use of color, texture, morphological features and craniofacial measurements that can be automatically extracted from a face image. As extensively discussed in Sect. 1, the ethnicity annotation requires a manual procedure that takes into account the somatic facial features which a human uses to distinguish the ethnicity categories.

According to the most recent trends, we choose to divide our dataset into the following four categories:

- African American (AA): the individuals of this ethnicity group typically have African, North American or South American origins and are characterized by dark skin color, full lips and wide nose.
- East Asian (EA): people belonging to this group have Chinese or other East and South East Asian origins. Their color skin is light, with shades from white to yellowish, and small nose, but the most distinctive feature is the almond shape of the eyes and the inclination between the medial and the lateral canthus, which make the eye look narrower.
- Caucasian Latin (CL): humans of such ethnicity have European, South American, Western Asian and North African origins and are characterized by a white or tanned skin, medium nose and lips and horizontally aligned eyes.
- Asian Indian (AI): folk belonging to this ethnicity group have Indian, South Asian or Pacific Island origins. They have characteristics in common with EAs and CLs, but we can distinguish them by noting very slight differences. They have a slightly darker skin color and eyes with more defined features with respect to EAs and CLs.

Examples of face images belonging to the four ethnicity categories are depicted in Fig. 1.

In order to avoid the other race effect, we asked three people belonging to different ethnicities, namely one African American, one Caucasian Latin and one Asian Indian, to annotate each identity with an ethnicity label among the considered four.

The results of the annotations fully confirm the importance of consulting multiple annotators. In fact, the three people fully agreed on only 85% of the dataset (7779 identities); in 14% of the cases (1278 identities), only two of the annotators gave unanimous labels, while in the remaining 1% (74 identities) they all provided conflicting opinions. The inter-rater agreement, computed with the Cohen's Kappa [9], is equal to 0.74 and confirms our hypotheses. This value confirms a good agreement between the annotators, but it also shows the necessity of averaging the annotations in order to avoid the other race effect.

To obtain the final annotations, we applied a majority voting rule, which allowed to determine the ethnicity label for 99% of the face images in the dataset; as for the remaining

1%, we employed a tie-break rule, by asking a fourth annotator the opinion about the ethnicity. Such annotator, unlike the others, was allowed to gather information about the identities (known the name and surname of the celebrity, it was possible to determine the birth place and so on) and the opinions of the other annotators, in order to take more into consideration the opinion of the person of the same ethnicity group, according to the ORE concept; despite this apparent advantage, the role of the latter was anything but simple, because the remaining 74 identities had characteristics common to different ethnicity groups, so confirming the difficulty of this task even for a human.

2.3 Dataset statistics

The face images have been then divided in training and test set, by preserving the identity partition provided by the original VGGFace2 authors. The training and the test sets are already split, and the ethnicity labels are available upon request at https://mivia.unisa.it. The downloadable package also contains the different annotation files produced by the three annotators.

The final VMER dataset, whose detailed statistics are reported in Table 2, consists of 3,309,742 face images of 9129 identities. There is no subject overlap between the training and the test sets, namely the samples of the subjects used for training the networks are not included in the test set. In face analysis, this separation is very important for evaluating the generalization capabilities of the neural networks.

The training and the test sets are unbalanced, since around 80% of the images belong to the Caucasian Latin category; this is not representative of the real distribution of ethnicities in the world. Nevertheless, we argue that the available samples are sufficient for obtaining a wide training set in which all the ethnicity categories are equally represented. The less represented class in the training set, namely the East Asian, includes 187,893 samples; if we randomly select 187,893 samples from each of the 4 classes, it is possible to obtain a balanced training set with more than 750.000 face images, that is by far the largest existing balanced training set for ethnicity recognition. In 3, we show how this procedure allows the convolutional neural networks to learn a set of features not specialized on the most represented ethnicities.

In Sect. 3, we perform different experiments with the original training and the balanced one in order to evaluate the impact of this aspect on the overall performance.

2.4 Pre-processing

Each face analysis algorithm requires a preliminary face detection step. Since we are interested in detecting frontal and non-frontal faces, we use the detector available in OpenCV based on ResNet-10, which is demonstrated to be a very effective CNN for face detection. In some cases, the algorithm may return more than one bounding box, so we carry out a non-maxima suppression retaining only the face proposal with the highest confidence. In addition, since the detector returns a rectangular box while most of the CNNs require a square input, we add to each detected face image a padding to make them square; at the same time, we annotate the coordinates of the original bounding box, that can be used for avoiding the costly face detection step.

Such pre-processing is very general and has at least two main advantages: (i) it allows to speed up the training process, since the coordinates of the faces in the images are already provided; (ii) thanks to the padding, it enables the possibility to perform further pre-processing operations such as face alignment and various types of data augmentation (histogram equalization, random cropping, image rotation, noise addition). Both these additional services, namely the coordinates of the detected faces and the code for performing data augmentation, can be obtained upon request at https://mivia.unisa.it.

3 Experiments

In this section, we describe the considered CNNs and the protocol adopted for our experimental analysis. We analyze the results from different points of view, evaluating the effect of the data augmentation, the possibility to balance the per-class error, the impact of the input size, the generalization capability and the learned features.

3.1 Deep network architectures

For our experimental analysis, we have chosen the CNNs that we consider the most promising and interesting among the modern deep network architectures, namely VGG-16, VGG-Face, ResNet-50 and MobileNet v2.

