Hindawi
Wireless Communications and Mobile Computing
Volume 2023, Article ID 9825958, 1 page
https://doi.org/10.1155/2023/9825958

Retraction

Retracted: Age Label Distribution Learning Based on Unsupervised Comparisons of Faces

Wireless Communications and Mobile Computing

Received 11 July 2023; Accepted 11 July 2023; Published 12 July 2023

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope
(2) Discrepancies in the description of the research reported
(3) Discrepancies between the availability of data and the research described
(4) Inappropriate citations
(5) Incoherent, meaningless and/or irrelevant content included in the article
(6) Peer-review manipulation

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] Q. Li, Z. Deng, W. Xu, Z. Li, and H. Liu, "Age Label Distribution Learning Based on Unsupervised Comparisons of Faces," *Wireless Communications and Mobile Computing*, vol. 2021, Article ID 1996803, 7 pages, 2021.

Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 1996803, 7 pages
https://doi.org/10.1155/2021/1996803

Research Article

Age Label Distribution Learning Based on Unsupervised Comparisons of Faces

Qiyuan Li,[1,2] **Zongyong Deng**,[1,3] **Weichang Xu**,[1,4] **Zhendong Li**,[1,4] **and Hao Liu**[1,4]

[1]*School of Information Engineering, Ningxia University, Yinchuan 750021, China*
[2]*School of Cyberspace Science and Technology, Beijing Institute of Technology, Beijing 100081, China*
[3]*College of Computer Science, Sichuan University, Chengdu 610065, China*
[4]*Collaborative Innovation Center for Ningxia Big Data and Artificial Intelligence Co-founded by Ningxia Municipality and Ministry of Education, Yinchuan 750021, China*

Correspondence should be addressed to Weichang Xu; xuwch@nxu.edu.cn and Hao Liu; liuhao@nxu.edu.cn

Received 22 August 2021; Accepted 16 October 2021; Published 13 November 2021

Academic Editor: Chi-Hua Chen

Although label distribution learning has made significant progress in the field of face age estimation, unsupervised learning has not been widely adopted and is still an important and challenging task. In this work, we propose an unsupervised contrastive label distribution learning method (UCLD) for facial age estimation. This method is helpful to extract semantic and meaningful information of raw faces with preserving high-order correlation between adjacent ages. Similar to the processing method of wireless sensor network, we designed the ConAge network with the contrast learning method. As a result, our model maximizes the similarity of positive samples by data enhancement and simultaneously pushes the clusters of negative samples apart. Compared to state-of-the-art methods, we achieve compelling results on the widely used benchmark, i.e., MORPH.

1. Introduction

Human face is a basic biological feature of human beings, and its image contains a lot of useful information, such as age, gender, identity, race, and emotion [1]. Face age estimation is aimed at using computer technology to predict the accurate age values for the given facial images. However, variations of the shape of the skull, the position of the facial features, wrinkles, lighting, expressions, and movements of videos likely give rises to bias prediction in the wild conditions [2]. Particularly when a small amount of training data is used, the accuracy of age prediction is generally not high.

Recently, although people have been working on age estimation research, the performance is still very limited. This is mainly affected by two factors. On the one hand, because the existing dataset is not complete enough, most methods are trained in a supervised way, which requires manual annotations. On the other hand, the relationship of face data and age labels is usually complexly heterogeneous and nonlinear [3, 4]. Hence, this urgently prompts us to propose robust and accurate facial age estimation particularly under unconstrained environments.

