Pattern Recognition 72 (2017) 15–26

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Effective training of convolutional neural networks for face-based gender and age prediction

Grigory Antipov[a,b,*], Moez Baccouche[a], Sid-Ahmed Berrani[a], Jean-Luc Dugelay[b]

[a] Orange Labs, 4 rue Clos Courtel, 35512 Cesson-Sévigné, France
[b] Eurecom, 450 route des Chappes, 06410 Biot, France

ARTICLE INFO

Article history:
Received 23 December 2016
Revised 2 May 2017
Accepted 25 June 2017
Available online 27 June 2017

MSC:
68T10
68T45

Keywords:
Gender recognition
Age estimation
Convolutional neural network
Soft biometrics
Deep learning

ABSTRACT

Convolutional Neural Networks (CNNs) have been proven very effective for human demographics estimation by a number of recent studies. However, the proposed solutions significantly vary in different aspects leaving many open questions on how to choose an optimal CNN architecture and which training strategy to use. In this work, we shed light on some of these questions improving the existing CNN-based approaches for gender and age prediction and providing practical hints for future studies. In particular, we analyse four important factors of the CNN training for gender recognition and age estimation: (1) the target age encoding and loss function, (2) the CNN depth, (3) the need for pretraining, and (4) the training strategy: mono-task or multi-task. As a result, we design the state-of-the-art gender recognition and age estimation models according to three popular benchmarks: *LFW*, *MORPH-II* and *FG-NET*. Moreover, our best model won the ChaLearn Apparent Age Estimation Challenge 2016 significantly outperforming the solutions of other participants.

1. Introduction

Automatic recognition of human demographic traits like gender and age has a number of immediate applications in multiple domains. Indeed, intelligent security systems can locate a person of interest based on a specific set of soft biometric attributes. Automatic age estimation algorithms can prevent minors from purchasing alcohol or tobacco from vending machines. The content of the advertising billboards can be adapted depending on the demographics of pedestrians. In general, large face datasets can be easily managed and organized based on the demographics of humans. Therefore, many research efforts have been devoted to design an automatic system which can estimate these essential human characteristics from face images [1].

Recently, deep neural networks, and in particular Convolutional Neural Networks (CNNs) [2], have boosted nearly all domains of computer vision (e.g. object detection and recognition [3], image super-resolution [4], image captioning [5] and many others). Since 2012, the prestigious *ImageNet* challenge on object recognition and localization [6] has been won uniquely by CNNs with constantly increasing depths [3,7–9]. The Face Recognition (FR) domain has also experienced a very significant breakthrough due to CNNs [10–12].

Unsurprisingly, CNNs have been widely used for both Gender Recognition (GR) and Age Estimation (AE) problems in recent years. For example, all winning solutions of the two last editions of the AE challenge [13,14] are based on CNNs. More generally, Tables 8 and 9 (which are discussed in details in Section 5) demonstrate that since 2014, the majority of studies both on GR and AE have been either integrally based on CNNs or at least, have used CNN-learned features as part of their models. These observations underpin the practical interest of the present research, the goal of which is to find optimal ways of design and training of CNNs for GR and AE.

In particular, we have identified that existing CNN-based approaches for GR and AE vary in principal in the following 4 axes: (1) the target encoding and the loss function used for AE, (2) the depth of the used CNN architecture, (3) presence and type of pretraining (General Task (GT)[1] or FR), and (4) the way how the networks are trained: separately for GR and/or AE or simultaneously for both tasks. In this paper, we evaluate the importance of each of the presented axes on the resulting performances providing prac-

* Corresponding author.
E-mail addresses: grigory.antipov@orange.com (G. Antipov), moez.baccouche@orange.com (M. Baccouche), sidahmed.berrani@orange.com (S.-A. Berrani), jean-luc.dugelay@eurecom.fr (J.-L. Dugelay).

[1] Here and further in this work, by the "GT pretraining" we understand the pretraining on the *ImageNet* dataset [6] of 1000 classes.

http://dx.doi.org/10.1016/j.patcog.2017.06.031
0031-3203/

tical hints for researchers and practitioners who will choose CNNs for addressing GR and AE problems.

Our main contributions can be summarized as following:

1. We identify Label Distribution Age Encoding [15] as an optimal way to represent the target age when training a CNN.
2. We conclude that AE requires deeper CNN architectures than GR when CNNs are trained from scratch.
3. We show that FR pretraining allows effective training of deep gender and age CNNs. In addition, we show that FR pretraining is more suited for the target problems than the GT one.
4. We demonstrate that CNNs benefit from multi-task training for GR and AE when trained from scratch. However, this positive effect is encapsulated by FR pretraining.
5. We report the state-of-the-art results on three popular benchmarks: *LFW, MORPH-II* and *FG-NET*.
6. Based on the designed AE model, we have won the ChaLearn Apparent Age Estimation Challenge 2016 [14,16].

The rest of the article is organised as following: in Section 2, we present the related studies on automatic GR and AE; in Section 3, which is the central one of the present work, we identify the optimal CNN design and training parameters for GR and AE; in Section 4, we use the conclusions of the previous Section as the basis to design the state-of-the-art CNNs for GR and AE; in Section 5, we evaluate our best models on popular benchmarks against the state-of-the-art; and finally, Section 6 provides the final conclusions of this work and the directions for the future research.

2. Related work

In this Section, we briefly present the most relevant works on automatic GR and AE from face images. More detailed bibliography studies can be found in the following surveys: [17,18] (on GR), [19,20] (on AE) and [1,21] (both on GR and AE). Following the categorization from Han et al. [22], we can roughly split non-CNN GR and AE methods into (1) shape-based, (2) texture-based and (3) appearance-based methods. According to the mentioned categorization, CNNs are closer to the texture-based approaches. However, in our opinion, CNNs form a 4th separate group. Below, we provide examples of approaches in each of the four categories.

2.1. Shape-based methods

In shape-based (or anthropometry-based) approaches, GR and AE are performed using distances between predefined facial landmarks to describe the topological differences between male and female faces or between faces of different ages. For example, Poggio et al. [23] measured 15 distances (pupil to eyebrow separation, nose width etc.) while Fellous [24] selected 24 horizontal and vertical distances in a human's face to recognize gender. In case of AE, Kwon and Lobo [25] computed 6 metric proportions on frontal face images and used them to separate babies from adults. Similarly, Ramanathan and Chellappa [26] used 8 proportions to model age progression among children and teenagers up to 18 years old. In general, in AE, the anthropometric features are mainly useful to distinguish children from adults, since the facial shape becomes quite stable for adults [19]. A common downside of anthropomorphic methods is the fact that they are very sensitive to precise estimation of the facial landmarks [22].

2.2. Texture-based methods

Many studies on GR and AE from face images are based on extraction of the image-based texture features from the processed images. The most straightforward way to extract texture information from images is to directly use pixel intensities. This simple approach was employed by several GR works [27–29] with various classification algorithms. Raw pixels contain a lot of redundant information which can be removed using dimensionality reduction methods. To this end, Khan et al. [30] used Principal Component Analysis (PCA) while Jain and Huang [31] employed Independent Component Analysis (ICA) in the context of GR. AE is a more sophisticated problem than GR, and Guo et al. [32] found that unsupervised dimensionality reduction methods like PCA, ICA or Locally Linear Embedding (LLE) are not able to project face images to sufficiently discriminative subspaces. Instead, the authors successfully employed Orthogonal Locality Preserving Projections (OLPP) which is a supervised manifold learning algorithm. This promising idea of using manifold learning for the supervised dimensionality reduction method was later further developed in subsequent works of the same research group [33–35] using respectively Marginal Fisher Analysis (MSA), Locality Sensitive Discriminant Analysis (LSDA), Kernel Partial Least Squares (KLPS) and Correlation Component Analysis (CCA).

General-purpose hand-crafted features were also successfully used for estimation of human demographics. Thus, Local Binary Patterns (LBP) are one of the most basic and popular hand-crafted features. They were broadly utilised for GR [36–38] and AE [39,40]. Biologically Inspired Features (BIF) were used for GR in [22], but they proved to be particularly effective for AE [21] which is confirmed in a number of works [34,35,41]. Some other hand-crafted features were also tried for GR and AE, though less frequently than LBP and BIF. For example, Wang et al. [42] employed Scaled Invariant Feature Transforms (SIFT) for GR, Gabor filters were used by Xia et al. [43] for GR and by Liu and Wechesler [44] for AE, and Haar-like features allowed Zhou et al. [45] to train a boosting model for AE.

