Computer Vision and Image Understanding

journal homepage: www.elsevier.com

Age estimation from faces using deep learning: a comparative analysis

Alice **OTHMANI**[a,**], Abdul Rahman **TALEB**[b], Hazem **ABDELKAWY**[a], Abdenour **HADID**[c]

[a] *Université Paris-Est, LISSI, UPEC,*
94400 Vitry sur Seine, France.
[b] *Sorbonne Université,*
75006 Paris, France.
[c] *Center for Machine Vision and Signal Analysis (CMVS)*
University of Oulu, Finland.

ABSTRACT

Automatic Age Estimation (AAE) has attracted attention due to the wide variety of possible applications. However, it is a challenging task because of the large variation of facial appearance and several other extrinsic and intrinsic factors. Most of the proposed approaches in the literature use hand-crafted features to encode ageing patterns. Deeply learned features extracted by Convolutional Neural Networks (CNNs) algorithms usually perform better than hand-crafted features. The main contribution of this paper is an extensive comparative analysis of several frameworks for real AAE based on deep learning architectures. Different well-known CNN architectures are considered and their performances are compared. MORPH, FG-NET, FACES, PubFig and CASIA-web Face datasets are used in our experiments. The robustness of the best deep estimator is evaluated under noise, expression changes, ”crossing” ethnicity and ”crossing” gender. The experimental results demonstrate the high performances of the popular CNNs frameworks against the state-of-art methods of automatic age estimation. A Layer-wise transfer learning evaluation is done to study the optimal number of layers to fine-tune on AAE task. An evaluation framework of Knowledge transfer from face recognition task across AAE is performed. We have made our best-performing CNNs models publicly available that would allow one to duplicate the results and for further research on the use of CNNs for AAE from face images.

© 2020 Elsevier Ltd. All rights reserved.

1. Introduction

Human face conveys a significant information about identity, age, gender, emotion, and ethnicity. It is a key demographic and a soft biometric trait for human identification. Ages are also important in the face-to-face communication between humans. Facial features influence one person’s attraction to another. They can signal cues to fertility and health. Therefore, these factors can increase a person’s productivity and success.

Age is one of the facial attributes which plays a significant role in helping or hindering communication. Like culture, beliefs, experience, language, age can affect both how we say what we mean, as well as how we interpret what others mean. It is a factor that influences how we communicate with each other, and can act as a barrier, along with many other factors. In a study which came out of the university of Pennsylvania and which analyzes the vocabulary of 75000 facebook users, researchers showed that the vocabulary of a person can predict his/her age

[**] Corresponding author:
e-mail: alice.othmani@u-pec.fr (Alice **OTHMANI**)

(Schwartz et al. (2013)).

Raising the ability of a machine to recognize and interpret faces and facial traits such as age in real time can improve the interaction between humans and machines. Many researchers pay attention to the automatic interpretation of face images. Consequently, systems to identify faces and gender, estimate age and recognize emotions, have been developed.

However, faces change with age: as we get older, the skin becomes thicker and its color and texture change, the tissue composition begin to be more sub-cutaneous and the facial skeleton lines and wrinkles appear. The process of ageing is very complicated and varies greatly for different individuals.

Thus, Automatic Age Estimation (AAE) from face images is a challenging topic because of the large facial appearance variations. It is due to a mixture of extrinsic and intrinsic factors. The extrinsic factors are mainly determined by living environment, health conditions, lifestyle, etc., while intrinsic factors include physiological elements, such as genes. Robust AAE systems based on facial images should deal with facial expressions and appearance changes.

AAE systems have a wide range of applications in Human-Computer Interaction (HCI), in surveillance and web content filtering and in Electronic customer relationship management (E-CRM). They are needed mainly because humans fail to perform age estimation accurately. Thus, it is crucial to develop AAE systems that outperform human performance.

2. Motivations and paper organization

Different shallow learning surveys for AAE exist (Fu et al., 2010; Ramanathan et al., 2009). In fact, to the best of our knowledge, there is no comparative study that combines the most popular deep learning models, with the existing state-of-the-art CNN architectures for AAE. The main contributions of this work are as follows:

- In this paper, an extensive comparative analysis of several frameworks for real AAE based on deep learning architectures is given. Several well-known CNNs and public datasets are used. The best configuration for each architecture based on Morph dataset is stated and different configurations are tested.

- The used CNNs are pre-trained on ImageNet to solve object category classification. This first initialization leverages and transfers the knowledge from object recognition domain to facial age estimation domain. An evaluation framework of knowledge transfer from face recognition task across AAE is performed to study the performance of knowledge transfer from related tasks.

- The AAE from face images is a challenging topic because of a mixture of extrinsic and intrinsic factors. Thus, the robustness of the best architecture is evaluated under expression changes, “crossing” ethnicity and “crossing” gender.

The structure of our proposed work is organized as follows: the next section illustrates the related work. Section 4 represents the global schema and the several studied CNN architectures. In Section 5, more details about the datasets and the different experiments are given. Then, the performance of real age estimation of the several frameworks are studied. After discussing the different results and presenting a comparison with the existing approaches, section 6 concludes the paper.

3. Automatic age estimation

3.1. Traditional shallow age learning

In this section, we briefly review automatic age estimation (AAE) using traditional computer vision techniques. These techniques refer to the use of hand-crafted features, they lean towards a human-driven approach. Ageing patterns extraction is the first step in AAE. In the second step, often these feature descriptors are combined with traditional machine learning classification or regression (Support Vector Machine, Random Forest, K-Nearest Neighbors KNN, etc.). The second step concerns a shallow age estimator of the extracted ageing patterns from the first step.

Ageing patterns extraction: this step consists of extracting a set of features to represent the age patterns of the face. Many hand-crafted features have been proposed to describe the shape and the texture patterns of the face (Guo et al. (2009)). Geometric or shape-based ageing patterns are sensitive to pose variations and insufficient for age estimation in adults. Therefore, appearance models have been proposed to capture texture information of the face along with its shape (Georgopoulos et al. (2018)).

Global and local features have been used to describe texture and shape in appearance models for automatic age estimation (Fu et al (2010)). The appearance-based methods include mainly the Local Binary Patterns (LBP) (Gunay and Nabiyev (2008)), Biologically Inspired Features (BIF) (Han et al. (2013)) and Active Appearance Models (Lanitis et al. (2002)).

The texture characteristics of the face have been described using the LBP (Gunay and Nabiyev (2008)). The skin texture regularity determined based on the distribution shape of the LBP histogram is a good age descriptor. The active appearance model (AAM) is also used for face ageing by learning a linear model for shape and intensity from images and a set of landmarks (Lanitis et al. (2002)). Fifty parameters of the AAM are fed to a classifier and the problem is formulated as a regression problem. The BIF features have been investigated for age estimation in (Han et al. (2013)). The ageing subtly on faces are encoded using standard deviation (STD) operator. The normalization with the STD operator reveals local variation capturing vital ageing information like wrinkles and eyelid bags. A series of local descriptors and their combinations which fuse and exploit texture as LBP and SURF and local appearance-based descriptors as HOG have been evaluated under a diversity of settings and the extensive experiments carried out on two large databases: Morph and FRGC (Huerta et al. (2014)).

To better explore the connections between facial features and age labels, distance metric learning and dimensionality reduction are performed in addition to the traditional two-step framework of age estimation algorithms (ageing patterns learning + age estimator learning). Reducing the dimensionality of the extracted features can alleviate the over-fitting problem (Chao et al. (2013)). The number of samples for each age is not balanced in different datasets. To overcome the potential data imbalance problems, a label-sensitive concept and several imbalance analysis are introduced in (Chao et al. (2013)).

Age estimation accuracy is improved through a combination of the proposed hybrid features and the hierarchical classifier. The wrinkle feature is extracted using a set of region-specific Gabor filters, each of which is designed based on the regional direction of the wrinkles, and the skin feature is extracted using a local binary pattern (LBP), capable of extracting the detailed textures of skin (Choi et al. (2011)).

Shallow age estimator: the age estimator learns from the ageing patterns how to predict the age. It can be considered as a classification problem, when each age is taken as a class label. On the other hand, age estimation can also be considered as a regression problem, where each age is used as a regression value. Other approaches combine regression and classification. In these approaches, the age estimator can be modeled as a classifier and a regressor respectively, and then the two models are complementary fused for better performance (Liu et al. (2015)). In fact, hybrid approaches which combine the classification and the regression methods improve the accuracy of age estimation systems by taking advantage of the merits from both (Fu et al. (2010)).

