Expert Systems With Applications 85 (2017) 76–86

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Age and gender classification from speech and face images by jointly fine-tuned deep neural networks

Zakariya Qawaqneh[a], Arafat Abu Mallouh[a], Buket D. Barkana[b,*]

[a] Computer Science and Engineering Department, University of Bridgeport, Bridgeport, CT 06604, USA
[b] Electrical Engineering Department, University of Bridgeport, Bridgeport, CT 06604, USA

ARTICLE INFO

Article history:
Received 24 January 2017
Revised 14 May 2017
Accepted 15 May 2017
Available online 16 May 2017

Keywords:
Age and gender classification
Fine tuning
Deep neural networks
Cost function
Face recognition

ABSTRACT

The classification of human's age and gender from speech and face images is a challenging task that has important applications in real-life and its applications are expected to grow more in the future. Deep neural networks (DNNs) and Convolutional neural networks (CNNs) are considered as one of the state-of-art systems as feature extractors and classifiers and are proven to be very efficient in analyzing problems with complex feature space. In this work, we propose a new cost function for fine-tuning two DNNs jointly. The proposed cost function is evaluated by using speech utterances and unconstrained face images for age and gender classification task. The proposed classifier design consists of two DNNs trained on different feature sets, which are extracted from the same input data. Mel-frequency cepstral coefficients (MFCCs) and fundamental frequency (F0) and the shifted delta cepstral coefficients (SDC) are extracted from speech as the first and second feature sets, respectively. Facial appearance and the depth information are extracted from face images as the first and second feature sets, respectively. Jointly training of two DNNs with the proposed cost function improved the classification accuracies and minimized the over-fitting effect for both speech-based and image-based systems. Extensive experiments have been conducted to evaluate the performance and the accuracy of the proposed work. Two publicly available databases, the Age-Annotated Database of the German Telephone Speech database (aGender) and the Adience database, are used to evaluate the proposed system. The overall accuracy of the proposed system is calculated as 56.06% for seven speaker classes and overall exact accuracy is calculated as 63.78% for Adience database.

© 2017 Elsevier Ltd. All rights reserved.

1. Introduction

Classifying the age and gender from speech and face images is a challenging task that has an increasing number of applications in real-life. Recent developments in hardware and software enabled new methods and techniques to be investigated to enhance the performance of age and gender classification systems. Media tools used for recording and noisy recording environments affect the performance of the age and gender classification systems from speech utterances. Age estimation from face images is affected by the pose and variant illumination and resolution of face images. As a result, age and gender classification from speech signals or face images require efficient classifiers and feature extractors in order to achieve satisfactory classification accuracies.

Deep neural networks (DNNs) and convolutional neural networks (CNNs) have shown remarkable success in machine learning and computer vision fields. DNNs and CNNs are capable to extract and classify variant feature sets efficiently after careful training. DNNs and CNNs are most successful when the training set is characterized by a complex feature space that requires high level representation. Currently, most of the speech and image databases are relatively big and variant. In this work, the DNNs and the CNNs are utilized as feature extractors and classifiers for age and gender classification from speech and face images. A new cost function is proposed to fine-tune two DNNs jointly. The proposed cost function is tested and verified by using the aGender speech database and Adience image database. Mel-frequency cepstral coefficients (MFCCs), fundamental frequency (FO), and shifted delta cepstral coefficients (SDC) are used for speaker age and gender classification. Facial and superpixels based feature sets are used for age classification from face images. Two DNNs are trained concurrently on different feature sets that are extracted from the same training sample. The jointly fine-tuned DNNs use the cost function to accommodate the concurrent training and the cost function

* Corresponding author.
E-mail addresses: zqawaqne@my.bridgeport.edu (Z. Qawaqneh), aabumall@my.bridgeport.edu (A.A. Mallouh), bbarkana@bridgeport.edu, bdbarkana@aol.com (B.D. Barkana).

http://dx.doi.org/10.1016/j.eswa.2017.05.037
0957-4174/© 2017 Elsevier Ltd. All rights reserved.

calibrates the network weights and biases. The main contributions of this paper can be listed as below.

- A new cost function is developed for the concurrent and jointly fine-tuned DNNs to optimize the networks parameters.
- A new classifier architecture is proposed. Two DNNs are trained concurrently and fine-tuned jointly where each DNN have different feature set.
- Extensive performance analysis is completed to evaluate the performance of the proposed classifiers and the cost function. Depth features based on image superpixels and their relations are used for age classification from face images. MFCCs and SDC features are used for speaker age and gender classification.

The remainder of the paper is organized as follows: Section 2 reviews the related work. Section 3 presents the proposed work while the experimental results and discussion are given in Section 4. Finally, Section 5 concludes the paper.

2. Related work

Feature extraction and classification are the most important stages in any classification problem. In this section we briefly review related works for speaker age and gender classification and age classification from face images.

2.1. Speaker age and gender classification

Many frequency and time domain feature sets have been evaluated for speaker age and gender classification such as MFCCs, Relative Spectral Transform - Perceptual Linear Prediction (RASTA-PLP), Linear predictive coefficients (LPCs), fundamental frequency (FO), pitch-range features (PR), and energy (Barkana & Zhou, 2015). MFCCs set is the most commonly used spectral feature sets. The advantages of MFCCs come from its ability to model the vocal tract filter in a short time power spectrum (Cerrato, Falcone, & Paoloni, 2000; Gałka et al., 2015). Although the SDCs set was used to analyze speech signals, it is not employed for age and gender classification from speech utterances in previous works. The SDC features capture phonetic information over a long time duration by deriving a delta cepstral pattern over multiple utterance frames. Barkana and Zhou developed a feature set called PR set in time-domain and reported promising classification accuracies (Barkana & Zhou, 2015).

In Deepawale, Bachu, and Barkana (2008), they developed a method to identify speaker's gender by using formants, pitch, and energy features. They combined these features using a classifier and showed that the energy feature between adjacent formants for female speakers is smaller than that of male speakers.

Metze et al. examined multiple classifiers for speaker age and gender classification based on telephone applications. They compared the classification results with human performance for the same data (Metze et al., 2007). Four automatic classification methods were compared: a parallel phone recognizer, a dynamic Bayesian networks to combine prosodic features, a linear prediction analysis, and a GMM based on MFCC features. Accuracies were 54%, 40%, 27%, and 42%, respectively. Overall classification accuracy of human listeners was reported as 55% for the aGender corpus.

SVM and Decision Tree (DT) classifiers were used in Lee and Kwak (2012). MFCCs were extracted in order to build an age and gender classification system for a human robot. They conducted their research on a private corpus. The achieved overall accuracies using MFCC-SVM and MFCC-DT for age classification were 91.39% and 88.37%, respectively. The accuracies for gender classification using the same systems were reported as 93.16% and 91.45%. Bocklet, Stemmer, Zeissler, and Nöth (2010) studied multiple systems with different combinations. They used a combination of several glottal, spectral, and prosodic feature sets. They achieved an overall accuracy of 42.2% by their GMM- universal back- ground model (UBM) classifier.

Bahari used weighted Supervised Non-Negative Matrix Factorization (WSNMF) and general regression neural networks (GRNNs) to design an age and gender classification system. They achieved an accuracy of 96% for gender recognition on a Dutch speech database. For age estimation, the achieved mean absolute error was 7.48 years (Bahari, 2011). Meinedo and Trancoso (2010) applied the fusion technique on four classification models. Short and long term acoustic and prosodic features were extracted. The highest classification accuracy was achieved by the linear logistic regression classifier among the four proposed models. Dobry, Hecht, Avigal, and Zigel (2011) proposed a speech dimension reduction method for age-group classification and precise age estimation. After deploying an SVM with RBF kernel, they noticed that the classifier's performance was improved by using their dimension reduction method. The SVM classifier was faster and less affected by the problem of over-fitting. Bahari, McLaren, and van Leeuwen (2014) proposed an i-vector model for each speech utterance and utilized Least Squares Support Vector Regression (LSSVR) for the estimation of a speaker's age. Their work was tested on telephone conversations from the National Institute for Standards and Technology (NIST-2010).

