LEARNING NEIGHBORHOOD-REASONING LABEL DISTRIBUTION (NRLD) FOR FACIAL AGE ESTIMATION

Zongyong Deng[1,*], *Mo Zhao*[1,*], *Hao Liu*[1,2], *Zhenhua Yu*[1,2], *Feng Feng*[1,2]

[1]School of Information Engineering, Ningxia University, Yinchuan, China
[2]Collaborative Innovation Center for Ningxia Big Data and Artificial Intelligence
Co-founded by Ningxia Municipality and Ministry of Education, Yinchuan, China
{zongyongdeng_nxu, zhaomo_nxu}@outlook.com {liuhao, zhyu, feng_f}@nxu.edu.cn

ABSTRACT

In this paper, we propose to learn a neighborhood-reasoning label distribution (NRLD) for facial age estimation. Unlike conventional label distribution methods with fixed-structural aging patterns, in this work, our NRLD aims to reason about more resilient and adaptive label distribution by disentangling the graph of face neighbors. In particular, our model holds the assumption on that the sample-specific age label distribution is principally influenced by a mixture of interpretable and meaningful factors, which typically cause plausible edges connected to the anchors. Under the scenario of each factor, we specifically collect the subset of graph edges and then convolute them with face samples to regress a mean-variance label distribution. During the training process, the mixture hyperparameters of our label distribution are iteratively optimized by following the Expectation-Maximization schema. Extensive experimental results on three challenging widely-evaluated datasets indicate the superiority in comparisons with most state of the arts.

Index Terms— Facial age estimation, label distribution learning, subspace learning, causal learning

1. INTRODUCTION

Human age estimation, aiming at predicting the exact biological age values based on the given facial images, plays a vital role in many applications of visual attribute analysis [1, 2]. Recently, although efforts have been devoted to age estimation [3–5], the performance still remains limited especially under such challenging cases when face samples undergo large variations including diverse expressions, cross populations and genders, partial occlusions, etc. This is mainly due to two-fold reasons. On one hand, the relationship of face data and age labels is usually complexly heterogeneous and nonlinear [5, 6]. On the other hand, the neighboring age labels undergo ambiguity and correlation [7, 8], since ages are usually arranged as an order in practice. Hence, this urgently motivates us to propose robust and accurate facial age estimation particularly versus unconstrained environments.

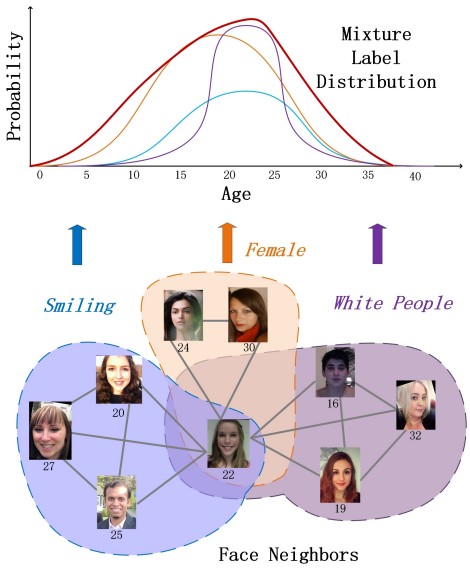

Fig. 1. Demonstration of Our Insight. Our model aims to learn a mixture of label distributions which are actually caused by semantic factors (*e.g.*, expressions, gender, ethnicity, etc.) based on the face neighborhood space. It significantly exploits flexible and real-world facial age label patterns for robust age estimation. Best viewed in the color pdf file.

Conventional age estimation methods could be roughly categorized into two major ingredients: feature representation and age predictor. Feature representation-based methods [10–12] aim to seek discriminative feature descriptors for ages based on the face images. Respectively, age predictor-

∗ indicates equal contribution, the corresponding author is Hao Liu. This work was supported in part by the Scientific Research Projects of Colleges and Universities of Ningxia under Grant NGY2018050, in part by the National Science Foundation of China under Grant 61806104, in part by the Natural Science Foundation of Ningxia under Grant 2018AAC03035, and in part by the Youth Science and Technology Talents Enrollment Projects of Ningxia under Grant TJGC2018028.

