Single View Facial Age Estimation Using Deep Learning with Cascaded Random Forests

Imad Eddine Toubal(✉), Linquan Lyu, Dan Lin, and K. Palaniappan

Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO 65211, USA
itdfh@umsystem.edu

Abstract. The task of estimating a person's real age using unconstrained facial images has been actively studied in biometrics research. We developed several deep learning architectures and supervision methods for facial age estimation and evaluate the impact of different preprocessing and face alignment (or normalization) methods on the feature embedding subspace. The proposed novel two-stage supervised learning model utilizes ResNeXt as a backbone combined with a two-layer random forest (TLRF) to estimate age. Our deep architectures are trained using a custom loss function to handle variations in gender, pose, illumination, ethnicity, expression and context, on the *VGG-Face2 MIVIA Age Dataset* with over 575K images, as part of the Guess the Age (GTA) contest. Surprisingly, face alignment using FANet during training did not improve accuracy. We were able to achieve an Age Accuracy and Regularity score $AAR = 7.02$ with a variance $\sigma = 1.16$ using only ResNeXt. The proposed ResNeXt+TLRF model improved age-class generalizability with a smaller variance of $\sigma = 0.98$ and a second best $AAR = 6.97$.

Keywords: Age estimation · Face recognition · Face verification · Face alignment · Deepfakes · Deep learning · Random forest · Biological age

1 Introduction

Age estimation has many real-world applications including social robotic interaction, biometrics, demographics, business intelligence, online advertising, item recommendation, identity verification, video surveillance, access control, human-computer interaction, privacy and security, crowd behavior, law enforcement, and many more [1–3]. Single facial image age prediction is highly challenging [4–7], due to the variability in how individuals age based on their "ageotype" [8]. Everyone ages at different rates and biological age is influenced by genetics, diet, exercise, stress and environment. Moreover, visual cues about an individual's chronological age can vary due to pose, lighting, gender, scale, cosmetics, accoutrements, race, height, weight, health, emotion, occlusion, etc. [1,3,9,10]. Facial age feature embeddings can also be used to improve face recognition [11] and distinguish between real and synthetic (Deepfake) faces [12].

N. Tsapatsoulis et al. (Eds.): CAIP 2021, LNCS 13053, pp. 285–296, 2021.
https://doi.org/10.1007/978-3-030-89131-2_26

In the field of facial age estimation, there is an absence of large, reliable annotated datasets due to the difficulty in establishing ground truth ages. The LAP 2016 dataset [13] is reliable but only contains 7,591 images. Large datasets, like IMDB-Wiki [14], CACD [15], and UTK [16] are annotated with the age information based on online web crawling and social networks, therefore, reliability is not guaranteed. Some of the datasets for face aging prediction (prediction of a person's appearance at a younger or older time period) do not have enough diversity because many pictures are from the same individual at different times; like the FG-NET [17] dataset which contains 1,002 images of only 82 people. MORPH Album 2 [18] is another longitudinal dataset that contains 55,134 images of 13,618 subjects, but with a limited age distribution that ranges between 16 to 77. The CAIP Guess the Age (GTA) Contest [19], uses the *VGG-Face2 MIVIA Age Dataset* [2]. It consists of 575,073 images of more than 9,000 identities, collected at different ages. The images are extracted from the VGGFace2 [20] and annotated with the person's age by means of a knowledge distillation technique [2]. The *VGG-Face2 MIVIA Age Dataset* is the most accurate facial age dataset currently available at this scale in terms of sample size and heterogeneity. Despite the lack of precise age data, several machine learning and data driven age estimation models have emerged [1]. DLDL-v2 (ThinAgeNet) [21] currently stands as the state-of-the-art on the MORPH Album 2 and ChaLearn 2015 and 2016 [22] datasets.

Guess the Age (GTA) Contest considers the biometric task of estimating a person's age using only their facial image as input [1,2]. Although there are over 575K age labeled images in the *VGG-Face2 MIVIA Age Dataset* covering gender, ethnicity, varying poses, scale and illumination, there is a high degree of age class imbalance. The four age groups covering, 1 to <20 and ≥ 60, the two youngest and two oldest groups (out of eight categories) constitute less than 10% of the data; the youngest and oldest age categories make up less than 1%.

