Face Anthropometry Aware Audio-visual Age Verification

Pavel Korshunov
pavel.korshunov@idiap.ch
Idiap Research Institute
Martigny, Switzerland

Sébastien Marcel*
sebastien.marcel@idiap.ch
Idiap Research Institute
Martigny, Switzerland

ABSTRACT

Protection of minors against destructive content or illegal advertising is an important problem, which is now under increasing societal and legislative pressure. The latest advancements in an automated age verification is a possible solution to this problem. There are however limitations of the current state of the art age verification methods, specifically, the lack of approaches focusing on video-based or even solely audio-based approaches, since the image domain is the one with the majority of publicly available datasets. In this paper, we consider the problem of age verification as a multimodal problem by proposing and evaluating several audio- and image-based models and their combinations. To that end, we annotated a set of publicly available videos with age labels, with a special focus on the children age labels. We also propose a new training strategy based on the adaptive label distribution learning (ALDL), which is driven by facial anthropometry and age-based skin degradation. This adaptive approach demonstrates the best accuracy when evaluated across several test databases.

CCS CONCEPTS

• **General and reference** → **Verification**; • **Computing methodologies** → *Supervised learning by classification*.

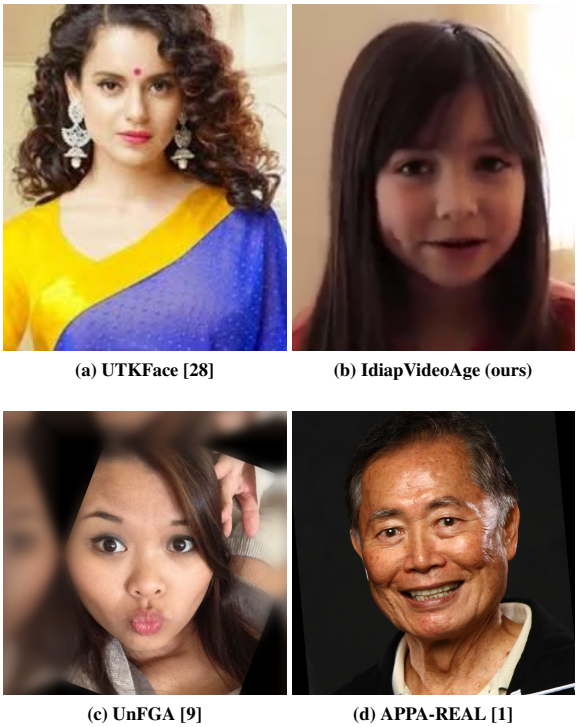
KEYWORDS

Age verification, video database, audio-visual evaluation

ACM Reference Format:
Pavel Korshunov and Sébastien Marcel. 2022. Face Anthropometry Aware Audio-visual Age Verification. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), October 10–14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3503161.3548434

1 INTRODUCTION

The children are consuming online-based content at an increasingly younger age. Multiple instances are reported when children are exposed to unsolicited, destructive, or even illegal content, ranging from an advertisement not taking into account the age of the consumers[1] to online predators pretending being young kids in online games and chat rooms[2]. This unguarded interaction of kids with Internet created a growing pressure by the society, and now, several countries are starting to adopt legislation initiatives enforcing the protection of minors against online harm[3]. The effective enforcement of such protection, however, would require some sort of a mechanism that, preferably, in a non-intrusive and a private manner, can verify someone's age. Automated on-device age verification using personal audio recordings or facial images is one such way of doing that.



(a) UTKFace [28] **(b) IdiapVideoAge (ours)**

(c) UnFGA [9] **(d) APPA-REAL [1]**

Figure 1: Examples of faces from image/video databases.

The main issues with currently available methods for automated age verification include their specific focus on image domain [3, 6, 11, 15, 18, 23, 26], lack of the evaluation on data with children labels, and concerns about privacy and biometric data sharing. The datasets used by the state of the art research, typically, contain only images

*Also with University of Lausanne, Switzerland.

MM '22, October 10–14, 2022, Lisboa, Portugal

ACM ISBN 978-1-4503-9203-7/22/10…$15.00
https://doi.org/10.1145/3503161.3548434

[1] https://www.childrenandscreens.com/findings/advertising-and-marketing-findings/
[2] https://www.nytimes.com/interactive/2019/12/07/us/video-games-child-sex-abuse.html
[3] https://www.gov.uk/government/news/landmark-laws-to-protect-children-and-stop-abuse-online-published

Table 1: Databases with age labels used in the experiments.

Database	Modality	Number of samples	Label type
UTKFace [28]	image	24 361	true age
APPA-REAL [1]	image	7591	true age
UnFGA [9]	image	26 580	apparent age, short intervals
TIDIGITS [14]	audio	25 019	true age
IdiapVideoAge[4] (created by us)	video	4260	apparent precise age

and they fall into one of the following categories: large datasets with true age labels but without kids, such as Morph-II dataset [19] of incarcerated individuals of ages 16 years and older, datasets built by scrapping the web with unreliable age labels, of which IMDB-WIKI [20] is the largest representative, databases with apparent (i.e., appearing as such to a human observer) age labels such as APPA-REAL [1], and the smaller datasets but with precise age labels that span all ages, such as UTKFace [28]. For audio domain, the datasets with age labels are few and far between, with the most notable examples including a very large CommonVoice [2] albeit with very imprecise labels in ten years ranges and TIDIGITS [14], a relatively small set of short audio samples but with reliable age information.

