Automatic Age Estimation from Face Images via Deep Ranking

Huei-Fang Yang[1]
hfyang@citi.sinica.edu.tw
Bo-Yao Lin[2]
boyaolin@iis.sinica.edu.tw
Kuang-Yu Chang[2]
kuangyu@iis.sinica.edu.tw
Chu-Song Chen[2]
song@iis.sinica.edu.tw

[1] Research Center for Information
Technology Innovation
Academia Sinica
Taipei, Taiwan

[2] Institute of Information Science
Academia Sinica
Taipei, Taiwan

Abstract

Automatic age estimation (AAE) from face images is a challenging problem because of large facial appearance variations resulting from a number of factors, *e.g.*, aging and facial expressions. In this paper, we propose a generic, deep ranking model for AAE. Given a face image, our network first extracts features from the face through a scattering network (ScatNet), then reduces the feature dimension by principal component analysis (PCA), and finally predicts the age via category-wise rankers. The robustness of our approach comes from the following characteristics: (1) The scattering features are invariant to translation and small deformations; (2) the rank labels encoded in the network exploit the ordering relation among labels; and (3) the category-wise rankers perform age estimation within the same group. Our network achieves superior performance on a large-scale MORPH dataset and two expression ones, Lifespan and FACES.

1 Introduction

A human face reveals great amounts of information about an individual, including identity, age, gender, and emotion. This paper focuses on automatic age estimation (AAE) from face images, which amounts to determining the exact age or age group of a face image according to features from faces.

Although great effort has been devoted to AAE [2, 15, 25], it remains a challenging problem. The main difficulties come from the large facial appearance variations because of a mixture of extrinsic and intrinsic factors. Among all the factors, aging is one main cause of the facial appearance changes. The aging process is complicated and varies greatly for different individuals. Consequently, differences in aging patterns are observed between males and females and between different races. Facial expression is another one. It introduces facial changes similar to the ones caused by aging. For example, the smiling expression of a young person creates skin folds from each side of the nose to the corners of the mouth, and such nasolabial folds are commonly noticeable in an elder adult. AAE algorithms need to overcome heterogeneity in facial appearance changes due to these two factors to provide accurate age estimates.

Pages 55.1-55.11
DOI: https://dx.doi.org/10.5244/C.29.55

1.1 Related Work

Human Age Estimation Features are of importance for recognition and detection problems in computer vision. Earlier popular face descriptors are local binary patterns (LBP) [1] and Gabor features. Later, Geng et al. [9] proposed the AGing pattErn Subspace (AGES) to model aging patterns for AAE. Guo et al. [13] investigated biologically inspired features (BIF) obtained from a model that imitates the functionality of simple and complex cells in the visual cortex and showed that the BIFs have greatly reduced the estimation error on several face datasets. More recently, deep learning approaches capable of learning descriptors from data have received much attention. Yi et al. [25] proposed a multi-scale CNN that learns age estimation, gender classification, and ethnicity classification from raw global face images and local patches. Wang et al. [22] used representations from different layers in a CNN as facial features. These extracted features are coined the deep learned aging pattern (DLA), designed on the basis of that each layer's features exhibit particular activity patterns for a given face image

After the feature extraction, age prediction is the next step. Generally, the age prediction can be cast as a regression or a classification problem. Support vector regressions (SVRs) and support vector machines (SVMs) are the popular regression and classification models, respectively. Nevertheless, neither regression nor classification models take into account the order information between the data labels; hence, ranking approaches have received much attention recently. Yang et al. [24] employed the RankBoost to build the ranking model from pairwise samples for feature selection on the Haar features. Chang et al. [5] proposed the Ordinal Hyperplanes Ranker (OHRank), which decomposes the age estimation problem into a set to binary classifications according to the relative order among labels and aggregates the results of these binary classifiers in the age estimation phase.

