2018 24th International Conference on Pattern Recognition (ICPR)
Beijing, China, August 20-24, 2018

Revised Contrastive Loss for Robust Age Estimation From Face

Hongyu Pan[1,2], Hu Han[*,1], Shiguang Shan[1,2,3] and Xilin Chen[1,2]
[1]Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, China
[2]University of Chinese Academy of Sciences, Beijing 100049, China
[3]CAS Center for Excellence in Brain Science and Intelligence Technology
hongyu.pan@vipl.ict.ac.cn, {hanhu, sgshan, xlchen}@ict.ac.cn

***Abstract*—Age estimation has broad applications in many fields, such as video surveillance, social networking, and human-computer interaction. Many of the existing approaches treat age estimation as a classification problem; however, the individual age values are not independent classes; they have an ordinal relationship. Classification loss such as softmax is not able to model such kind of relationship. In this paper, we propose a new loss, called revised contrastive loss, to model the ordinal relationship of individual ages. Specifically, the revised contrastive loss is proposed to penalize the distance between two face images in the feature space according to their age difference, which makes the learned features more discriminative for the age estimation task. We embed the proposed revised contrastive loss and softmax loss into a Convolutional Neural Network (CNN), and optimize the networks via Stochastic Gradient Descent (SGD) in an end-to-end fashion. Experimental results on a number of challenging face aging databases (FG-NET, MORPH Album II, and CLAP2016) show that the proposed approach outperforms the state-of-the-art methods by a large margin using a single model.**

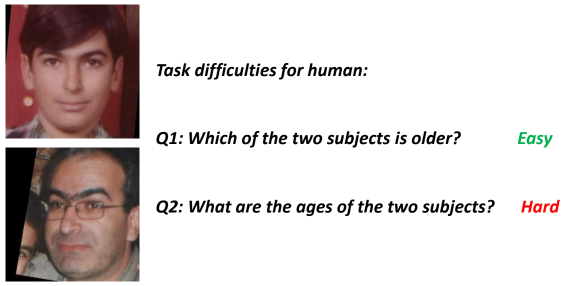

Fig. 1. Given two face images from the FG-NET dataset, we can see that estimating the exact age of each subject is more difficult than determining which subject is older than the other. The ordinal relationship of the individual ages could be exploited by the age estimation models for learning features that are discriminative for individual face images with different ages.

I. Introduction

Age estimation from facial images has broad applications in different fields, such as video surveillance, social networking, and human-computer interaction. Existing approaches for age estimation can be grouped into three categories: classification-based method, regression-based method, and ranking-based method. Classification-based methods are commonly used for age group classification of face images [1][2]. The classification-based methods do not consider the ordinal relationship of different age group, and thus the costs of classifying a young subject as middle-aged subject and old subject are the same. Apparently, such a modeling method is not optimum for the age estimation task.

Regression-based methods are widely used to estimate the exact age of the subject in a face image [3][4][5]. Many of the existing regression-based methods use a Euclidean loss, which is able to reflect the ordinal relationship of the age values. However, this type of loss defined based on a single image does not necessarily retain the relative order among samples.

In recent years, several ranking-based methods have been proposed for age estimation from face images [6][7]. These approaches treat the age values as a rank order data. And use multiple binary classifiers to determine the rank of a test face image. Different from the L_2 loss based regression methods, these methods could explicitly make use of the ordinal relationship among the samples.

We observed that it is very difficult to estimate the exact age of each subject; by contrast, it is relatively easy to determine which subject is older than the other given two subjects' face images (see Fig. 1). Inspired by this observation and relative attribute learning [8], in this paper, we propose a revised contrastive loss that makes use of the pair-wise ordinal relationship between individual face images to achieve more age informative feature learning. As shown in Fig. 2, in addition to the softmax loss for age classification, a revised contrastive loss is proposed to penalize the distance between two face images in their feature space according to their age difference. This makes the learned features more informative for the age estimation task. The contributions of our work are as follows:

- We propose a new loss, named revised contrastive loss, to model the pair-wise ordinal relationship among individual samples in order to learn more age informative features.
- The proposed function can be easily embedded into different CNN networks, and optimized via SGD in an

*H. Han is the corresponding author.