Li, Han, and Narayanan (2013) proposed a system which combined five classifiers: the Gaussian Mixture Model (GMM) based on MFCC features, the GMM-SVM mean supervector, the GMM-SVM maximum likelihood linear regression (MLLR) supervector, the GMM-SVM Tandem Posterior Probability (TPP) supervector, and the SVM baseline subsystems by using 450-dimensional feature vectors including prosodic features. In addition, they combined two or more systems using a fusion technique in order to increase accuracy. The overall accuracies of the five classifiers were 43.1%, 42.6%, 36.2%, 37.8%, and 44.6%, respectively. The combined GMM based-and-GMM-SVM mean supervector system, the combined GMM-SVM MLLR supervector-and-GMM-SVM TPP supervector system, the combination of the first four classifiers, and the combination of all five classifiers achieved 45.2%, 40.3%, 50.4%, 52.7% of overall accuracies, respectively.

One of the most successful classifiers is deep neural networks (DNNs) in the literature. DNNs have been used by a wide range of applications such as computer vision (Krizhevsky, Sutskever, & Hinton, 2012; Nguyen, Yosinski, & Clune, 2015; Zeiler, 2013), image processing and classification (Ciresan, Meier, & Schmidhuber, 2012; Simonyan & Zisserman, 2014), conversational speech recognition (Hinton, 2012), natural language recognition (Richardson, Reynolds, & Dehak, 2015), speaker verification (Variani, Lei, McDermott, Lopez Moreno, & Gonzalez-Dominguez, 2014) and audio processing (Graves, Mohamed, & Hinton, 2013). DNNs could be used for feature extraction and classification purposes.

Qawaqneh, Mallouh, and Barkana (2017) proposed transformed MFCCs features (T-MFCCs) with a shared labels technique. The T-MFCCs were extracted from the original MFCCs by using the tied-state triphones. After training a DNN on the tied-state triphones as labels, a bottleneck feature (BNF) extractor was used to extract the T-MFCCs set. Mallouh, Qawaqneh, and Barkana (2017) proposed a DNN bottleneck feature set that was generated from the MFCCs set to build a GMM model for each seven age and gender classes by using GMM-UBM technique. Both studies verified the effectiveness of the DNNs in speaker age and gender classification.

2.2. Age classification from face images

Some of the studies that were carried out on the age estimation/classification from face images have focused on feature

extraction methods while the others have focused on classification methods.

Farkas studied the changes of the head size from the infancy to adulthood (Farkas, 1994). Based on these changes, a mathematical model was developed to estimate the age from face images. Hayashi et al. proposed a system, which uses texture and shape features (Hayashi, Yasumoto, Ito, Niwa, & Koshimizu, 2002). Their proposed system was tested on a Japanese database of 500 images.

Different variations of a manifold aging system have been studied and developed by several groups (Fu, Xu, & Huang, 2007; Geng, Zhou, & Smith-Miles, 2007; Scherbaum, Sunkel, Seidel, & Blanz, 2007). The manifold aging pattern of human face, called AGES, constructs a representative face subspace and models the face pattern in a time order as a sequence of the person face images (Geng et al., 2007). One of the main disadvantages of this model is the need for face images of a person at different ages. Fu et al. proposed a method to overcome this problem (Fu et al., 2007). They developed a new manifold space that analyzes and learns the face manifold for different age stages to create a low-dimensional embedding space. In their proposed system, they compensated the absence of a person's face images by using the different persons' face images.

Local binary patterns (LBP) feature set and appearance features were extracted by using an effective texture descriptor (Gunay & Nabiyev, 2008). The feature sets for age estimation were evaluated on the FERET and PIE datasets by k-nearest neighbor (KNN) and AdaBoost classifiers. The KNN classifier achieved 80% of overall accuracy on the FERET database. The AdaBoost classifier achieved 80% and 90% of overall accuracies on the FERET and PIE databases, respectively.

Gabor features and the fuzzy linear discriminate analysis (FLDA) were used to estimate the age from the face image (Gao & Ai, 2009). They showed that Gabor features were more efficient than the LBP features. To handle some of the challenges in face images such as occlusions and blur, the spatial flexible patches (SFP) were proposed as a feature descriptor by Yan, Liu, and Huang (2008). The proposed feature descriptor was tested on the FG-NET database and achieved MAE of 4.94 years.

Bio-inspired features (BIF) were extracted by a pyramid of Gabor filters at all directions of the input image in Mu, Guo, Fu, and Huang (2009). BIF succeeded to handle the small rotations and the scale changes in the input images. MAE of 4.77 was achieved by using an SVM classifier and BIF on the FG-NET database. A combination of BIF and age manifold features were evaluated by Guo et al. and achieved MAE of 2.61 years for females and 2.58 for males on the YGA database by an SVM classifier (Guo, Mu, Fu, Dyer, & Huang, 2009).

In Shan (2010), LBP and Gabor features were extracted from unconstrained face images. A discriminative pattern of the LBP-histogram was learned by using Adaboost. The system achieved 55.9% of overall accuracy on the Group Photos benchmark by using SVM.

CNNs and DNNs have been recently used for age estimation from face images. In Levi and Hassner (2015), a simple CNN architecture was used as a feature extractor and a classifier to avoid overfitting problem. They evaluated their work on the Adience benchmark and achieved 50.7% of overall accuracy. Chen et al. (2016) and Ranjan et al. (2015) proposed new systems for age regression based on facial identification features by using a relatively small deep CNN model. Both works used the same face identification model (Chen, Patel, & Chellappa, 2016) for feature extraction. In Ranjan et al. (2015), a cascaded classification and regression system based on a coarse age classifier was proposed. They introduced an age regressor for each age group based on the extracted features from the coarse age classifier. Then, they used an error correcting method for correcting the regression error for subjects. Chen et al. proposed a system, where the features were extracted from a pre-trained CNN for face identification. The extracted features were fed to a small neural network to regress the age of the subject (Chen et al., 2016).

3. Proposed work

A DNN architecture contains three layers: input, hidden, and label layers. DNN feeds forward the input features through the hidden layers to reach to the cost function, which calculates the difference (error) between the classified label and the real one. After calculating the network error, the back-propagation algorithm is applied to minimize the propagated error by finding the derivative of the cost function with respect to the network parameters; weights and biases. Once the error is minimized at the output layer, it is back-propagated to the previous layer to be used for minimizing the layer error. The same process of layer-wise error minimization is performed through the network layers. The same cycle of feed-forward and back-propagation is repeated until the error between the classified and real value is acceptable.