Cross Expression Age Estimation A systematic study has shown that expression changes have a prominent influence on age estimation [12]. Guo and Wang [12] proposed to learn the correlations between neutral and other expressions for cross expression age estimation. Their method assumes that pairs of expressions from the same individuals are available in the training samples, making it inappropriate for cases where this assumption is violated. For improvement, Zhang and Guo [26] proposed a weighted random subspace framework that overcomes the aforementioned limitation of [12]. However, their framework is limited in estimating age across only two expressions. Observing this limitation, Alnajar et al. [2] introduced a graphical model with a latent layer to jointly learn the age and all expressions. The latent layer is designed to capture the relationship between ages, expressions, and features, thereby achieving expression invariant age estimation.

1.2 Proposed Approach

We propose a generic, deep network model, DeepRank+, for inferring human age from a face image (see Figure 1). Our model consists of a wavelet scattering network (ScatNet) [3] with 3 layers (L_{1-3}), followed by a dimensionality reduction component by PCA and 3 fully-connected layers (L_{4-6}). ScatNet extracts face representations robust to translations and small deformations; PCA is to reduce the high dimension of the concatenated scattering coefficients (SCs, denoted by the black arrows) produced by each node in ScatNet; the fully-connected network learns to predict the age rank. We construct the output layer with the category and rank labels so that each category has its own age estimator. If the category labels are not used in the training, the output layer can be constructed with only the rank labels, and we term it DeepRank.

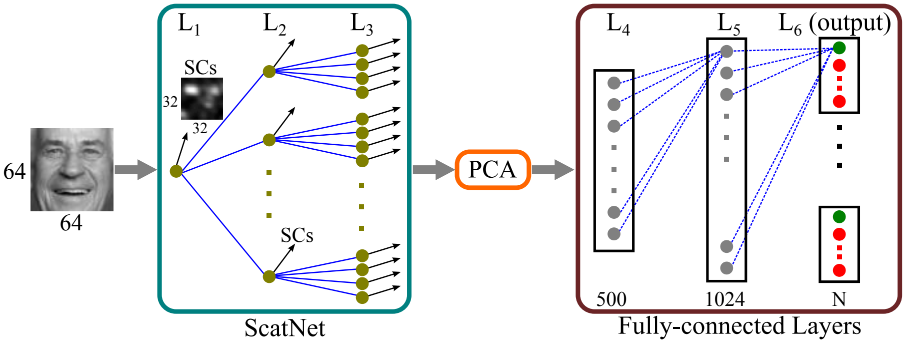

Figure 1: Our network model for human age prediction from face images. Our DeepRank+ comprises a 3-layer ScatNet, a dimensionality reduction component by principal component analysis (PCA) that reduces the feature dimension to 500, and a 3-layer fully-connected network with an architecture of 500-1024-N. The number of nodes N is application dependent. We use ReLUs in layer L_5 and sigmoids in layer L_6. The output layer is constructed with the category labels (the green nodes) and ranking ones (the red). When the category information is not used in training, the output layer can be constructed simply with the rank labels; such a network is coined as DeepRank.

Our contributions are: (1) We develop a generic, deep ranking model that combines ScatNet and the ordinal ranking for AAE from face images. Our ranking method is point-wised and thus is easily scaled up to large-scale datasets; (2) our deep ranking model is general and can be applied to age estimation from faces with large facial appearance variations as a result of aging or facial expression changes; and (3) we show that the high-level concepts learned from large-scale neutral faces can be transferred to estimating ages from faces under expression changes, leading to improved performance. We notice that the idea of transfer learning has been adopted in the recognition task through deep convolutional networks [19]; however, to our knowledge, this is the first time that transfer learning via deep networks is used in human age estimation.

2 Deep Ranker

We discuss the main components of our network in this section, including (1) ScatNet that extracts face representations, (2) DeepRank that estimates age from a face image, and (3) DeepRank+ that performs category-wise age estimation.