VGG-16 [42] is one of the most experimented CNN architectures for facial soft biometrics analysis. It achieved a significant success thanks to its shallow architecture (around 130K parameters, 13 convolutional layers and 3 fully connected layers), that allows to better generalize even in presence of small training sets. It achieves state-of-the-art age estimation accuracy [6] and it is not a case, since there are not very large datasets for training deep networks for age estimation. Being one of the most popular CNNs, we include it in our performance analysis. As evident from the name, it consists of 16 levels and requires an input of 224×224 pixels.

VGG-Face [33] is VGG-16 trained from scratch for face recognition on almost 1,000,000 images. This CNN is probably the most adopted architecture for facial soft biometrics analysis. Indeed, the availability of weights pre-trained on a very large number of face images and not for general image

classification task (ImageNet) makes it very suited for transfer learning. VGG-Face achieved an impressive accuracy in face recognition [33] and age estimation [6], and it has been successfully experimented also for gender recognition [15]; for this reason, we believe it can be effective also for ethnicity recognition purposes.

ResNet-50 is one of the residual networks [21] proposed by the Microsoft research group, which won the ILSVRC and COCO 2015 competitions. The most important feature of such architecture is the introduction of the residual blocks, which allow ResNet to require less processing time for training and less extra parameters for increasing the depth of the network. Consequently, various versions of ResNet have been proposed with increasing number of layers (18, 34, 50, 101, 152). However, it has been demonstrated that ResNet-50 is very effective for other facial soft biometrics analysis, namely age group classification [6] and emotion recognition [26]. For this reason, in this paper we use its version with 50 layers, which takes as input an image of 224 × 224 pixels.

MobileNet v2 is one of the MobileNets architectures [23], very suited for mobile and embedded vision applications; think, as an example, to a cognitive robot or a smart camera that is able to perform face analysis in real time with a good accuracy even using normal CPUs [15]. The software optimization which allows this network to significantly reduce the processing time is the transformation of the convolutional layers in depthwise and pointwise operations, without paying a lot in terms of accuracy. In our opinion, this network architecture can be useful for real-time ethnicity recognition applications running on low cost devices. In this paper, we use the most popular v2 version, which consists of 17 layers and that, in its original version trained with ImageNet, requires an input of 224 × 224 pixels.

3.2 Experimental protocol

For each considered CNN, we apply the same experimental protocol. First of all, we start from the models and the weights already available for all the networks; in particular, we use the implementations of the CNNs already available in Keras with Tensorflow backend. Then, we train them by starting from the pre-trained weights, performing a fine tuning of all the layers.

We use the Adam optimizer and start from a learning rate equal to 0.0005, applying a learning rate decay of 0.5 every 6 epochs; moreover, we setup a weight decay equal to 5e−5. We impose a batch size equal to 64 and build the batch in order to preserve the a priori distribution of the training set. In more details, since the training set (see Table 2) includes 7.7% of African American, 6.0% of East Asian, 79.9% of Caucasian Latin and 6.4% of Asian Indian face images, we build each batch with 5 African American, 4 East Asian, 51 Caucasian Latin and 4 Asian Indian faces.

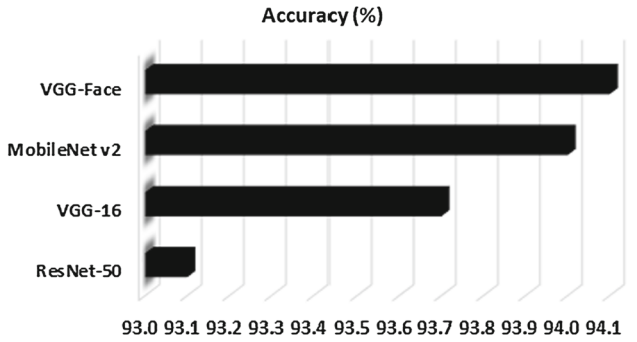

Fig. 2 Ethnicity recognition accuracy (%) of the considered CNNs on the VMER dataset. In this experiment, the CNNs are trained without data augmentation

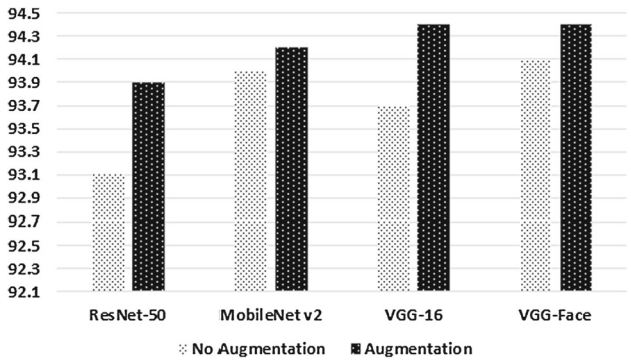

Fig. 3 Ethnicity recognition accuracy (%) of the considered CNNs without and with data augmentation on the VMER dataset. The positive effect of the data augmentation is evident for all the CNNs

We fix the maximum number of training epochs to 20 and implement an early stopping mechanism: if the accuracy on the validation set does not improve for 3 consecutive epochs, the training is stopped.

3.3 Results

The results achieved by the considered convolutional neural networks, trained with the above mentioned protocol, are reported in Fig. 2. As noticed in other face analysis tasks [6], VGG-Face is the most effective CNN also on the VMER dataset, obtaining 94.1% of accuracy. However, the gap with the other networks is not so wide, being all able to achieve an accuracy greater than 93% (MobileNet v2 94.0%, VGG-16 93.7%, ResNet-50 93.1%).

