Age interval and gender prediction using PARAFAC2 and SVMs based on visual and aural features

Evangelia Pantraki[1], Constantine Kotropoulos[1*], Andreas Lanitis[2]

[1]Department of Informatics, Aristotle University of Thessaloniki, Thessaloniki, Greece

[2]Department of Multimedia & Graphic Arts , Cyprus University of Technology , Limassol, Cyprus

[*]Corresponding author: Tel: +30-231099.8225; E-mail:costas@aiia.csd.auth.gr

Abstract: Parallel Factor Analysis 2 (PARAFAC2) is employed to reduce the dimensions of visual and aural features and provide ranking vectors. Subsequently, score level fusion is performed by applying a Support Vector Machine classifier to the ranking vectors derived by PARAFAC2 to make gender and age interval predictions. The aforementioned procedure is applied to the Trinity College Dublin Speaker Ageing database, which is supplemented with face images of the speakers and two single-modality benchmark datasets. Experimental results demonstrate the advantage of using combined aural and visual features for both prediction tasks.

1. Introduction

A biometric characteristic is a measurable, physical characteristic or personal behavioural trait used to recognize the identity, or verify the claimed identity, of an individual. If a time lapse between the enrolment and test phase exists, the identification/verification decision is heavily influenced [1, 2]. However, ageing related effects on biometric features provide the necessary clues for age prediction or estimation.

Here, a contribution to the field of soft biometrics for gender and age interval prediction relying on Parallel Factor Analysis 2 (PARAFAC2) [3] is investigated. We extend the approach in [4], which was based solely on speech utterances, by proposing a bimodal verification system that takes into account both speech utterances and face images in order to predict gender and age intervals. The bimodal system proposed here is motivated merely by the need to improve age estimation ability in cases that both speech and face images are available. The consideration of both modalities could potentially be useful when one of two modalities provides noisy/corrupted input. Typical applications where the combination of aural and visual cues is useful include the analysis of video sequences of crime scenes captured by surveillance cameras/microphones. In such cases, either speech signals or face images of suspects may be corrupted with noise or occlusions, hence the combination of speech and face data could yield better age estimates of the persons appearing in the video, supporting in that way the process of suspect identification. Another application domain is the automatic detection of under aged actors in videos, in an attempt to support the development of forensic tools that can be used for dealing with the escalating problem of child pornography [5]. Both approaches, the unimodal one in [4] and bimodal one here, utilize the powerful decomposition properties of PARAFAC2, but differ significantly in the procedure followed in order to make predictions. That is, the proposed bimodal method incorporates Support Vector Machines (SVMs) in the decision making process. To the best of our knowledge this is the first time a bimodal method involving aural and visual features is used to address age interval and gender prediction.

The aforementioned framework is applied first to the Trinity College Dublin Speaker Ageing (TCDSA) database [6]. Since the TCDSA database includes only speech samples for a set of speakers, the dataset is supplemented with face images of each speaker. For the purposes of this paper, an extended version of the TCDSA dataset is created that includes face images of the TCDSA speakers that are contemporary of their speech recordings. When the proposed framework is applied to the aforementioned extended TCDSA dataset, using a Leave-One-Person-Out (LOPO) evaluation scheme, promising results are demonstrated, when either noise-free or noisy speech utterances are employed along with face images. Further experiments have been conducted in two single-modality widely used benchmark datasets, namely the facial dataset FG-NET [7] and the NIST 2008 Speaker Recognition Evaluation (SRE) Test Set [8].

2. Related work

Ageing has various effects on human face and voice. An evaluation of speaker verification on the TCDSA database with a Gaussian Mixture Model - Universal Background Model (GMM-UBM) system revealed that the verification scores of genuine speakers decreased progressively as the time span between training and testing increased, while the imposter scores were less affected [6]. The addition of temporal information to the mel frequency cepstral coefficients (MFCCs) caused an increase in the rate of degradation [9]. The performance of the i-vector system in terms of both discrimination and calibration was found to degrade progressively as the absolute age difference between the training and test samples increased [10]. In [11], Linear Discriminant Analysis was performed to reduce the dimension of i-vectors and Support Vector Regression was utilized for automatic age estimation. A Partial Least Squares based ranker was proposed for age estimation in [12].

Many methods have been developed for the automatic prediction of biometric characteristics based on facial characteristics and the extraction of facial ageing patterns [13]. Recent surveys on soft biometrics based on facial features can be found in [14, 15]. Classification for age interval prediction is also demonstrated in [16], where Principal Component Analysis is adopted to reduce the feature dimensions, and an SVM is utilized for decision making. Current trends in facial age estimation employ the use of biologically inspired features (BIFs) [17] and adaptive age label distributions [18]. In [19], convolutional neural networks were deployed for the task of joint age interval and gender classification, while SVMs were deployed in [20].

3. Datasets

The longitudinal TCDSA database [6] has been used in the first set of experiments. The database contains recordings spanning a year range per speaker varying between 30 and 60 years at irregular intervals between 1 to 10 years. The duration of speech recordings in the TCDSA database varies from 25 seconds to 35 minutes. The database includes a different number of recordings per speaker, varying from 4 to 47 recordings per speaker. The total number of speakers included in the TCDSA dataset is 26, including 15 males and 11 females.

Face images were collected for each speaker of the dataset by locating publicly available visual material portraying the speakers included in the TCDSA database. Effort has been devoted so that the face images were captured close to the speakers' age. Since the exact matching was difficult, a 3-year tolerance was allowed between the age of a person when his/her face was captured and the age associated to his/her utterance. Such a 3-year tolerance is not expected to affect the exactness of speech and face image matching, because 5-year age intervals are typical to age group classification.

A total duration of 30 seconds is kept from each recording or less if the recording's duration is shorter than 30 seconds. If many face images of the person at the age of the speech recording have been collected, more than one segments of 30 seconds long are kept. A total of 227 recordings could be matched with contemporary face images. Finally, the total number of speakers included in the extended TCDSA audio-visual dataset was 25, including 14 males and 11 females. The collected face images were resized to 60×60 pixels and the face was cropped in order to remove background. All face images were converted to grayscale. Some examples of the collected face images for four speakers of the extended TCDSA dataset are depicted in Figure 1. Pose and illumination vary greatly over the collected face images, as can be observed by the sample face images depicted in Figure 1. To the best of our knowledge, the extended TCDSA dataset that combines age separated speech samples and face images, is a unique dataset that supports bimodal age interval prediction experiments using aural and visual features.

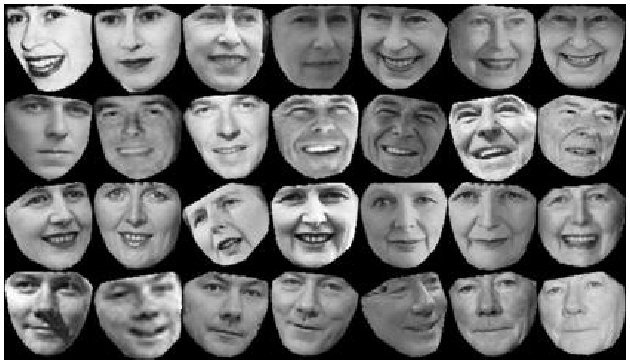


Fig. 1 Face images depicting four speakers of the extended TCDSA dataset at ascending ages.