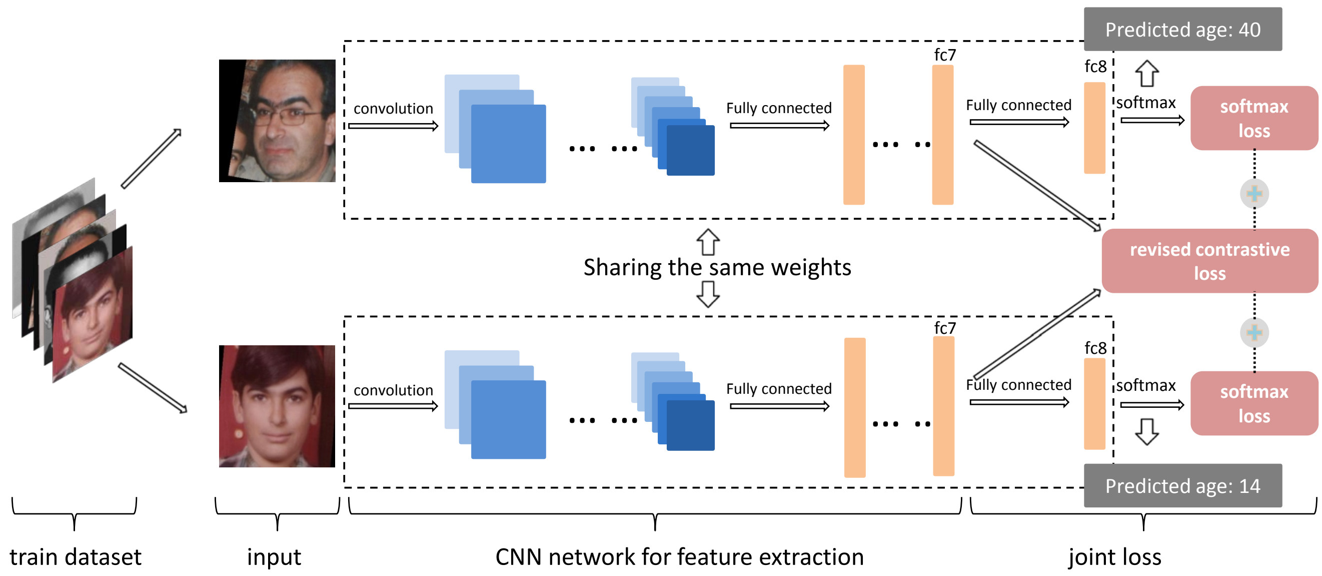

Fig. 2. Overview of the architecture of the proposed approach for age estimation. The input is a pair of images, which are passed through two CNNs sharing the same weights for feature extraction. A joint loss consisting of softmax and revised contrastive loss is then used during backpropagation, in which softmax loss penalizes the differences between the estimated age and the ground-truth age, and the revised contrastive loss penalizes the incorrect ordinal relationship of each pair of images.

end-to-end fashion.

- The proposed approach is evaluated on a number of challenging databases (FG-NET [9], MORPH Album II [10] and CLAP2016 [11]), and achieves better results than the state-of-the-art methods.

II. Related Work

A. Age Estimation

Kwon *et al.* [12] did the very early work on age estimation from a face, in which the ages are divided into only three groups, babies, young adults and senior adults. Later, accurate age estimation from a face image attracted increasing attentions. Guo *et al.* [13] used multi-directional and multi-scale Gabor filters followed by feature pooling to extract BIF features for age estimation. BIF based age estimation methods [13][14] reported promising age estimation results on a number of public-domain face databases such as FG-NET [9], MORPH II [10] databases.

With the success of deep learning methods, deep learning architectures are also being used in age estimation. Similar to [12], Yi *et al.* [15] used CNN models to extract features from different facial regions, and used a square loss for age to do age estimation. The deep feature shows its advantages over the previous hand-crafted features. Ordinal information was used in [7] to train multiple binary CNN networks and aggregated these outputs as predicted age. Yang *et al.* [16] proposed an age distribution learning method based on softmax, in which the label is an age distribution, instead of an age value.

B. Relative Attributes Learning

Typical visual classification approaches usually map low-level image features to object category classes directly. Nevertheless, relative attributes [8] aimed at learning a ranking function by utilizing the information of how object/scene categories relate according to different attributes. The relative attribute has been studied in a number of ways. For example, Wang *et al.* [17] proposed an approach to learn attributes and object classes together, utilizing explicit similarity-based supervision to share training samples with limited. A hierarchically structured approach was proposed in [18] to learn some rankers for each facial attribute and combined all of the features and ranks per attribute as a new feature for a global ranking function to classify these attributes.