Conventional age estimation methods could be roughly categorized into two major ingredients: feature representation and age predictor. Feature representation-based methods [5–7] are aimed at seeking discriminative feature descriptors for ages based on the face images. Respectively, age predictor-based methods [8, 9] basically learn to classify the age ranker based on the input feature representation. Apart from that, label distribution has emerged as the widely employed and state-of-the-art methods such as [10–12]. The algorithm typically encodes a range of age labels to a symmetrical distribution, e.g., Gaussian or triangle distribution, reflecting the smoothness for high-performance age estimation. Nevertheless, they are constrained to take only fixed-structural form to model the ambiguous properties of age labels, which are usually nonrobust to complex cross-

population face data domains. In order to solve this problem, most scholars usually adopt feature fusion methods, such as [13, 14], but these methods seldom pay attention to the high correlation between adjacent samples and often require a lot of annotation data to achieve. Therefore, we propose a flexible unsupervised comparison of label distribution learning age estimation method, which can solve the above problems.

Similar to the wireless sensor network in the space to monitor and record the physical conditions of the environment and organize the collected data in a central location. In this article, we propose a label distribution learning method based on unsupervised comparison, dubbed UCLD, which typically models heterogeneous face aging data for robust face age estimation. Compared with the traditional fixed and inflexible label distribution methods, our method not only takes into account the high correlation between adjacent samples but also reduces the dependence of the model on the data. In this article, we believe that the learned distribution is determined by the relationship between the samples, as shown in Figure 1. Technically, we first construct the embedding space of each anchored sample based on the facial appearance information. Then, the age feature is extracted through the constraints of the two projection layers and the contrast loss. Our network structure uses the improved VGG-16 [15] for effective feature learning. Figure 2 illustrates the flow chart. In order to further evaluate the effectiveness of our proposed method, we conduct extensive experiments on two field datasets. Compared with the existing facial age estimation methods, it achieves significantly superior performance.

2. Methodology

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

Considering the size and efficiency of the model, the convolutional neural network used in this article is an improved network from four aspects based on the VGG-16 [15] architecture. First, the three fully connected layers of the VGG-16 [15] architecture contain approximately 90% of the parameters of the entire model. In this paper, only two fully connected layers are used and the dimensionality is reduced sequentially, and the mixed layer constructed by the maximum pooling layer and the global average pooling layer is retained. Second, in order to further reduce the model size, the number of filters in each convolutional layer is reduced by half to make it thinner than the original VGG-16 [15] architecture. Third, in order to speed up the training, a batch normalization layer is added after each convolutional layer [17]. Finally, the pretraining model is obtained through the comparison learning module, and then, the label distribution learning module and the expectation regression module are added to jointly learn the age distribution. The algorithm will be described in detail in the following.

2.1. Problem Setting. Assume the input space $X = R^{h*w*c}$, where h, w, and c represent the height, width, and number of channels of the input image, respectively. The label $Y = R$ represents the actual age value. On the training set $\mathscr{D} = \{(x^i, y^i)\}_{i=1}^{N}$ with the number of samples N, define $x^i \in X$ as the ith input image, and $y^i \in Y$ as the corresponding age. The age estimation problem is to learn the mapping function $\mathscr{F} : X \longrightarrow Y$ in order to make the error between the predicted value $\hat{y}$ and the true value y as small as possible on a given input image x.

Gao et al. [18] defined $l = [0 : \triangle l : 100]$ as an ordered label vector, where $\triangle l$ is a fixed real number. Using an equal step size $\triangle l$ to quantify y, the probability density function of the normal distribution that generates the true value p through y and σ is

$$p^k = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(l_k - y)^2}{2\sigma^2}\right), \tag{1}$$

where σ is a hyperparameter and p^k is the probability that the true age is l_k years old. This article is aimed at maximizing the similarity between the true value p and the predicted value $\hat{p}$ generated by the convolutional neural networks.

2.2. Contrastive Loss. For a set of N randomly sampled sample pairs $\{x_k, y_k\}_{k=1\cdots N}$, the corresponding batch used for training consists of $2N$ sample pairs $\{x_l, y_l\}_{l=1\ldots 2N}$, where x_{2k} and y_{2k-1} are two random enhanced views of $x_{k(k=1\cdots N)}$ and $y_{2k-1} = y_{2k} = y_k$.

