Computers and Electrical Engineering 68 (2018) 337–347

Contents lists available at ScienceDirect

Computers and Electrical Engineering

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

Age estimation using deep learning☆

Soumaya Zaghbani[a,*], Noureddine Boujneh[b], Med Salim Bouhlel[c]

[a] SETIT, ENIG, Gabes, Tunisia
[b] FSG, Gabes, Tunisia
[c] SETIT, High Institute of Biotechnology, Sfax, Tunisia

ARTICLE INFO

Keywords:
Age estimation
Deep learning
Supervised autoencoder
Softmax classifier
Features extraction

ABSTRACT

Age has always been an important attribute of identity. It also has been an important factor in social interaction. The posture, vocabulary, facial wrinkles and the intonation are all elements that facilitate the prediction of the user's age. Age estimation from the face by numerical analysis finds many potential applications such as the development of intelligent human-machine interfaces and improvement of safety and protection in various sectors such as transport, security and medicine. In many works, researchers are particularly interested in the face's features to regress the age. Recent advances in Artificial Intelligence (AI) and particulary Deep Learning (DL) techniques increase motivations to use this methods to estimate age. In this work, we present a novel method for age estimation from a facial images based on autoencoders. Autoencoder is an artificial neural network used for unsupervised learning of efficient coding. Its aim is to learn a representation for a set of data. The purpose of this work is to exploit the performance of autoencoders to learn features in a supervised manner to estimate user's age. We use MORPH, FG-NET datasets to test the performance of our proposed method. Experimental results show the robustness and effectiveness of the proposed method through the MAE (Men Average Error) rate showing a value of 3.34% for MORPH dataset and 3.75% for FG-NET.

1. Introduction

In recent decades, with the growing need to automate recognition and surveillance systems, analysing human faces become the most topics of interest in computer vision field such as face detection, gender classification, and facial expression recognition. In this framework, researches in age estimation techniques have been progressed [1]. Recently, human age presents an important indication on the target user in many application such as age-based access control, age estimation in crime investigation, age adaptive targeted marketing, electronic customer, relationship management, entertainment for recommended systems, security access control, biometrics, precision advertising, intelligent surveillance, internet access control, ethnicity classification, social media analysis, medicine, psychology, and Human Computer Interaction (HCI) [2,3]. In fact, the challenge of novel HCI now is not only to provide natural and simple interaction with the interfaces but also to adapt itself to the user's cognition, sensory and cultural abilities. For the purpose to improve reliability and easiness, these new interactive interfaces search to automatically learn and adapt itself to the user's natural abilities and features [4]. As a result, the interface needs to gather the maximum information about the user to improve the quality of experience. In this framework, age estimation is deemed to be an important indicator of user identity. Indeed, age present a principal channel of nonverbal information which could help researchers in this domain to improve the performance and the efficiency of novel

☆ Reviews processed and recommended for publication to the editor-in-chief by guest editor Dr. Y. Zhang.

* Corresponding author.
E-mail address: soumaya.zaghbani@gmail.com (S. Zaghbani).

https://doi.org/10.1016/j.compeleceng.2018.04.012
Received 29 October 2017; Received in revised form 17 April 2018; Accepted 18 April 2018
Available online 25 April 2018
0045-7906/

interfaces [4]. In general, the automatic estimation of age by a machine is useful in applications where the objective is to determine the age of an individual without previously identifying it [5]. Estimating age from facial images is a complex task: firstly, an age label can be considered as an individual class, so we talk about an important number of classes. Secondly, on the learning side, this problem is much more difficult than those of face detection and gender estimation, and binary classifiers, well studied, may not be applied directly. Also, two different people can grow differently, since the aging process is determined not only by the genes of the person, but also by many external factors such as health, lifestyle, environment and conditions. Another problem is that from adulthood to old age, the most noticeable change is the aging of the skin (change in texture) and changes in the form of continues wrinkles [6]. Finally, men and women can age differently. In general, age estimation can be seen as a classification or a regression problem depending on whether the age groups are considered to be classes or age as a continuous variable [7].

Since the 1970s, AI researchers have been trying to unravel the secrets of how the human brain works in the hope that one day it will be possible to replicate it virtually and transfer it to machines that would become as intelligent and autonomous as humans. This scenario allowed researchers to start designing more intelligent machines [19]. In fact, before 2006, we did not know how to train a deep architecture: the iterative optimization converged towards local minima of poor quality. Neural networks with more than two hidden layers randomly initiated therefore give worse results than ordinary networks. Deep Neural Networks (DNN) are no more than multilayered networks with classical architecture, they have several hidden layers. Indeed, even if the approximation theorems of the late 1980s assert that a single hidden layer is sufficient for the approximation of any sufficiently regular function, there is nothing to prevent a prioritize the implementation of learning by propagation in networks with several hidden layers. In this context, we present in this work an age regression system witch exploit the robustness and effectiveness of DNN to estimate the user's age with high accuracy by learning robust and discriminative features from the data [8]. The work consists of two modules. The first one is to extract from the image the facial area and adjust the user's head position to improve the classification task. The second module provides an estimation of age based on the facial features as well as an off-line learned model. It is obvious that the overall performance of the system depends on the performance of the two modules. This paper is organized as follows: in the next section, related works are described. In Section 3, we present the autoencoders models existing in literature. In Section 4, proposed method is described. In Section 5 results are reported. Finally, conclusions are drawn in Section 6.