The hand-crafted ageing patterns are fed to shallow classifiers or regressors for age estimator learning. A locally adjusted robust regressor (LARR) is used for learning and predicting the human age using Support Vector Regressor (SVR) (Guo et al. (2008)). An improved hierarchical classifier based on support vector machine (SVM) and support vector regressor (SVR) is proposed in (Choi et al. (2011)). In order to avoid the over-fitting, the dropout-SVM approach is used for face attribute learning (Eidinger et al. (2014)). An age-oriented local regression algorithm named KNN-SVR (K nearest neighbors support vector regression) definitely outperforms SVR thanks to the use of the local regression after performing manifold learning (Chao et al. (2013)).

3.2. Deep age learning

Recently, high-level semantic features are designed based on deep Neural Networks architectures for AAE. The multi-level Neural Networks perform a series of transformations on the face image. On each transformation, a denser representation of the face is learnt. More and more abstract features are learnt in the deeper layers and can allow a better prediction of the class of the ageing patterns. High-level semantic features extracted by deep learning algorithms usually perform better than hand-crafted features. The several proposed frameworks present the same pipeline starting with face detection then face alignment and finally a deep features representation using CNNs to estimate the age of a person (Rothe et al., 2016; Ranjan et al., 2015). These frameworks can be divided into two categories depending on the input data fed to the CNNs. Some approaches feed the full image to the CNN (Wang et al., 2015; Rothe et al., 2016; Yang et al., 2013; Ranjan et al., 2015; Liu et al., 2015; Pan et al., 2018) and other approaches crop the input face image into many local patches (Yi et al., 2014; Dong et al., 2016). All the patches are fed to independent convolutional sub-networks. The response of each patch is combined at the fully connected layer to estimate the age. Less deep architectures are used when several sub-networks learn the ageing patterns from several patches (Yi et al. (2014)). The proposed CNN architectures for extracting deep learned age features present the same components in different order and with different depths: convolutional layers, pooling layers, normalization layers and fully connected layers. The different structures of the CNNs are summarized in Table 2. In (Yang et al. (2013)), the network model is slightly different. Their DeepRank+ comprises a 3-layer wavelet scattering network (ScatNet) (Bruna and Mallat (2013)), a dimensionality reduction component by principal component analysis (PCA) and a 3-layer fully connected network. It is a multi-task ranker: first, it performs between category classification and then within category age estimation. In the ICCV2015 Look for People Apparent Age Estimation, the first ranked approach of CVL-ETH employs 20 VGG deep neural networks and the second ranked approach employs 8 GoogLeNets (Liu et al. (2015)). The deeper architecture which leads to much more computation cost gives the best results.

Different strategies of age encoding are proposed and AAE is approached in different ways: classification with coarse categories, per-year classification, regression, label distribution learning or even ranking (Antipov et al. (2017)). The ordinal regression problem for AAE by using CNN is adressed by Niu et al. (2016) to simultaneously conduct feature learning and regression modelling. For that, a series of binary classification sub-problems are solved with multiple output CNNs learning algorithms. Ranking-CNN, which consists on a series of basic CNNs trained with ordinal age labels, is more likely to get smaller estimation errors when compared with multi-class classification approaches (Chen et al. (2017)). Label Distribution Age Encoding (LDAE) is an intermediate approach between the discrete classification and continuous regression. It encodes a notion of neighbourhood between different age classes. Thus, AAE becomes a label distribution learning problem where Kullback-Leibler divergence between the predicted and ground truth age distributions (Gao et al.

(2017), Huo et al. (2016)) or a mean-variance loss (Pan et al. (2018)) are considered. In a recent work, Antipov et al. (2017) studied the optimal way of training CNNs for AAE by analyzing experimentally: (1) the age encoding and the loss functions, (2) CNN depths and (3) pretraining and training strategies. They have concluded that first, Label Distribution Age Encoding (LDAE) is more effective for CNN training for AAE than pure classification and regression age encoding. Second, AAE requires deep CNN architectures when trained from scratch. Third, Face Recognition pretraining improves the robustness of deep CNNs for AAE and it is more suitable for the target task than the general task. Finally, multi-task training helps AAE when CNN is trained from scratch. Deep Multi-task learning (DMTL) is also studied by Wang et al. (2017) and Han et al. (2018). The DMTL learns jointly multiple CNN models to handle various attributes. It addresses the prediction of one category of homogeneous attributes on two stages: the features extraction stage shared by all attributes categories contains five convolutional layers and two fully connected layers. The multi-task estimation stage contains two sub-networks, each is designed to fine-tune the shared features for attribute category-specific prediction. Different loss functions are designed for multi-task learning. Errors of related tasks are back-propagated jointly for shared features learning. An overview of the most important methods and result on Age Estimation is tabulated in Table 1.

3.3. Cross-domain Age estimation

Human ageing is determined by genes and influenced by intrinsic and extrinsic factors. Previous studies have demonstrated that ageing among populations is different and that learning age jointly with gender and/or ethnicity and/or expression is a more challenging task than learning age independently from these factors (Guo et Mu (2010), Guo and Zhang (2014), Bhattarai at al. (2016), Georgopoulos et al. (2018)).

The influence of gender and ethnicity on AAE on the large Morph dataset has been studied in (Guo et Mu (2010)). It has been shown that cross either race or gender or both decreases performance in AAE. Guo and Zhang (2014) introduced Cross-population discriminant analysis for AAE. The ageing patterns are projected into a common space using Linear Discriminant Analysis and they are correlated with different populations. The low dimensional projection for cross domain age estimation has also been studied in (Bhattarai at al. (2016)) followed by a regression. The projected ageing patterns are fed to a regressor which predicts the age from the domain aligned features. Only few examples from the target domain are used in the training, along with more examples from the source domain and it has been demonstrated that it could be sufficient to predict very well ages from the target domain.

The influence of facial expressions on AAE has been studied in (Alnajar et al. (2014), Lou et al. (2017)). Facial expressions cause changes in facial muscles and an overlap with ageing patterns. The age and the expression are jointly learned using a graphical model with a latent layer between the expression/age and the ageing patterns (Alnajar et al. (2014), Lou et al. (2017)). The latent layer encodes changes in face appearance by learning the relationship between the age and the expression from a training data. The proposed expression-invariant age predictor predicts the age across different facial expression without prior-knowledge of the expression labels (Alnajar et al. (2014), Lou et al. (2017)). More recently, learning the age jointly with the expression is studied by combining scattering and convolutional neural networks (Yang et al. (2018)). The CNN model includes two parallel columns composed on ConvNet and ScatNet , two fully connected layers and an output layer. To better extract the ageing patterns, the CNN is followed by a dimensionality reduction technique for more compact representation.

Therefore, ageing patterns are directly affected by gender, race and expressions. Thus, it is more challenging to design an age estimator which generalizes for different

Table 1: Overview of Age Estimation Methods. * The IMDB-WIKI was used to pre-train the model. ** different protocols and pre-training on half cleaned IMDB-WIKI.

Paper	Method	Datasets	MAE / Accuracy
Hand-crafted based methods			
Lanitis et al. (2002)	AAM + regression	private	4.3 (case2)
Guo et al. (2008)	Locally adjusted robust regressor(LARR)	FG-NET UIUC-IFP-Y/F UIUC-IFP-Y/M	5.07 5.25 5.30
Gunay and Nabiyev (2008)	LBP	FERET + private	80% (6 class)
Guo et al. (2009)	Bio-inspired features (BIF)	FG-NET Private YGA:F Private YGA:M	4.77 3.91 3.47
Choi et al. (2011)	Gabor + LBP + hierarchical classifier based on SVM	FG-NET BERC PAL	4.65 4.68 4.32
Chao et al. (2013)	Label-sensitive relevant component analysis	FG-NET	4.4
Han et al. (2013)	component and holistic BIF	FG-NET MORPH 2 PCOS	4.6 4.2 5.1
Edinger et al. (2014)	LBP + FPLBP + dropout-SVM	Adience collection GALLAGHER Benchmark	45.1% (8 class) 66.6% (7 class)
Huerta et al. (2014)	LBP + SURF + HOG + CCA	MORPH FRGC	4.25 4.17
Deep-learning based methods			
Yang et al. (2013)	ScatNet + PCA + Fully connected layers	MORPH II LIFESPAN FACES	3.49 5.19 7.04
Yi et al. (2014)	multi-scale CNN, sub-network per patch	MORPH II	3.63
Wang et al. (2015)	CNN + dimensionality reduction + classification (SVR, PLS, CCA)	MORPH II FG-NET	4.77 4.26
Niu et al. (2016)	Multiple output CNN learning algorithm	MORPH II AFAD	3.27 3.34
Han et al. (2018)	Deep Multi-task learning (DMTL)	MORPH II LFW+	3.0 4.5
Rothe et al. (2016) *	VGG-16 architecture	MORPH II FG-NET	2.68 3.09
Pan et al. (2018)	mean-variance loss + CNN	MORPH II FG-NET	2.41 4.10
Pan et al. (2018) *	mean-variance loss + CNN	MORPH II FG-NET	2.16 2.68
Antipov et al. (2017)	VGG-16 + LDAE	MORPH II FG-NET	2.99/2.35 ** 2.84

categories.