978-1-7281-1331-9/20/$31.00

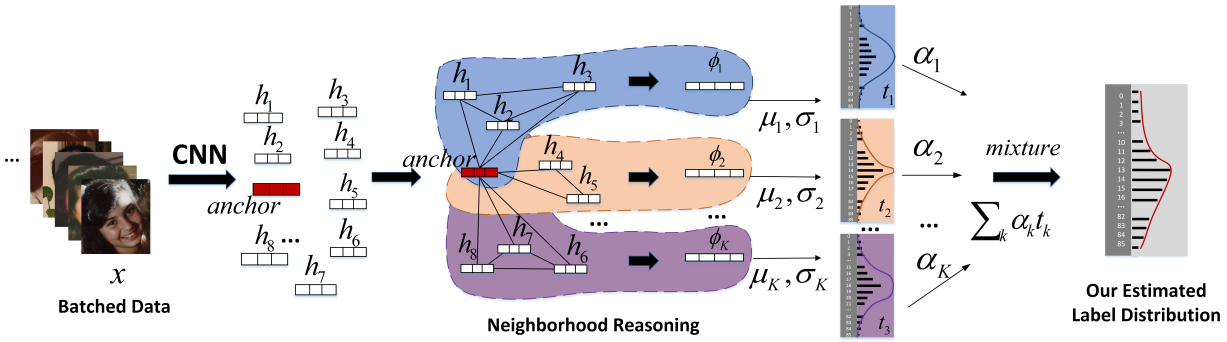

Fig. 2. Flowchart of Our NRLD. Our architecture starts with the inputting faces feeding to the prelearned CNN. Having obtained these deep features, we first construct a neighborhood graph based on the appearance similarity with all possible pairs. Then our model reasons out K folds of plausible edges within the graph, thus giving rise to K disentangled routings with respect to the K latent factors. Under each factor, we aggregate the features ϕ in this routine and project them to a specific mean-variance label distribution $\boldsymbol{t}$. Lastly, we learn to fuse with these latent distributions and estimate the flexible and adaptive label distribution for age estimation. The network parameters are optimized via back-propagation. In the meantime, the mixture hyperparamter of mean-variance label distributions are iteratively by the widely-employed Expectation-Maximization optimization [9].

based methods [13, 14] basically learn to classify the age ranker based on the input feature representation. However, both procedures of seeking feature representation and learning age predictors are separately optimized, which is likely trapped into sub-optima and thus leads to limited generation capacity particularly when the target sample is quite discrepant from the training ones. To circumvent this limitation, deep learning has been adopted recently [4, 8], which allows the training of an end-to-end system that attempts to alleviate the above drawbacks. Apart from that, label distribution has been emerged as the widely-employed and state-of-the-art methods such as [7, 15, 16]. The algorithm typically encodes a range of age labels to a symmetrical distribution, *e.g.*, Gaussian or triangle distribution, reflecting the smoothness and correlation for high-performance age estimation. Nevertheless, they are constrained to take only fixed-structural form to model the ambiguous properties of age labels, which is usually non-robust to complex cross-population face data domains and even hardly explainable. Therefore, we propose a flexible label distribution learning approach for age estimation which is able to address the aforementioned issues.

In this paper, we propose a neighborhood-reasoning label distribution learning method, dubbed NRLD, typically modeling the heterogeneous face aging data for robust facial age estimation. In comparison to the conventional label distribution approaches taking a fixed and inflexible form, we intend to infer a sample-specific and adaptive label distribution, which flexibly introduces meaningful and semantic variations in the cross-population aging pattern. In this paper, we argue that the real-world age label distribution should be explicitly disentangled as a mixture of Gaussian-like distributions. Moreover, our label distribution is relaxed in a non-symmetric but convex form. Hence, the learned distributions are reasoned by a set of interpretable latent factors, *e.g.*, facial expressions, human genders and populations, as viewed in Fig. 1. Technically, we first construct the neighborhood graph adjacent with each anchored sample based on the facial appearance information. Then our model reasons out different routings on the neighborhood graph, where each routing attributes to one mean-variance distribution for practical aging pattern. Our proposed module is readily to be plugged in the modern very deep feature extractor, *e.g.*, VGG-Face Net [17] for efficient feature learning. During training process, we make inferences for the hyperparameters of the mixture distribution iteratively with the Expectation-Maximization optimization method [18]. Fig. 2 illustrates the flowchart. To further evaluate the effectiveness of our proposed method, we conduct extensive experiments on three in-the-wild datasets, which significantly show the superior performance compared with existing facial age estimation methods.