In this paper, we are alleviating the issues related to age verification by first providing a publicly available video database with age labels, and then proposing several image, audio, and video-based methods for age verification and assessing them on this database. We address the concern of data security and privacy by assuming that the age verification methods are run locally on mobile devices without sending biometric data outside of the device. Therefore, we consciously limit the methods we explore in this paper to those that are capable to be run on a mobile device.

We built the video dataset with age labels, referred to as IdiapVideoAge[4], by taking more than 4000 videos from two existing video databases VoxCeleb2 [7] and Child speech dataset from Google[5] and then assigning age labels to the videos by employing human annotators. It is important to note that we have ensured that a large portion (about half) of videos contained kids with ages that are less than 18 years old. We believe that this large number of videos with children age labels makes this dataset unique and very valuable to the research and industrial communities, which are increasingly interested in developing an accurate detection of minors in videos.

Using the proposed IdiapVideoAge[4] video dataset and publicly available UTKFace [28], APPA-REAL [1], UFGA [9], and TIDIGITS [14] datasets (see Table1 for details), we trained and evaluated several state of the art image and audio-based age verification approaches and their multi-modal combinations. For image-based methods, to keep the evaluations comparable, we fixed an underlying model to a relatively simple and small MobileNetV2 architecture [13]. This choice of architecture is justified since we focus on age verification for mobile or VR set scenarios, which are dictated by the current industry interests. Once architecture is fixed, we compare different state of the art regression and classification-based techniques and losses. We also propose our variant of label distribution learning [11, 15] based on face anthropometry, when we use the heuristic based on facial growth rates [17, 25] and age-based skin degradation [10] rates to determine the standard deviation of Gaussian distribution of labels during training. The main aim of this proposed training method, which we coined adaptive label distribution learning (ALDL), is to estimate the ages of children better, while tolerating larger errors in older age categories. The logic is that determining whether a person is below or over 18 years old is critical for many practical applications, while whether the person is 35 or 40 years old is not that important.

For the audio-based age verification, we proposed and evaluated two types of approaches, a simple one based on a combination of handcrafted Mel-frequency cepstral coefficients (MFCCs) and a two layers of long short-term memory (LSTM) layers and an end-to-end approach based on SincNet with four bidirectional LSTMs that was developed for speaker diarization [4] using Pyannote [5]. We evaluated these approaches using TDIGITS and the audio tracks of the proposed IdiapVideoAge datasets. Also, to leverage the multimodal nature of videos from IdiapVideoAge dataset, we have assessed three different ways to combine audio and image-based approaches by i) score fusion when scores from audio and video tracks are combined into one score, ii) features fusion when the embeddings from two architectures are joined and then classified with a support vector machine (SVM), and iii) a joint training when both audio and image-based architectures are joined at the embedding level and trained together with the same loss.

Therefore, the main contributions of this paper are:

(1) A new publicly available dataset of videos with age labels, labeled with human annotators. The dataset contains about 4000 videos with about half containing labels of children.
(2) An extensive evaluation of existing audio- and image-based models and training approaches for age verification. Consideration of several approaches to audio-visual joint modeling.
(3) A new label distribution learning method based on face anthropometry allowing it to be more adaptive to different age labels, especially, to children age categories.

To allow researchers to verify, reproduce, and extend our work, we provide all presented age verification models and code to run and evaluate them as an open-source Python package[6].

2 DATABASES WITH AGE LABELS

In our evaluations, we used three databases with images, one purely audio database, and one video database that we annotated ourselves (see Table 1 for details and some examples in Figure 1). Please note that we did not use a popular Morph-II database [19] of facial shots

[4]https://www.idiap.ch/dataset/idiapvideoage
[5]https://research.google.com/audioset/dataset/child_speech_kid_speaking.html
[6]Source code: https://gitlab.idiap.ch/bob/bob.paper.age_verification

made in prison because it does not contain any images of children (the youngest is 16 years old). We also did not use IMDB-WIKI [20] because, from our visual inspection, the age labels in IMDB-WIKI have very poor accuracy, with about 30% of images being mislabeled, so we deemed this database to be of little use in training our age verification models.

2.1 UTKFace image dataset

UTKFace dataset [28] has one of the most precise true age labels and contains over 20 000 face images with ages from a few months old babies all the way to 116 years olds. All images in the database are labeled by age, gender, and ethnicity. The images cover large variation in pose, facial expression, illumination, occlusion, resolution, etc.

2.2 APPA-REAL image dataset

APPA-REAL database [1] contains 7591 images with associated real and apparent age labels. All images are split into 4113 train, 1500 validation, and 1978 test images. Although the database contains apparent age labels collected via crowdsourcing, in our evaluations, we only use real true ages provided by the authors.