3.4. Datasets for Automatic Age Estimation

Several datasets have been used for AAE in the literature: Morph (Wang et al., 2015; Yi et al., 2014; Niu et al., 2016; Rothe et al., 2016; Yang et al., 2013), FG-NET (Wang et al., 2015; Rothe et al., 2016), AFAD (Niu et al. (2016)), Lifespan (Yang et al. (2013)), Faces (Yang et al. (2013)), ICCV-2015 challenge dataset (Ranjan et al., 2015; Liu et al., 2015), IMDB-WIKI (Rothe et al. (2016)), AgeDB (Moschoglou et al. (2017)), Audience collection (Eidinger et al. (2014)), CACD (Chen et al. (2015)), Web Image Mining DB (Ni et al. (2009)), FERET (Phillips et al. (1998)) et PIE (Sim et al. (2002)). A comparison of the different cited datasets is given in Table 3.
CNNs require large training datasets. Morph is the most popular dataset for AAE from face images. It contains more than 55000 face images and can overcome the overfitting problem when training deep networks.

4. Proposed Methodology

In this comparative analysis, the framework for age estimation contains five steps which are illustrated in Figure 1. Given a full color image, the face is first detected then aligned. Then, the image is resized to 224 × 224 to have a unique input size. Each aligned face is then passed through a deep CNN to compute ageing patterns required for age estimation. Finally, a 1-layer neural network performs a regression on these patterns to estimate the apparent age. The details of the different steps are presented in the following sections.

4.1. Face Detection and Alignment

The first step concerns the detection and the alignment of the faces mainly because many datasets used in this work do not show centered faces. For better accuracy of AAE, the face images fed to the CNNs must be with minimum background, centered and aligned to a normalized position. A facial landmark detector is used in this work (Sagonas et al., 2013; Kazemi and Sullivan, 2014). Facial landmarks are used to localize and represent salient regions of the face (eyes, eyebrows, nose, mouth and jawline). A fully discriminative model based on a cascade of boosted decision forests to regress the position of landmarks from a sparse set of pixel intensities is performed and it provides accurate landmarks in the majority of cases.

Since faces are detected, facial alignment is realized. The CNNs can tolerate small alignment errors and copes well with such level of precision. The alignment is simply a transformation from an input coordinate space to an output coordinate space such that all the faces are centered, eyes lie on a horizontal line, and faces are scaled such that the faces sizes are nearly identical. Facial landmarks show better performance for face alignment than Haar cascades or HOG detectors since the bounding box provided less precision to estimate the eye location as compared to landmarks indexes.

4.2. Deep ageing patterns learning

Once the face is aligned, ageing patterns are learned. Different CNN architectures are used and different frameworks are performed: VGG16 (Simonyan et al. (2014)), VGG19 (Simonyan et al. (2014)), ResNet50 (He et al. (2016)), InceptionV3 (Szegedy et al. (2016)) and Xception (Chollet (2017)).
The first CNN architecture is the VGG-16 which is formed of 13 convolutional layers, and 3 top fully connected layers. Another architecture that we used is the VGG-19 architecture, which is similar to the VGG-16 architecture, but with 3 more convolutional layers. With the third CNN architecture, a deep residual learning is applied (He et al. (2016)) to train deeper networks with lower complexity. The residual network is called ResNet. It is formed of blocks of convolutional layers, with the addition of the residual shortcut connections. We also studied the inceptionV3 neural network (Szegedy et al. (2016)), which is an improvement of the original inception model. This network uses inception modules. An inception module is a

Table 2: Comparison of different CNN architectures for AAE from face images. The different proposed architectures present the same components in different order and with different depths: convolutional layers, pooling layers, normalization layers and fully connected layers. (*): category-specific feature learning

Ref	Architecture				Number of layers	Input size
	Convolutional	Pooling	FC	Others		
Wang et al. (2015)	3	2	1	-	6	60 × 60
Yi et al. (2014)	1/sub-net	1/sub-net	-	1 local-connected/sub-net	3 × 23 sub-net + 1	48 × 48
Dong et al. (2016)	4	3	1	-	8	39 × 31
Niu et al. (2016)	3	2	1	3 normalization	9	60 × 60
Rothe et al. (2016)	13	-	3	-	16	256 × 256
Yang et al. (2013)	-	-	3	3-layer ScatNet + PCA	6	64 × 64
Ranjan et al. (2015)	10	5	2	4 normalization	21	100 × 100
Han et al. (2018)	5	3	2	5 normalization + (*)	15	-

Table 3: Datasets for AAE from 2D face images. More datasets with age labels exist in the literature. In this table, we mention the most known datasets with large number of images which are required to train deep networks.

Public dataset	**Ref**	**Age range**	**Images Nb**	**Subjects Nb**	**Emotion Expression**	**Gender**	**Ethnicity**
MORPH II	Ricanek and Tesafaye (2006)	16-77	55,134	13,618	Neutral	Unbalanced	Unbalanced
FG-NET	Panis et al. (2016)	0-69	1,002	82	with expressions	Unbalanced	Unbalanced
AFAD	Niu et al. (2016)	15-40	160k	-	with expressions	Unbalanced	Unbalanced
Lifespan	Minear and Park (2004)	18-93	1,046	580	Neutral and expressions	Unbalanced	Unbalanced
Faces	Ebner et al. (2010)	19-80	2,052	171	6 expressions	Balanced	Unbalanced
IMDB-WIKI	Rothe et al. (2016)	-	523,051	100,000	with expressions	Unbalanced	Unbalanced
2015 ICCV challenge	Ranjan et al. (2015) Liu et al. (2015)	0-100	4,699	-	with expressions	Unbalanced	Unbalanced
AgeDB	Moschoglou et al. (2017)	1-101	16,458	568	without expressions	Unbalanced	Unbalanced
Adience collection	Eidinger et al. (2014)	0-60+	26,580	2,284	without expressions	Unbalanced	Unbalanced
CACD	Chen et al. (2015)	-	163,446	2000	without expressions	Balanced	Unbalanced
Web Image Mining DB	Ni et al. (2009)	1-80	219,892	-	without expressions	Unbalanced	Unbalanced
FERET	Phillips et al. (1998)	-	14,126	1,199	without expressions	Unbalanced	Unbalanced
PIE	Sim et al. (2002)	-	41,638	68	with expressions	Unbalanced	Unbalanced

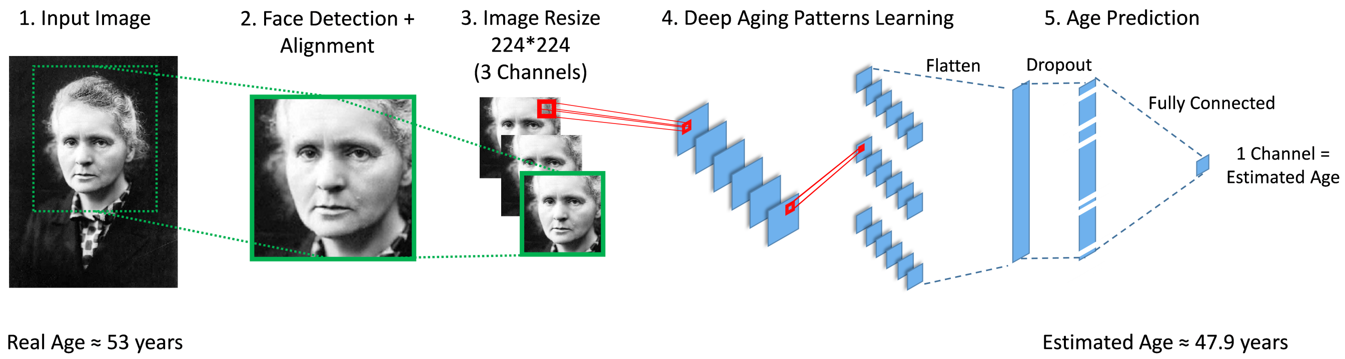

1. Input Image
2. Face Detection + Alignment
3. Image Resize 224*224 (3 Channels)
4. Deep Aging Patterns Learning
5. Age Prediction
Flatten
Dropout
Fully Connected
1 Channel = Estimated Age
Real Age ≈ 53 years
Estimated Age ≈ 47.9 years

Fig. 1: Overview of the proposed approach for AAE. For each image, the face is detected and aligned. Then, the image is resized to 224 × 224. Each aligned face is passed through a CNN for features extraction. Finally, a regression output layer estimates the apparent age. In this example, the Xception architecture is used in step 4.

block of different convolutional sequences, performed separately on their unique given input, and concatenated at the end into the block's output. The last studied architecture is Xception network, it is based on depthwise separable convolution layers. A depthwise separable convolution consists in performing convolution separately on each channel of the input.