A second set of experiments was conducted on well known benchmark datasets albeit the latter ones are unimodal ones. These datasets were i) the FG-NET dataset, which comprises of 1002 face images that belong to 82 unique persons (48 male and 34 female) at various ages [7] and ii) the NIST 2008 SRE set [8]. In particular, we used a subset of 1016 recordings that belonged to 458 speakers, 172 of which were male and the rest were female. Similar to the TCDSA database, a segment of 30 seconds duration was kept from each recording.

4. Proposed method

For each speech recording and face image, a feature vector is extracted as follows. Auditory cortical representations are extracted from speech utterances. These descriptors are inspired by the way sound is perceived and processed by the human auditory system [21]. For their extraction, a number of parameters needs to be determined. Following [22], 128 filters are employed, which cover 8 octaves between 44.9 Hz and 11 kHz. Each utterance is described by a vector $\mathbf{x_1} \in \mathbb{R}_+^{F_1 \times 1}$ where $F_1 = 7680$ (i.e., 128 frequency channels $\times$ 10 rates $\times$ 6 scales), where $\mathbb{R}_+$ is the set of non-negative real numbers. BIFs were extracted from each face image following the procedure proposed in [17] for human age estimation. These features are actually a pyramid of Gabor filters and are similar to the way the human visual system processes visual stimulus. Each face image is described by a vector $\mathbf{x_2} \in \mathbb{R}_+^{F_2 \times 1}$, where $F_2 = 13188$, following the parameter values used in [17].

Here, our goal is to exploit the powerful decomposition properties of PARAFAC2 to jointly predict speaker's age interval and gender. A PARAFAC2 model is trained on an irregular fourth-order tensor $\mathcal{X}$ having four slices (i.e., matrices). Let $\mathbf{X}^{(1)} \in \mathbb{R}_+^{F_1 \times I^{tr}}$ be the training speech utterance feature matrix, where F_1 denotes the number of audio features and I^{tr} is the number of training speech utterances. Similarly, let $\mathbf{X}^{(2)} \in \mathbb{R}_+^{F_2 \times I^{tr}}$ be the training face image feature matrix, where F_2 denotes the number of image features and I^{tr} is the number of training face images. Each speech utterance is matched with a face image allowing for a tolerance of ±3 years of its capturing date from speaker's age. To represent the person's age, indicator vectors of dimension L are employed, where L is the number of levels persons' age is quantized to. The age matrix is denoted as $\mathbf{X}^{(3)} \in \mathbb{R}_+^{L \times I^{tr}}$. Its li element $X_{li}^{(3)}$ is 1 if the ith person's age falls into the domain of the lth quantization level and 0 otherwise. For example, let us consider $L = 10$ age intervals. The age intervals are carefully chosen in order to have an adequate (ideally, the same) number of observations in each interval and to cover the entire age range of all speakers in a dataset. Since in our case we have only few utterances of speakers aged less than 25 years old or more than 81 years old, the first age interval represents speakers aged less than 25 and the last interval speakers aged greater than 81. The 2nd to 9th age intervals have a range of 7 years. Let us denote the fourth matrix as $\mathbf{X}^{(4)} \in \mathbb{R}_+^{M \times I^{tr}}$, where M denotes the number of persons. Its mi element $X_{mi}^{(4)}$ is 1 if the ith speech recording is uttered by the mth speaker and likewise ith face image belongs to mth person, too. The persons are grouped according to gender as follows. The first $M_1 = 11$ rows of matrix $\mathbf{X}^{(4)}$ are assigned to female persons, while the remaining $M_2 = 14$ rows are assigned to male persons. Clearly, $M = M_1 + M_2 = 25$ speakers.

Since $\mathcal{X}$ has four slices, the PARAFAC2 seeks a decomposition of the form:

$$\mathbf{X}^{(n)} = \mathbf{U}^{(n)}\,\mathbf{H}\,\mathbf{S}^{(n)}\,\mathbf{W}^{\mathsf{T}}, \quad n = 1, 2, \ldots, 4 \tag{1}$$

where $\mathbf{U}^{(n)} \in \mathbb{R}^{I_n \times k}$, $n = 1, 2, \ldots, 4$ is an orthogonal matrix for each slice, $\mathbf{H} \in \mathbb{R}^{k \times k}$ is a square matrix, $\mathbf{S}^{(n)} \in \mathbb{R}^{k \times k}$ is a diagonal matrix of weights for the nth slice of $\mathcal{X}$, and $\mathbf{W} \in \mathbb{R}^{I^{tr} \times k}$ is a coefficient matrix. Clearly, $I_1 = F_1$, $I_2 = F_2$, $I_3 = L$, and $I_4 = M$. Parameter k denotes the number of latent variables to be extracted from each utterance and face image, respectively. To achieve uniqueness, the square matrix $(\mathbf{U}^{(n)}\mathbf{H})^{\mathsf{T}}\,(\mathbf{U}^{(n)}\mathbf{H})$ is kept constant over n [3]. The decomposition (1) subject to the aforementioned orthogonality constraints for $\mathbf{U}^{(n)}$ can be obtained by solving the optimization problem:

$$\underset{\mathbf{U}^{(n)},\,\mathbf{H},\,\mathbf{S}^{(n)},\,\mathbf{W}}{\operatorname{argmin}} \sum_{n=1}^{4} \|\mathbf{X}^{(n)} - \mathbf{U}^{(n)}\,\mathbf{H}\,\mathbf{S}^{(n)}\,\mathbf{W}^{\mathsf{T}}\|_F^2. \tag{2}$$

The optimization problem (2) can be effectively solved with the algorithm described in [23]. Having solved the optimization problem (2), one computes the matrix $\mathbf{B}_1 \triangleq \mathbf{U}^{(1)}\,\mathbf{H}\,\mathbf{S}^{(1)} \in \mathbb{R}^{F_1 \times k}$. $\mathbf{B}_1$ spans a speech feature space of dimension k, where the semantic relations between the speech feature vectors and their associations with speaker's face image features, age, and gender are retained. Similarly, $\mathbf{B}_2 \triangleq \mathbf{U}^{(2)}\,\mathbf{H}\,\mathbf{S}^{(2)} \in \mathbb{R}^{F_2 \times k}$ spans a face image feature space of reduced dimension k, where the semantic relations between the image feature vectors and their associations with person's speech features, age, and gender are retained. Indeed, the semantic relations between the age vectors as well as the gender vectors are propagated to the feature spaces through the common matrix of right singular vectors $\mathbf{W}$.

Having derived the speech and the face image feature spaces of reduced dimensions (spanned by $\mathbf{B}_1$ and $\mathbf{B}_2$, respectively), we proceed to a validation stage aiming to tune the parameters of an SVM classifier applied to validation sketches, i.e., reduced dimension feature vectors, in order to predict the gender and the age interval. During validation, for each audio feature vector $\mathbf{x}_1^v$, a sketch $\tilde{\mathbf{x}}_1^v$ is derived by pre-multiplying the feature vector $\mathbf{x}_1^v \in \mathbb{R}^{F_1 \times 1}$ with $\mathbf{B}_1^\dagger$, i.e., $\tilde{\mathbf{x}}_1^v = \mathbf{B}_1^\dagger \, \mathbf{x}_1^v \in \mathbb{R}^{k \times 1}$. Similarly, for each face image feature vector $\mathbf{x}_2^v$, another sketch $\tilde{\mathbf{x}}_2^v = \mathbf{B}_2^\dagger \, \mathbf{x}_2^v \in \mathbb{R}^{k \times 1}$ is computed. Needless to say that both $\tilde{\mathbf{x}}_1^v$ and $\tilde{\mathbf{x}}_2^v$ bear information from all slices through the bottleneck model matrix $\mathbf{H}$, which is present in both $\mathbf{B}_1$ and $\mathbf{B}_2$.

