Apparent Age Estimation from Face Images Combining General and Children-Specialized Deep Learning Models

Grigory Antipov[1,2], Moez Baccouche[1], Sid-Ahmed Berrani[1], Jean-Luc Dugelay[2]
[1]Orange Labs – France Telecom, 4 rue Clos Courtel, 35512 Cesson-Sévigné, France
[2]Eurecom, 450 route des Chappes, 06410 Biot, France
{grigory.antipov,moez.baccouche,sidahmed.berrani}@orange.com, jean-luc.dugelay@eurecom.fr

Abstract

This work describes our solution in the second edition of the ChaLearn LAP competition on Apparent Age Estimation. Starting from a pretrained version of the VGG-16 convolutional neural network for face recognition, we train it on the huge IMDB-Wiki dataset for biological age estimation and then fine-tune it for apparent age estimation using the relatively small competition dataset. We show that the precise age estimation of children is the cornerstone of the competition. Therefore, we integrate a separate "children" VGG-16 network for apparent age estimation of children between 0 and 12 years old in our final solution. The "children" network is fine-tuned from the "general" one. We employ different age encoding strategies for training "general" and "children" networks: the soft one (label distribution encoding) for the "general" network and the strict one (0/1 classification encoding) for the "children" network. Finally, we highlight the importance of the state-of-the-art face detection and face alignment for the final apparent age estimation. Our resulting solution wins the 1st place in the competition significantly outperforming the runner-up.

1. Introduction

Historically being one of the most challenging topics in facial analysis [13], automatic age estimation from face images has numerous practical applications such as demographic statistics collection, customer profiling, search optimization in large databases and assistance of biometrics systems. There are multiple reasons why automatic age estimation is a very challenging task. The most principal among them are an uncontrolled nature of the ageing process, a significant variance among faces in the same age range and a high dependency of ageing traits on a person.

Recently, deep neural networks have significantly boosted many computer vision domains including unconstrained face recognition [26, 19, 24] and facial gender recognition [2]. However, the progress in unconstrained facial age estimation is much slower, due to the difficulty of collecting and labelling large datasets which is essential for training deep networks.

The vast majority of existing age estimation studies deals with the problem of estimation of a person's *biological age* (i.e. objective age defined as the elapsed time since the person's birth date). However, in 2015, the first ChaLearn Looking at People (LAP) competition on *apparent age* estimation (i.e. subjective age estimated from a visual appearance of a person) was conducted [6]. The organizers collected a dataset of face images and developed a web service where people could annotate these images with an apparent age. More than 100 teams have participated in the competition and the 5 best approaches were based on deep Convolutional Neural Networks (CNNs).

In 2016, the second edition of the ChaLearn LAP Apparent Age Estimation (AAE) competition has been organized [7]. We have participated in this competition and have won the 1st place outperforming all other participants by a significant margin. Our final solution is mainly inspired by the solution of the previous year's winners [21]. We improve the approach of [21] by using: (1) a combination of "general" apparent age estimation model with soft age encoding and "children" model with 0/1 age encoding, and (2) precise face alignment prior to age estimation. In this paper, we detail our winning solution in the ChaLearn LAP AAE competition motivating the selected design choices.

The rest of the paper is organized as follows: in Section 2, we present related works on biological and apparent age estimation and existing age encoding strategies, in Section 3, we describe external image datasets which we have used for training in addition to the competition dataset, in Section 4, we detail our data preprocessing and age estimation approaches, in Section 5, we highlight the importance of certain design choices in our solution by experimenting on the validation dataset of the competition, in Section 6, we present the final results of the competition, and we summarize our contributions and conclusions in Section 7.

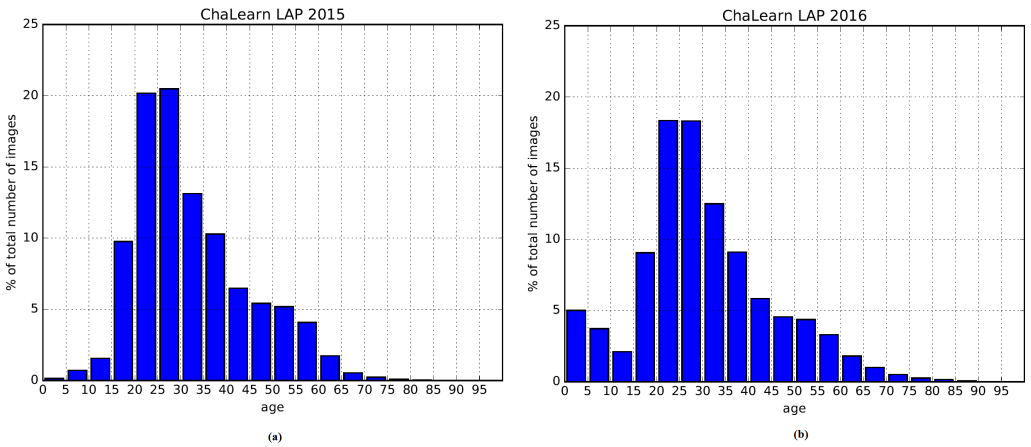

Figure 1. Apparent age distribution in the ChaLearn LAP AAE competition datasets (training+validation): (a) 2015, (b) 2016.