4.3. Age estimation

The pre-trained CNNs for ImageNet classification task have an output softmax layer of 1000 channels, one for each of the object classes. However, the age estimation is a non-linear regression problem. We want to predict the age which is a continuous value rather than a set of discrete classes.

Thus, the last softmax layer from each CNN architecture is removed and replaced with 1-layer neural network to learn the age regression function. In this work, the real age estimation is considered as a regression and not a classification problem.

To avoid overfitting due to the small number of training data comparing to the high dimension of the features, a dropout layer is added before the last layer. The regression is learned by optimizing the mean squared error (MSE) loss function defined by:

$$L = 1/N \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \tag{1}$$

where L is the average loss for all the training samples, Y_i is the estimated age and $\hat{Y_i}$ is the real age.

5. Experiments and Results

5.1. Datasets

In our experiments, several datasets are used:

- **MORPH Dataset:** is the largest publicly available longitudinal face database (Ricanek and Tesafaye (2006)). Album 2 of the dataset is used. It contains 55134 images of more than 13000 subjects, spanning from 2003 to late 2007. Ages range from 16 to 77 years. The average number of images per subject is 4. However, the ethnicity and gender distributions are very unbalanced as shown in Table 4.
- **FACES Dataset:** is a database of facial expressions in younger, middle-aged and older men and women, conceived between 2005 and 2007 (Ebner et al. (2010)). It contains face images of 171 subjects with ages ranging from 19 to 80 years old. Each individual has faces in two sets of six facial expression (neutrality, sadness, disgust, fear, anger, and happiness), resulting in a total of 2052 images. For more details, refer to Table 3.
- **FG-NET Dataset:** is a public ageing database, released in 2004 (Panis et al. (2016)). It contains 1002 images from 82 different subjects with ages ranging from 0 to 69 years old. It displays considerable variability in resolution, illumination, viewpoint and expressions. This variability is due to the fact that images were collected by scanning photographs of subjects found in personal collections. Some of the images present occlusion problems, hats and bears.

5.2. Experimental settings

CNNs require a lot of training data due to the large number of parameters in the model to learn. Contrariwise, not all the datasets used in our experiments contain enough images to train the deep CNNs. Besides, the training process is very time-consuming, it can take from hours to months on the optimization. Meanwhile, using a GPU with its parallel architecture can considerably reduce the computational time. The larger the number of GPUs, the less the computational time of training. To overcome these problems, a transfer learning strategy (Pan and Yang (2010)) is set up with two steps:

- pre-training step: the randomly initialized networks are first trained by a related task that owns enough labeled images,

Table 4: MORPH Dataset distribution by Gender and Ethnicity.

	African	**European**	**Asian**	**Hispanic**	**"Other"**	**Total**
Male	36832	7961	141	1667	44	46645
Female	5757	2598	13	102	19	8489
Total	42589	10559	154	1769	63	55134

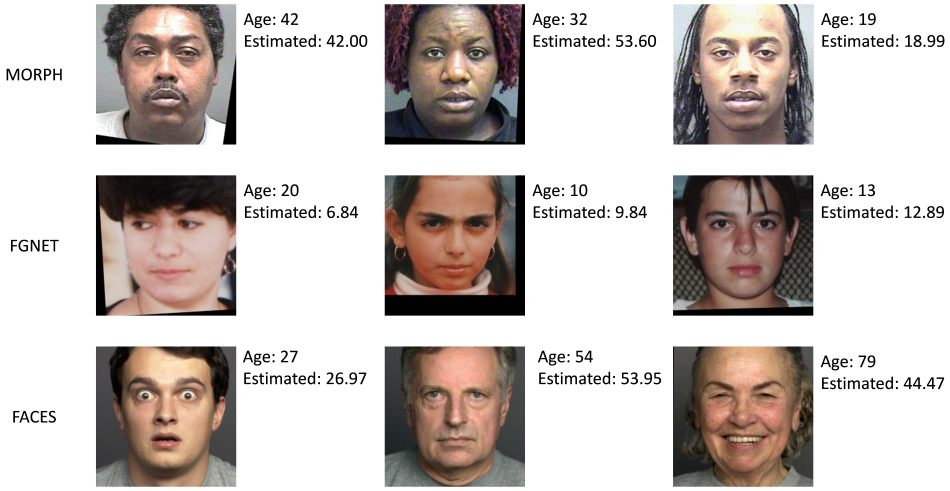


Fig. 2: Illustration of good and bad results of AAE from different datasets using the Xception architecture. Row 1: examples from MORPH dataset, row 2: examples from FGNET dataset, row 3: examples from FACES dataset.

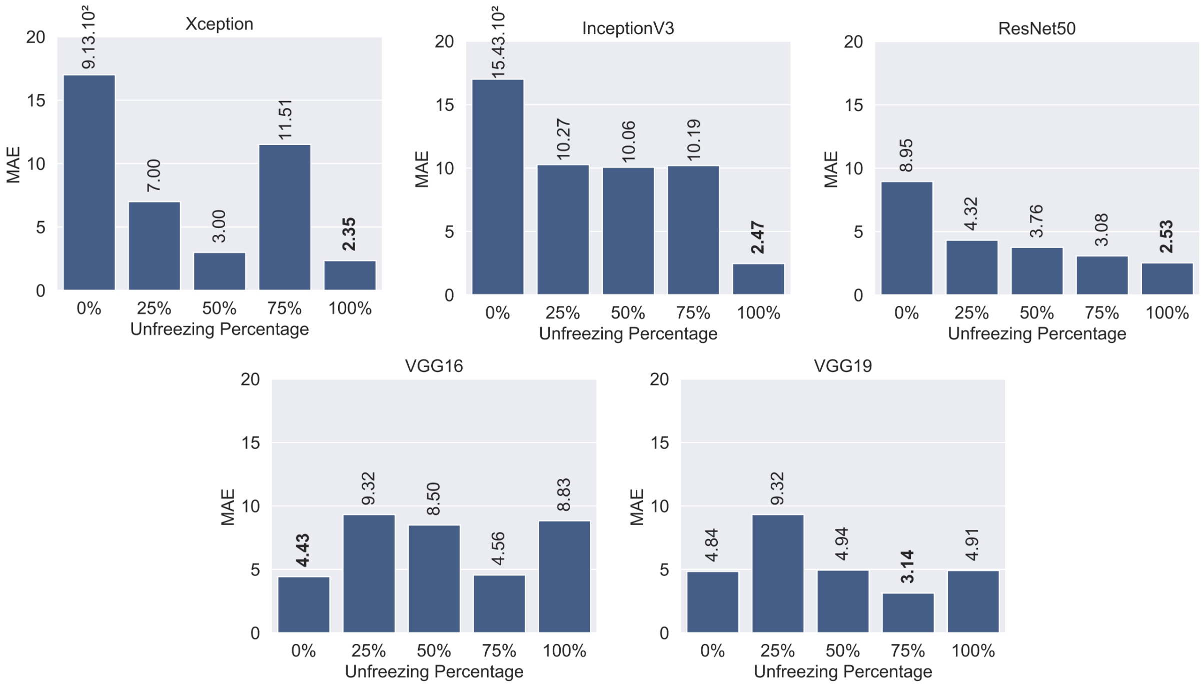


Fig. 3: Performance comparison of CNNs on AAE under the percentage of unfrozen layers. The unfreezing concerns the trainable layers which are mostly the convolution layers. The different networks are pre-trained on ImageNet and fine-tuned on Morph dataset.

- fine-tuning step: the parameters learnt in the pre-training step are used as initialization for new task.

All of the networks have been initialized with weights from training on ImageNet for classification. Then, the CNNs are fine-tuned on the dataset to test. Adam optimizer with its default configuration is used in the training of all the CNNs. The training rate is set experimentally to 0.001. This rate is reduced when the loss value stops decreasing. The batch size is fixed to 32 samples. An early stopping is performed when the accuracy stops improving after 10 epochs. The train-test split was performed using a stratified approach that ensures the same distribution of ages in each of the training and testing dataset. Like in (Niu et al., 2016; Rothe et al., 2016), the MORPH dataset is randomly divided into 80% for training and 20% for testing. During the training phase, 90% of the training set is used for learning the weights and 10% is used for validation. For the FG-NET dataset, we do not perform the same split percentage as for the MORPH dataset because there is much fewer data. Consequently, a split of 90% for training and 10% for testing is chosen.