Next, ranking vectors for age interval and gender prediction are derived. In particular, the ranking vector for age interval prediction from validation speech sketch $\tilde{\mathbf{x}}_1^v$ is obtained as $\mathbf{a}_1^v = \mathbf{U}^{(3)} \, \mathbf{H} \, \mathbf{S}^{(3)} \, \tilde{\mathbf{x}}_1^v$. Likewise, the ranking vector for age interval prediction from validation face sketch $\tilde{\mathbf{x}}_2^v$ is found as $\mathbf{a}_2^v = \mathbf{U}^{(3)} \; \mathbf{H} \; \mathbf{S}^{(3)} \; \tilde{\mathbf{x}}_2^v$. By concatenating the two age ranking vectors $\mathbf{a}_1^v \in \mathbb{R}_+^{L \times 1}$ and $\mathbf{a}_2^v \in \mathbb{R}_+^{L \times 1}$, the augmented ranking vector $\mathbf{a}^v = [\mathbf{a}_1^{v\top} | \mathbf{a}_2^{v\top}]^\top \in \mathbb{R}_+^{2L \times 1}$ is formed. Let us denote by $\mathbf{A}^v \in \mathbb{R}_+^{2L \times I^v}$ the matrix whose columns are the age ranking vectors for all validation measurements. Subsequently, an SVM employing a linear kernel is trained for age interval prediction. The SVM is fed by the columns of $\mathbf{A}^v$.

A similar procedure is followed for gender prediction. Starting from a ranking vector for gender prediction from validation speech sketch $\tilde{\mathbf{x}}_1^v$ and face sketch $\tilde{\mathbf{x}}_2^v$, i.e., $\mathbf{g}_i^v = \mathbf{U}^{(4)} \, \mathbf{H} \mathbf{S}^{(4)} \, \tilde{\mathbf{x}}_i^v, \;\; i = 1, 2$, the augmented ranking vector $\mathbf{g}^v = [\mathbf{g}_1^{v\top} | \mathbf{g}_2^{v\top}]^\top \in \mathbb{R}_+^{2M \times 1}$ is formed. A second SVM employing a linear kernel is trained for gender prediction. This SVM is applied to the columns of matrix $\mathbf{G}^v$ associated to the ranking vectors of all validation measurements.

During the test phase, first the sketches from a test speech utterance and its associated test face image are computed and then the augmented ranking vectors $\mathbf{a}^{te} = [\mathbf{a}_1^{te\top} | \mathbf{a}_2^{te\top}]^\top$ and $\mathbf{g}^{te} = [\mathbf{g}_1^{te\top} | \mathbf{g}_2^{te\top}]^\top$ are derived. The trained SVM is applied to $\mathbf{a}^{te}$ for age interval prediction. Gender prediction is obtained by the second trained SVM, which is fed by $\mathbf{g}^{te}$.

The proposed PARAFAC2+SVM method performs score level fusion, because the elements of the ranking vectors $\mathbf{a}_1^v$ or $\mathbf{a}_2^v$ if sorted in descending order can be interpreted as follows. The largest element of $\mathbf{a}_1^v$ satisfies $j^* = \operatorname{argmin}_{j=1}^{L} \; \|\mathbf{e}_j - \mathbf{a}_1^v\|_2^2 = \operatorname{argmax}_{j=1}^{L} \; \mathbf{e}_j{}^\top \mathbf{a}_1^v$, where $\mathbf{e}_j$ is an $L \times 1$ indicator vector for age interval prediction. Similarly, the second largest element of $\mathbf{a}_1^v$ is the second best prediction of age interval, and so on. The same procedure applies for gender prediction where $\mathbf{e}_j$ is an $M \times 1$ indicator vector, accordingly. We resort to the SVM to perform multi-biometric score level fusion, when is fed by the aforementioned ranking vectors.

5. Evaluation protocol and metrics

5.1. Machine-based evaluation protocol and metrics

In order to assess the performance of the proposed framework in bimodal age interval and gender prediction, we conducted a first set of experiments on the extended TCDSA dataset, which comprises of $I = 227$ observations. During the evaluation, the Leave-One-Person-Out (LOPO) evaluation protocol was applied. Successively, the observations (speech recordings and face images) of each speaker were included into the test set while the observations belonging to the remaining speakers of the dataset were used for training and validation.

As described before, SVM classifiers were applied to the ranking vectors for age interval and gender prediction derived by PARAFAC2. The prediction of gender constitutes a binary classification problem, while the prediction of age interval was treated as a multi-class classification problem, where the number of age classes equals the considered age intervals. In the TCDSA dataset, the age classes are 10, since we considered 10 different age intervals.

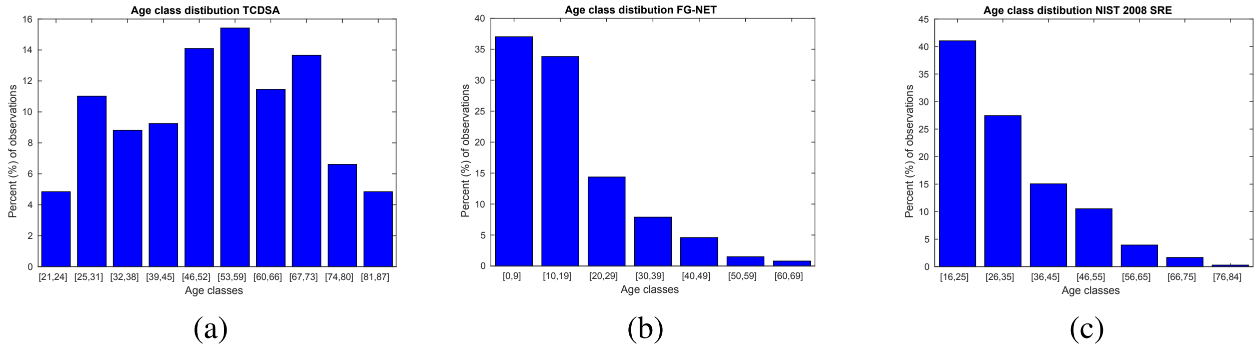
For running SVMs, we used the LIBSVM package [24]. The type of classifier is C-Support Vector Classification with a linear kernel. The best value for parameter C, i.e., the cost parameter of SVM, was selected based on the performance in the validation set. More specifically, in each fold of LOPO, the 20% of the observations that did not serve as test samples were exploited for validation and the remaining 80% composed the train set. In order to achieve a balanced train and validation set in each fold, the observations were assigned to train and validation sets by applying stratified sampling. To this end, at each fold, we examined each age interval separately and the observations belonging to each interval were randomly partitioned by 80% into the train set and 20% into the validation set.