2.3 UnFGA image dataset

The 26 580 of images in UnFGA database [9] were collected from Flickr, which were uploaded using iPhone 5 (or later) smart-phone devices, and released by their authors to the general public under the Creative Commons (CC) license. The data included in the collection is intended to be as true as possible to the challenges of real-world imaging conditions. In particular, it attempts to capture all the variations in appearance, noise, pose, lighting and more, that can be expected of images taken without careful preparation or posing. The ages of faces in the database were labeled manually by the authors and the labels are actually short intervals of three or five years. Therefore, in our evaluations, we have used the midpoints of the label intervals as ground truth age labels.

2.4 TIDIGITS audio dataset

TIDIGITS dataset contains speech which was originally designed and collected at Texas Instruments, Inc. for the purpose of designing and evaluating algorithms for speaker-independent recognition of connected digit sequences. The dataset contains speech from 326 speakers (111 men, 114 women, 50 boys and 51 girls) each pronouncing 77 digit sequences of 1 to 4 seconds long. Each speaker group is partitioned into test and training subsets. The dataset contains prices true age labels of each person.

2.5 IdiapVideoAge video dataset

We built IdiapVideoAge[4], by taking more than 4000 videos from two existing video databases VoxCeleb2 [7] and Child speech dataset from Google[5] and then label the age of people in the videos using three different human annotators. We asked annotators to provide a valid age label if a person's face is visible within more than 80% of the video frames and it is clear that the audible speech matches the person in the video. As the age label, we used the average of the three annotators. We will release it as a publicly available dataset to advance the research in multimodal age verification.

3 AGE VERIFICATION METHODS

We consider three types of age verification methods: i) image-based, ii) audio-based, and iii) video-based. For the image-based methods, we evaluate several state of the art approaches and propose our own variant of 'distribution learning' approach based on facial anthropometry. For the audio-based methods, we propose one simple method that utilizes hand-crafted MFFC features, but which can still perform *on par* with a state of the art method that is based on a speech segmentation [4]. For video-based methods, we consider three well-known techniques with various levels of complexity of combining two audio- and image-based models: i) fuse scores of separately trained models to get one score per a video sample, ii) extract the features/embeddings from the two models and train another simple classifier, such as SVM to produce resulted scores, and iii) joint both models at the embedding level and train them jointly using the same loss.

It is important to note that an age verification can be considered as either a classification or regression problem, which will dictate the way the model will be trained and evaluated. If the goal is to detect an actual age of a person, then regression is a logical choice, although a classification, where each age is a class, is another possible approach. In practical applications, especially, when we focus on detecting ages of children, the classification can be viewed as either a binary problem, e.g., detect whether a person is below or above 18 years old, or a categorical problem when we detect, for instance, whether the person is a child (below 8 years old), of a puberty age (between 8 and 13), an adolescent (between 13 and 18), a young adult (between 18 and 25), an adult (between 25 and 35), of a middle age (between 35 and 50), and a senior (above 50) [8].

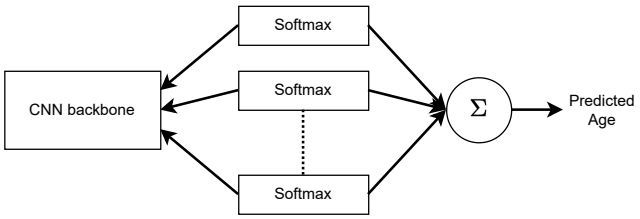


Figure 2: Diagram of regression via classification approach.

3.1 Image-based approaches

As an underlying architecture for image-based age verification approaches, we use MobileNetV2 [13]. MobileNetV2 is definitely not the latest and not the most efficient architecture, but we assume that it needs to be running on a mobile device for privacy and data protection reasons. Moreover, the choice of the underlying architecture is not really that important when we are comparing such techniques as training strategies and losses. The logic is that if one specific technique performs relatively well with MobileNetV2, it will also perform as well with another more efficient architecture.

It is also important to note that in all experiments, we use MobileNetV2 with weights pre-trained for face recognition problem on MS-Celeb database [12] of facial images. We have also tried models pre-trained on a generic ImageNet [21] dataset, but we noticed a consistently better accuracy if the model is pre-trained on faces.

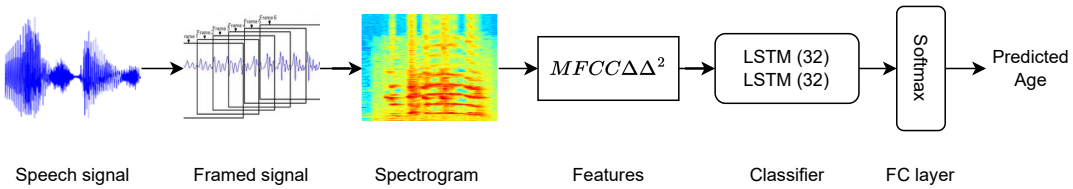


Figure 3: Diagram of MFCCs with LSTM approach for audio-based age classification.

All image-based models we compare share the following parameters and training techniques. Faces in all images (and video frames) are detected by MTCNN [27] and cropped and aligned to 112×112 size. We trained with 32-size batches, using Adam optimizer with learning rate 0.0002, and using such standard image augmentations like randomly changing brightness, hue, contrast, JPEG compression, etc. We trained for 200 epochs with early stopping based on the validation loss.