Such results suggest that the proposed dataset allows to effectively train CNN architectures for ethnicity recognition. However, it is worth to deepening the analysis by applying data augmentation or specific design choices for balancing the errors and optimizing the processing time, in order to evaluate the impact of these factors.

3.4 Effect of data augmentation

Data augmentation on the training set is a strategy which demonstrated to be very effective for improving the generalization capabilities of the neural networks; this is definitely true for face analysis, since the possible face variations in terms of pose, orientation, resolution, image quality and occlusions, require the adoption of techniques for making the training set more representative of the real facial variability.

Therefore, we performed a new training of the considered CNNs by applying data augmentation. To make a fair comparison, we have not increased the number of training samples, but we defined a pseudo-random procedure to synthetically add variations to the available face images. In particular, we randomly applied the following data augmentation techniques: gaussian noise addition, brightness change, image rescaling and random flip. It is worth mentioning that these operations are not mutually exclusive, since we randomly combined also 2 or more augmentation strategies.

To reproduce the effect of motion blur or low image quality, we add gaussian noise, produced with a zero-mean gaussian distribution, by fixing $sigma = 12$. To complete the transformation, we normalize the image to have values between 0 and 255.

To simulate overexposure and underexposure, which can be present depending on how the camera is installed with respect to the light source, we randomly add or subtract brightness to the original image. In particular, we subtract 30% of the pixel intensity values to reduce the brightness of the original image, while doing the opposite to reproduce the overexposure. Also in this case, we finally normalize the image to have values in the range [0, 255].

In real environments, the distance between the face and the camera is always variable; if the person is far from the camera, the resulting face image can have a very low resolution. To reproduce this effect, we randomly subsample the original image by resizing it with a random scaling factor of 2 or 4. Of course, this transformation is applied before rescaling the face image to 224×224 pixels, namely the size required by the target CNNs.

Finally, we further augment the dataset by randomly applying a flip of the original image.

The results of this experiment, shown in Fig. 3, demonstrate the effectiveness of the data augmentation, since all the CNNs benefit from the application of this strategy. VGG-Face and VGG-16 achieve an accuracy of 94.4%, while MobileNet v2 and ResNet-50 94.2% and 93.9%, respectively. Among them, ResNet-50 and VGG-16 obtain a more relevant performance improvement (0.8% and 0.7%), while VGG-Face and MobileNet v2 achieve a smaller increase (0.4% and 0.2%), probably because they start from a higher accuracy.

It is important to clarify that we applied the data augmentation only on the training set and not on the test set, to make a fair comparison; thus, the performance improvement is even more significant and the augmentation techniques demonstrated their effectiveness also on the original (non-synthetic) samples.

Considering the improvement achieved with the data augmentation, the other two experiments reported in the following are carried out by applying this technique.

3.5 Balanced dataset

As evident from the results reported in Table 3, the CNNs are more specialized in the recognition of Caucasian Latin individuals. This accuracy imbalance is probably due to the different number of samples available for the various ethnicity groups, which implies an unbalanced prior distribution of the training set and a specialization of the neural networks in the classification of the most represented classes. Performance imbalance is not necessarily a negative factor, since some real problems have intrinsic imbalance; in fact, the ability to recognize more effectively the most representative categories could be a desired behavior. Think, as an example, to a self-service petrol station, which must automatically recognize each type of banknote; the ability to recognize more reliably small denominations, which are presented with higher probability by the customers, is certainly a desired feature.

Nevertheless, especially when the dataset is not balanced due to lack of samples and not for a choice, it might be interesting to balance the accuracy on each class and reduce the imbalance introduced by the a priori distribution of the dataset. To this aim, we investigate this aspect performing a new experiment with a balanced version of the dataset, which consists of around 750, 000 samples. In particular, we train all the CNNs by using a batch composed by the same number of samples for each ethnicity group. In this way, for each epoch the neural networks do not perceive the imbalance and assign the same weight to each class. The results of this experiment are reported in Table 3.

We expect on all the CNNs a gain in terms of average error per class and error variance, paying in terms of overall accuracy. In fact, MobileNet v2 and ResNet-50 are able to break down the average error and its standard deviation ($9.9 \pm 4.4\%$ and $10.8 \pm 4.6\%$), but they have a drop in the overall accuracy (93.2% vs 94.2% and 92.7% vs 93.9%). On the other hand, VGG-16 and VGG-Face reduce the error and its standard deviation less ($11.9 \pm 6.7\%$ and $12.6 \pm 7.7\%$), but have also a smaller decrease in accuracy; in particular, VGG-Face retains its 94.4% overall accuracy, achieving the double goal of balancing the error per class while maintaining a stable ethnicity recognition performance.