In this paper, we propose a novel age estimation approach that uses a two-layer classification-plus-regression random forest trained on deep feature embeddings from the ResNeXt50 architecture [23]. We show that an ensemble of weak decision trees trained on deep features has smaller variance than a pure deep neural model with end-to-end optimization.

2 Deep Learning Methods for Age Estimation

2.1 Pre-processing

We used z-normalization to normalize the intensity value of the pre-cropped input face images. *VGG-Face2 MIVIA Age dataset* contains images with already cropped single faces, and hence a face detection step was not necessary.

Our experience with incorporating a face-alignment step using FANet produced mixed results [24]. FANet was used to estimate 68 facial key-points in the cropped face image. These extracted key-points are matched with a template (standard face pose) set of key-points to estimate the 2D alignment transformation matrix. We then apply this transformation to warp the original face image to realign the face. Sample results from this step are shown in Fig. 1.

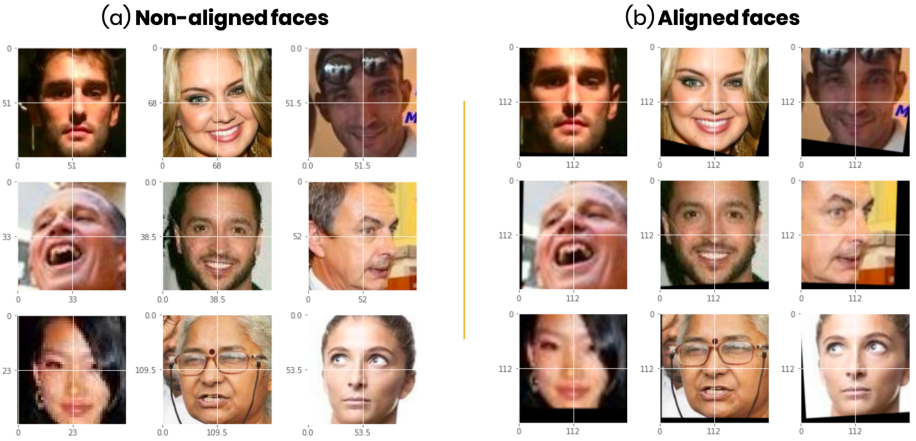

Fig. 1. Sample face-alignment using FANet [24] applied to face images from the *VGG-Face2 MIVIA Age Dataset*. Note that not all faces are warped when they are side profiles or have up-down tilts. Our final age estimates using ResNeXT+TLRF (vs actual) for these subjects from left to right and top to bottom are: 27 (27), 27 (30), 29 (29), 34 (35), 31 (31), 56 (57), 27 (29), 58 (59), 28 (28).

We evaluated the potential benefit of face-alignment since this can reduce the learning complexity for age estimation when faces are in similar poses. However, we found that face-alignment reduced the diversity in the training dataset which could lead to overfitting, and reduce the performance of deep neural networks. For this reason, we trained the deep neural models on non-aligned faces to ensure better generalizability. Although face-alignment of training images had limited benefit in our initial testing, several approaches are being studied to better incorporate face-alignment as a data augmentation approach to improve performance during inference.

2.2 ResNeXt CNN

Architecture. We use the ResNeXt architecture [23] for extracting feature descriptors due to its advantages over the classical ResNet architecture. ResNet uses residual blocks [25] that make use of sequential convolution layers with an added skip connection. This simple modification led to a breakthrough in performance when compared to classical CNNs (such as VGG [26]). The ResNeXt architectural insight was the notion of *cardinality*, that many parallel small convolutions are better than a single deep sequence of convolutions with wider kernels. This is done by using parallel convolution streams with fewer channels instead of a single sequential stream with more channels. Using the cardinality property, they experimentally demonstrated an improvement in accuracy on the ImageNet benchmark [27] by simply increasing the cardinality without adding more parameters. This is crucial when dealing with smaller class sizes where over-fitting is more likely.

Hyper-parameters. We train a single output regression version of ResNeXt using the Adam [28] optimizer that is a variation of the Stochastic Gradient Descent algorithm [29]. We use an initial learning rate $\alpha = 10^{-4}$. The model weight initialization is based on transfer learning with pre-trained weights from the ImageNet classification dataset. Additionally, we adopt warm restart scheduling during training using the cosine annealing method [30]. Learning rate is one of the most important hyper-parameters in training neural networks. For this reason, adaptively selecting a learning rate and/or scheduling are crucial for a more robust training [31–33].