2. Related works

Several methods have been reported in the literature to automatically estimate the user's age. A typical schema of the existing methods for age estimation usually consists of two phases [7]: the first one is extracting image features and representations for age, and the second one is learning an age estimator with these image features. The procedure of age estimation focus to label a facial image automatically with the exact age or the age group of the individual face and extracting significant features from facial faces improve the accuracy of age recognition. In 1998 Kwon and all [9] used a small dataset with 47 images to estimate ages using the anthropometric models and wrinkle pattern. In 2002 Lanitis [10] used Active Appearance Models (AAM) for an age representation, according to the experiment results some part of the face and in particular the area around the eyes present a significant source of information for the task of automatic age estimation. In 2007 Geng and al [11], used the MORPH and FG-Net dataset for Linear Discriminant Analysis (LDA) and KNN (k-Nearest Neighbors) age classification. The method named AGES presented in this work uses the representative subspace to model the aging pattern. In 2015 Jhony K. and al [12] use also the AAM, the MORPH and FG-Net datasets to predict ages of four groups between 1 and 77 years. They proposed a framework of AAM, Local Binary Patterns (LBP) and Gabor wavelet to extract highly discriminating features. One year later Dat Tien and al [3] used Weighted Multi-level Local Binary Pattern (WMLBP) based on a fuzzy logic system for features extraction and Support Vector Machine (SVM) to classify ages of 637 participants, ranging from 18 to 93 years old. The most disadvantage of AAM in previous works was the influence of eliminating nose in areas of hair, background and non-uniform on their performance. They proposed in their article a novel method based on WMLBP to avoid these disadvantages cited previously (such us illumination background). Since 2005 the expression DL was introduced to the Machine Learning. DL is a set of automatic learning of methods, attempting to model with a high level of abstraction of data. Many researchers exploited the performance of DL in facial analysis to regress and estimate the user's age [15]. Ivan Huerta and al [13] presented in their work a framework based on two phases: in the first one they used (Histogram of Oriented Gradients) HOG, LBP and Speeded Up Robust Features (SURF) for features extraction and in the second one they used the Convolutional Neural Network (CNN) to incorporate types of layers (convolutional and pooling layers) to extract important features and other layer (locally and fully connected neural) to relate globally information and learn data. In [14] Yuan Dong and all used the Deep Convolutional Neural Network (DCNN) for features extraction to predict age of the facial image. They used the transfer learning strategy to label face image. In [15] Paul Rodrõguez, and al presented a novel method for age and gender recognition. The purpose was to increase the CNN's performance and robustness to image deformation. They presented a pipeline of an attention network which estimates the most informative patches in the low-resolution image. Experiment results had shown the robustness of this method to clutter and deformation. In [16] Hao Liu and al used Ordinal Deep Features Learning (ODFL) architecture to estimate the user's age. They used the DCNN to learn features directly from the image pixels. Experiment results showed the effectiveness of this method. Xiao long Wang and al [17] proposed a new framework for image-based age estimation, they used an unsupervised learned aging feature via convolutional sparse coding .To improve the robustness of this method, standard derivation pooling was applied to the features maps generated by convolutional sparse coding. Experimental results showed better results than traditional Max pooling schemes. Zakariya Qawaqneh and al proposed [18] a new algorithm for the classification of human's age and gender throw the speech and face images using CNN for feature extracting and classification. Some related works, methods, age ranges and the most used dataset in age estimation could be summarized by the Table 1.

Table 1
Related works, most used datasets and age groups.

Year	Authors	Datasets	Age groups	Methods	Facial pose
1998	Kwon and Al [9]	Small private dataset	Child, young and senior adults	Anthropometric models and wrinkle patterns	Frontal face
2002	Lanitis and Al [10]	250 images for training and 80 for testing	Range A(0–18) Range B (19–30)	AMM	Frontal face
2007	Xin and Al [11]	FG-NET and MORPH	0–69 for FG-NET and 15–69 for MORPH	KNN and SVM	Frontal face
2010	Dehshibi and Al [1]	498 gray-scale images. 298 for training and 200 for testing	(AG1), under15 (AG2), 16–30 (AG3), 31–50 (AG4). over 51	The geometric ratios of facial components	Anthropometric models for young ages
2015	Jhony and Al [5]	FG-Net and MORPH	A (0–13), B (14–21), C (22–39), D (40–77)	AAM, LBP, GW, SVM	Frontal face
2015	Ivan and Al [13]	MORPH and FRGC	20–55	DCNN	Frontal face
2016	Dat and Al [3]	MORPH and Pal	17–65	WMLBP,GWand SVM	Frontal face
2017	Yuan and Al [14]	Private Dataset a dataset contains 8025 persons.	0–2, 3–7, 8–12, 13–19, 20–36, 37–65, 65 +	DCNN	Frontal face
2017	Pau and Al [15]	Adience and IoG datasets.	Adience: (0–2 4–6 8–13 15–20 25–32 {38–43, 48–53} 60 +) IOG: 0–2 3–7 8–12 13–19 20–36 37–65 66 +	CNN	Centered, unconcluded faces with common background
2017	Hao and Al [16]	MORPH, FG-NET And FACES	16 to 77 (MORRPH), 0 to 69 (FG-NET), 19 to 80 (FACES)	DCNN	Frontal face
2017	Xiaolong and Al [17]	MORPHFG-NET FG-NET	0–9,10–19,20–29,30–39, 40–49,50–59,60–69	DCNN	Frontal face
2017	Zakariya and Al [18]	Adience	0–2, 4–6, 8–13, 15- 20, 25- 32, 38- 43, 48- 53, 60	DNN	Frontal face

3. Deep neural networks and autoencoders

DNN have already proven to be very effective in various domains, including image recognition, language processing, speech recognition and robotics. The term DL refers to a treatment performed by a large number of artificial neurons, through their interactions allow the system to learn progressively from images, video, audio or other data. In the DL, supervised algorithms such as CNN and DNN, neural networks used labeled examples for features extraction. A typical DNN is constructed by combining a stacked autoencoder, which comprises a desired number of cascaded autoencoder layers with a softmax classifier. For autoencoders networks, features learning phase is unsupervised since it's not using labeled data. The basic architecture of an unsupervised autoencoder is a move forward with an input layer, often one hidden layers and an output layer. Deep autoencoders (DAE), is a DL model, have been widely researched and applied to many domains [15,29], it is built by stacked sparse autoencoders, and the softmax classifier is generally selected as the output layer for classification problem. An autoencoder may be used for pre-training or for dimensionality reduction when the architecture takes the form of a bottleneck. For simplicity, consider an autoencoder with a hidden layer; autoencoder can then learn several levels of representations by stacking hidden layers. It is a features extraction algorithm; it helps to find a representation of data. The features generated by the autoencoders represent the data point better than the points themselves. The purpose is to minimize reconstruction error based on loss function such as the square error [27]. The traditional autoencoders are composed by two stages the first one is the encoding phase; we consider the input vector X transformed into a hidden representation as following:

$$z = f(X) = \mathrm{sigm}(WX + b) \tag{1}$$

Where W and b present the weight matrix and the bias between the input and hidden layer, respectively and the mapping was through sigmoid function or tanh function:

$$\theta(y) = \mathrm{sigm}(y) = \frac{1}{1 + e^{-y}} \quad \text{or} \quad \theta(y) = \tanh(y) = \frac{e^{y} - e^{-y}}{e^{y} + e^{-y}} \tag{2}$$