3.1. Architecture of jointly fine-tuned DDNs and the proposed cost function

In this paper, the proposed approach is based on fine-tuning two DNN networks. The first network, DNN1, is trained on the first feature set, while the second network, DNN2, is trained on the second feature set. It is illustrated in Fig. 1. Both networks are trained simultaneously and their output functions interact with each other. DNN2 is trained with a sigmoid output function (σ) and the loss mean squared error function (L_{DNN2}) as given by Eqs. (1) and (2). DNN1 is trained with a Softmax output function and the loss mean squared error function (L_{DNN1}) which are given by Eqs. (4) and (5).

$$\bar{y} = \sigma = \frac{1}{1 + e^{-z}} \tag{1}$$

$$L_{DNN2} = \frac{1}{n} \times \sum_{j=1}^{n} (y - \overline{y})^2 \tag{2}$$

z is the input vector for the output layer, n is the number of labels, y is the output vector values of the true label, and $\overline{y}$ is the output vector values of the sigmoid function. The derivative (d) of the loss function is defined in Eq. (3), where $\odot$ represents the element-wise product.

$$d(L_{DNN2}) = -(y - \bar{y}) \odot (\bar{y} \odot (1 - y)) \tag{3}$$

DNN1 is trained with the Softmax output function and its cross entropy error function is given by Eqs. (4) and (5).

$$\bar{y} = Softmax = \frac{e^{z_j}}{\sum e^{z_j}} \tag{4}$$

$$L_{DNN1} = -\sum_{j} y_j \log \bar{y}_j \tag{5}$$

y_j is the output vector of values of the true label and $\bar{y}_j$ is the output vector values of the Softmax function. The derivative (d) of the loss function is defined in Eq. (6).

$$d(L_{DNN1}) = -(y - \bar{y}) \tag{6}$$

The DNN2 is trained and jointly fine-tuned with the DNN1 as shown in Fig. 2. The jointly fine-tuned network is trained with the loss function defined in Eq. (7).

$$L_{joint} = L_{DNN1} + L_{DNN2} \tag{7}$$

The Softmax and the Sigmoid are the two parts of our proposed joint fine-tuned loss function. The error on the output layer for

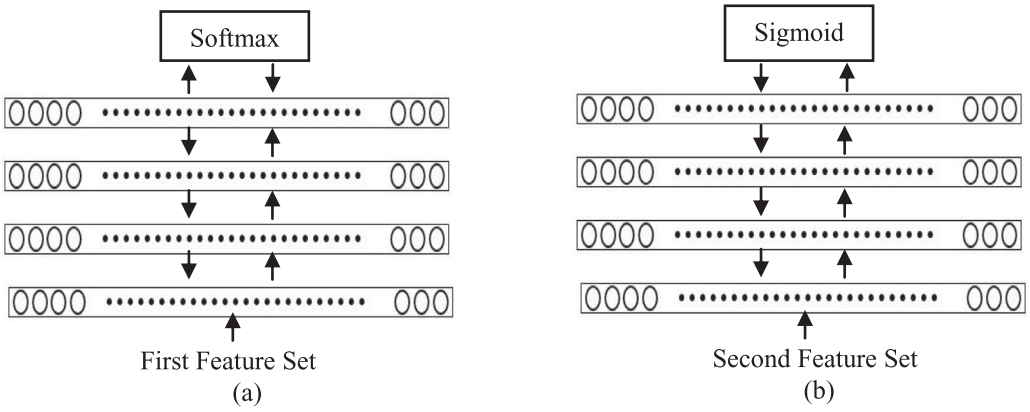

Fig. 1. (a) DNN1 with Softmax output function. (b) DNN2 with sigmoid output function.

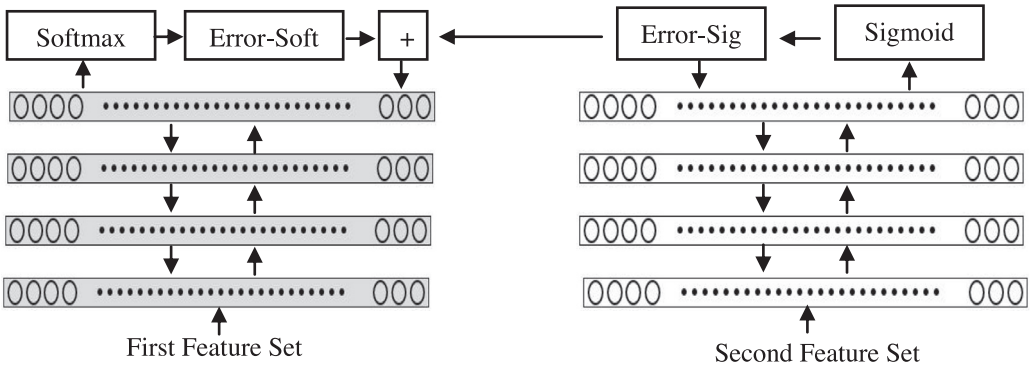

Fig. 2. Architecture of the proposed jointly fine-tuned DNN1 and DNN2 with the proposed cost function.

the jointly fine-tuned network can be calculated by summing the derivatives of both loss function errors of the DNN1 and DNN2 as in Eq. (8)

$$d\left(L_{Joint}\right) = (-(y-\bar{y}) \odot (\bar{y} \odot (1-y))) + (-(y-\bar{y})) \tag{8}$$

Finally, classifying the age and gender of any speaker from any speech utterance or estimating age from face images is done by computing the Softmax output values vector, $(\overline{y_j})$, of the jointly fine-tuned DNN as in Eq. (9).

$$S = \arg max_j \overline{y_j}, \tag{9}$$

Two loss functions are used jointly to fine-tune the newly proposed method by using two different feature sets extracted from the same database. The generated error from the first feature set is different than the generated error from the second set during training. It means that fine-tuning will affect both DNNs differently. By correlating the generated error of the two feature sets on the same input simultaneously, we aim to compute an error that provides more accurate update on the weights and biases of the network during the jointly fine-tuning process. Moreover, the proposed jointly fine-tuning approach uses the proposed cost function. The Softmax function models the joint distribution over the output variables, which means increasing the value for some outputs leads to a decrease in the probability of other outputs. The Sigmoid function models the marginal distributions over the outputs so that increasing or decreasing one of the output values will not affect the other outputs. In this work, the different nature of each output function is merged by adding the error generated from the Sigmoid to the Softmax function in order to identify different error sources.

Over-fitting is a problem in machine learning and it can lead the network parameters to overfit the training data. Different techniques such as dropout and weight decay are used to reduce the effect of overfitting (Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov, 2014). The proposed cost function helps to minimize the effect of over-fitting by jointly fine-tuning the error of DNN1 and DNN2.

3.2. Score level fusion

Let n be the number of the labels of the output layer and the output posterior probabilities of the DNN1 and DNN2 are out1 and out2, respectively. The fused vector S for a given input j can be written by Eq. (10).

$$S_j = \beta \; x \; out_1 + (1 - \beta) \; x \; out_2 \tag{10}$$

The final scoring, (S_j), is considered to be the index of the maximum value of the system fusion output vector. β is the controlling parameter used for fusing the output results of the two networks. Their values are set based on the accomplishment of each network. After conducting extensive experiments, β is set to be 0.8 in this work.

4. Experiments and discussions

In this section, we present the experimental results and compare our approach with the baseline DNN architectures and with other state-of-the-art methods.

4.1. Speaker age and gender classification

4.1.1. Feature sets

Well known MFCCs and SDC feature sets are used as the first and the second feature sets for the DNN1 and DNN2 architectures, respectively. While MFCCs set models the vocal tract filter in a

Table 1
Network settings.

	DNN1	DNN2	Joint fine-tuned DNNs
No. of hidden layers	4	4	4
No. of nodes/hidden layer	1024	1024	1024
Learning rate	Start at 0.1 then decreased by 0.2 every 6 epochs	Start at 0.1 then decreased by 0.2 every 6 epochs	Start at 0.1 then decreased by 0.2 every 3 epochs
No. of epochs	15	20	20
Cost function	Softmax	Sigmoid	L_{joint}
Input features	SDC	MFCCs+F0	SDC + MFCCs + F0
Activation function	Rectified	Rectified	Rectified

Table 2
Overall classification accuracies of the eight networks on the aGender database (%).

	C1(C)	C2(YF)	C3(YM)	C4(MF)	C5(MM)	C6(SF)	C7(SM)	**Overall accuracy**
DNN1 with SDC	54.33	52.60	44.80	25.13	42.33	46.13	55.87	**45.89**
DNN2 with MFCCs+F0	57.47	45.73	44.33	35.40	33.67	35.73	49.67	**43.14**
DNN1 with MFCCs+F0+SDC	62.80	60.23	41.08	38.15	38.31	51.23	56.85	**49.81**
DNN2 with MFCCs+F0+SDC	58.62	57.31	38.08	35.38	37.08	50.15	56.69	**47.62**
DNN1 with SDC and dropout	61.23	61.77	40.92	36.54	36.31	48.92	58.15	**49.12**
DNN2 with MFCCs+F0 and dropout	58.38	44.31	40.77	36.62	33.69	49.23	57.54	**45.79**
Fusion Model	59.69	65.54	39.69	37.00	39.77	52.54	59.92	**50.59**
Jointly Fine-Tuned	65.70	60.38	50.85	46.62	49.62	57.51	61.74	**56.06**

short time power spectrum, the SDC set captures phonetic information over a long-time duration.

