IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 28, 2022, accepted February 16, 2022, date of publication February 24, 2022, date of current version March 8, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3154403

Cross-Dataset Learning for Age Estimation

BEICHEN ZHANG AND YUE BAO, (Member, IEEE)
Department of informatics, Tokyo City University, Setagaya Campus, Tokyo 158-8557, Japan
Corresponding author: Beichen Zhang (g1991804@tcu.ac.jp)

ABSTRACT Age estimation from a single human face image has been an important yet challenging task in computer vision and multimedia. Due to the large individual differences in human faces, including the differences in races and genders, the performance of a learning model depends largely on training data. The existing learning methods are challenged by insufficient numbers of images and poor-quality images in datasets, as well as by new low-precision data that are dissimilar to existing training data. In this paper, we propose a learning method called the cross-dataset training convolutional neural network (CDCNN), which uses a general framework for cross-dataset training in age estimation. We adopted convolutional neural networks (CNNs) with VGG-16 architectures pretrained on ImageNet and treated the age estimation problem as a classification problem. For the classification results, softmax is utilized to map the output and provide value refinement. We conducted a series of experiments on the Craniofacial Longitudinal Morphological Face Database (MORPH), Cross-Age Celebrity Dataset (CACD), and Asian Face Age Dataset (AFAD). The results show that simultaneous training on multiple datasets using additional labeled data achieves a more impressive performance when compared to training on a single, independent dataset. Our proposed cross-dataset training model achieves state-of-the-art results on both the AFAD and CACD age estimation benchmarks with great generalizability.

INDEX TERMS Age estimation, deep learning, cross-dataset training.

I. INTRODUCTION

Human face images contain many characteristics, such as gender, age, race, expression, and health condition characteristics. Among all these characteristics, age estimation has become a very challenging and important issue, as it is used in various fields, such as human-computer interactions [1], [14], [39], identification [2], security [45], and precision advertising [5].

Studies on the age estimation problem using facial images can be traced to 1994 [44]. In the beginning, researchers only used skin wrinkles and parts of regions on a human face to estimate an approximate age range. It took many years until a widely used representation method named biologically inspired features (BIFs) [13] was proposed in 2009; this method achieved much better accuracy than traditional methods, such as the local binary pattern (LBP) [33], scale invariant feature transform (SIFT) [9], and Gabor [8] methods. In recent years, convolutional neural networks (CNNs) have demonstrated impressive performance in various fields of computer vision, such as object detection [35], segmentation [23], and face recognition [46]. CNNs have also become a mainstream approach for the age estimation problem [29], [42], and it has shown inspiring progress at the ChaLearn looking at people challenge [30].

Despite extensive studies on the age estimation problem, the performance of the existing methods is still far from meeting real life demands in terms of accuracy and reliability. The factors that make this problem so difficult can be divided into two categories: extrinsic appearance factors, including illumination, pose, and expression factors [15], and intrinsic human factors, including race, gender, and health condition factors [13]. Many previous works have focused on extrinsic factors because intrinsic factors are difficult to resolve. This difficulty stems from inhomogeneous features caused by 1) people of the same age having a very large variation in facial appearance (Fig. 1) and 2) the human face changing in different ways according to age, e.g., fast bone growth in childhood will result in fewer facial changes in adulthood [26] (Fig. 2). It is a major challenge to design an age estimator that can accurately estimate facial images from different populations for various age groups because of the large disparities in aging patterns and populations.

The associate editor coordinating the review of this manuscript and approving it for publication was Zhenhua Guo.

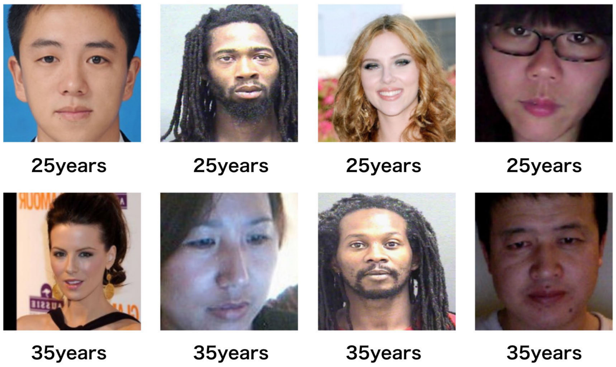


FIGURE 1. Differences between different people of the same age.

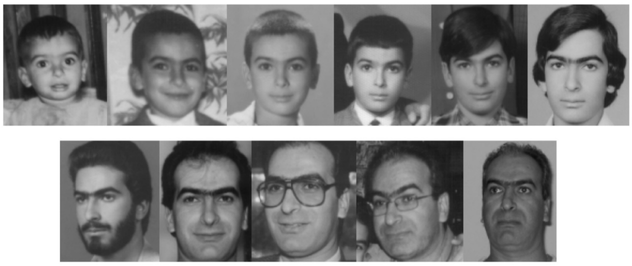


FIGURE 2. Changes in facial appearance from childhood to adulthood.