2.1 ScatNet as Facial Descriptors

ScatNet can eliminate variability resulting from translations, rotations or scaling such that the scattering descriptors it produces are adequate for face images with displacements and deformations [4, 6]. Furthermore, the scattering representations have been proven to be effective in recognizing handwritten digits and discriminating textures [3]. In general, ScatNet is a deep convolutional network of specific characteristics. First, it uses predefined wavelets. The scattering representations are then computed by a cascade of wavelet transforms and modulus pooling operators from shallow to deep layers. As such, unlike standard CNNs, ScatNet requires no learning of the parameters. Moreover, each node outputs the scattering coefficients (SCs) that can preserve high-frequency information. With the nonlinear modulus and averaging operators, ScatNet can produce representations that are discriminative as well as invariant to translation and small deformations. Readers are advised to refer to [3] for more details. As ScatNet provides fundamentally invariant representations for discriminating feature extraction, only the weights of the fully-connected layers (L_{4-6}) are learned in our network model, which considerably reduces the training time.

In our design, ScatNet uses predefined Morlet wavelets of 5 scales and of 6 orientations. The SCs (the black arrows in Figure 1) yielded by each node are concatenated, forming a face representation of 130,048 ($32 \times 32 \times 127$) dimensions per image. Note that our age ranker relies merely on the SCs from holistic images for age prediction. SCs from parts of a face are not used in our settings. Then, we apply PCA to reduce the dimension of SCs to 500, which is fed to layer L_4.

2.2 DeepRank: An Ordinal Regression Ranker

Unlike the methods of [21] that every pair of images serves as an input to neural networks and then the results are aggregated for ranking, we design a method that infers the ranks directly from a single image. Our method does not rely on pairwise sample sets and thus is easily scaled up to large-scale datasets.

Rank Encoding We employ the reduction framework [17] to conduct a deep ranker. The merits of the reduction frameworks are that (1) it has a theoretical guarantee of the cost bound, and that (2) when it is applied in a cost-sensitive setting where the errors are measured by the mean absolute error (MAE), the cost is 1 for all data samples. Given a set of training samples $X = \{(x_i, y_i), i = 1, \ldots, N\}$, let $x_i \in R^D$ be the input image and y_i be a rank label ($y_i \in \{1, \ldots, K\}$), respectively, where K is the number of age ranks. For rank k, we separate X into two subsets, X_k^+ and X_k^-, as follows:

$$\begin{aligned} X_k^+ &= \{(x_i, +1) | y_i > k\} \\ X_k^- &= \{(x_i, -1) | y_i \leq k\}. \end{aligned} \tag{1}$$

Next, we use the two subsets to learn a binary classifier from the network, which then conducts an answer to the query: “Is the image with a ranking score higher than k?” To each query, we simply make a binary decision between the positive and negative sides. Hence, each query reduces the age rank estimation task to a binary classification problem. A series of query results imply the ordinal relationship between the rank labels, where each query identifies the preferred classes.

According to the above reason, because the k-th binary classifier focuses on determining whether the age rank of an image is greater than k, $K-1$ such binary classifiers are required for K age ranks. Therefore, the output layer can be encoded in an intuitive-to-understand structure: For a face image with true age k, the teaching vector with length $K-1$ is designed as $\left[1, \ldots, 1, -1, \ldots, -1\right]$, where the first $k-1$ values are 1 and the remaining -1.

Prediction Our ranking approach follows a divide-and-conquer strategy: ranking estimation is divided into a set of binary classification problems; when each binary classifier performs well on its own task, the ranking performance of the aggregated results is ensured. Accordingly, to predict the age from a face image x_i, we use the following aggregation rule suggested in [5, 17] for integrating the network's binary outputs into a rank $r(x_i)$:

$$r(x_i) \equiv 1 + \sum_{k=1}^{K-1} [\![O_k(x_i) > 0]\!], \tag{2}$$

where $O_k(x_i)$ is the output of the kth node, and $[\![\cdot]\!]$ is 1 if the condition is met and 0 otherwise.

We notice that Cheng et al. [7] encoded the rank labels in a way similar to ours. However, their prediction is based on a decision rule that scans the output nodes in an order and selects the last one whose output is greater than a threshold T. Such a method may yield inaccurate estimates when inconsistencies in the outputs occur. In contrast, our aggregation rule gives the estimates through a summation of the outputs and thus does not rely on any particular order of the outputs.