Moreover, a very promising approach is combining several feature representations strategies in one model. Thus, in the recent work [46], Castrillón-Santana et al. analysed and compared different methods of fusion of various hand-crafted features, including LBP, Histogram of Oriented Gradients (HOG), Weber Local Descriptors (WLD) and others, in one GR model. Similarly, Moeini et al. [47] combined LBP features and raw pixel intensities extracted from different regions of faces to learn a regression dictionary for GR and AE. In the same spirit, Liu et al. [48] combined LBP, HOG and BIF features to train a hierarchical AE model obtaining the state-of-the-art performances.

2.3. Appearance-based methods

The appearance-based approaches for automatic GR utilize both texture and shape information from face images. A typical appearance-based method is Active Appearance Models (AAM) which was initially proposed for image coding [49]. Using the training dataset, AAM separately learns a statistical shape model and an intensity model of face images. Lanitis et al. [50] extended AAM for age modelling by proposing an aging function to explain variations in ages. Later AAM was independently applied for GR by Xu et al. [51] and by Shih [52]. The famous AGing pattErn Subspace (AGES) algorithm for AE [53] also uses AAM. The basic idea of AGES is to model the aging pattern, which is defined as a sequence of a particular individual's face images sorted in time order, by constructing a representative subspace. The proper aging pattern for a previously unseen face image is determined by the projection in the subspace that can reconstruct the face image with minimum reconstruction error, while the position of the face image in that aging pattern will then indicate its age. In AGES, each face is firstly encoded with AAM.

Similarly to anthropometry-based approaches, appearance-based algorithms suffer from imprecise estimation of facial landmarks.

2.4. CNNs-based methods

Recently, many studies on GR and AE have employed CNNs (cf. Tables 8 and 9). In this Subsection, we make an attempt to organize these works highlighting the differences between them.

One of the most evident difference between various CNN models is the choice of the network architecture. CNNs can be roughly split into shallow networks (up to $5-6$ convolutional layers) and deep networks with more convolutional layers. We have observed that all studies [54–60] which train gender/age CNNs from scratch use shallow architectures, while the works employing deeper architectures (like AlexNet [7] or VGG-16/19 [61]) fine-tune already pretrained CNNs [62–66]. Moreover, two pretraining types (the GT one and the FR one) were used for the demographics estimation. Their fitness for the target problems was studied by Ozbulak et al. [66]. However, the results of Ozbulak et al. are difficult to interpret given the fact that two types of pretraining are compared on two different architectures: *AlexNet* and *VGG-16*.

Loss functions and age encoding strategies are another source of variation between different AE CNNs. Some papers address AE as an ordinal regression problem [58,67]. Others define custom loss functions [65,68]. However the vast majority of CNN-based age models were trained either with pure classification [59,62,66] or with pure metric regression objectives [56,63,64].

Finally, several studies [56,58] compared mono-task training for GR and AE versus simultaneous multi-task training. Results are apparently contradictory: Yi el al. [56] reported no difference between mono-task and multi-task training, and Yang et al. [58] obtained an improvement in AE performance from the multi-task training.

3. CNN design and training strategy

As outlined in Section 2.4, the existent CNN-based solutions for GR and AE mainly differ in the following aspects: (1) the age encoding and the loss function for the age CNNs, (2) the depth of the employed CNN architectures, (3) the use of pretraining, and (4) the training strategy: mono-task vs. multi-task. The objective of this Section is to study each of these design and training parameters and to evaluate their relative impact on the resulting gender and age prediction accuracies. In particular, in Section 3.1, we detail the mentioned CNN parameters highlighting their importance, and in Section 3.2, we experimentally compare the selected configurations. The conclusions of this Section are used to optimally train the deep top performing gender and age CNNs in the following Section 4.

3.1. Studied CNN parameters

Table 1 summarizes the CNN parameters which are evaluated in the present Section. Below, we subsequently define each of them highlighting their importance for GR and AE CNNs.

3.1.1. Target age encoding and loss function

Target encoding defines how the target labels (in our case, genders and ages) are represented in a neural network. Both the information which is given (or not) to the neural network during training and the choice of the loss function for optimization depend on target encoding.

GR is a binary classification problem which does not leave much liberty for the choice of the target encoding and the loss function to optimize. Binary classification problems are solved by neural networks with one or two neurons at the output layer. In the first case, the logistic regression loss function is employed for optimization and in the second case, the cross-entropy one. Cross-entropy loss is mathematically equivalent to logistic one in case of binary classification, so there is no need for experimental comparison of the two losses. If not said otherwise, we train GR CNNs with two neurons at the output layer and the cross-entropy loss.

Table 1
CNN design and training parameters for GR and AE CNNs which are evaluated in Section 3. GR = Gender Recognition. AE = Age Estimation. FR = Face Recognition.

Parameter	Tested values	
	Gender CNN	Age CNN
Target Age Encoding	N/A	0/1-CAE RVAE LDAE
CNN Depth	2 conv. layers 4 conv. layers 6 conv. layers 8 conv. layers	
Pretraining/Multi-task Learning	No pretraining, mono-task FR pretraining, mono-task No pretraining, multi-task FR pretraining, multi-task	

Table 2
Age encodings and corresponding loss functions. N denotes the number of images in a mini-batch, t denotes the targets and p denotes the predictions of CNNs.

Encoding	Loss function
0/1-CAE	$L_{CAE} = -\frac{1}{N}\sum_{k=1}^{N}\sum_{i=1}^{100} t_i^k \log p_i^k$
RVAE	$L_{RVAE} = \frac{1}{N}\sum_{k=1}^{N}(t^k - p^k)^2$
LDAE	$L_{LDAE} = -\frac{1}{N}\sum_{k=1}^{N}\sum_{i=1}^{100}(t_i^k \log p_i^k + (1-t_i^k)\log(1-p_i^k))$

Contrary to GR, the AE problem can be approached in many different ways: classification with coarse categories, per-year classification, regression or even ranking (cf. Section 2.4). Each case imposes particular age encoding and loss function. In this Section, we compare three strategies which proved to be the most effective during the 1st edition of ChaLearn Apparent Age Estimation Challenge [13]: pure per-year classification (employed by the 1st place winner [62]), pure regression (employed by the runner-ups [63,64]) and soft classification (employed by the participants who got the 4th place [69]). It is important to highlight that the results of the ChaLearn Challenge cannot be regarded as a fair comparison between the mentioned age encoding strategies because many other factors influence the final performances of AE methods (each team used different CNN architectures, pretraining types, training datasets etc.)

Table 2 presents the compared age encodings as well as the corresponding loss functions, and Fig. 1 provides an example on how they are used to encode an age of an example face image. Below, we detail each of the three encodings.

In **pure per-year classification**, each age (with a precision up to one year) is treated as a separate class which implies that the age label is encoded as a one-hot 1D-vector. The size of this vector corresponds to the number of classes (in this work, we use 100 classes for ages between 0 and 99 years old). We further refer to this encoding as *0/1 Classification Age Encoding (0/1-CAE)*.

Pure regression has real numbers as targets, therefore real age values are used as labels in this case. This straightforward age encoding is referred as *Real-Value Age Encoding (RVAE)* in our work.

Finally, **soft classification** can be seen as an intermediate case between the discrete classification and continuous regression. As in pure classification, in soft classification ages are encoded by vectors of the dimension which corresponds to the number of classes. However, instead of being binary, the values in the vector are encoded with Gaussian distribution centred at the target age. This allows to encode a notion of neighbourhood between different age

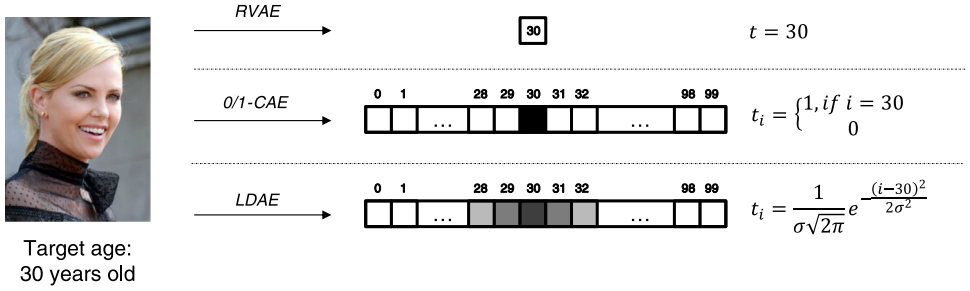

Fig. 1. Example of age encodings. t denotes the resulting encoding. σ is a hyper-parameter of LDAE. In this work, we use $\sigma = 2.5$ (by experimenting with various $\sigma \in [1, 4]$, we have not experienced a significant impact of the σ value on the resulting performance).

classes (which is present in RVAE but does not exist in 0/1-CAE): LDAE for example, encodes that the age of 20 years old is closer to the age of 21 years old than to the age of 80 years old. Following the work where the encoding was introduced [15], we refer to it as *Label Distribution Age Encoding (LDAE)*.