III. Proposed Method

As shown in Fig. 2, we propose a revised contrastive loss for learning features that are discriminative for individual ages. We embed the revised contrastive loss and softmax loss together into CNN, and optimized the network via SGD [19] in an end-to-end fashion. We provide the details of our approach in the following sections.

A. Revised Contrastive Loss

Contrastive loss was originally proposed for dimensionality reduction in [20] aiming to map similar input samples to nearby points and dissimilar samples to distant points in the low dimensional feature space. Contrastive loss has been found to be useful for image classification tasks, face recognition, etc. The fundamental idea of contrastive loss can be represented as

$$\ell c(x_i,x_j) = \begin{cases} \frac{1}{2}\left\|f(x_i)-f(x_j)\right\|_2^2, & y_i = y_j \\ \frac{1}{2}max(0,\mathscr{M}-\left\|f(x_i)-f(x_j)\right\|_2)^2, & y_i \neq y_j, \end{cases} \tag{1}$$

where x_i and x_j are the input samples, y_i and y_j are the age labels, $f(x\cdot)$ is the feature learning function, and $\mathscr{M}$ is a margin. We can see that the standard contrastive loss includes two parts: one part is the distance between a pair of samples and the other part is the distance margin. A pair of samples contributes to the loss only when their distance is within the margin.

While contrastive loss aims to retain the intra-class and inter-class distances, it does not retain the ordinal relationship between individual samples. However, for age estimation task, the ordinal relationship of individual ages can be very important. For example, classifying a 20-year old subject as 30-year old and classifying a 20-year old subject as 50-year old are both poor age estimates, but apparently, these two estimates should have different penalties.

Let (x,y) denote a face image and the corresponding age label. Given three face images and their ages, i.e., (x_i,y_i), (x_j,y_j), and (x_k,y_k), without loss of generality, we assume $y_i > y_j$ and $y_j > y_k$. Thus, we can get $y_i > y_j > y_k$ based on the transitive property. In addition, we can derive

$$y_i - y_k > y_i - y_j, and\ y_i - y_k > y_j - y_k. \tag{2}$$

Based on the three face images, we can build three image pairs, i.e., $\{(x_i,y_i),(x_j,y_j)\}$, $\{(x_i,y_i),(x_k,y_k)\}$ and $\{(x_j,y_j),(x_k,y_k)\}$. For the three pairs of face images, we expect that the feature space learned by $f(\cdot)$ is able to retain their ordinal relationships as derived in Eq. 2, i.e.,

$$\begin{aligned} \|f(x_i) - f(x_k)\|_2 &> \|f(x_i) - f(x_j)\|_2 \\ \|f(x_i) - f(x_k)\|_2 &> \|f(x_j) - f(x_k)\|_2 \end{aligned} \tag{3}$$

Eq. 3 suggests that in order to retain the ordinal relationship between individual face images in the learned feature space, the margin used for image pair $\{x_i,\ x_k\}$ should be larger than either the margin used for image pair $\{x_i,\ x_j\}$ or the margin used for image pair $\{x_j,\ x_k\}$. In another word, the margin should NOT be a constant; by contrast, it should change w.r.t. the age difference between a pair of face images. Therefore, in the proposed revised contrastive loss, we determine the margin of an image pair $\{x_i,x_j\}$ based on their age difference, i.e.,

$$\mathscr{M}_{ij} = \alpha(y_i - y_j). \tag{4}$$

where α is a scalar, and we use $\alpha = 1$ by default. Then, when $yi \neq y_j$, the revised contrastive loss ℓr can be computed as

$$\ell r(x_i,x_j) = \frac{1}{2} max(0, \mathscr{M}_{i,j} - \|f(x_i) - f(x_j)\|_2)^2. \tag{5}$$

When $y_i = y_j$, similar to the standard contrastive loss, the two samples in the feature space are expected to be as close as possible

$$\|f(x_i) - f(x_j)\|_2 \to 0. \tag{6}$$

Finally, our revised contrastive loss ℓr can be formulated as

$$\ell r(x_i,x_j) = \begin{cases} \frac{1}{2}\|f(x_i) - f(x_j)\|_2^2, & y_i = y_j \\ \frac{1}{2} max(0, \mathscr{M}_{ij} - \|f(x_i) - f(x_j)\|_2)^2, & y_i > y_j, \end{cases} \tag{7}$$

Compared with the standard contrastive loss, the revised contrastive loss is able to determine the margin according to the age difference between a pair of face images. As a result, the learned feature space is expected to retain the ordinal relationship between individual face images with different ages.