2. METHODOLOGY

In this section, we present a detailed description of our problem formulation, the proposed NRLD model and finally its alternatively associated optimization procedure.

Problem Formulation. Let $X = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$ be the training set which contains N samples in total, where $\boldsymbol{x}_i \in \mathbb{R}^d$ specifies the i-th face sample consisting of $W \times H$ pixels. As demonstrated in Fig. 2, our architecture starts with batches of raw face samples. These batches are then fed to the CNN feature extractor, VGG-Face Net [17], thus performing the immediate deep feature representation $\boldsymbol{h}$ as follows:

$$\boldsymbol{h} = \mathrm{CNN}(\boldsymbol{x}), \tag{1}$$

where the $\mathrm{CNN}(\cdot)$ function is integrated with a sequence of

convolution operations, nonlinear *ReLU* functions and fully connections and the parameters are sequentially adjusted during the modeling learning phase. Note that we employ the modern deep architecture, which is learned by amounts of face samples with personal identifies, making robust initialization and fast model convergence for network parameters.

Based on one batch of these extracted deep feature $\boldsymbol{h}$s, our method reasons about a topological graph with face neighbors. For a clear clarification on the graph-based notations, we let $G = \{V, E\}$ denote the constructed graph, integrated with all possible pairs (m, n) of the face samples in the input batch. We assume m is fixed as the anchored sample, which is associated with other samples denoted by n. Consequently, all nodes in the graph are subjected to the relation $(m, n) \in E$. It means that there exist more than one edges from anchor m to the neighboring face n. One possible manner is to learn data-dependent label distribution from the simple neighborhood graph [16], which likely ignores the intrinsic various facial attributes and performs the black-box predictions for pseudo age label patterns.

To address the above issue, we focus on disentangling the neighborhood graph in K channels with an explainable way, where each channel can extract different feature ϕ principally determined by the latent factors. We let $\mathcal{N}(\cdot)$ to specify the core function of our approach, which aims to estimate the specific mean-variance label distributions by exploiting the disentangled neighborhood information. Hence, our objective can be formulated as minimizing the following optimization:

$$\min_{\mathcal{N}} \sum_{i=1}^{N} \|\mathcal{N}(\boldsymbol{x}_i) - y_i^*\|. \tag{2}$$

Obviously, the core step in the formulation (2) mainly lies on learning the parameters of $\mathcal{N}(\cdot)$, typically disentangling the neighborhood under different latent factors in a fully-supervised manner.

Our architecture interacts with all possible graph edges and then results in K plausible routings where each routing only selects a subset of plausible connections. With these connected relations, our function $\mathcal{N}(\cdot)$ disentangles K factors based on the whole neighborhood graph, thus leading to K aggregated features $\boldsymbol{\phi}_k$ (we ignore the anchor index) where k is valued from 1 up to K. By doing this, we derive the feature $\boldsymbol{\phi}$ into K parts: $\boldsymbol{\phi} = [\phi_1, \cdots, \phi_k, \cdots, \phi_K]$. From another perspective, we view the neighborhood graph into K subspaces where each subspace provides semantic and meaningful factor. Under the scenario of each factor, we project the aggregated feature denoted ϕ to one specific Gaussian-like label distribution by $\boldsymbol{t}$. Ultimately, we specify mixture hyperparameter by a set of $\boldsymbol{\alpha}$ and our architecture will estimate these αs to fuse the estimated distributions as a flexible age label distribution.

Model Learning. In our approach, we assume that the real-world age label distribution is associated with the factorized face neighborhood containing diverse semantic facial attributes. Based on this sense, our model seeks folds of latent factors to be disentangled, which cause latent label distributions. Given the constructed graph G with face neighbors (m, n), we develop a feature aggregation schema denoted by $f(\cdot)$ (assumed to be anchored by m) to perform $\boldsymbol{\phi}_m$:

$$\boldsymbol{\phi}_m = f(\boldsymbol{h}_m, \{\boldsymbol{h}_n \in \mathbb{R}^{d_{in}} : (m, n) \in E\}), \tag{3}$$

where the dimension of d_{in} is consistent with the outputting dimension of the employed deep CNN feature extractor.