4.1.2. Database

The aGender database is used to test the performance of the proposed work. It consists of 47 hours of prompted and free text (Schuller et al., 2010). The number of speakers in the database is 954, and it includes seven categories: Children (C, 7–14 years old), young-females (YF, 15–24 years old), young-males (YM, 15–24 years old), mature-females (MF, 25–54 years old), mature-males (MM, 25–54 years old), senior-females (SF, 55–80 years old), and senior-males (SM, 55–80 years old). The number of utterances in the database is 65,364, and the average length of each utterance is 2.58 seconds. The database was divided into two parts: the training set contains 53,076 utterances (770 speakers), while the test set contains 17,332 utterances (25 speakers/class). The nature of speech content is short commands, single words, and numbers.

4.1.3. Network architectures

The number of nodes in the input layer of the DNN1and DNN2 represent (n x 39) and (n x 14) features, respectively. The target frame is concatenated with its neighboring preceding (n-1)/2 frames and its neighboring following (n-1)/2 frames. The number of frames, n, is set to 11 after rigorous trial and error process. The 11 sequence frames contain the target frame and the previous and the next (n-1)/2 frames. Table 1 shows the network settings of the DNN1, DNN2, and jointly fine-tuned DNNs. The DNN1 and DNN2 with dropout, –a dropout has been added after each layer in the supervised phase–, have the same settings with that of the DNN1 and DNN2. In the fusion model, the DNN1 and DNN2 with dropout are fused by using the score level fusion. It has the same network settings with that of the DNN1 and DNN2.

4.1.4. Results and comparisons

Table 2 shows the overall classification accuracy and individual class accuracies of the DNN1, DNN2 with and without dropout by using individual and concatenated feature sets, score fusion model, and jointly fine-tuned DNNs. The jointly fine-tuned DNNs achieved the highest accuracies among all methods. The confusion matrices of the six models are given in Tables 3–8. For each table the left-most column represents the prediction, and the top row represents the actual label. Bold values indicate the percent of samples correctly classified. The highest misclassification rate occurred among the female-classes groups and the male-classes groups.

In Fig. 3, the performance of the eight networks is compared for female and male classes. All networks achieved the lowest classification accuracies for the middle-aged female (MF) and middle-aged male (MM) classes. However, the jointly fine-tuned network

Table 3
Confusion matrix for seven class age and gender task by DNN1 with SDC (%).

	C	YF	YM	MF	MM	SF	SM
C	**54.33**	22.88	2.67	6.13	0.73	11.13	2.13
YF	13.00	**52.60**	0.40	16.47	0.20	16.93	0.40
YM	0.87	1.00	**44.80**	2.13	26.20	4.60	20.40
MF	4.40	26.47	1.73	**26.13**	1.53	37.67	2.07
MM	1.07	0.80	30.93	1.40	**42.33**	2.60	20.87
SF	4.27	16.20	3.00	23.27	1.20	**46.13**	5.93
SM	1.07	0.47	10.98	0.67	26.87	4.07	**55.87**

Table 4
Confusion matrix for seven class age and gender task by DNN2 with MFCCs+F0 (%).

	C	YF	YM	MF	MM	SF	SM
C	**57.47**	18.73	6.33	6.27	2.33	5.40	3.47
YF	16.93	**45.73**	0.80	23.73	1.20	11.40	0.20
YM	3.20	1.80	**44.33**	2.60	27.60	4.07	16.40
MF	9.20	19.53	2.67	**35.40**	2.40	27.60	3.20
MM	1.73	0.73	31.27	2.47	**33.87**	2.87	27.07
SF	10.00	14.87	2.27	31.40	1.93	**35.73**	3.80
SM	0.87	0.60	16.53	1.33	27.27	3.93	**49.47**

Table 5
Confusion matrix for seven class age and gender task by DNN1 with MFCCs+F0+SDC (%).

	C	YF	YM	MF	MM	SF	SM
C	**62.80**	18.62	5.08	4.00	2.69	4.89	1.92
YF	12.85	**60.23**	0.31	14.31	0.54	11.69	0.08
YM	1.69	1.00	**41.08**	2.54	30.08	1.85	21.77
MF	6.38	28.62	1.31	**38.15**	0.77	24.31	0.46
MM	0.77	0.46	27.69	0.62	**38.31**	1.69	30.46
SF	7.31	14.92	2.15	20.38	1.62	**51.23**	2.38
SM	0.31	0.38	14.15	1.54	24.31	2.46	**56.85**

Table 6
Confusion matrix for seven class age and gender task by DNN2 with MFCC+F0+SDC (%).

	C	YF	YM	MF	MM	SF	SM
C	**58.62**	20.85	6.00	3.69	2.08	6.69	2.08
YF	12.31	**57.31**	0.54	16.69	0.54	12.15	0.46
YM	1.00	0.92	**38.08**	2.08	36.85	1.92	19.15
MF	6.77	28.54	2.00	**35.38**	0.69	26.08	0.54
MM	0.92	0.23	31.54	0.92	**37.08**	1.38	27.92
SF	7.15	15.31	2.62	21.15	1.69	**50.15**	1.92
SM	0.62	0.77	15.54	0.92	22.92	2.54	**56.69**

Table 7
Confusion matrix for seven class age and gender task by jointly fine-tuned network (%).

	C	YF	YM	MF	MM	SF	SM
C	**65.69**	14.85	5.69	3.23	2.08	6.15	2.31
YF	11.69	**60.38**	0.54	17.85	0.69	8.31	0.54
YM	2.15	0.85	**50.85**	2.54	26.92	2.92	13.77
MF	8.77	17.85	1.23	**46.62**	1.23	23.77	0.54
MM	1.15	0.08	21.54	0.31	**49.62**	1.31	26.00
SF	8.11	9.08	1.23	21.08	1.46	**57.51**	1.54
SM	1.62	0	7.69	1.11	22.15	5.69	**61.74**

Table 8
Confusion matrix for seven class age and gender task by fusion model (%).

	C	YF	YM	MF	MM	SF	SM
C	**59.69**	19.92	8.00	2.84	2.00	5.61	1.92
YF	11.15	**65.53**	0.92	14.30	0.53	7.15	0.38
YM	1.15	0.46	**39.69**	3.07	33.61	3.07	18.92
MF	8.84	26.84	2.07	**37.00**	1.61	23.08	0.53
MM	1.15	0	25.92	0.46	**39.77**	1.00	31.69
SF	5.53	13.84	1.76	23.31	1.46	**52.53**	1.53
SM	1.30	0.38	10.84	0.53	23.15	3.84	**59.92**

achieved higher classification accuracies for these two classes. 46.62% and 49.62% are calculated for the MF and MM classes, respectively. The jointly fine-tuned network achieved the best results except for the young female (YF) class, where the DNN1 with dropout and the fusion model achieved better results than the jointly fine-tuned network (Fig. 3). The jointly fine-tuned network outperformed other networks for male classes (Fig. 3). Compared to the results of the DNN1 and DNN2, the fusion model and DNN1 with dropout showed better performance for the female classes. The jointly fine-tuned network outperformed the other networks because of the new cost function that used the errors generated from two feature sets to adjust network parameters. It helped to reduce the effect of over-fitting.