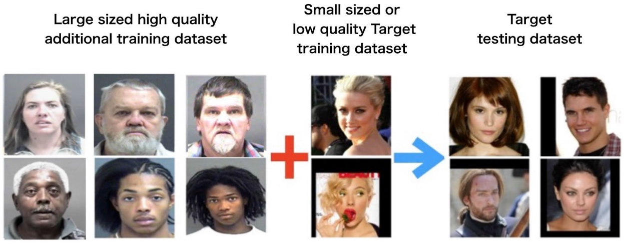


FIGURE 3. Cross-dataset age estimation problem.

Almost all existing facial image datasets with age labels contain facial images from a single race. Some previous works suggested estimating age for each ethnicity separately [11], [12]. However, separate models face a new problem that most of the open datasets have insufficient data for each ethnicity, and some data have incorrect labels. The option of creating a dataset that is specific to a model would be expensive, and the data would be difficult to manually collect or label sufficiently. Instead of labeling more data to address this problem, a large high-quality dataset of one ethnicity has been used to improve the performance of age estimation on small-sized or low-quality datasets with images of faces of another ethnicity [34]. This method can also improve the generalizability of a trained model so that it is more reliable in real-life applications.

The problem with this cross-dataset age estimation approach (Fig. 3) is that the target dataset only has a small set of high-quality training data, which causes instability in the learning process and leads to low-precision learning results. In this paper, we propose an end-to-end learning model named the cross-dataset training CNN (CDCNN) to achieve more accurate age estimation. Previous studies using cross-dataset training have not been applied to age estimation with facial images. The reason is that cross-dataset training usually uses multiple datasets with different labels of different classes, in which the labels may be replicated from different datasets. This is a complex problem that causes conflicts between positive/negative data samples from different datasets and needs to be solved in cross-dataset training. However, we noticed that facial images with different age labels are not heterogeneous because they are naturally labeled with different integers. This led us to the idea that the age estimation problem can be easily handled as a classification problem where different classes represent different ages and there are no label conflicts.

Our CDCNN model uses a CNN that extracts features from facial images for age estimation. The features extracted with the learning method can be more discriminative and robust to variations in appearance compared to handcrafted aging features, such as those used in the active appearance model (AAM) [6] or BIF model [13]. This study treats age estimation as a classification problem rather than a regression problem. Softmax is utilized to map the output and provide expected value refinement. To obtain higher performance on the target dataset, the model also uses additional labeled data from multiple datasets.

Experiments are conducted based on three standard benchmarks for real age estimation: the Asian Face Age Dataset (AFAD) [48], Cross-Age Celebrity Dataset (CACD) [4], and Craniofacial Longitudinal Morphological Face Database (MORPH) [25]. Moreover, cross-dataset training is performed on a combined dataset, and the performance result is compared to the result on one dataset. The experimental results demonstrate that our proposed model outperforms several state-of-the-art methods on ADAD and CACD. The rest of this paper is organized as follows: Section 2 reviews the related works, Section 3 introduces our proposed learning method and implementation details, Section 4 presents the experimental results and discussions, and the conclusion and discussion of future issues are in Section 5.

II. RELATED WORK

A. AGE ESTIMATION

Kwon and Lobo completed some of the earliest work on age estimation [44]; they divided the age distribution into three groups (babies, young adults, and senior adults). Subsequently, the age estimation problem became more attractive. Lanitis *et al.* employed AAM [6] to combine both texture (e.g., wrinkles) and shape features of the human face and made impressive progress in the age estimation problem. Other works using AAM, such as [1], introduced different classifiers, including quadratic functions and artificial neural networks. Furthermore, the advanced aging BIF method was proposed [13] using multidirectional and multiscale Gabor filters. Although notable results have been reported using BIF on mainstream databases of facial images for age estimation (MORPH and FG-NET), the limitation of handcrafted features makes BIF and similar methods hardly optimal solutions for the age estimation problem.

In recent years, learning methods have been widely used in many computer vision tasks, such as object detection [35], segmentation [23], and face recognition [46], because of their great success. In age estimation tasks, CNNs perform better than previous methods [10], [29], [44], [48]. Instead of mapping a full image to a certain age, as in manifold learning, CNNs aim to automatically learn efficient age-related features.

B. CROSS-DATASET TRAINING

Integrated face analytic networks use a cross-dataset training method [18], in which various tasks with different datasets are trained together. In this study, multiple features, such as facial landmarks and facial emotions, are integrated to create a generic facial model; therefore, it is unnecessary to label all features for every data sample. This method significantly improves the performance by modeling the interaction among features as an element of the scope of multitask learning.

A more closely related approach was proposed by Kuang *et al.* [47] at the ChaLearn Looking at People (LAP) challenge 2015. They proposed a method for leveraging publicly available labeled facial age datasets to estimate ages from unconstrained face images and learned discriminative age-related representations on multiple publicly available age datasets using a deep CNN. CNN training was supervised by rich binary codes and thus modeled as a multilabel classification problem. The codes represent different age group partitions at multiple granularities and gender information. Then, they trained a regressor from the deep representation to the age on the small training dataset provided by the LAP organizer by fusing random forest and quadratic regression methods with a local adjustment method.