2.3 DeepRank+: A Multi-task Ranker

The performance of age estimation algorithms can be affected by several factors such as race, gender, or expression. Guo and Mu [10] showed that crossing race or gender results in significant errors in age estimation, and better performance can be obtained by performing age estimation within the same race/gender category. Motivated by this, we introduce DeepRank+, a generalization of DeepRank, that learns to perform category-wise age estimation via embedding the category information in the label encoding.

Category-wise Label Encoding Category-wise age estimation is commonly carried out in a hierarchical manner, first performing between-category classification and then within-category age estimation [10, 15]. Employing a hierarchical strategy in our network involves introducing more layers, which may result in more computational costs. We instead concatenate the encoding of multiple category-wise rankers altogether in the output layer. This design allows our network to learn category-wise age estimation while making all estimators learn from the shared high-level representations.

Assume there are C category-wise rankers. The encoding for each ranker consists of two constituents: the category label(s) and the age rank. That is, for category j, its encoding is given as $e_j = [g_j, r_j]$, where g_j denotes the desired output of the category and r_j denotes the teaching vector of the rank presented in Sec 2.2. Concatenating the C encoding sets forms a final encoding: $E = \left[g_0, r_0, \ldots, g_j, r_j, \ldots, g_C, r_C\right]$.

We employ different encoding strategies for the category and rank labels in E. The former is regarded as a classification problem, and each face image contributes to the learning. For a face image in category j, we set g_j to 1 and the remaining $g_{m\neq j}$ to -1. As we aim at designing a network that performs category-wise age estimation, the latter is learned from the face images within the same category. That is to say, only the face images in category j contribute to the learning of the ranker for category j. To this end, we employ the “don’t care” in the rank labels. Specifically, when a face image with age k belongs to category j, the teaching vector is defined as:

$$\left[\overbrace{-1}^{g_1},\underbrace{\overbrace{0,\ldots,0}^{r_1}}_{K-1},\ldots,\overbrace{1}^{g_j},\overbrace{\underbrace{1,\ldots,1}_{k-1},\underbrace{-1,\ldots,-1}_{K-k}}^{r_j},\ldots,\overbrace{-1}^{g_C},\underbrace{\overbrace{0,\ldots,0}^{r_C}}_{K-1}\right], \tag{3}$$

where 0s denote “don’t care”.

We implement our approach by using CAFFE [16], which provides the “don’t care” in the label encoding. Our network is trained through a minimization of the cross entropy loss.

Prediction To predict the age from a face image x_i, we first determine to which category the face image belongs, which amounts to examining the category outputs and selecting the one with the highest probability. Then, we use Equation (2) to aggregate the ranking results of the determined category.

Table 1: Statistics on data splits of MORPH for the experiments. The MORPH dataset is divided into 3 non-overlapping subsets, S_1, S_2, and S_3 (Other).

Ethnicity	S_1		S_2		S_3 (Other)	
	Female	Male	Female	Male	Female	Male
Black	1285	3980	1285	3980	3187	28843
White	1285	3980	1285	3980	31	39
Others*	—	—	—	—	129	1843

*Races include Hispanic, Asian, and Other.

3 Experimental Results

We conducted two sets of experiments on three face datasets. One is the experiments on age estimation from faces obtained from different gender and races on a large-scale MORPH dataset (Section 3.1). The other is the experiments on age estimation from faces under expression changes from the Lifespan (Section 3.2) and FACES datasets (Section 3.3). The performance is evaluated by the mean absolute error (MAE). It is defined as $\frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$, where y_i is the true age, $\hat{y}_i$ the estimated age, and N the number of test images. The lower the MAE is, the better performance an age estimator attains.

3.1 Age Estimation on MORPH

Dataset The MORPH Album 2 (a.k.a. MORPH) [20] comprises more than 55,000 face images of 13,000 individuals aged from 16 to 77 years (and thus there are 62 age ranks). On average, each individual has 3∼4 images. MORPH is the largest corpus of publicly available longitudinal images with detailed age, gender, and ethnicity (*e.g.*, Black, White, Hispanic, Asian, and Other) labels.