Here are the different image-based training strategies we evaluate in this paper:

- *classification*: A typical classification training strategy with a cross entropy loss. A fully-connected layer of size equal to the number of classes is added at the top of MobileNetV2 architecture and the full model is trained on faces with age labels.
- *regression*: A typical regression training strategy using mean squared error (MSE) loss and a single node added on top of MobileNetV2 model.
- *rvc*: Regression via classification training strategy (RVC) [3] (see Figure 2 for an illustration) when an age range is split into several sets of classes using sliding window. The network has several heads (fully-connected layers), one for each split. At the inference, the average of the expected values on the outputs from several network heads is taken and is considered to be the predicted age. There are different ways to split the age range and there can be different number of network heads. Following the insights from [3], we used 5 network heads of 17 classes each with the sliding window moving within the range from 1 to 61 years. Please note that the authors of the method [3] trained their models using a subset of the images from UTKFace dataset with ranges from 21 to 60 years old only (to have a more balanced data distribution). Since we used the whole dataset, our experimental results are not directly comparable with the results in [3].
- *distribution*: Distribution based training strategy [11], where instead of using one-hot encoding for true labels, as it is in a typical classification (with a classification layer added at the top of a MobileNetV2 model), a normal distribution with a specified sigma is used. It means that the ground truth label instead of a strict class becomes a distribution with the center at that true label. The model predicts a distribution, so as a loss, Kullback-Leibler divergence to compare predicted and true distributions is used. For inference, an expectation of the predicted distribution is used as the final predicted age. To compute normal distribution for each label during training, sigma 2.0 is recommended by the authors of [11].
- *adaptive*: Our proposed variant of a distribution-based strategy, where the sigma of the normal distribution is not fixed but is dynamic, depending on the true age label. For people younger than 18 years old, sigma is computed using the craniofacial growth rates [17], for middle ages, a more moderate growth rates are used [16], and for more senior ages, the changes in skin texture are taken into account [10]. The actual values of sigma follow a linear approximations of the changes in face of skin corresponding to the age ranges. The logic behind choosing sigma depending on the true age label is for the training to choose a small value of sigma below 1.0 (narrow normal distribution) when the difference between two ages is critical, e.g., for children, and a large sigma larger than 2.0 (wide distribution) for age ranges when faces of people do not change that fast, like for middle ages. For the exact implementation, see our open source package[6].

3.2 Audio-based approaches

We propose to use two different audio-based approaches: a long short-term memory (LSTM) based model that uses Mel-frequency cepstral coefficients (MFCC) features and a end-to-end model adapted from the model created for speaker segmentation task [4] of the diarization problem and implemented with *pyannote* toolkit [5].

For the first model, which we refer to as *simple* and illustrated by Figure 3, we compute MFCC features from a given audio sample by first splitting it into overlapping 20ms-long speech frames with 10ms overlap. The frames are pre-emphasized with 0.97 coefficient and pre-processed by applying Hamming window. MFCC features are obtained from a power spectrum (512-sized FFT) by applying mel-scale filter of size 20. A discrete cosine transform (DCT-II) is applied to the filtered values and first 20 coefficients are taken with their deltas and delta-deltas [24] resulting in 60-long feature vectors.

We use sliding windows of length 25 with overlap 5 of these MFCC feature vectors as input to two stacked LSTM layers (each with 16 hidden units) and fully-connected layer on top as an embedding. The final classification (or regression) layer is added with a cross entropy (or MSE) loss.

For the second model, which we refer to as *pyannote*, we used model pre-trained on DIHARD3 speech corpus [22] created for the diarization problem. The model is proposed in [4] to be used for a segmentation task but we modified it for the age classification task. The architecture of the model remains the same as proposed by the authors and basically consists of a SincNet network which takes audio samples as input and four bi-directional LSTM layers (with 128 units) stacked on top with two additional fully connected layers. Please see [4] for more details.

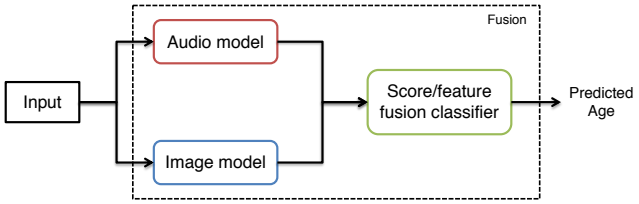


Figure 4: Diagram of score or feature fusion to join audio and image-based age classification models.

Table 2: Evaluation of image-based models on the test set of UTKFace dataset in seven age categories scenario.

	Training Data	Training strategy	f1-score
0	UTKFace	*adaptive*	0.599
1	Several	*rvc* [3]	0.596
2	Several	*adaptive*	0.591
3	Several	*distribution* [11]	0.589
4	UTKFace	*classification*	0.581
5	UTKFace	*rvc* [3]	0.581
6	Several	*classification*	0.559
7	APPA-REAL	*classification*	0.516
8	UnFGA	*classification*	0.483

3.3 The multimodal approach

We evaluated three different ways to combining image and audio modalities (see an example diagram in Figure 4), including the following:

- Score fusion. Train separate image-based and audio-based models (can be trained on different databases) and then use them to compute scores for each video independently. Using a support vector machine (SVM), we trained a simple classifier on the score pairs from both models to combine them in the most efficient way and lead to one score value, representing the output of both modalities.
- Feature fusion. Use the same pretrained models but instead of scores, combine the last layer embeddings and train an SVM classifier to produce the one final score that leverages both modalities.
- Co-training of two modalities. Combine two models by joining their last embedding layers together and co-train them using the same loss function. This joint co-training needs to be done on the same samples, hence only a database with videos can be used.