In line with these features $\boldsymbol{\phi}$ to be disentangled with K latent factors, we further learn K specific Gaussian label distributions denoted by $\{\boldsymbol{t}_k\}_{k=1}^K$ which are parameterized by $\{(\mu_k, \sigma_k)\}_{k=1}^K$. The disentangled label distributions $\boldsymbol{t}$ for the k-th routing is computed as follows:

$$\boldsymbol{t}_k = \text{MLP}(\frac{\varphi(\boldsymbol{W}_k^T \phi_k + \boldsymbol{b}_k)}{\|\varphi(\boldsymbol{W}_k^T \phi_k + \boldsymbol{b}_k)\|_2}), \tag{4}$$

where $\{\boldsymbol{W}_k, \boldsymbol{b}_k\}$ specifies the parameters representing K latent factors to be disentangled, $\text{MLP}(\cdot)$ provides the stacking layers for deeper configuration, and $\varphi(\cdot)$ is the nonlinear function, respectively. In parallel, we literately estimate the mixture coefficient $\boldsymbol{\alpha}$ of these latent label distributions and finally perform the mixture distribution as

$$\hat{\boldsymbol{y}} = \sum_{k=1}^{K} \alpha_k \boldsymbol{t}_k. \tag{5}$$

With the learned mixture distribution, we reason out the disentangled age label distribution under different explainable factors. Each factor facilitates the discriminative neighborhood information, thus making reinforced robust age label pattern particularly regarding with faces captured in wild conditions. (refer to visualization in Section 3)

To efficiently optimize (5), we maximize the likelihood function with respect to the parameters including the mean-variance coefficients and the hyperparameters via EM schema [9]. In detail, we first initialize μ_p and σ_p by the feedforward execution function in (4) and all mixing hyperparameters are assigned to $\frac{1}{K}$.

For *E-step*, we compute the current parameters by using the following responsibilities:

$$\gamma\left(z_{np}\right) = \frac{\alpha_p \boldsymbol{t}\left(\boldsymbol{x}_n | \mu_p, \sigma_p\right)}{C}, \tag{6}$$

where $C = \sum_{j=1}^{P} \alpha_j \boldsymbol{t}\left(\boldsymbol{x}_n | \mu_j, \sigma_j\right)$ means a constant variable.

Accordingly, for *M-step*, we re-estimate other parameters by evaluating those responsibilities as follows:

$$\mu_p^{new} = \frac{1}{N_p} \sum_{n=1}^{N} \gamma\left(z_{np}\right) x_n, \tag{7}$$

$$\sigma_p^{new} = \frac{1}{N_p} \sum_{n=1}^{N} \gamma\left(z_{np}\right) \left(x_n - \mu_p^{new}\right) \left(x_n - \mu_p^{new}\right)^\top, \tag{8}$$

$$\alpha_p^{new} = \frac{N_p}{N}. \quad (9)$$

The above EM optimization checks for the convergence of all parameters of iterations. During inference stage, we first estimate the sample-specific age label distribution in the testing set. Then we search the maximal age value based on the probability over every dimension in the projected label distribution.

3. EXPERIMENTS

To evaluate the effectiveness of the proposed method, we conducted results on three widely-used datasets including MORPH [19], FGNET [1] and ChaLearn [20]. The face samples in these datasets usually undergo challenging cases due to wild conditions. To further reinforce the performance, we incorporated one large scale face aging database IMDB-WIKI [21] for model pretraining.

Datasets. *MORPH [19]:* This dataset standardizes different face aging databases, which mainly contains 55608 facial images with 13000 identities with various races. All samples with age annotations are particularly in the range from 16 to 77 years old. Averagely, each personal identity has at least six samples. For fair comparisons, we strictly followed the experimental settings employed in [5].

FGNET [1]: This dataset collects 1002 face images with 82 personal identities in total. The age range of this dataset covers from 0 to 69 years old. Due to sparse and imbalance of this dataset, we only have about 12 samples captured in unconstrained environments due to facial aspect ratios, illumination and diverse expressions. For fair evaluation setting, we employed the leave-one-person-out (LOPO) protocol by following [14].

ChaLearn [20]: This dataset collects 4699 images from Internet. For its age annotation, the competition organizer recruited around ten volunteers to manually label the appearance ages. Hence, the final ground-truth for each image incorporates with the mean value and the standard derivation. For experiment setting, we utilized 2476 images for training. Note that, we only use 1136 images for validation by following the standard dataset configuration [20].