For the decoding stage, the hidden representation is mapped back to the first representation as following:

$$\widehat{X} = g(z) = \mathrm{sigm}(W'' z + b'') \tag{3}$$

Where W' and b' denote the weight matrix and the bias between the hidden and output layer respectively And g is the reciprocol function of f. The reconstruction error is defined by minimizing the Euclidean cost:

$$\mathrm{argmin}_{W,W'} \| X - \widehat{X} \|_2^2 \tag{4}$$

3.1. Stacked autoencoders

Several derivations of autoencoders were proposed in literature, such as stuck autoencoder; denoising autoencoders, progressive autoencoders and sparse autoencoder [23]. Generally, thr architecture of all these autoencoders is composed by an input layer, over than one hidden layer and output layer. As described previously autoencoders dispose two phases: the first one is the encoding; it is a determinist mapping which transform the input vector X in a hidden representation (here $X = \{x1|x2|x3|....|xn\}$ consists of all the training samples stacked as columns). It takes as input the train or the test vector. This layer is connected to the first hidden layer, for the autoencoder description we suppose that input vector X include the bias term.:

$$f(X) = \theta(WX) \tag{5}$$

Where W presents the weight's matrix and WX is the input vector to the hidden layer. Between the input and output the mapping is a nonlinear function, it stacked input data. In the decoding stage, the hidden representation is mapped back to the reconstruction of input features; the decoder changes obtained data to the initial space using the following equation:

$$\widehat{X} = W''\theta(WX) \tag{6}$$

Stacked autoencoders involving of using multiple layers of autoencoders in which the output of each layer presents the inputs of the successive layer. Once the first layer is pre-trained it can be used as an input of the next autoencoder. The final layer can deal with traditional supervised classification and the pretrained neural network can be fine-tuned using backpropagation. Formally, the stack architectures have multiples layers and could be resolved by minimized the loss function as follow [26]:

$$argmin_{w_1 \ldots W_{l-1},\, w'_1 \ldots W'_l} \| X - (w'_1\, \theta(w'_2 \ldots w'_l(f(x)) \|_F^2 \tag{7}$$

Where f(X): $\theta\,(w_{l-1}\theta\ (w_{l-2} \ldots \theta\ w_1(x)))$ and F refers to Frobenius norm.

3.2. Stacked denoising autoencoder

Denoising autoencoder is an unsupervised feature learning algorithm, and the corrupted data are generated by manually corrupting the clean data with predefined noises. The trained model is robust to these predefined noises. As the name denoising autoencoder is a variant of autoencoders with the ability to learn the distributions of the noise present in the input data, and to reproduce its corresponding uncorrupted version and could be resolved by minimized the loss function as follow [26,27]:

$$argmin_{w_{1\ldots}W_{l-1},\ w'_{1\ldots}W'_{l}}\ \|X-(w'_1\,\theta(w'_2\ldots w'_l(\mathrm{f(x)})\|_F^2+R(W,X) \tag{8}$$

Where f(X): θ (w_{l-1}θ (w_{l-2}…θ $w_1(X)$)) and R is the regularization term. Deep denoising autoencoder approved their interest in many domains such as biology, image processing and speech recognition.

3.3. Stacked progressive autoencoder

The progressive autoencoders aim at mapping the images at large pose to virtual images at smaller pose, while mapping those already at smaller poses to themselves. The minimized function presented as next [28]:

$$argmin_{w_k,w'_k}\ \|X-\ (w'_1\,\theta(w'_2\ldots w'_l(\mathrm{f}(G^{k-1}))\|_F^2 \tag{9}$$

Where K = (1,2,…., L) and G^{k-1} is the representation from the hidden layer of the (k-1) progressive autoencoder and G^0 = X. Stacked progressive autoencoders composed of multiple shallow autoencoders was presented to learn pose-invariant features by smoothly mapping faces to nearby frontal views [28].

3.4. Optimisation algorithms

In their article, Abdullah and al [29] affirme that “the most important factors affecting the classification performance of DNNs is their training”. Indeed, optimization in DNN is a difficut task due to the importance number of parameters specially for autoencoders, therefore, the choice of optimization algorithm is very delicate. Many algorithms for training autoencoders have been proposed, but the most popular ones are gradient-based. Gradient Descent (GD) algorithm applied when we seek the minimum of a function whose analytical expression is known, which is differentiable. It is a fundamental algorithm used under derived forms. The main issue of these algorithms that are not robust where the dimension of the search space is very high [29]. The Stochastic Descent Gradient (SDG) algorithm is a simple algorithm but it has an effective approach for learning linear classifiers only under convex cost conditions. This is particularly the case of SVM [30]. Although this algorithm has been around for a few decades, it has recently again attracted researchers in the context of large-scale learning, it has been successfully applied. Conjugate Gradient (CG) is another algorithm of optimization based on gradient, the methods comprise a class of unconstrained optimization algorithms which are characterized by low memory requirements and strong local and global convergence properties. The main problem of these algorithms is their sensetivity to the noise. Another popular algorithm used by many researchers is the (L-BFGS), it is used for large scale optimization problems. Its is more stable and faster because it needs much less memory than the other algorithms by utilizing the history of gradient evaluations. But they are very complex in implimentation [19,26].