4 METRICS AND EVALUATION PROTOCOLS

We consider three main scenarios when evaluating age verification methods: i) a *binary* classification, when we evaluate whether the person is less than 18 years old or older, ii) a categorical classification, when we classify a person in one of *seven age categories* such as childhood (below 8 years old), puberty (between 8 and 13), adolescence (between 13 and 18), early adulthood (between 18 and 25), adulthood (between 25 and 35), middle aged (between 35 and 50), and seniority (above 50) [8], and iii) detection of the *exact age* of a person.

To evaluate the accuracy of the age detection, especially in classification scenarios, we use *f1-score*, which is defined as f1-score = $\frac{2(P*R)}{P+R}$, where P precision and R is recall. The *f1-score* allows us to compare two different classifiers in a balanced way. To ensure the balanced f1-score value for data with unbalanced number of different labels (e.g., the number of samples in different age categories can vary a lot), we used a weighted variant of the metric. Also note that the higher the *f1-score* value is the better.

To compare models that predict exact ages, e.g., a regression trained model, it is more common to use mean absolute error (MAE) metric, which measures an average error of the predicted ages from the ground truth values. We also use this metric to compare our results with state of the art methods, which typically focus on predicting exact ages [3, 11, 15]. Note that the lower the MAE value is the better.

5 EXPERIMENTAL RESULTS

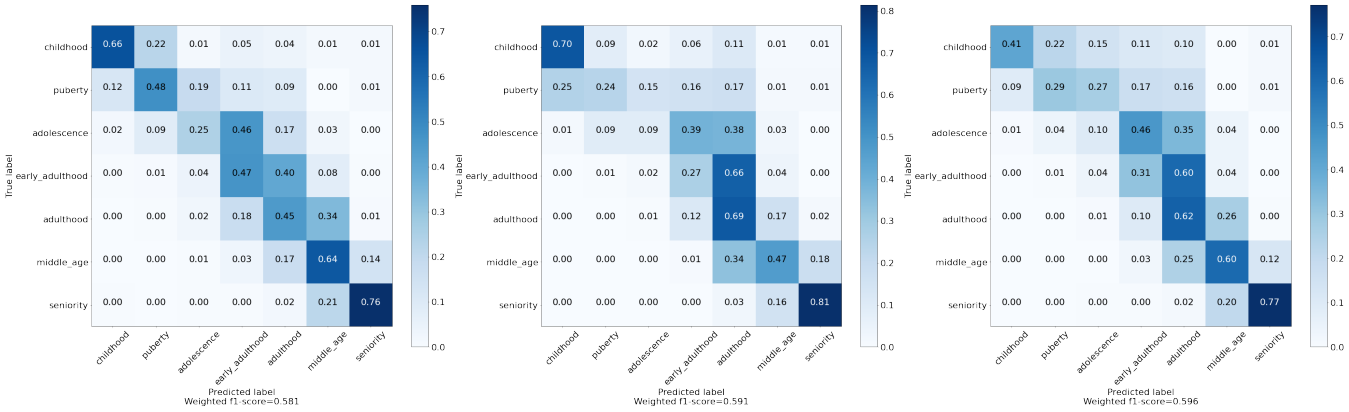
We have implemented all our models using Tensorflow 2.0[7] except for the end-to-end audio-based approach that used pyannote [5], which is implemented using PyTorch[8].

We conducted extensive intra- and inter-database experiments, however, we focus on a selected subset of the experiments in order to save space. For the evaluation code and all of the experimental results, please refer to our open source package[6].

The results for the image-based models (see Section 3.1) for seven age categories scenario (see Section 4) are presented in Table 2. The table shows the overall *f1-scores* computed on the test set of UTKFace database. For training, we either used a training set of UTKFace, a training set of another database (like APPA-REAL or UnFGA), or a combination of training sets from four different databases, denoted as 'Several' in the column 'Training Data' of the table: the three image-based databases plus the video frames from our IdiapVideoAge database. From Table 2, we can note that the proposed 'adaptive' approach that is based on facial anthropometry performs well compared to a simple classifier, regression via classification, or a distribution-based approaches. The top three approaches are very similar in terms of *f1-score* accuracy with differences in third decimals. What is important to note is that training and testing on the same UTKFace database can lead to overfitting and misleadingly better results, as it is demonstrated when we compare results of the simple classifier when it is trained on different databases (all are tested on the same test set of UTKFace). It performs well when trained on the same database (row 4) but its performance degrades significantly when trained on either APPA-REAL or UnFGA databases. Training on a combination of datasets and also using a more advanced training technique will however lead to a considerable improvement in accuracy and generalization capabilities.

To better understand the top performed models from Table 2 and compare them with a simple approach, we have plotted confusion matrices for *classification*, *adaptive*, and *rvc* models in Figure 5. From the plots, we can see that the most challenging age categories are puberty, adolescence, and early-adulthood, where the models are

[7] https://www.tensorflow.org/
[8] https://pytorch.org/



(a) ***classification*** **model trained on UTKFace** **(b)** ***adaptive*** **model trained on Several datasets** **(c)** ***rvc*** **[3] model trained on Several datasets**

Figure 5: Confusion matrices for selected image-based models tested on UTKFace.