As manifested in [21], learning nonstationary kernels for a regression problem is usually difficult since it easily causes overfitting in the training process. In our work, we propose the CDCNN method, which treats the age estimation problem as a classification problem instead of a regression problem. Additionally, the CDCNN uses data with the same integer labels from different datasets to solve a single classification task instead of a multitask learning problem, leading to a more stable and simpler model.

III. PROPOSED METHOD

Fig. 4 illustrates the overview of our approach and outlines the proposed method. Each step is explained in detail in this section.

A. FACE CROPPING

The performance can change when the background of a face image changes. Different face alignments in datasets can also lead to performance variations. To prevent the influence of surrounding pixels, all the facial images were squeezed to 256 × 256 pixels and then randomly cropped to 224 × 224 pixels to be input into the neural network. Face image cropping ensures that every face is randomly placed in different locations regardless of the original dataset. This method also makes our model robust enough to handle different face alignment situations in real applications.

Although we did not align every dataset with pixelwise precision in a consistent manner, this cropping approach was satisfactory for our work.

B. CNN ARCHITECTURE

A CNN is introduced to estimate the age from facial images. The network uses facial images as input and outputs the age estimation results. Several facial image datasets with age labels are used for training the CNN.

The CNN VGG-16 [31] architecture (Fig. 4) was chosen because (i) it has a deep but manageable architecture, (ii) previous work using VGG-16 has achieved impressive results on the ImageNet challenge [27], and (iii) pretrained models using VGG-16 are available, which easily facilitates training.

The VGG-16 architecture has more layers than early networks, such as AlexNet [20]; VGG-16 consists of 16 layers, of which 13 are convolutional and 3 are fully connected. Smaller filters of (3 × 3) pixels are adopted in VGG-16; they have simpler compositions compared to those of early CNNs, but they also cause the model to be more complex with the deeper network. For all experiments in this study, the model begins training with CNN models pretrained on the ImageNet dataset [31]. Then, we fine-tune the CNNs with images from each training facial dataset adapted to age estimation. Fine-tuning allows the network to obtain the features of the target dataset, therefore optimizing the performance of the estimation results.

C. OUTPUT LAYER AND EXPECTED VALUE

We used the VGG-16 network pretrained on ImageNet for classification with 1,000 outputs. Each neuron from the output layers represents an object class. In contrast, age estimation is a regression rather than a classification problem since the values of ages are continuous.

In training for regression, only 1 neuron of output is needed in the last layer of VGG-16. Additionally, the Euclidean loss function is used. However, training such a regression model directly is relatively unstable, as outliers can result in substantial errors. This creates large gradients that make it difficult for the whole network to converge, leading to unstable prediction results.

To address this issue, we use a classification method to solve the age estimation problem and therefore discretize different values of ages into $|X|$ classes. Each x_i covers one age value. In this way, the CNN is trained as a classification model where the predicted value is computed by the output

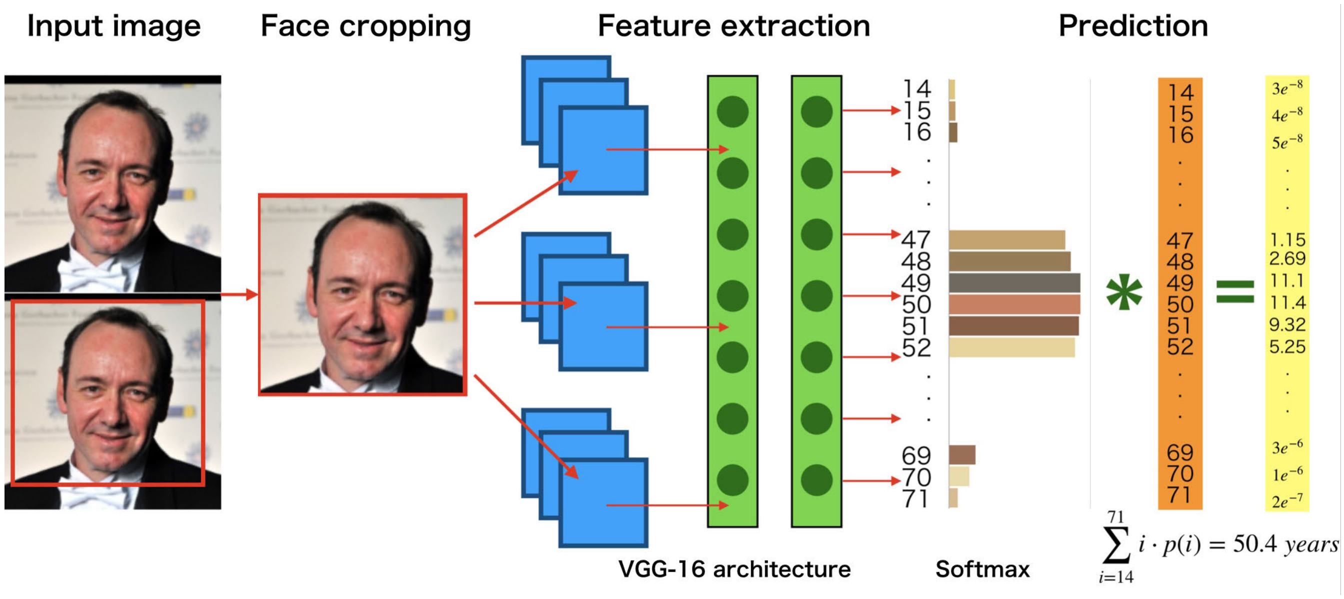


FIGURE 4. Overview of the architecture for the proposed age estimation method.