Fig. 4 shows the cross entropy error of the jointly fine-tuned network, the DNN1 with dropout, and the mean squared error of the DNN2 with weight decay. The change in error value of the DNN2 is slightly decreasing in time. A single DNN is exposed to over-fitting and local optima convergence. On the other hand, the change in the error of the jointly fine-tuned network oscillates however it progressively decreases over time. This oscillation permits the jointly fine-tuned DNNs to avoid local optima and allow the network to increase the margin of the learned weights and biases, which in turn avoids over-fitting. DNN1 with the dropout is also analyzed in Fig. 4. The same conclusion can be made, where the error of the DNN1 with dropout as a single network decreases over time but the classification accuracies are affected by over-fitting. Fig. 5 shows the performance of the fusion model for different β values. The network achieved better accuracies when the β value is set to 0.8.

The overall accuracies of some of the previous systems on the aGender database by using the MFCCs set are listed in Table 9. The best performance by using MFCCs is achieved by a DNN classifier with regularized weights in literature. Transformed MFCCs were developed and evaluated on aGender database in Mallouh et al. (2017) and Qawaqneh et al. (2017). The highest reported overall classification accuracy by using the original MFCCs is 45.89%, which was achieved by a DNN classifier with regularized weights (Qawaqneh et al., 2017). The proposed cost function and DNN architectures by this study improved the classification accuracies about 10% for the MFCCs. The results of our proposed work are comparable with the results of the work presented by Mallouh et al. (2017) and Qawaqneh et al. (2017). Although the accuracies of the proposed work are slightly less than the results in Mallouh et al. (2017) and Qawaqneh et al. (2017), one should note that the original MFCCs are used in this work instead of the transformed MFCCs.

Jointly fine-tuned DNN1 and DNN2 by using the proposed cost function achieved significantly higher accuracies than the DNN1 and DNN2 architectures, individually. Jointly fine-tuned architecture achieved 56.06% of overall accuracy while DNN1 and DNN2 architectures by using full set of features achieved 49.81% and 47.62% of overall accuracies, respectively. These results demonstrate that improved performance by the proposed method is due to the new architecture of the network and not because of the discriminative power of the features used. Training and testing of the DNNs on high-dimensional feature sets alleviate the curse-of-dimensionality

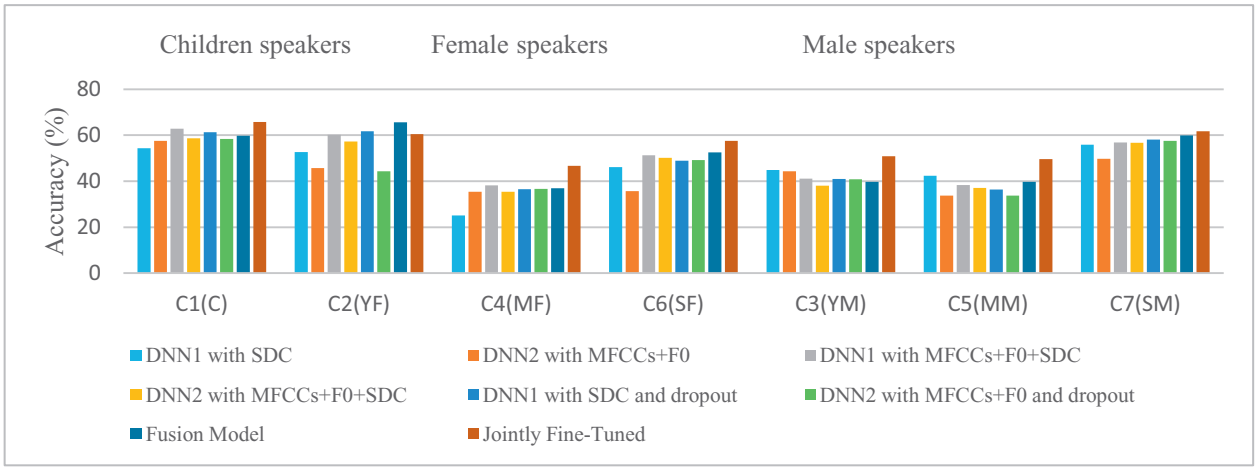

Fig. 3. Classification accuracies of the eight networks for female and male speakers.

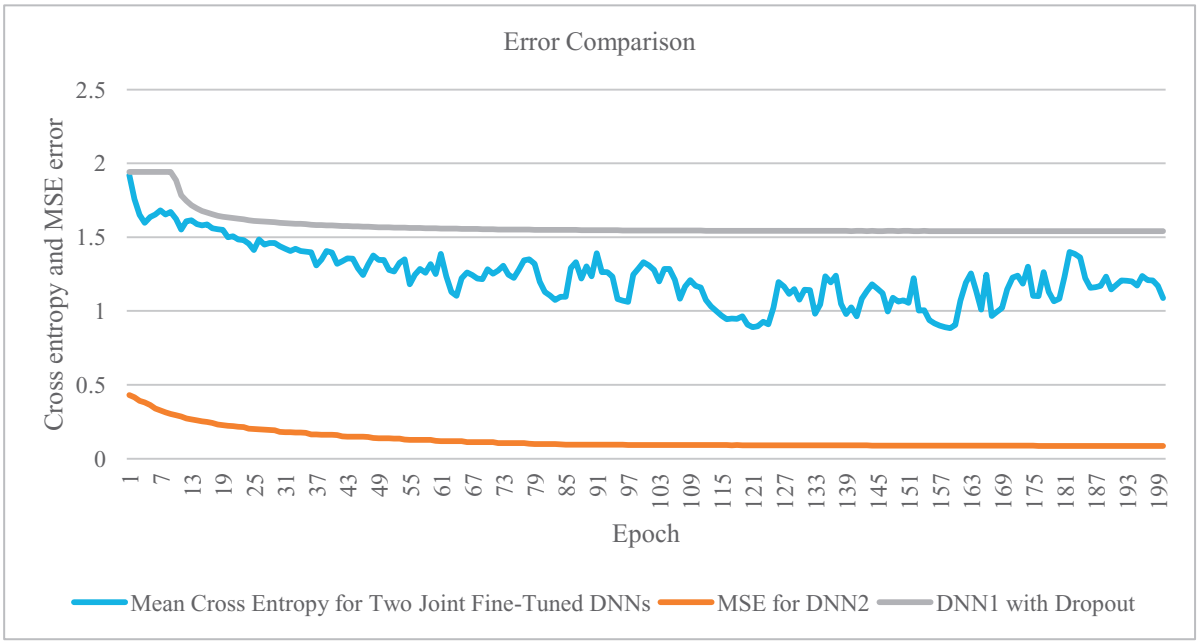

Fig. 4. Cross entropy error over 200 epochs for the proposed work, DNN1 with dropout, and the mean squared error for DNN2 with weight decay.

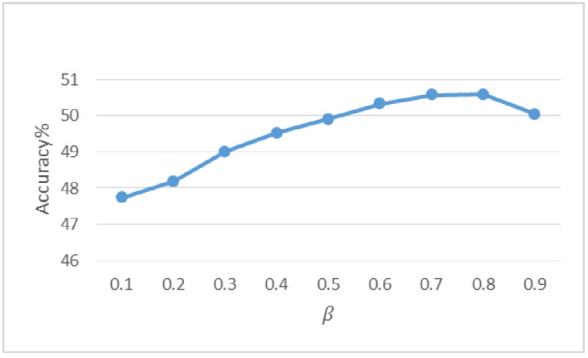

Fig. 5. The performance accuracy of fusing DNN1 and DNN2 with respect to β value.

problem. Such systems require dimensionality reduction. Our experimental results support that the proposed cost function can be successfully used to jointly fine-tune multiple DNNs, which are trained on different feature sets. The fusion model and the jointly fine-tuned networks have the same feature sets however the jointly fine-tuned network significantly outperformed the fusion model.