B. Embedding into CNNs

As shown in Fig. 2, we use CNN (e.g., AlexNet [21], VGG [22]) as the feature extraction function f, and embed the proposed revised contrastive loss into CNN to learn features that are discriminative for individual ages. Similar to [23], we also use softmax loss together with our revised contrastive loss to make the convergence of the network stable [23].

In the training phase, given a pair of input images $\{x_i,x_j\}$, we obtain the features $\{f(x_i), f(x_j)\}$ using the CNN network. Then, we could compute the revised contrastive loss ℓr for these two features following Eq. 7. At the same time, for each feature in $\{f(x_i), f(x_j)\}$, we calculate the softmax loss, and get ℓs_i and ℓs_j, respectively.Therefore, the joint loss during training can be written as

$$\begin{aligned} \ell &= \ell s_i + \ell s_j + \lambda \ell r \\ &= \sum_{i=1}^{N} -log p_{iy_i} + \sum_{j=1}^{N} -log p_{jy_j} \\ &+ \frac{\lambda}{2} \sum_{i=1}^{N} \begin{cases} \|f(x_i) - f(x_j)\|_2^2, & y_i = y_j \\ max(0, \alpha(y_i - y_j) - \|f(x_i) - f(x_j)\|_2)^2, & y_i > y_j. \end{cases} \end{aligned} \tag{8}$$

where λ is a parameter balancing the impact of the revised contrastive loss and the softmax loss. We set $\lambda = 0.01$ based on empirical results. We perform SGD [19] to optimize the weights of the network. The gradients of the joint loss used for updating the network weights are calculated as

$$\frac{\partial \ell}{\partial f(x_i)} = \frac{\partial \ell s_i}{\partial f(x_i)} + \lambda \frac{\partial \ell r}{\partial f(x_i)}, \tag{9}$$

and

$$\frac{\partial \ell}{\partial f(x_j)} = \frac{\partial \ell s_j}{\partial f(x_j)} + \lambda \frac{\partial \ell r}{\partial f(x_j)}. \tag{10}$$

The gradients w.r.t. x_i and x_j of the revised contrastive loss in Eqs. 9 and 10 can be discussed in two cases..

If $y_i = y_j$, the gradients are

$$\frac{\partial \ell r}{\partial f(x_i)} = f(x_i) - f(x_j), \tag{11}$$

$$\frac{\partial \ell r}{\partial f(x_j)} = -(f(x_i) - f(x_j)). \tag{12}$$

If $y_i > y_j$ and $\|f(x_i) - f(x_j)\|_2 > \mathscr{M}_{i,j}$, the gradients of the revised contrastive loss are

$$\frac{\partial \ell r}{\partial f(x_i)} = -\frac{(\alpha(y_i - y_j) - \|f(x_i) - f(x_j)\|_2)(f(x_i) - f(x_j))}{\|f(x_i) - f(x_j)\|_2}, \tag{13}$$

and

$$\frac{\partial \ell r}{\partial f(x_j)} = \frac{(\alpha(y_i - y_j) - \|f(x_i) - f(x_j)\|_2)(f(x_i) - f(x_j))}{\|f(x_i) - f(x_j)\|_2}. \tag{14}$$

When $y_i < y_j$, the gradients of the revised contrastive loss are with the same form as Eqs. 13 and 14. We should point out that the two images in the image pair $\{x_i,x_j\}$ go through two

CNN branches, but these two CNN branches share exactly the same network weights.

While the softmax loss in the joint loss makes the training of the model stable, the revised contrastive loss utilizes the ordinal relationship to make the feature space more discriminative for individual ages. In the inference phase, the input is a single face image not a pair of face images. So it only needs to go through one branch of the network to compute the per class probabilities, i.e., $\{p_1, p_2, ..., p_C\}$ using for softmax loss. C is the total number of distinct ages in the training set. The age with the maximum probability is used as the final age estimate for the test face image.