Section 3.2 provides an experimental evaluation of the compared age encodings.

3.1.2. CNN depth

The depth of a neural network (i.e. number of its hidden layers) has a fundamental importance in Deep Learning as it allows to learn a hierarchy of image descriptors for a particular problem starting from the elementary features in the early hidden layers until the high-level problem-dependent features in the last ones [8,70]. Informally, the more complicated is a particular problem, the deeper CNN architecture is required to address it.

Several recent works [9,71] have shown that fully-convolutional CNNs (i.e. CNNs composed of only convolutional layers with no fully-connected ones) perform almost identically to classical CNNs while having much less trainable parameters. It suggests that the discriminative power of a CNN depends rather on its convolutional layers than on its fully-connected ones.

Therefore, in this work, we evaluate the impact of the number of convolutional layers on the quality of gender/age CNNs. In particular, CNNs from 2 and up 8 convolutional layers are compared in Section 3.2.

3.1.3. Pretraining and multi-task training

Despite pretraining and multi-task training may seem as two completely independent techniques at the first sight, both of them are considered as particular cases of the so called "Transfer Learning" [72]. Indeed, the idea of Transfer Learning is that the knowledge learned from one problem can be reused for the other one. It is both reflected in pretraining and multi-task learning.

In case of pretraining, CNN is initialized by a training on a separate complex problem for which there is a lot of training data. The rich internal CNN representations which are learned during pretraining facilitate the further CNN training (also called "fine-tuning") for a problem of interest.

Thus, in this work, we have selected Face Recognition (FR) as a pretraining task due to the following two intuitions. Firstly, contrary to GR and AE problems, FR allows training very deep CNNs from scratch as in [10–12] which proves that this problem is difficult enough to serve as a strong CNN regularizer during the training. Secondly, being a face-related task, FR is close to our target problems. Indeed, gender is a part of a person's identity, therefore GR can be seen as an elementary sub-problem of FR. Though age is clearly independent of a person's identity, it was shown that the face representation learned by a FR CNN implicitly encodes elementary age information [73].

A multi-task CNN is trained to resolve several problems (in our case, GR and AE) at the same time. This way, the CNN learns to extract more information from input images than in case of mono-task training which also results in richer internal CNN representations.

In Section 3.2, FR pretraining and multi-task training are evaluated both separately and simultaneously in the frame of GR and AE problems.

3.2. Experiments

3.2.1. Experimental protocol

We firstly define the experimental protocol which is used for evaluation of all tested CNN parameters in this Section. The protocol consists of the set of the CNN architectures with varying number of convolutional layers as well as of the training and test datasets.

fast_CNN architecture. These CNN architectures are presented in Table 3. For our experiments, we have designed a set of compact CNN architectures of varying depths: *fast_CNN_2, fast_CNN_4, fast_CNN_6* and *fast_CNN_8* with 2, 4, 6 and 8 convolutional layers, respectively. All of them are used for evaluation of the impact of the CNN depth on GR and AE, while the experiments on target age encoding and Transfer Learning are performed with the middle-size architecture *fast_CNN_4*, which is further referred as *fast_CNN* for simplicity.[2]

We have empirically observed that when trained from scratch, using significantly more complex CNN architectures than *fast_CNN* results in poor convergence of the early layers of the network (especially in case of GR). As shown below, this is due to the relative simplicity of the GR problem with respect to FR problem, for example. Thus, we have opted for quite compact CNN architectures (comparing to the state-of-the-art ones like *VGG-16* [61], GoogLeNet [9] and ResNet [3]), as the goal of this Section is the objective comparison of the presented above CNN parameters rather than the design of the best performing gender/age CNNs. The latter is done in Section 4, where we train the very deep state-of-the-art CNNs of 16 and 50 layers for AE and GR based on the conclusions of the present Section.

fast_CNN follows the same basic design principles as *VGG-16* CNN [61]. In particular, (1) all convolutional layers are composed of square feature maps with kernels of size 3x3 pixels, and (2) max-pooling layers reduce both heights and widths of feature maps in 2 times. In order to facilitate convergence and to prevent overfitting, we employ a batch normalization module [74] before ReLU activations and a 0.5-dropout module [75] on the fully connected layer. *fast_CNN* is fed with 64x64 face RGB-images. The retina size of 64x64 for *fast_CNN* has been chosen in conformance with previous work on GR and AE [55,58]. The design of the output (fully-

[2] "Conv: N@MxM" denotes a convolutional layer with N filters of size MxM. "MaxPool: MxM" means that input maps are downsampled by a factor of M using Max-Pooling. "FC: N" denotes a fully-connected layer with N neurons.

Table 3
CNN architectures which are used in experiments of Section 3.

fast_CNN_2	*fast_CNN_4*	*fast_CNN_6*	*fast_CNN_8*
Input: 64x64	Input: 64x64	Input: 64x64	Input: 64x64
Conv1_1: 32@3x3	Conv1_1: 32@3x3	Conv1_1: 32@3x3	Conv1_1: 32@3x3
—	Conv1_2: 32@3x3	Conv1_2: 32@3x3	Conv1_2: 32@3x3
—	—	Conv1_3: 32@3x3	Conv1_3: 32@3x3
—	—	—	Conv1_4: 32@3x3
MaxPool: 2x2	MaxPool: 2x2	MaxPool: 2x2	MaxPool: 2x2
Conv2_1: 32@3x3	Conv2_1: 32@3x3	Conv2_1: 32@3x3	Conv2_1: 32@3x3
—	Conv2_2: 32@3x3	Conv2_2: 32@3x3	Conv2_2: 32@3x3
—	—	Conv2_3: 32@3x3	Conv2_3: 32@3x3
—	—	—	Conv2_4: 32@3x3
FC: 512	FC: 512	FC: 512	FC: 512
Experiment-specific output layer			

connected) layer as well as the corresponding loss function depend on the particular experiment.

Training dataset: IMDB-Wiki_cleaned. In this work, all GR and AE CNNs have been trained on the internal *IMDB-Wiki_cleaned* dataset which is a subset of the public *IMDB-Wiki* dataset [76] collected in 2015. *IMDB-Wiki_cleaned* contains about 250K images (2 times less than *IMDB-Wiki*).

The original *IMDB-Wiki* dataset suffers from a large number of wrong gender and age annotations, so in *IMDB-Wiki_cleaned* we have left only those images for which we are sure that the corresponding annotations are correct (the details are explained in [16]).

Test dataset: private balanced gender age (PBGA). The common problem of public benchmark datasets (like *LFW, MORPH-II* and *FG-NET* used in Section 5 for comparison with state-of-the-art) is the fact that they are not well-balanced. For example, the ratio of men and women both in *LFW* and *MORPH-II* is almost 80% to 20%. Similarly, about 50% of images in *FG-NET* belong to children while *MORPH-II* dataset contains almost 0 images of people over 60 and below 18 years old.

The performances measured on these benchmarks are prone to be biased. This is not critical for comparing the final best gender and age estimators with other state-of-the-art models (anyway, almost all GR and AE studies evaluate their algorithms on one of the three listed benchmarks).

However, in this Section, where our goal is to make important design and training choices for gender and age CNNs, we want to minimize the possible bias due to the evaluation dataset. To this end, we use a private internal dataset of non-celebrities. For each age in the interval between 12 years old and 70 years old, the dataset contains 30 images of men and 30 images of women. Thus, 3540 images in total. Below, we refer to this dataset as Private Balanced Gender Age dataset or simply the *PBGA* dataset. All results reported on the *PBGA* dataset in this Section are calculated according to the cross-dataset protocol (i.e. without fine-tuning on *PBGA*).