IMDB-WIKI [21]: This dataset contains more than half a billion labeled images of celebrities, which are crawled from IMDB and Wikipedia. Most of the images contain types of noise, so it is not suitable for evaluation. However, it is still a good choice to use this dataset for pretraining. We selected about 200 thousand images by following the database setting [21] to pretrain our network.

Evaluation Metric. In the experiments, we leveraged the mean absolute error (MAE) [10] which computes the discrepancy between the estimated age values and the groundtruths. Obviously, the lower the MAE value, the better performance it achieves. We also used the Gaussian error to evaluate on the ChaLearn dataset by following [20]. In particular, the Gaussian error is computed via the following equation: $1-\sum_{i=1}^{N} \exp\left(-\frac{(\hat{y}_i-y_i^*)^2}{2\sigma^{*2}}\right)$, where $\hat{y}$ is the predicted age value, y^* is labeled mean age apparent age, σ^* is the annotated standard deviation and N specifies the number of testing samples, respectively.

Table 1. Comparisons of MAEs of our approach compared with different state-of-the-art methods on MORPH dataset. We achieve the best performance compared with others.

Method	MAE	Year
BIF+KNN	9.64	-
LDL [15]	5.69	2013
CPNN [15]	5.67	2013
CS-LBMFL [12]	4.37	2015
ODFL [3]	3.12	2017
M-LSDML [8]	2.89	2018
SADAL [22]	2.75	2019
BridgeNet [5] †	2.38	2019
NRLD	**2.35**	**-**
NRLD †	**1.81**	**-**

Table 2. The results on MORPH dataset. The performance of two different settings and their average are reported. Our method achieves the state-of-the-art performance.

Method	Train	Test	MAE	Avg	Year
BIF+KCCA	S1	S2+S3	4.00	3.98	2013
	S2	S1+S3	3.95		
DRFs [6]	S1	S2+S3	-	2.98	2018
	S2	S1+S3	-		
BridgeNet [5] †	S1	S2+S3	2.74	2.63	2019
	S2	S1+S3	2.51		
NRLD	**S1**	**S2+S3**	**2.48**	**2.47**	**-**
	S2	**S1+S3**	**2.46**		
NRLD †	**S1**	**S2+S3**	**2.35**	**2.34**	**-**
	S2	**S1+S3**	**2.33**		

Implementation Details. For each input image, we first detected the whole face with MTCNN [23]. Then we aligned it based on the detected facial landmarks. We augmented all images with horizontal flipping. We employed the VGG Face Net as the backbone network and modified the last four layers with our proposed disentangle module. On the MORPH and FGNET datasets, the initial learning rate of CNN part was set to 0.001. To accelerate training convergence, the initial learning rate of the disentangle module was set to 0.01. Notably, for the ChaLearn dataset, the learning rate was reduced by ten times to avoid overfitting. We additionally prelearned the network by IMDB-WIKI [21] data. Note that we used † as the final performance with auxiliary IMDB-WIKI dataset.

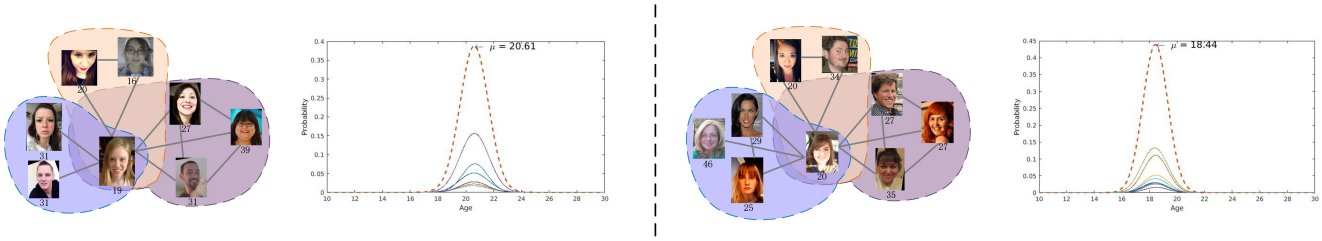

Fig. 3. Examples of face images and NRLD results. For each anchor face image, we obtained its neighbors based on appearance similarity and then reasoned out K (here we set to 8) different label distributions with respect to the factors. Here, we only visualized three factors thus leading to a better clarification. Finally, we fused these latent distributions and estimated the flexible and adaptive label distribution for age estimation. (Best viewed on a monitor when zoomed in.)

Table 3. Comparisons of MAEs of our approach compared with state-of-the-art methods on FGNET dataset. Our NRLD yields compelling performance compared with BridgeNet [5].