Loss. We define a new loss function $\mathcal{L}_{AAR}$ inspired by the Age Accuracy and Regularity (AAR) metric from the GTA contest. For a set of predicted ages $\hat{\mathbf{y}}$ and real ages $\mathbf{y}$ of size N, the loss function equation is given as:

$$\mathcal{L}_{AAR}(\mathbf{y}, \hat{\mathbf{y}}) = \gamma \mathcal{L}_1(\mathbf{y}, \hat{\mathbf{y}}) + \lambda \sigma \tag{1}$$

where:

$$\mathcal{L}_1(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{N} \sum_{i=1}^{N} \ell_1(y_i, \hat{y}_i) \tag{2}$$

with:

$$\ell_1(y, \hat{y}) = \begin{cases} \frac{1}{2\beta}(y - \hat{y})^2, & \text{if } |y - \hat{y}| < \beta \\ |y - \hat{y}| - \frac{1}{2}\beta, & \text{otherwise} \end{cases} \tag{3}$$

and:

$$\sigma = \sqrt{\frac{1}{8} \sum_{j=1}^{8} [\mathcal{L}_1(y^j, \hat{y}^j) - \mathcal{L}_1(y, \hat{y})]^2} \tag{4}$$

where y is the set of true ages, $\hat{y}$ is the set of predicted ages, y^j and $\hat{y}^j$ are the true and predicted ages respectively that belong to j^{th} age group. $\mathcal{L}_1$ is the smooth L1 norm (mean absolute error), σ is a regularization term to reduce the model's sensitivity to the dataset imbalance. γ and λ are coefficients terms for two parts of the loss function. The loss parameters used in this work are $\gamma = 0.7$, $\lambda = 0.3$, and $\beta = 1.0$.

Note that there are two main differences between our loss function and the AAR metric. First, we use *smooth L1* distance ℓ_1 as opposed to MAE. *Smooth L1* was proven less sensitive to outliers and less prone to exploding gradients [34,35]. Second, we do not clip $\mathcal{L}_1$ and σ components to a maximum value; instead, we give them different weights to emphasize one over the other.

Label Distribution Smoothing (LDS). To tackle the challenge of age imbalance in the dataset, the Label Distribution Smoothing (LDS) method was evaluated [36]. LDS convolves a symmetric 1-D Gaussian smoothing kernel k with the label distribution (histogram) $p(y)$ to produce a kernel-smoothed version that interpolates information of data samples with nearby labels. A symmetric kernel

is a kernel that satisfies: $k(y+\Delta y) = k(y-\Delta y)$ and $\nabla_y k(y+\Delta y)+\nabla_y k(y-\Delta y) = 0, \forall y \in Y$. The smoothed label distribution, $p'(x)$, is a convolution between the distribution $p(y)$ and the kernel $k(y)$:

$$p'(y) = k(y) * p(y) \tag{5}$$

where $*$ is the convolution operator. The loss function is then reweighted by scaling the estimates with the inverse of the label frequency for each sample:

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} \frac{\ell(y_i, \hat{y}_i)}{p'(\hat{y}_i)}. \tag{6}$$

2.3 Two-Layer Random Forest (TLRF)

ResNeXt takes an RGB input image x_i and uses a series of convolutional blocks to produce a feature embedding of 2,048 dimensions $f_i \in \mathbb{R}^{2,048}$. It then uses a single-layer perceptron (fully connected neural regressor) with learned weights to make a final prediction of age, $\hat{y}_i \in \mathbb{R}$. We replace the neural regressor with our two-layer random forest (TLRF) combining classification-plus-regression to make a final prediction. In the first stage, TLRF uses ResNeXt features as input to the first layer (random forest classifier) to make a classification of the given sample's age group in the form of a probability vector $p_i(G) \in \mathbb{R}^8$ where:

$$G \in \{[1,9], [10,19], [20,29], [30,39], [40,49], [50,59], [50,59], [60,+\infty]\}$$

We concatenate the 8-dimensional predicted probability vector for all eight of the age groups with the learned 2048-dimensional deep embedding feature vector f_i into an augmented vector. We, then use that as input to the second random forest regressor layer in our TLRF. The final regression output is then rounded up to the nearest integer. A visual diagram of our approach is shown in Fig. 2.

