applied sciences

MDPI

Article

A New Cost Function Combining Deep Neural Networks (DNNs) and *l*2,1-Norm with Extraction of Robust Facial and Superpixels Features in Age Estimation

Arafat Abu Mallouh [1,†], Zakariya Qawaqneh [2,†] and Buket D. Barkana [3,*]

[1] Computer Science Department, Manhattan College, Riverdale, NY 10471 USA; aabumallouh01@manhattan.edu
[2] Department of Computing Sciences, The College at Brockport State University of New York, Brockport, NY 14420, USA; zqawaqneh@brockport.edu
[3] Electrical Engineering Department; University of Bridgeport, Bridgeport, CT 06604, USA
[*] Correspondence: bbarkana@bridgeport.edu; Tel.: +1-203-576-4577
[†] These authors contributed equally to this work.

Received: 21 August 2018; Accepted: 25 September 2018; Published: 16 October 2018

Featured Application: Facial images provide variety of information about a person such as personal identity, age, gender, color, ethnicity, head pose, eye gaze, and emotion. Age estimation can be used to build real-world applications in biometrics, law enforcement, surveillance, and human-computer interaction (HCI).

Abstract: Automatic age estimation from unconstrained facial images is a challenging task and it recently has gained much attention due to its wide range of applications. In this paper, we propose a new model based on convolutional neural networks (CNNs) and *l*2,1-norm to select age-related features for the age estimation task. A new cost function is proposed. To learn and train the new model, we provide the analysis and the proof for the convergence of the new cost function to solve minimization problem of deep neural networks (DNNs) and the *l*2,1-norm. High-level features are extracted from the facial images by using transfer learning, since there are currently not enough large age databases that can be used to train a deep learning network. Then, the extracted features are fed to the proposed model to select the most efficient age-related features. In addition, a new system that is based on DNN to jointly fine-tune two different DNNs with two different feature sets is developed. Experimental results show the effectiveness of the proposed methods and achieved the state-of-art performance on a public database.

Keywords: age estimation; face images; DNNs; facial and superpixel features; *l*2,1-norm

1. Introduction

Age estimation from face images is considered as predicting or classifying the real or the apparent age of a person. Facial image estimation is an important step in a wide range of academic and commercial applications [1]. Recently, finding semantic information from images has gained the attention of the research studies due to the vast number of images, which are added on the internet daily or being stored on personal phones and computers. Facial images provide variety of information about a person, such as personal identity, age, gender, color, ethnicity, head pose, eye gaze, and emotion.

Designing a robust age estimation system has many challenges. People do age at different rates and they have different aging patterns that can be affected by genetic factors, social conditions, race, ethnicity, and life style [2]. Some people can look years younger than their chronological age while some can look years older. In most cases, the factors that affect aging are personal and difficult to control [3,4]. There are dissimilarities between aging rates of men and women. Women may put makeup to hide aging marks and may wear accessories, resulting in a younger look [5]. Building an efficient large database containing millions of face images for age estimation is relatively very difficult, since it requires an access to participants' private information and the collected images need to be manually labeled.

Age estimation can be used to build real-world applications in biometrics, law enforcement, surveillance, and human-computer interaction (HCI). There are two types of biometric systems that are unimodal and multimodal [6]. In a multimodal biometric system, a combination of several traits is collected and one of the traits in this system is the age [7]. Recently, with the astronomical growth of internet content and the spread of the smart machines, age estimation can be used in law enforcement, surveillance, and control applications. For instance, a person's access to cigarettes or alcohol from vending machines can be restricted based on his/her estimated age. Companies, such as Google [8] and Amazon [9], offer products and services based on customers' personalized experience and surfing history. Information of customers' age can be used as one of the distinctive factors of the customer's profiles.

Up until a few years ago, low-level features were extracted in most cases and they were used to represent visual image information to design a model for age estimation. Currently, complex and high-level features can be extracted by using deep neural networks (DNNs) and convolutional neural networks (CNNs) to represent the visual information about one's age. In this paper, two high-level feature sets are extracted by using CNNs. Existing benchmarks for age classification task are relatively small when compared to the other benchmarks that are used for other classification tasks, such as image classification, semantic segmentation, and face recognition. Therefore, in this work, pre-trained CNNs from other image tasks are employed to extract the feature sets for age estimation. We propose a regularization framework that is based on a new cost function for selecting the robust features from the original feature sets by combining the DNN and the $l2,1$-norm. The analysis and the proof of the convergence of the new cost function to solve the minimization problem of the DNN and the $l2,1$-norm are presented. Efficiency of the selected robust features are studied. The jointly fine-tuned CNNs based on the Softmax and the Sigmoid (JFN-SS) in [10] are used and a new jointly fine-tuned framework based on amplifying the output of two DNNs (JFN-A) is introduced.

Several methods have been proposed in the literature to represent the age information from the facial images. One of the early models is the anthropometric model that uses the size and proportions of the human face. Kwon and Lobo [11] studied the cranio-facial changes in feature-position ratios and skin wrinkles as features for three age groups; baby, young adult, and senior adult. The ratios between the features and different face regions were computed. Farkas [12] presented a mathematical model to estimate the growth of the head from infancy to adulthood. Ramanthan and Chellappa [13] computed eight ratios of distance measures for modeling age progression. They proposed a craniofacial growth model by illustrating how the age-based anthropometric constraints on facial proportions translate into linear and non-linear constraints on facial growth parameters.

The active appearance model (AAM) was proposed by Cootes et al. [14]. This model is a statistical shape model to represent the face in the image by capturing the shape and the grey-level information. Lanitis et al. [15,16] studied the aging effects on face images and described the effects of aging on facial appearance. They built a statistical-based face model. By their proposed shape intensity face model and automatic age simulation, statistically significant improvement was reported in the performance of the age classification system. Luu et al. [17] proposed an AAM with 68 facial land marks for age estimation. A variation of the AAM model, called the Contourlet Appearance Model (CAM), was proposed in [18] to calculate the landmarks.

Geng et al. [19,20] defined an ageing pattern subspace (AGES) model for representing the ageing process as the sequence of an individual's face images sorted in time order by constructing a representative subspace. The images, which were not available for some age groups, are synthesized by using an expectation-maximization (EM) -like iterative learning algorithm. The AGES model builds an aging pattern for different age stages. On the other hand, the manifold model, as proposed by Fu et al. [21,22], was used to handle the ageing process in [23,24]. This model is more flexible than the AGES model, since images of different persons can be used for unavailable images of some ages. The manifold models use several linear regression functions to learn the low-dimensional aging trend from a group of face images for each age.

The appearance model (AM) is an alternative way to represent the age-related features. It focuses on texture, pattern analysis, and wrinkles. Hayashi et al. [25,26] extracted local and global facial features. Texture and shape features were calculated by using a semantic-level description of the face. Gunay and Nabiyev [27] used effective texture descriptor for appearance feature extraction and utilized local binary patterns (LBP). Gao and Ai [28] used the Gabor features with fuzzy-LDA. Their work showed that Gabor features are more effective than the LBP. Yan et al. [29,30] employed spatial flexible patches (SFP) as a feature descriptor in order to handle images with small undesirable defects such as occlusions and head pose. Mu et al. [31] proposed bio-inspired features (BIF) that have the ability to handle small rotations and scale changes effectively. Shan [32] exploited the LBP and Gabor features. Adaboost was used to learn the discriminative LBP-Histogram bins for age estimation. Lu et al. [33] proposed a cost sensitive local binary feature learning by using facial images. They extracted low-dimensional binary codes face patches from the raw pixels while using several hashing functions.

Recently CNNs and DNNs have been started to be used for feature extraction and classification for the facial age estimation due to their success in several computer vision fields. Levi and Hassner [5] used CNNs in age estimation for the first time. A simple CNN architecture was used as a feature extractor and a classifier to avoid overfitting problem. Ranjan et al. [34] proposed a cascaded classification and regression system based on a coarse age classifier. They introduced an age regressor for each age group based on the features extracted from the coarse age classifier. Then they used an error correcting method for correcting the regression errors for subjects. Chen et al. [35] proposed an age classification system. In their system, the feature set was extracted by using a pre-trained CNN for face identification task. The extracted features were fed to a small neural network to regress the age of the subject. Yang et al. [36] proposed a generic deep network model that extracted facial features by using a convolutional scattering network. The dimension of these features was reduced by PCA. They estimated the age using three fully connected layers that act as category-wise rankers. Yi et al. [37] extracted local aligned patches while using several facial landmarks. For each face image, 23 patch pairs were extracted in total. Each patch was trained in separate CNN and their final fully connected layer outputs were fused to estimate the age of a person. Liu et al. [38] proposed the AgeNet model to estimate the age apparent for the ChaLearn 2015 Apparent Age database. Two different CNN models were trained and fused to estimate the apparent age. Qawaqneh et al. [10] proposed a new model to jointly fine-tune two DNNs that were based on a new cost function. The two DNNs were trained on different feature sets, which were extracted from the same input data.