IV. EXPERIMENTS

We provide extensive evaluations of the proposed age estimation approach and comparisons with the state-of-the-art methods on several public-domain face aging databases including MORPH Album II [10], FG-NET [9], and CLAP2016 [11].

A. Datasets

MORPH Album II is one of the largest longitudinal face databases in the public domain, which contains 55,134 face images of 13,617 subjects and the range from 16 to 77 [24]. We use two types of widely used testing protocols in our evaluations. One is the five-fold random split (RS) protocol for all the images [7][6][25][26]; the other is the five-fold subject-exclusive (SE) protocol [14]. The latter testing protocol is more challenging since it assures the images of one subject only appear in one fold.

FG-NET database was a very early database used for age estimation, which contains 1,002 face images from 82 individuals and the ages range from 0 to 69 [9]. We follow a widely used leave-one-person-out (LOPO) protocol [6], [26], [14] in our experiments.

CLAP2016 dataset was released in 2016 at the ChaLearn Looking at people challenge, which contains 4,113, 1,500, and 1,979 face images in the training set, validation set, and testing set, respectively [11]. Different from the MORPH II and FG-NET databases, the ages provided in the CLAP2016 dataset are apparent ages collected via crowdsourcing, so there are a mean age and a variance for each face image.

B. Evaluation Metrics

We report the mean absolute error (MAE) [27] and cumulative score (CS) [13] on the MORPH II and FG-NET databases. MAE is defined as the mean absolute error between the estimated age ($\hat{y}_i$) and ground-truth age (y_i): $MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$, where N is the number of testing images. CS measures the age estimation accuracy given a tolerance of absolute error θ: $CS(\theta) = \frac{\sum_{i=1}^{N}[|y_i-\hat{y}_i| \leq \theta]}{N} \times 100\%$, where $[\cdot]$ equals 1 if the expression is true; otherwise it equals 0.

For the CLAP2016 dataset, we use the ε-error [11] provided with the standard testing protocol $\varepsilon = 1 - \frac{1}{N}\sum_{i=1}^{N} e^{-\frac{(y_i-\mu_i)^2}{2\sigma_i^2}}$, where μ and σ are the ground-truth mean age and standard deviation, respectively.

C. Experiment Settings

We use Caffe [28] to implement age estimation network. We align all the face images based on five facial landmarks detected using an open source SeetaFaceEngine [1], and resize all the face images into 256 × 256 × 3.

Two CNNs, e.g., AlexNet [21] with batch normalization [29] and VGG-16 [22] are used in our approach. Both models are pre-trained on ImageNet 2012 [30]. Besides, the VGG-16 model is also pre-trained using IMDB-WIKI, which is a large scale face database with age and gender labels [26].[2] We use an initial learning rate of 0.001 and a batch size of 64 for both AlexNet and VGG-16, and reduce the learning rate by multiplying 0.1 for every 10 epochs (AlexNet), and 15 epochs (VGG-16). The input face images are randomly cropped to 224×224 and 227×227, respectively. In our experiments, all the pairs or triples are random sampled.

D. Age Estimation Results

Comparisons of Different Losses. To validate the effectiveness of our revised contrastive loss, we first compare it with the traditional softmax loss and the joint of softmax loss and contrastive loss, triplet loss by performing age estimations on the MORPH II database using both the RS and SE protocols. The MAE and CS (θ = 5) of the four different losses are shown in Table I. We can see that jointly using softmax loss and contrastive loss, triplet loss outperforms using softmax loss alone. This is reasonable because although the traditional contrastive and triplet loss does not enforce the ordinal relationship during feature learning, it still considers the pair-wise relationship between samples, and leads to more robust features. However, joint use of softmax loss and the revised contrastive loss performs the best. This shows that retaining the ordinal relationship is helpful for learning more robust features in age estimation tasks.

Loss	RS		SE	
	MAE	CS(θ=5)	MAE	CS(θ=5)
Softmax loss	3.324	78.96%	4.043	73.24%
Softmax and contrastive loss	3.219	80.12%	3.832	75.33%
Softmax and triplet loss	3.249	80.48%	3.750	76.21%
Softmax and Proposed loss	**3.075**	**81.83%**	**3.513**	**78.57%**

TABLE I
COMPARISONS OF THE AGE ESTIMATION MAEs AND CS(5) BY DIFFERENT LOSSES ON THE MORPH II DATABASE.