5.3. Performance of age estimation

To evaluate the performance of the different frameworks in terms of Mean Absolute Error (MAE), a set of empirical experiments have been conducted. The experiments consists of progressive unfreezing of the deep CNNs layers ranging from 0% to 100% with 25% gradually increasing, as shown in Fig. 3. The progressive unfreezing allows to train a subset of the hidden layers while freezing the others, which allows a detailed evaluation of the model in the different configurations.

5.3.1. Layer-wise Transfer Learning Evaluation

Total Transfer Learning Effect. To evaluate the ability of each model to transfer the learned features from the Image-Net initial weights, 0% unfreezing of the hidden layers is applied. As shown in Fig. 3, VGG (VGG16, VGG19) deep models family gets the best results while the Inception (InceptionV3, Xception) deep models family gets the worst results. Consequently, the learned features from VGG (VGG16, VGG19) as well as ResNet can be transferred from the domain of object recognition to the domain of facial ageing estimation and they are more relevant for AAE than the features learned by the more complex convolutional layers of the Inception (InceptionV3, Xception). Furthermore, the best MAE achieved by the VGG16 is 4.43 when 0% unfreezing of the hidden layers is applied. Thus, only the late fully connected layer of the VGG16 need to be fine-tuned for AAE.

Unfreezing Effect [Partial Transfer Learning Effect]. To study the unfreezing effect [Partial Transfer Learning] through different convolutional layers, 25%, 50%, 75% unfreezing of the hidden layers is applied, as shown in Fig. 3. With regard to ResNet model, the more layers being unfrozen, the more accurate results the model can get, which means that the residual connections helps the ResNet model to fuse/transfer more relevant features from the frozen layers to the unfrozen layers. On another hand, the VGG (VGG16, VGG19) and the Xception models do not improve the performance gradually for unfreezing effect and the behavior of the layer-wise transfer learning is non-smooth and changes for the different groups of unfrozen convolutional layers. With regard to InceptionV3 model, compared to 0% unfreezing, the model becomes much more accurate, although the accuracy of the model is almost constant through the different partial unfreezing configurations. Besides, the best MAE of the VGG19 is 3.14 and it is achieved when unfreezing 75% of the hidden layers. Thus, for AAE, only 75% of the hidden layers of the VGG19 requires to be fine-tuned.

NO Transfer Learning . In this phase, each model is trained with 100% unfrozen layers, trying to grasp the best features to get the best accuracy. As shown in Fig. 3, the Xception, the InceptionV3 and the ResNet50 obtain the best MAEs compared to the previous unfreezing configurations of 2.35, 2.47 and 2.53 respectively. Thus, for these three models, all layers should be fine-tuned for AAE task.

Conclusion and discussion about the layer-wise transfer learning experiment . In convolutional neural networks, the early layers learn low-level features which are abstract and general. While, the later layers are more specific to the application. Thus, we can expect that fine-tuning the last few layers should be sufficient. However, in our work, progressively unfreezing the convolutional layers of the different deep neural networks shows different and non-smooth behavior with respect to the unfreezing percentage for different networks. In fact, the optimal number of layers to unfreeze depends on several factors as shown in previous studies in the literature:

- **the size of the training set:** In Khan et al. (2019), for different size of the training set, the behavior of the layer-wise transfer learning for alzheimer diagnosis from MRI images changes.
- **the training data:** In Khan et al. (2019), the trained VGG16 using entropy-based selected images outperforms both training from scratch and random selection with transfer learning.
- **the application or the classification task:** the required layers to be fine-tuned differs from one application to another. For example, only the late fully connected layer needs to be fine-tuned in pulmonary embolism detection, while the late and the middle layers should be fine-tuned in colonoscopy frame classification. On another hand, all layers require fine-tuning for better accuracy for interface segmentation and polyp detection (Tajbakhsh et al. (2016)).

In this work, two others factors are considered in the layer-wise transfer learning: different depths of the networks and different layers types. Thus, we can conclude from our study that the optimal number of layers to fine-tune for age estimation from face images task depends on the depth and the convolutional layers type of the network.

5.4. Best configuration and Robustness

Best configuration for each architecture based on Morph Dataset. From the previous evaluations, in case of No Transfer Learning, we can conclude that the Xception deep network is the best model followed by InceptionV3, ResNet, VGG19, and VGG16 respectively for AAE. In case of Transfer learning, VGG16 model is the best followed by VGG19, ResNet, Xception, and InceptionV3 respectively. In case of partial Transfer learning, ResNet model gets the best performance followed by, VGG19, VGG16, Xception, and InceptionV3 respectively. An illustration of some of the good and bad results of AAE from MORPH dataset is shown in Figure 2, first row.

Computational Complexity. For the experiments, we used a Tesla K80 GPU with a 12 Gb memory and 61 Gb RAM. The computation complexity for the different methods is summarized in Figure 4. We can conclude from Figure 4 that the Inception family has the best performance with reasonable computation time which satisfies the real time applications.

Robustness to noise of the best deep age estimator. The experiments in Figure. 5 demonstrate the robustness of the best deep age estimator based on the Xception architecture to two types of noises with different levels. In the first three experiments, the test images are generated by adding the Gaussian noise of mean 0 and variance 0.01, 0.05 and 0.1 respectively. In the last three experiments, the test images are blurred by convolving them with a PSF. The angle of the blur is fixed in the different experiments to 45% and three lengths of the blur in pixels are considered 10, 20 and 30 respectively. The simulated blur is one of the frequent image degradation that can be caused by many factors like motion during the image capture or an out-of-focus.

It can be seen that, when Gaussian noise is less (variance is small), the Xception model can estimate accurately the human age. However, with the increasing of variance, the accuracy of the age estimation get worse. In general, satisfactory results have been obtained by using the deep age estimator based on Xception architecture for the images added with Gaussian noise and the resulting MAE is varying from 2.52 to 4.63 for a variance from 0.01 to 0.1 . On the contrary, this model is more sensitive to blur noise than Gaussian noise and the Xception based model shows an MAE varying from 2.96 to 12.73 for a blur length from 10 to 30 pixels. Thus, the age estimation framework based on Xception model is robust to small amounts of Gaussian noise and motion blur.

5.5. Generalization of the pre-trained CNNs on MORPH to other datasets

To evaluate the performance of the different CNN architectures trained on MORPH dataset, the model is tested on another dataset. MORPH is a standard dataset and the images are taken in the same conditions. We tested the CNN architectures on FG-NET dataset which displays a big variability. The face images in FG-NET are collected from personal photographs. Thus, the images present different illuminations, different viewpoints and expressions. The results are shown in Table 7. Testing on FG-NET dataset without any fine-tuning presents an MAE varying from 10.8 to 13.2, VGG16 outperforms other architectures but the error increases by more than 7 years in the different tests. When a fine-tuning is applied, the accuracy of AAE increases and Xception CNN outperforms the other trained CNNs and achieves an MAE of 3.6 years. The transition from ideal conditions of faces acquisition (MORPH dataset) to faces in the wild conditions (FG-NET dataset) shows an increased error rate, a fine-tuning improved considerably the results.

Based on our results, regarding the experiments on the two different datasets (MORPH and FG-NET), we can conclude that the Xception model is the best model for AAE. An illustration of some of the good and bad results of AAE from FG-NET dataset is shown in Figure 2, second row.

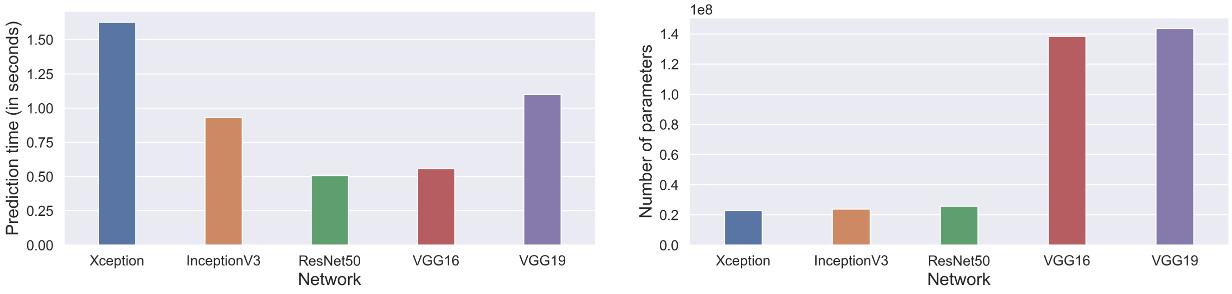
5.6. Knowledge transfer for AAE

The used CNNs are pre-trained on ImageNet, this is not optimal as large scale face recognition datasets are now available and facial age estimation task can benefit from pre-training with face related tasks (Antipov et al. (2017)). Thus, an evaluation framework of knowledge transfer from related task (face recognition) and from general task (ImageNet) for AAE is performed. Two more experiments are then realized. In the first experiment, the Xception model is pre-trained using the Public Figures Face (PubFig) dataset and in the second experiment, it is pre-trained using CASIA-Web face dataset. An MAE of 2.01 years is obtained when pre-training with CASIA-WEB dataset and it outperforms the MAE of 2.35 when pre-training with ImageNet. However, the MAE reaches 4.07 years when pre-training with PubFig dataset. It can be explained by the fact that PubFig dataset contains less images (58K) than CASIA-Web face dataset (500k).