In this work, we propose a new cost function and a new model to find the most significant age-related features from the high-level features that are extracted by a CNN for the age estimation task from face images. In addition, the JFN-A is proposed to combine more multiple feature sets in order to enhance the estimation process.

2. Materials and Methods

Deep learning is being used in a wide range of applications and the $l2,1$-norm intersects with different deep learning applications. Sections 2.1 and 2.2 provide a brief overview of deep learning and the $l2,1$-norm. Section 2.3 introduces the proposed work. Section 2.4 explains the feature sets used in this work. Section 2.5 summarizes the jointly fine tuning of different DNNs. Section 2.6 elaborates on the specifications of the used benchmark.

2.1. Deep Learning

DNN is considered as the second generation of the artificial neural network (ANN). ANN is composed of one input layer, one hidden layer, and one output layer. It showed limited success in feature extraction and classification tasks. This simple architecture is constrained by the low computational power and the limited learning algorithms [39]. On the last decade, the advances in hardware and learning algorithms made it possible to design deep learning architectures. DNN, which was introduced by [40], consists of one input layer, multiple hidden layers, and one output layer. DNN is capable of extracting efficient features from vast datasets for different classification tasks [41]. DNNs have shown significant improvement in many computer vision fields, such as face recognition [42,43], image classification [44–46], object detection [47], and semantic segmentation [48]. Nowadays, CNN is considered as one of the most successful deep architectures for feature extraction from images. CNNs are composed of a number of convolutional layers and a number of fully connected layers. The convolutional layers use filters to extract distinctive features, while the fully connected layers are used for classification.

2.2. l2,1-Norm

Recently, the space and the size of the available data for different machine learning fields are vast. One of the key stages of learning models from such vast data sets is to select features that permit the machine to learn the embedded models. The main goal is to select the most significant features from the available features for a better machine learning performance. It reduces the dimension of the feature space, allows for the machine to learn faster, and it obtains a generalizable model. Several number of selection methods have been developed, but recently the *l*2,1-norm has been reported as one of the most efficient techniques to develop models that can select features across all data points with joint sparsity [49–52]. The *l*2,1-norm has been used in different machine learning fields and was also used for tensor factorization. The matrix *l*2,1-norm was proposed by [53] as an invariant of the *l*2-norm. The *l*2,1-norm can be defined, as in (1).

$$||\mathrm{M}||_{2,1} = \sum_{i=1}^{n} \sqrt{\sum_{j=1}^{m} x_{ij}^{2}} = \sum_{i=1}^{n} \left|\left|x^{i}\right|\right|_{2} \tag{1}$$

M is a matrix and x_{ij} is the element in the *i*th row and *j*th column. In this paper, we combined the *l*2,1-norm with a DNN in one model for selecting the age-related features.

2.3. Proposed Model: DNNs and l2,1-Norm Regularization Framework for Robust Features Selection

A regularization framework is proposed for selecting features, by combining the DNN and *l*2,1-norm with a new cost function. For a given data, $[X = x_1, x_{2, \ldots,} x_m] \in \mathrm{R}^{\mathrm{fxm}}$, let $[Y = y_1, y_{2, \ldots,} y_m]$ $\in \mathrm{R}^{\mathrm{mxc}}$, where *m* is the number of training samples, *f* is the number of feature dimension, and *c* is the number of classes. The goal is to learn a projection matrix $W \in \mathrm{R}^{\mathrm{fxc}}$. W is used to select the most robust features from the common space, which is defined by the class labels. The new cost function uses the mean squared error (MSE), sigmoid as an output layer function, and *l*2,1-norm as a regularizer. The minimization cost function is given, as in Equation (2).

$$E(y,x) = \left\{ \frac{1}{2m} \left(sigm\left(X^{T} W \right) - Y \right)^{2} + \lambda ||W||_{2,1} \right\} \tag{2}$$

where λ is a controlling parameter, $||.||_{2,1}$ is the *l*2,1-norm, and $sigm(z) = \left(\frac{1}{(1+e^{-z})} \right)$.

2.3.1. The Architecture

Figure 1 shows the new model for robust feature selection based the proposed cost function. The model consists of two hidden layers with f nodes in each layer. The f equals to the number of features in the input layer.

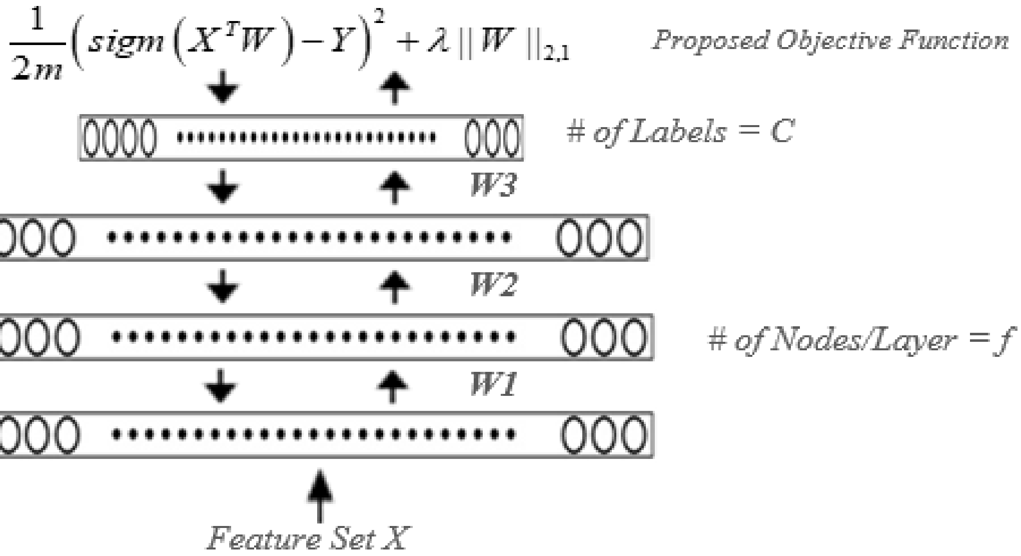
$\frac{1}{2m}\left(sigm\left(X^TW\right)-Y\right)^2+\lambda\,||W||_{2,1}$
Proposed Objective Function
# of Labels = C
W3
W2
# of Nodes/Layer = f
W1
Feature Set X

Figure 1. Model architecture for finding the projection matrix *W*. The weights between the layers (*W*1, *W*2, and *W*3) are modified based on the error propagated from the new cost function.

The number of nodes in the output layer is set to be the number of labels, C. The architecture is chosen to be compatible with the dimension of the projection matrix (*W*), whose dimension ought to be compatible with the number of features in the training samples and the number of labels. Therefore, the number of nodes in the last hidden layer and the number of labels is set to be the number of features in the training samples (f) and the number of the classes in the training samples (C), respectively, since W3 is taken as the projection matrix (*W*). Note that the process of extracting the features by using the output weights W3 depends on the weights *W*1 and *W*2. *W*1 and *W*2 weights must be calculated first in order to calculate the weights in the output layer *W*3. In DNNs that, as we move from the lower layers to the higher layers, the resulted features by using their corresponding weights become more abstract and high-level features. In addition, we use the output layer weights (*W*3) as the projection matrix to match the size of the features in the training set.

The optimization of the number of hidden layers and the nodes in each hidden layer is practically determined by experiments. The number of row in *W* equals to the number of columns in the training data and the number of columns in *W* equals to the number of class labels. As shown in Figure 2, the new robust features for each training sample can be calculated by multiplying each training sample with the projection matrix *W*.

Enhancing the capability of the DNN as a feature extractor by merging the *l*2,1-norm regularization into a new cost function will result in more robust features. The new cost function calculates the resulted error for each batch of the training data and it calibrates the weights between the layers more accurately.