Table 3 Per-class and overall ethnicity recognition accuracy achieved by the considered network architectures, by varying the training set

CNN	Training set	AA (%)	EA (%)	CL (%)	AI (%)	Overall (%)	Error (%)
VGG-Face	Unbalanced	79.2	90.3	97.8	71.9	94.4	15.2 ± 10.0
VGG-Face	Balanced	82.7	93.0	96.7	77.4	94.4	12.6 ± 7.7
VGG-16	Unbalanced	80.0	91.0	97.8	69.2	94.4	15.5 ± 10.9
VGG-16	Balanced	84.8	92.9	96.0	78.9	94.1	11.9 ± 6.7
MobileNet v2	Unbalanced	80.9	87.6	97.8	71.2	94.2	15.6 ± 9.7
MobileNet v2	Balanced	89.3	93.7	94.1	83.2	93.2	9.9 ± 4.4
ResNet-50	Unbalanced	80.5	88.0	97.7	66.4	93.9	16.9 ± 11.4
ResNet-50	Balanced	87.9	93.1	93.7	82.2	92.7	10.8 ± 4.6

The unbalanced training set allows to preserve the a priori distribution of the dataset, while the balanced version is built taking the same number of images of the different categories and allows to reduce the per-class error variance. The last column reports the average error and its standard deviation

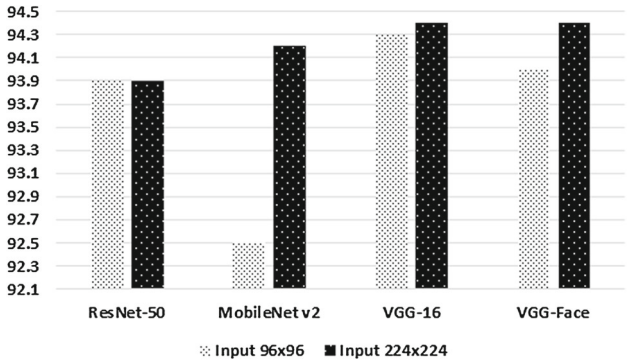

Fig. 4 Ethnicity recognition accuracy (%) of the considered CNNs by varying the input size (96 × 96 and 224 × 224) on the VMER dataset. A significant performance decrease affects only MobileNet v2, while the others are more or less independent on the input size

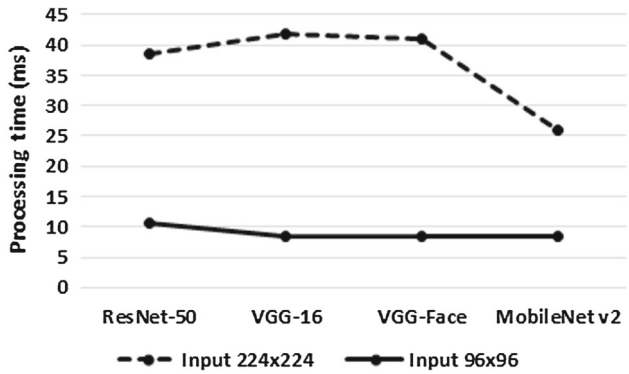

Fig. 5 Processing time (ms) for a batch of 64 images of the considered CNNs by varying the input size (96 × 96 and 224 × 224) on a NVIDIA Titan Xp GPU. The processing time is reduced for all the CNNs of a factor between 3 and 5

Therefore, VGG-Face demonstrates its robustness even in this experiment.

3.6 Impact of the input size

All the considered networks require an input size of 224 × 224, while the size of more than 85% of the face images available in the dataset is less than 200 × 200 pixels. Considering this aspect, we hypothesize that a reduction of the input size should not significantly affect the ethnicity recognition accuracy; on the other hand, it surely implies a gain in terms of training and inference time due to the reduction of parameters and operations.

To this aim, we modify all the considered CNNs to accept an input size equal to 96×96, that is the average size of all the face images in the dataset. Then, we re-train them by applying the same experimental protocol described in Sect. 3.2 with data augmentation.

The results of this experiment are reported in Fig. 4 and confirm our hypotheses. In fact, only MobileNet v2 has a significant drop of the performance with respect to the original CNN (92.5% vs. 94.2%), while ResNet-50, VGG-16 and VGG-Face have a drop of less than 0.5%. On the other hand, such design choice allows to break down the processing time of a factor between 3 and 5, as shown in Fig. 5.

Hence, the idea of reducing the input size of the CNNs, adapting the whole architecture to this choice, can be useful whether there are strict constraints in terms of processing time and memory.[1]

3.7 Generalization capability

In this section, we perform a cross-dataset experiment to verify whether the proposed VMER allows to improve the generalization capability of the convolutional neural networks trained with its images and labels.

To this aim, we follow the experimental protocol described in [24]. We train the same network architecture with two

[1] We used the VGG-Face model fine tuned on FairFace available in the DeepFace library: https://github.com/serengil/deepface.

Table 4 Per-class and overall ethnicity recognition accuracy achieved by ResNet-34 and VGG-Face on the test sets of VMER, FairFace and UTKFace, by varying the training set

Test set	CNN	Training set	AA (%)	EA (%)	CL (%)	AI (%)	Overall (%)
VMER	ResNet-34	VMER	76.6	87.9	97.8	59.8	93.4
	ResNet-34	FairFace	69.3	55.1	96.7	11.5	85.9
	VGG-Face	VMER	79.2	90.3	97.8	71.9	**94.4**
	VGG-Face[1]	FairFace	67.9	83.0	84.1	50.7	81.3
FairFace	ResNet-34	VMER	87.5	81.3	85.0	55.3	80.2
	ResNet-34	FairFace	81.9	88.9	89.9	59.7	**84.3**
	VGG-Face	VMER	86.1	81.5	83.1	56.5	79.4
	VGG-Face[1]	FairFace	81.1	85.0	83.3	43.3	77.6
UTKFace	ResNet-34	VMER	82.7	90.3	96.7	64.3	**89.5**
	ResNet-34	FairFace	69.9	90.2	96.9	31.4	83.5
	VGG-Face	VMER	81.7	90.2	96.4	64.8	89.3
	VGG-Face[1]	FairFace	50.0	73.5	88.7	29.1	75.0