3.2.2. Experimental results

Target age encoding. Table 4 compares AE accuracies of *fast_CNN*s trained with different target age encodings presented in Section 3.1. In Table 4 and further in this work, AE CNNs are compared according to Mean Absolute Errors (MAEs). MAE is simply defined as a mean value of absolute differences between predicted ages p and real (biological) ages t (the averaging is done on N testing examples): $MAE = \frac{1}{N}\sum_{i=1}^{N}|p_i - t_i|$.

For 0/1-CAE- and LDAE-based CNNs, we explore two possibilities to predict an age given 100 activations of the output layer. On the one hand, one can select the class (i.e. the age) which corresponds to the neuron with the highest activation — we denote this approach as "ArgMax" in Table 4. On the other hand, the age can be estimated as the expected value of all output activations: $age = \sum_{i=1}^{100} i * p_i$, where p_i is the activation of the ith output neuron (here, we assume that $\sum_{i=1}^{100} p_i = 1$).

Table 4
Comparison of target age encodings. Age Estimation (AE) MAEs are reported on the *PBGA* dataset. Experiments are performed using the *fast_CNN* architecture.

Age encoding	Age prediction type	Age MAE
0/1-CAE	ArgMax	7.00
	Expected value	6.42
RVAE	N/A	7.19
LDAE	ArgMax	6.58
	Expected value	**6.05**

Table 5
Impact of the CNN's depth on Gender Recognition (GR) and Age Estimation (AE). Results are reported on the *PBGA* dataset.

CNN	Gender		Age MAE
	CA	AUC	
fast_CNN_2	92.2%	0.9833	6.65
fast_CNN_4	92.8%	0.9867	6.05
fast_CNN_6	92.9%	0.9862	5.95
fast_CNN_8	92.3%	0.9859	5.89

The results in Table 4 have at least two conclusions. Firstly, we observe that AE by expected values significantly outperforms AE by "ArgMax" both for 0/1-CAE and for LDAE. This result conforms with the similar findings by Rothe et al. [62]. Secondly, the results demonstrate the general superiority of the CNN trained with LDAE over CNNs trained with 0/1-CAE and RVAE. Indeed, LDAE combines the strong points of two other encodings: the similarity of the neighbouring ages (as in RVAE) and the robustness of AE (as in 0/1-CAE). Based on the obtained results, in the rest of the paper, we use LDAE encoding and the expected value approach for AE.

CNN depth. As explained in Section 3.1, four CNN architectures of different depths: *fast_CNN_n*, where $n \in \{2, 4, 6, 8\}$ is the number of convolutional layers, are compared for GR and AE tasks. GR CNNs are evaluated according to their Classification Accuracies (CAs) and also according to their Areas Under ROC Curves (AUCs) for the sake of more balanced evaluation between two classes (men and women).

The results presented in Table 5 highlight the difference between GR and AE. Indeed, in case of GR (columns 2 − 3 of Table 5), we observe that the best performances are already obtained with

Table 6
Effect of Transfer Learning (FR pretraining and multi-task learning) for Gender Recognition (GR) and Age Estimation (AE) CNNs. Results are reported on the *PBGA* dataset using *fast_CNN*. FR = Face Recognition.

Training type	Pretraining	Gender		Age MAE
		CA	AUC	
Mono-task	None	92.8%	0.9867	6.05
Multi-task	None	93.9%	0.9891	5.96
Mono-task	FR	**95.0%**	**0.9917**	**5.96**
Multi-task	FR	94.5%	0.9874	5.96

Table 7
Our best performing Gender Recognition (GR) and Age Estimation (AE) CNNs. Results are reported on the *PBGA* dataset. FR = Face Recognition. GT = General Task.

CNN	Pretraining	Gender		Age MAE
		CA	AUC	
VGG-16	GT	96.8%	0.9958	4.50
VGG-16	FR	97.1%	0.9967	**4.26**
ResNet-50	FR	**98.7%**	**0.9991**	4.33

only four convolutional layers. Increasing the depth up to six layers has almost no impact on GR results, while *fast_CNN_8* of eight convolutional layers performs even worse than shallower networks overfitting on the training dataset. At the same time, the column 4 of Table 5 clearly indicates a positive correlation between the depth of AE CNNs and their performances. *fast_CNN_4* outperforms *fast_CNN_2* by almost 10% while *fast_CNN_6* and *fast_CNN_8* subsequently improve the AE by more than 1% each.

These findings illustrate that AE is a more complex and demanding problem than GR. Indeed, the performed experiments show that contrary to AE, GR training does not provide CNNs with the information which is discriminative enough to take the full advantage of the CNN's depth.

FR pretraining and multi-task training. Experiments presented in Table 6 evaluate the impacts of the FR pretraining and multi-task training on the performances of gender and age *fast_CNN*s. Thus, both FR pretraining and simultaneous learning for the two tasks increase the GR and AE accuracies with respect to the mono-task *fast_CNN* which is trained from scratch (lines (1, 3) and (1, 2) of Table 6, respectively).

The relative improvement of Transfer Learning on gender *fast_CNN* is more important than that on age *fast_CNN*. This perfectly makes sense as GR training itself is not challenging enough to take the full advantage of deep CNNs (cf. the results of the CNN depth experiments). Hence, FR pretraining and multi-task training work as regularizers during the GR training making *fast_CNN* to learn richer and more expressive internal CNN representations. At the same time, AE is a more complicated problem than GR which rather requires more sophisticated deep CNN architectures than an explicit help of Transfer Learning (though the latter also remains useful for age *fast_CNN* as shown in Table 6).

Moreover, while the two Transfer Learning approaches have exactly the same impact on age *fast_CNN* (MAE improvement from 6.05 to 5.96), FR pretraining is more effective than multi-task training for gender *fast_CNN*. Indeed, as already mentioned above, GR can be considered as a sub-problem of FR because gender is a part of a person's identity. Thus, the internal CNN representations of input faces which are learned during FR pretraining contain information which can be directly used to predict gender.

Finally, the lines (3, 4) of Table 6 demonstrate that FR pretraining and multi-task training for GR and AE are not complimentary. Combining the two approaches together does not improve age MAEs and even leads to a slight decrease of gender CAs. This result indicates that the CNN regularization arising from the multi-task training for GR and AE has already been obtained during the FR pretraining. So we can conclude that FR pretraining encompasses the positive effects of the multi-task training for GR and AE being a more general regularization approach.

4. State-of-the-art CNNs for gender and age prediction

In this Section, we design the top performing gender and age prediction CNNs. The idea is to employ some of contemporary deep CNN architectures which have proven to be the most effective for other problems (such as *ImageNet* classification [6]) and to train them for GR and AE according to the conclusions of Section 3. These conclusions can be summarized as following: (1) LDAE should be employed as the age encoding strategy; (2) AE is a more complex problem than GR, and both GR and AE trainings can be improved with help of Transfer Learning; (3) FR pretraining is particularly effective for GR; and (4) FR pretraining encompasses multi-task training meaning that the two Transfer Learning strategies should not be used together. In particular, we adopt two recent CNN architectures:*VGG-16* [61] and *ResNet-50* [3] of 16 and 50 layers, respectively. *VGG-16* is a natural choice because the design of *fast_CNN*, which is used in Section 3, is inspired from *VGG-16*, so this architecture can be considered as a very deep extension of *fast_CNN*. At the same time, *ResNet*s of different depths are currently the state-of-the-art CNN architectures. As shown in [77], *ResNet-50* is a very good trade-off between the running time and the resulting performances.

More precisely, the following strategy is used to train both CNNs (*VGG-16* and *ResNet-50*) for GR and AE:

1. GR and AE CNNs are firstly pretrained for FR.
2. CNNs for GR and AE are trained separately (mono-task training).
3. LDAE is used to encode ages for AE CNNs.

In this work, we employ *VGG-16* from [11] which is pretrained for FR and obtains 97.2% of face verification accuracy on the standard *LFW* benchmark [78]. We have pretrained *ResNet-50* for FR following the same training strategy as in [11], and the resulting CNN reaches 99.3% on *LFW*. Hence, *ResNet-50* has proven to be much more effective than *VGG-16* for FR.