4.2. Age classification from face images

4.2.1. Feature sets

Two different feature sets extracted from face images to validate the efficiency of the proposed cost function for age estimation from face images. The pre-trained VGG-Face is used to extract the first features set (Parkhi, Vedaldi, & Zisserman, 2015). In this pre-trained model, the last convolutional layer is used for feature extraction. The high dimensional feature vector that is obtained from each training image is reduced to a size of 4096 by using the PCA. The second feature set is extracted by dividing the image to homogeneous superpixels and then finding information from these superpixels. The relationship between each superpixel and its neighboring superpixels is used to get the depth information. For this purpose, we used the pre-trained model that finds the features depending on the depth information (Liu, Shen, & Lin, 2015). The feature vector that is obtained from the last convolutional layer for each training image is reduced to a size of 512 by using the PCA.

4.2.2. Database

Adience database is the most recent benchmark of facial images. This database is designed for age estimation (Eidinger, Enbar, & Hassner, 2014). The Adience database consists of 23 K of unconstrained face images collected from the Flicker website. It is a challenging database due to the type of face images, which are unfiltered wild images. The database has 2284 subjects and eight age labels.

4.2.3. Network architecture

The input feature set for the DNN1 is the facial information obtained from the pre-trained VGG-Face model for face recognition. The number of nodes in the input layer is 4096. The input feature set for the DNN2 is the depth information and the number of nodes in the input layer is 512. The network settings for the DNN1, DNN2, and the jointly fine-tuned DNNs are given in Table 10.

4.2.4. Results and comparisons

Several experiments have been carried out for age classification from face images. The DNN1 by using the facial features, DNN2 by using the depth from superpixels features, and the jointly fine-tuned DNNs by using the proposed cost function, are evaluated. Table 11 presents the exact and 1-off (when a person belongs to his/her exact group or the group immediately before or after his/her exact group) classification accuracies of these networks. The overall accuracy of the jointly fine-tuned network outperforms the DNN1 and DNN2 in terms of the exact and the 1-off accuracy.

Tables 12–16 show the confusion matrices for the DNN1, DNN2, and the jointly fine-tuned DNNs architectures, respectively. For each table the leftmost column represents the prediction, and the top row represents the actual label. Bold values indicate the percent of samples correctly classified.

It can be observed that the DNN1 architecture with facial features outperformed the DNN2 architecture with superpixel based features by about 4% in exact accuracy and about 13% in 1-off accuracy. This might be due to the fact that the facial features used

Table 9
Overall performance comparison in speaker age and gender classification. (*Bold values represent the performances of the proposed systems by this work).*

Work	System	Features	Overall accuracy (%)
Li et al. (2013)	GMM base	13 MFCCs, 450-dimensional acoustic features, prosodic features (F0, F0 envelop, jitter, and shimmer)	43.1
	Mean Super Vector		42.6
	MLLR Super Vector		36.2
	TPP Super Vector		37.8
	SVM Base		44.6
	MFuse 1 + 2		45.2
	MFuse 3 + 4		40.3
	MFuse 1 + 2 + 3 + 4		50.4
	MFuse 1 + 2 + 3 + 4 + 5		52.7
Qawaqneh et al. (2017)	i-vector	MFCCs	43.60
		T-MFCCs	56.13
	DNN with regularized weights	MFCCs	45.89
		T-MFCCs	58.98
	DNN with random weights	MFCCs	43.67
		T-MFCCs	55.04
Mallouh et al. (2017)	GMM-UBM	MFCCs	43.55
	GMM–UBM	T-MFCCs	57.63
This work	**DNN1**	**SDCs**	**45.89**
	DNN2	**MFCCs+FO**	**43.14**
	DNN1	MFCCs+F0+SDC	**49.81**
	DNN2	MFCCs+F0+SDC	**47.62**
	DNN1 with dropout	**SDCs**	**49.12**
	DNN2 with dropout	**MFCCs+FO**	**45.79**
	Fusion model	**SDCs+ MFCCs+FO**	**50.59**
	Joint Fine-Tuned	**SDCs+ MFCCs+FO**	**56.06**

Table 10
Proposed network architectures and settings.

	DNN1	DNN2	Joint fine-tuned DNNs
No. of hidden layers	2	2	2
No. of nodes/hidden layer	1024	1024	512
Learning rate	0.1	0.1	Start at 0.1 then decreased by 0.2 every 3 epochs
Dropout	0.7	0.7	0.5
Weight decay	10^{-3}	10^{-3}	10^{-4}
No. of epochs	15	15	20
Cost function	Softmax	Sigmoid	L_{joint}
Input features	Facial	Superpixel	Facial+Superpixel
Activation function	Rectified	Rectified	Rectified

Table 11
Overall classification accuracies on Adience database (%).

	0–2	4–6	8–13	15–20	25–32	38–43	48–53	60-	**Accuracy%**	**1-off Acc**
DNN1with facial features	88.41	60.18	39.12	43.61	67.14	43.79	14.52	57.20	**57.45**	**94.32**
DNN2 with superpixels features	86.34	58.82	34.18	15.01	80.78	11.05	10.88	53.31	**53.62**	**81.40**
DNN1with facial and superpixels features	88.82	66.32	44.12	54.63	61.55	40.43	21.58	74.71	**59.22**	**91.63**
DNN2 with facial and superpixels features	91.10	53.68	44.41	29.07	75.28	37.08	2.07	81.71	**58.71**	**89.02**
Jointly Fine-Tuned DNNs	85.92	62.28	45.29	36.12	76.70	44.97	14.52	84.44	**62.37**	**94.46**

by the DNN1 have more significant information about persons' age. It is also observed that the performance of the DNN1 is the same or better than the performance of the jointly fine-tuned DNNs for two age groups, (15–20) and (48–53), which the DNN2 performed poorly for these two age groups. On the other hand the DNN1 and DNN2 architectures achieved simialar performance with full set of features, which are facial features and superpixels based features. Both architectures achieved about 59% of exact accuracy and about 90% of 1-off accuracy. Jointly training of two networks by using the proposed cost function improved the classification accuracies significantly. We compare our proposed work with the recent noteworthy state-of-the-art methods on Adience dataset in Table 17. The jointly fine-tuned network outperformed the other methods in terms of the exact and 1-off accuricies.

Eidinger et al. used the droupout-SVM approach to avoid overfitting and proposed face aligment techniuqe to solve the uncertainties of the facial feature extractor (Eidinger et al., 2014). In our work, the faces are not aligned for extracting the features in the pre-processing stage. Levi and Hassner employed DNNs for the first time to estimate age and gender from face images (Levi & Hassner, 2015). They used a relatively simple and shallow network for feature extracting and classifying stages. They proposed an oversampling technique to partially overcome face misalignment in the images. Chen et al. used thousands of images to train face identification model to be used as base model for extracting image features in their proposed age classification system (Chen et al., 2016). Their system requires high computation time. In contrast, we utilized facial features that were already extracted and evaluated by a well-trained model for face recognition task. This way, our system does not require any extra training so that the required computation time of our system is considerably low.

Table 12
Confusion matrix of the DNN1 (Facial features) (%).

	0–2	4–6	8–13	15–20	25–32	38–43	48–53	60-
0–2	**88.41**	10.56	0.21	0.00	0.83	0.00	0.00	0.00
4–6	24.04	**60.18**	12.46	2.46	0.88	0.00	0.00	0.00
8–13	0.88	11.18	**39.12**	36.47	11.18	0.59	0.59	0.00
15–20	0.88	0.44	9.69	**43.61**	43.17	2.20	0.00	0.00
25–32	0.00	0.09	1.89	9.66	**67.14**	19.13	1.61	0.47
38–43	0.20	0.20	0.79	2.96	37.67	**43.79**	11.83	2.56
48–53	0.00	0.00	0.83	0.00	7.88	58.51	**14.52**	18.26
60-	0.00	0.00	0.00	1.56	1.56	10.12	29.57	**57.20**

Table 13
Confusion matrix of the DNN2 (Depth features based on superpixels and their relations) (%).