3.5. Softmax classifier

A frequent use of the softmax function appears in the domain of machine learning, especially in logistic regression: we associate with each output possibility a score, which is converted into probability with the softmax function. When a classification task has more than two classes, it is standard to use a softmax output layer. It is the latest layer and it offers a way to predict a discrete probability distribution over the classes [29].

4. Proposed method

The highlight of DAE is the effectiveness to extract useful features by unsupervised learning, it only reserves crucial information of the data in robust and discriminative representations after detecting and removing input redundancies. An autoencoder can simply be defined as a network of artificial neurons that learns a hidden representation in order to reconstruct its inputs. It is therefore an unsupervised architecture where model inputs are similar to outputs, and where the model generaly, optimizes its parameters according to the mean square error [26,27,29]. In this work, we propose a novel method of age estimation based on DL architecture precisely the autoencoder, as described in Fig. 1, the work is divided into two phases; the first one is the data preparation; face detection and in plane rotation of the head. The second one is the classification using supervised autoencoder.

4.1. Face detection and in-rotation compensation

The first step in our work is to extract the face from the images, for this reason we used the AdaBoost framework of Viola P. and Jones M [20] published on July 13, 2001. This algorithm is based on the principle of “boosting” which consists in assembling several weak classifiers in order to obtain a system with high capacity. Rather to work directly on the pixels of the image, Viola P. and Jones M., introduce characteristics called “pseudo-Haar” strongly inspired by wavelets. The result is shown in Fig. 2.

The head position presents an important factor in age estimation. In fact, due to the emotion or the pose when capturing the image head pose could influence the algorithm performance, it is necessary to adjust the orientation. To resolve this problem we propose to adjust the head orientation through the eyes-position. So we extract the eyes position and rotate the image throw this level. After detecting left and right eyes position we apply the following formula to have the rotation angle [15,21].

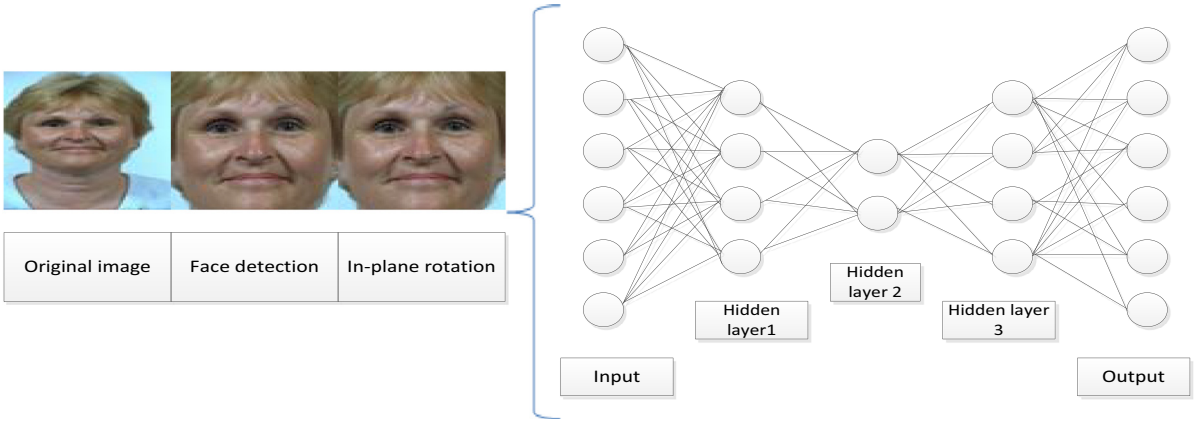

Fig. 1. General pipeline of our proposed method: in this architecture we used the Deep Sparse Supervised Autoencoder (DSSAE) with three hidden layers.

Fig. 2. Cropped face after extraction with AdaBoost algorithm.

$$\Phi = tan^{-1}\left(\frac{R_j - L_j}{R_i - L_i}\right) \tag{10}$$

Where (R_i,R_j) and (L_i,L_j) present the detected positions of right and left eyes respectively. Fig. 3 presents the eyes position, initial head position and the adjusted one.

4.2. Proposed autoencoder: Deep Supervised Sparse Autoencoder (DSSAE)

The goal of an autoencoder is to find some correlation between the input data in order to reduce the size of these data and to classify them. The sparse coding method is another model that seeks a good representation. The term sparse is used to indicate that we want hidden neurons to have the same probability of activation. The number of neurons of the hidden layers is smaller than that of the input and output layers, it will be necessary to compress the data and thus to find a correlation between the data and thus classify them according to this correlation. The minimized function of sparce autoencoder presented as following [22]:

Fig. 3. Eyes detection and cropped face after in plane rotation.

$$\text{argmin}_{W,W'} \|X - W'\theta(WX)\|_F^2 + \beta \sum_{j=1}^{m} KL(\rho \| \hat{\rho}_j) \tag{11}$$

Where m is the number of hidden nodes, and β is a coefficient that determines the weight of sparsity penalty item. ρ is the sparsity parameter, in other words, it stands for the target average activation of hidden units, which is generally a small value nearing zero: $\hat{\rho}_j = \frac{1}{N} \sum_{i=1}^{N} h_i(x_i)$ denotes the average activation of hidden node j, and the Kullback-Leibler divergence is defined as:

$$KL(\rho \| \hat{\rho}_j) = \rho \log \frac{\rho}{\hat{\rho}_j} + (1-\rho) \log \frac{1-\rho}{1-\hat{\rho}_j} \tag{12}$$