Our experiments showed that TLRF improves the performance and stability of ResNeXt. For each layer of TLRF, we utilize a random forest of 100 decision trees trained in parallel on the ResNeXt embedding feature vectors, and each decision tree uses a maximum of 128 randomly selected features.

2.4 Training the Deep Architectures

Dataset Split. The *VGG-Face2 MIVIA Age Dataset* consists of 575,073 example cases [1,2]. We used 90% of this dataset (517,562) for training and the remaining 10% for evaluation (57,511). As the dataset is not uniformly distributed in terms of age, we sample 10% from each age group j for evaluation; rather than 10% uniformly sampled across the entire set. We then divide the training data further into a training and validation split of 90% and 10% sizes respectively.

Pre-processing. Face images are normalized as explained in the pre-processing section. In addition, we resize the images to 224×224 resolution to match the expected input of ResNeXt network. Additionally, for better network stability, we normalize the age to range between 0 and 1.

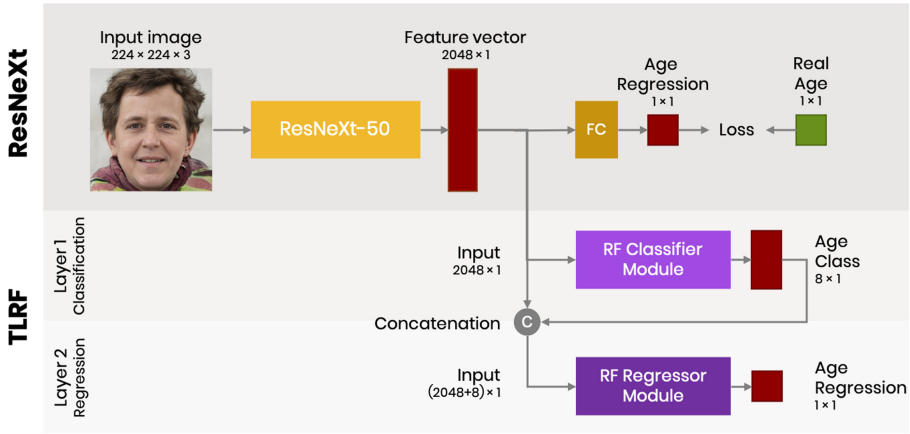

Fig. 2. Proposed ResNeXt+TLRF facial age estimation pipeline using ResNeXt-50 feature embedding vector with ImageNet transfer learning plus *VGG-Face2 MIVIA Age* training. A dual stage random forest estimates both class labels and age estimates.

Data Augmentation. For data augmentation, we use random horizontal flipping. Since our experiments showed that face-alignment hurts training, random rotations and distortions could also be applied in future work.

3 Experimental Results

In addition to the mean absolute error, we report the Age Accuracy and Regularity (AAR metric). GTA contest defined the AAR performance measure as:

$$\text{AAR} = max(0; 7 - \text{MAE}) + max(0; 3 - \sigma) \tag{7}$$

with a maximum score of 10; with $\sigma = \sqrt{\frac{1}{8}\sum_{j=1}^{8}(\text{MAE}_j - \text{MAE})^2}$ where MAE is the mean absolute error and MAE_j is the mean absolute error for the j^{th} age group. The mean average error (MAE) is given as $MAE = \frac{1}{N}\sum_i |y_i - \hat{y}_i|$, where i is the sample index over all age categories. All evaluations are performed on the evaluation set that we described in Sect. 2.4 unless specified otherwise.

3.1 ResNeXt

A performance comparison of ResNet vs. ResNeXt in terms of MAE is given in the Table 1. Additionally, the table shows the difference in performance of our custom soft AAR loss compared to the mean squared error (MSE) loss $\mathcal{L}_{MSE}$ where, for a given batch of size n:

$$\mathcal{L}_{MSE} = \frac{1}{n}\sum_{i=1}^{n} ||y_i - \hat{y}_i||^2 \tag{8}$$

Table 1 shows how using ResNeXt over its predecessor ResNet improves performance. Additionally, our custom AAR loss consistently improves the AAR metric in both networks. We can also note that LDS did not help in improving the performance; hence, we choose to move forward with ResNeXt trained using the AAR Loss. The LDS-trained ResNeXt network was trained using both MSE and AAR loss functions. The LDS-trained ResNeXt trained using AAR loss shows better performance.