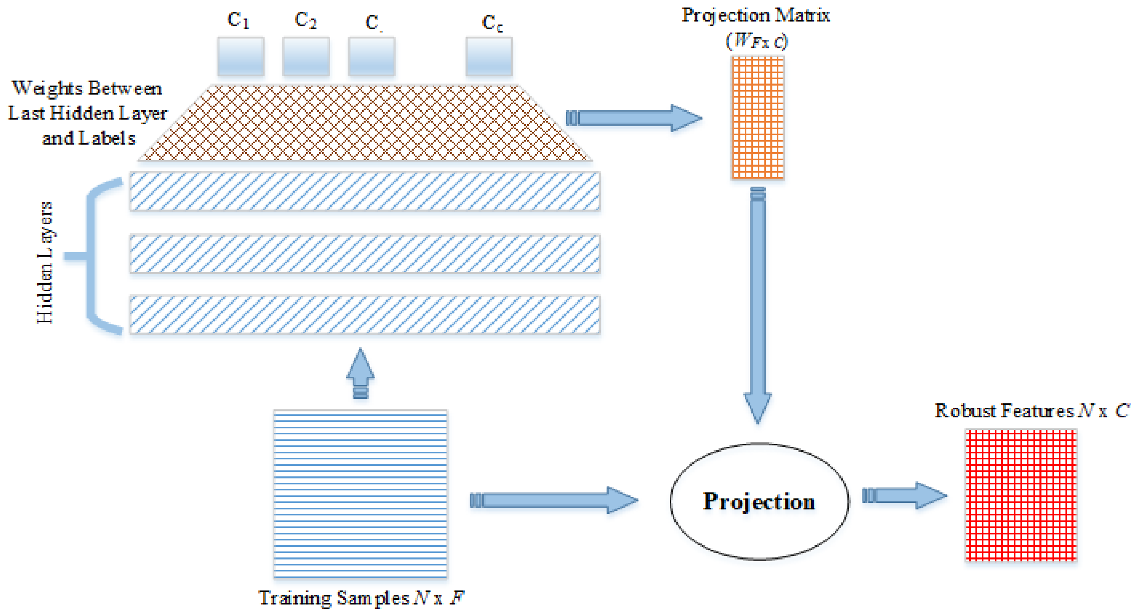

Figure 2. Robust feature selection process. The dimension of the resulted robust features for the training samples is mxc.

2.3.2. Learning

In the learning process, it is aimed to minimize the proposed cost function with respect to the projection matrix W, as in Equation (3).

$$min_{W}^{E(y,x)} = \left\{ \left(sigm\left(X^{T}W \right) - Y \right)^{2} + \lambda ||W||_{2,1} \right\} \tag{3}$$

The cost function in Equation (2) is not easy to minimize with the presence of the $l2,1$-norm. In [52,54,55], the half-quadratic minimization method was used to solve the minimization of the $l2,1$-norm. One should know that the minimizer function of the $l2,1$-norm is unpredictable near the origin. Therefore, according to $l2,1$-norm analysis in [55], a $\varnothing(x) = \sqrt{\epsilon + x^2}$ can be defined to solve this problem, where ϵ is chosen to be a decreased value to ensure that the function in Equation (3) with the $l2,1$-norm is converged. Additionally, $\varnothing$ should satisfy all of the conditions in Equation (4).

$$\begin{aligned} & x \rightarrow \varnothing(x) \ is\ convex\ on\ R, \\ & x \rightarrow \varnothing\left(\sqrt{x}\right) \ is\ concave\ on\ R_{+}, \\ & \varnothing(x) = \varnothing(-x),\ \forall x \in R, \\ & \varnothing(x) \ is\ \mathrm{C}^{1}\ on\ R, \\ & \varnothing''\left(0^{+}\right) > 0,\ \underset{x\rightarrow\infty}{lim} \varnothing(x)/x^{2} = 0. \end{aligned} \tag{4}$$

Lemma 1. *Let be a function satisfying all conditions in Equation (4). There exists a conjugate function,$\varphi(.)$, as in Equation (5) such that*

$$\varnothing(\left|\left|w^{i}\right|\right|_{2} = \underset{p\in R}{inf} \left\{ p \left|\left|w^{i}\right|\right|_{2}^{2} + \varphi(p) \right\} \tag{5}$$

p is determined by the minimizer function $\delta(.)$ with respect to $\varnothing(.)$. Based on $\varnothing(x)$, $\lambda||W||_{2,1}$ is replaced with $\lambda \sum_{i}^{f} \sqrt{\epsilon + ||w^{i}||_{2}^{2}}$, then the Equation (3) is reformulated, as in Equation (6).

$$\underset{W}{\min E(y,x)} = \left\{ \frac{1}{2m} \left(sigm\left(X^{T}W \right) - Y \right)^{2} + \lambda \sum_{i}^{f} \sqrt{\epsilon + ||w^{i}||_{2}^{2}} \right\} \tag{6}$$

According to Lemma 1, the function of $\lambda \sum_i^f \sqrt{\epsilon + ||w^i||_2^2}$ can be reformulated, as in Equation (7).

$$\lambda Tr\left(W^T QW\right) \tag{7}$$

$q = \delta\left(||w^i||_2\right) \in \mathrm{R}^{\mathrm{f}}$ is an auxiliary vector and $Q = (q)$. The operator (.) puts a vector q on the main diagonal of Q. q is computed by using the optimizer function, as in Equation (8).

$$q_i = \frac{1}{\sqrt{||w^i||_2^2 + \epsilon}} \tag{8}$$

According to Equation (7), the minimization function can be written, as in Equation (9).

$$\underset{W}{\min E(y,x)} = \frac{d}{dW}\left\{\frac{1}{2m}\left(sigm\left(X^T W\right) - Y\right)^2 + \lambda Tr\left(W^T QW\right)\right\} \tag{9}$$

The analytic minimization solution of Equation (9) with respect to W is finally written by Equation (10).

$$\underset{W}{\min E(y,x)} = \left(sigm\left(X^T W\right) - Y\right)\left(sigm\left(X^T W\right)\left(1 - \frac{1}{sigm(X^T W)}\right)\right) + \lambda QW \tag{10}$$

2.4. Feature Sets

In this work, two feature sets are used to evaluate the efficiency of the proposed model. Both feature sets are extracted from the training set in the Adience benchmark. The first feature set is extracted from a pre-trained model for face recognition, namely the VGG-Face [43]. This model was trained to extract distinctive facial features that can be valuable in age estimation from facial images. The details of extracting the age-related facial features from the training samples by using the VGG-Face model are available in [10]. The second feature set represents the depth of the superpixels in a facial image and the relations between the superpixels and their neighbors. In our previous work [10], we showed the significance of the superpixels depth features in age estimation. The superpixels depth and their relations were extracted by using a pre-trained model in depth estimation [56]. More details on the extraction of the superpixels features are available in [10].

2.5. Jointly Fine-Tuning of Two DNNs

Two methods are used for jointly fine-tuning of two DNNs with two feature sets to show the efficiency of the proposed model.

2.5.1. By Amplified Features (JFN-A)

In this section, we propose a jointly fine-tuned model that is based on the amplification of the features that are resulted from the element-wise summation of the last hidden layers of the two DNNs with two different feature sets. The input features of the first DNN are the facial features which were extracted from pre-trained model on face recognition, while the input features of the second DNN are the superpixels and their relations, which were extracted from a pre-trained model for depth estimation.

As shown in Figure 3, the proposed JFN-A consists of three parts. The first and second parts are two supervised DNNs, and the third part is a supervised neural network that jointly fine-tunes the first and second parts. The first and second parts have two hidden layers and one output layer. The third part has one input layer, one hidden layer, and one output layer. There are eight labels in the output layer. The input of the third part is the element-wise summation of the last hidden layers of the first and second parts.