As already observed in Section 3, FR pretraining has a direct influence on GR because the latter can be considered as a particular sub-problem of the former. This is further confirmed by the results in Table 7. Indeed, being more accurate for FR, *ResNet-50* also outperforms *VGG-16* for GR by 1.6 CA points.

On the contrary, AE and FR are two independent problems, and while FR pretraining has a very important regularization role to facilitate AE training, the particular FR accuracy is not a decisive aspect for AE as in the case of GR. Thus, as presented in Table 7, the AE accuracies of *ResNet-50* and *VGG-16* are almost the same. Actually, the fact that a much deeper *ResNet-50* does not improve *VGG-16* for AE reveals the limits of the *IMDB-Wiki_cleaned* dataset which is used for AE training. Indeed, *ResNet-50* CNN model is so complex that it overfits on 250K of training images just after about 5 training epochs (while *VGG-16* does not overfit even after 50 full epochs). That said, we believe that *ResNet-50* would outperform *VGG-16* on AE if more training images with age annotations were available. For example, in order to effectively pretrain *ResNet-50* for FR, we have used a dataset of several millions of face images.

Summarizing the results in Table 7, we select *ResNet-50* CNN as our best model for GR, and *VGG-16* CNN as our best model for AE.

Finally, as a side remark of this Section, it is interesting to measure the particular impact of FR pretraining with respect to General Task (GT) pretraining. Indeed in Section 3, we only intuitively mo-

Table 8
Comparison of our best Gender Recognition (GR) CNN with the state-of-the-art works on *LFW* and *MORPH-II* datasets.

Reference	Year	Used approach	CA	
			LFW	*MORPH-II*
[41]	2010	BIF + OLPP	N/A	97.8%
[34]	2011	BIF + kPLS	N/A	98.2%
[36]	2012	LBP + SVM	94.8%	N/A
[37]	2013	Multiscale LBP + SVM	95.6%	N/A
[35]	2014	BIF + kCCA	N/A	98.4%
[56]	2014	Multi-scale CNN	N/A	97.9%
[58]	2015	Ranking CNN	N/A	97.9%
[22]	2015	BIF + hierarchical SVM	N/A	97.6%
		Human Estimators	N/A	96.9%
[81]	2015	FIS + SVM/RBF	93.35%	N/A
[38]	2015	LBP + C-Pegagos	96.86%	N/A
[82]	2016	Local CNN	94.5%	N/A
[55]	2016	Compact CNN	97.3%	N/A
[54]	2016	LBP/HOG/CNN + SVM	98.0%	N/A
[47]	2017 (in press)	SLCDL + CRC	96.4%	N/A
This Work	**2017**	**ResNet-50 CNN**	**99.3%**	**99.4%**

tivate the choice of FR as a pretraining task. In order to quantitatively confirm this intuition, we also train *VGG-16* from [61] (pretrained on *ImageNet*) for GR and AE, and the resulting scores are presented in line 1 of Table 7. As one may observe, the FR pretraining is more effective than the GT one both for GR and AE (lines (1,2) of Table 7).

Moreover, the difference between FR and GT CNNs can be also perceived qualitatively. To this end, we visualize the mean activations of the intermediate convolutional layers of GT and FR *VGG-16* CNNs when human faces are given at the inputs of the two networks in Fig. 2. More precisely, *VGG-16* is composed of five blocks of $2-3$ convolutional layers in each of them, and in Fig. 2, we present the mean activations at the last convolutional layers of each of these block. In general, early convolutional layers of a deep CNN are activated by elementary parts of input images: like edges, corners etc. Thus, activations in the early layers *conv1_2* and *conv2_2* of the FR and GT *VGG-16* CNNs are similar, and they focus on the most salient face parts (i.e. eyes, mouth, and nose). FR *VGG-16* further consolidates these activations in the deeper layers *conv3_3, conv4_3* and *conv5_3* targeting its attention on the face region. Therefore, the last convolutional layer of the FR CNN is a high-level face descriptor which can be potentially used for GR and AE. On the contrary, the mean activations of the last *conv5_3* layer of GT *VGG-16* are uniformly dispersed all over the map demonstrating that GT *VGG-16* is not trained to focus on human faces (there are few human faces among *ImageNet* images). Thus, FR pretraining allows a CNN to better extract high-level information from face images than GT pretraining making the former more suitable for face-related problems such as GR and AE.

5. Benchmark evaluation

In previous Section 4, we have designed the top performing deep CNNs: *ResNet-50* for GR and *VGG-16* for AE. In this Section we evaluate these two CNNs on three popular benchmark datasets: *LFW* (for GR), *FG-NET* (for AE) and *MORPH-II* (for both tasks).

5.1. Benchmark datasets

Below, we present the benchmark datasets and the corresponding evaluation protocols.

5.1.1. LFW (gender recognition)

The Labelled Faces in the Wild (*LFW*) dataset [78] containing 13,233 photos was collected in 2007. Today, it is the standard benchmark for face and gender recognition systems. In this Section, we employ it for the comparison of our best GR CNN with the state-of-the-art GR models. Most of the recent studies reporting GR results on *LFW* do not fine-tune their models on the target dataset. Therefore, we also follow this cross-dataset protocol for *LFW*.

5.1.2. MORPH-II (gender recognition and age estimation)

The *MORPH-II* dataset [79] is the biggest public dataset of non-celebrities with both gender and age annotations. The dataset which was collected by American law enforcement services contains more than 50K face images.

Guo et al. [41] proposed an evaluation protocol on *MORPH-II* which was later adopted by a large part of the research community. The protocol is the following: the *MORPH-II* dataset is split into three non-overlapping parts S_1, S_2 and S_3 with predefined proportions on gender and ethnicity distributions in each of the parts.

Table 9
Comparison of our best Age Estimation (AE) CNN with the state-of-the-art works on *FG-NET* and *MORPH-II* datasets. (*) different protocol (80% of dataset for training, 20% for test).

Reference	Year	Used approach	MAE	
			FG-NET	*MORPH-II*
[45]	2005	Boosting + Regression	7.48	N/A
[53]	2007	AGES	6.77	8.83
[32]	2008	OLPP + regression	5.07	N/A
[83]	2009	AAM + SVR	4.37	N/A
[84]	2010	MTWGP	4.83	6.28
[41]	2010	BIF + OLPP	N/A	4.33
[85]	2011	CAM + SVR	4.12	N/A
[86]	2011	OHRANK	4.85	5.69
[34]	2011	BIF + kPLS	N/A	4.18
[35]	2014	BIF + kCCA	N/A	3.92
[56]	2014	Multi-scale CNN	N/A	3.63
[22]	2015	BIF + hierarchical SVM	4.8	3.8
		Human Estimators	4.7	6.3
[57]	2015	Unsupervised CNN	4.11	3.81
[58]	2015	Ranking CNN	N/A	3.48
[67]	2015	Ordinal CNN	N/A	3.27
[48]	2015	Hierarchical grouping and fusion	2.81-3.55	2.97-3.63
[65]	2016	Group-aware CNN	3.93	3.25
[62]	2016	ImageNet VGG-16 CNN + regression	3.09	2.68*
This Work	**2017**	**VGG-16 CNN + LDAE**	**2.84**	**2.99/2.35***

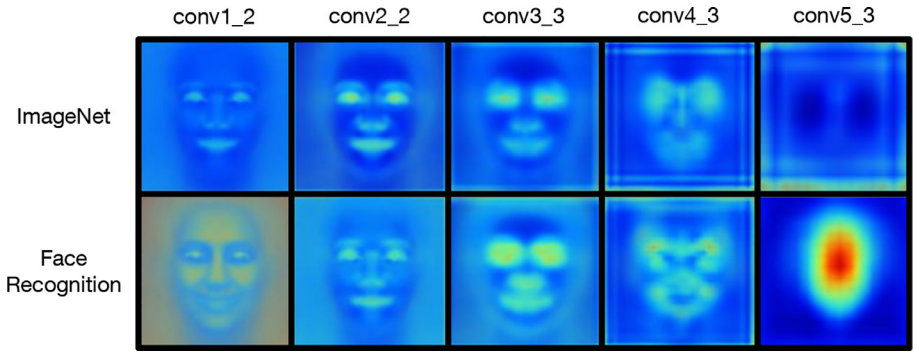

Fig. 2. Heat maps of mean activations of convolutional layers in two *VGG-16* CNNs: the one trained for General Task (GT) classification on *ImageNet* (top), and the one trained for Face Recognition (FR) (bottom). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

GR and AE systems are firstly trained on S_1 and tested on $S_2 \cup S_3$, and secondly trained on S_2 and tested on $S_1 \cup S_3$. Mean CA and MAE over these two experiments are reported as the final ones. We follow this protocol to evaluate both our best GR and AE CNNs.