Pre-training on ImageNet outperforms pre-training using PubFig on age estimation task. Xception has a large number of parameters (22 millions) and a lot of data is needed to train it. Therefore, PubFig dataset does not contain enough data to train the upper layers of the deep network. Consequently, training the lower layers of Xception model using ImageNet recognizing shapes and sizes and refining the upper layers using faces achieves better results. Thus, AAE can benefit more from general task when several millions of images are used in pre-training than face related task when the face related task dataset is small.

5.7. Comparison with existing methods of AAE from face images

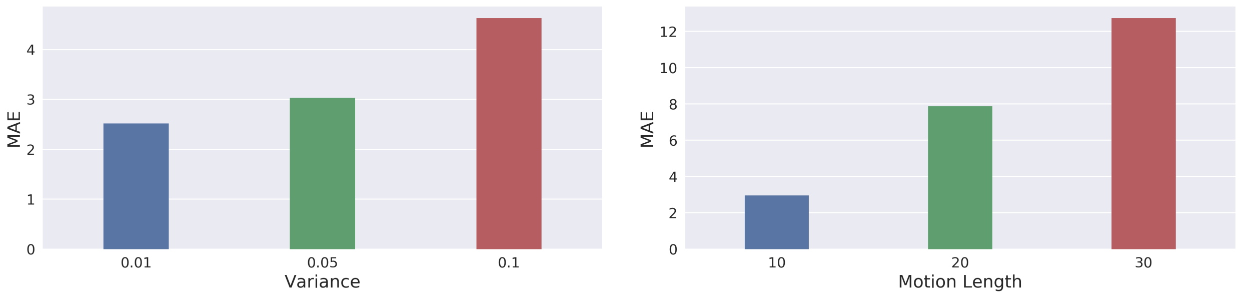
The proposed frameworks with well-known CNN architectures are compared with existing shallow and deep learning based methods tested on Morph II dataset (see Table 6). The state-of-the-art approaches based on hand-crafted features are tested mainly on small datasets and few approaches are tested on Morph II. The reported MAEs of the hand-crafted based methods are between



(a) The average computation time.

(b) Number of trainable parameters per network.

Fig. 4: Computation Complexity



(a) Gaussian Noise.

(b) Blur Noise.

Fig. 5: Robustness of the age estimation framework based on inception model to noise. (a) When adding Gaussian noise of zero mean and variance 0.01 , 0.05 and 0.1. (b) When adding a blur noise of a fixed angle of 45% and a blur length of 10, 20 and 30 pixels.

Table 5: Performance of Xception CNN under extrinsic and intrinsic factors. Three factors are tested. First, AAE under facial expression changes (training on neutral faces and testing on faces with expressions and vice versa). Second, AAE under crossing gender (training on male and testing on female faces and vice versa). Third, AAE under crossing ethnicity (testing one ethnic group after training on the remaining ethnic groups)

Factors	**Datasets**		**Subfactors**	**MAE**	**Average MAE**
	Train	**Test**			
Facial Expressions Crossing	MORPH	FACES	Neutral → With Expressions	10.98	13.43
	FACES	MORPH	With Expressions → Neutral	15.89	
Gender Crossing	MORPH	MORPH	Male → Female	4.33	4.25
			Female → Male	4.18	
Ethnicity Crossing	MORPH	MORPH	H, B, W, O → Asian	2.62	3.47
			A, B, W, O → Hispanic	3.11	
			A, B, H, O → White	3.43	
			A, H, W, O → Black	4.73	

6.49 years (Chang et al. (2010)) and 3.6 years (Han et al. (2015)). Deep models outperform traditional shallow age learning methods and thus high-level semantic ageing patterns extracted by deep learning algorithms perform better than hand-crafted features. With more data and deep networks, AAE can achieve better performances.

As shown in Table 6, Xception architecture performs the best among all the existing approaches for AAE from face images on Morph II dataset. The VGG-16 architecture presented by Antipov et el. (2017) pretrained on half cleaned IMDB-WIKI dataset with the optimal CNN training strategies for AE and on their best protocol of split train/test performs exactly like our Xception based framework without FR transfer learning. The Xception, InceptionV3 and ResNet50 architectures present very good performances comparing to the existing methods. All three architectures present deep networks and could learn more information about the ageing process than other shallow networks.

To the best of our knowledge, our proposed framework based on Xception outperforms all the existing deep CNN architectures in the literature for AAE from face images and achieves an MAE of 2.35 when pre-trained on ImageNet and an MAE of 2.01 when pre-trained on CASIA-Web Face dataset. The approach proposed by Pan et al. (2018) presents the second best MAE of 2.41 because it benefits from not only distribution learning but also the additional constraints introduced to the distribution via mean-variance loss. Pre-training the same model on IMDB-WIKI improves the performance of the network and achieves an MAE of 2.16 which is previously the best reported MAE for AAE in the literature. The work reported in (Rothe et al. (2016)) presents an MAE of 2.68 years and it is achieved due to the additional fine-tuning on the IMDB-WIKI dataset before the fine-tuning on the MORPH dataset. With almost the same number of images used in the pre-training with CASIA-Web Face dataset, our framework based on Xception network achieves a better MAE of 2.01 years. Rothe et al. (2016) approached AAE as 101 age classification problem and to improve the accuracy of the prediction, the softmax expected value is computed.

We can conclude: (1) AAE is complex task which requires a lot of training data; and (2) age encoding strategy is important for AAE but it is possible to reach the best accuracy without using distribution learning based age encoding and its corresponding loss function if the deep network is pre-trained on face related task with very large dataset.

5.8. Robustness to extrinsic and intrinsic factors

Ageing is a non-uniform and non-linear process, and the facial patterns varies from one person to another. Several intrinsic and extrinsic factors influence the age progression. Intrinsic, or chronological ageing, is the inevitable genetically determined process that naturally occurs. The extrinsic, or preventable environmental factors may magnify intrinsic ageing. Most premature ageing is caused mainly but not only by repetitive facial expressions, sun exposure or smoking. The AAE may be influenced by these extrinsic and intrinsic factors. In this section, the impact of three factors on the ageing process is studied using deep learning scenarios. First, we study the AAE under facial expressions changes. Then, the influence of crossing ethnicity and crossing gender on AAE.

AAE under facial expression changes: in previous studies based on shallow learning methods, the AAE under facial expressions changes is studied (Guo and Wang, 2012; Nguyen et al., 2014). It is demonstrated that facial expressions affect the age estimation accuracy. In this study, the cross-expression age estimation is studied and the robustness of the best deep age estimator is verified under facial expression changes. The crossing situations from neutral faces of MORPH to all other expressions available in FACES dataset are considered. The reverse scenario from faces with expressions to neutral faces is also evaluated.

For the first scenario, the Xception architecture is fine-tuned on the neutral faces of MORPH dataset. The trained model is tested over the six face expressions of FACES dataset. The experimental results on FACES dataset is shown in Table 5. An MAE of 10.98 is obtained. The error increases by 8.63 years in average. The performance of the deep age estimator decreases considerably with expression changes.

For the second scenario, the Xception architecture is fine-tuned on the Faces of FG-NET and tested over the neutral faces of MORPH. An MAE of 15.8 years is reached (Table 5). The error is even more important in the second scenario than the first one. Thus, it is more challenging for a deep age estimator trained on faces with expressions to estimate correctly the age of neutral faces.

The cross-expression age estimation using shallow learning framework by considering crossing situations from neutral to only one of the five expressions in FACES dataset (happy, disgust, fearful, sad and angry) shows an MAE varying from 8.66 to 11.87 years (Guo and Wang (2012)). The shallow age estimator gets a big error as in a deep age estimator under facial expression changes. Consequently, the age estimation under facial expression changes is a hard task, the face ageing patterns under facial expression changes is different from the face ageing patterns under neutral faces and they need to be learned for an accurate age estimator.