In the data processing of $2N$ extended samples, let $i \in I \equiv \{1\ldots 2N\}$ be the index of an arbitrary augmented sample, and let $j(i)$ be the index of the other augmented sample originating from the same source sample. In unsupervised contrastive learning [19–21], the loss takes the following form:

$$\mathscr{L}^{\text{self}} = \sum_{i\in I} \mathscr{L}_i^{\text{self}} = -\sum_{i\in I} \log \frac{\exp\left(Z_i \cdot Z_{j(i)}/\tau\right)}{\sum_{a\in A(i)} \exp(Z_i \cdot Z_a/\tau)}. \tag{2}$$

Here, $Z_l = \text{Proj}(\text{Enc}(x_l)) \in R^{\mathscr{D}_p}$, the $\cdot$ symbol denotes the inner product, $\tau \in R^+$ is a scalar temperature parameter, and $A(i) \equiv I \setminus \{i\}$. The index i is called the anchor, index $j(i)$ is called the positive, and the other $2(N-1)$ indices $(k \in A(i) \setminus \{j(i)\})$ are called the negatives. Note that for each anchor i, there is 1 positive pair and $2N-2$ negative pairs. The denominator has a total of $2N-1$ terms (the positive and negatives).

2.3. Label Distribution Learning. If the true ages of the two input images are similar, the two images are considered similar. In other words, input images with similar outputs are theoretically highly correlated. In order to use the features extracted from these correlations, the label distribution learning module quantifies the range of possible y values into labels in l.

Specifically, given the input image x and the corresponding label distribution p, it is assumed that $f = \mathscr{F}(x;\theta)$ is the activation of the last layer of the convolutional neural network, where θ represents the parameters of the convolutional neural network. A fully connected layer passes f to $x \in R^K$ through

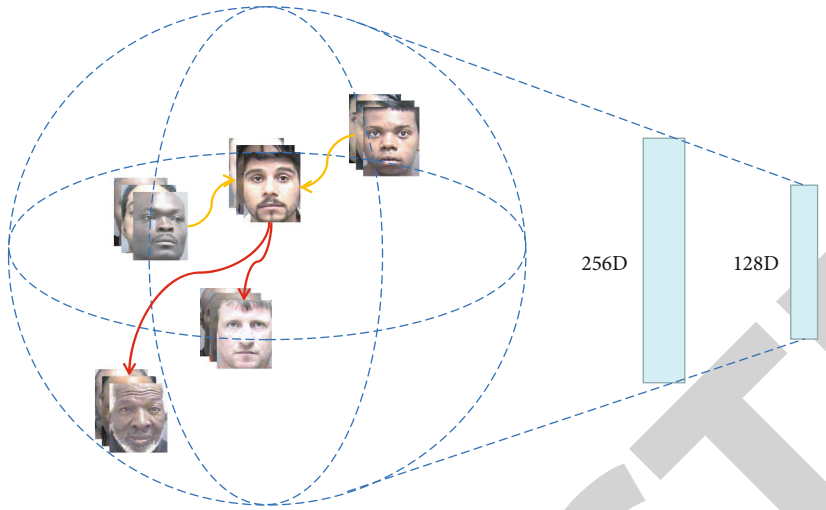

Figure 1: Demonstration of our insight. Our model is aimed at constructing a balanced embedding space, so that the anchor is closer to similar samples and farther away from different samples. Then, the age characteristics of the samples are extracted through two projection layers to make a robust age estimation.