Table 3: Evaluation of image-based models on the test set of UTKFace dataset in the exact age scenario.

	Training Data	Training strategy	Mean absolute error (MAE) overall	puberty	childhood	adolescence	early adulthood	above 25 years
0	Several	*adaptive*	5.969	5.707	7.973	7.748	3.970	6.063
1	UTKFace	*adaptive*	5.999	6.976	6.707	6.113	3.577	6.230
2	Several	*distribution* [11]	6.069	5.915	9.613	9.033	5.129	5.933
3	Several	*rvc* [3]	6.195	7.976	7.267	7.596	4.814	6.203
4	UTKFace	*rvc* [3]	6.577	9.780	7.560	7.099	5.390	6.569
5	Several	*regression*	7.666	11.476	10.947	10.179	6.177	7.511

Table 4: Evaluation of audio-based models on the test set of IdiapVideoAge dataset in binary (18-years old) scenario.

	Training Data	Training strategy	f1-score
0	IdiapVideoAge	*pyannote* [4]	0.976
1	IdiapVideoAge	*simple*	0.961
2	TIDIGITS	*pyannote* [4]	0.921
3	TIDIGITS	*simple*	0.801

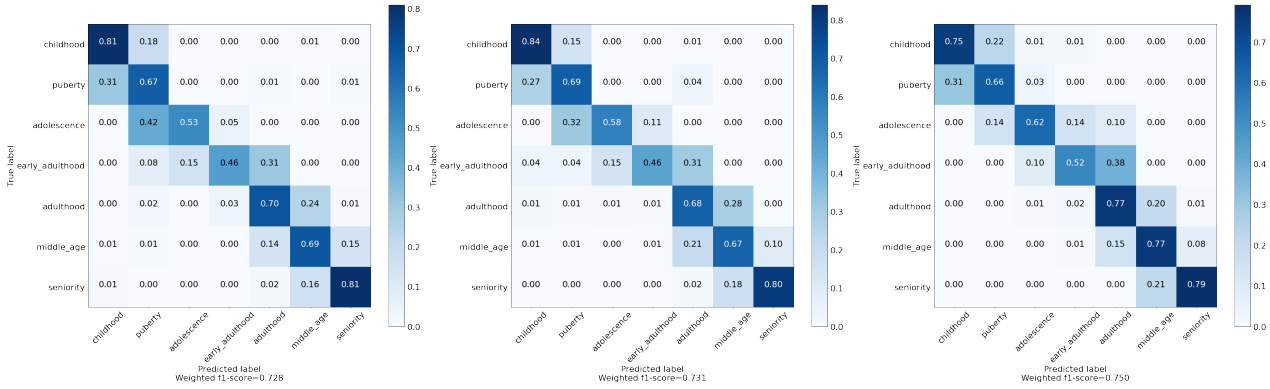
the most 'confused'. Note that the *adaptive* model is significantly more accurate in the childhood age category compared to *rvc* model (with slightly higher overall f1-score), which means the intended focus on detecting children more accurately has worked, however, the lack of detection in the puberty and adolescence means there is still an obvious room for improvement. Indeed, these confusion matrices indicate that age verification is still a difficult problem.

The majority of the state of the art approaches focus on detecting the exact age of a person from the facial image, therefore, we evaluated the image-based models (see Section 3.1) in this scenario as well. The evaluation results are shown in Table 3. The table contains the same models as in the evaluation scenario when we had seven age categories (see Table 2), but a *classification* model is replaced by a *regression* model instead. Also, mean absolute error (MAE) is used in this scenario (since we are actually measuring the error of detecting correct age) instead of *f1-score*. In both scenarios, the same test set of UTKFace database was used as the evaluation set.

From Table 3, we can note that the proposed *adaptive* approach has the smallest overall MAE value. We also computed MEA metric for some of the 'younger' categories to compare the methods in terms of their performance in each of the children categories. This by-category-comparison shows that the *adaptive* approach performs well for the childhood age category (below 8 years old) but the puberty and adolescence ages are the most challenging, which is consistent with the confusion matrices from Figure 5.

To understand how the image-based model perform when used directly on videos, we assessed *adaptive* and *distribution* models on our IdiapVideoAge video database in seven age categories scenario (see top two rows of Table 5). When compared with the results of testing on UTKFace database shown in Table 2, we can notice that IdiapVideoAge database seem less challenging for image-based models, since the overall *f1-score* 0.75 on IdiapVideoAge database is higher than 0.591 obtained on UTKFace dataset for the same *adaptive* model.

Compared to the image-based models, both of the audio-based models (see Section 3.2 for details) perform poorer on IdiapVideoAge dataset as is demonstrated by the lowest *f1-scores* in the two bottom rows of Table 5, with *pyannote* performing considerably better than a *simple* model (based on MFCC features with LSTM networks). The main reason for a relatively poor performance of *simple* audio-based



(a) *classification* model trained on IdiapVideoAge **(b) Audio-visual model co-trained on IdiapVideoAge** **(c) *adaptive* model trained on Several datasets**

Figure 6: Confusion matrices for image-based and audio-visual joint models tested on IdiapVideoAge in seven age categories scenario.