Table 1. Accuracy comparison of ResNet, ResNeXt and ResNeXt with label distribution smoothing LDS using the evaluation data.

Architecture	Loss	MAE_1	MAE_2	MAE_3	MAE_4	MAE_5	MAE_6	MAE_7	MAE_8	**MAE**↓	$\sigma\downarrow$	**AAR**↑
ResNet	$\mathcal{L}_{MSE}$	1.74	1.94	1.58	2.00	2.05	1.83	1.77	1.94	1.87 ± 2.05	0.14	7.99
	$\mathcal{L}_{AAR}$	1.79	1.96	1.47	1.83	1.94	1.77	1.69	1.99	1.75 ± 1.98	0.17	8.08
ResNeXt	$\mathcal{L}_{MSE}$	2.11	1.85	1.53	1.95	2.02	1.81	1.72	1.93	1.82 ± 2.06	0.17	8.01
	$\mathcal{L}_{AAR}$	1.91	1.89	1.46	1.80	1.93	1.72	1.74	1.78	1.73 ± 1.97	0.15	8.12
ResNeXt (LDS)	$\mathcal{L}_{MSE}$	1.92	2.11	1.61	1.97	2.04	1.84	1.85	1.88	1.88 ± 2.06	0.15	7.98
	$\mathcal{L}_{AAR}$	1.98	2.02	1.51	1.90	2.01	1.79	1.76	1.87	1.81 ± 2.05	0.17	8.03

3.2 Two-Layer Random Forest (TLRF)

TLRF Classifier Module. Although classifying a face's age group is an easier task than the exact age, it is still challenging due to age class imbalance. Table 2 shows the performance of the TLRF classifier module on our evaluation set. The F1 measure is much lower for underrepresented age groups due to lower recall.

TLRF Regressor Module. Several regression random forest topologies were evaluated against our proposed TLRF. First, a traditional regression random forest (RRF) was trained and evaluated with different number of trees using the ResNeXt 2048-dimensional feature descriptor. Then, we compare the single-layer random forest approach (RRF) to the proposed TLRF. Table 3 summarizes the RF ablation study experimental results using ResNeXt in combination with different RF configurations. For this part, a ResNeXt-50 was trained using MSE loss. The ResNeXt-50 residual deep network with a fully connected final regression layer performed well with an AAR of 8.01 on the held out evaluation set and 8.16 on the combined training and evaluation sets. Incorporating a random

Table 2. TLRF age group classifier module performance (using ResNeXt descriptor) on evaluation data. Support is the subset of data in each of the eight age categories used for evaluation.

Age group	<10	10–19	20–29	30–39	40–49	50–59	60–69	>69	Overall
Precision	0.85	0.75	0.87	0.83	0.82	0.83	0.82	0.85	0.83
Recall	0.63	0.58	0.89	0.83	0.83	0.84	0.77	0.54	0.83
F1	0.72	0.66	0.88	0.83	0.82	0.84	0.80	0.66	0.83
Support	185	1960	14153	15001	13226	9341	3298	347	57,511

Table 3. Experimental results showing accuracy on *training+evaluation* (T+E) and *evaluation* (E) sets with different random forest learning methods (number of trees and number of layers). Last row ResNeXt+TLRF is our final result. RRF refers to Regression Random Forest. All ResNeXt networks were trained using the MSE loss.