2. Related work

2.1. Biological age estimation

As already mentioned in Section 1, the existing age estimation studies mainly focus on biological age estimation.

There are 2 publicly available datasets which are mostly used in the context of biological age estimation: FG-NET dataset [1] and MORPH-II dataset [20]. FG-NET dataset contains about 1000 images obtained mainly from scanning old photos. MORPH-II dataset is bigger than FG-NET containing about $55,000$ images. This dataset was collected by American law enforcement services.

The most used metric for evaluating systems of automatic estimation of a biological age is Mean Absolute Error (MAE). MAE is simply defined as a mean value of absolute differences between predicted ages $\hat{x}$ and real (biological) ages x: $MAE = \frac{1}{N}\sum_{i=1}^{N}|\hat{x_i} - x|$.

The problem of biological age estimation has been studied for a long time. The very first works (notably [15] (1999)) focused mainly on cranio-facial development theory using geometrical ratios between different face regions to identify a person's biological age. Age estimation was treated as a classification problem with coarse classes (babies, young adults, adults and seniors). Later studies approached biological age estimation from face images in a conventional computer vision manner: designing of feature representations for input images and training regression functions or classifiers on the obtained representations. In that context, feature designing to describe the ageing pattern proved to be of particular importance. For example, in 2007, [9] proposed to model the ageing pattern defined as the sequence of a particular individual's face images sorted in time order by constructing a representative subspace: AGES (AGeing pattErn Subspace). Authors obtained MAEs of 6.8 and 8.8 on FG-NET and MORPH-II, respectively. While in 2009, [12] investigated a possibility of applying Biologically Inspired Features (BIF) for age estimation. Authors proposed the "STD" operator for encoding the ageing subtlety on faces. They obtained the MAE of 4.8 on FG-NET dataset. This result was further improved by [10] in 2011 who proposed to combine BIF with Kernel Partial Least Square regression (KPLS) and reached the MAE of 4.2 on FG-NET dataset.

Finally, the recent development of deep learning methods (where feature designing and age estimation stages are combined into one neural model) has allowed to further improve automatic age estimation quality. Thus, [29] (2014) is one of the first works to apply CNNs for age estimation. Authors employed several shallow multiscale CNNs on different face regions and obtained the MAE of 3.6 on MORPH-II dataset. The most recent work of [28] (2015) is also based on CNNs. Authors proposed using a ranking encoding for age and gender and reported the state-of-the-art MAE of 3.5 on MORPH-II dataset.

2.2. Apparent age estimation

Despite being strongly correlated with each other, an apparent age of a person can be very different from her (his) biological age [6]. The first edition of the ChaLearn LAP AAE competition [6] boosted the research in apparent age estimation by making public the first dataset with apparent age annotations of 4691 images. In the second edition of the competition [7], this dataset has been extended to 7591 images (4113 images for training, 1500 for validation and 1978 for test). Not only the number of images has increased, but also the age distribution has changed with respect to the first edition of the competition (see Figure 1). In particular, the percentage of children images has significantly increased in the second edition of the competition.

Each image of the competition dataset is annotated with a mean age μ and a corresponding standard deviation σ (these statistics are calculated based on at least 10 human votes per image). The metric which has been selected by the competition organizers to evaluate apparent age estimation systems is quite different from MAE which is used for biological age estimation. The competition metric ϵ is defined as the size of the tail of the normal distribution with the mean μ and the standard deviation σ with respect to the predicted value $\hat{x}$: $\epsilon = 1 - e^{-\frac{(\hat{x}-\mu)^2}{2\sigma^2}}$. Therefore, the apparent age estimation errors on examples with a small standard deviation (i.e. on examples on which human votes are close to each other) are penalized stronger than the same errors on examples with a high standard deviation (i.e. on examples on which human votes disagree between each other).

Below, we present 3 winning entries of the first edition of the competition. All of them are based on CNNs.