probabilities with softmax from $|X|$ neurons.

$$E(P) = \sum_{i=1}^{|X|} x_i \cdot p_i \tag{1}$$

where $P = 1, 2, 3, \ldots, |X|$ is the output of the last layer and $p_i \in P$ represents the probability of class i with softmax normalization. The experimental results show that the probability weighted average values can increase the performance and robustness compared to using classifications directly [28].

D. CROSS-DATASET LEARNING

Merging datasets is a key operation for cross-dataset learning. A frequent requirement for merging is joining across columns of different labels for the same entity, for instance, the ethnicity and gender of a person. A traditional cross-dataset learning method that uses multiple datasets with different labels for different classes has the problem that the labels may be replicated in different datasets. However, since age labels are integers, even in different datasets used for age estimation, our cross-dataset learning method combines different datasets from the beginning and trains the mixed datasets using one CNN with one loss function.

Therefore, our cross-dataset learning method has simple structure and achieves higher accuracy since the new cross-dataset model is able to grasp knowledge embedded independently in different datasets. The only thing to note is the balance of data from different datasets. In our work, we ensure that the number of images from the smallest dataset is more than half of the number of images from the largest dataset.

E. EVALUATION PROTOCOL

Two different measures are used for quantitative evaluations in these experiments.

MAE. The popular evaluation metric, the mean absolute error (MAE), is used to measure the performance of different age estimation algorithms. As the definition, the absolute error between the ground truth and the predicted age is averaged as MAE, which is defined as follows:

$$MAE = \frac{1}{K} \sum_{i=1}^{K} |\tilde{y}_i - y_i| \tag{2}$$

where K is the total number of data samples, y_i represents the ground-truth age, and $\tilde{y}_i$ represents the predicted age of the i-th sample. In general, better age estimation approaches have smaller MAE values in the test results.

CS5. Another evaluation metric, the cumulative score (CS), is also used to evaluate different age estimation approaches. CS is the percentage of samples in which the error e (MAE in age estimation) is smaller than a given number i (the year in age estimation) and is defined as

$$CS(i) = K_{e \leq i}/K \tag{3}$$

Related papers give various values of CS from 0 to 10 or simply set a fixed value for CS. Here, $i = 5$ is always used, similar to the approaches used in [21], [22], and [17]. CS values can only be provided for some competitors because not every paper reports the CS values.

IV. EXPERIMENTS

In the following section, the datasets and quantitative and qualitative results are presented; then, a conclusion with a discussion on the findings is given.

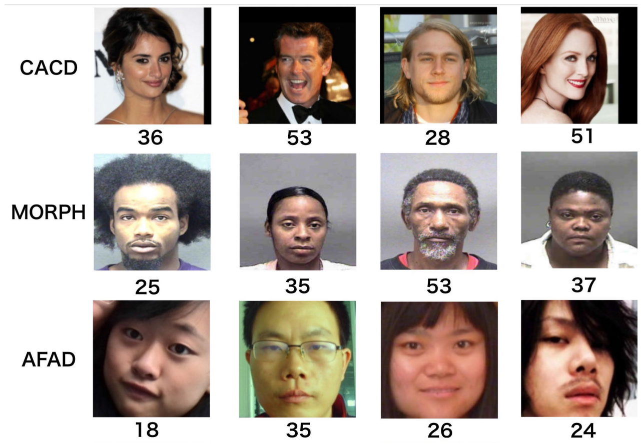


FIGURE 5. Examples of CACD [4], MORPH [25], and AFAD [48]. The number below each image is the ground truth age.

TABLE 1. The number of images, including the split distribution, for each dataset.

Dataset	Number of images
CACD-MORPH-AFAD	**136,473**
CACD	30896(200 celebs)
MORPH	55128
AFAD	50449
[4pt/5pt] **CACD-MORPH**	**86,024**
CACD	30896(200 celebs)
MORPH	55128
[4pt/5pt] **AFAD-MORPH**	**105,577**
MORPH	55128
AFAD	50449
[4pt/5pt] **CACD** [4]	**163,446(2000 celebs)**
train	15448(170 celebs)
test	2723(30 celebs)
[4pt/5pt] **MORPH** [25]	**55,134**
train	46855
test	8273
[4pt/5pt] **AFAD** [48]	**164,432**
train	50449
test	8895

A. DATASETS

In this paper, we use 3 basic datasets for age estimation and 3 different combinations of them for cross-dataset training. Fig. 5 shows some samples from each basic dataset. Table 1 demonstrates the size, including the split distribution for training and testing, of each dataset.