5.1.3. FG-NET (age estimation)

FG-NET [80] is a tiny dataset containing 975 face images of 82 persons with age annotations. Despite its small size, *FG-NET* is still broadly used in AE community. The Leave One Person Out (LOPO) (i.e. 82-fold Cross-Validation) protocol has been widely adopted for evaluating AE models on *FG-NET*. We follow this protocol to compare our AE CNN with the state-of-the-art.

5.2. Quantitative evaluation

For convenience, Tables 8 and 9 regroup the scores of our best GR *ResNet-50* and AE *VGG-16*, respectively comparing them with the state-of-the-art.

Many of the works from Tables 8 and 9 are discussed in Section 2, but for all reported results we provide short descriptions of the employed methods in the dedicated column. To the best of our knowledge, the current best results for GR were obtained by [54] and [35] on *LFW* and *MORPH-II*, respectively. Castrillon et al. [54] combined hand-crafted features (LBP and HOG) with the features from a compact CNN (comparable by size to *fast_CNN*) and used an SVM classifier above. Guo et al. [35] used BIF features (which are somewhat similar to the features from early layers of deep CNNs) and a kernel-based Canonical Correlation Analysis for simultaneous estimation of gender and age. We improve the results of these two works from 98.0% to 99.3% and from 98.4% to 99.4%, respectively. For both datasets, the improvements are statistically significant with $p < 0.01$ according to the proportions test. We believe that the key reason for the success of our model is the usage of FR as pretraining which has allowed us to train a much deeper CNN than those which were employed by previous CNN-based approaches for GR.

The state-of-the-art AE models were reported by the recent works Liu et al. [48,65] and Rothe et al. [62]. The study from Liu et al. [48] is extremely interesting because despite the authors employed a fusion of very basic hand-crafted features with a standard SVR, they managed to obtain excellent AE results by a meticulous selection of a hierarchical structure of their model (i.e. by first predicting an age group and then estimating the precise age inside the group) and by proposing several feature fusion algorithms. However, the choice of an optimal combination of features to fuse depends on the dataset, therefore it is difficult to evaluate the real AE scores from their work (thus, in Table 9, we provide intervals from their paper rather than a single score). Liu et al. [65] used a hierarchical age grouping to train an AE CNN reporting the currently best score on *MORPH-II* following the well-established protocol from Guo and Mu [41]. Rothe at al. [62] did not follow this protocol on *MORPH-II* so their score on *MORPH-II* cannot be compared to others in Table 9 (for the sake of fair comparison, we evaluate our age CNN both according to the protocols from [41] and [62]). Rothe et al. [62] also obtained the best MAE of 3.09 on the *FG-NET* dataset. The approach of Rothe et al. [62] is very similar to ours: the same *VGG-16* CNN architecture and *IMDB-Wiki* training dataset. However, the principal difference between our solutions is the fact that we use LDAE instead of regression encoding and FR pretraining instead of GT pretraining. The results in Table 9 confirm the validity of the training choices made in Section 3.

5.3. Qualitative evaluation

GR and AE by our best CNNs can be qualitatively assessed in Fig. 3. We provide both examples of successful predictions and examples on which our models fail (in the case of AE, the failed prediction means that the corresponding MAE is significantly bigger than the average one).

For the sake of better understanding of the designed models, we perform a simple ablation analysis to estimate the relative importance of various regions of human faces for the designed GR and AE CNNs. The idea is to blur these regions in input images (using Gaussian filter with $\sigma = 7$) and to evaluate the resulting impacts on gender CAs and age MAEs. The amount of impact indicates the importance of each tested region for the respective tasks. We use three types of occlusions: 49 small square regions, 7 vertical stripes and 7 horizontal stripes. The results are presented in Fig. 4.

Globally, we observe that both networks are sensitive to the salient regions of the face: eyes, eyebrows, nose and mouth. The gender CNN is more sensitive to the centre of the mouth and to the periocular region conforming with previous studies [87,88], while the age CNN more equally depends on all salient face parts. Fig. 4 also demonstrates that the two CNNs quite precisely follow the horizontal symmetry of faces.

5.4. ChaLearn competition on apparent age estimation

In order to further validate the selected approach, in 2016, we participated in the 2nd edition of the ChaLearn Apparent Age Estimation challenge [14]. Our submission [16] was based on the *VGG-16* CNN pretrained for FR and fine-tuned for AE using LDAE encoding (as described in this work). Starting from the basic CNN, we additionally fine-tuned a separate network for AE of children from 0 to 12 years old (because there were many children in the competition dataset). We won the challenge significantly outperforming

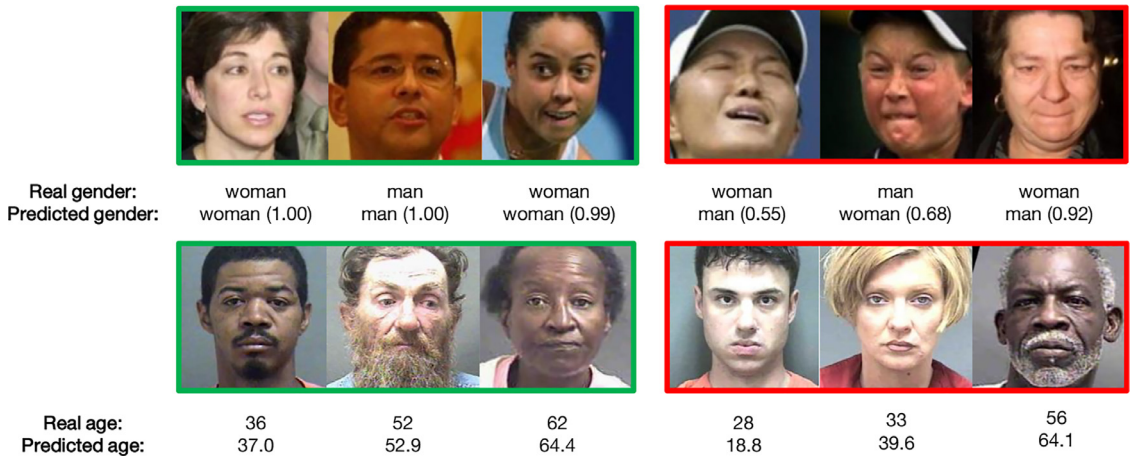

Fig. 3. Examples of Gender Recognition (GR) (on *LFW*) and of Age Estimation (AE) (on *MORPH-II*) by our best models. Both successful and failed cases are presented. For GR, the maximum softmax activation is provided. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

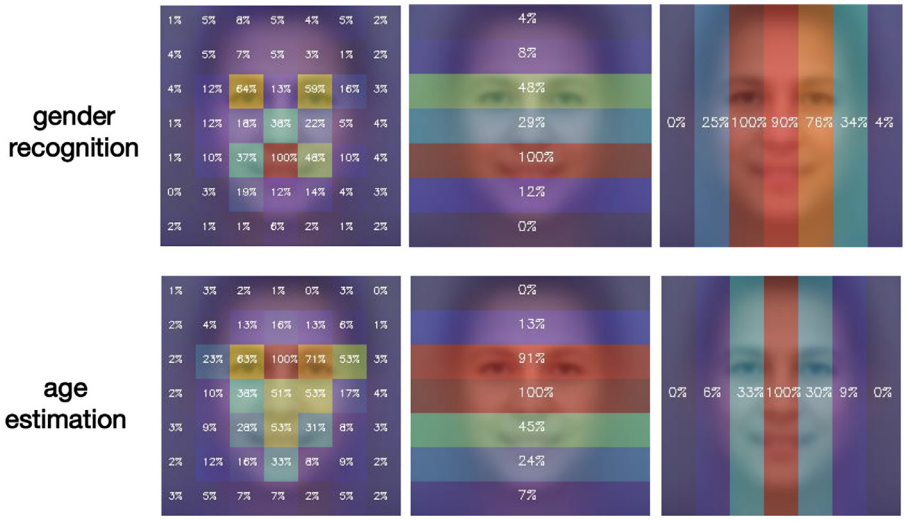

Fig. 4. Sensitivity to occlusions of our best CNNs. Percentages and heat maps indicate the relative losses in performances after blurring the corresponding image parts. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 10
Final results of the ChaLearn Apparent Age Estimation challenge 2016 [14].