Table 5: Evaluation of image, audio, and joint models on the test set of IdiapVideoAge audio-visual database in seven age categories scenario.

	Training Data	Image model	Audio model	Fusion Type	*f1-score*
0	Several	*adaptive*	none	no fusion	0.750
1	Several	*distribution* [11]	none	no fusion	0.738
2	IdiapVideoAge	*classification*	*simple*	embeddings	0.733
3	IdiapVideoAge	*classification*	*simple*	co-joint training	0.731
4	IdiapVideoAge	*classification*	*pyannote* [4]	score	0.729
5	IdiapVideoAge	*classification*	none	none	0.728
6	IdiapVideoAge	*classification*	*simple*	score	0.711
7	IdiapVideoAge	none	*pyannote*	no fusion	0.614
8	IdiapVideoAge	none	*simple*	no fusion	0.521

model is the fact that it is trained from scratch on the age-labelled data without any pre-training stage. All image-based and *pyannote* models were pre-trained on external large databases, so this fact needs to be considered when judging all of the performances.

If we train and evaluate the audio-based models in a binary scenario, when we are focusing solely on detecting whether a person is below or above 18 years old, the performance becomes more appealing, with accuracy nearing 100%, as illustrated by Table 4. From the table, it is clear that *pyannote* model generalizes the best, since even if it is trained on TIDIGITS database and evaluated on IdiapVideoAge, its *f1-score* is still well above 0.9, which shows that this state of the art model has better learning capacity compared to a simple one.

As the last important question we want to answer is to understand whether the audio and video data is complimentary to each other and whether we can leverage from both modalities to go above their individual accuracies when combining them into a joint approach. To evaluate that, we took one image-based *classification* model, we chose it due to its simplicity compared to other state of the art models, and combined it with both *simple* and *pyannote* audio-based models using score fusion, feature fusion, and co-joint training. All these methods are compared with representatives of an individual modality in Table 5. This table shows that all variants of multimodal approaches lead to better performance compared to single modality methods, as rows 2, 3, and 4, which are multimodal approaches, have higher *f1-scores* compared to a corresponding *classification* image-based model in row 5 and both audio-based model at the bottom. However, Table 5 also demonstrates that a more advanced training techniques like *adaptive* and *distribution* can lead to better performance (the top two rows of the table) without resorting to a much more complex joint models.

To analyze this observation in more details, we plot confusion matrices for three models (see Figure 6): a simple image-based *classification* model, this model combined with a *simple* audio-based model jointly trained on IdiapVideoAge dataset, and the best performing image-based *adaptive* model. The figure shows that the situation is not as simple as it appears, if we only look at Table 5, since we can clearly notice that an addition of audio modality leads to a significant improvement in both childhood, puberty, and adolescence categories compared to a single *classification* model. It is also interesting that a joint audio-visual model is better than a more advance *adaptive* model for childhood and puberty categories, which is a significant observation, if the main focus is on detection of children. A more work and investigation can be done in this direction but the obtained results are nevertheless telling that a co-joint training of an audio-visual model could be a viable practical alternative.

6 CONCLUSION

In this paper, we considered the problem of age verification in image, audio, and multi-domain contexts with the main focus on detecting the age of children. We built a video database with age labels that can be used to train and test both single and multi-domain models. We also proposed an image-based approach based on facial anthropometry that shown to advance the state of the art, especially, when applied to children age categories. This study has also demonstrated that the problem of age verification is very challenging and more research work needs to be done to bring the field to a practically useful state.

ACKNOWLEDGMENTS

This work was funded by InnoSuisse 43595.1 IP-ICT project and Swiss Center for Biometrics Research and Testing.

REFERENCES

[1] Eirikur Agustsson, Radu Timofte, Sergio Escalera, Xavier Baro, Isabelle Guyon, and Rasmus Rothe. 2017. Apparent and Real Age Estimation in Still Images with Deep Residual Regressors on APPA-REAL Database. In *12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*. 87–94. https://doi.org/10.1109/FG.2017.20

[2] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber. 2020. Common Voice: A Massively-Multilingual Speech Corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*. 4211–4215.

[3] Axel Berg, Magnus Oskarsson, and Mark O'Connor. 2020. Deep ordinal regression with label diversity. In *25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2740–2747.

[4] Hervé Bredin and Antoine Laurent. 2021. End-to-end speaker segmentation for overlap-aware resegmentation. In *Proc. Interspeech 2021*. Brno, Czech Republic, 3111–3115.

[5] Hervé Bredin, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and Marie-Philippe Gill. 2020. pyannote.audio: neural building blocks for speaker diarization. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2020)*. Barcelona, Spain.