Settings We followed the training/testing settings of Guo and Mu [10] in the experiments. MORPH is partitioned into 3 non-overlapping subsets, namely, S_1, S_2, and S_3 (Other). Table 1 shows the detailed distributions of these splits. S_1 and S_2 can be employed in either training or testing whereas S_3 is used only for testing. That is, when a model is trained on S_1 (S_2), its performance is evaluated on $S_2 \cup S_3$ ($S_1 \cup S_3$). These two setups are referred to as settings A and B, respectively. Besides, all the face images are aligned on the basis of the positions of eyes and resized to 64×64.

Network Models When compared to other methods, DeepRank and DeepRank+ are set as follows. For DeepRank, we set the number of nodes in L_6 to 61 (*i.e.*, $62-1$). For DeepRank+, we consider the combinations of ethnicity (Black (B) and White (W)) and gender (Female (F) and Male (M)) the category labels, which results in 4 categories, namely, BF, BM, WF, and WM. Thus, layer L_6 consists of 248 (*i.e.*, $(1+61)\times 4)$) nodes.

Quantitative Comparison Table 2 shows the comparison of DeepRank and DeepRank+ to other alternatives, including CNN-based [22, 23, 25] and non-CNN-based [11, 15] approaches. On age prediction, our DeepRank attains a MAE of 3.57 years, performing favorably against the state-of-the-arts. When category information is also considered, DeepRank+ further reduces the MAE to 3.49 years. This indicates that performing category-wise age estimation can be useful in determining facial appearance differences between categories due to the aging process. Note that although our approach and other CNNs take advantage of deep learning, the underlying features are of broad differences. The CNNs of [22, 23, 25] learn face representations from raw global and/or local regions whereas ours relies on the

Table 2: Performance comparison of various methods on age estimation, gender classification, and ethnicity classification on MORPH. The MAE (in years) and accuracy (%) are evaluation metrics for age estimators and gender/ethnicity classifiers, respectively.

	Age (MAE)			Gender (%)			Ethnicity (%)		
Method	A[a]	B	Mean	A	B	Mean	A	B	Mean
KCCA [11]	4.00	3.95	3.98	98.5	98.4	98.5	98.9	99.0	99.0
Han et al. [15][b]	—	—	3.60	—	—	97.6	—	—	99.1
DLA [22]	—	—	4.77	—	—	—	—	—	—
Baseline CNN [25][c]	4.64	4.55	4.60	—	—	—	—	—	—
Multi-task CNN [25]	3.72	3.54	3.63	98.0	97.8	97.9	99.1	98.1	98.6
DeepRank	3.57	3.57	3.57	—	—	—	—	—	—
DeepRank+	3.48	3.49	3.49	97.9	97.8	97.9	98.7	98.6	98.7

[a] A (B) refers to the experiment setting that uses S_1 (S_2) for training and $S_2 \cup S_3$ ($S_1 \cup S_3$) for testing.

[b] They used a larger set of MORPH (78,207 images), different from our settings that used 55,132 images.

[c] Yi et al. [25] implemented the CNN model proposed by Yang et al. [23] and tested it on MORPH.

scattering descriptors that need not to be learned. These results indicate that the scattering descriptors are effective facial features.

In addition to age estimation, our DeepRank+ can be applied to gender and ethnicity classification. As shown in Table 2, we can see that DeepRank+ yields accuracies comparable to other approaches. Joint learning of age estimation, gender classification, and race classification not only boosts the performance on age estimation but also allows our network to accurately obtain other facial attributes from a face image.

Performance vs. # of Nodes in Layer L_5 We also investigated the performance of DeepRank with respect to the number of nodes in layer L_5. The number of nodes is set to 1024, 2048, and 4096, and the resulting MAEs are 3.57, 3.59, and 3.60 years, respectively. The networks with different number of nodes in L_5 yield comparable performance. This stable performance suggests that our proposed ranking model does not depend on the deliberate tuning of the network architecture.