The purpose of training in sparse autoencoder is to find convenient parameters to minimize the objective function of equation Eq. (11), where backpropagation algorithms are generally used to train the model. Especially, selecting a suitable number of hidden nodes in the model would play a much more important role in achieving higher performance than the learning algorithm or the depth of the model. Supervised learning is one of the most powerful algorithms in AI, it was very robust in image processing, speech recognition, zip code recognition and pattern recognition; it improves understanding human genomes such as facial detection, gender recognition and facial expression analysis. Many researchers considered that autoencoders may be supervised because they could be more performant when they specify classes, their researchers denotes the performance of this method and results shown the robustness of supervised autoencoders thein the ordinary networks [28,23,26,21]. Specifically, most applications of it still require that we manually specify the input features x given to the algorithm. Once a good feature representation is given, a supervised learning algorithm can do well. The supervised autoencoder makes the features consist to the same age similaraty, in fact, it extracts more efficient features for age face representation. The solution proposed for the classification process in this work is the use of the sparse supervised autoencoder. Firstly, we propose to learn features such us they have the same sparsity signature across the class. In their work Majumdar and al propose l1 norm for regularization presented as follow:

$$\text{argmin}_{W,W'} \|X - W'\theta(WX)\|_F^2 + \lambda \|WX\|_1 \tag{13}$$

Where λ, presents the regularization term using l1-norm regularization.

However, in scenarios where labeled training data is available and can be used for training, the proposed formulation can be applied to encode discriminative information. To pass from the unsupervising to the supervising manner, firstly the training data know should be presented us following:

$$X = \left[x_{1,1} | \ldots | x_{1,n_1} x_{2,1} | \ldots | x_{2,n_2} \ldots x_{c,1} | \ldots | x_{c,n_c}\right]$$

Where the training data is divided into classes (c). The idea is to learn features into common sparse support. and WXi will be row sparse. This is achieved by incorporating $l_{2,1}$ norm regularization as following [28]:

$$\text{argmin}_{W,W'} \|X - W'\theta(WX)\|_F^2 + \lambda \sum_{c=1}^{c} \|WX_c\|_{2,1} + \beta \sum_{c=1}^{c} KL(\rho \| \hat{\rho}_c) \tag{14}$$

Where $v_{2,1} = \sum_j v^{j\,\to}{}_2$ is the sum of l_2-norms of the rows. We indicate that the input X belong to class c during the classification phases, then takeing into consederaion all classes we optimize w and w' leading to a minumum of objectif function. The inner l_2-norm promotes a dense (non-zero) solution within the selected rows, but the outer l1-norm (sum) enforces sparsity in selecting the rows [28]. The proposed formulation shown in Eq. (14) enforces row-sparsity within each group. This makes the optimization supervised i.e, the information regarding the class labels is required to formulate equation. The formulation enforces supervision by constraining that the features from the same group should have the same sparsity signature. To train the DSSAE we use the Scaled Conjugate Gradient (SCG) algorithm because it is efficient, in terms of speed and in terms of capacity control. After the learning phase of the autoencoder layers, the supervised fine-tuning of the overall deep architecture is performed by stacking the autoencoders and the softmax classification layers. The proposed DSSAE described in Fig. 4.

5. Experiment results

This section is dedicated to evaluate the implemented method, and perform the various comparisons with other methods. The implementation method was based on the use of FG-NET and MORPH datasets. MORPH database contains 55,000 unique images of more than 13,000 individuals, spanning from 2003 to late 2007. Ages range from 16 to 77 with a median age of 33. The average number of images per individual is 4 and the average time between photos is 164 days, with the minimum being 1 day and the maximum being 1681 days. The standard deviation of days between images is 180. The evaluation of the performance of our proposed methods is giving in the evaluation of features learning using supervised autoencoder using the FG-NET datasets also [30,24]. The Face and Gesture Recognition Research Network (FG-NET) aging database contains on average 12 pictures of varying ages between 0 and 69, for each of its 82 subjects. Altogether there is a mixture of 1002 color and greyscale images, which were taken in totally uncontrolled environments. Each was manually annotated with 68 landmark points. In addition there is a data file for every image, containing type, quality, size of the image and information about the subject such as age, gender, spectacles, hat, mustache, beard and pose. One particular problem with this dataset is the fact, that images are not equally distributed over age and thus only few images of persons older than 40 are available. The illustration of the facial images used in the test is described in Table 3; the table summary the number of male and female facial images used in the two databases and the interval of age. Fig. 5 presents selected

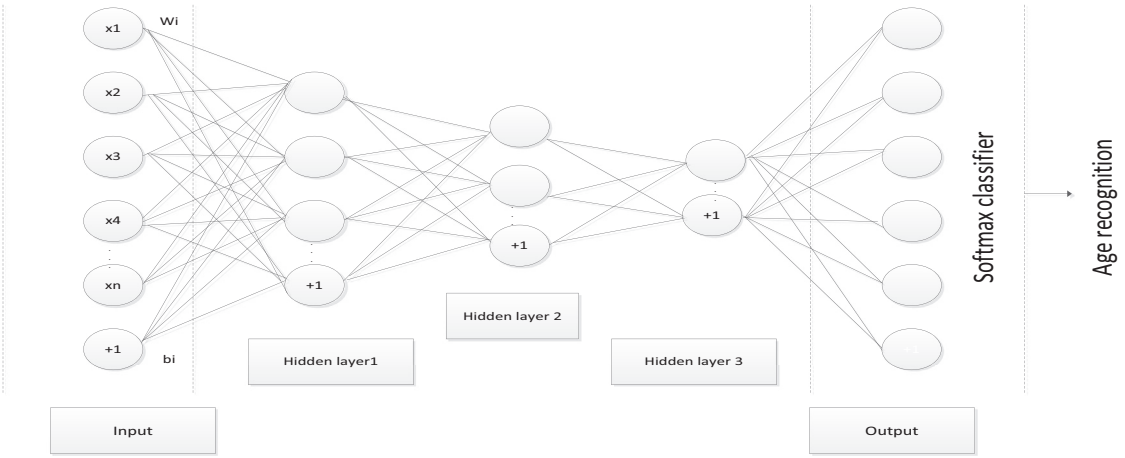

Fig. 4. Architecture of the proposed DSSAE.