[6] Wenzhi Cao, Vahid Mirjalili, and Sebastian Raschka. 2020. Rank consistent ordinal regression for neural networks with application to age estimation. *Pattern Recognition Letters* 140 (2020), 325–331. https://doi.org/10.1016/j.patrec.2020.11.008

[7] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. 2018. VoxCeleb2: Deep Speaker Recognition. In *Proc. Interspeech 2018*. 1086–1090. https://doi.org/10.21437/Interspeech.2018-1929

[8] Akhmet Dyussenbayev. 2017. Age Periods Of Human Life. *Advances in Social Sciences Research Journal* 4 (03 2017). https://doi.org/10.14738/assrj.46.2924

[9] Eran Eidinger, Roee Enbar, and Tal Hassner. 2014. Age and Gender Estimation of Unfiltered Faces. *IEEE Transactions on Information Forensics and Security* 9, 12 (2014), 2170–2179. https://doi.org/10.1109/TIFS.2014.2359646

[10] F. Flament, R. Bazin, S. Laquieze, V. Rubert, E. Simonpietri, and B. Piot. 2013. Effect of the sun on visible clinical signs of aging in Caucasian skin. *Clin Cosmet Investig Dermatol* 6 (2013), 221–232.

[11] Bin-Bin Gao, Hong-Yu Zhou, Jianxin Wu, and Xin Geng. 2018. Age Estimation Using Expectation of Label Distribution Learning. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI 2018)* (Stockholm, Sweden). 712–718. https://doi.org/10.5555/3304415.3304517

[12] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. 2016. MS-Celeb-1M: A Dataset and Benchmark for Large-Scale Face Recognition. In *European Conference on Computer Vision (ECCV)*, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (Eds.). Cham, 87–102.

[13] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. 2017. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *ArXiv* abs/1704.04861 (2017).

[14] G. R. Leonard and G. Doddington. 1993. TIDIGITS LDC93S10. In *Linguistic Data Consortium*. Philadelphia.

[15] Yiming Lin, Jie Shen, Yujiang Wang, and Maja Pantic. 2021. FP-Age: Leveraging Face Parsing Attention for Facial Age Estimation in the Wild. *CoRR* abs/2106.11145 (2021). arXiv:2106.11145 https://arxiv.org/abs/2106.11145

[16] H.S. Matthews, Richard L. Palmer, Gareth S. Baynam, Oliver W. Quarrell, Ophir D. Klein, Richard A. Spritz, Raoul C. Hennekam, Susan Walsh, Mark Shriver, Seth M. Weinberg, Benedikt Hallgrimsson, Peter Hammond, A.J. Penington, Hilde Peeters, and P. D. Claes. 2021. Large-scale open-source three-dimensional growth curves for clinical facial assessment and objective description of facial dysmorphism. *Scientific Reports* 11 (2021). Issue 1.

[17] H.S. Matthews, A.J. Penington, R. Hardiman, Y. Fan, J. G. Clement, Kilpatrick N. M., and P. D. Claes. 2018. Modelling 3D craniofacial growth trajectories for population comparison and classification illustrated using sex-differences. *Scientific Reports* 8, 4771 (2018).

[18] Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. 2016. Ordinal Regression with Multiple Output CNN for Age Estimation. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4920–4928. https://doi.org/10.1109/CVPR.2016.532

[19] K. Ricanek and T. Tesafaye. 2006. MORPH: a longitudinal image database of normal adult age-progression. In *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*. 341–345. https://doi.org/10.1109/FGR.2006.78

[20] Rasmus Rothe, Radu Timofte, and Luc Van Gool. 2015. DEX: Deep EXpectation of apparent age from a single image. In *IEEE International Conference on Computer Vision Workshops (ICCVW)*.

[21] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)* 115, 3 (2015), 211–252. https://doi.org/10.1007/s11263-015-0816-y

[22] Neville Ryant, Kenneth Church, Christopher Cieri, Jun Du, Sriram Ganapathy, and Mark Liberman. 2020. Third DIHARD Challenge Evaluation Plan. *arXiv preprint* (2020). https://arxiv.org/abs/2006.05815

[23] Omry Sendik and Yosi Keller. 2019. DeepAge: Deep Learning of face-based age estimation. *Signal Processing: Image Communication* 78 (2019), 368–375. https://doi.org/10.1016/j.image.2019.08.003

[24] F. K. Soong and A. E. Rosenberg. 1988. On the use of instantaneous and transitional spectral information in speaker recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 36, 6 (1988), 871–879.

[25] Sonja Windhager, Philipp Mitteroecker, Ivana Bešlić Rupić, Tomislav Lauc, Ozren Polašek, and Katrin Schaefer. 2019. Facial aging trajectories: A common shape pattern in male and female faces is disrupted after menopause. *American Journal of Physical Anthropology* 169 (2019), 678–688.

[26] Ke Zhang, Na Liu, Xingfang Yuan, Xinyao Guo, Ce Gao, Zhenbing Zhao, and Zhanyu Ma. 2020. Fine-Grained Age Estimation in the Wild With Attention LSTM Networks. *IEEE Transactions on Circuits and Systems for Video Technology* 30, 9 (Sept. 2020), 3140–3152. https://doi.org/10.1109/TCSVT.2019.2936410

[27] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint Face Detection and Alignment using Multi-task Cascaded Convolutional Networks. *CoRR* abs/1604.02878 (2016). arXiv:1604.02878 http://arxiv.org/abs/1604.02878

[28] Zhifei Zhang, Yang Song, and Hairong Qi. 2017. Age Progression/Regression by Conditional Adversarial Autoencoder. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4352–4360. https://doi.org/10.1109/CVPR.2017.463