Deeply-learned Hybrid Representations for Facial Age Estimation

Zichang Tan1,2, Yang Yang1,2, Jun Wan1,2,*, Guodong Guo3,4, Stan Z. Li1,2,5
1CBSR&NLPR, Institute of Automation, Chinese Academy of Sciences, Beijing, China;  
2University of Chinese Academy of Sciences, Beijing, China;  
3Institute of Deep Learning, Baidu Research, Beijing, China;  
4National Engineering Laboratory for Deep Learning Technology and Application, Beijing, China;  
5Faculty of Information Technology, Macau University of Science and Technology, Macau, China.  
{zichang.tan, yang.yang, jun.wan, szli}@nlpr.ia.ac.cn, guoguodong01@baidu.com

Abstract

In this paper, we propose a novel unified network named Deep Hybrid-Aligned Architecture for facial age estimation. It contains global, local and global-local branches, which are jointly optimized and thus can capture multiple types of features with complementary information. In each sub-network of each branch, we employ a separate loss to extract the independent region features and use a recurrent fusion to explore correlations among them. Considering that pose variations may lead to misalignment in different regions, we design an Aligned Region Pooling operation to generate aligned region features. Moreover, a new large private age dataset named Web-FaceAge owning more than 120K samples is collected under diverse scenes and spanning a large age range. Experiments on five age benchmark datasets, including Web-FaceAge, Morph, FG-NET, CACD and Chalearn LAP 2015, show that the proposed method outperforms the state-of-the-art approaches significantly.

1 Introduction

Nowadays, accurate age estimation remains a challenging problem due to the variations including the intrinsic factors in facial aging (e.g., various genes) and the complex variations in face images (e.g., different face poses and camera viewing angles). Most of existing facial age estimation methods either learn the global representations from the whole face [Rothe et al., 2018; Niu et al., 2016; Tan et al., 2018; Gao et al., 2018; Li et al., 2018; Pan et al., 2018] or extract local features by using local hand-crafted descriptors like BIF [Guo and Mu, 2013]. However, we consider it is sub-optimal to learn global or local features: 1) with only global context, it may ignore crucial details, like beard and wrinkles; 2) with only local details, it may lose the structural information, and hardly achieve a robust description. Motivated by our human visual system which often leverages both the global context and local details to capture the effective information [Li et al., 2017b], we consider both global and local information for facial age estimation.

*Corresponding Author

To that end, we propose a new Deep Hybrid-Aligned Architecture (DHAA) for age estimation. It consists of global, local and global-local branches, which learns coarse-grained global context, fine-grained local details (e.g., wrinkles) and integrated configuration (global-local information), respectively. In each branch, we use several sub-networks to extract features on different regions (global, local or global-local). To fully explore the information among them and extract reliable features for age estimation, we adopt the following strategies: firstly, we force a separate loss function in each sub-network to learn region features independently and use one loss function in the final layer to guide the whole network; then, we employ the recurrent fusion based on Long Short Term Memory (LSTM) [Hochreiter and Schmidhuber, 1997] to explore the spatial correlation among different face regions in each branch; finally, all branches are jointly optimized to capture the complementary information among them. Our DHAA is different from previous works [Yi et al., 2014; Li et al., 2017a; Singh and Lee, 2016] of global and local features learning, which neglect the correlation among different regions and do not involve the integrated features learning.

Considering that our human face has a good structure, we can select some interested regions according to the facial landmarks, e.g., eyes, nose and mouth regions. Most existing studies [Wei et al., 2017; Zhao et al., 2018] use ROI pooling to generate the features from the interested regions. However, it may hardly deal with the region misalignment caused by large face pose variations. Thus, we propose Aligned Region Pooling (ARP) to generate aligned region features based on the affine transformation with given some source and target key points. It can reduce the misalignment in the selected regions, which helps the network to extract more effective region features.

Moreover, the lack of large-scale dataset may hinder the development of facial age estimation. Currently, there are two large age datasets, IMDB-WIKI [Rothe et al., 2018] and CACD [Chen et al., 2014], which contain more than 100K face images. However, as pointed out by the previous work [Tan et al., 2018], IMDB-WIKI contains many noises like inaccurate labels and images containing no face or multiple faces. Thus, it is usually used for pretraining rather than evaluation. In CACD, all images come from only 2000 celebrities, which is relatively small. In addition, the age ranges in the existing datasets are also relatively small, i.e., the age
ranges of Morph II and CACD are [16, 77] and [14, 62], respectively. Therefore, we decide to collect a new large age dataset with a large number of samples and instances, diverse scenes, and a large age range.