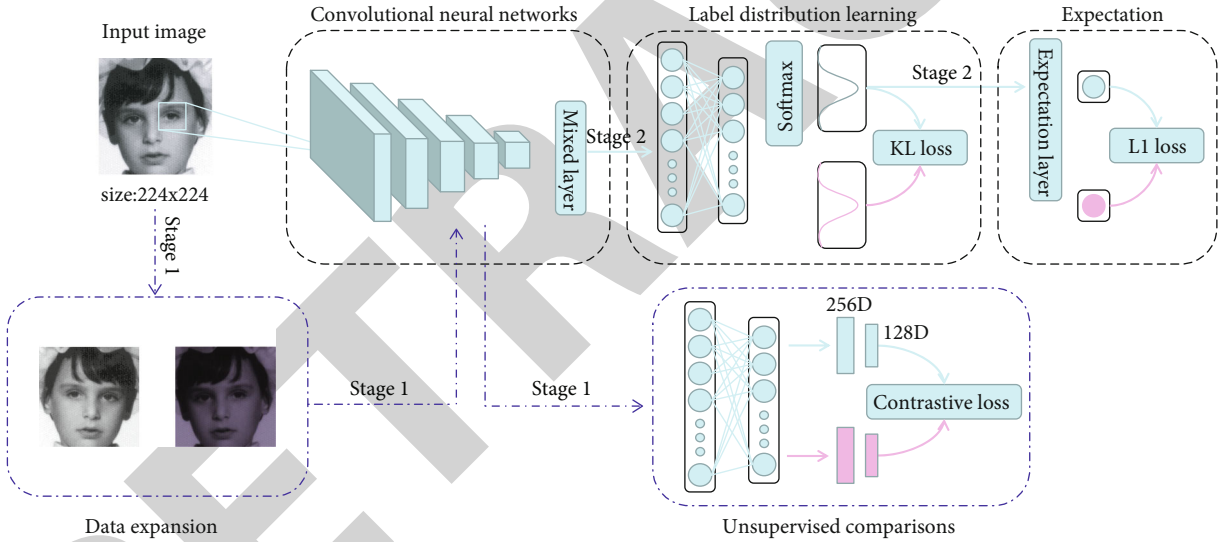

Figure 2: Flowchart of our UCLD. Our structure is divided into two stages. In the first stage, after data expansion of the image, the age samples are input into the preset CNN to get the normalized embedding of the image and then the vector embedded through the two projection layers is calculated and compared to the loss to obtain the ConAge model, which is the basis for the algorithm proposed in this paper. In the second stage, after obtaining these relevant depth features, they are projected into the average variance label distribution through a small linear layer, and the network parameters are optimized through backpropagation. At the same time, the mixed hyperparameters of the average variance label distribution are iterated through the widely used expectation-maximization optimization [16].

$$x = W^T f + b. \tag{3}$$

Then, we use the softmax function to convert x into a probability distribution as follows:

$$\widehat{p}_k = \frac{\exp(x_k)}{\sum_t \exp(x_t)}. \tag{4}$$

For a given input image, the goal of label distribution learning is to find θ, W, and b to generate $\widehat{p}$ similar to p.

Finally, the KL divergence is used as a measure of the difference between the real label and the predicted label. Therefore, the following loss function is defined on the training sample:

$$L_{ld} = \sum_k p_k \ln \frac{p_k}{\widehat{p}_k}. \tag{5}$$

2.4. Expectation Regression. Using only the label distribution learning module cannot accurately predict the age of the

Table 1: Face age estimation result table.

Method	Network	Dataset	MAE
DLDL-v2 (baseline)	TinyAge	FGNET	4.4676
	ThinAge		4.1322
UCLD	*ConAge*		*3.6046*
DLDL-v2 (baseline)	TinyAge	MORPH	2.5118
	ThinAge		2.3440
UCLD	*ConAge*		*2.2142*

Table 2: Comparison result table of different settings.

	ConAge*	ConAge	ConAge1	ConAge2	ConAge3
Label	Exist	None	None	None	None
Linear	2	2	1	2	2
Batch size	24	24	24	64	72
Epoch	120	120	120	120	120
Temp	0.07	0.07	0.07	0.07	0.07
MAE	2.3477	*2.2142*	2.2627	2.3185	2.3436

face. Therefore, this paper uses the expected regression module proposed in the DLDL-v2 [18] framework to improve the accuracy of face age prediction.