[21] are the winners of the first edition of the competition. Their approach is based on pretraining of the VGG-16 CNN [23] on the ImageNet dataset [22], training this network on the IMDB-Wiki dataset for the biological age estimation task (this dataset has been collected and made public by the authors) and, finally, fine-tuning for the apparent age estimation task on the competition data. Authors trained their CNN for a classification with 101 classes (ages between 0 and 100 years old) and used the expected value of 101 neurons as an age estimation at the test phase. The resulting ϵ is 0.2650. [16] are the runners-up of the competition. Authors used the GoogLeNet CNN [25] as their basic model. Authors pretrained the GoogLeNet CNN for face recognition task on the CASIA WebFace dataset [30], then the CNN was trained on CACD [4], WebFaceAge [18] and Morph-II datasets for the biological age estimation task, and finally, the CNN was fine-tuned on the competition data for the apparent age estimation task. Authors combined CNNs trained for age regression and for age classification with distributed labelling. As a result, they obtained ϵ of 0.2707. The third result in the competition was achieved by [32]. Their approach is very similar to the one by [30]: also pretraining of the GoogLeNet CNN on the CASIA WebFace dataset, training for biological age estimation on publicly available age datasets and the final fine-tuning for apparent age estimation on the competition data. However, the particularity of the solution by [32] is the usage of the cascade approach for age classification: firstly, a coarse classification in one of 10 age groups and then a fine-grained intra-group regression. The final result of [32] is ϵ of 0.2948.

Summarizing the approaches of the 3 winners of the first edition of the ChaLearn LAP AAE competition, the following common strategies can be highlighted:

1. All 3 winners use deep CNN architectures (either VGG-16 or GoogLeNet) pretrained on large image datasets (either ImageNet or CASIA-WebFace).
2. All 3 winners employ the same pipeline for training their CNN: firstly, training on large datasets for biological age estimation and secondly, fine-tuning on the competition dataset for apparent age estimation.

Relying on the success of these 2 strategies in the first edition of the competition, we also follow them in our solution in the second edition of the competition.

2.3. Age labels encoding

In literature, there are 3 commonly used age labels encodings for automatic age estimation systems. These encodings are presented below:

1. **Real number encoding**. This is a pure regression approach. In real number encoding, the age labels are encoded just as real numbers.
2. **0/1 classification encoding**. This is a pure classification approach. In 0/1 classification encoding, we predefine a certain number of classes (for example, 100 classes for ages between 0 and 99 years old) and the age labels are encoded as binary vectors containing a single non-zero value corresponding to the class to which a certain example belongs to.
3. **Label distribution encoding**. Label distribution encoding [8] can be seen as the soft version of 0/1 classification encoding. In label distribution encoding, on the one hand, we predefine a certain number of classes (as in case of 0/1 classification encoding) but on the other hand, the age labels are encoded not with binary vectors but with real-valued vectors representing the probability distributions of belonging to corresponding classes. More precisely, assuming that we encode an age $x \in \mathbb{R}$ with a label vector L of length N (N classes), the label vector L will be defined as follows: $L_i = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(i-x)^2}{2\sigma^2}}; i = 1, \ldots, N$, where σ is a predefined parameter. In other words, in order to encode an age x, we fit a normal distribution with an expected value of x and a standard deviation of σ. The advantage of label distribution encoding with respect to 0/1 classification encoding is the fact that apart from storing the information to which class a certain example belongs to, a label vector also stores the information about the neighbouring classes (i.e. neighbouring ages). This additional information can be useful during training. In particular, label distribution encoding provides a machine learning model with the information that, for example, it is better to predict 20 years old instead of 21 years old, than 100 years old instead of 21 years old. This information is missing in 0/1 classification encoding. Finally, it is worth noting that

0/1 classification encoding is an extreme case of label distribution encoding when $\sigma \to 0$.

3. External data

In this section, we present the datasets which we have used for biological age estimation training in our work.

IMDB-Wiki Inspired by the success of the 1st place winners of the first edition of the ChaLearn LAP AAE competition [21], we have decided to use the IMDB-Wiki dataset collected and used by them for the biological age estimation training. Authors made this dataset public in 2016.

The dataset consists of $523,051$ images collected from 2 sources: IMDb[1] ($460,723$ images) and Wikipedia[2] ($62,328$ images). The distribution of ages in the IMDB-Wiki dataset is presented in Figure 2.

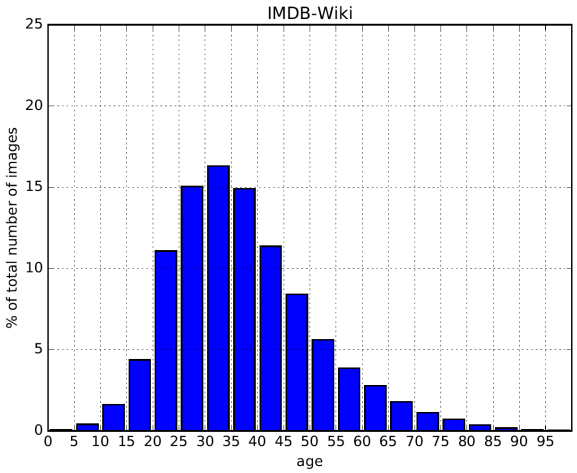

Figure 2. Biological age distribution in the IMDB-Wiki dataset.

Due to the fact that each image contains a celebrity (whose identity, gender and birth date are known) and a timestamp, authors managed to automatically annotate all images in the IMDB-Wiki dataset with biological ages.