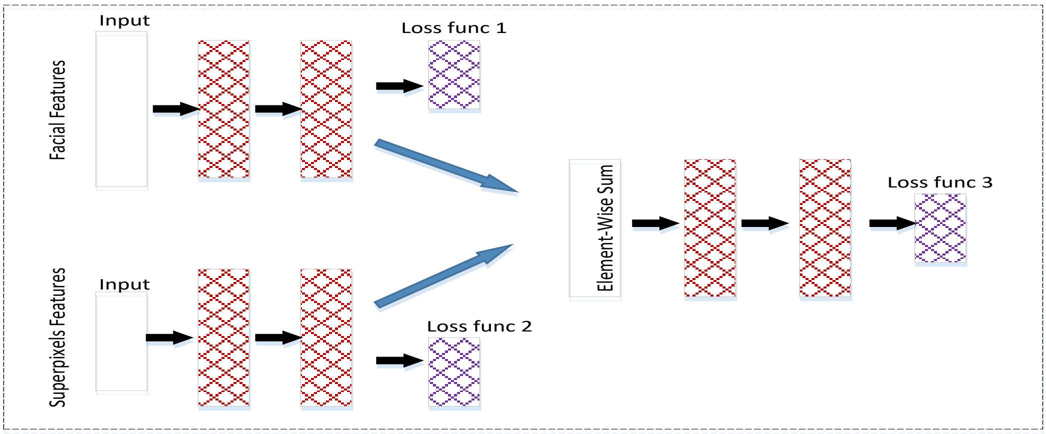

Figure 3. JFN-A: Jointly fine-tuning of two deep neural networks (DNNs) by element-wise summation of the outputs of the last hidden layers of the first and second parts.

The learning process of the JFN-A is explained as follows:

Step 1: Three loss functions are used to train the three networks. All three loss functions are the softmax cross entropy function, as in (11).

$$L_i = -\sum_{j=1}^{c} y_j \log\left(\overline{y}_{i,j}\right) \tag{11}$$

L_i is the loss function of the network i, y_j is the jth value of the label, and $\overline{y}_{i,j}$ is the jth output value of the network i.

Step 2: The first and second parts of the model are trained by using first batch of their corresponding features. Then, the outputs of the last hidden layer of both networks are element-wise summed to form the input of the third part, as in Equation (12).

$$x_{3,j} = \ Relu\left(l_{1,j}\right) + \ Relu\left(l_{2,j}\right) \tag{12}$$

$x_{3,j}$ is the input of the third part, $l_{1,j}$ and $l_{2,j}$ are the outputs of the last hidden layer of the first and second parts, and *Relu* is the rectified activation function.

Step 3: Feedforwarding, error calculation, and backpropagation are performed on the third part.

Step 4: The steps from 1 to 3 are repeated for the rest of the training batches.

Step 5: After the training is complete, the softmax output of the third part, $\overline{s}$, is obtained as the final decision, as in Equation (13).

$$\overline{s} = \mathrm{arg}max_j\, \overline{y}_{3,j} \tag{13}$$

2.5.2. By the Softmax and the Sigmoid (JFN-SS)

This method is based on our previous work in [10]. It consists of two DNNs that use different feature sets. These feature sets are extracted from the same input images. The first network is trained on the first feature set. The cross-entropy is used as the loss function. The Softmax function is used at the output layer. The second network is trained on the second feature set. The sigmoid function is used to calculate the output layer probabilities, the mean squared error loss function is used to calculate the DNN2 output error. Both of the networks are fine-tuned by the derived cost function that uses the generated errors in both DNNs and later calibrates the weights in the first DNN, accordingly. This method is explained in detail in [10].

2.6. Database

The Adience benchmark is used in this work. It contains 26K face images of 2284 subjects who are divided into eight age groups that are called labels. More details about the Adience benchmark are available in [57]. Standard five-fold, subject-exclusive cross-validation protocol is applied for dividing the database into train and test sets. The Adience is a challenging database, since it consists of unfiltered face images, which were uploaded to the Flicker website while using smart phones. The images are not filtered with any manual filtering techniques. Images in the database reflect real-world conditions of uncontrolled environments such as significant variations in pose, expression, lighting, image quality, and resolution. The Adience is not designed for face recognition task so that the number of images per subject is not balanced. Around 80 percent of the subjects in the database have only one image, while the rest have around 100 to 400 images. When the number of images per subject is small for a label, while it is bigger for other labels, the classifier will be biased for the labels with more images.

3. Results

3.1. Robust Feature Selection Method

In this work the Adience database is used to test and evaluate the proposed work. For each image sample, the facial and the superpixels feature sets are extracted. Then, the proposed robust features are extracted for each feature set. For each image sample, the facial, superpixels, and their derived robust features are concatenated and two DNNs and one NN are trained for age classification task. The first DNN is used for finding the projection matrix and the second DNN is used for training the original features concatenated with their robust features. Both DNNs consist of two hidden layers and one input layer of 4096 and 512 nodes. The size of the input features for the first DNN is 4096 and the size of the input features for the second DNN is 4104. Dropout rate of 0.7, learning rate of 0.01, and weight decay of 10^{-4} are used in both networks. Since the resulted robust feature sets are relatively small in size, a neural network with one hidden layer (NN) is used to report the accuracy results as the robust feature set is used as input features (Lines 3 and 6 in Table 1). The input for this network consists of eight features, which is the robust feature set size for each image sample in the database. The hidden layer consists of 50 nodes. The dropout rate is chosen to be 0.5 and the learning rate is set to 0.01. Table 1 shows the overall accuracies using all feature sets, the facial, superpixels, derived robust for facial, derived robust for superpixels, facial concatenated with its robust, and superpixels concatenated with its robust features. As it can be seen in Table 1, the robust features achieved classification accuracies that are comparable to the performance of the original features (facial and superpixels feature sets). As an advantage the dimension of the robust features is much smaller than the dimension of the original features. These obtained results by using the robust features verify the effectiveness of the proposed model in finding efficient and distinctive features (robust features) for the age classification from wild facial images. Moreover, using NN with the robust feature sets instead of using DNN at the classification stage reduces the computational time. DNN is used to find the accuracy results for the other feature sets (lines 2, 4, 5, 7 in Table 1). DNN consists of several hidden layers, while NN consists of one hidden layer. As well as, the number of nodes in each DNN layer much bigger than the number of nodes in the NN single hidden layer. This means that the number of connection between the nodes in the DNN is much higher than those in the NN, and this increases computation complexity in the DNN.

Table 1. Overall classification accuracies for facial robust features on Adience database (%). Appendix A presents the confusion matrices for the robust facial and robust superpixels features.

Age Groups / Feature Sets	0–2	4–6	8–13	15–20	25–32	38–43	48–53	60–	Accuracy	1-Off Acc
Facial	88.41	60.18	39.12	43.61	67.14	43.79	14.52	57.20	57.45	94.32
Robust-Facial	82.19	69.12	39.71	12.78	76.42	19.53	6.64	38.52	53.25	81.18
Facial + Robust-Facial	86.96	65.96	45.88	35.24	78.98	41.22	15.35	83.66	63.22	94.38
Superpixels Features	86.34	58.82	34.18	15.01	80.78	11.05	10.88	53.31	53.62	81.40
Robust-Superpixels	78.14	52.37	36.46	20.28	67.61	17.77	17.38	57.15	49.95	78.89
Superpixels + Robust-Superpixels	87.56	56.82	43.57	31.05	78.31	24.81	29.70	60.96	58.31	86.17

The concatenation of the original features and their robust features improve the classification accuracies by a wide margin. Based on the results in Table 1, it can be observed that features based on the facial feature set performed better than the superpixels feature sets. There are two main reasons. Firstly, the facial feature set was extracted from a pre-trained model on VGG-Face for face identification and classification problem; whereas the superpixels feature set was extracted from a model that was pre-trained for depth estimation problem. Facial and robust facial feature sets contain more age-related information and more relevant to age estimation task. Secondly, the database (about two million unconstrained face images) used in pre-training the VGG-Face model is much larger than the database (about tens of thousands of images that not all are face images) used in pre-training the depth estimation model.

The concatenation of both feature sets with their robust features improves the classification accuracy significantly. A classification process depends on two main factors: feature extraction and classification process. The facial and superpixels features have shown significant improvements in age classification. Moreover, the newly derived robust features with the same settings and classifier enhanced the model's performance by reducing the computation time.

The proposed framework combines the efficiency of DNNs for extracting distinctive feature and the powerfulness of the *l*2,1-norm for selecting robust features. The *l*2,1-norm is well known for its ability to deal with the outliers in images. Since the unconstrained image database contains various outlier images, the *l*2,1-norm based sigmoid cost function enables our model to focus on finding robust age-related features and reducing the negative effects of the outliers. The contribution of *l*2,1-norm for selecting robust features can be controlled by tuning the λ parameter. Figure 4 shows the impact of choosing different values for λ on the classification accuracy.