As shown in Figure 2, when the predicted value and label are obtained, the expected value is output:

$$\hat{y} = \sum_k \hat{p}_k l_k, \tag{6}$$

where $\hat{p}_k$ represents the predicted probability that the input image belongs to label l_k. Given the input image, the error between the expected value $\hat{y}$ and the true value y is minimized. The error metric uses the l_1 loss function, as shown in the following:

$$L_{er} = |\hat{y} - y|, \tag{7}$$

where $|\cdot|$ represents the absolute value.

2.5. Optimization. By jointly learning the label distribution and expected regression, the values of θ, W, and b can be obtained in a given data set $\mathscr{D}$. The final loss function is defined as a weighted combination of two loss functions L_{ld} and L_{er}.

$$L = L_{ld} + L_{er}, \tag{8}$$

where λ is the weight that weighs the importance of the two losses. Substituting (5), (6), and (7) into (8), we get

$$L = -\sum_k p_k \ln \hat{p}_k + \lambda \left| \sum_k \hat{p}_k l_k - y \right|. \tag{9}$$

In this framework, optimization variables include θ, W, and b. First, backpropagation through the network, and then use the stochastic gradient descent algorithm to optimize the parameters. The derivative of L with respect to $\hat{p}_k$ is

$$\frac{\partial L}{\partial \hat{p}_k} = -\frac{p_k}{\hat{p}_k} + \lambda l_k \operatorname{sign}(\hat{y} - y). \tag{10}$$

For any k and j, the derivative of the softmax function (4) is as follows:

$$\frac{\partial \hat{p}_k}{\partial x_j} = \hat{p}_k \left(\delta_{(k=j)} - \hat{p}_j \right). \tag{11}$$

Among them, if $k = j$, then $\delta_{(k=j)}$ is 1; otherwise, it is 0. Then,

$$\frac{\partial L}{\partial x} = \hat{p} - p + \lambda \operatorname{sign}(\hat{y} - y) \hat{p} \circ (l - \hat{y}). \tag{12}$$

Applying the chain rule to (3) again, the derivative of L with respect to θ, W, and b can be easily obtained

$$\frac{\partial L}{\partial W} = \frac{\partial L}{\partial x} f, \frac{\partial L}{\partial b} = \frac{\partial L}{\partial x}, \frac{\partial L}{\partial \theta} = \frac{\partial L}{\partial x} W^T \frac{\partial \mathscr{F}}{\partial \theta}. \tag{13}$$

Once θ, W, and b are known, in the forward network calculation, the age prediction value $\hat{y}$ of any face image x can be generated by (6), and finally, the age estimation of the face image is realized.

3. Experiments

In order to evaluate the effectiveness of this method, we conducted research results on two widely used datasets, including FGNET [22] and MORPH [23]. Due to wild conditions, face samples in these datasets often experience challenging situations. In order to illustrate the advantages of this model, we only use the MORPH dataset for model pretraining.

3.1. Datasets. The FG-NET dataset was constructed by Professor Lanitis of the University of Cyprus in Europe while studying the age estimation algorithm for faces. This dataset collected a total of 1002 facial images of 82 people through scanning. Each image provides 68 key points of face information, ranging from 0 to 69 years old. It is currently one of the most open real age datasets of the young people. For fair evaluation setting, we employed the leave-one-person-out (LOPO) protocol by following [9].