However, for the majority of images from the IMDB-Wiki dataset, the provided annotations are not directly usable. The problem comes from the fact that a lot of images contain more than one person. Assuming that all faces in the image are detected automatically, it is not obvious how to automatically select a face to which the given annotation corresponds to. To circumvent this problem, we have pursued the 2 following approaches:

1. We have used those images for which the "Head Hunter" face detector [17] has detected only one face (a similar approach was employed by [21]). In this case, we can be sure that the detected face corresponds to the provided age annotation. This approach has resulted in $182,019$ images.

2. We have developed a simple web interface for the manual annotation of the remaining images. Given an input image and a corresponding annotation (the person identity, gender and age), a user has to simply select a face in the image to which the given annotation corresponds to. By crowdsourcing the annotation process via the described interface, we have managed to annotate $68,548$ images (26 persons participated in the annotation campaign which lasted for 4 days).

Thus, in total, $250,367$ images from the IMDB-Wiki dataset have been used in our experiments. In order to avoid ambiguity with the whole IMDB-Wiki dataset, below, we refer to this subset of $250,367$ images of the IMDB-Wiki dataset as the "cleaned" IMDB-Wiki dataset.

Collected dataset with images of children As it is seen in Figure 2, there are very few images of children younger than teenage (i.e. 12 years old and younger) in the IMDB-Wiki dataset. Therefore, an age recognition model which is trained on this dataset is likely to perform poorly for age estimation of children. This was not a major problem in the first edition of the ChaLearn LAP AAE competition given that there were very few children in the competition dataset (see Figure 1(a)). However, this problem becomes very important in the second edition of the competition where children occupy almost 10% of all images (see Figure 1(b)).

It should also be noticed that according to the competition dataset annotations, the average standard deviation of human votes for images of children (between 0 and 12 years old) is about 1, while the average standard deviation for all other images is about 5. Thus, according to the competition data, humans estimate an age of a child almost 5 times more precisely than an age of an adult. As it is mentioned in Section 2.2, the competition metric ϵ is defined in the way that the same absolute error in age estimation is penalized more for images with small standard deviation of human votes.

The above observation shows the importance of predicting ages for children images with a very high precision and the need of training children images with precise biological age annotations. Therefore, we have manually collected a private dataset of 5723 children images in the 0-12 age category using the Internet search engines.

4. Proposed solution

ChaLearn LAP AAE competition is an "end-to-end" competition meaning that given at input raw real-life images (from Wikipedia, social networks etc.), participants have to output corresponding apparent age estimations.

[1]The Internet Movie Database: `www.imdb.com`
[2]The free Internet encyclopaedia: `www.wikipedia.org`

Required image preprocessing (e.g. face detection and face alignment) is considered as a part of the challenge. Therefore, our solution is split into 2 logical steps: image preprocessing and apparent age estimation itself. In this section, we present the mentioned steps one by one.

4.1. Image preprocessing

Face detection We have used the open source "Head Hunter" face detector [17]. In particularly, we have employed the fast implementation by [19]. In order to detect faces regardless of an image orientation, we rotate each input image at all angles in the range [-90°, 90°] with the step of 5°. We then select the rotated version of the input image which gives the strongest output of the face detector for the face alignment step. If no face is detected in all rotated versions of the input image, the initial image is upscaled and the presented algorithm is repeated until a face is detected. 2 upscaling operations has been enough to detect at least one face in all images of the competition dataset. As recommended in [21] (and also confirmed by our own experiments), we extend the face area detected by the "Head Hunter" face detector and take 40% of its width to the left and to the right and 40% of its height above and below.

Face alignment We have integrated the state-of-the-art face alignment solution by [27] in our image preprocessing pipeline. The solution of [27] is based on the multi-view facial landmark detection. There are 5 landmark detection models: a frontal model, 2 profile models and 2 half-profile models. Each of these models is tuned to work on one of the corresponding facial poses. The face alignment follows the face detection and requires running of all 5 landmark models on the detected face. Each model reports a confidence score which shows how well the corresponding landmarks are detected in the given face. We then select the model with the highest confidence score and perform an affine transformation from the detected landmarks to the predefined optimal positions of these landmarks with respect to the detected facial pose.

We have also tried to use an older commercial solution for face detection and face alignment which is based on [31] and [3] respectively. Our experiments presented in Section 5 compare the 2 approaches and clearly demonstrate the merits of the open-source state-of-the-art solutions.

4.2. Apparent age estimation

Following the winning solution from the previous edition of the ChaLearn LAP AAE competition [21], we also employ the 2-steps strategy of CNN-training for apparent age estimation: firstly, we train our CNNs for biological age estimation on external datasets, and secondly, we fine-tune them for apparent age estimation on the competition data.

However, there are several key novelties in our approach with respect to the approach of [21]. We highlight these novelties below:

1. As it is mentioned in Section 3, the precision of the apparent age estimation on children images has a very high influence on the final score in the second edition of the ChaLearn LAP AAE competition. Therefore, we have trained a separate model for estimating apparent ages of children (0-12 years old) using the external data described in Section 3. The gain of integrating this separate CNN in the final solution is quantitatively evaluated in Section 5.