Comparisons with the State-of-the-art. We then provide comparisons with the state-of-the-art age estimation methods

[1]https://github.com/seetaface/SeetaFaceEngine

[2]This database is large, but the age labels can be quite noisy because they are calculated based on the date of birth of the public figures and the timestamps of the photos crawled from the Internet.

on the MORPH II, FG-NET, and CLAP2016 databases. A number of the state-of-the-art methods are used for comparisons such as Rank-CNN [7], DEX [26], RED-SVM [31], and DIF [14] and so on. Table. II shows the MAEs on the MORPH II and FG-NET datasets with RS and SE protocols. The results suggest that ranking-based methods, such as [7][6][25] perform better than classification or regression-based methods [26][31]. This is reasonable because the ordinal relationship is considered by ranking-based methods, which improves the robustness of the age estimation models. The proposed method performs the best among all the approaches, because our method not only considers the ordinal relationship, but also explicitly quantizes the distance between an ordinal image pair. We also calculate the MAEs for males and females on the MORPH II, which shows that the MAE of males is lower, because of the imbalance of the distribution of gender. Even if the MAE for females is larger, the MAE of female is close to the overall performance of state-of-the-arts.

Method	MORPH II		FG-NET	
	MAE	Protocol	MAE	Protocol
Rank-CNN [7]	2.96	RS	-	-
OHRank [6]	6.07	RS	4.48	LOPO
DIF [14]	3.60	SE	3.80	LOPO
OR-CNN [25]	3.27	RS	-	-
DEX [26]	3.25	RS	4.63	LOPO
RED-SVM [31]	6.49	RS	5.24	LOPO
Ours	**2.46/2.88**	RS/SE	**4.21**	LOPO
male/female	2.31/3.22	RS	-	-
	2.73/3.61	SE	-	-

TABLE II
COMPARISONS OF THE AGE ESTIMATION MAES BY THE PROPOSED APPROACH AND THE STAGE-OF-THE-ART METHODS ON THE FG-NET AND MORPH II DATABASES.

Fig. 3 show the entire CS curves on the FG-NET and MORPH II databases using the LOPO and RS protocols, respectively. The proposed method achieves higher accuracy than all the baseline methods on average.

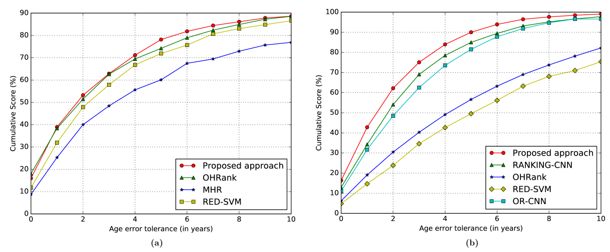

Fig. 3. Age estimation cumulative scores by the proposed approach and the state-of-the-art methods on (a) FG-NET with a LOPO protocol, and MORPH II (b) with a random split (RS) protocol.

On the CLAP2016, our age estimation method achieves an ε-error of 0.3171, which is the best result among single model age estimation methods (see Table. III). The first place method in the CLAP competition reported a lower error, but they used

Rank	Team Name	ε-error	Single model?
1	OrangeLabs [32]	0.2411	NO
*	**Ours**	**0.3171**	**YES**
2	Palm_seu	0.3214	NO
3	CMP+ETH	0.3361	NO
4	WYU_CVL	0.3405	NO
5	ITU_SiMiT	0.3668	NO
6	Bogazici	0.3740	NO
7	MIPAL_SNU	0.4565	NO
8	DeepAge	0.4573	YES

TABLE III
COMPARISONS OF THE ε-ERROR BY THE PROPOSED APPROACH AND STAGE-OF-THE-ART METHODS ON THE CLAP2016 DATABASE. THE RESULTS OF STATE-OF-THE-ART METHODS ARE FROM [11].

the score-level fusion of multiple CNNs. Such a fusion is likely to have a larger memory and computational costs than our method. In addition, they collected some children' face images from the Internet to train their age estimation models.

Fig. 4 shows some age estimation results by the proposed approach on the MORPH II, FG-NET, and CLAP2016 databases. We observe that the proposed approach is robust to most of the common facial appearance variations such as small poses, partial occlusions, and expressions. The age estimation accuracy may decrease when the face images have very bad illumination, partial occlusion, and blurring (see the bottom row of Fig. 4).