CACD. CACD [4] contains 163,446 human images of 2,000 celebrities. These images were collected online using search engines by year and name of each celebrity. The age labels, which are obtained by computing the date on which the photos were taken and the celebrity's birth date information, are therefore noisy data with partial labeling or image errors. To better evaluate performance, the CACD was divided into three parts according to the different celebrities: 1800 for training, 80 for validation, and 120 for testing. Among them, the validation and testing parts are clean datasets obtained by manually removing noisy images. In our experiments, only a manually cleaned subset with 18,171 photos was used.

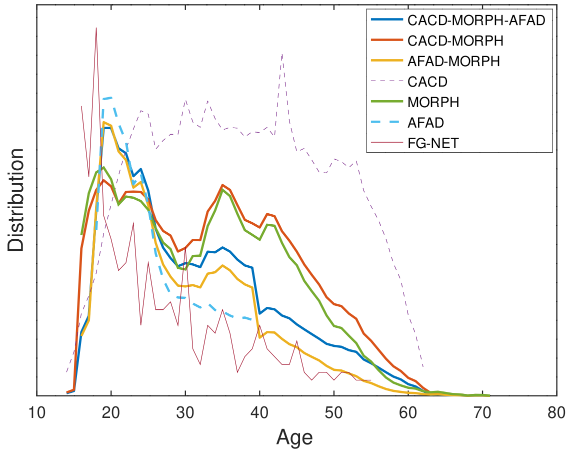


FIGURE 6. The age distribution of the used datasets.

MORPH. MORPH [25] contains 55,000 unique images of more than 13,000 individuals and is the largest publicly available face database with manual data collection. The ages in the dataset range from 16 to 77, with a median age of 33. We only adopted a subset for our experiments, with ages ranging from 16 to 71, and it consists of approximately 55,000 photos to balance the dataset.

AFAD. AFAD [48] contains more than 160K facial images with the corresponding age and gender labels, and it is a new dataset used for age estimation. This is an Asian face-oriented dataset; therefore, all facial images are collected from the RenRen Social Network (RSN), which is widely used by Asian students and graduate students. The ages in the dataset ranged from 15 to more than 40 years old. Only a subset, named AFAD-LITE, which contains 59,344 images of individuals of 22 continuous ages from 18 to 39, was used in our experiments to balance the dataset.

Fig. 6 shows the various age distributions of all the datasets used. CACD has a balanced age distribution between 20 and 60 but only has few samples outside this range. MORPH obviously has 2 peaks on the age distribution curve at approximately 20 and 40 since the data are collected from two sources. AFAD only covers ages 18-40 and has a peak at 20 because most of the images come from high school and college students. FG-NET has an even younger age distribution than AFAD, and most samples are ages younger than 30. CACD-MORPH has a similar distribution as MORPH, and CACD-MORPH-AFAD and AFAD-MORPH have similar distributions as AFAD.

B. IMPLEMENTATION DETAILS

As mentioned above, the CNN in our method is trained for classification. The predicted age is calculated by the probability weighted output with softmax normalization.

For all the experiments, we use the initialized weights trained on ImageNet. This model is then fine-tuned with output neurons on the target dataset for classification and uses the same number of neurons as the number of age classes.

Before training, we initialized all the experiments with the same weights from training on ImageNet [31]. During the training phase, every dataset is split, where 65% of the data is used for learning the weights, 20% is used for validation, and the last 15% is used for testing. We terminate the training process after the CNN overfit on the validation set. The training batch size is 64, and the drop out ratio is 0.5. We train the network using stochastic gradient descent (SGD) with an initial learning rate of 0.01 and reduce it to 1/10 per 15k iterations. The number of output neurons is changed for different datasets because they have different numbers of age classes. The Caffe framework [19] is used to train the networks on Nvidia GTX1080 GPUs.

C. RESULTS AND COMPARISON

First, the performance of our architecture is compared with those of other works on three baseline benchmarks. Then, the results of the proposed cross-dataset method trained on the AFAD and CACD benchmarks is compared to baseline results and the results of the other works. The CACD-MORPH and AFAD-MORPH cross-dataset training settings combine data from the MORPH, CACD and AFAD datasets. Second, the same network and training process are used to ensure a fair comparison. The MAE and CS5 values are also evaluated for all these settings.

1) BASELINES

Three baseline benchmarks are used in the experiment. Fig. 7 shows the learning curves of all three benchmarks. The performance of our proposed method on age estimation is presented in this section.

On MORPH Our CNN structure achieves an MAE value of 2.76 when fine-tuning the CNN on the MORPH training dataset. MORPH was randomly split (85%/15%) into training/testing sets similar to that in [3], [40], [41], and [37]. To reduce experimental errors, the experiment was repeated five times with different random splittings. The average value of the 5 experiments was used as the final result, which indicated robust performance. The quantitative results are summarized in Table 2 and the CS curves are showed in Fig. 8. This 0.6-year margin is comparable to the state-of-the-art model because MORPH is the most popular dataset for many researchers studying age estimation. Since our CDCNN method is used for improving the performance of datasets that have insufficient numbers of images and poor-quality images by grasping features from other datasets, this study uses MORPH only in cross-data learning and mainly focuses on datasets lacking high-quality labeled facial images.