3.2 Age Estimation on Lifespan

Dataset The Lifespan [18] contains 1,046 images of 580 individuals in various expressions (*e.g.*, anger, annoyed, disgusted, grumpy, happy, neutral, sad, and surprise). Each participant has a neutral expression photo, and part of participants have photos taken in one or more other expressions. The ages range from 18 to 93, resulting in a total of 76 age ranks. This dataset is created for providing faces representative of age groups seen in real life and across the lifespan, with an emphasis on older adults.

Settings Following the experimental settings in [12], we employed a 5-fold cross validation scheme in evaluating the performance of our method. In Lifespan, the number of images in each expression is unbalanced and varies greatly; like [2, 12], faces in neutral and happy expressions are selected for evaluation, which contains 580 neutral and 258 happy faces. In addition, as Lifespan is a relatively smaller dataset compared with MORPH, we expanded the data by including the mirrors of the original face images in the training process.

Network Models As only 838 face images from Lifespan were used in the evaluation, we adopted the idea of transfer learning and fine-tuned a pre-trained network on Lifespan.

Table 3: Performance of different age estimators and expression classifiers evaluated on Lifespan. The MAE (in years) and accuracy (%) are evaluation metrics for age estimators and expression classifiers, respectively. The mean is computed as the weighted average.

	Age Estimation (MAE)			Expression Recognition (%)		
Method	Neutral	Happy	Mean	Neutral	Happy	Mean
Guo et al. [14][a]	—	—	8.85	—	—	91.05
Guo et al. [12]	6.19	6.34	6.24	—	—	—
Zhang et al. [26]	5.99	6.26	6.07	—	—	—
Alnajar et al. [2]	5.72	4.14	5.26	—	—	93.91
DeepRank	5.01	2.72	4.31	—	—	—
DeepRank+	5.64	4.18	5.19	97.93	90.98	95.81

[a] The results are from [2], in which they implemented the method of Guo et al. [14] for a fair comparison.

This pre-trained model (*i.e.*, DeepRank with 1024 nodes in L_5 and 62 nodes in L_6) has learned to estimate ages from neutral faces on MORPH. We set the number of nodes in L_6 of DeepRank to 75 because of the 76 age groups in Lifespan. The facial expressions are deemed category labels (*i.e.*, 2 labels) so the number of nodes in L_6 of DeepRank+ is set to 152 (*i.e.*, $(1+75)\times 2$).

Quantitative Comparison Table 3 shows the performance of various age estimators and facial expression classifiers. Our DeepRank achieves superior performance, a MAE of 4.31 years, performing favorably against other alternatives. Notably, our DeepRank attains a MAE of 2.72 years on age estimation from happy faces, a 34% error reduction compared to the best known MAE of 4.14 years in [2]. DeepRank+, on the other hand, yields a MAE of 5.19 years on age estimation, better than other approaches but worse than DeepRank. Deep learning techniques require large amounts of data, but Lifespan provides relatively fewer training samples in each expression. This could be the reason why DeepRank+ performs slightly worse than DeepRank on small datasets. Though DeepRank+ does not benefit much from category information on age estimation, it demonstrates an improved performance on expression recognition, a 1.90% higher overall accuracy than [2].

3.3 Age Estimation on FACES

Dataset The FACES [8] contains face images of 171 subjects ranging from ages 19 to 80 (62 age ranks). Each individual has faces in two sets of six facial expressions (neutrality, sadness, happiness, disgust, fear, and anger), resulting in 342 images in each expression and 2,052 in total.

3.3.1 Training on FACES

Settings We employed a 5-fold cross validation [12] in the evaluation. In FACES, each expression is of equal number of images, and thus all individual images are used.

Network Models DeepRank has 61 nodes in L_6 due to the 62 age groups in FACES, and DeepRank+ has 372 nodes because 62 (1 category and 61 rank labels) nodes are required for each category-wise ranker and there are 6 expressions. Like the experiments on Lifespan, the weights of the networks are initialized as those of a network pre-trained on MORPH.