A range of different values for parameter C was examined for both SVMs; the one trained on gender ranking vectors and the one trained on age ranking vectors. The value of parameter C that yielded the best result on the validation set was used for training the SVM and subsequently, predicting the gender and the age interval of the test observations. For age interval prediction, the "one-against-one" approach was followed.

The F_1 measure was employed as metric to assess the predictions made by the proposed method. The F_1 measure is the averaged harmonic mean of precision and recall. Since age prediction is a multi-class classification problem, the F_1 measure is calculated for each age class and micro-averaging is performed to yield a collective figure of merit. Micro-averaging pools per-measurement decisions across classes, and then computes the evaluation metrics on the pooled contingency table.

Furthermore, the performance of the proposed PARAFAC2+SVM framework was compared to the performance of the Random model and SVM classifiers. The Random model gives a sense of the lowest expected value for the metric under consideration on a given dataset. Let us describe the Random model for gender prediction [25]. Apparently, a similar procedure was applied to age interval prediction. The Random model samples the gender class (without replacement) from a multinomial distribution parameterized by the gender prior distribution estimated using the observed gender in the training set [25]. Accordingly, the most frequent gender in the training set is more likely to be chosen for a test observation. Moreover, SVM classifiers with linear kernel were applied to speech features, face image features and concatenated speech and face image features. Clearly, when the baseline SVM classifier is applied to the concatenated speech and face image features, feature level fusion is performed. For the SVMs, a validation stage similar to the one used for the ranking vectors of the proposed method was performed. Furthermore, in an alternative experiment, SVM regression was applied to the ranking vectors derived by PARAFAC2 in order to obtain regression values for age estimation. Here, a linear kernel was employed for the ϵ-Support Vector Regression included in the LIBSVM package [24]. The best values for cost parameter C and parameter ϵ of regression SVMs were selected based on results on validation set.

In order to acquire a clear view of the proposed method performance, we conducted experiments on FG-NET and NIST 2008 SRE datasets, using the LOPO evaluation protocol as well. According to previous studies [17][26][27], the age classes considered for the FG-NET dataset were: [0,9], [10,19], [20,29], [30,39], [40,49], [50,59], [60,69]. The age classes that we considered for the NIST 2008 SRE dataset were: [16,25], [26,35], [36,45], [46,55], [56,65], [66,75], [76,84]. The distribution of observations across age classes for the extended TCDSA dataset, the FG-NET dataset, and the subset of NIST 2008 SRE dataset are depicted in Figure 2.



(a) (b) (c)

Fig. 2 Distribution of observations across age classes for: (a) the extended TCDSA dataset, (b) the FG-NET dataset, and (c) the subset of NIST 2008 SRE dataset used in the experiments.

5.2. Human-based evaluation protocol and metrics

The performance of the proposed method on age interval prediction was compared to human performance on predicting age interval based on speakers' utterances and face images. To this end, an on-line questionnaire[1] was designed that included 25 samples, one for each of the TCDSA speakers. For each sample person, the respondent was asked to determine the age interval based solely on speech information (speech recording of the sample person), based solely on image information (face image of the sample person) and based on a combination of speech and image information (contemporary speech recording and face image of the sample person). The age intervals used in the questionnaire were the same age intervals that were considered in machine-based experiments. Our goal was to investigate the human performance on age interval prediction based on different modalities. Since the questionnaire already consisted of 75 questions specific to age interval prediction and gender prediction based on face images and speech recordings was considered less challenging for humans than age interval prediction, the human performance for gender prediction was not evaluated in the questionnaire for brevity.

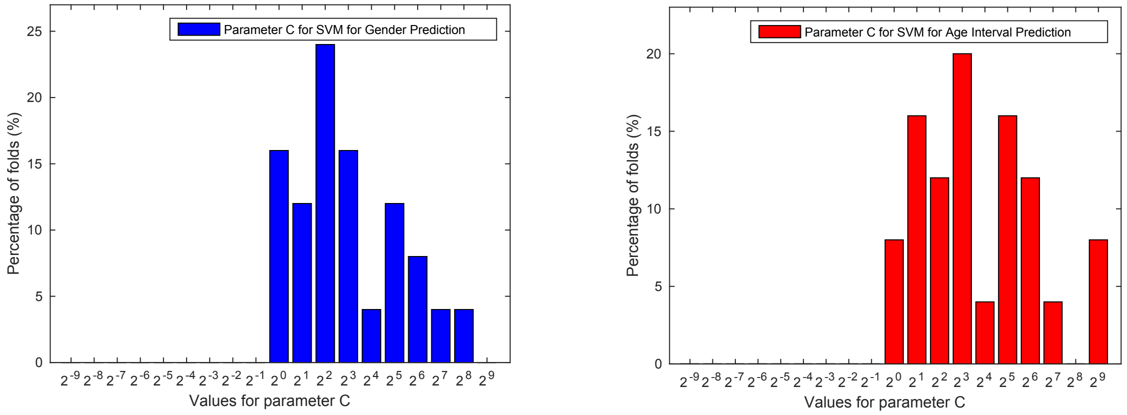
6. Experimental Results

Firstly, PARAFAC2 was applied and yielded ranking vectors for gender and age interval prediction by jointly processing speech utterances and contemporary face images of the extended TCDSA dataset described in Section 3. A number of $k = 10$ latent dimensions were extracted via PARAFAC2 for each of the 4 slices in the TCDSA dataset. The value of k was chosen, so that the orthogonality constraint required for $\mathbf{U}_n,\ n = 1, 2, \ldots, 4$ is satisfied. Following the same restrictions, a number of $k = 7$ latent dimensions was extracted via PARAFAC2 from the FG-NET and the NIST 2008 SRE datasets. Secondly, the augmented ranking vectors $\mathbf{g}^{te}$ and $\mathbf{a}^{te}$ were fed to the dedicated SVM for either gender or age interval prediction using the LOPO protocol detailed in Section 5.1.

LOPO defines $M = 25$ folds in the TCDSA dataset, since the audio-visual TCDSA dataset includes 25 persons. In each fold, a grid searching was performed to determine the value of C that yielded the top F_1 measure for each prediction task in the validation dataset associated to

[1] http://vmclab.polldaddy.com/s/age-estimation-test

the fold. The histograms of the selected values for parameter C during validation across the 25 folds for either gender or age interval prediction are depicted in Figure 3. It is seen that the most frequent top performing value of C for gender prediction was 4. For age interval prediction, the most frequent top performing value of C was 8.



(a)

(b)

Fig. 3 Distribution of the values admitted by parameter C of SVMs applied on the augmented ranking vectors derived by PARAFAC2 during the validation process for: (a) gender prediction and (b) age interval prediction.