V. CONCLUSION

In this paper, we proposed revised contrastive loss to utilize the ordinal relationship in a pair of face images. By jointly use our revised contrastive loss and softmax loss, an explicit ordinal constraint is introduced to achieve more robust feature representations than using softmax or the traditional contrastive loss alone. Experimental results on the MORPH II, FG-NET, and CLAP2016 databases show that the proposed approach outperforms the state-of-the-art age estimation approaches by a large margin. In our future work, we would like to generalize the proposed approach towards cross-database testing scenarios. In addition, age estimation from imbalanced data will also be studied.

VI. ACKNOWLEDGMENT

This research was supported in part by the National Basic Research Program of China (grant 2015CB351802), Natural Science Foundation of China (grants 61732004, 61672496, and 61650202), and Youth Innovation Promotion Association CAS (2018135).

REFERENCES

[1] A. Lanitis, C. Draganova, and C. Christodoulou, "Comparing different classifiers for automatic age estimation," *IEEE Trans. Syst. Man, Cybern. B, Cybern*, vol. 31, no. 1, pp. 621–628, Jan. 2004.

[2] Z. Yang and H. Ai, "Demographic classification with local binary patterns," in *Proc. ICB*, 2007.

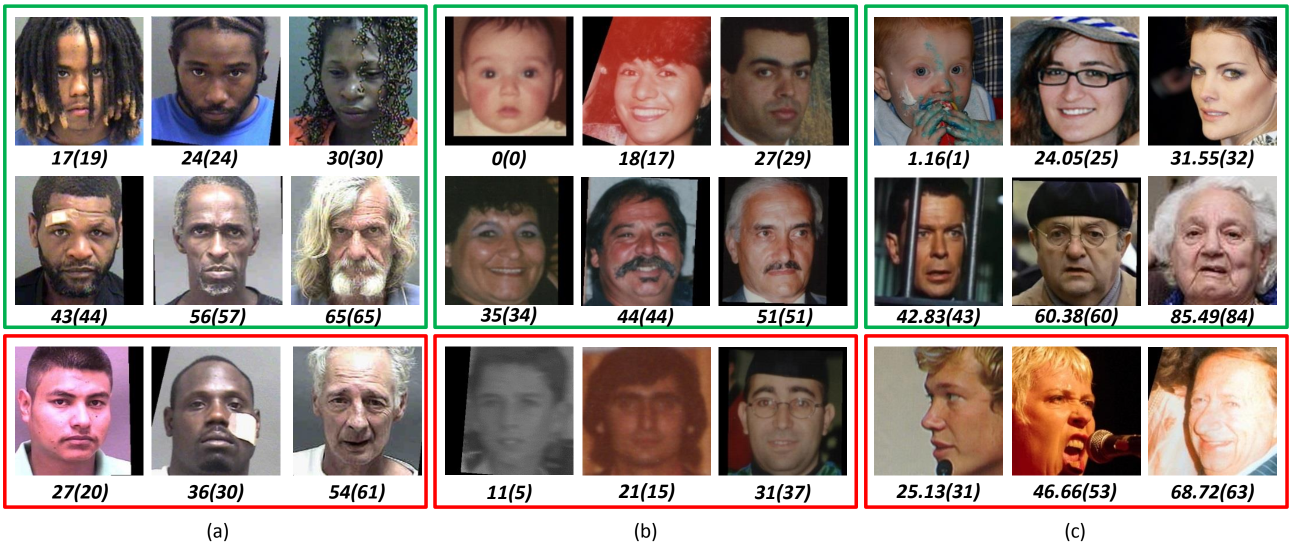

Fig. 4. Examples of age estimation results by the proposed approach on the (a) MORPH II, (b) FG-NET and (c) CLAP2016. The top two rows show good age estimation results, and the third row shows poor age estimation results. The numbers below each image show the real age and predicted age of the subject, i.e., real age (predicted age).

[3] G. Guo, Y. Fu, C. Dyer, and T. Huang, "Image-based human age estimation by manifold learning and locally adjusted robust regression," *IEEE Trans. Image Process*, vol. 17, no. 7, pp. 1178–1188, May 2008.

[4] Y. Fu and T. Huang, "Human age estimation with regression discriminative aging manifold," *IEEE Trans. Multimedia*, vol. 10, no. 4, pp. 578–584, May 2008.