Table 4: Performance of different age estimators and expression classifiers evaluated on FACES. The MAE (in years) and accuracy (%) are evaluation metrics for age estimators and expression classifiers, respectively. Neu., ang., dis., fea., and hap. denote neutral, angry, disgusted, fearful, and happy, respectively. FER denotes facial expression recognition.

	Age Estimation (MAE)							FER (%)
Method	Neu.	Ang.	Dis.	Fea.	Hap.	Sad	Mean	Mean
Guo et al. [14][a]	—	—	—	—	—	—	9.94	84.68
Guo et al. [12]	—	—	—	—	—	—	9.27	—
Alnajar et al. [2]	5.97	8.21	8.17	8.25	6.77	7.07	7.41	92.19
Zhang et al. [26] w/ Joint	—	—	—	—	—	—	7.13	—
DeepRank	6.19	8.43	8.20	6.71	8.00	6.86	7.40	—
DeepRank+	6.10	8.26	7.82	7.27	8.41	6.97	7.47	93.92
DeepRank w/ Joint	5.99	7.12	8.15	6.35	7.77	6.68	7.01	—
DeepRank+ w/ Joint	5.85	7.87	7.80	6.66	7.49	6.59	7.04	94.12

[a] The results are from [2], in which they implemented the method of Guo et al. [14] for a fair comparison.

Quantitative Comparison We can see from Table 4 that DeepRank yields better performance than most of the approaches on age estimation. The performance of DeepRank+ is comparable but slightly worse than DeepRank. This is because DeepRank+ requires more training samples in each expression to learn category-wise age estimation. These results motivate us to investigate whether including additional datasets in the training can improve the estimation performance on FACES, which is presented in the following.

3.3.2 Training on Extended Data

Settings We enriched the training data by including the 580 neutral and 258 happy faces from Lifespan in the experiment of FACES for performance comparison.

Network Models DeepRank has 75 nodes in L_6 because of the 76 age groups in the samples with additional data from Lifespan, and DeepRank+ has 456 (*i.e.*, (1 (category) + 75 (ranks)) × 6 (expressions)) nodes. In the following, we denote the networks trained on the FACES and Lifespan by DeepRank w/ Joint and DeepRank+ w/ Joint for clarity. The weights of all the networks are initialized as those of a network pre-trained on MORPH.

Quantitative Comparison As shown in Table 4, the performance of both DeepRank and DeepRank+ have been further improved by including the extended training samples from Lifespan. On age estimation, DeepRank w/ Joint and DeepRank+ w/ Joint attain MAEs of 7.01 and 7.04 years, respectively; both networks yield comparable results, achieving state-of-the-art performance. On recognizing facial expressions, DeepRank+ and DeepRank+ w/ Joint achieve 1.73% (93.92% vs. 92.19%) and 1.93% (94.12% vs. 92.19%) higher accuracies than the best results reported in [2], respectively. Hence, with sufficient amounts of training samples, our proposed approach is better on estimating ages and recognizing expressions.

4 Conclusions

We have presented a generic, deep ranking model for automatic age estimation from a single face image. Our model employs ScatNet, a deep CNN of specific characteristics, for extracting the facial features insensitive to translation and local deformation. Our encoding of the rank labels takes advantage of ordinal regression. In combination with category information in the label encoding, our network jointly learns the categories and ages and thus performs age estimation in a category-wise manner. Our approach attains superior performance to several state-of-the-arts on the MORPH, Lifespan, and FACES datasets.

References

[1] Timo Ahonen, Abdenour Hadid, and Matti Pietikäinen. Face description with local binary patterns: Application to face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(12):2037–2041, 2006.

[2] Fares Alnajar, Zhongyu Lou, José Manuel Álvarez, and Theo Gevers. Expression-invariant age estimation. In *Proc. BMVC*, 2014.

[3] Joan Bruna and Stéphane Mallat. Invariant scattering convolution networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1872–1886, 2013.