The F_1 measure for gender prediction in several experiments is summarized in Table 1. The performance of the bimodal PARAFAC2+SVM framework proposed in Section 4, is listed in the 6th column of Table 1. For comparison purposes, the F_1 measure in four additional experiments is also presented. The name of each experiment is coded as follows. 1) The modality exploited in PARAFAC2 model is denoted by Speech, Image, or Speech+Image. More specifically, speech is exclusively used in the experiments of the 2nd and 8th column. PARAFAC2 was applied to a third-order tensor having 3 slices, namely the speech feature, the age interval, and gender indicator matrices, as in [4]. Similarly, for experiments in the 3rd and 7th column, a unimodal image-based system was created, where face image features were included in the PARAFAC2 model instead of speech features. A bimodal system was considered in the experiments listed in the 4th, 5th and 6th column. Here, a fourth-order irregular tensor with four slices, namely the speech feature, the image feature, the age interval, and the gender indicator matrices was decomposed by PARAFAC2. 2) The modality of the ranking vectors derived by PARAFAC2 and fed to SVMs is indicated by speech rv, image rv, or augmented rv in Table 1. Of course, in the unimodal speech system shown in the 2nd and 8th column, only speech ranking vectors were extracted by PARAFAC2 and the SVM was applied solely on these speech ranking vectors. Similarly, only image ranking vectors were derived by the unimodal image system presented in the 3rd and 7th column of Table 1. In the bimodal speech+image systems presented in the 4th, 5th, and 6th column of Table 1 respectively, ranking vectors of different modalities were utilized. In the 4th column experiment, gender was predicted by an SVM trained only on speech ranking vectors $\mathbf{g}_1^{tr}$ and tested on $\mathbf{g}_1^{te}$ ranking vectors respectively. Similarly, in the 5th column experiment, gender was predicted by an SVM trained only on face image ranking vectors $\mathbf{g}_2^{tr}$ and tested on the $\mathbf{g}_2^{te}$ ranking vectors, respectively. Finally, bimodal gender prediction by applying the method described in Section 4 to the augmented ranking vectors was assessed in the 6th column of Table 1.

The F_1 measure for gender prediction of the SVM classifier and the Random model, which are used as baseline models, is also reported in Table 1. For the speech modality (2nd and 8th column), the SVM classifier was applied to speech features. In the 3rd and 7th column, the SVM was applied to image features instead. In the speech+image modality, an SVM classifier was applied to the concatenated speech and image features. Moreover, Table 1 summarizes the F_1 measure for gender prediction on three datasets, namely the audio-visual TCDSA dataset, the FG-NET dataset, and the subset of NIST 2008 SRE dataset. FG-NET comprises of face images, therefore only the image modality was exploited. On the contrary, only the speech modality was exploited for the NIST 2008 SRE dataset.

Table 1 F_1 measure for gender prediction achieved by the proposed PARAFAC2+SVM method using LOPO protocol on various datasets.

Gender prediction results							
	TCDSA					**FGNET**	**NIST**
	Speech	**Image**	**Speech+Image**			**Image**	**Speech**
Random model	0.4581					0.5	0.5059
PARAFAC2 + SVM	0.9075	0.5066	$0.9207^{\text{speech rv}}$	$0.5154^{\text{image rv}}$	$0.8987^{\text{augmented rv}}$	0.5190	0.8268
SVM	0.9427	0.6123	0.9251			0.6747	0.8691

It is seen from Table 1 that the proposed method clearly outperformed the Random model for gender prediction in all experiments. The proposed PARAFAC2+SVM framework yielded its best performance (4th column) in the TCDSA dataset, when both speech and image modalities were included in the PARAFAC2 model. On the contrary, the baseline SVM classifier yielded its best performance based solely on speech features (2nd column). When each modality was separately exploited, gender prediction based on speech was found to be more accurate than that based on face images (2nd and 3rd column) for both PARAFAC2+SVM and baseline SVM. Additionally, when the PARAFAC2 model included both speech and face image features, the SVM based on speech ranking vectors yielded better results than the SVM based on face image ranking vectors (4th and 5th column). The inclusion of image ranking vectors into the augmented ranking vector yielded a small performance degradation (4th and 6th column). Reading the 3rd and 6th columns of Table 1 from the point of view of face image features, predictions based solely on face image were drastically improved (i.e., increase in F_1 by 0.3921) when speech features and the associated gender ranking vectors were included. Moreover, the results presented in Table 1 indicate that the performance of the proposed bimodal PARAFAC2+SVM method for gender prediction is comparable to that of the bimodal SVM when either augmented or speech ranking vectors are utilized (4nd and 6th column). Furthermore, the proposed method and the baseline SVM classifier demonstrate comparable performance when only speech features are utilized (2nd and 8th column). Nevertheless, the SVMs seem to perform better than the proposed method when the image modality is employed (3rd and 7th column). Some examples of face images of the TCDSA dataset where gender was misclassified by the proposed method are shown in Figure 4.

The figures of merit for age interval prediction are collected in Table 2. In addition, Table 2 includes the F_1 measure values when one age class difference between the predicted and the actual age class of each observation is allowed. That corresponds to a tolerance of 7 years on average allowed to age interval prediction in the TCDSA dataset. In the FG-NET and the NIST 2008 SRE datasets, the average tolerance allowed is 10 years, since 10-year age intervals were considered in these datasets. The results depicted in Table 2 demonstrate a great performance improvement when the aforementioned tolerance on age interval prediction is allowed for the three evaluated methods,

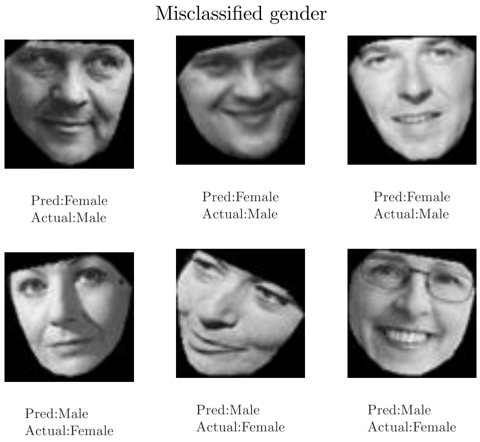


Fig. 4 Face images of the TCDSA dataset where gender was misclassified by PARAFAC2+SVM.

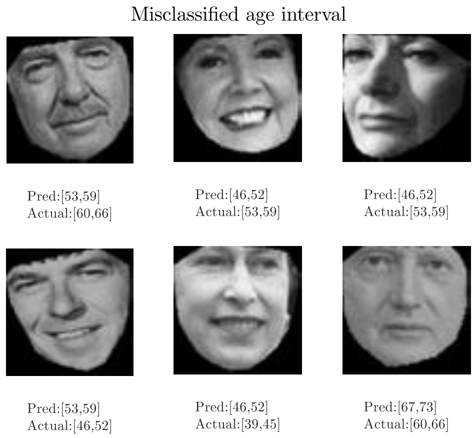


Fig. 5 Face images of the TCDSA dataset where age interval was misclassified by PARAFAC2+SVM.

namely the proposed PARAFAC2+SVM, the Random model, and the SVM classifier. In Table 2, the proposed PARAFAC2+SVM method always outperforms the Random model. Similar to gender prediction, the proposed PARAFAC2+SVM framework yielded its best performance for age interval prediction in the TCDSA dataset, when both modalities were included in the PARAFAC2 model and either speech ranking vectors (4th column) or augmented ranking vectors (6th column) were employed. Contrastingly, the baseline SVM classifier yielded its best performance in the TCDSA dataset based exclusively on speech features (2nd column), which was also observed for gender prediction. When each modality was separately exploited, the age interval prediction based on speech was found to be more accurate than that based on face images (2nd and 3rd column) for both proposed PARAFAC2+SVM and baseline SVM in TCDSA dataset. The same was observed for gender prediction (Table 1), as well. When tolerance in age interval prediction is allowed, the top F_1 measure of 0.4273 for the TCDSA dataset was measured for PARAFAC2+SVM that classifies augmented ranking vectors driven by both sketch speech features and sketch image features. The performance gain against predictions based on speech features exclusively (i.e., 0.0484) and image features exclusively (i.e., 0.0969) is worth noticing. The results obtained by the SVM classifier are similar to the results obtained by the proposed method in all datasets examined. The F_1 measure in the FG-NET and the NIST 2008 SRE datasets are numerically better than the ones admitted in the TCDSA dataset for all evaluated methods, even the Random model. Some examples of face images of the TCDSA dataset where age interval was misclassified by the proposed method are presented in Figure 5.