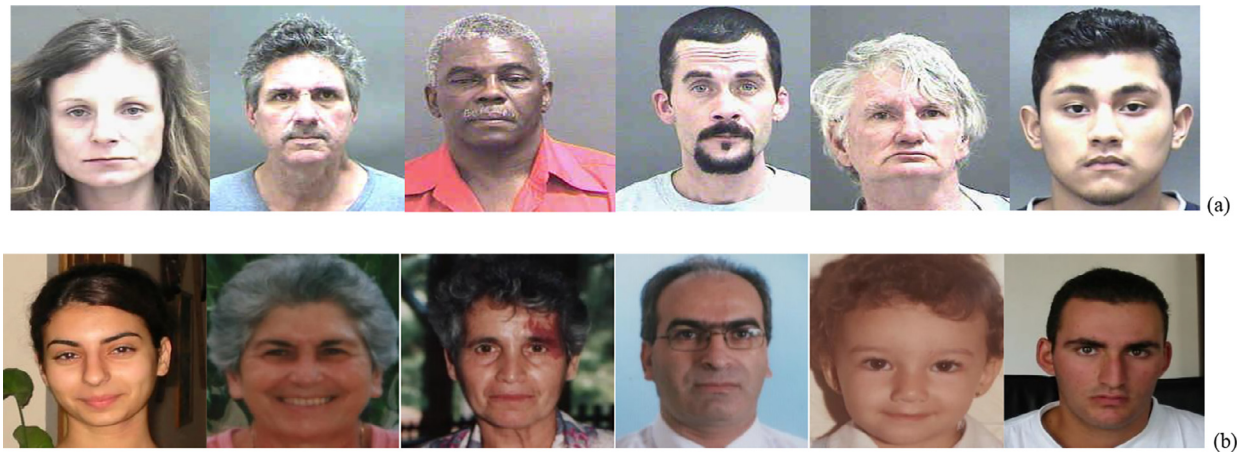

Fig. 5. Examples of face images with different age values from MORPH (a) and FG-NET (b).

examples from the apparent age estimation datasets from the FG-NET and MORPH databases. In this work we use three Layers of size L, L/2 and L/16 full frontal face image with size of 32 × 32. All parameters used in these experiments detailed in Table 2.

The proposed approach requires training and parameter regularization to learn the classifier. In this research, representation learned by using over 50,000 frontal images from the MORPH and 1002 from the FG-NET.

The training sets provided with the respective databases are used to learn the classifier. The contribution of this work is presented into many points; the first one is the effectiveness of the method in age estimation and the second one is the performance of supervised autoencoders compared with unsupervised one. Also, evaluate the performance of the SSAE on the two datasets (MORPH and FG-NET), finally perform extensive testing and analysis of the proposed algorithm in the age estimation domain. For testing and evaluating the proposed approach we use firstly the Mean Absolute Error to define the algorithm performance in age estimation it is calculated as following [12–15,18]:

$$E = \frac{||y' - y||2}{N} \tag{15}$$

Where y' and y present the predicted and real age value respectively and N denotes the number of the testing facial images.

Table 4 lists the MAE results of some methods using different architecture of DL compared with our method. The effectiveness of

Table 2
Parameters used for the proposed work.

Parameters	Layers size	Regularization term (λ)	Sparsity parameter (ρ)	Weight sparsity penalty (β)	Maximum epochs
Layer 1	50	0.002	0.5	5	1000
Layer 2	25	0.003	0.8	5	
Layer 3	5	0.001	0.1	5	

Table 3
Illustration of the used datasets.

	Female	Male	Ages	Number of images
MORPH	5,757	36,832	16–77	55,134
FG-NET	395	607	0–69	1,002

Table 4
Comparison of MAE with the state-of-the-art methods on MORPH and FG-NET datasets.

Method	Year	FG-NET	MORPH
CSC + Max Pooling [19]	2017	4.10	3.78
CSC + STD Pooling [19]	2017	4.01	3.66
ODLF [18]	2017	3.89	3.12
Flexible overlapped AAM + LPQ [12]	2015	4.87	5.68
Sharad Kohli and al [4]	2013	3.83	–
IvanHuerta and al [13]	2015	–	4.25
D2C [24]	2017		3.06
Net VGG [25]	2017		2.96
Hybrid			
GA-DFL [22]	2016	4.16	3.37
Proposed method	–	**3.75**	**3.34**

DL architecture was presented in many previous works such as the work of Ivan Huert and al [13] when authors presented two frameworks for age estimation. In the first one they used HOG and LBP for features extraction and in the second they used the DL. The difference marge in the MAE was very interesting in the deep classification. The result in Table 4 shown the robustness of our method compared with other works, our method presents an important rate of MAE as shown in the table. The proposed method was tested also by age groups; we classify ages into four ranges, Child (0–14) youth (16–30), senior (31–50) and elderly (51–over). Table 5 presents the evaluation of our method in every category using the MAE rate. The approach shown the performance on child and senior groups, the MAE presents the best result with less than 2.88 and 3.92 for the FG-NET dataset and 3.26 for the MORPH one. (MORPH dataset doesn't contain child group).

The proposed method evaluated also using the cumulative score presented as follow [22,24]:

$$\mathrm{CS}(£) = \frac{N_E \leq £}{N} \times 100\% \tag{16}$$

Where $N_E \leq £$ is the number of images on which the error £ is no less than E. basically, £ starts from 0. This cumulative score has been a standard evaluation metric for the age estimation problem since the very beginning of age estimation researches. This evaluation method has been employed in many previous works. Fig. 6 shows the CS curves of facial age estimation method for the ages from 0 to 20 years for the FG-NET database. According to the results, we inferred that our model performed a very competitive performance, which shows the effectiveness of the proposed method.