Method	MAE↓		$\sigma\downarrow$		AAR↑	
Dataset	T+E	E	T+E	E	T+E	E
ResNeXt	**1.35** ± 1.75	**1.82** ± 2.06	0.14	0.17	8.16	**8.01**
ResNeXt + RRF (64 trees)	1.65 ± 1.57	1.90 ± 2.10	0.16	0.15	8.19	7.95
ResNeXt + RRF (100 trees)	1.66 ± 1.56	1.89 ± 2.09	0.15	0.15	8.20	7.96
ResNeXt + RRF (200 trees)	1.66 ± 1.56	1.89 ± 2.09	0.17	0.15	8.17	7.96
ResNeXt + TLRF (2×100 trees)	1.66 ± 1.56	1.88 ∓ 2.08	**0.13**	**0.14**	**8.21**	7.98

forest learning component improves the overall AAR accuracy using 100 trees to 8.20 on the combined training and evaluation sets and reduces AAR to 7.96 on the held out evaluation set. Increasing the number of trees to 200 did not improve performance on the evaluation set and decreased AAR performance to 8.17 on the combined T+E sets. *Using the two-layer classification plus regression random forest with the same ResNeXt-50 feature embedding vector results in the best AAR of 8.21 on the combined training and evaluation sets and improves the score to 7.98 comparing to traditional regression random forests.* This model also had the smallest class standard deviation (σ), on the held out evaluation set and the combined set.

3.3 Generalizability Performance Using the Withheld GTA Data

Based on the results described previously, we submitted the ResNeXt+TLRF as our single official submission to the GTA Challenge competition. After our official submission to the GTA contest, we continued to explore the generalization capability of the different architectures on the unseen hidden dataset with assistance from the MIVIA Lab at the University of Salerno.

Experimental results in Table 4 show that using the proposed custom AAR loss function consistently improves the generalizability of face estimation MAE accuracy in both our evaluation and the GTA hidden test set for all methods. ResNeXt trained using the AAR loss function has the highest AAR score of 8.12 on the heldout evaluation data and score of 7.02 on the hidden test set. The submitted ResNeXt+TLRF method also generalizes well on new unseen faces and has the lowest age group variance of 0.98. We notice that apart from the two underrepresented age groups (MAE_1 and MAE_8 with $< 1.0\%$ samples), the MAE scores are quite consistent between our evaluation split data and the GTA challenge's hidden test data. It is important to note that the standard deviation, σ, is more than eight times higher in the hidden dataset than the held out evaluation data due to larger deviations in MAE_1 and MAE_8. The lowest variation in the hidden or withheld data is the ResNeXt (MSE)+TLRF method, that we submitted to the GTA contest, and is italicized in Table 4.

Table 4. Results on the Guess the Age (GTA) contest hidden (or withheld) test dataset. Column labeled D indicates dataset used: T, for our separate evaluation set (see Sect. 2.4); H, for the unseen hidden GTA challenge test set.

Method	D	MAE_1	MAE_2	MAE_3	MAE_4	MAE_5	MAE_6	MAE_7	MAE_8	**MAE**↓	σ ↓	**AAR**↑
ResNeXt (MSE)	T	2.11	1.85	1.53	1.95	2.02	1.81	1.72	1.93	1.82	0.17	8.01
ResNeXt (AAR)	T	1.91	1.89	1.46	1.80	1.93	1.72	1.74	1.78	**1.73**	0.15	**8.12**
ResNeXt (LDS MSE)	T	1.92	2.11	1.61	1.97	2.04	1.84	1.85	1.88	1.88	0.15	7.98
ResNeXt (LDS AAR)	T	1.98	2.02	1.51	1.90	2.01	1.79	1.76	1.87	1.81	0.17	8.03
ResNeXt + RF	T	1.81	2.04	1.57	2.01	2.08	1.88	1.88	1.97	1.89	0.15	7.96
ResNeXt (MSE) + TLRF	*T*	*1.89*	*1.92*	*1.57*	*2.01*	*2.08*	*1.87*	*1.87*	*1.92*	*1.88*	***0.14***	*7.98*
ResNeXt (AAR) + TLRF	T	1.61	1.94	1.51	1.85	2.00	1.78	1.79	1.79	1.79	0.15	8.06
ResNeXt (MSE)	H	5.92	2.52	1.71	1.86	1.96	1.86	2.37	3.21	1.91	1.31	6.78
ResNeXt (AAR)	H	5.35	2.37	1.59	1.80	1.89	1.78	2.23	3.09	**1.82**	1.16	**7.02**
ResNeXt (LDS MSE)	H	5.87	2.63	1.86	2.03	2.02	1.85	2.07	2.60	2.00	1.26	6.74
ResNeXt (LDS AAR)	H	5.42	2.24	1.74	1.94	1.92	1.80	2.24	3.55	1.91	1.19	6.90
ResNeXt + RF	H	5.35	2.38	1.66	1.83	1.94	1.85	2.38	3.82	1.88	1.20	6.92
ResNeXt (MSE) + TLRF	*H*	*4.84*	*2.45*	*1.87*	*2.05*	*2.10*	*1.94*	*2.44*	*3.67*	*2.05*	***0.98***	*6.97*
ResNeXt (AAR) + TLRF	H	5.29	2.35	1.67	1.83	1.94	1.83	2.34	3.64	1.87	1.17	6.96