	0–2	4–6	8–13	15–20	25–32	38–43	48–53	60-
0–2	**86.34**	12.84	0.21	0.00	0.41	0.00	0.00	0.21
4–6	24.30	**58.82**	8.25	3.86	3.19	0.35	0.00	1.23
8–13	1.76	12.06	**34.18**	10.00	37.00	2.35	0.00	2.65
15–20	0.00	4.41	15.42	**15.01**	57.23	4.41	1.32	2.20
25–32	0.57	1.33	4.17	5.59	**80.78**	4.07	1.23	2.27
38–43	0.20	2.56	2.96	3.75	62.13	**11.05**	6.90	10.45
48–53	0.00	3.32	2.49	5.81	47.62	14.94	**10.88**	14.94
60-	0.39	1.56	2.72	5.45	24.90	7.78	3.89	**53.31**

Table 14
Confusion matrix of the DNN1 (Facial features + superpixels and their relations features) (%).

	0–2	4–6	8–13	15–20	25–32	38–43	48–53	60-
0–2	**88.82**	10.97	0.00	0.00	0.00	0.00	0.00	0.21
4–6	22.81	**66.32**	8.42	1.40	0.35	0.18	0.00	0.53
8–13	0.59	7.06	**44.12**	33.53	12.06	2.06	0.59	0.00
15–20	0.00	0.00	9.25	**54.63**	31.72	2.20	1.32	0.88
25–32	0.19	0.09	3.41	12.31	**61.55**	19.03	2.94	0.47
38–43	0.00	0.59	0.59	9.86	26.63	**40.43**	12.03	9.86
48–53	0.00	0.00	0.83	0.41	6.22	49.79	**21.58**	21.16
60-	0.00	0.00	0.00	0.78	1.95	7.00	15.56	**74.71**

Table 15
Confusion matrix of the DNN2 (Facial features + superpixels and their relations features) (%).

	0–2	4–6	8–13	15–20	25–32	38–43	48–53	60-
0–2	**91.10**	8.28	0.41	0.00	0.00	0.21	0.00	0.00
4–6	26.14	**53.68**	15.44	2.98	1.75	0.00	0.00	0.00
8–13	0.59	6.76	**44.41**	20.29	25.59	2.06	0.00	0.29
15–20	1.32	0.44	15.42	**29.07**	52.86	0.88	0.00	0.00
25–32	0.00	0.19	4.26	4.64	**75.28**	12.88	1.52	1.23
38–43	0.59	0.79	0.99	7.69	40.24	**37.08**	0.99	11.64
48–53	0.00	0.41	0.83	1.66	13.69	61.00	**2.07**	20.33
60-	0.00	0.00	0.39	0.39	3.50	13.23	0.78	**81.71**

Table 16
Confusion Matrix of the jointly fine-tuned network (Facial and depth features).

	0–2	4–6	8–13	15–20	25–32	38–43	48–53	60-
0–2	**85.92**	12.63	0.41	0.00	1.04	0.00	0.00	0.00
4–6	23.16	**62.28**	12.11	1.75	0.53	0.00	0.00	0.18
8–13	0.59	7.65	**45.29**	30.59	14.71	0.88	0.00	0.29
15–20	0.00	0.44	11.45	**36.12**	50.66	1.32	0.00	0.00
25–32	0.00	0.09	1.61	3.13	**76.70**	17.14	0.95	0.38
38–43	0.20	0.00	0.79	0.99	37.67	**44.97**	7.50	7.89
48–53	0.00	0.00	0.00	0.00	9.13	39.83	**14.52**	36.51
60-	0.00	0.00	0.00	1.17	0.78	5.45	8.17	**84.44**

4.3. General discussion and implications

In this work, we evaluated the performance of the proposed cost function to fine tune two DNNs jointly for age and gender classification from speech signals and age estimation from face images. We conducted extensive experiments on the aGender speech database and the Adience image database. Experimantal results showed that the jointly fine-tuned network with the proposed cost function outperformed the previous systems. Although the age estimation from face images and speaker age and gender classification are two different fields, our proposed cost function succesfuly implemented in both fields and improved the classification accuricies significantly. It can be inferred that the proposed cost function can be considered as feature set independent and it can be used to fine-tune two DNNs for any classification task.

Table 17
Comparison with the state-of-the-art methods on Adience database (%).

Method		Exact accuracy	1-off accuracy
Eidinger et al. (2014)		45.1	79.5
Levi and Hassner (2015) using single crop		49.5	84.6
Levi and Hassner (2015) using over-sample		50.7	84.7
Chen et al. (2016)		52.88	88.45
This work	**DNN1 with facial features**	**57.45**	**94.32**
	DNN2 with superpixels features	**53.62**	**81.40**
	DNN1 with facial and superpixels features	59.22	91.63
	DNN2 with facial and superpixels features	58.71	89.02
	Jointly fine-tuned network	**63.78**	**93.70**

One of the benefits of the proposed cost function is the ability to reduce the effect of the overfitting problem. This is achieved by involving the propagated errors generated from two networks that perform simultaneous learning process on two feature sets. These feature sets were extracted from the same speech utterances or face images in this work. The two networks calculated the error depending on two different cost functions, where each one has a different approach to calculate the error. Another advantage of the proposed work is to involve more and different feature sets in order to optimize the network parameters and to minimize the overfitting problem further. Moreover, the jointly fine-tuning of two networks provides a platform to combine and extract the distinctive features of two different feature spaces by coupling the learning process of two networks using the proposed cost function.

Using fixed and unified learning rate for both networks leads to propagate different and incompatible error rates without reflecting the actual joint learning. The later happens when the used feature sets are loosely related and each set requires different network parameters in order to extract the desired patterns of data that contains higher representations than the initial form of the feature sets. In this work, different learning rates have been calibrated automatically for the jointly fine-tuned network to reach stable learning ratios between the two DNNs. In our case, calibrating the learning rates of the two DNNs was quite fast and non-problematic, hence the jointly fine-tuned networks converged in a reasonable time.

A possible implication is that the number of networks and feature sets will affect the performance of the proposed work. For instance, if three networks with three feature sets are used, the cost function should be modified in order to calculate the effect of the third feature set especially if the new set is unrelated to the other two feature sets.

5. Conclusions and future work

In this paper, it is shown that two DDNs can be successfully jointly fine-tuned by the proposed cost function for speaker age and gender classification and age estimation from face images. Two DNN architectures with two different feature sets are presented and jointly fine-tuned. This approach lessens the possibility of the dimensionality and overfitting problem during training and testing while utilizing more features to represent age and gender information. The jointly fine-tuned networks and the new cost function calculates the network parameters in an accurate way that reflect the learning using two different feature sets in two networks. The first network, DNN1, is trained using the first feature set. The cross-entropy is used as the cost function. The Softmax function is used at the output layer. The second network, DNN2 is trained on another feature set. The sigmoid function is used to calculate the output layer probabilities, and the mean squared error loss function is used to calculate the propagated error in the DNN2.

The contributions of this work can be summarized as: (1) the proposed cost function is derived from the well-known softmax and sigmoid functions. After extensive experimental work, the effectiveness of the function is verified in two fields for the same task, speaker age and gender classification and age estimation from face image; (2) jointly fine-tuned DNNs with different feature sets show great potential to improve the accuracies in age and gender classification from speech or face images; (3) the SDC feature set, which is extracted from speech utterances and the depth features, which are extracted from the superpixels and their relations in face images are employed for the first time for the task.

The framework of the proposed work can be applied to any classification problem such as language identification, speaker verification, or image classification, where there are different feature sets. As future work, we plan to extend the proposed method to incorporate more feature sets and jointly fine-tune more than two DNNs. As an example we plan to generate transformed MFCCs and SDC sets using bottleneck feature extractor and jointly fine-tune two DNNs for speaker age and gender classification.

No conflict of interest.

Acknowledgments

We thank the maintainers of the Age-Annotated Database of German Telephone Speech and the Adience database.