On the AFAD Our CNN structure achieves an MAE value of 3.30 when training the network on the AFAD training dataset. AFAD was randomly split into training/testing (85%/15%) sets, and the experiments were repeated 5 times. The average value of the 5 experiments is used as the final result. The quantitative results are summarized in Table 3. A state-of-the-art performance is achieved, although there is no obvious margin.

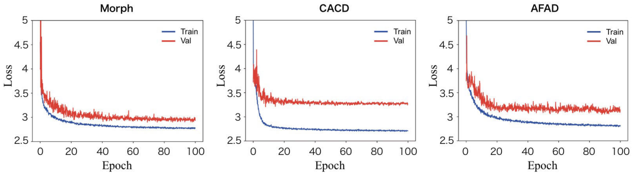


FIGURE 7. Learning curves of MORPH, CACD and AFAD.

TABLE 2. Performance comparison on MORPH [25] (*: the values are estimations from the CS curves in the papers).

Method	MAE	CS
SVR [16]	3.48	78.8%*
OR-CNN [48]	3.27	73.0%*
Ranking-CNN [29]	2.96	85.0%*
ours	**2.76**	**84.9%**
MA-SFV2 [43]	2.68	90.0%*
CORAL [36]	2.64	N/A
DRFs [38]	2.17	91.3%

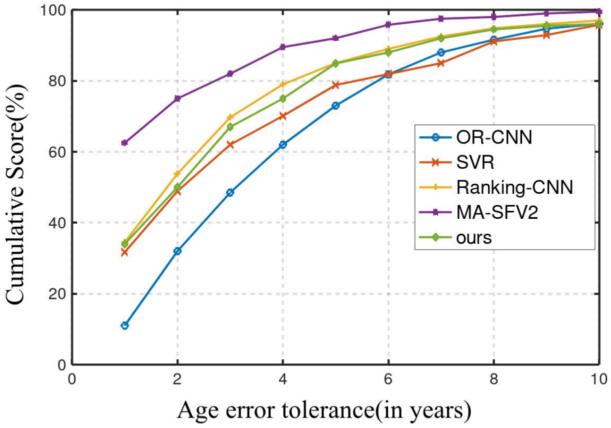


FIGURE 8. CS curve on MORPH.

TABLE 3. Performance (MAE) comparison on AFAD [25].

Method	MAE
BIFS + OHRank [21]	3.84
CORAL [36]	3.48
OR-CNN [48]	3.34
ours	**3.30**

On CACD Our CNN structure achieves an MAE value of 4.58 when fine-tuning the CNN on the CACD training dataset. CACD was randomly split into training/testing (85%/15%) sets, and the experiments were repeated 5 times. The average value of the 5 experiments is used as the final result. The quantitative results are summarized in Table 4. Compared with MORPH and AFAD, CACD has more data but is noisy with partial labeling or image errors. Therefore, only the manually cleaned subset with 18,171 photos was used instead of the more than 100,000 photos in CACD. A state-of-the-art performance is achieved, although there are no obvious margins.

TABLE 4. Performance (MAE) comparison on CACD [4].

Method	MAE
CORAL [36]	5.35
dLDLF [37]	4.73
DRFs [38]	4.64
RNDF [32]	4.60
ours	**4.58**

TABLE 5. Performance (MAE) comparison on AFAD [25] with cross-dataset training.

Method	MAE
BIFS + OHRank [21]	3.84
CORAL [36]	3.48
OR-CNN [48]	3.34
ours (without cross-dataset training)	**3.30**
ours (CDCNN)	**3.11**

TABLE 6. Performance (MAE) comparison on CACD [4] with cross-dataset training.

Method	MAE
CORAL [36]	5.35
dLDLF [37]	4.73
DRFs [38]	4.64
RNDF [32]	4.60
ours (without cross-dataset training)	**4.58**
ours (CDCNN)	**3.96**

2) CROSS-DATASET TRAINING

The regression methods are less stable, especially on datasets lacking high-quality labeled facial images, such as CACD and AFAD, than the classification method because of the large gradients, which make it difficult for the whole network to converge. Therefore, our simple classification model can obtain better performance even compared with more complex regression models on the datasets shown in Tables 3 and 4.

It is difficult to improve performance using different complicated CNN structures because there are not enough high-quality labeled facial images in datasets such as AFAD and CACD. We propose a CDCNN method to solve this problem and improve the quality of the datasets for age estimation instead of improving the learning model. The AFAD-MORPH and CACD-MORPH datasets are used for training, and the results show great improvement.

AFAD with MORPH Our CNN structure achieves an MAE value of 3.11 when training the network on the AFAD-MORPH training dataset and testing on the AFAD test set. The AFAD-MORPH dataset only uses the data from the split training portion of AFAD to ensure that no testing data are trained. The quantitative results are summarized in Table 5. This cross-dataset training method improves the results by 0.2 years over the results of the previous state-of-the-art methods reported in [48].