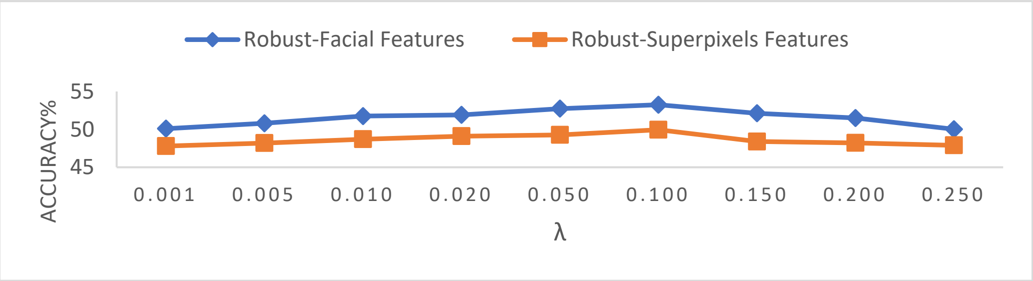

Figure 4. The effect of λ on the accuracy results achieved by the robust features. The best accuracy result is achieved with λ = 0.1 for both feature sets, while the classification accuracy drops when the λ value drifts away from 0.1 from both sides.

3.2. Jointly Fine-Tuning Robust Feature Sets

The extracted robust facial and superpixels feature sets are used as input for JFN-A and JFN-SS. The performance of the jointly fine-tuning networks based on the amplified feature sets is evaluated by using the proposed robust feature sets. The network is depicted in Figure 5.

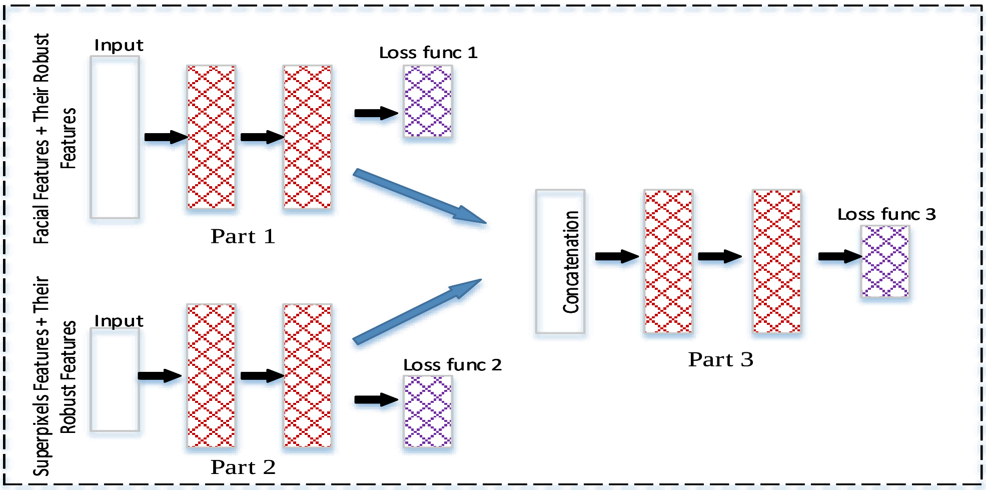
Facial Features + Their Robust Features
Input
Loss func 1
Part 1
Superpixels Features + Their Robust Features
Input
Loss func 2
Part 2
Concatenation
Loss func 3
Part 3

Figure 5. JFN-A is fed with facial and superpixels feature sets concatenated with their robust features. The facial features that are concatenated with their corresponding robust features are fed to the first part of the network, while the superpixels features and their corresponding robust features are fed to the second part of the network.

The network settings for Part 1 and Part 2 of the JFN-A are chosen as two hidden layers of size 1024 and input layer size is 4096 and 512, respectively. The learning rate, dropout rate, and weight decay are 0.1, 0.7, and 10^{-3}, respectively, for both parts. Both parts are jointly fine-tuned by using Part 3. Two more hidden layers of size 512 are added on the top of the last hidden layers of Part 1 and Part 2. Then, one output layer with eight age labels is added. The learning rate for Part 3 is set to 0.01, with dropout rate of 0.8, and weight decay of 10^{-4}. The training is stopped when there is no improvement in the validation set results. The performance of the second joint fine-tuning method [10] and the proposed robust feature sets are evaluated. The two feature sets that are concatenated with their robust features are trained and tested, as shown in Figure 6.

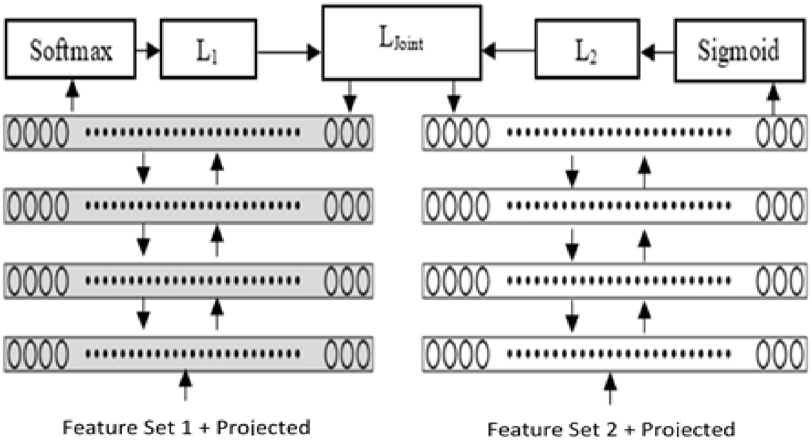

Figure 6. The proposed cost function with facial and superpixels and their robust features as input.

The JFN-A and JFN-SS with the two feature sets concatenated with their robust features outperform the accuracies that are observed by using each feature set alone. The results are given in Table 2. The performance of the networks that are using the facial features concatenated with their robust features is better than that of the superpixels features and their robust features for some classes for the same reasons that are explained in the previous section.

Table 2. Overall classification accuracies for facial and superpixels concatenated with their robust features by JFN-A and JFN-SS on Adience database (%).

Age Groups **Features**	**0–2**	**4–6**	**8–13**	**15–20**	**25–32**	**38–43**	**48–53**	**60–**	**Accuracy**	**1-Off Acc**
(1) Superpixels Features concatenated with their robust features	87.56	56.82	43.57	31.05	78.31	24.81	29.70	60.96	58.31	86.17
(2) Facial Features concatenated with their robust features	86.96	65.96	45.88	35.24	78.98	41.22	15.35	83.66	63.22	94.38
(1) + (2) as input for JFN-A	88.5	63.44	49.84	37.97	82.55	43.29	27.92	81.61	65.55	94.86
(1) + (2) as input for JFN-SS	88.72	68.58	48.23	34.86	82.56	35.89	28.46	83.67	65.20	91.39

The proposed jointly fine-tuned networks by using the proposed robust features enhanced the overall accuracy in the age estimation. Moreover, it is noticed that both JFN-A and JFN-SS achieve significant results. One-off accuracy of the JFN-A is about 3% higher than that of the JFN-SS. This might be due to the nature of the sigmoid function that is used in the JFN-SS. The sigmoid cost function is mostly used in binary classification and focuses on the target label. It does not care about how close the other labels to the target label. The JFN-A uses the softmax cost function only. The softmax cost function evaluates the probabilities of all the labels, so that it increases the one-off accuracy of the model. These models take the advantage of using two different feature sets and also take the advantage of training the data set on two different DNNs. The JFN-A model trains and minimizes the error by using three different cross-entropy functions as cost functions.

The effects of hidden nodes per layer and the number of hidden layers on the classification accuracy of the JFN-A are shown in Figure 7.

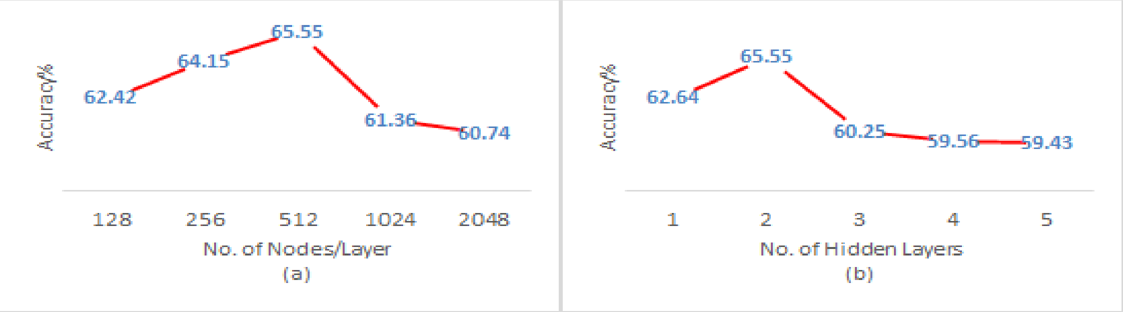

Figure 7. The accuracy of the JFN-A with respect to (**a**) the number of hidden nodes and (**b**) the number of hidden layers.