2. We combine 2 age labels encoding strategies which are presented in Section 2.3. On the one hand, we employ a label distribution age encoding for training the "general" CNNs which allows our neural networks to better capture the concept of an apparent age (which is rather a range of values than a precise real value). On the other hand, we employ a 0/1 classification encoding for the "children" CNNs because for children, a possible range of apparent age values is very narrow and, therefore, it is meaningful to encode each year as a completely separate class.[3] Our experiments have shown that using this combined age labels encoding strategy is advantageous with respect to using only distributed age encoding or only 0/1 classification encoding for both "general" and "children" CNNs.

3. Our experiments in Section 5 demonstrate that the quality of image preprocessing has a very strong impact on the final ϵ-score. Therefore, we employ the state-of-the-art open source solution from [27] for face alignment in our final approach.

4.2.1 Training pipeline

The integral training pipeline of all apparent age estimation CNNs is presented in Figure 3. Starting with the pretrained VGG-16 CNN from [19], we train a "general" CNN for biological age estimation of all ages between 0 and 99 years old on the "cleaned" IMDB-Wiki dataset using the label distribution age encoding. From the obtained network, we fine-tune a "children" CNN for biological age estimation of children between 0 and 12 years old. This time, the 0/1 classification age encoding is used. The next step is fine-tuning of 2 resulting CNNs (the "general" one and the "children" one) for apparent age estimation. In case of the "general" CNN, we combine all training and validation images from the competition dataset (5613 images in total)

[3] Here and below, we refer to the CNNs which estimate all ages between 0 and 99 years old as the "general" ones, while to the CNNs which estimate only ages of children between 0 and 12 years old as the "children" ones.

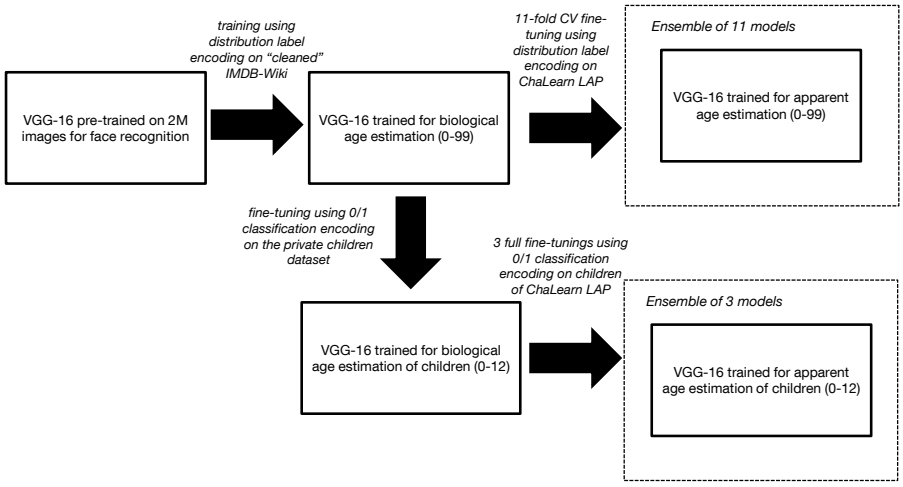

Figure 3. Training pipeline.

and fine-tune 11 "general" CNNs for apparent age estimation using 11-fold cross-validation where the size of each of 11 training datasets is 5113 images and the size of each of 11 non-overlapping validation datasets is 500 images. In case of the "children" CNNs, we combine all images of children between 0 and 12 years old from the training and validation parts of the competition dataset (there are 543 of them). Due to the small number of available images, we fine-tune the "children" CNNs for apparent age estimation without any validation saving the CNN weights at 3 predefined points which have been chosen by experimenting on the validation dataset[4]. As a result, we obtain 3 "children" CNNs for apparent age estimation.

4.2.2 Testing pipeline

The pipeline of our final solution at test stage is presented in Figure 4. An input image is firstly processed by a face detector which defines a face box and rotates the image accordingly. Then the detected face is aligned and the resulting image is resized to 224x224 pixels (the size of an input to the VGG-16 CNN). From the obtained image, we generate its 7 modified versions: the mirrored one, the ones rotated at $\pm 5^\circ$, the ones shifted by 5% on the left/right and the ones scaled in/out by 5%. This is done in order to compensate a negative impact from minor face alignment errors (which are inevitable given the difficulty of the competition dataset). In total, there are 8 images including the original one. All these images are processed by 11 "general" CNNs. We take the values of 100 output neurons after each of $88 (= 8 * 11)$ CNN forward passes, average them and normalize them to sum up to 1. Thus, we obtain a vector p of 100 values representing probabilities of belonging to ages between 0 and 99 years old. The final "general" age prediction is calculated as an expected value of these probabilities: $general\_age = \sum_{i=0}^{99} i * p_i$. If the predicted "general" age is superior to 12, it is considered as the final apparent age estimation and the algorithm stops. In the opposite case, we process the same 8 images as before by 3 "children" CNNs. We take the values of 13 output neurons after each of $24 (= 8 * 3)$ CNN forward passes, average them and normalize them to sum up to 1. Thus, we obtain a vector p of 13 values representing probabilities of belonging to ages between 0 and 12 years old. The final "children" age prediction is calculated as an expected value of these probabilities: $children\_age = \sum_{i=0}^{12} i * p_i$. The predicted "children" age is considered as the final apparent age estimation.