Position	Team	ϵ-score[a]
1	**OrangeLabs (our team)**	**0.2411**
2	palm_seu	0.3214
3	cmp+ETH	0.3361
4	WYU_CVL	0.3405
5	ITU_SiMiT	0.3668
6	Bogazici	0.3740
7	MIPAL_SNU	0.4569
8	DeepAge	0.4573

[a] The AE used in the Challenge. It evaluates how far are the model predictions from the estimations made by humans. The lower the better.

other competitors (see Table 10) which confirms the effectiveness of the selected training strategy.

6. Conclusion and future work

In this work, we have been looking for an optimal way of training CNNs for Gender Recognition (GR) and Age Estimation (AE) problems. To this end, we have analysed and experimentally compared (1) different target age encodings (and loss functions), (2) CNN architectures of various depths, and (3) two Transfer Learning strategies, namely: pretraining and multi-task training. As a result, we have obtained the state-of-the-art CNN models for GR and AE.

Below, we highlight the key conclusions of our work:

1. Label Distribution Age Encoding (LDAE) is more effective for AE CNN training than pure classification and regression encodings.
2. AE is a more complex problem than GR. Therefore, when no pretraining is used, AE requires deeper CNN architectures than GR.
3. Face Recognition (FR) pretraining is essential for training deep gender and age CNNs and more suited for the target tasks than the General Task (GT) pretraining.
4. Multi-task training for GR and AE helps only when a CNN is trained from scratch.
5. We have obtained the state-of-the-art results on popular benchmarks: 99.3% of CA on *LFW*, 2.99 of MAE and 99.4% of CA on *MORPH-II*, and 2.84 of MAE on *FG-NET*.
6. The trained AE *VGG-16* CNN is used as a starting point in our winning submission [16] in the ChaLearn Apparent Age Estimation Challenge 2016 [14].

In our future work, we plan to explore the effectiveness of hierarchical approach for GR and AE. The idea is to firstly classify

images into coarse age categories and then to separately train GR and AE CNNs for each category. The recent work [48] as well as our own study [16] (where we train a separate model for children images) demonstrate the high potential of this approach. It might also be interesting to extend our mono-task vs. multi-task study from 2 (gender and age) to k demographic characteristics.

References

[1] A. Dantcheva, P. Elia, A. Ross, What else does your biometric data reveal? A survey on soft biometrics, TIFS 11 (3) (2016) 441–467.
[2] Y. LeCun, B. Boser, J.S. Denker, D. Henderson, et al., Handwritten digit recognition with a back-propagation network, NIPS, 1990.
[3] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, CoRR abs/1512.03385 (2015).
[4] C. Dong, C.C. Loy, K. He, X. Tang, Image super-resolution using deep convolutional networks, TPAMI 38 (2) (2016) 295–307.
[5] J. Johnson, A. Karpathy, L. Fei-Fei, Densecap: fully convolutional localization networks for dense captioning, CVPR, 2016.
[6] O. Russakovsky, J. Deng, H. Su, J. Krause, et al., ImageNet large scale visual recognition challenge, IJCV 115 (3) (2015) 211–252.
[7] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, NIPS, 2012.
[8] M.D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, ECCV, 2014.
[9] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, et al., Going deeper with convolutions, CVPR, 2015.
[10] Y. Taigman, M. Yang, M. Ranzato, L. Wolf, Deepface: closing the gap to human-level performance in face verification, CVPR, 2014.
[11] O.M. Parkhi, A. Vedaldi, A. Zisserman, Deep face recognition, BMVC, 2015.
[12] Y. Sun, X. Wang, X. Tang, Deeply learned face representations are sparse, selective, and robust, CVPR, 2015.
[13] S. Escalera, J. Fabian, P. Pardo, X. Baro, et al., Chalearn looking at people 2015: apparent age and cultural event recognition datasets and results, ICCVW, 2015.
[14] S. Escalera, M. Torres, B. Martinez, X. Baro, et al., Chalearn looking at people and faces of the world: face analysis workshop and challenge 2016, CVPRW, 2016.
[15] X. Geng, C. Yin, Z.-H. Zhou, Facial age estimation by learning from label distributions, TPAMI 35 (10) (2013) 2401–2412.
[16] G. Antipov, M. Baccouche, S.-A. Berrani, J.-L. Dugelay, Apparent age estimation from face images combining general and children-specialized deep learning models, CVPRW, 2016.
[17] E. Mäkinen, R. Raisamo, An experimental comparison of gender classification methods, PRL 29 (10) (2008) 1544–1556.
[18] C.B. Ng, Y.H. Tay, B.-M. Goi, Recognizing human gender in computer vision: a survey, PRICAI, 2012.
[19] Y. Fu, G. Guo, T.S. Huang, Age synthesis and estimation via faces: a survey, TPAMI 32 (11) (2010) 1955–1976.
[20] G. Panis, A. Lanitis, N. Tsapatsoulis, T.F. Cootes, Overview of research on facial ageing using the fg-net ageing database, IET Biom. 5 (2) (2016) 37–46.
[21] G. Guo, Human age estimation and sex classification, in: VABI, 2012, pp. 101–131.
[22] H. Han, C. Otto, X. Liu, A.K. Jain, Demographic estimation from face images: human vs. machine performance, TPAMI 37 (6) (2015) 1148–1161.
[23] B. Poggio, R. Brunelli, T. Poggio, Hyberbf networks for gender classification, 1992.
[24] J.-M. Fellous, Gender discrimination and prediction on the basis of facial metric information, Vision Res. 37 (14) (1997) 1961–1973.
[25] Y.H. Kwon, N. da Vitoria Lobo, Age classification from facial images, CVIU 74 (1) (1999) 1–21.
[26] N. Ramanathan, R. Chellappa, Modeling age progression in young faces, CVPR, 2006.
[27] S. Gutta, H. Wechsler, P.J. Phillips, Gender and ethnic classification of face images, FG, 1998.
[28] B. Moghaddam, M.-H. Yang, Learning gender with support faces, TPAMI 24 (5) (2002) 707–711.
[29] S. Baluja, H.A. Rowley, Boosting sex identification performance, IJCV 71 (1) (2007) 111–119.
[30] A. Khan, A. Majid, A.M. Mirza, Combination and optimization of classifiers in gender classification using genetic programming, IJKBIES 9 (1) (2005) 1–11.
[31] A. Jain, J. Huang, Integrating independent components and linear discriminant analysis for gender classification, FG, 2004.
[32] G. Guo, Y. Fu, C.R. Dyer, T.S. Huang, Image-based human age estimation by manifold learning and locally adjusted robust regression, TIP 17 (7) (2008) 1178–1188.
[33] G. Guo, G. Mu, Y. Fu, C.R. Dyer, T.S. Huang, A study on automatic age estimation using a large database., ICCV, 2009.
[34] G. Guo, G. Mu, Simultaneous dimensionality reduction and human age estimation via kernel partial least squares regression, CVPR, 2011.
[35] G. Guo, G. Mu, A framework for joint estimation of age, gender and ethnicity on a large database, IVC 32 (10) (2014) 761–770.
[36] C. Shan, Learning local binary patterns for gender classification on real-world face images, PRL 33 (4) (2012) 431–437.
[37] J.E. Tapia, C.A. Perez, Gender classification based on fusion of different spatial scale features selected by mutual information from histogram of lbp, intensity, and shape, TIFS 8 (3) (2013) 488–499.
[38] S. Jia, N. Cristianini, Learning to classify gender from four million images, PRL 58 (2015) 35–41.
[39] Z. Yang, H. Ai, Demographic classification with local binary patterns, ICB, 2007.
[40] A. Gunay, V.V. Nabiyev, Automatic age classification with lbp, ISCIS, 2008.
[41] G. Guo, G. Mu, Human age estimation: what is the influence across race and gender? CVPRW, 2010.
[42] J.-G. Wang, J. Li, W.-Y. Yau, E. Sung, Boosting dense sift descriptors and shape contexts of face images for gender recognition, CVPRW, 2010.
[43] B. Xia, H. Sun, B.-L. Lu, Multi-view gender classification based on local gabor binary mapping pattern and support vector machines, IJCNN, 2008.
[44] C. Liu, H. Wechsler, Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition, TIP 11 (4) (2002) 467–476.
[45] S.K. Zhou, B. Georgescu, X.S. Zhou, D. Comaniciu, Image based regression using boosting method, ICCV, 2005.
[46] M. Castrillón-Santana, M. De Marsico, M. Nappi, D. Riccio, Meg: texture operators for multi-expert gender classification, CVIU (2016).
[47] A. Moeini, H. Moeini, A.M. Safai, K. Faez, Regression facial attribute classification via simultaneous dictionary learning, PR 62 (2017) 99–113.
[48] K.-H. Liu, S. Yan, C.-C.J. Kuo, Age estimation via grouping and decision fusion, TIFS 10 (11) (2015) 2408–2423.
[49] T.F. Cootes, G.J. Edwards, C.J. Taylor, et al., Active appearance models, TPAMI 23 (6) (2001) 681–685.
[50] A. Lanitis, C.J. Taylor, T.F. Cootes, Toward automatic simulation of aging effects on face images, TPAMI 24 (4) (2002) 442–455.
[51] Z. Xu, L. Lu, P. Shi, A hybrid approach to gender classification from face images, ICPR, 2008.
[52] H.-C. Shih, Robust gender classification using a precise patch histogram, PR 46 (2) (2013) 519–528.
[53] X. Geng, Z.-H. Zhou, K. Smith-Miles, Automatic age estimation based on facial aging patterns, TPAMI 29 (12) (2007) 2234–2240.
[54] M. Castrillón-Santana, J. Lorenzo-Navarro, E. Ramón-Balmaseda, Descriptors and regions of interest fusion for gender classification in the wild. comparison and combination with cnns., CoRR abs/1507.06838v2 (2016).
[55] G. Antipov, S.-A. Berrani, J.-L. Dugelay, Minimalistic cnn-based ensemble model for gender prediction from face images, PRL 70 (2016) 59–65.
[56] D. Yi, Z. Lei, S.Z. Li, Age estimation by multi-scale convolutional network, ACCV, 2014.
[57] X. Wang, R. Guo, C. Kambhamettu, Deeply-learned feature for age estimation, WCACV, 2015.
[58] H.-F. Yang, L. B-Y, C. K-Y, C. C-S, Automatic age estimation from face images via deep ranking, BMVC, 2015.
[59] G. Levi, T. Hassner, Age and gender classification using convolutional neural networks, CVPRW, 2015.
[60] A. Ekmekji, Convolutional Neural Networks for Age and Gender Classification, Technical Report, Stanford University, 2016. URL http://cs231n.stanford.edu/reports2016/003_Report.pdf.
[61] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, CoRR abs/1409.1556 (2014).
[62] R. Rothe, R. Timofte, L. Van Gool, Deep expectation of real and apparent age from a single image without facial landmarks, IJCV (2016).
[63] X. Liu, S. Li, M. Kan, J. Zhang, et al., Agenet: deeply learned regressor and classifier for robust apparent age estimation, ICCVW, 2015.
[64] Y. Zhu, Y. Li, G. Mu, G. Guo, A study on apparent age estimation, ICCVW, 2015.
[65] H. Liu, J. Lu, J. Feng, J. Zhou, Group-aware deep feature learning for facial age estimation, PR (2016).
[66] G. Ozbulak, Y. Aytar, H.K. Ekenel, How transferable are cnn-based features for age and gender classification? BIOSIG, 2016.
[67] Z. Niu, M. Zhou, L. Wang, X. Gao, G. Hua, Ordinal regression with multiple output cnn for age estimation, CVPR, 2016.
[68] Y. Dong, Y. Liu, S. Lian, Automatic age estimation based on deep learning algorithm, Neurocomputing 187 (2016) 4–10.
[69] X. Yang, B.-B. Gao, C. Xing, Z.-W. Huo, et al., Deep label distribution learning for apparent age estimation, ICCVW, 2015.
[70] Y. Bengio, et al., Learning deep architectures for ai, Found. Trends Mach. Learn. 2 (1) (2009) 1–127.
[71] M. Lin, Q. Chen, S. Yan, Network in network, ICLR, 2014.
[72] S.J. Pan, Q. Yang, A survey on transfer learning, IEEE TKDE 22 (10) (2010) 1345–1359.
[73] Z. Liu, P. Luo, X. Wang, X. Tang, Deep learning face attributes in the wild, ICCV, 2015.
[74] S. Ioffe, C. Szegedy, Batch normalization: accelerating deep network training by reducing internal covariate shift, ICML, 2015.
[75] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, JMLR 15 (1) (2014) 1929–1958.
[76] R. Rothe, R. Timofte, L.V. Gool, Dex: deep expectation of apparent age from a single image, ICCVW, 2015.
[77] A. Canziani, A. Paszke, E. Culurciello, An analysis of deep neural network models for practical applications, CoRR abs/1605.07678 (2016).
[78] G.B. Huang, M. Ramesh, T. Berg, E. Learned-Miller, Labeled faces in the wild: a database for studying face recognition in unconstrained environments, Technical Report, University of Massachusetts, Amherst, 2007.