CACD with MORPH Our CNN structure achieves an MAE value of 3.96 when training the network on the CACD-MORPH training dataset and testing on the CACD test set. The CACD-MORPH dataset only uses the data from the split training portion of CACD to ensure that no testing data are trained. The quantitative results are summarized in Table 6. To date, this is the first reported work with an MAE value below 4 years on CACD, and it improves the results by nearly 0.7 years compared to those of the state-of-the-art method.

V. CONCLUSION

To address the problem of the insufficient numbers of images and poor-quality images in facial datasets for age estimation, we propose the CDCNN method that jointly trains multiple datasets labeled with different features for age estimation; this method demonstrates state-of-the-art results on CACD and AFAD.

The main contributions of this paper are as follows: (1) To the best of our knowledge, we are the first to jointly train multiple datasets for age estimation through cross-dataset learning. (2) Treating the age estimation problem as a classification task instead of a traditional regression task leads to better performance on facial datasets with insufficient numbers of images and poor-quality images. (3) The proposed CDCNN model for age estimation achieves state-of-the-art accuracy on CACD with an MAE value of 3.96 and on AFAD with an MAE value of 3.11.

However, the accuracy of the baseline network can still be improved, and more facial data could be used for training. In future work, more facial images with age labels and deeper networks, such as residual networks [24], could be used to improve performance. Another option is to make a small-scale efficient model, such as C3AE [7], with only 39.7K parameters, unlike VGG-16, which has 138M. Ultimately, our CDCNN method could also be used for other facial feature estimation tasks, such as gender, race, expression, or health condition estimation.

REFERENCES

[1] A. Lanitis, C. Draganova, and C. Christodoulou, "Comparing different classifiers for automatic age estimation," *IEEE Trans. Syst. Man, Cybern. B, Cybern.*, vol. 34, no. 1, pp. 621–628, Feb. 2004.

[2] 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, Apr. 2002.

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

[4] B.-C. Chen, C.-S. Chen, and W. H. Hsu, "Face recognition and retrieval using cross-age reference coding with cross-age celebrity dataset," *IEEE Trans. Multimedia*, vol. 17, no. 6, pp. 804–815, Jun. 2015.

[5] C. Shan, F. Porikli, T. Xiang, and S. Gong, "Video analytics for busniess intelligence," in *Studies in Computational Intelligence*, vol. 409. Springer, 2012, pp. 1–373. [Online]. Available: https://link.springer.com/book/10.1007/978-3-642-28598-1

[6] T. Cootes and G. Edwards, "Active appearance models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 681–685, Oct. 2001.

[7] C. Zhang, S. Liu, X. Xu, and C. Zhu, "C3AE: Exploring the limits of compact model for age estimation," in *Proc. CVPR*, 2019, pp. 12587–12596.

[8] D. Gabor, "Theory of communication," *J. Inst. Elect. Eng. III, Radio Commun. Eng.*, vol. 93, no. 3, pp. 429–457, Nov. 1946.

[9] D. Lowe, "Object recognition from local scale-invariant features," in *Proc. ICCV*, 1999, pp. 1150–1157.

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

[11] G. Guo and G. Mu, "Human age estimation: What is the influence across race and gender?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jan. 2010, pp. 71–78.

[12] G. Guo, G. Mu, Y. Fu, C. Dyer, and T. Huang, "A study on automatic age estimation using a large database," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep. 2009, pp. 1986–1991.

[13] G. Guo, G. Mu, Y. Fu, and T. S. Huang, "Human age estimation using bio-inspired features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 112–119.

[14] H. Han, C. Otto, and A. K. Jain, "Age estimation from face images: Human vs. Machine performance," in *Proc. Int. Conf. Biometrics (ICB)*, Jun. 2013, pp. 1–8.

[15] H. Han, C. Otto, X. Liu, and A. K. 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.

[16] H. Liao, Y. Yan, W. Dai, and P. Fan, "Age estimation of face images based on CNN and Divide-and-Rule strategy," *Math. Problems Eng.*, vol. 2018, Jun. 2018, Art. no. 1712686.

[17] I. Huerta, C. Fernandez, and A. Prati, "Facial age estimation through the fusion of texture and local appearance descriptors," in *Proc. ECCV Workshops*, 2014, pp. 667–681.

[18] J. Li, S. Xiao, F. Zhao, J. Zhao, J. Li, J. Feng, S. Yan, and T. Sim, "Integrated face analytics networks through cross-dataset hybrid training," in *Proc. 2017 ACM Multimedia Conf.*, 2017, pp. 1531–1539.

[19] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "CAFFE: Convolutional architecture for fast feature embedding," in *Proc. Int. Conf. Multimedia*, 2014, pp. 675–678.

[20] A. Krizhevsky, I. Sutskever, and G. Hinton, "'Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2012, pp. 1097–1105.

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

[22] K. Chen, S. Gong, T. Xiang, and C. C. Loy, "Cumulative attribute space for age and crowd density estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2467–2474.