Figure 7a presents the accuracies for different number of hidden nodes for part 3 of the JFN-A. The highest accuracy is observed when the number of nodes was 512. Moreover, Figure 7b shows how the accuracy significantly decreases when the number of hidden layers exceeds 2. Since we have two level of classification, the inputs of the Part 3 in the JFN-A network are high level features, which do not need deep architecture to enhance these features further, while the purpose is to jointly fine-tune the features together. The top row in Figure 8 shows a set of images that are incorrectly classified, and the bottom row shows a set of images that are classified correctly by the proposed networks.

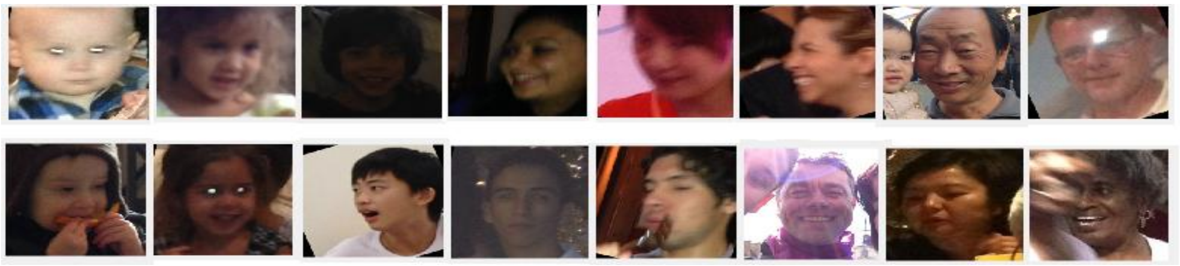

Figure 8. Challenging images in the Adience database. Images in the top row were classified incorrectly by the proposed networks. Images in the bottom row were classified correctly by the proposed work.

4. Discussion

The state-of-the-art on the Adience database is listed in Table 3. The facial image alignment was studied to improve the classification accuracies along with the effects of dropout by using the SVM based classifier [57]. The work in [5] was the first step in age and gender classification from face images by using DNNs. It employed a relatively simple and shallow network for feature extraction and classification stages and it proposed an over-sampling technique to solve the misalignment challenge partially. Hebda and Kryjak proposed a compact DCNN architecture for age and gender estimation [58]. The aim of their work is to train a fast DCNN as a base for the DNN in video training with the input image size of 32 × 32. Zhu et al. studied a DCNN for a multitask learning to train the shared features for age and gender tasks in end-to-end manner [59]. The works in [5,58,59] built simple CNNs without utilizing any pre-trained models. Based on their reported results, one can notice that the methods are not efficient for extracting age-related features, since the models were trained by using the Adience database, which is considered to be a relatively small database.

A cascaded CNN is introduced in age estimation in [35]. It consists of three modules: age group classifier, DCNN bases regressors, and erroneous age prediction. Rothe et al. utilized a deep CNN architecture that is based on the VGG-16 model that was pre-trained for image classification. It does not rely on facial landmarks to extract facial age-related features. The pretrained DNN was used as a facial feature extractor [60]. The models are mainly built and trained for apparent age estimation. Since apparent age estimation and real age estimation are different tasks, the reported improvements are limited when compared with other works [35,60].

A new face descriptor model is presented in [61] that is based on three attributes: (1) The age primitives, which finds the crucial texture primitives; (2) The latent second direction to keep the structural information; and, (3) The global adaptive threshold to discriminate in the flat and textured region. The new descriptor was used to extract facial features for age classification. In work [62], the label-sensitive deep metric learning (LSDML) is proposed to learn a discriminative feature similarity in facial age estimation. The goal was to exploit the label correlation between the training face samples in term of the labels (subspace) to achieve balanced training samples. In [63], a new hybrid architecture is developed based on CNN and extreme learning machine (ELM). Both techniques combined together to deal with age and gender classification. The CNN is used for feature extraction, while the ELM is used as a classifier. Different techniques for facial age estimation are proposed in [60–63].

A new DCNN model that was based on an attention network is introduced in [64]. This model estimates the most informative patches in low-resolution images, which are further processed in a patch network in higher resolution.

Table 3. Comparison of state-of-the-art results (%).

	Method	Exact Accuracy	1-Off Accuracy
[5]	Shallow CNNs Using Single Crop	49.5	84.6
	Shallow CNNs Using Over-Sample	50.7	84.7
[10]	DNN1 with Facial Features	57.45	94.32
	DNN2 with Superpixels Features	53.62	81.40
	Facial and Superpixels as input for JFN-SS	62.37	94.46
[35]	Cascaded Convolutional Neural Network	52.88	88.45
[57]	LBP	41.4	78.2
	LBP + FPLBP	44.5	80.7
	LBP + FPLBP + Dropout 0.5	44.5	80.6
	LBP + FPLBP + Dropout 0.8	45.1	79.5
[58]	Compact DCNN architecture as base for video learning	42.0	-
[59]	Employ light weight DCNN for a multitask learning scheme (age + gender), best model single-6-conv	49.7	-
[60]	DCNNs based on VGG-16 architecture + softmax expected function for refinement	55.6	89.7
[61]	Subject-Exclusive DAPP	54.9	-
	Subject-Inclusive DAPP	62.2	-
[62]	LSDML: w/o data augmentation	56	-
	LSDML: random cropping + horizontal flipping	56.9	-
	M-LSDML: w/o data augmentation + 3 DB	58.2	-
	M- LSDML: random cropping + horizontal flipping + 3 DB	60.2	-
[63]	Proposed CNN-ELM + Dropout 0.5	51.4	-
	Proposed CNN-ELM + Dropout 0.7	52.3	-
[64]	VGG-16-Faces + Attention Network	61.8	95.1
This work	Facial and Superpixels as input for JFN-A	63.78	93.70
	(1) Superpixels Features Concatenated with Their Robust Features	58.31	86.17
	(2) Facial Features Concatenated with Their Robust Features	63.22	94.38
	(1) + (2) as input for JFN-A	65.55	94.86
	(1) + (2) as input for JFN-SS	65.20	91.39

In Table 3, the proposed work outperforms the state-of-the-art methods because the proposed method extracts more effective age-related feature sets and uses more capable classifier architectures, while all of the previous works focused on only one of these tasks. Previous works, which did not use DNNs or used a shallow DNN that was trained on one relatively small database, did not achieve satisfactory results [5,35,57,59]. Previous works that used relatively deep NNs trained on several age databases with different augmentations to increase the training samples achieved slightly improved results, such as the work in [60], subject-exclusive DAPP in [61], LSDML and M-LSDML in [62], and [63]. The highest accuracy reported on the Adience database is 62.2%. It was found by using the subject-inclusive protocol [61]. Same subjects are allowed to appear as test and train samples in subject-inclusive protocol. Similar classification accuracy is reported in [64] as 61.8%. Although the reported accuracy rate is relatively competitive, the work relied on several very deep CNNs (Res-50 DNNs) to extract the feature set and it resulted in high computation time. Our proposed work achieved the highest accuracies in the literature of age and gender classification from unconstrained facial images by following the subject-exclusive protocol. Among our proposed methods, the JFN-A and JFN-SS achieved the highest exact accuracies of 65.55% and 65.20%, respectively. In general, the following points should be considered in order to perform an efficient and successful age estimation:

- The availability of large databases is essential to perform efficient age classification from facial images. It is especially important when the model utilizes deep and very deep NNs. Otherwise, the model is liable to overfitting problem. Database should contain sufficient examples of reflecting real environment challenges, such as pose, illumination, resolution, and other real conditions. In the case of relatively small databases, transferable learning could be used to compensate the limited facial images. For example, pre-trained models, which were trained over large databases for different yet related tasks, could be used and customized.