Another experiment is realized to study the impact of transferring knowledge from facial expression task to age estimation task and its capacity to improve the performance of the deep age estimator under facial expression changes. For that, the Xception architecture trained on the neutral faces of MORPH dataset is fine-tuned on the facial expression images of the extended Cohn-Kanade (CK+) dataset and tested on the FACES dataset. The CK+ includes 593 sequences from 123 subjects. The image sequence varies in duration between 10 and 60 frames and goes from the onset (neutral face) to peak facial expression (Lucey et al. (2010)). In our experiment, we chose to consider the first three frames as neutral faces and from the fourth to the last frame as facial expression. An MAE of 8.86 is achieved and it demonstrated that transferring knowledge from facial expression recognition task helps to learn age jointly with the expression.

An illustration of some of the good and bad results of AAE from FACES dataset is shown in Figure 2, third row.

The influence of crossing ethnicity on AAE: the individuals of the same ethnic group may share mutual face characteristics: skin color, skin texture and facial shape traits. The age estimation performance under variation across ethnicity has been studied based on shallow learning approaches (Guo and Mu, 2010; Ricanek et al., 2009). In (Guo and Mu (2010)), the MAE for "no cross" age estimation is 4.96 years. It increases to 7.41 years in "crossing ethnicity" age estimation. Thus, AAE is affected by ethnicity significantly as its crossing causes large error increase. In this paper, we investigate the performance of deep learning approaches in AAE across ethnic groups. We use MORPH dataset, four tests have been realized:

- training on Asian, Black, White and others and testing on Hispanic,
- training on Asian, black, Hispanic and others and testing on White,
- training on Asian, white, Hispanic and others and testing on black,
- training on black, white, Hispanic and others and testing on Asian.

The results of the different experiments are shown in Table 5. MAEs between 2.62 and 4.73 years are obtained. Testing on Hispanic or white gives very close MAEs, while testing on black gives the higher MAE. Thus, we can empirically conclude that hispanic, white and asian races share some close ageing features. On the other hand, blacks have more unique ageing features that cannot be learned from other ethnicities.

The influence of crossing gender on AAE: patterns in adult human faces reflect the masculinization or feminization that occurs at puberty: larger jawbones, more prominent cheekbones and thinner cheeks and lips are the patterns of male faces that differentiate them from female faces (Little et al. (2011)). The influence of crossing gender on AAE based on shallow learning has been studied in (Guo and Mu (2010)). The average MAE changes to 8.38 years in “cross gender” age estimation. It has been demonstrated that gender affects the AAE in shallow learning scenario. In this paper, we studied the cross-gender age estimation using CNN architecture trained on MORPH dataset. Two experiments are performed:

- training on man face images and testing on women face images. An MAE of 4.33 years is obtained (Table 5),
- training on women face images and testing on man face images. An MAE of 4.18 years is obtained (Table 5).

The deep age estimator performs well across gender and better than shallow learning scenario.

6. Conclusion and Future Works

In this paper, we studied the performance of several frameworks based on CNN architectures. The framework based on Xception network outperforms the state-of-the-art methods based on deep or shallow learning for automatic age estimation with an MAE of 2.35 years when pre-trained on ImageNet and an MAE of 2.01 when pre-trained on CASIA-Web face dataset. The knowldge transfer evaluation demonstrates that AAE can benefit more from general tasks when several millions of images is used in pre-training than face related task when the face related task dataset is small. The layer-wise transfer learning evaluation demonstrates that the optimal number of layers to fine-tune on AAE task depends on the depth of the network, on its architecture and on its layers types. Using the best deep age estimator, we investigate the effects of gender, expression changes and ethnicity on AAE. Despite the fact that deep CNNs have improved the performance of AAE, we confirmed that facial expressions crossing affects considerably age estimation. On another hand, the deep estimator performs well and almost equally across gender. The experimental results showed that testing on black faces and training on other ethnicity present the highest error in the crossing ethnicity tests. We can conclude that they have a more unique ageing patterns than others. In future work, we are interested in studying age estimation from face images under occlusion and different illumination conditions.

References

Wang, X., Guo, R., Kambhamettu, C., 2015. Deeply-learned feature for age estimation. 2015 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 534-541.

Yi, D., Lei, Z., Li, S. Z., 2014. Age estimation by multi-scale convolutional network. Asian conference on computer vision, pp. 144-158.

Dong, Y., Liu, Y., Lian, S., 2016. Automatic age estimation based on deep learning algorithm. Neurocomputing, 187, 4-10.

Niu, Z., Zhou, M., Wang, L., Gao, X., Hua, G., 2016. Ordinal regression with multiple output cnn for age estimation. IEEE conference on computer vision and pattern recognition, pp. 4920-4928.

Rothe, R., Timofte, R., Van Gool, L., 2016. Deep expectation of real and apparent age from a single image without facial landmarks. International Journal of Computer Vision, 126(2-4), pp. 144-157.

Yang, H. F., Lin, B. Y., Chang, K. Y., Chen, C. S., 2013. Automatic age estimation from face images via deep ranking. networks, 35(8), pp. 1872-1886.

Ranjan, R., Zhou, S., Cheng Chen, J., Kumar, A., Alavi,

Table 6: Comparison of the age estimation MAEs by the proposed CNN architectures and the state-of-the-art methods on MORPH II dataset. * The IMDB-WIKI used for pre-training. ** The CASIA-WEB Face used for pre-training. § The PubFig used for pre-training. † Half cleaned IMDB-WIKI used for pre-training. TL: transfer Learning, MV loss: mean-variance loss, DR: Dimensionality Reduction

Methods	**Approach**	**MAE**
Chang et al. (2010)	Ranking	6.49
Chang et al. (2011)	OHRANK	6.07
Guo and Mu (2011)	KPLSR	4.18
Han et al. (2013)	BIF	4.2
Han et al. (2015)	Boosting	3.6
Yang et al. (2013)	ScatNet+PCA+FC	3.49
Yi et al. (2014)	Multi-scale CNN	3.63
Wang et al. (2015)	CNN+DR+classif	4.77
Niu et al. (2016)	Multiple CNNs	3.27
Han et al. (2018)	DMTL	3.0
Rothe et al. (2016) *	VGG-16	2.68
Pan et al. (2018)	MV loss+CNN	2.41
Pan et al. (2018) *	MV loss+CNN	2.16
Antipov et al. (2017) †	VGG-16 + LDAE	2.99/2.35
Proposed (Xception)	Xception	2.35
Proposed **	Xception + TL	**2.01**
Proposed §	Xception + TL	4.07
Proposed	InceptionV3	2.47
Proposed	ResNet50	2.53
Proposed	VGG-16	4.43
Proposed	VGG-19	3.14

Table 7: The generalization of the pre-trained CNNs on Morph dataset to other dataset. To evaluate the performance of the different CNN architectures trained on Morph, the model is tested on FG-NET dataset.

Network	**Unfreezing percentage**	**MAE**	
		Without fine-tuning	**With fine-tuning**
Xception	100%	12.03	**3.67**
ResNet50	100%	11.64	3.77
InceptionV3	100%	13.26	4.08
VGG19	75%	14.35	4.98
VGG16	0%	**10.82**	6.02

A., Patel, V. M., Chellappa, R., 2015. Unconstrained age estimation with deep convolutional neural networks. IEEE international conference on computer vision workshops, pp. 109-117.

Liu, X., Li, S., Kan, M., Zhang, J., Wu, S., Liu, W., ..., Chen, X., 2015. Agenet: Deeply learned regressor and classifier for robust apparent age estimation. IEEE International Conference on Computer Vision Workshops, pp. 16-24.

Guo, G., Mu, G., Fu, Y., Huang, T. S., 2009. Human age estimation using bio-inspired features. IEEE Conference on Computer Vision and Pattern Recognition CVPR, 2009, pp. 112-119.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. ICLR, 2015.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. IEEE conference on computer vision and pattern recognition, pp. 770-778.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2016. Rethinking the inception architecture for computer vision. IEEE conference on computer vision and pattern recognition, pp. 2818-2826.

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. arXiv preprint, 1610-02357.

Bruna, J., Mallat, S., 2013. Invariant scattering convolution networks. IEEE transactions on pattern analysis and machine intelligence, 35(8), pp. 1872-1886.

Ebner, N. C., Riediger, M., Lindenberger, U., 2010. FACES - A database of facial expressions in young, middle-aged, and older women and men: Development and validation. Behavior Research Methods, 42(1), pp. 351-362. doi:10.3758/BRM.42.1.351.

Ricanek, K., Tesafaye, T., 2006, April. Morph: A longitudinal image database of normal adult age-progression. Automatic Face and Gesture Recognition, 2006, pp. 341-345.

Panis, G., Lanitis, A., Tsapatsoulis, N., Cootes, T. F.,

2016. Overview of research on facial ageing using the FG-NET ageing database. Iet Biometrics, 5(2), pp. 37-46.

Minear, M., Park, D. C., 2004. A lifespan database of adult facial stimuli. Behavior Research Methods, Instruments and Computers, 36(4), pp. 630-633.