The MORPH dataset was constructed by Karl Ricanek Jr. of North Carolina State University and others when they studied face aging. The dataset consists of two parts: Album1 and Album2, which contain 1724 and 55608 face images, respectively. Album1 was collected from 1962 to 1998, and the age span was 15-68 years; Album2 was collected from 2003 to 2007, and the age span was 16-77. Since the number of collections of Album2 is significantly more than that of Album1, most scholars use Album2 for facial age estimation research. In order to make fair comparisons, we also use the

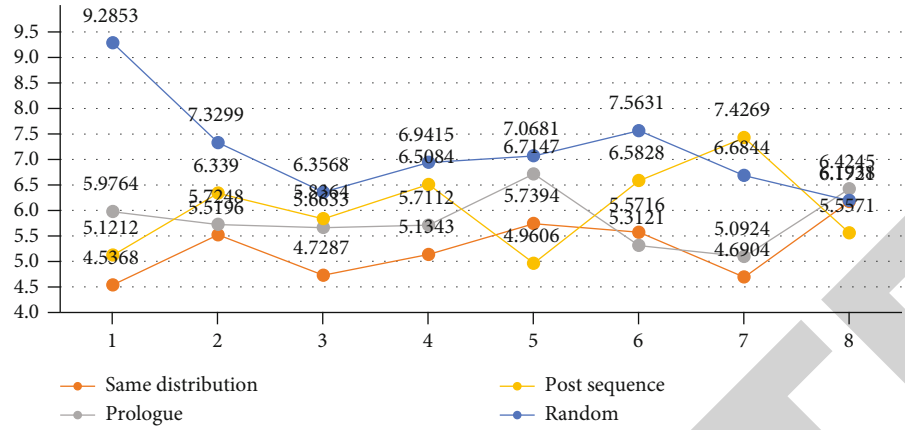

Figure 3: The comparison results of the four weakly supervised sampling methods on the TinyAge network architecture and the FG-NET dataset in 8 experiments.

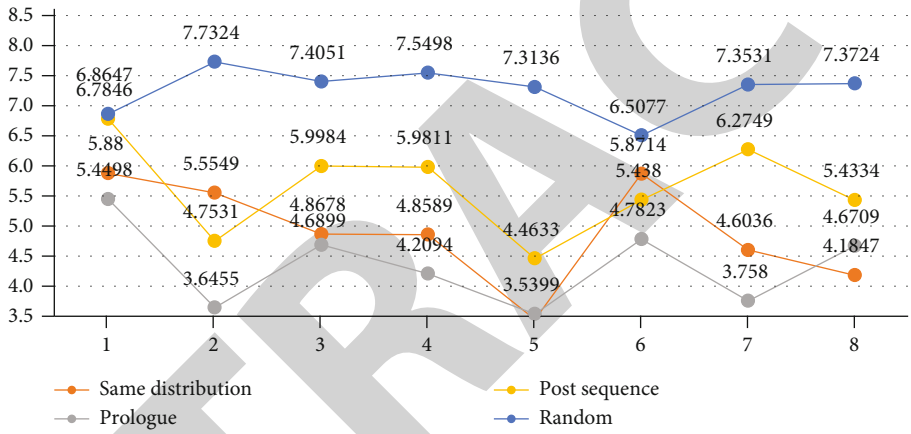

Figure 4: The comparison results of the four weakly supervised sampling methods on the ThinAge network architecture and the FG-NET dataset in 8 experiments.

Table 3: Weakly supervised face age estimation result table.

Method	Network	Dataset	MAE
DLDL-v2 (baseline)	ThinAge	FGNET (25%)	6.4146
UCLD	*ConAge*		*6.3342*
DLDL-v2 (baseline)	ThinAge	MORPH (25%)	2.8834
UCLD	*ConAge*		*2.6545*

Album2 dataset, where 80% of the data is used as the training set and 20% of the data is used as the test set.

3.2. Evaluation Metric. In the experiment, we use Mean Absolute Error (MAE) [24] to calculate the difference between the estimated age value and the true age value. Obviously, the smaller the value of MAE, the smaller the error between the predicted age and the true age, and the better the performance of the model, as shown in Table 1.

Please note that the DLDL-v2 [18] mentioned in this article is all source codes released by them. Compare our method with the experimental results of DLDL-v2 on the FGNET and MORPH datasets. Obviously, our method is more advantageous. In addition, we also changed the experimental settings several times as shown in Table 2.