Method	MAE	Year
BIF+KNN	8.24	-
OHRanker [14]	4.48	2011
LDL [15]	5.77	2013
CPNN [15]	4.76	2013
CS-LBMFL [12]	4.36	2015
M-LSDML [8]	3.74	2018
SADAL [22]	3.67	2019
BridgeNet [5] †	2.56	2019
NRLD	**3.23**	**-**
NRLD †	**2.55**	**-**

Table 4. Comparisons of MAEs and Gaussian errors of our NRLD compared with state-of-the-art methods on ChaLearn. It can be seen that our approach outperforms existing models.

Method	MAE	Gaussian Error	Year
BIF+KNN	7.19	0.620	-
l_2 Regression	5.05	0.456	-
Han *et al.* [24]	5.2	0.449	2018
DEX [21]	3.44	0.299	2018
DLDL-v2 [25]	3.14	0.272	2018
ODL [3]	3.95	0.312	2019
BridgeNet [5] †	2.98	0.26	2019
NRLD	**3.64**	**0.312**	**-**
NRLD †	**2.78**	**0.233**	**-**

Comparisons with State-of-the-art Methods. We compared our approach with folds of state-of-the-art methods. We carefully conducted the experiments of the state-of-the-art methods [12, 14, 21] by their released source codes. For other methods, we strictly reported their performance by cropping the results from the original papers. Besides, we carefully followed the settings in [5] and [3].

Table 1, Table 2 show the MAEs of our approach on MORPH dataset with different settings and Table 3 shows the results on FGNET dataset, respectively. From the results, we see that our model achieves competitive performance compared with the state-of-the-art methods and even achieves better performance than recent label distribution learning methods. Moreover, we made three-fold conclusion: (1) The traditional methods such as DEX [21] and ODFL [3], treat each age label independently without taking ordinal relation into account. However, our label distribution learning method reconsiders the age labels by introducing the correlation information across a set of adjacent age labels. Consequently, the label distribution accurately and flexibly simulates the real-world age patterns. (2) Unfortunately, some label distribution learning methods such as LDL [15] and CPNN [15] which only enforce a fixed-structural pattern on the age label distribution (*i.e.*, "Gaussian" or "Triangle"), likely result in inflexibly adaptive to real-world face aging data. Benefiting from the data-dependent manner, our method captures more semantic information and achieves diverse and explainable distributions. (3) Particularly from the results on FGNET, we see that our NRLD outperforms most state of the arts even with limited training data. This achievement is due to that the factorized neighborhood space complements various age-related semantic information to reinforce our label distribution. Thus, our model slightly relies on large-scale data. Besides, we conducted experiment on the very challenging ChaLearn. Table 4 tabulates the MAEs and Gaussian errors compared with the state-of-the-arts on the ChaLearn dataset. From the results, we figure out that our NRLD achieves comparable performance in contrast to other methods, which also reflects the effectiveness of our approach.

Visualization of our mixture distributions. To better demonstrate the insight of our NRLD intuitively, we visualized some examples of our learned distributions. As shown in Fig 3, we observed that the learned label distributions based on the face neighborhood space are various. The factors for

being a neighbor are expressions, gender, ethnicity and so on. These examples indicate that most reasoned factors contribute to the diversity of age patterns and make the sample-specific age label distribution better fit the real-world face data.

4. CONCLUSIONS

In this paper, we have proposed a neighborhood-reasoning label distribution (NRLD) learning for facial age estimation. The proposed NRLD has explicitly illustrated that the sample-specific age label distribution is principally influenced by a mixture of interpretable and meaningful factors. Experiments on three datasets have shown the effectiveness of the proposed approach. In the future work, we will focus on few-shot learning to resolve the class imbalance.

5. REFERENCES

[1] Andreas Lanitis, Christopher J. Taylor, and Timothy F. Cootes, "Toward automatic simulation of aging effects on face images," *TPAMI*, vol. 24, no. 4, pp. 442–455, 2002.

[2] Xiangbo Shu, Jinhui Tang, Hanjiang Lai, Luoqi Liu, and Shuicheng Yan, "Personalized age progression with aging dictionary," in *ICCV*, 2015, pp. 3970–3978.

[3] Hao Liu, Jiwen Lu, Jianjiang Feng, and Jie Zhou, "Ordinal deep learning for facial age estimation," *TCSVT*, vol. 29, no. 2, pp. 486–501, 2019.