The DL presents an important success in the machine learning algorithm and researches improve the performance of these machines such as the autoencoders witch was used with supervised manner in this proposed framework. To evaluate this work we compare the result of MAE rate between the proposed DSSAE and Denoising stacked AutoEncoders (DSAE) as noted in the Fig. 7. Results demonstrate that the supervision of autoencoders improve their performance. Since DL models have demonstrated superior performance for many classification problems such as speech recognition, facial recognition, emotion recognition, in this section, we compare the proposed framework with several popular models of classification in the literature. To test the proposed method we used the same framework with both KNN and Wavelet Network (WN) classification. The result shown in Table 6 demonstrates the efficiency of the DL algorithm compared with the classic classification algorithms. DL presents an important rate of MAE comparing with classical classifiers. DNN demonstrate better performance for age estimation than shallow neural networks. Many architecture of autoencoders [26,28] were presented in literature; our proposed method is different from the existing methods in the following points: 1) the above classification methods using autoencoders are unsupervised ones for representation learning while our method is a supervised model for classification; 2) the presented method aim to add various regularization terms to learn better feature

Table 5
MAE results by age groups on MORPH and FGNET datasets.

Method	MORPH	FG-NET
0–14	–	2.88
16–30	3.26	3.92
31–50	3.13	4.16
51-over	3.62	3.69

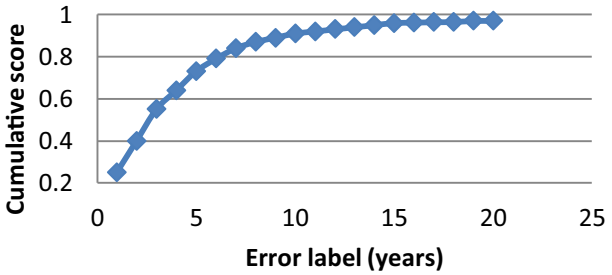

Fig. 6. Cumulative scores of proposed age estimation at error levels from 0 to 20 years on the FG-NET dataset.

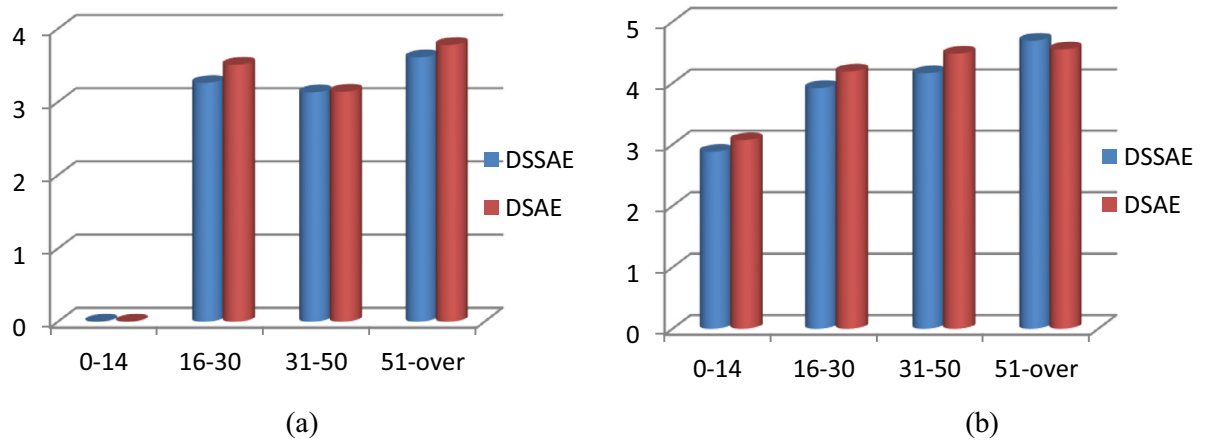

Fig. 7. Comparison between DSSAE and DSAE in ages groups of MORH dataset (a) and FG-NET dataset (b).

Table 6
Comparison of MAE results with KNN and WN classifier.

	FG-NET			MORPH		
Method	KNN	WN	DSSE	KNN	WN	DSSE
0–14	5.91	3.76	2.88	-	-	–
16–30	4.88	4.92	3.92	5.05	4.29	3.26
31–50	5.12	4.91	4.16	4.86	4.53	3.13
51-over	6.87	5.22	4.69	5.49	4.25	3.62

representation. Like shown in Table 4 We can see that our DSSAE outperforms all the rest of the methods, and it achieves the best performance, which proves the effectiveness of SSAE for extracting robust features for age estimation. In fact, besides using more appropriate reconstruction error term, our DSSAE also enforces extracted features with the same age to be similar. Therefore, the performance when the model is trained and tested on the same age or same group of age is generally higher than when the model is trained and tested on different ages. In addition, our DSSAE also outperforms major above methods in the datasets (MORPH and FG-NET). Experimental results clearly demonstrate the superiority of DSSAE for recognition tasks over other DNN methods. In addition The performance with smaller λ (0.01,0.02, or 0.03) also demonstrates the importance of the similarity preservation term. In our objective function optimization, we simply used the SCG to optimize the objective function. Many existing works have shown that different optimization methods will greatly affect the performance of DNN.

6. Conclusion

In this paper, we have presented a novel approach for age estimation. We investigate the autoencoder in a supervised manner using DSSAE to extract high-level features and learn facial features. The proposed formulation aims to learn supervised feature vectors that maximize the intra-class similarity. According to the experiment results and comparison with existing approaches on the FG-Net and MORPH datasets, presented method showcase the effectiveness of the proposed algorithm. On the classification task we considered, our supervised pre-training approach provides better results than the unsupervised approach. The performances obtained by the proposed methods are better than the results in the literature. Our future development axes will concern the combination of speech and facial features using DSSAE for age estimation.