The networks trained with the proposed dataset achieves the best performance over the UTKFace test set, demonstrating that VMER allows to improve the generalization capability
The best result for each dataset is highlighted in bold

training sets, namely VMER and FairFace, and evaluate its performance on a third test set, e.g., UTKFace. As done in [24], we use an ImageNet pretrained version of ResNet-34 and run the training procedure with an Adam optimizer and a learning rate of 0.0001 for 100 epochs, until the validation accuracy stops improving. In addition, we perform a similar experiment with VGG-Face, by comparing the performance of the network trained on VMER with the same architecture trained by using FairFace[1].

Since FairFace includes seven ethnicity categories, we follow the instructions given in [24] for reducing the classes to the same four available in VMER. In particular, they propose the following mapping: Indian and Black are trivially mapped on *Asian Indian* and *African American*, East Asian and Southeast Asian are grouped in the *East Asian* class and the remaining categories (Middle Eastern, White and Latino) are considered *Caucasian Latin*.

The results of this experiment are summarized in Table 4. The ResNet-34 and the VGG-Face networks trained on VMER achieve on UTKFace an overall accuracy of 89.5% and 89.3%, respectively; the corresponding CNNs trained on FairFace obtain a substantially lower performance, namely 83.5% and 75.0%. This result shows that the networks trained on VMER have a greater generalization capability, while those trained with FairFace are more specialized on their training set.

In fact, ResNet-34 trained on FairFace achieves on the test set of the same dataset the best accuracy (84.3%), but the performance is significantly lower than the one obtained by the corresponding CNN trained with VMER on the other test sets. Looking at Table 4, we notice that VMER allows to better generalize on the Asian Indian samples, while the ResNet-34 trained with FairFace have a dramatic decrease in the accuracy on this category (31.4% for UTKFace, 11.5% for VMER). This difference is probably due to the greater number of samples available in VMER for each category and to the high accuracy of the ethnicity annotations.

3.8 Feature visualization

The last experiment we present has the aim of visualizing the discriminative features learned by a CNN trained with VMER. To achieve this goal, we firstly compute the class activation maps [49] to determine the regions in the image which are relevant for recognizing a specific ethnicity category. A class activation map is a heat map computed for each pixel of the input image; its pixels with red color gradations correspond to the regions of the image most used by the neural network to recognize the specific class to which the sample belongs. Since this technique is designed for network architectures having an average pooling and a linear dense layer after the final convolutional layer, we applied it on VGG-Face by using the tool available in keras-vis[2].

Figure 6 shows the average class activation maps obtained when VGG-Face is applied on African American, East Asian, Caucasian Latin and Asian Indian samples. It is evident that the considered CNN recognizes the African American samples by analyzing the region of the lips and the nose, which are discriminative features for this ethnicity. The neuron that recognizes East Asian faces focuses its attention on the region including the eyes and the nose, whose particular shape and size are distinctive facial traits. As evident in Fig. 6, the average images of Caucasian Latins and Asian Indians are quite similar and the distinction between the two ethnicity categories is harder. The average class activation maps show that the CNN focus its attention on the lower part of the face for Caucasian Latins, while for the Asian Indians is arguably more sparse, including the forehead and the cheekbones.