From the results reported in Tables 1 and 2, it is seen that the performance of both proposed PARAFAC2+SVM and baseline SVM on gender prediction is more solid than that on age interval prediction. Therefore, a further investigation on the age interval prediction was conducted. In order to examine whether the proposed PARAFAC2+SVM method is robust to noise corruption, we conducted the same experiments presented in Table 2 having added street noise to the speech recordings [28]. The experimental findings after the addition of noise to the speech recordings are presented in Table 3. Clearly, the proposed PARAFAC2+SVM framework (6th column) outperformed the baseline SVM on bimodal age interval prediction. As expected, the F_1 measure of PARAFAC2+SVM based solely on speech has been decreased (from 0.1498 to 0.1366), but interestingly, the performance of the bimodal approach has not deteriorated (from 0.1586 to 0.1542)and

Table 2 F_1 measure for age interval prediction achieved by the proposed PARAFAC2+SVM method using LOPO protocol on various datasets. The abbreviation Approx refers to approximate age interval prediction with one age class tolerance.

Age prediction results							
	TCDSA					FG-NET	NIST
	Speech	Image	Speech+Image			Image	Speech
Random model	0.0749					0.2804	0.2904
PARAFAC2 + SVM	0.1498	0.1322	$0.1586^{\text{speech rv}}$	$0.1322^{\text{image rv}}$	$0.1586^{\text{augmented rv}}$	0.4950	0.3917
SVM	0.1762	0.1454	0.1322			0.5529	0.3612
Approximate Age prediction results							
	TCDSA					FG-NET	NIST
	Speech	Image	Speech+Image			Image	Speech
Approx Random	0.2819					0.6407	0.6289
Approx PARAFAC2 + SVM	0.3789	0.3304	$0.3965^{\text{speech rv}}$	$0.3744^{\text{image rv}}$	$0.4273^{\text{augmented rv}}$	0.8283	0.7156
Approx SVM	0.4405	0.3744	0.3921			0.8743	0.7293

Table 3 F_1 measure for age interval prediction achieved by the proposed PARAFAC2+SVM method using LOPO protocol on the TCDSA dataset. Here, street noise has been added to the speech recordings.

Age prediction results after the addition of noise to the speech recordings					
	TCDSA				
	Speech	Image	Speech+Image		
Random model	0.0749				
PARAFAC2 + SVM	0.1366	0.1322	$0.1542^{\text{speech rv}}$	$0.1189^{\text{image rv}}$	$0.1542^{\text{augmented rv}}$
SVM	0.1806	0.1454	0.1410		
Approximate Age prediction results after the addition of noise to the speech recordings					
	TCDSA				
	Speech	Image	Speech+Image		
Approx Random	0.2819				
Approx PARAFAC2 + SVM	0.3436	0.3304	$0.3965^{\text{speech rv}}$	$0.3216^{\text{image rv}}$	$0.4361^{\text{augmented rv}}$
Approx SVM	0.4802	0.3744	0.4009		

has even been improved for approximate age interval prediction (from 0.4273 to 0.4361).

Moreover, performance assessment was conducted with respect to the quality of face images. Here, the quality of the face images of the extended TCDSA dataset was evaluated by 3 persons as "high", "medium", or "low". The quality characterization for each image was based on majority. In total, 78 face images were characterized as of "low" quality, 73 as of "medium" quality, and 76 as of "high" quality. Examples of "low" and "high" quality face images are depicted in Figure 6. The

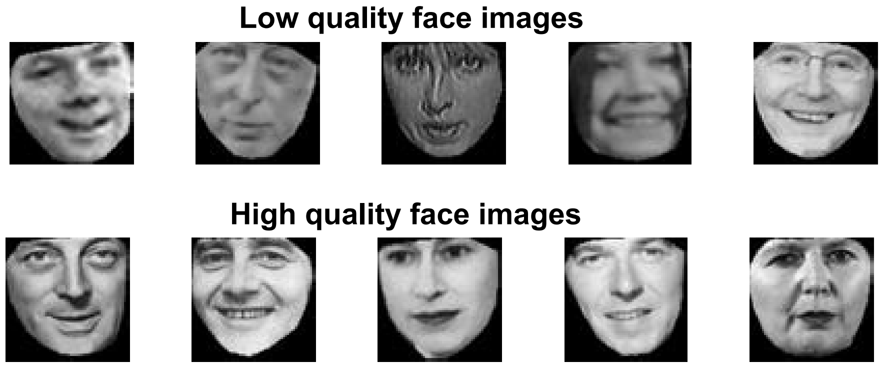


Fig. 6 Examples of "low" and "high" quality images of the extended TCDSA dataset.

Table 4 F_1 measure for age interval prediction achieved by the proposed PARAFAC2+SVM method using LOPO protocol on the TCDSA dataset. Here, the results with respect to the quality of face images in the extended TCDSA dataset are presented.

Age prediction results with respect to face image quality						
		TCDSA				
Image quality		**Speech**	**Image**	**Speech+Image**		
Low	**Random model**	0.2051				
	PARAFAC2 + SVM	0.1282	0.1410	$0.1538^{\text{speech rv}}$	$0.1154^{\text{image rv}}$	$0.1667^{\text{augmented rv}}$
	SVM	0.2308	0.1154	0.1026		
High	**Random model**	0.1053				
	PARAFAC2 + SVM	0.1842	0.1842	$0.1579^{\text{speech rv}}$	$0.1579^{\text{image rv}}$	$0.1711^{\text{augmented rv}}$
	SVM	0.1316	0.2105	0.1711		
Approximate Age prediction results with respect to face image quality						
		TCDSA				
Image quality		**Speech**	**Image**	**Speech+Image**		
Low	**Approx Random model**	0.3462				
	Approx PARAFAC2 + SVM	0.3974	0.2821	$0.4359^{\text{speech rv}}$	$0.3590^{\text{image rv}}$	$0.3974^{\text{augmented rv}}$
	Approx SVM	0.5128	0.2692	0.3718		
High	**Approx Random model**	0.2237				
	Approx PARAFAC2 + SVM	0.3553	0.3684	$0.3289^{\text{speech rv}}$	$0.3816^{\text{image rv}}$	$0.4737^{\text{augmented rv}}$
	Approx SVM	0.3816	0.4474	0.4211		

F_1 measure results for the "low" and "high" quality images are shown in Table 4. It is apparent that both PARAFAC2+SVM and the baseline SVM performed better on "high" quality images (4th column) for age and approximate age interval prediction. Interestingly, PARAFAC2+SVM outperformed SVM on "low" quality images for age prediction (4th to 7th column). Both methods demonstrated a comparable performance for bimodal age interval prediction on "high" quality images. For age interval prediction where one age class tolerance is allowed, the image-based PARAFAC2+SVM outperformed the image-based SVM on "low" quality images (4th column), while the bimodal PARAFAC2+SVM outperformed the bimodal SVM on both "low" and "high" quality face images (7th column).