References

[1] MahdiDehshibi M, Bastanfard A. A new algorithm for age recognition from facial images. Sig Process 2010;90(October (2010)):2431–44.
[2] EunChoi S, YounJooLee S, KangRyoungPark J. Age estimation using a hierarchical classifier based on global and local facial features. Pattern Recognit 2010;44(October(2011)):1262–81.
[3] Nguyen DT, Park KR. Enhanced Age Estimation by Considering the Areas of Non-skin and the Non-uniform Illumination of Visible Light Camera Sensor. Expert Syst Appl 2016(March).
[4] Kohli S, Prakash S, Gupta P. Hierarchical age estimation with dissimilarity-based classification. Neurocomputing 2012;120(August (2013)):164–76.
[5] Pontes, J.K., Britto, A.S., Fookes, C., Koerich, A.L., "A flexible hierarchical approach for facial age estimation based on multiple features," Pattern Recognit S0031-3203(15)00445-8, December 2015.
[6] Fergusonab E, Wilkinson C. Juvenile age estimation from facial images. Sci Justice 2016(August).
[7] Sai PK, Wang J-G, Teoh E-K. Facial age range estimation with extreme learning machines. Neurocomputing 2014(March).
[8] Tian Q, Chen S. E cross-heterogeneous-database age estimation through correlation representation learning. Neurocomputing 2016(July).
[9] Kwon YH, da Vitoria Lobo N. Age classification from facial images. Comput Vis Image Understanding 1999;74(1):1–21.
[10] Lanitis, A., "On the significance of different facial parts for automatic age estimation,' Department of Computer Science and Engineering, Cyprus College, P.O. Box 22006, Nicosia, Cyprus. 2002.
[11] Geng X, Zhou Z-H, Smith-Miles K. Automatic age estimation based on. IEEE Trans Pattern Anal Mach Intell 2007;29(December (12)).
[12] Pontes JK, Britto AS, Fookes C, Koerich AL. A flexible hierarchical approach for facial age estimation based on multiple features. Pattern Recognit 2015(December).
[13] Huertaa I, Fernándezb C, Segurab C, Hernandoc J, Pratia A. A deep analysis on age estimation. Pattern Recognit Lett 2015;68(June):239–49.
[14] Dong Y, Liu Y, Lian S. Automatic age estimation based on deep learning algorithm. Expert Syst Appl 2016;187(April).
[15] Rodrõguez P, Cucurull G, Gonfaus JM, Xavier Roca F, Gonzalez J. Age and gender recognition in the wild with deep attention. Pattern Recognit 2017;72(December):563–71.
[16] Liua H, Lua J, Fenga,b J, Zhoua J. Age and gender recognition in the wild with deep attention. Pattern Recognit 2017(October).
[17] Wang1 X, Li1 R, Zhou1 Y, Kambhamettu C. A study of convolutional sparse feature learning for human age estimation. 12th IEEE international conference on automatic face & gesture recognition. 2017.
[18] Qawaqneh Z, Mallouh AA, Barkana BD. Age and gender classiþcation from speech and face images by jointly fine-tuned deep neural networks. Expert Syst Appl 2017(January).
[19] Zaghbani S, Boujneh N, Bouhlel MS. Real time hand gesture recognition using features extraction. International conference on machine vision. 2016.
[20] Viola P, Jones M. Robust real-time face detection. Int J Comput Vis 2004;57(2):137–54. 2004.
[21] Lu J, Tan YP. Ordinary preserving manifold analysis for human age and head pose estimation. IEEE Trans Hum Mach Syst 2013;43(MARCH (2)).
[22] Majumdar A, Singh R, Vatsa M. Face verification via class sparsity based supervised encoding. Trans Pattern Anal Mach Intell 2016(c):0162–8828.
[23] Guo Y, Liu Y, Oerlemans A, Lao S, Wu S, Lew MS. Deep learning for visual understanding: a review. Neurocomputing 2015;187(November (2016)):27–48.
[24] Li K, Xing J, Hu W, Maybank SJ. D2C: deep cumulatively and comparatively learning for human age estimation. Pattern Recognit 2017(January). S0031-3203(17)30009-2.
[25] Xing J, Li K, Hu W, Yuan C, Ling H. Diagnosing deep learning models for high accuracy age estimation from a single image. Pattern Recognit 2017(January). S0031-3203(17)30007-9.
[26] Soto Vega PJ, Feitosa RQ, Ayma Quirita VH, Happ PN. Single sample face recognition from video via stacked supervised auto-encoder. 2016 29th SIBGRAPI conference on graphics, patterns and images. 2016.
[27] Vincent P, Larochelle H, Lajoie I, Bengio Y, Manzagol P-A. Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion. J Mach Learn Res 2010;11(2010):3371–408.
[28] Kan M, Shan S, Chang H, Chen X. Stacked progressive auto-encoders (SPAE) for face recognition across poses. Computer vision and pattern recognition (CVPR), IEEE conference. 2014.
[29] Caliskan, A., Yuksel, M.E., Badem, H., Basturk, A., "Performance improvement of deep neural network classifiers by a simple training strategy", Eng Appl Artif Intell 67 (2017) 14–23.
[30] Badem H, Basturk A, Yuksel ACME. A new efficient training strategy for deep neural networks by hybridization of artificial bee colony and limited–memory BFGS optimization algorithms. Neurocomputing 2017;266(November 29):506–26.

Soumaya Zaghbani is currently a PhD student at National Engineering School of Gabes. She is attached to the research Lab SETIT. She received the Diploma of Master on Computer sciences and Multimedia from the Higher Institute of computer sciences and multimedia of Gabes Tunisia (2011). Currently, she is a research member in the Research Unit of Sciences and Technologies of Image and Telecommunications (SETIT).

Noureddine Boujnah is actually assistant professor at Gabes university, he held his PhD from Polytechnic of Turin Italy in satellite communication in 2011, matser of science in signal processing from Higher school of communication of Tunis (Sup'Com-) and Engineering degree in telecommunication from Sup'com too, he carried out his postdoctoral research activities at Lodz technical university-Poland.

Mohamed Salim BOUHLEL he is a full professor at Sfax University Tunisia. He is the Head of the Research Lab SETIT since 2003. He was the director of the higher Institute of Electronics and Communications of Sfax Tunisia (ISECS) 2008-201. He received the golden medal with the special appreciation of the jury in 1999 on the occasion of the first International Meeting of Invention, Innovation and Technology (Dubai, UAE). He was the vice president and founder member of the Tunisian Association of the specialists in Electronics and the Tunisian Association of the experts in Imagery. He is the president the Tunisian association in Human- Machine Interaction since 2013.