[5] B. Xiao, X. Yang, H. Zha, Y. Xu, and T. Huang, "Metric learning for regression problems and human age estimation," in *Proc. PCM*, 2009.

[6] K. Y. Chang, C. S. Chen, and Y. P. Hung, "Ordinal hyperplanes ranker with cost sensitivities for age estimation," in *Proc. CVPR*, 2011, pp. 585–592.

[7] S. Chen, C. Zhang, M. Dong, J. Le, and M. Rao, "Using ranking-cnn for age estimation," in *Proc. CVPR*, 2017.

[8] D. Parikh and K. Grauman, "Relative attributes," in *Proc. ICCV*, 2011, pp. 503–510.

[9] G. Panis, A. Lanitis, A. Lanitis, and T. F. Cootes, "Overview of research on facial ageing using the FG-NET ageing database," *IET Biom.*, vol. 5, no. 2, pp. 37–46, May 2015.

[10] K. R. Jr. and T. Tesafaye, "MORPH: A longitudinal image database of normal adult age-progression," in *Proc. FG*, 2006.

[11] S. Escalera, M. Torres, B. Martinez, X. Bar, H. J. Escalante, I. Guyon, I. Guyon, I. Guyon, M. Oliu, M. A. Bagheri, and M. A. Bagheri, "Chalearn looking at people and faces of the world: Face analysis workshop and challenge 2016," in *Proc. CVPR Workshops*, 2016.

[12] Y. Kwon and N. Lobo, "Age classification from facial images," in *Proc. CVPR*, 1994, pp. 762–767.

[13] G. Guo, G. Mu, Y. Fu, and T. Huang, "Human age estimation using bio-inspired features," in *Proc. CVPR*, 2009, pp. 112–119.

[14] H. Han, C. Otto, X. Liu, and A. Jain, "Demographic estimation from face images:human vs. machine performance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 6, pp. 1148–1161, Jun. 2015.

[15] D. Yi, Z. Lei, and S. Z. Li, "Age estimation by multi-scale convolutional network," in *Proc. ACCV*, 2015, pp. 144–158.

[16] X. Yang, B. Gao, C. Xing, Z. Huo, X. Wei, Y. Zhou, J. Wu, and X. Gen, "Deep label distribution learning for apparent age estimation," in *Proc. ICCV Workshops*, 2015, pp. 102–108.

[17] G. Wang and D. Forsyth, "Joint learning of visual attributes, object classes and visual saliency," in *Proc. ICCV*, 2009.

[18] A. Datta, R. Feris, and D. Vaquero, "Hierarchical ranking of facial attributes," in *Proc. FG*, 2011, pp. 36–42.

[19] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[20] R. Hadsell, R. Hadsell, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proc. CVPR*, 2006.

[21] A. Krizhevsky, S. Ilya, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. NIPS*, 2012, pp. 1097–1105.

[22] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, 2015.

[23] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," in *Proc. NIPS*, 2014.

[24] A. Lanitis, C. J. Taylor, and T. F. Cootes, "Toward automatic simulation of aging effects on face images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 4, pp. 442–455, Aug. 2002.

[25] Z. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua, "Ordinal regression with multiple output CNN for age estimation," in *Proc. CVPR*, 2016.

[26] R. Rothe, R. Timofte, and L. V. Gool, "Deep expectation of real and apparent age from a single image without facial landmarks," *Int. J. Comput. Vis.*, pp. 1–14, Aug. 2016.

[27] Z. Huo, X. Yang, C. Xing, Y. Zhou, P. Hou, J. Lv, and X. Geng, "Deep age distribution learning for apparent age estimation," in *Proc. CVPR Workshops*, 2016.

[28] J. Yangqing, S. Evan, D. Jeff, K. Sergey, L. Jonathan, G. Ross, G. Sergio, and D. Trevor, "Caffe: Convolutional architecture for fast feature embedding," *arXiv preprint arXiv:1408.5093*, 2014.

[29] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducinginternal covariate shift," *arXiv:1502.03167v3*, 2015.

[30] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and F. Li, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[31] K. Y. Chang, C. S. Chen, and Y. P. Hung, "A ranking approach for human age estimation based on face images," in *Proc. ICPR*, 2010.

[32] G. Antipov, M. Baccouche, S. Berrani, and J. Dugelay, "Apparent age estimation from face images combining general and children-specialized deep learning models," in *Proc. CVPR Workshops*, 2016.