[23] K. He, G. Gkioxari, and R. Girshick, "Mask R-CNN," in *Proc. ICCV*, 2017, pp. 2980–2988.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016, pp. 770–778.

[25] K. Ricanek and T. Tesafaye, "MORPH: A longitudinal image database of normal adult age-progression," in *Proc. 7th Int. Conf. Autom. Face Gesture Recognit.*, 2006, pp. 341–345.

[26] N. Ramanathan, R. Chellappa, and S. Biswas, "Age progression in human faces: A survey," *J. Vis. Lang. Comput.*, vol. 15, pp. 3349–3361, Oct. 2009.

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

[28] 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.*, vol. 126, no. 2, pp. 1–14, Aug. 2016.

[29] S. Chen, C. Zhang, M. Dong, J. Le, and M. Rao, "Using ranking-CNN for age estimation," in *Proc. ICCV*, 2017, pp. 5183–5192.

[30] S. Escalera, J. Fabian, P. Pardo, X. Baro, J. Gonzalez, H. J. Escalante, and I. Guyon, "ChaLearn 2015 apparent age and cultural event recognition: Datasets and results," in *Proc. Int. Conf. Comput. Vis., ChaLearn Looking People Workshop*, 2015, pp. 1–9.

[31] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–5.

[32] S. Li and K. Cheng, "Visualizing the decision-making process in deep neural decision forest," in *Proc. CVPR*, 2019, pp. 114–117.

[33] T. Ojala, M. Pietikainen, and D. Harwood, "Performance evaluation of texture measures with classification based on Kullback discrimination of distributions," in *Proc. 12th Int. Conf. Pattern Recognit.*, 1994, pp. 582–585.

[34] T. Perrett and D. Damen, "Recurrent assistance: Cross-dataset training of LSTMs on kitchen tasks. in *Proc. Int. Conf. Comput. Vis. Workshop (ICCVW)*, 2017, pp. 1354–1362.

[35] C. Wang and C. Zhong, "Adaptive feature pyramid networks for object detection," *IEEE Access*, vol. 9, pp. 107024–107032, 2021.

[36] W. Cao, V. Mirjalili, and S. Raschka, "Rank consistent ordinal regression for neural networks with application to age estimation," *Pattern Recognit. Lett.*, vol. 140, pp. 325–331, Dec. 2020.

[37] W. Shen, K. Zhao, Y. Guo, and A. Yuille, "Label distribution learning forests," in *Proc. NIPS*, 2017, pp. 834–843.

[38] W. Shen, Y. Guo, Y. Wang, K. Zhao, B. Wang, and A. Yuille, "Deep regression forests for age estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2304–2313.

[39] X. Geng, Z.-H. Zhou, Y. Zhang, G. Li, and H. Dai, "Learning from facial aging patterns for automatic age estimation. in *Proc. ACM Int. Conf. Multimedia*, pp. 307–316, 2006. 1.

[40] X. Geng, "Label distribution learning," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 7, pp. 1734–1748, Jul. 2016.

[41] X. Geng, K. Smith-Miles, and Z. Zhou, "Facial age estimation by learning from label distributions," in *Proc. AAAI*, 2010, pp. 451–456.

[42] X. Liu, S. Li, M. Kan, J. Zhang, S. Wu, W. Liu, H. Han, S. Shan, and X. Chen, "Agenet: Deeply learned regressor and classifier for robust apparent age estimation," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2015, pp. 16–24.

[43] X. Liu, Y. Zou, H. Kuang, and X. Ma, "Face image age estimation based on data augmentation and lightweight convolutional neural network," *Symmetry*, vol. 12, no. 1, p. 146, Jan. 2020.

[44] Y. H. Kwon and D. V. Lobo, "Age classification from facial images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Mar. 1994, pp. 762–767.

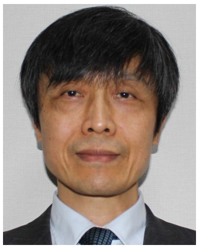
[45] Z. Song, B. Ni, D. Guo, T. Sim, and S. Yan, "Learning universal multi-view age estimator using video context," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Nov. 2011, pp. 241–248.

[46] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.

[47] Z. Kuang, C. Huang, and W. Zhang, "Deeply learned rich coding for cross-dataset facial age estimation," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW)*, Dec. 2015, pp. 338–343.

[48] Z. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua, "Ordinal regression with multiple output CNN for age estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4920–4928.

BEICHEN ZHANG received the B.S. degree in electronic engineering from Tsinghua University, China, in 2014, and the M.S. degree in integrated circuit engineering from the Institute of Semiconductors, Chinese Academy of Sciences, China, in 2017. He is currently pursuing the Ph.D. degree in information engineering with the Graduate School of Integrative Science and Engineering, Tokyo City University, Japan. His research interests include artificial intelligence, computer vision, and image processing.



YUE BAO (Member, IEEE) received the Ph.D. degree in systems innovation from Kanazawa University, in 1996. He is currently working as a Professor with the Graduate School of Integrative Science and Engineering, Tokyo City University, Japan. His research interests include 3-D display, computer graphics, and image processing.