[4] Kuang-Yu Chang and Chu-Song Chen. A learning framework for age rank estimation based on face images with scattering transform. *IEEE Trans. Image Processing*, 24(3): 785–798, 2015.

[5] Kuang-Yu Chang, Chu-Song Chen, and Yi-Ping Hung. Ordinal hyperplanes ranker with cost sensitivities for age estimation. In *Proc. CVPR*, pages 585–592, 2011.

[6] Kuang-Yu Chang, Chu-Song Chen, and Yi-Ping Hung. Intensity rank estimation of facial expressions based on a single image. In *Proc. SMC*, pages 3157–3162, 2013.

[7] Jianlin Cheng, Zheng Wang, and Gianluca Pollastri. A neural network approach to ordinal regression. In *Proc. IJCNN*, pages 1279–1284, 2008.

[8] Natalie C. Ebner, Michaela Riediger, and Ulman Lindenberger. FACES–a database of facial expressions in young, middle-aged, and older women and men: Development and validation. *Behavior Research Methods*, 42(1):351–362, 2010.

[9] Xin Geng, Zhi-Hua Zhou, and Kate Smith-Miles. Automatic age estimation based on facial aging patterns. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(12):2234–2240, 2007.

[10] Guodong Guo and Guowang Mu. Human age estimation: What is the influence across race and gender? In *Proc. CVPRW*, 2010.

[11] Guodong Guo and Guowang Mu. Joint estimation of age, gender and ethnicity: CCA vs. PLS. In *Proc. FG*, pages 1–6, 2013.

[12] Guodong Guo and Xiaolong Wang. A study on human age estimation under facial expression changes. In *Proc. CVPR*, pages 2547–2553, 2012.

[13] Guodong Guo, Guowang Mu, Yun Fu, and Thomas S. Huang. Human age estimation using bio-inspired features. In *Proc. CVPR*, pages 112–119, 2009.

[14] Guodong Guo, Rui Guo, and Xin Li. Facial expression recognition influenced by human aging. *IEEE Trans. Affective Computing*, 4(3):291–298, 2013.

[15] Hu Han, Charles Otto, Xiaoming Liu, and Anil K. Jain. Demographic estimation from face images: Human vs. machine performance. *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(6):1148–1161, 2015.

[16] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross B. Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proc. ACM MM*, pages 675–678, 2014.

[17] Hsuan-Tien Lin and Ling Li. Reduction from cost-sensitive ordinal ranking to weighted binary classification. *Neural Computation*, 24(5):1329–1367, 2012.

[18] Meredith Minear and Denise C. Park. A lifespan database of adult facial stimuli. *Behavior Research Methods, Instruments, & Computers*, 36(4):630–633, 2004.

[19] Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *Proc. CVPR*, pages 1717–1724, 2014.

[20] Karl Ricanek and Tamirat Tesafaye. MORPH: A longitudinal image database of normal adult age-progression. In *Proc. FGR*, pages 341–345, 2006.

[21] Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, and Ying Wu. Learning fine-grained image similarity with deep ranking. In *Proc. CVPR*, pages 1386–1393, 2014.

[22] Xiaolong Wang, Rui Guo, and Chandra Kambhamettu. Deeply-learned feature for age estimation. In *Proc. WACV*, pages 534–541, 2015.

[23] Ming Yang, Shenghuo Zhu, Fengjun Lv, and Kai Yu. Correspondence driven adaptation for human profile recognition. In *Proc. CVPR*, pages 505–512, 2011.

[24] Peng Yang, Lin Zhong, and Dimitris N. Metaxas. Ranking model for facial age estimation. In *Proc. ICPR*, pages 3404–3407, 2010.

[25] Dong Yi, Zhen Lei, and Stan Z. Li. Age estimation by multi-scale convolutional network. In *Proc. ACCV*, pages 144–158, 2014.

[26] Chao Zhang and Guodong Guo. Age estimation with expression changes using multiple aging subspaces. In *Proc. IEEE BTAS*, pages 1–6, 2013.