Pan, S. J., Yang, Q., 2010. A survey on transfer learning. IEEE Transactions on knowledge and data engineering, 22(10), pp. 1345-1359.

Guo, G., Wang, X., 2012. A study on human age estimation under facial expression changes. IEEE Conference onComputer Vision and Pattern Recognition, pp. 2547-2553.

Nguyen, D. T., Cho, S. R., Shin, K. Y., Bang, J. W., Park, K. R., 2014. Comparative study of human age estimation with or without preclassification of gender and facial expression. The Scientific World Journal, 2014.

Guo, G., Mu, G., 2010. Human age estimation: What is the influence across race and gender?. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 71-78.

Ricanek, K., Wang, Y., Chen, C., Simmons, S. J., 2009. Generalized multi-ethnic face age-estimation. IEEE 3rd International Conference on Biometrics: Theory, Applications, and Systems, pp. 1-6.

Little, A. C., Jones, B. C., DeBruine, L. M., 2011. Facial attractiveness: evolutionary based research. Philosophical Transactions of the Royal Society of London B: Biological Sciences, 366(1571), pp. 1638-1659.

Sagonas, G. Tzimiropoulos, S. Zafeiriou, M. Pantic, 2013. A semi-automatic methodology for facial landmark annotation. IEEE conference on computer vision and pattern recognition workshops, pp. 896-903.

Kazemi, V., Sullivan, J., 2014. One millisecond face alignment with an ensemble of regression trees. IEEE Conference on Computer Vision and Pattern Recognition, pp. 1867-1874.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., ..., Ungar, L. H., 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. PloS one, 8(9), e73791.

Trained CNNs: `https://gitlab.com/TalebAbdulRahman/CNNs_aging_project.git`

Fu, Y., Guo, G., Huang, T.S., 2010. Age synthesis and estimation via faces: A survey. IEEE transactions on pattern analysis and machine intelligence, 32(11), pp. 1955-1976.

Ramanathan, N., Chellappa, R., Biswas, S., 2009. Age progression in human faces: A survey. Journal of visual languages and computing, 15, pp. 3349-3361.

Georgopoulos, M., Panagakis, Y., Pantic, M., 2018. Modeling of facial aging and kinship: A survey. Image and Vision Computing, 80, pp. 58-79.

Gunay, A., Nabiyev, V. V., 2008. Automatic age classification with LBP. IEEE International Symposium on Computer and Information Sciences, pp. 1-4.

Han, H., Otto, C., Jain, A. K., 2013. Age estimation from face images: Human vs. machine performance. IEEE International Conference on Biometrics (ICB), pp. 1-8.

Huerta, I., Fernández, C., Prati, A., 2014. Facial age estimation through the fusion of texture and local appearance descriptors. European Conference on Computer Vision, pp. 667-681.

Lanitis, A., Taylor, C. J., Cootes, T. F., 2002. Toward automatic simulation of aging effects on face images. IEEE Transactions on pattern Analysis and machine Intelligence, 24(4), pp. 442-455.

Fu, Y., Guo, G., Huang, T. S., 2010. Age synthesis and estimation via faces: A survey. IEEE transactions on pattern analysis and machine intelligence, 32(11), pp. 1955-1976.

Chao, W. L., Liu, J. Z., Ding, J. J., 2013. Facial age estimation based on label-sensitive learning and age-oriented regression. Pattern Recognition, 46(3), pp. 628-641.

Choi, S. E., Lee, Y. J., Lee, S. J., Park, K. R., Kim, J., 2011. Age estimation using a hierarchical classifier based on global and local facial features. Pattern Recognition, 44(6), pp. 1262-1281.

Guo, G., Fu, Y., Dyer, C. R., Huang, T. S., 2008. Image-based human age estimation by manifold learning and locally adjusted robust regression. IEEE Transactions on Image Processing, 17(7), pp. 1178-1188.

Eidinger, E., Enbar, R., Hassner, T., 2014. Age and gender estimation of unfiltered faces. IEEE Transactions on Information Forensics and Security, 9(12), pp. 2170-2179.

Guo, G., Mu, G., 2010. Human age estimation: What is the influence across race and gender?. IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, pp. 71-78.

Alnajar, F., Lou, Z., Álvarez, J. M., Gevers, T., 2014. Expression-Invariant Age Estimation. In BMVC.

Guo, G., Zhang, C., 2014. A study on cross-population age estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4257-4263.

Bhattarai, B., Sharma, G., Lechervy, A., Jurie, F., 2016. A joint learning approach for cross domain age estimation. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1901-1905.

Yang, H. F., Lin, B. Y., Chang, K. Y., Chen, C. S., 2018. Joint Estimation of Age and Expression by Combining Scattering and Convolutional Networks. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 14(1), 9.

Pan, H., Han, H., Shan, S., Chen, X., 2018. Mean-variance loss for deep age estimation from a face. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5285-5294.

Huo, Z., Yang, X., Xing, C., Zhou, Y., Hou, P., Lv, J., Geng, X., 2016. Deep age distribution learning for apparent age estimation. In Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp. 17-24.

Han, H., Jain, A. K., Wang, F., Shan, S., Chen, X., 2018. Heterogeneous face attribute estimation: A deep multi-task learning approach. IEEE transactions on pattern analysis and machine intelligence, 40(11), pp. 2597-2609.

Wang, F., Han, H., Shan, S., Chen, X., 2017. Deep multi-task learning for joint prediction of heterogeneous face attributes. In 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), pp. 173-179.

Moschoglou, S., Papaioannou, A., Sagonas, C., Deng, J., Kotsia, I., Zafeiriou, S., 2017. Agedb: the first manually collected, in-the-wild age database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 51-59.

Chen, B. C., Chen, C. S., Hsu, W. H., 2015. Face recognition and retrieval using cross-age reference coding with cross-age celebrity dataset. IEEE Transactions on Multimedia, 17(6), pp. 804-815.

Ni, B., Song, Z., Yan, S., 2009. Web image mining towards universal age estimator. In Proceedings of the 17th ACM international conference on Multimedia, pp. 85-94.

Phillips, P. J., Wechsler, H., Huang, J., Rauss, P. J., 1998. The FERET database and evaluation procedure for face-recognition algorithms. Image and vision computing, 16(5), pp. 295-306.

Sim, T., Baker, S., Bsat, M., 2002. The CMU pose, illumination, and expression (PIE) database. In Proceedings of Fifth IEEE International Conference on Automatic Face & Gesture Recognition, pp. 53-58.

Chang, K. Y., Chen, C. S., Hung, Y. P., 2010. A ranking approach for human ages estimation based on face images. In 2010 20th International Conference on Pattern Recognition, pp. 3396-3399.

Chang, K. Y., Chen, C. S., Hung, Y. P., 2011. Ordinal hyperplanes ranker with cost sensitivities for age estimation. In CVPR 2011, pp. 585-592.

Han, H., Otto, C., Liu, X., Jain, A. K., 2015. Demographic estimation from face images: Human vs. machine performance. IEEE transactions on pattern analysis and machine intelligence, 37(6), pp. 1148-1161

Guo, G., Mu, G., 2011. Simultaneous dimensionality reduction and human age estimation via kernel partial least squares regression. In CVPR 2011, pp. 657-664.

Antipov, G., Baccouche, M., Berrani, S. A., Dugelay, J. L., 2017. Effective training of convolutional neural networks for face-based gender and age prediction. Pattern Recognition, 72, pp. 15-26.

Gao, B.-B., Xing, C., Xie, C.-W., Wu, J., Geng, X., 2017. Deep label distribution learning with label ambiguity. IEEE Trans. on Image Processing, 26(6), pp. 2825-2838.

Chen, S., Zhang, C., Dong, M., Le, J., Rao., M., 2017. Using ranking-CNN for age estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5183-5192.

Lou, Z., Alnajar, F., Alvarez, J. M., Hu, N., Gevers, T., 2017. Expression-invariant age estimation using structured learning. IEEE transactions on pattern analysis and machine intelligence, 40(2), pp. 365-375.

Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I., 2010. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression., in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, pp. 94–101.

Tajbakhsh, N., Shin, J. Y., Gurudu, S. R., Hurst, R. T., Kendall, C. B., Gotway, M. B., Liang, J., 2016. Convolutional neural networks for medical image analysis: Full training or fine tuning?. IEEE transactions on medical imaging, 35(5), pp. 1299-1312.

Khan, N. M., Abraham, N., Hon, M., 2019. Transfer learning with intelligent training data selection for prediction of Alzheimer's disease. IEEE Access, 7, pp. 72726-72735.