Running the final age estimation system on all 1987 test images takes about 3.5 hours.

4.3. Training details

In this work, on multiple occasions, we initialize the weights of a CNN A with the weights of the trained CNN B, while CNNs A and B have different output layers. In all such cases, we initialize all layers but the output (last) one of the CNN A with the corresponding layers of the CNN B, while the output layer of the CNN A is initialized randomly.

"General" and "children" VGG-16 CNNs have identical architectures (the one from [23] without the softmax layer of 1000 neurons) with the exception of the output fully-connected layer. In "general" CNNs, the output layer

[4]We do not guarantee that 3 is an optimal number of "children" networks. Due to time constraints, we have not tested an ensemble of more than 3 "children" networks.

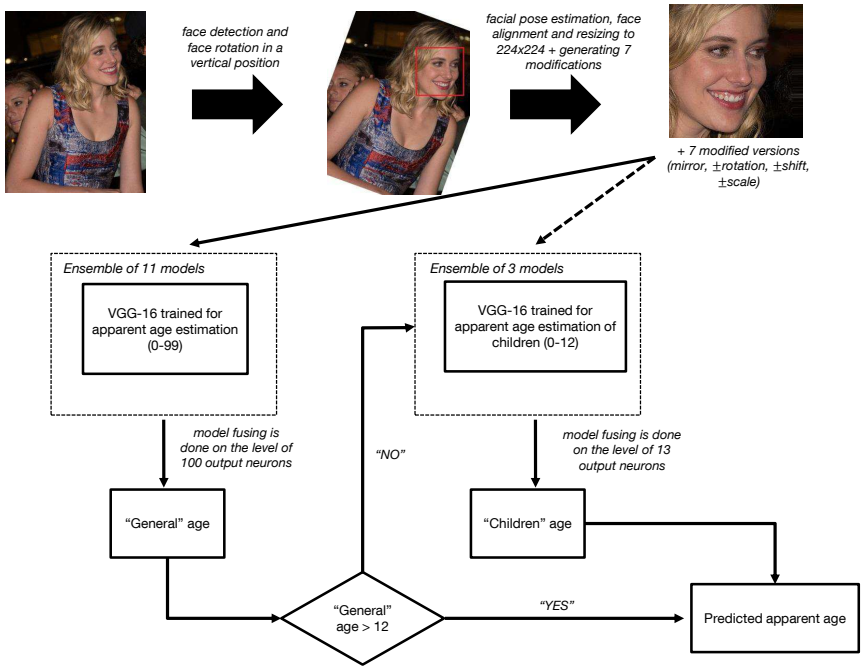

Figure 4. Testing pipeline.

CNN	Learning rate	# training iterations	Training time
"general" for biological age	10^{-3}	$5 * 10^4$	≈ 1.5 days
"children" for biological age	10^{-4}	$3 * 10^3$	≈ 2 hours
"general" for apparent age	10^{-2}	$7.5 * 10^3$	≈ 5 hours
"children" for apparent age	10^{-3}	10^3	≈ 40 minutes

Table 1. Training details.

contains 100 neurons corresponding to ages between 0 and 99 years old and sigmoid activations, while in "children" CNNs, the output layer contains 13 neurons corresponding to ages between 0 and 12 years old and a global softmax activation. All CNNs are optimized by the gradient descent with momentum of 0.9 using the mini-batches of 32 images (other optimization details are given in Table 1). All CNNs in this work have been trained using Caffe deep learning framework [14] on the Tesla K40c GPU.

We use the 5-times data augmentation when fine-tuning "general" and "children" CNNs for apparent age estimation on the competition data. Apart from the original images, we use their mirrored versions, randomly rotated versions (the absolute rotation angle is no more than $5°$), randomly shifted versions (the absolute shift length is no more than 5% of the image size) and randomly scaled versions (the scaled size is between 95% and 105% of the original size).

5. Experiments

In this section, we present the results of our experiments on the validation dataset of the competition illustrating the impacts of certain design choices from Section 4 on the apparent age estimation quality.

The experimental results on the validation dataset of the competition are regrouped in Table 2.