Additionally, using MSE loss, our TLRF method outperformed LDS. Although ResNeXt (AAR) without a TLRF module achieves a slightly better AAR score than ResNeXt+TLRF, it actually has a higher (worse) σ score, which indicates less generalizability across underrepresented age groups. Other methods of augmentation may help with enhancing the generalization capability of the architectures by pretraining with automatic face aging methods which provide a large amount of ground truth across age categories [37], selectively augmenting the lowest representated groups more, incorporating augmentation in feature space during the random forest training, etc.

4 Conclusions

Accurate unconstrained age estimation or categorization, using images or video, is useful in a number of applications including face recognition, age appropriate advertising and retail, venue access, detecting deep fakes, health and exercise monitoring, emotion analysis, forensics, privacy and security applications [38]. Our proposed two-stage supervised learning pipeline for facial age estimation using a ResNeXt deep learning stage followed by a two-layer random forest (TLRF) was able to estimate age with a mean absolute error of about 2 years across all eight age categories with a standard deviation of less than one year. Despite the significant class imbalance in the training data, we were able to achieve an AAR score of 6.97±0.98 (ResNeXt+TLRF) and 7.02±1.16 (ResNeXt) out of 10.0 on the hidden test data of the *VGG-Face2 MIVIA Age Dataset* as part of the Guess the Age (GTA) contest. The most challenging age categories are the youngest and oldest groups at the two extremes of the age distribution for which there was the least amount of training data (less than 1%). The experimental results demonstrate that a distribution adaptive (AAR) loss

function is effective for training with class imbalance. Face alignment did not improve performance and test time data augmentation had limited benefit. For facial age estimation, an ensemble of weak learners trained on deep features is less sensitive to under-represented age groups compared to a purely deep neural regression model trained in an end-to-end fashion.

Acknowledgments. Research partially supported by U.S. National Science Foundation award 2114141, Army Research Laboratory cooperative agreement W911NF1820285 and Army Research Office DURIP W911NF-1910181. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the U.S. Government or agency.

References

1. Carletti, V., Greco, A., Percannella, G., Vento, M.: Age from faces in the deep learning revolution. IEEE Trans. Pattern Anal. Mach. Intell. **42**(9), 2113–2132 (2020)
2. Greco, A.A., Vento, S.M., Vigilante, V.: Effective training of convolutional neural networks for age estimation based on knowledge distillation. Neural Comput. Appl. 1–16 (2021)
3. Abdolrashidi, A., Minaei, M., Azimi, E., Minaee, S.: Age and gender prediction from face images using attentional convolutional network. arXiv preprint arXiv:2010.03791 (2020)
4. Park, U., Tong, Y., Jain, A.K.: Age-invariant face recognition. IEEE Trans. Pattern Anal. Mach. Intell. **32**(5), 947–954 (2010)
5. Angulu, R., Tapamo, J.R., Adewumi, A.O.: Age estimation via face images: a survey. EURASIP J. Image Video Process. **2018**(1), 1–35 (2018)
6. Ranjan, R., et al.: Unconstrained age estimation with deep convolutional neural networks. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 109–117 (2015)
7. Han, H., Otto, C., Liu, X., Jain, A.K.: Demographic estimation from face images: Human vs. machine performance. IEEE Trans. Pattern Anal. Mach. Intell. **37**(6), 1148–1161 (2015)
8. Ahadi, S., et al.: Personal aging markers and ageotypes revealed by deep longitudinal profiling. Nat. Med. **26**(1), 83–90 (2020)
9. Greco, A., Saggese, A., Vento, M., Vigilante, V.: A convolutional neural network for gender recognition optimizing the accuracy/speed tradeoff. IEEE Access **8**, 130771–130781 (2020)
10. Yolcu, G., Oztel, I., Kazan, S., Oz, C., Palaniappan, K., Lever, T.E., Bunyak, F.: Facial expression recognition for monitoring neurological disorders based on convolutional neural network. Multimed. Tools Appl. **78**(22), 31581–31603 (2019)
11. Wang, M., Deng, W.: Deep face recognition: a survey. Neurocomputing **429**, 215–244 (2021)
12. Lewis, J.K., et al.: Deepfake video detection based on spatial, spectral, and temporal inconsistencies using multimodal deep learning. In: IEEE Applied Imagery Pattern Recognition Workshop (AIPR), pp. 1–9 (2020)
13. Escalera, S., et al.: ChaLearn looking at people and faces of the world: face analysis workshop and challenge 2016. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 1–8 (2016)