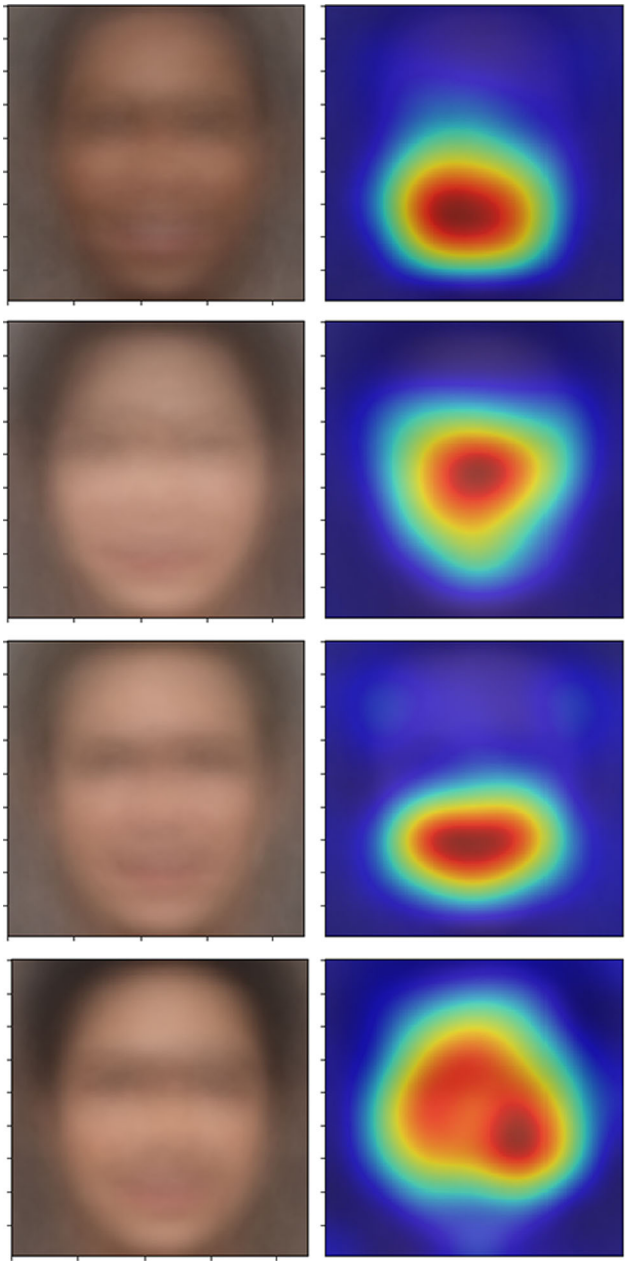

Fig. 6 Average face images and class activation maps obtained by applying our VGG-Face trained on VMER over all the African American (first row), East Asian (second row), Caucasian Latin (third row) and Asian Indian (fourth row) samples. The parts in red correspond to the face regions more relevant for determining the ethnicity

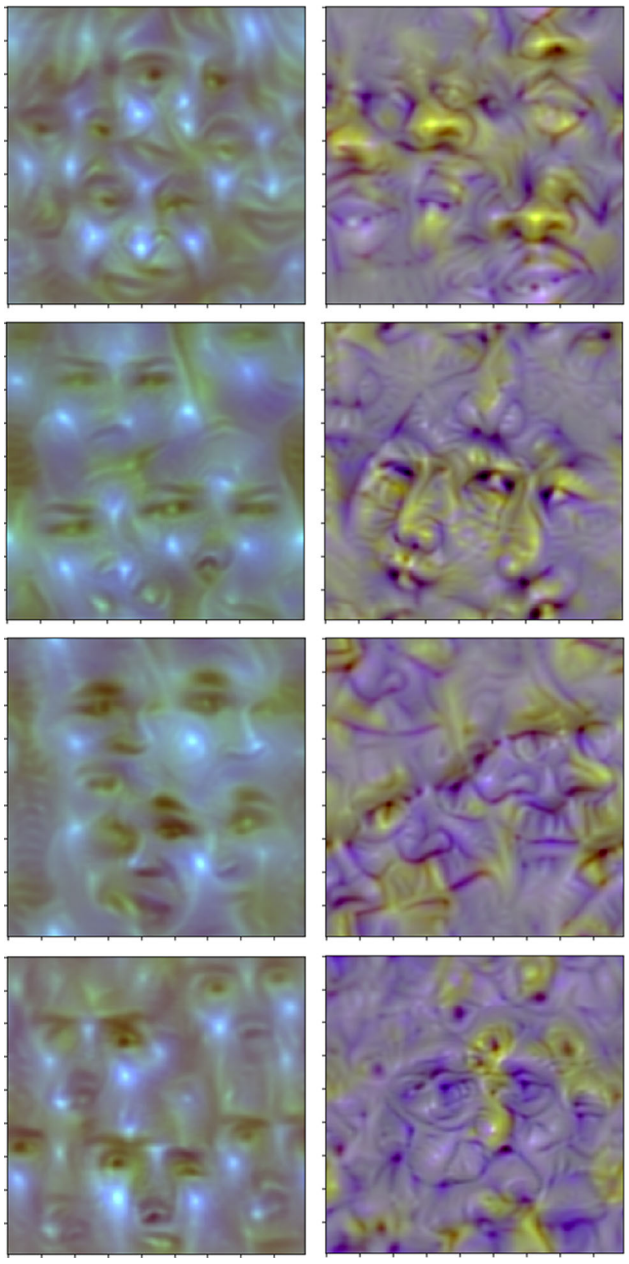

Fig. 7 Result of the Activation Maximization applied on four output neurons of the original VGG-Face trained for face recognition (first column) and of the one fine-tuned for ethnicity recognition (second column). The neurons of the original CNN are sensitive to the whole face image, while the ones belonging to our version are activated by more specific patterns. Full lips and wide nose are the patterns more relevant for African Americans (first row), almond eyes and small nose are significant for East Asians (second row), thin lips and particular nose and eye shapes are distinctive for Caucasian Latins (third row). The lack of strong patterns activating the neuron responsible for Asian Indians (fourth row) partially explains the low accuracy in the recognition of this ethnicity group

To find deeper insights about the features learned with our procedure, we apply the Activation Maximization (AM) method [14] over the four neurons in the output layer of our VGG-Face network fine tuned on the VMER dataset and over four neurons of the last layer of the original one pre-trained for face recognition. This method allows to iteratively generate the image patterns that maximize the activation of the considered neuron; therefore, we can infer the distinctive

facial traits for each ethnicity and the differences with respect to the original features.

Figure 7 shows the results of the Activation Maximization. We can note that the image patterns which maximize the activation of the output neurons of the original VGG-Face, optimized for face recognition, include more or less the whole face. On the other hand, the output neurons of the VGG-Face fine-tuned for ethnicity recognition are sensitive to more specific facial traits, consistently with respect to the class activation maps.