In the first line of Table 2, we present the ϵ-score of the model which has been trained for biological age estimation on the "cleaned" IMDB-Wiki dataset. This score (0.3927) is to be compared with the line 3 score (0.2986) which represents the performance of our model fine-tuned on the training dataset of the competition for the apparent age estimation task. The large gap of almost 0.1 of ϵ-score (i.e. 24%) between these 2 results clearly demonstrates the difference between apparent and biological age estimations as well as the importance of fine-tuning on the competition data.

Lines 2 and 3 of Table 2 highlight the impact of the quality of face detection and alignment on the competition results. Using the state-of-the-art open-source face detection and face alignment solutions by [17] and [27] respectively has allowed us to gain 0.01 of ϵ-score (i.e. 3%) with respect to the older commercial solution based on [31] and [3].

The data augmentation during the fine-tuning for appar-

Biological age training	Apparent age fine-tuning	Image preprocessing (face detection + face alignment)	Data augmentation during apparent age fine-tuning	Data augmentation during testing	Children model	ϵ-score
Yes	No	**[17] + [27]**	No	No	No	0.3927
Yes	**Yes**	[31] + [3]	No	No	No	0.3086
Yes	**Yes**	**[17] + [27]**	No	No	No	0.2986
Yes	**Yes**	**[17] + [27]**	**Yes**	No	No	0.2825
Yes	**Yes**	**[17] + [27]**	**Yes**	**Yes**	No	0.2782
Yes	**Yes**	**[17] + [27]**	**Yes**	**Yes**	**Yes**	**0.2609**

Table 2. Experimental results of a single model on the competition validation dataset.

ent age estimation (line 4 of Table 2) has proved to be very efficient gaining us about 0.015 of ϵ-score (i.e. 5%) with respect to fine-tuning without data augmentation. The data augmentation during the test stage (as explained in Section 4.2) has been efficient as well: the gain of about 0.005 in terms of ϵ-score i.e. 2% (line 5 of Table 2).

Finally, the last line of Table 2 proves the importance of the accurate age estimation of children. Adding a separate model for this age category has improved our validation score by about 0.017 of ϵ-points (i.e. 6%).

6. Competition results

Position	Team	ϵ-score
1	**OrangeLabs**	**0.2411**
2	palm_seu	0.3214
3	cmp+ETH	0.3361
4	WYU_CVL	0.3405
5	ITU_SiMiT	0.3668
6	Bogazici	0.3740
7	MIPAL_SNU	0.4569
8	DeepAge	0.4573

Table 3. Final results of the second edition of the ChaLearn LAP AAE competition.

The final results of the second edition of the ChaLearn LAP AAE competition are presented in Table 3.

Our team (**OrangeLabs**) has won the 1st place largely outperforming all other participants. Our final score on the test dataset ($\epsilon = 0.2411$) improved our best result obtained on the validation dataset ($\epsilon = 0.2609$) by about 0.02 of ϵ-points (i.e. 8%). As in the solutions of the previous year's competition [21, 16, 32], we have experienced a significant gain of performance due to merging of multiple models which have been trained using cross-validation.

In Figure 5, we present some examples of apparent age estimation by our solution on images from the competition test dataset.

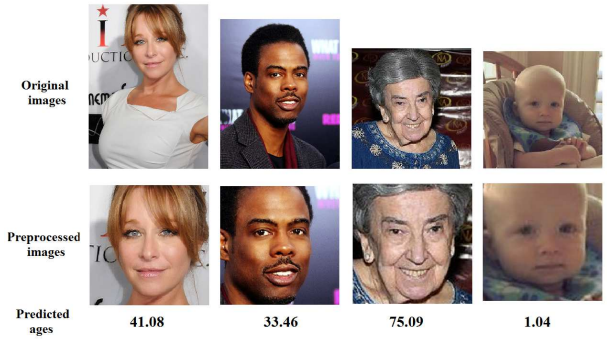

Figure 5. Apparent age estimation examples.

7. Conclusions and future work

In this work, we have presented our winning solution for the second edition of the ChaLearn LAP AAE competition.

The starting point of our approach is the training pipeline from the winning solution by [21] of the first edition of the competition: firstly, training the VGG-16 CNN for biological age estimation and then fine-tuning it for apparent age estimation. However, we have managed to improve the previous year's results by (1) using a separate age estimation model for images of children between 0 and 12 years old, (2) combining age encoding strategies: label distribution encoding for the "general" model and 0/1 classification encoding for the "children" model, and (3) integrating the state-of-the-art solution for face alignment by [27].

Our results are fully reproducible as we make the source codes and the trained CNN models publicly available[5].

Several works [11, 5] have shown the existence of certain interdependency between different soft biometrics traits (as age, gender and others) which has not been yet explored in this work due to time constraints. This path will be studied in our future work.

[5]Our final solution can be downloaded at `https://cactus.orange-labs.fr/apparent-age-estimation/`

References

[1] Fg-net aging dataset. http://fipa.cs.kit.edu/433_451.php.

[2] G. Antipov, S.-A. Berrani, and J.-L. Dugelay. Minimalistic cnn-based ensemble model for gender prediction from face images. *Pattern Recognition Letters*, 70:59–65, 2016.