To test whether the F_1 measure differences between the PARAFAC2+SVM framework and the baseline SVM classifier are statistically significant, we applied the probabilistic approach presented in [29]. To this end, the probability distributions of F_1 measure for the PARAFAC2+SVM and the SVM were inferred, samples were taken from these distributions and their differences were evaluated. Subsequently, based on the observed differences, the probability that the F_1 measure value for PARAFAC2+SVM is higher than the F_1 measure value for SVM was computed. The computed probabilities rely on how many of the correct predictions made by PARAFAC2+SVM were misclassified by SVMs and vice versa. So, the two methods can have similar F_1 measure, but if one method is correct whenever the other makes mistakes, the probability that the first method outperforms the latter is expected to be high. The proposed bimodal PARAFAC2+SVM framework that employs augmented ranking vectors was found to outperform bimodal SVM for age interval and approximate age interval prediction on the TCDSA dataset. Moreover, PARAFAC2+SVM outperformed SVM for age prediction on the NIST 2008 SRE dataset. On the other hand, SVMs appeared to perform better for gender prediction on the FG-NET and the NIST 2008 SRE datasets. Another favourite case, was the use of 4 slices in PARAFAC2 and age interval prediction based on speech ranking vectors in TCDSA (4th column of Table 2). In the experiments, where noise was added to the speech recordings, the proposed bimodal PARAFAC2+SVM framework (6th column of Table 3) outperformed the baseline SVM for age and approximate age interval prediction based on the aforementioned probabilistic approach. Moreover, in the experiments presented in Table 4, image-based PARAFAC2+SVM (4th column) outperformed the SVM on age interval and approximate age interval prediction on “low” quality face images. In addition, the proposed bimodal PARAFAC2+SVM framework with augmented ranking vectors (7th column of Table 4) outperformed the SVM on age interval prediction when “low” quality images are exploited. For approximate age interval prediction, PARAFAC2+SVM admitted a higher probability its F_1 measure outperforms that of the SVM on both “low” and “high” quality images (7th column of Table 4).

Furthermore, in order to compare the performance of the proposed method to that of other age estimation approaches applied to the FG-NET dataset, the Mean Absolute Error (MAE) was also measured as an evaluation metric. MAE is a regression metric and is the average of the absolute errors between the predicted age value and the actual age value. Here, we performed age interval prediction, but, in order to calculate MAE, the predicted age label for each test observation was converted to the mean age of the training observations in each age class. The MAEs for age prediction on the FG-NET dataset that were calculated following the aforementioned procedure are shown in 3rd and 5th column of Table 5 for proposed PARAFAC2+SVM and baseline SVM, respectively. Moreover, in Table 5, the MAEs of different age estimation approaches applied to the FG-NET dataset are, also, presented. Since the PARAFAC2+SVM method was developed for addressing the age-group classification problem rather than the age estimation problem, a direct comparison between the MAE obtained by the proposed method and that attained by other methods is not totally fair, but such a comparison helps to draw a rough assessment of the proposed method’s potential. To facilitate comparison, the MAE results for age estimation when SVM regression was applied to the ranking vectors derived by PARAFAC2 are presented in the 4th column of Table 5. In addition, the MAE results obtained by SVM regression applied to image features are presented in 6th column of Table 5. In each case, a linear kernel was employed by the SVMs and the LOPO protocol was followed. It is seen that both PARAFAC2+SVM and SVM attained their best MAEs when classification was applied.

Table 5 MAE results (years) for age prediction across age classes on the FG-NET dataset.

Age estimation results - FGNET									
Range	**#images**	**PAR+SVM Clas**	**PAR+SVM Reg**	**SVM Clas**	**SVM Reg**	**BIF [17]**	**RUN [26]**	**QM [27]**	**MLP [27]**
0-9	371	2.67	6.27	2.13	5.37	2.99	2.51	6.26	11.63
10-19	339	4.03	3.77	3.82	4.01	3.39	3.76	5.85	3.33
20-29	144	10.32	7.35	8.51	6.40	4.30	6.38	7.10	8.81
30-39	79	20.49	16.71	15.36	13.80	8.24	12.51	11.56	18.46
40-49	46	29.21	25.28	22.66	19.28	14.98	20.09	14.80	27.98
50-59	15	39.10	33.84	26.79	27.47	20.49	28.07	24.27	49.13
60-69	8	48.70	42.42	36.55	36.23	31.62	42.50	37.38	49.13
Total	1002	7.77	7.98	6.25	6.94	4.77	5.78	7.57	10.39

The performance of humans in age prediction was assessed using the questionnaire detailed in Section 5.2. In total, 43 persons answered the questionnaire. 40% of the respondents were female and 60% were male. The age classes considered in the questionnaire were the same as for the proposed method. The micro-averaged F_1 measure for age interval prediction and approximate age interval prediction (with one age class tolerance) are summarized in Table 6. Of course, the questionnaire included questions regarding a subset of the TCDSA dataset, but a rough comparison between the questionnaire results and the results obtained by the proposed method is feasible.

From the results summarized in Table 6 it is seen that human performance in age interval prediction based on images was better than that resorting to either speech utterances or both face images and speech utterances. Moreover, if we compare the human-based performance with the machine-based one for age interval prediction, it is seen that face images were proven more supportive to humans in making predictions. By simultaneous exposition to face images and speech recordings, the F_1 measure improved than that measured when the person listened to only speakers' utterances. In machine-based experiments on TCDSA, large performance discrepancies were not identified across the modalities.

Table 6 F_1 measure for the performance of humans on questionnaire for age and approximate age prediction on the TCDSA dataset.

Human-based age prediction results			
	Speech	**Image**	**Speech+Image**
Age prediction	0.1228	0.2949	0.2595
Approximate age prediction	0.3981	0.6884	0.6651

7. Conclusions

Experimental results demonstrate that using two sources of information gives a clear advantage to the proposed method. The proposed PARAFAC2+SVM framework demonstrated its best performance for both classification tasks when both speech and face image modalities were included in the PARAFAC2 model. Interestingly, the combination of the two modalities of information leads to an improvement on the proposed method performance, while the human-based performance was not increased by employing two modalities. Due to the successful semantically oriented feature reduction performed by PARAFAC2, the performance of PARAFAC2+SVM is comparable to that of the SVM applied to raw high dimensional features. Apparently, the great dimensionality reduction performed via PARAFAC2 speeds the SVM training and testing and leads to compelling profits in computational efficiency (e.g., memory, time). By conducting experiments with noisy speech utterances and face images of low quality, is is demonstrated that the bimodal PARAFAC2+SVM framework deals more efficiently with noise and compensates the low quality in one of the two modalities. The ability of the bimodal PARAFAC2+SVM method to deal effectively with noisy input of one modality is a key finding as in most real life applications involving audio and visual input, one of the two modalities is likely to be corrupted with noise.