- Extracting distinctive feature set(s) is also a very important step. Distinctive features allow for the classifier to differentiate between different age groups efficiently. Finding such feature sets is not an easy task and it needs a thorough investigation and study.
- Choosing a classifier to utilize and customize for the age classification task from facial images plays a major role in reaching a satisfactory and competitive classification accuracies.

5. Conclusions

In this work, a new method for enhancing the ability of the DNN as a feature selector is proposed. The new method embeds the *l*2,1-norm into the cost function of the DNN forming a new cost function. The DNN minimizes the new cost function to find the robust features from the input feature set. Several experiments are carried out to evaluate the performance of the proposed cost function. The new method is tested over a publicly available bench mark of facial images for age estimation (Adience). The new method is applied on two feature sets: the facial features and the superpixels features, both of the features were extracted while using pre-trained models. In addition, the robust features extracted using the proposed method are tested over two joint fine-tuning techniques, JFN-A and JFN-SS. The performance of the proposed method outperforms the state-of-the-art results by almost 3% in terms of the overall accuracy. Moreover, utilizing the resulted robust features by using the proposed work improved the performance of the joint fine-tuning methods.

Author Contributions: Conceptualization, A.A.M., Z.Q. and B.D.B.; Formal analysis, A.A.M., Z.Q. and B.D.B.; Investigation, A.A.M. and Z.Q.; Methodology, A.A.M. and Z.Q.; Resources, A.A.M. and Z.Q.; Software, A.A.M. and Z.Q.; Supervision, B.D.B.; Validation, A.A.M., Z.Q. and B.D.B.; Visualization, A.A.M., Z.Q. and B.D.B.; Writing—original draft, A.A.M. and Z.Q.; Writing—review & editing, A.A.M., Z.Q. and B.D.B.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. Confusion matrix for the robust facial features (%).

Predict Actual	0–2	4–6	8–13	15–20	25–32	38–43	48–53	60–
0–2	82.19	15.73	0.21	0.00	1.86	0.00	0.00	0.00
4–6	13.51	69.12	12.98	1.23	2.81	0.18	0.00	0.18
8–13	0.59	5.88	39.71	5.59	44.12	3.82	0.00	0.29
15–20	0.44	0.88	15.42	12.78	65.64	3.96	0.00	0.88
25–32	0.19	0.38	5.49	1.52	76.42	11.65	0.76	3.60
38–43	0.59	0.39	1.97	0.59	62.73	19.53	3.55	10.65
48–53	0.83	0.41	3.73	2.49	52.70	23.24	6.64	9.96
60–	0.00	0.39	1.17	0.00	29.97	27.24	2.72	38.52

Table A2. Confusion matrix for the robust superpixels features (%).

Predict Actual	0–2	4–6	8–13	15–20	25–32	38–43	48–53	60–
0–2	78.14	17.35	2.24	0.32	1.41	0	0	0.54
4–6	28.27	52.37	9.42	4.64	2.92	1.16	0.64	0.58
8–13	2.31	7.67	36.46	8.54	38.12	5.03	0	1.87
15–20	0	4.99	14.08	20.28	53.24	3.92	1.85	1.64
25–32	1.88	2.43	6.04	7.93	67.61	9.51	1.23	3.37
38–43	1.63	2.66	3.49	6.7	55.44	17.77	5.08	7.23
48–53	0	3.96	5.65	3.72	44.29	11.89	17.38	13.11
60–	0.07	2.07	3.69	7.81	20.83	5.46	2.92	57.15

References

1. Fu, Y.; Guo, G.; Huang, T.S. Age synthesis and estimation via faces: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *32*, 1955–1976. [PubMed]
2. Barer, B.M. Men and women aging differently. *Int. J. Aging Hum. Dev.* **1994**, *38*, 29–40. [CrossRef] [PubMed]
3. Sveikata, K.; Balciuniene, I.; Tutkuviene, J. Factors influencing face aging. Literature review. *Stomatologija Baltic Dent. Maxillofacial J.* **2011**, *13*, 113–115.
4. Ramanathan, N.; Chellappa, R. Face verification across age progression. *IEEE Trans. Image Process.* **2006**, *15*, 3349–3361. [CrossRef] [PubMed]
5. Levi, G.; Hassner, T. Age and gender classification using convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Boston, MA, USA, 7–12 June 2015; pp. 34–42.
6. Hong, L.; Jain, A.K.; Pankanti, S. Can multibiometrics improve performance? In Proceedings of the AutoID '99, Summit, NJ, USA, 28–29 October 1999; pp. 59–64.
7. Jain, A.K.; Dass, S.C.; Nandakumar, K. Soft biometric traits for personal recognition systems. In *Biometric Authentication*; Springer: Berlin/Heidelberg, Germany, 2004; pp. 731–738.
8. Brin, S.; Page, L. Reprint of: The anatomy of a large-scale hypertextual web search engine. *Comput. Netw.* **2012**, *56*, 3825–3833. [CrossRef]
9. Linden, G.; Smith, B.; York, J. Amazon. com recommendations: Item-to-item collaborative filtering. *IEEE Internet Comput.* **2003**, *7*, 76–80. [CrossRef]
10. Qawaqneh, Z.; Mallouh, A.A.; Barkana, B.D. Age and gender classification from speech and face images by jointly fine-tuned deep neural networks. *Expert Syst. Appl.* **2017**, *85*, 76–86. [CrossRef]
11. Kwon, Y.H.; Lobo, N.D. Age classification from facial images. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 21–23 June 1994; pp. 762–767.
12. Farkas, L.G. *Anthropometry of the Head and Face*; Raven Press: New York, NY, USA, 1994.
13. Ramanathan, N.; Chellappa, R. Modeling age progression in young faces. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), New York, NY, USA, 17–22 June 2006; pp. 387–394.
14. Cootes, T.F.; Edwards, G.J.; Taylor, C.J. Active appearance models. *IEEE Trans. Pattern Anal. Mach. Intell.* **2001**, *23*, 681–685. [CrossRef]
15. Lanitis, A.; Taylor, C.J.; Cootes, T.F. Modeling the process of ageing in face images. In Proceedings of the Seventh IEEE International Conference on Computer Vision, Kerkyra, Greece, 20–27 September 1999; pp. 131–136.
16. Lanitis, A.; Taylor, C.J.; Cootes, T.F. Toward automatic simulation of aging effects on face images. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 442–455. [CrossRef]
17. Luu, K.; Ricanek, K.; Bui, T.D.; Suen, C.Y. Age estimation using active appearance models and support vector machine regression. In Proceedings of the IEEE 3rd International Conference on Biometrics: Theory, Applications, and Systems, Los Angeles, CA, USA, 22–25 October 2009; pp. 1–5.
18. Luu, K.; Seshadri, K.; Savvides, M.; Bui, T.D.; Suen, C.Y. Contourlet appearance model for facial age estimation. In Proceedings of the International Joint Conference on Biometrics, Washington, DC, USA, 11–13 October 2011; pp. 1–8.
19. Geng, X.; Zhou, Z.-H.; Zhang, Y.; Li, G.; Dai, H. Learning from facial aging patterns for automatic age estimation. In Proceedings of the 14th ACM International Conference on Multimedia, Santa Barbara, CA, USA, 23–27 October 2006; pp. 307–316.
20. Geng, X.; Zhou, Z.-H.; Smith-Miles, K. Automatic age estimation based on facial aging patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 2234–2240. [CrossRef] [PubMed]
21. Fu, Y.; Xu, Y.; Huang, T.S. Estimating human age by manifold analysis of face pictures and regression on aging features. In Proceedings of the IEEE International Conference on Multimedia and Expo, Beijing, China, 2–5 July 2007; pp. 1383–1386.
22. Fu, Y.; Huang, T.S. Human age estimation with regression on discriminative aging manifold. *IEEE Trans. Multimed.* **2008**, *10*, 578–584. [CrossRef]