References

Bahari, M. H. (2011). Speaker age estimation and gender detection based on supervised non-negative matrix factorization. In *Paper presented at IEEE workshop on the biometric measurements and systems for security and medical applications (BIOMS)* (pp. 1–6).

Bahari, M. H., McLaren, M., & van Leeuwen, D. A. (2014). Speaker age estimation using i-vectors. *Engineering Applications of Artificial Intelligence, 34*, 99–108.

Barkana, B. D., & Zhou, J. (2015). A new pitch-range based feature set for a speaker's age and gender classification. *Applied Acoustics, 98*, 52–61.

Bocklet, T., Stemmer, G., Zeissler, V., & Nöth, E. (2010). Age and gender recognition based on multiple systems-early vs. late fusion. In *Paper presented at the INTERSPEECH* (pp. 2830–2833).

Cerrato, L., Falcone, M., & Paoloni, A. (2000). Subjective age estimation of telephonic voices. *Speech Communication, 31*(2), 107–112.

Chen, J.-C., Kumar, A., Ranjan, R., Patel, V. M., Alavi, A., & Chellappa, R. (2016). A cascaded convolutional neural network for age estimation of unconstrained faces. In *Paper presented at IEEE 8th International conference on the biometrics theory, applications and systems (BTAS)* (pp. 1–8).

Chen, J.-C., Patel, V. M., & Chellappa, R. (2016). Unconstrained face verification using deep cnn features. In *Paper presented at IEEE winter conference on the applications of computer vision (WACV)* (pp. 1–9).

Ciresan, D., Meier, U., & Schmidhuber, J. (2012). Multi-column deep neural networks for image classification. In *Paper presented at the IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 3642–3649).

Deepawale, D. S., Bachu, R., & Barkana, B. D. (2008). Energy estimation between adjacent formant frequencies to identify speaker's gender. In *Paper presented at the fifth international conference on information technology: new generations (ITNG)* (pp. 772–776).

Dobry, G., Hecht, R. M., Avigal, M., & Zigel, Y. (2011). Supervector dimension reduction for efficient speaker age estimation based on the acoustic speech signal. *IEEE Transactions on Audio, Speech, and Language Processing, 19*(7), 1975–1985.

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

Farkas, L. G. (1994). *Anthropometry of the head and face* (p. 1994). New York: Raven Press.

Fu, Y., Xu, Y., & Huang, T. S. (2007). Estimating human age by manifold analysis of face pictures and regression on aging features. In *Paper presented at the IEEE international conference on multimedia and expo* (pp. 1383–1386).

Gałka, J., Grzybowska, J., Igras, M., Jaciów, P., Wajda, K., Witkowski, M., et al. (2015). System supporting speaker identification in emergency call center. In *Paper presented at the sixteenth annual conference of INTERSPEECH* (pp. 724–725).

Gao, F., & Ai, H. (2009). Face age classification on consumer images with gabor feature and fuzzy lda method. In *Paper presented at the third international conference on biometrics* (pp. 132–141).

Geng, X., Zhou, Z.-H., & Smith-Miles, K. (2007). Automatic age estimation based on facial aging patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 29*(12), 2234–2240.

Graves, A., Mohamed, A.-R., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. In *Paper presented at the IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 6645–6649).

Gunay, A., & Nabiyev, V. V. (2008). Automatic age classification with LBP. In *Paper presented at the 23rd international symposium on computer and information sciences (ISCIS'08)* (pp. 1–4).

Guo, G., Mu, G., Fu, Y., Dyer, C. R., & Huang, T. S. (2009). A study on automatic age estimation using a large database. In *Paper presented at the IEEE 12th international conference on in computer vision (ICCV)* (pp. 1986–1991).

Hayashi, J., Yasumoto, M., Ito, H., Niwa, Y., & Koshimizu, H. (2002). Age and gender estimation from facial image processing. In *Paper presented at the proceedings of the 41st annual conference (SICE): Vol. 1* (pp. 13–18).

Hinton, G. (2012). A practical guide to training restricted boltzmann machines. In *Neural networks: tricks of the trade: Vol. 7700* (pp. 599–619).

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Paper presented at 26th annual conference on neural information processing systems (NIPS)* (pp. 1097–1105).

Lee, M.-W., & Kwak, K.-C. (2012). Performance comparison of gender and age group recognition for human-robot interaction. *International Journal of Advanced Computer Science and Applications (IJACSA), 3*(12), 207–211.

Levi, G., & Hassner, T. (2015). Age and gender classification using convolutional neural networks. In *Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 34–42).

Li, M., Han, K. J., & Narayanan, S. (2013). Automatic speaker age and gender recognition using acoustic and prosodic level information fusion. *Computer Speech & Language, 27*(1), 151–167.

Liu, F., Shen, C., & Lin, G. (2015). Deep convolutional neural fields for depth estimation from a single image. In *Paper presented at the proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5162–5170).

Mallouh, A. A., Qawaqneh, Z., & Barkana, B. D. (2017). New transformed features generated by deep bottleneck extractor and a GMM–UBM classifier for speaker age and gender classification. *Neural Computing and Applications*, 1–13. doi:10.1007/s00521-017-2848-4.

Meinedo, H., & Trancoso, I. (2010). Age and gender classification using fusion of acoustic and prosodic features. In *Paper presented at the INTERSPEECH* (pp. 2818–2821).

Metze, F., Ajmera, J., Englert, R., Bub, U., Burkhardt, F., Stegmann, J., et al. (2007). Comparison of four approaches to age and gender recognition for telephone applications. *Paper presented at the IEEE international conference on acoustics, speech and signal processing (ICASSP)*: Vol. 4 IV–1089.

Mu, G., Guo, G., Fu, Y., & Huang, T. S. (2009). Human age estimation using bio-inspired features. In *Paper presented at the IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 112–119).

Nguyen, A., Yosinski, J., & Clune, J. (2015). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Paper presented at the IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 427–436).

Parkhi, O. M., Vedaldi, A., & Zisserman, A. (2015). Deep face recognition. *Paper presented at the British machine vision conference*.

Qawaqneh, Z., Mallouh, A. A., & Barkana, B. D. (2017). Deep neural network framework and transformed MFCCs for speaker's age and gender classification. *Knowledge-Based Systems, 115*, 5–14.

Ranjan, R., Zhou, S., Cheng Chen, J., Kumar, A., Alavi, A., Patel, V. M., et al. (2015). Unconstrained age estimation with deep convolutional neural networks. In *Paper presented at the proceedings of the ieee international conference on computer vision workshops* (pp. 109–117).

Richardson, F., Reynolds, D., & Dehak, N. (2015). Deep neural network approaches to speaker and language recognition. *IEEE Signal Processing Letters, 22*(10), 1671–1675.

Scherbaum, K., Sunkel, M., Seidel, H. P., & Blanz, V. (2007). Prediction of individual non-linear aging trajectories of faces. In *Paper presented at the computer graphics forum: 26* (pp. 285–294).

Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C. A., et al. (2010). The INTERSPEECH 2010 paralinguistic challenge. In *Paper presented at the INTERSPEECH* (pp. 2795–2798).

Shan, C. (2010). Learning local features for age estimation on real-life faces. In *Paper presented at the proceedings of the 1st ACM international workshop on multimodal pervasive video analysis* (pp. 23–28).

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *Presented at third international conference on learning representations (ICLR)*.

Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research, 15*(1), 1929–1958.

Variani, E., Lei, X., McDermott, E., Lopez Moreno, I., & Gonzalez–Dominguez, J. (2014). Deep neural networks for small footprint text-dependent speaker verification. In *Paper presented at the IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 4052–4056).

Yan, S., Liu, M., & Huang, T. S. (2008). Extracting age information from local spatially flexible patches. In *Paper presented at the IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 737–740).

Zeiler, M. D. (2013). *Hierarchical convolutional deep learning in computer vision* Ph.D. Thesis, Janary 2014. New York University.