Comparing the unimodal systems, the proposed unimodal framework based on speech utterances reported better results than that based on face images. This may be attributed to the large variability of the images' recording conditions (i.e., lighting, pose or appearance). Moreover, the classification task of gender prediction yielded more accurate results than the age interval classification. Ultimately, age prediction can be generally considered as a demanding task since biological age may differ drastically from chronological age. It is also worth noting that the task under study involves age prediction using narrow range intervals.

The work presented in this paper is one of the first attempts ever to combine speech and face images for addressing age interval and gender prediction. A limiting factor for this type of experiments is the availability of suitable datasets. In this particular effort, we attempted to address the problem by augmenting the TCDSA with face images. However, the images used for our experiments were limited in terms of quantity and quality as the limited number of face images retrieved, included sources of variation similar to the ones encountered in the "wild". Despite the aforementioned limitations, the results obtained for age interval and gender prediction show an actual advantage of using combined aural and visual features. Therefore, we believe that the direction of using bimodal methods in age estimation and gender prediction needs to be investigated further. For example, we shall adopt dedicated face image normalization techniques, allowing the standardization of the appearance of faces and explore ways for increasing the samples available in the datasets, so that the training/testing process is enhanced.

8. References

[1] Lanitis, A.: 'A survey of the effects of aging on biometric identity verification', Int. Journal of Biometrics, 2010, **2**, (1), pp. 34–52

[2] Kinnunen, T., Li, H.: 'An overview of text-independent speaker recognition: From features to supervectors', Speech Communication, 2010, **52**, (1), pp. 12–40

[3] Harshman, R. A.: 'PARAFAC2: Mathematical and technical notes', UCLA Working Papers in Phonetics, 1972, **22**, pp. 30–47

[4] Pantraki, E., Kotropoulos, C., Lanitis, A.: 'Age interval and gender prediction using PARAFAC2 applied to speech utterances'. Proc. Int. Workshop Biometrics and Forensics, Limassol, Cyprus, March 2016, pp. 1–6

[5] Polastro, M. C., Eleuterio, P. M. S.: 'Nudetective: A forensic tool to help combat child pornography through automatic nudity detection'. Proc. IEEE Int. Workshop Database and Expert Systems Applications, Bilbao, Spain, August 2010, pp. 349–353

[6] Kelly F., Drygajlo A., Harte N.: 'Speaker verification with long-term ageing data'. Proc. IARP Int. Conf. Biometrics, New Delhi, India, March 2012, pp. 478–483

[7] Panis, G., Lanitis, A., Tsapatsoulis, N., *et al.*: 'Overview of research on facial ageing using the FG-NET ageing database', IET Biometrics, 2016, **5**, (2), pp. 37–46

[8] NIST Multimodal Information Group: 'NIST 2008 Speaker Recognition Evaluation Test Set'. Linguistic Data Consortium, Philadelphia, US, 2011

[9] Kelly F., Harte N.: 'Effects of long-term ageing on speaker verification', Biometrics and ID Management, 2011, **6583** of Lecture Notes in Computer Science, Springer Berlin/Heidelberg, pp. 113–124

[10] Kelly, F., Saeidi, R., Harte, N., *et al.*: 'Effect of long-term ageing on i-vector speaker verification'. Proc. Interspeech, Singapore, September 2014, pp. 86–90

[11] Sadjadi, S. O., Ganapathy, S., Pelecanos, J. W.: 'Speaker age estimation on conversational telephone speech using senone posterior based i-vectors'. Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing, Shanghai, China, March 2016, pp. 5040–5044

[12] Liu, H., Sun, X.: 'A partial least squares based ranker for fast and accurate age estimation'. Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing, Shanghai, China, March 2016, pp. 2792–2796

[13] Geng, X., Zhou, Z. H., Smith-Miles, K.: 'Automatic age estimation based on facial aging patterns', IEEE Trans. Pattern Analysis and Machine Intelligence, 2007, **29**, (12), pp. 2234–2240

[14] Nixon, M. S., Correia, P. L., Nasrollahi, K., *et al.*: 'On soft biometrics', Pattern Recognition Letters, 2015, **68**, pp. 218–230

[15] Arigbabu, O. A., Ahmad, S. M. S., Adnan, W. A. W., *et al.*: 'Recent advances in facial soft biometrics', The Visual Computer, 2015, **31**, (5), pp. 513–525

[16] Liu, L., Liu, J., Cheng, J.: 'Age-group classification of facial images'. Proc. IEEE Int. Conf. Machine Learning and Applications, Boca Raton, FL, US, December 2012, pp. 693–696

[17] Guo, G., Mu, G., Fu, Y., *et al.*: 'Human age estimation using bio-inspired features'. Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition, Miami, FL, US, June 2009, pp. 112–119

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

[19] Levi, G., Hassner, T.: 'Age and gender classification using convolutional neural networks'. Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition Workshops, Boston, MA, US, June 2015, pp. 34–42

[20] Bekhouche, S. E., Ouafi, A., Benlamoudi, A., *et al.*: 'Facial age estimation and gender classification using multi level local phase quantization'. Proc. IEEE Int. Conf. Control, Engineering & Information Technology, Tlemcen, Algeria, May 2015, pp. 1–4

[21] Mesgarani, N., Slaney, M., Shamma, S. A.: 'Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations', IEEE Trans. Audio, Speech, and Language Processing, 2006, **14**, (3), pp. 920–930

[22] Panagakis, Y., Kotropoulos, C. L., Arce, G. R.: ’Music genre classification via joint sparse low-rank representation of audio features’, IEEE/ACM Trans. Audio, Speech, and Language Processing, 2014, **22**, (12), pp. 1905–1917

[23] Chew, P. A., Bader, B. W., Kolda, T. G., *et al.*: ’Cross-language information retrieval using PARAFAC2’. Proc. ACM Int. Conf. Knowledge Discovery and Data Mining, San Jose, CA, US, August 2007, pp. 143–152

[24] Chang, C. C., Lin, C. J.: ’LIBSVM: A library for support vector machines’, ACM Trans. Intelligent Systems and Technology, 2011, **2**, (3), pp. 27:1-27:27.

[25] Turnbull, D., Barrington, L., Torres, D., *et al.*: ’Semantic annotation and retrieval of music and sound effects’, IEEE Trans. Audio, Speech, and Language Processing, 2008, **16**, (2), pp. 467–476

[26] Yan, S., Wang, H., Tang, X., *et al.*: ’Learning auto-structured regressor from uncertain nonnegative labels’. Proc. IEEE Int. Conf. Computer Vision, Rio de Janeiro, Brazil, October 2007, pp. 1–8

[27] Lanitis, A., Draganova, C., Christodoulou, C.: ’Comparing different classifiers for automatic age estimation’, IEEE Trans. Systems, Man, and Cybernetics, Part B (Cybernetics), 2004, **34**, (1), pp. 621–628

[28] Loizou, P. C.: Speech enhancement: Theory and practice (Boca Raton, FL: CRC Press, 2007, 2nd edn. 2013)

[29] Goutte, C., Gaussier, E.: ’A probabilistic interpretation of precision, recall and F-score, with implication for evaluation’. Proc. Eur. Conf. Information Retrieval Research, Santiago de Compostela, Spain, March 2005, pp. 345–359