[4] K. Li, J. Xing, C. Su, W. Hu, Y. Zhang, and S. Maybank, "Deep cost-sensitive and order-preserving feature learning for cross-population age estimation," in *CVPR*, 2018, pp. 399–408.

[5] Wanhua Li, Jiwen Lu, Jianjiang Feng, Chunjing Xu, Jie Zhou, and Qi Tian, "Bridgenet: A continuity-aware probabilistic network for age estimation," in *CVPR*, 2019.

[6] Wei Shen, Yilu Guo, Yan Wang, Kai Zhao, Bo Wang, and Alan L. Yuille, "Deep regression forests for age estimation," in *CVPR*, 2018, pp. 2304–2313.

[7] Bin-Bin Gao, Chao Xing, Chen-Wei Xie, Jianxin Wu, and Xin Geng, "Deep label distribution learning with label ambiguity," *TIP*, vol. 26, no. 6, pp. 2825–2838, 2017.

[8] H. Liu, J. Lu, J. Feng, and J. Zhou, "Label-sensitive deep metric learning for facial age estimation," *TIFS*, vol. 13, no. 2, pp. 292–305, 2018.

[9] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B*, vol. 39, no. 1, pp. 1–38, 1977.

[10] Xin Geng, Zhi-Hua Zhou, and Kate Smith-Miles, "Automatic age estimation based on facial aging patterns," *TPAMI*, vol. 29, no. 12, pp. 2234–2240, 2007.

[11] Yun Fu, Guodong Guo, and Thomas S. Huang, "Age synthesis and estimation via faces: A survey," *TPAMI*, vol. 32, no. 11, pp. 1955–1976, 2010.

[12] Jiwen Lu, Venice Erin Liong, and Jie Zhou, "Cost-sensitive local binary feature learning for facial age estimation," *TIP*, vol. 24, no. 12, pp. 5356–5368, 2015.

[13] Yu Zhang and Dit-Yan Yeung, "Multi-task warped gaussian process for personalized age estimation," in *CVPR*, 2010, pp. 2622–2629.

[14] Kuang-Yu Chang, Chu-Song Chen, and Yi-Ping Hung, "Ordinal hyperplanes ranker with cost sensitivities for age estimation," in *CVPR*, 2011, pp. 585–592.

[15] Xin Geng, Chao Yin, and Zhi-Hua Zhou, "Facial age estimation by learning from label distributions," *TPAMI*, vol. 35, no. 10, pp. 2401–2412, 2013.

[16] Z. He, X. Li, Z. Zhang, F. Wu, X. Geng, Y. Zhang, M.-H. Yang, and Y. Zhuang, "Data-dependent label distribution learning for age estimation," *TIP*, vol. 26, no. 8, pp. 3846–3858, 2017.

[17] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *BMVC*, 2015.

[18] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977.

[19] Karl Ricanek Jr. and Tamirat Tesafaye, "MORPH: A longitudinal image database of normal adult age-progression," in *FG*, 2006, pp. 341–345.

[20] Sergio Escalera, Junior Fabian, Pablo Pardo, Xavier Baró, Jordi Gonzàlez, Hugo Jair Escalante, Dusan Misevic, Ulrich Steiner, and Isabelle Guyon, "Chalearn looking at people 2015: Apparent age and cultural event recognition datasets and results," in *ICCV*, 2015, pp. 243–251.

[21] Rasmus Rothe, Radu Timofte, and Luc Van Gool, "Deep expectation of real and apparent age from a single image without facial landmarks," *IJCV*, vol. 126, no. 2-4, pp. 144–157, 2018.

[22] Hao Liu, Penghui Sun, Xing Wang, Zhenhua Yu, Suping Wu, and Sun Xuehong, "Similarity-aware and variational deep adversarial learning for facial age estimation," in *TMM, accepted.*

[23] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, 2016.

[24] Hu Han, Anil K. Jain, Fang Wang, Shiguang Shan, and Xilin Chen, "Heterogeneous face attribute estimation: A deep multi-task learning approach," *TPAMI*, vol. 40, no. 11, pp. 2597–2609, 2018.

[25] Bin-Bin Gao, Hong-Yu Zhou, Jianxin Wu, and Xin Geng, "Age estimation using expectation of label distribution learning," in *IJCAI*, 2018, pp. 712–718.