Among them, linear represents the number of projection layers used. Despite using different settings, the experimental results of our method on the MORPH dataset still maintain the most advanced performance.

3.3. Implementation Details. For each face image, the size is adjusted to 224×224 before being input to the network. Then, select one of the five data enhancement methods: random horizontal flip, random zoom, random rotation, color distortion, and Gaussian blur to process the image. The comparative learning module of the network is used to generate a pretraining model on the MORPH dataset. The initial learning rate is set to 0.001, and it is reduced by 10 times every 30 iterations. After the pretraining is completed, delete the contrast learning module of the network and add the label distribution learning module and the expectation regression module to test the face age dataset. During the

test, the test image and its flipped copy are fed to the network, and its predicted value is averaged as the final age estimate.

In order to further evaluate the performance of the method proposed in this paper, the following weakly supervised experiments are completed. Regarding the sample order in fully supervised training as the original order, five sampling methods are proposed as follows:

(i) Sampling with the same distribution: that is, the probability of taking out 25% of the labeled data in the original sample interval is equal.

(ii) Preorder sampling: take the first 25% of the labeled data in the order of the original sample.

(iii) Postsampling: take the last 25% of the labeled data in the order of the original sample.

(iv) Random sampling: 25% of the labeled data is randomly selected from the original sample.

(v) Single sampling: that is, only different labeled data are retained in the original sample.

The TinyAge and ThinAge network architectures were applied to these five sampling methods, respectively, and eight tests were performed on the first face data file in the FG-NET dataset. The average MAE after 8 tests on the two networks with a single sampling method are 16.81 and 13.03, respectively. The test results of the other four sampling methods are shown in Figures 3 and 4.

Change the training dataset to a weakly supervised training dataset, and use only 25% of the labeled data to test the optimal ThinAge network architecture in DLDL-v2 and the ConAge network architecture proposed in this article. The experimental results are shown in Table 3.

It can be seen from the experimental results that our method has better performance than the DLDL-v2 framework regardless of whether it is fully supervised or weakly supervised. In addition, we have reached three conclusions: (1) traditional methods, such as DEX [25] and ODFL [25], process each age label independently without considering their previous correlation. Our unsupervised comparison method simulates the way humans observe things and can flexibly consider the relationship between age samples. (2) Some label distribution learning methods, such as LDL [11] and CPNN [11], only implement a fixed structural model on the age label distribution, which may lead to rigid adaptation to real-world facial aging data. Thanks to the comparative learning module, our method obtains more accurate semantic information, making subsequent test results more accurate. Particularly in a weakly supervised experimental setting, it can be seen that even if only a quarter of the data is used, the performance of our UCLD is better than most technical levels. This achievement is mainly because our model is less dependent on data.

4. Conclusion

In this article, in view of the high correlation between adjacent age samples and the strong dependence of existing methods on data, we combine contrast loss and label distribution learning to learn abstract representations in an unsupervised manner. An unsupervised contrast label distribution (UCLD) learning method is proposed, which is similar to the processing form of wireless sensor networks. Extensive experiments on two datasets have proved the effectiveness of the method, especially the MORPH dataset reflects the advanced nature of the method. In future work, we will focus on efficiently distinguishing similar images to solve the problem of age prediction accuracy.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

This work was supported in part by the National Science Foundation of China under Grants 61806104 and 62076142, in part by the West Light Talent Program of the Chinese Academy of Sciences under Grant XAB2018AW05, and in part by the Youth Science and Technology Talents Enrolment Projects of Ningxia under Grant TJGC2018028.

References

[1] R. Angulu, J. R. Tapamo, and A. O. Adewumi, "Age estimation via face images: a survey," *EURASIP Journal on Image and Video Processing*, vol. 2018, no. 1, 2018.