23. Scherbaum, K.; Sunkel, M.; Seidel, H.P.; Blanz, V. Prediction of Individual Non-Linear Aging Trajectories of Faces. In Proceedings of the Computer Graphics Forum, Oxford, UK, 12 October 2007; pp. 285–294.
24. Guo, G.; Fu, Y.; Dyer, C.R.; Huang, T.S. Image-based human age estimation by manifold learning and locally adjusted robust regression. *IEEE Trans. Image Process.* **2008**, *17*, 1178–1188. [PubMed]
25. Hayashi, J.; Yasumoto, M.; Ito, H.; Niwa, Y.; Koshimizu, H. Age and gender estimation from facial image processing. In Proceedings of the 41st SICE Annual Conference, Osaka, Japan, 5–7 August 2002; pp. 13–18.
26. Hayashi, J.; Yasumoto, M.; Ito, H.; Koshimizu, H. Method for estimating and modeling age and gender using facial image processing. In Proceedings of the Seventh International Conference on Virtual Systems and Multimedia, Berkeley, CA, USA, 25–27 October 2001; pp. 439–448.
27. Gunay, A.; Nabiyev, V.V. Automatic age classification with LBP. In Proceedings of the 23rd International Symposium on Computer and Information Sciences (ISCIS), Istanbul, Turkey, 27–29 October 2008; pp. 1–4.
28. Gao, F.; Ai, H. Face age classification on consumer images with gabor feature and fuzzy LDA method. In Proceedings of the International Conference on Biometrics, Alghero, Italy, 2–5 June 2009; pp. 132–141.
29. Yan, S.; Liu, M.; Huang, T.S. Extracting age information from local spatially flexible patches. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Las Vegas, NV, USA, 30 March–4 April 2008; pp. 737–740.
30. Yan, S.; Zhou, X.; Liu, M.; Hasegawa-Johnson, M.; Huang, T.S. Regression from patch-kernel. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Anchorage, AK, USA, 23–28 June 2008; pp. 1–8.
31. Mu, G.; Guo, G.; Fu, Y.; Huang, T.S. Human age estimation using bio-inspired features. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Miami, FL, USA, 20–24 June 2009; pp. 112–119.
32. Shan, C. Learning local features for age estimation on real-life faces. In Proceedings of the 1st ACM International Workshop on Multimodal Pervasive Video Analysis, Firenze, Italy, 25–29 October 2010; pp. 23–28.
33. Lu, J.; Liong, V.E.; Zhou, J. Cost-sensitive local binary feature learning for facial age estimation. *IEEE Trans. Image Process.* **2015**, *24*, 5356–5368. [CrossRef] [PubMed]
34. Ranjan, R.; Zhou, S.; Chen, J.C.; Kumar, A.; Alavi, A.; Patel, V.M. Unconstrained age estimation with deep convolutional neural networks. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Santiago, Chile, 7–13 December 2015; pp. 109–117.
35. Chen, J.-C.; Kumar, A.; Ranjan, R.; Patel, V.M.; Alavi, A.; Chellappa, R. A cascaded convolutional neural network for age estimation of unconstrained faces. In Proceedings of the IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS), Angeles, CA, USA, 22–25 October 2016; pp. 1–8.
36. Yang, H.-F.; Lin, B.-Y.; Chang, K.-Y.; Chen, C.-S. Automatic Age Estimation from Face Images via Deep Ranking. In Proceedings of the British Machine Vision Conference (BMVC), Swansea, UK, 7–10 September 2015; p. 55.
37. Yi, D.; Lei, Z.; Li, S.Z. Age estimation by multi-scale convolutional network. In Proceedings of the Asian Conference on Computer Vision, Singapore, 1–5 November 2014; pp. 144–158.
38. Liu, X.; Li, S.; Kan, M.; Zhang, J.; Wu, S.; Liu, W.; Han, H.; Shan, S.; Chen, X. Agenet: Deeply learned regressor and classifier for robust apparent age estimation. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Santiago, Chile, 7–13 December 2015; pp. 258–266.
39. Schmidhuber, J. Deep learning in neural networks: An overview. *Neural Netw.* **2015**, *61*, 85–117. [CrossRef] [PubMed]
40. Hinton, G.E.; Osindero, S.; Teh, Y.-W. A fast learning algorithm for deep belief nets. *Neural Comput.* **2006**, *18*, 1527–1554. [CrossRef] [PubMed]
41. Hinton, G.; Deng, L.; Yu, D.; Dahl, G.E.; Mohamed, A.-R.; Jaitly, N. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Process. Mag.* **2012**, *29*, 82–97. [CrossRef]
42. Sun, Y.; Wang, X.; Tang, X. Deep learning face representation from predicting 10,000 classes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1891–1898.
43. Parkhi, O.M.; Vedaldi, A.; Zisserman, A. Deep face recognition. In Proceedings of the British Machine Vision Conference (BMVC), Swansea, UK, 7–10 September 2015; p. 6.

44. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In Proceedings of the Neural Information Processing Systems Conference, Lake Tahoe, NV, USA, 3–8 December 2012; pp. 1097–1105.
45. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Lake Tahoe, NV, USA, 26 June–1 July 2016; pp. 770–778.
46. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 1–9.
47. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
48. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 431–3440.
49. Argyriou, A.; Evgeniou, T.; Pontil, M. Multi-task feature learning. In Proceedings of the Neural Information Processing Systems Conference, Vancouver, BC, Canada, 3–6 December 2007; pp. 41–48.
50. Obozinski, G.; Taskar, B. Multi-task feature selection. In Proceedings of the 23rd International Conference on Machine Learning (ICML), Pittsburgh, PA, USA, 25–26 June 2006.
51. Wang, L.; Zhu, J.; Zou, H. Hybrid huberized support vector machines for microarray classification. In Proceedings of the 24th International Conference on Machine Learning (ICML), Corvalis, OR, USA, 20–24 June 2007; pp. 983–990.
52. Nie, F.; Huang, H.; Cai, X.; Ding, C. Efficient and robust feature selection via joint L21-norms minimization. In Proceedings of the Neural Information Processing Systems (NIPS) Conference, Vancouver, BC, Canada, 6–9 December 2010; pp. 1813–1821.
53. Ding, C.; Zhou, D.; He, X.; Zha, H. R 1-PCA: Rotational invariant L 1-norm principal component analysis for robust subspace factorization. In Proceedings of the 23rd International Conference on Machine Learning (ICML), Pittsburgh, PA, USA, 25–29 June 2006; pp. 281–288.
54. Gu, Q.; Li, Z.; Han, J. Joint feature selection and subspace learning. In Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), Barcelona, Spain, 16–22 July 2011; pp. 1294–1299.
55. He, R.; Tan, T.N.; Wang, L.; Zheng, W. L21 regularized correntropy for robust feature selection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Providence, RI, USA, 16–21 June 2012; pp. 2504–2511.
56. Liu, F.; Shen, C.; Lin, G. Deep convolutional neural fields for depth estimation from a single image. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 5162–5170.
57. Eidinger, E.; Enbar, R.; Hassner, T. Age and gender estimation of unfiltered faces. *IEEE Trans. Inf. Forensic Secur.* **2014**, *9*, 2170–2179. [CrossRef]
58. Hebda, B.; Kryjak, T. A compact deep convolutional neural network architecture for video based age and gender estimation. In Proceedings of the IEEE Federated Conference in Computer Science and Information Systems (FedCSIS), Gdańsk, Poland, 11–14 September 2016; pp. 787–790.
59. Zhu, L.; Wang, K.; Lin, L.; Zhang, L. Learning a lightweight deep convolutional network for joint age and gender recognition. In Proceedings of the IEEE 23rd International Conference on Pattern Recognition (ICPR), Cancun, Mexico, 4–8 December 2016; pp. 3282–3287.
60. Rothe, R.; Timofte, R.; van Gool, L. Deep expectation of real and apparent age from a single image without facial landmarks. *Int. J. Comput. Vis.* **2018**, 126–144. [CrossRef]
61. Iqbal, M.T.B.; Shoyaib, M.; Ryu, B.; Abdullah-Al-Wadud, M.; Chae, O. Directional Age-Primitive Pattern (DAPP) for Human Age Group Recognition and Age Estimation. *IEEE Trans. Inf. Forensic Secur.* **2017**, *12*, 2505–2517. [CrossRef]
62. Liu, H.; Lu, J.; Feng, J.; Zhou, J. Label-Sensiti00000ve Deep Metric Learning for Facial Age Estimation. *IEEE Trans. Inf. Forensic Secur.* **2018**, *13*, 292–305. [CrossRef]

63. Duan, M.; Li, K.; Yang, C.; Li, K. A hybrid deep learning CNN–ELM for age and gender classification. *Neurocomputing* **2018**, *275*, 448–461. [CrossRef]
64. Rodríguez, P.; Cucurull, G.; Gonfaus, J.M.; Roca, F.X.; Gonzàlez, J. Age and gender recognition in the wild with deep attention. *Pattern Recognit.* **2017**, *72*, 563–571. [CrossRef]