[79] K. Ricanek Jr, T. Tesafaye, Morph: a longitudinal image database of normal adult age-progression, FG, 2006.
[80] Fg-net aging dataset, (http://fipa.cs.kit.edu/433_451.php).
[81] T. Danisman, I.M. Bilasco, J. Martinet, Boosting gender recognition performance with a fuzzy inference system, ESwA 42 (5) (2015).
[82] J. Mansanet, A. Albiol, R. Paredes, Local deep neural networks for gender recognition, PRL 70 (2016) 80–86.
[83] K. Luu, K. Ricanek, T.D. Bui, C.Y. Suen, Age estimation using active appearance models and support vector machine regression, BTAS, 2009.
[84] Y. Zhang, D.-Y. Yeung, Multi-task warped gaussian process for personalized age estimation, CVPR, 2010.
[85] K. Luu, K. Seshadri, M. Savvides, T.D. Bui, C.Y. Suen, Contourlet appearance model for facial age estimation, IJCB, 2011.
[86] K.-Y. Chang, C.-S. Chen, Y.-P. Hung, Ordinal hyperplanes ranker with cost sensitivities for age estimation, CVPR, 2011.
[87] S.Y.D. Hu, B. Jou, A. Jaech, M. Savvides, Fusion of region-based representations for gender identification, IJCB, 2011.
[88] F. Juefei-Xu, E. Verma, P. Goel, A. Cherodian, M. Savvides, Deepgender: occlusion and low resolution robust facial gender classification via progressively trained convolutional neural networks with attention, CVPRW, 2016.

Grigory Antipov received his MSc degree from Polytechnic University of Catalonia (Spain) in 2014. He is currently pursuing a PhD on deep learning techniques for estimation of human semantic traits at Télécom ParisTech (France). His PhD is funded and co-supervised by Orange Labs.

Moez Baccouche received his PhD degree from National Institute of Applied Science (France) in 2013. His PhD work was funded and co-supervised by Orange Labs. He is currently a research scientist at Orange Labs in Rennes (France), and is working on deep learning and pattern recognition applied to face analysis.

Sid-Ahmed Berrani received his PhD degree from the University of Rennes (France) in 2004. His PhD work was funded and co-supervised by Technicolor within collaboration with INRIA Rennes and received the SPECIF Award from the French Society for Education and Research in Computer Science. He currently leads the "Multimedia Content Analysis Technologies" research team at Orange Labs in Rennes (France). His research activities focus on multimedia indexing, image/video analysis and TV stream structuring. He has authored or co-authored over 50 publications in journals and conference proceedings.

Jean-Luc Dugelay received his PhD degree from the University of Rennes (France) in 1992. His PhD work was funded and co-supervised by France Télécom Research. He is currently a professor in the department of digital security of EURECOM, Sophia Antipolis (France). His research activities focus in the domain of imaging security (image forensics, biometrics and video surveillance, mini drones), and facial image processing. He has authored or co-authored over 280 publications in journals and conference proceedings. He is a fellow member of IEEE and an elected member of the EURASIP Board of Directors. He is the founding Editor-in-Chief of the EURASIP journal on Image and Video Processing (SpringerOpen). In 2015, he served as general co-chair of IEEE ICIP.