[2] N. Ramanathan, R. Chellappa, and S. Biswas, "Age progression in human faces: a survey," *Journal of Visual Languages and Computing*, vol. 15, pp. 3349–3361, 2009.

[3] W. Li, J. Lu, J. Feng, C. Xu, J. Zhou, and Q. Tian, "Bridgenet: a continuity-aware probabilistic network for age estimation," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1145–1154, Long Beach, CA, USA, 2019.

[4] W. Shen, Y. Guo, Y. Wang, K. Zhao, B. Wang, and A. L. Yuille, "Deep regression forests for age estimation," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2304–2313, Salt Lake City, UT, USA, 2018.

[5] 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*, vol. 29, no. 12, pp. 2234–2240, 2007.

[6] Yun Fu, Guodong Guo, and T. S. Huang, "Age synthesis and estimation via faces: a survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 11, pp. 1955–1976, 2010.

[7] J. Lu, V. E. Liong, and J. Zhou, "Costsensitive local binary feature learning for facial age estimation," *EEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5356–5368, 2015.

[8] Z. Yu and D.-Y. Yeung, "Multi-task warped Gaussian process for personalized age estimation," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2622–2629, San Francisco, CA, USA, 2010.

[9] K.-Y. Chang, C.-S. Chen, and Y.-P. Hung, "Ordinal hyperplanes ranker with cost sensitivities for age estimation," in *CVPR*, pp. 585–592, Colorado Springs, CO, USA, 2011.

[10] B.-B. Gao, C. Xing, C.-W. Xie, J. Wu, and X. Geng, "Deep label distribution learning with label ambiguity," *IEEE Transactions on Image Processing*, vol. 26, no. 6, pp. 2825–2838, 2017.

[11] Xin Geng, Chao Yin, and Zhi-Hua Zhou, "Facial age estimation by learning from label distributions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 10, pp. 2401–2412, 2013.

[12] Z. He, X. Li, Z. Zhang et al., "Data-dependent label distribution learning for age estimation," *IEEE Transactions on Image Processing*, vol. 26, no. 8, pp. 3846–3858, 2017.

[13] Z. Deng, M. Zhao, H. Liu, Z. Yu, and F. Feng, "Learning neighborhood-reasoning label distribution (NRLD) for facial age estimation," in *2020 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, London, UK, 2020.

[14] A. Vaswani, N. Shazeer, N. Parmar et al., "Advances in Neural Information Processing Systems 30," in *Annual Conference on Neural Information Processing Systems 2017*, Long Beach, C, USA, December 4–9, 2017.

[15] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations, ICLR 2015*, San Diego, CA, USA, 2015.

[16] 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. 1–22, 1977.

[17] S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*, pp. 448–456, Lille, France, 2015.

[18] B. B. Gao, H. Y. Zhou, J. Wu, and X. Geng, "Age estimation using expectation of label distribution learning," *IJCAI*, pp. 712–718, 2018, ijcai.org.

[19] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*, pp. 1597–1607, Virtual Event, 2020.

[20] P. Khosla, P. Teterwak, C. Wang et al., "Supervised contrastive learning," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[21] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," in *Computer Vision–ECCV 2020: 16th European Conference*, pp. 776–794, Glasgow, UK, 2020.

[22] A. Lanitis, C. J. Taylor, and T. F. Cootes, "Toward automatic simulation of aging effects on face images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 4, pp. 442–455, 2002.

[23] K. Ricanek Jr. and T. Tesafaye, "MORPH: a longitudinal image database of normal adult ageprogression," in *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, pp. 341–345, Southampton, 2006.

[24] R. Rothe, R. Timofte, and L. Van Gool, "Deep expectation of real and apparent age from a single image without facial landmarks," *International Journal of Computer Vision*, vol. 126, no. 2-4, pp. 144–157, 2018.

[25] H. Liu, J. Lu, J. Feng, and J. Zhou, "Ordinal deep learning for facial age estimation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 2, pp. 486–501, 2019.