In particular, the output neuron responsible for the classification of African Americans is activated by full lips and wide noses, while the one that recognizes East Asians is sensitive to almond eyes and small noses. The output neuron specialized in the Caucasian Latin category uses thin lips and particular shapes of the nose and of the eyes to recognize face images belonging to this category. Finally, we are not able to find distinctive facial traits which activate the output neuron responsible for Asian Indians; the lack of focus on specific image patterns partially explains the difficulties of the CNN in recognizing samples of this class. We believe that the recognition of this particular ethnicity deserves further future investigations.[2]

4 Conclusion

The ethnicity is so far the less investigated facial soft biometric, due to the intrinsic difficulties of collecting a large, representative and reliably annotated dataset, which implies in turn the absence of a comprehensive benchmarking of the existing methods. To this aim, we proposed the new VMER dataset, annotated with four well-defined ethnicity categories (African American, East Asian, Caucasian Latin, Asian Indian) and composed by more than 3 millions of face images. We involved in the annotation procedure three individuals belonging to different ethnicities, in order to avoid the well-known problem of the own race bias and to provide reliable annotations. We used this dataset to evaluate the performance of 4 existing CNNs, namely VGG-Face, VGG-16, MobileNet v2 and ResNet-50, which demonstrated remarkable performance in other facial soft biometrics recognition. The experimental analysis proved the effectiveness of the proposed dataset, since all the considered neural networks were able to achieve a recognition accuracy close to 94%. The heterogeneous variations present within the dataset allowed to analyze the results from different points of view, evaluating the positive effect of data augmentation, the possibility to balance the per-class error and the impact of the input size. Among the various CNNs, VGG-Face has shown a slight superiority, confirming the excellent results already obtained by this neural network in face recognition, age estimation and gender recognition. We also demonstrated that the neural networks trained with VMER generalize better on different test sets than the corresponding models trained with FairFace, so far the largest and most representative ethnicity dataset. We finally visualize the features learned by the VGG-Face trained with VMER through the class activations maps and the Activation Maximization, showing that they correspond to distinctive facial traits typically used by humans to recognize the ethnicity.

Nevertheless, we consider these results as a baseline to encourage the research in this area and we do not exclude, indeed we hope, that further investigations may allow other researchers to achieve higher and more robust performance even in different scenarios.

[2] https://github.com/raghakot/keras-vis.

Funding Open access funding provided by Università degli Studi di Salerno within the CRUI-CARE Agreement.

References

1. Ahmed, A., Yu, K., Xu, W., Gong, Y., Xing, E.: Training hierarchical feed-forward visual recognition models using transfer learning from pseudo-tasks. In: European Conference on Computer Vision, pp. 69–82. Springer (2008)
2. Anwar, I., Islam, N.U.: Learned features are better for ethnicity classification. Cybern. Inf. Technol. **17**(3), 152–164 (2017)
3. Azzopardi, G., Greco, A., Saggese, A., Vento, M.: Fusion of domain-specific and trainable features for gender recognition from face images. IEEE Access **6**, 24171–24183 (2018)
4. Bastanfard, A., Nik, M.A., Dehshibi, M.M.: Iranian face database with age, pose and expression. Machine Vision pp. 50–55 (2007)
5. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: Vggface2: a dataset for recognising faces across pose and age. In: IEEE International Conference on Automatic Face & Gesture Recognition, pp. 67–74. IEEE (2018)
6. Carletti, V., Greco, A., Percannella, G., Vento, M.: Age from faces in the deep learning revolution. IEEE Trans. Pattern Anal. Mach. Intell. (2019)
7. Carletti, V., Greco, A., Saggese, A., Vento, M.: An effective real time gender recognition system for smart cameras. J. Ambient Intell. Human. Comput. 1–13 (2019)
8. Chen, J., Shan, S., He, C., Zhao, G., Pietikainen, M., Chen, X., Gao, W.: Wld: a robust local image descriptor. IEEE Trans. Pattern Anal. Mach. Intell. **32**(9), 1705–1720 (2010)