[3] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar. Localizing parts of faces using a consensus of exemplars. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2930–2940, 2013.

[4] B.-C. Chen, C.-S. Chen, and W. H. Hsu. Cross-age reference coding for age-invariant face recognition and retrieval. In *Proceedings of European Conference on Computer Vision*, 2014.

[5] A. Dantcheva, C. Velardo, A. Dangelo, and J.-L. Dugelay. Bag of soft biometrics for person identification. *Multimedia Tools and Applications*, 51(2):739–777, 2011.

[6] S. Escalera, J. Fabian, P. Pardo, X. Baro, J. Gonzalez, et al. Chalearn looking at people 2015: Apparent age and cultural event recognition datasets and results. In *Proceedings of IEEE International Conference on Computer Vision Workshops*, 2015.

[7] S. Escalera, M. Torres, B. Martinez, X. Baro, H. J. Escalante, et al. Chalearn looking at people and faces of the world: Face analysis workshop and challenge 2016. In *Proceedings of IEEE conference on Computer Vision and Pattern Recognition Workshops*, 2016.

[8] X. Geng, C. Yin, and Z.-H. Zhou. Facial age estimation by learning from label distributions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(10):2401–2412, 2013.

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

[10] G. Guo and G. Mu. Simultaneous dimensionality reduction and human age estimation via kernel partial least squares regression. In *Proceedings of IEEE conference on Computer Vision and Pattern Recognition*, 2011.

[11] G. Guo and G. Mu. Human age estimation: What is the influence across race and gender? In *Proceedings of IEEE conference on Computer Vision and Pattern Recognition Workshops*, 2015.

[12] G. Guo, G. Mu, Y. Fu, and T. S. Huang. Human age estimation using bio-inspired features. In *Proceedings of IEEE conference on Computer Vision and Pattern Recognition*, 2009.

[13] H. Han, C. Otto, and A. K. Jain. Age estimation from face images: Human vs. machine performance. In *Proceedings of IEEE International Conference on Biometrics*, 2013.

[14] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, et al. Caffe: Convolutional architecture for fast feature embedding. *CoRR*, abs/1408.5093, 2014.

[15] Y. H. Kwon and N. da Vitoria Lobo. Age classification from facial images. *Computer Vision and Image Understanding*, 74(1):1–21, 1999.

[16] X. Liu, S. Li, M. Kan, J. Zhang, S. Wu, et al. Agenet: Deeply learned regressor and classifier for robust apparent age estimation. In *Proceedings of IEEE International Conference on Computer Vision Workshops*, 2015.

[17] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool. Face detection without bells and whistles. In *Proceedings of European Conference on Computer Vision*, 2014.

[18] B. Ni, Z. Song, and S. Yan. Web image and video mining towards universal and robust age estimator. *IEEE Transactions on Multimedia*, 13(6):1217–1229, 2011.

[19] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *Proceedings of British Machine Vision Conference*, 2015.

[20] K. Ricanek Jr and T. Tesafaye. Morph: A longitudinal image database of normal adult age-progression. In *Proceedings of IEEE conference on Automatic Face and Gesture Recognition*, 2006.

[21] R. Rothe, R. Timofte, and L. V. Gool. Dex: Deep expectation of apparent age from a single image. In *Proceedings of IEEE International Conference on Computer Vision Workshops*, 2015.

[22] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, et al. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[23] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[24] Y. Sun, X. Wang, and X. Tang. Deeply learned face representations are sparse, selective, and robust. In *Proceedings of IEEE conference on Computer Vision and Pattern Recognition*, 2015.

[25] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, et al. Going deeper with convolutions. In *Proceedings of IEEE conference on Computer Vision and Pattern Recognition*, 2015.

[26] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of IEEE conference on Computer Vision and Pattern Recognition*, 2014.

[27] M. Uricár, V. Franc, D. Thomas, A. Sugimoto, and V. Hlavác. Real-time multi-view facial landmark detector learned by the structured output svm. In *Proceedings of IEEE conference on Automatic Face and Gesture Recognition*, 2015.

[28] H.-F. Yang, L. B-Y, C. K-Y, and C. C-S. Automatic age estimation from face images via deep ranking. In *Proceedings of British Machine Vision Conference*, 2015.

[29] D. Yi, Z. Lei, and S. Z. Li. Age estimation by multi-scale convolutional network. In *Proceedings of Asian Conference on Computer Vision*, 2014.

[30] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *CoRR*, abs/1411.7923, 2014.

[31] L. Zhang, R. Chu, S. Xiang, S. Liao, and S. Z. Li. Face detection based on multi-block lbp representation. In *Proceedings of IEEE International Conference on Biometrics*, 2007.

[32] Y. Zhu, Y. Li, G. Mu, and G. Guo. A study on apparent age estimation. In *Proceedings of IEEE International Conference on Computer Vision Workshops*, 2015.