14. Rothe, R., Timofte, R., Van Gool, L.: Deep expectation of real and apparent age from a single image without facial landmarks. Int. J. Comput. Vis. **126**(2), 144–157 (2018)
15. Chen, B.-C., Chen, C.-S., Hsu, W.H.: Cross-age reference coding for age-invariant face recognition and retrieval. In: European Conference on Computer Vision, pp. 768–783 (2014)
16. Zhang, Z., Song, Y., Qi, H.: Age progression/regression by conditional adversarial autoencoder. In: IEEE Conference on Computer Vision and Pattern Recognition pp. 5810–5818 (2017)
17. Lanitis, A., Taylor, C.J., Cootes, T.F.: Toward automatic simulation of aging effects on face images. IEEE Trans. Pattern Anal Mach. Intell. **24**(4), 442–455 (2002)
18. Rawls, A.W., Ricanek, K.: MORPH: Development and optimization of a longitudinal age progression database. In: European Workshop on Biometrics and Identity Management, pp. 17–24 (2009)
19. Guess The Age Contest 2021. https://gta2021.unisa.it/. Accessed 16 July 2021
20. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: VGGFace2: a dataset for recognising faces across pose and age. In: IEEE International Conference on Automatic Face & Gesture Recognition, pp. 67–74 (2018)
21. Gao, B.-B., Zhou, H.-Y., Wu, J., Geng, X.: Age estimation using expectation of label distribution learning. In: IJCAI, pp. 712–718 (2018)
22. Ponce-López, V., et al.: ChaLearn LAP 2016: first round challenge on first impressions - dataset and results. In: European Conference on Computer Vision Workshops, pp. 400–418 (2016)
23. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1492–1500 (2017)
24. Bulat, A., Tzimiropoulos, G.: How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks). In: IEEE International Conference on Computer Vision (2017)
25. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
26. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
27. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255 (2009)
28. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
29. Bengio, Y.: Practical recommendations for gradient-based training of deep architectures. In: Neural networks: Tricks of the Trade, pp. 437–478 (2012)
30. Loshchilov, I., Hutter, F.: SGDR: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016)
31. Xu, Z., Dai, A.M., Kemp, J., Metz, L.: Learning an adaptive learning rate schedule. arXiv preprint arXiv:1909.09712 (2019)
32. Schaul, T., Zhang, S., LeCun, Y.: No more pesky learning rates. In: International Conference on Machine Learning, pp. 343–351 (2013)
33. Zeiler, M.D.: ADADELTA: an adaptive learning rate method. arXiv preprint arXiv:1212.5701 (2012)
34. Girshick, R.: Fast R-CNN. In: IEEE International Conference on Computer Vision, pp. 1440–1448 (2015)

35. Philipp, G., Song, D., Carbonell, J.G.: The exploding gradient problem demystified-definition, prevalence, impact, origin, tradeoffs, and solutions. arXiv preprint arXiv:1712.05577 (2017)
36. Yang, Y., Zha, K., Chen, Y.-C., Wang, H., Katabi, D.: Delving into deep imbalanced regression. arXiv preprint arXiv:2102.09554 (2021)
37. Duong, C.N., et al.: Automatic face aging in videos via deep reinforcement learning. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 10013–10022 (2019)
38. Morris, J., Newman, S., Palaniappan, K., Fan, J., Lin, D.: Do you know you are tracked by photos that you didn't take: Location-aware multi-party image privacy protection. arXiv preprint arXiv:2103.10851 (2021)