9. Cohen, J.: A coefficient of agreement for nominal scales. Edu. Psychol. Meas. **20**(1), 37–46 (1960)
10. Dantcheva, A., Elia, P., Ross, A.: What else does your biometric data reveal? A survey on soft biometrics. IEEE Trans. Inf. For. Secur. 441–467 (2016)
11. Demirkus, M., Garg, K., Guler, S.: Automated person categorization for video surveillance using soft biometrics. In: Biometric Technology for Human Identification VII, vol. 7667, p. 76670P. International Society for Optics and Photonics (2010)
12. Ding, C., Tao, D.: A comprehensive survey on pose-invariant face recognition. ACM Trans. Intell. Syst. Technol. 37 (2016)
13. Dornaika, F., Arganda-Carreras, I., Belver, C.: Age estimation in facial images through transfer learning. Mach. Vis. Appl. **30**(1), 177–187 (2019)
14. Erhan, D., Bengio, Y., Courville, A., Vincent, P.: Visualizing higher-layer features of a deep network. Univ. Montr. **1341**(3), 1 (2009)
15. Foggia, P., Greco, A., Percannella, G., Vento, M., Vigilante, V.: A system for gender recognition on mobile robots. In: International Conference on Applications of Intelligent Systems, p. 9. ACM (2019)
16. Fu, S., He, H., Hou, Z.G.: Learning race from face: a survey. IEEE Trans. Pattern Anal. Mach. Intell. **36**(12), 2483–2509 (2014)
17. Fu, S.Y., Yang, G.S., Hou, Z.G.: Spiking neural networks based cortex like mechanism: a case study for facial expression recognition. In: International Conference on Neural Networks, pp. 1637–1642. IEEE (2011)
18. Gao, W., Cao, B., Shan, S., Chen, X., Zhou, D., Zhang, X., Zhao, D.: The cas-peal large-scale chinese face database and baseline evaluations. IEEE Trans. Syst. Man Cybern. A Syst. Humans **38**(1), 149–161 (2007)
19. Guo, G., Mu, G.: A framework for joint estimation of age, gender and ethnicity on a large database. Image Vis. Comput. **32**(10), 761–770 (2014)
20. Han, H., Jain, A.K., Wang, F., Shan, S., Chen, X.: Heterogeneous face attribute estimation: a deep multi-task learning approach. IEEE Trans. Pattern Anal. Mach. Intell. **40**(11), 2597–2609 (2017)
21. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
22. Hosoi, S., Takikawa, E., Kawade, M.: Ethnicity estimation with facial images. In: IEEE International Conference on Automatic Face and Gesture Recognition, pp. 195–200. IEEE (2004)
23. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: efficient convolutional neural networks for mobile vision applications. Preprint arXiv:1704.04861 (2017)
24. Kärkkäinen, K., Joo, J.: Fairface: Face attribute dataset for balanced race, gender, and age. Preprint arXiv:1908.04913 (2019)
25. Kumar, N., Berg, A., Belhumeur, P.N., Nayar, S.: Describable visual attributes for face verification and image search. IEEE Trans. Pattern Anal. Mach. Intell. **33**(10), 1962–1977 (2011)
26. Li, S., Deng, W.: Deep facial expression recognition: a survey. Preprint arXiv:1804.08348 (2018)
27. Lin, H., Lu, H., Zhang, L.: A new automatic recognition system of gender, age and ethnicity. In: Congress on Intelligent Control and Automation, vol. 2, pp. 9988–9991. IEEE (2006)
28. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3730–3738 (2015)
29. LoBue, V., Thrasher, C.: The child affective facial expression (cafe) set: validity and reliability from untrained adults. Front. Psychol. **5**, 1532 (2015)
30. Lyons, M.J., Budynek, J., Akamatsu, S.: Automatic classification of single facial images. IEEE Trans. Pattern Anal. Mach. Intell. **21**(12), 1357–1362 (1999)
31. Marx, K.: Encyclopedia britannica. Encyclopaedia Britannica Ultimate Reference Suite [M/CD]. Chicago: Encyclopsedia Britannica (2012)
32. Muhammad, G., Hussain, M., Alenezy, F., Bebis, G., Mirza, A.M., Aboalsamh, H.: Race classification from face images using local descriptors. Int. J. Artif. Intell. Tools **21**(05), 1250019 (2012)
33. Parkhi, O.M., Vedaldi, A., Zisserman, A., et al.: Deep face recognition. BMVC **1**, 6 (2015)
34. Peng, Y., Yin, H.: Facial expression analysis and expression-invariant face recognition by manifold-based synthesis. Mach. Vis. Appl. **29**(2), 263–284 (2018)
35. Phillips, P.J., Wechsler, H., Huang, J., Rauss, P.J.: The feret database and evaluation procedure for face-recognition algorithms. Image Vis. Comput. **16**(5), 295–306 (1998)
36. Ranjan, R., Patel, V.M., Chellappa, R., Castillo, C.D.: Deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition (2018). US Patent App. 15/746,237
37. Ricanek, K., Tesafaye, T.: Morph: a longitudinal image database of normal adult age-progression. In: International Conference on Automatic Face and Gesture Recognition, pp. 341–345. IEEE (2006)
38. Riccio, D., Tortora, G., De Marsico, M., Wechsler, H.: Ega-ethnicity, gender and age, a pre-annotated face database. In: IEEE Workshop on BIOMS, pp. 1–8. IEEE (2012)
39. Roomi, S.M.M., Virasundarii, S., Selvamegala, S., Jeevanandham, S., Hariharasudhan, D.: Race classification based on facial features. In: Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics, pp. 54–57. IEEE (2011)
40. Salah, S.H., Du, H., Al-Jawad, N.: Fusing local binary patterns with wavelet features for ethnicity identification. In: World Academy of Science, Engineering and Technology, 79, p. 471. World Academy of Science, Engineering and Technology (WASET) (2013)
41. Seidenari, L., Rozza, A., Del Bimbo, A.: Real-time demographic profiling from face imagery with fisher vectors. Mach. Vis. Appl. **30**(2), 359–374 (2019)
42. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. Preprint arXiv:1409.1556 (2014)
43. Thomaz, C.E., Giraldi, G.A.: A new ranking method for principal components analysis and its application to face image analysis. Image Vis. Comput. **28**(6), 902–913 (2010)
44. Wu, B., Ai, H., Huang, C.: Facial image retrieval based on demographic classification. In: International Conference on Pattern Recognition, vol. 3, pp. 914–917. IEEE (2004)
45. Xie, Y., Luu, K., Savvides, M.: A robust approach to facial ethnicity classification on large scale face databases. In: International Conference on Biometrics: Theory, Applications and Systems, pp. 143–149. IEEE (2012)
46. Yi, D., Lei, Z., Li, S.Z.: Age estimation by multi-scale convolutional network. In: Asian Conference on Computer Vision, pp. 144–158. Springer (2014)
47. Zawbaa, H., Aly, S.A.: Hajj and umrah event recognition datasets. Preprint arXiv:1205.2345 (2012)
48. Zhang, Z., Song, Y., Qi, H.: Age progression/regression by conditional adversarial autoencoder. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5810–5818 (2017)
49. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2921–2929 (2016)