Our contributions include:

1. A unified network for jointly learning the global context, local details and the integrated configuration, named Deep Hybrid-Aligned Architecture (DHAA), is proposed to solve the task of age estimation. To our best knowledge, there is no previous work that integrates these many cues for aging analysis.

2. A new pooling method, called Aligned Region Pooling (ARP), is proposed to generate aligned region features, which can cope with the facial region misalignment caused by large pose variations.

3. A new large private age dataset (Web-FaceAge) is collected, with diverse scenes, a large age range, and a large number of images and identities is collected.

4. Better results are achieved than the state-of-the-art on five age benchmark datasets, i.e., Web-FaceAge, Morph II, FG-NET, CACD, and Chalearn LAP 2015.

2 Our Approach

2.1 The Overall Design

The overview of the proposed network is shown in Fig. 1. The network is constructed based on ResNet-18 [He et al., 2016]. It contains three branches: global branch, local branch and global-local branch, which learn global context, local details and integrated features, respectively. For each branch, it consists of three sub-networks to extract the features on different regions, including global and local regions. In our implementation, the global regions mean the regions containing the entire face, and local regions are the regions around the eyes, nose and mouth. Considering the face symmetry and the image flipping in both training and test stage, we random select the local region around the left eye for experiments. The selected global and local regions can be found in Fig. 2.

Specially, the low level features are first shared for all branches, which allows the knowledge transfer between different branches and also helps to reduce the risk of overfitting.

Then, we generate the aligned region features by using the proposed ARP to cope with the misalignment caused by large pose variations. Moreover, each branch employs a recurrent fusion to explore the spatial correlation and the dependency between different face regions. Finally, all branches are jointly optimized and combined together to generate reliable and discriminative features.

2.2 Aligned Region Feature Generation

Although the input face image is aligned, the region parts may be not aligned. To address it, we propose Aligned Region Pooling (ARP) to generate aligned region features. The ARP is constructed based on the affine transformation, with given some source points and target points. We denote the source key points as \( R_s = \{p_1, \cdots, p_m\} \), where \( p_j \) is the coordinates of the \( j^{th} \) source point, and the target key points as \( R_t = \{q_1, \cdots, q_m\} \), where \( q_j \) denotes the coordinates of the \( j^{th} \) target point. Given \( R_s \) and \( R_t \), the parameter matrix \( M_A \) of the affine transformation \( T(\cdot; M_A) \) can be computed by minimizing Least Squares Error (LSE):

\[
\sum_{p_j \in R_s, q_j \in R_t} \|q_j - T(p_j; M_A)\|^2 \quad (1)
\]

In our experiments, we first calculate the \( M_A \) according to the source and target key points as shown in Fig. 2. Then,
based on the input features $X^s$ produced by the block $C^{B_i}$, the aligned features $X^t$ can be calculated as $X^t = T(X^s; M_A)$. The height and width coordinates of both source and target feature maps are normalized to $[-1, 1]$ according to the work [Jaderberg et al., 2015]. The aligned features $X^t$ are then adapted to the specific resolution (e.g., $56 \times 56$) before being fed into the subsequent blocks. To be specific, we use the ARP module to generate 5 aligned region features (see Fig. 2), including 2 global regions (the scales are different with one face is 1.2 times larger than the other) and 3 local regions (keep the left eye, nose and mouth in center). For example, the global features are aligned by using five key points: two eyes, nose tip and two mouth corners. Note that we simply use the entire features of $C^{B_i}$ as the first global feature, because the face is aligned in advance.

2.3 Feature Extraction

In our network, global and local branches employ the raw features or the aligned region features from $C^{B_i}$ block as input, while global-local branch learns the integrated configuration by using the concatenation of global and local region features in the second block layer as the input (see Fig. 1). To be specific, for the block $C^{B_4}_{c_k}$ in the global-local branch, it takes the concatenated features of $C^{B_4}_{c_6}$ and $C^{B_4}_{c_2}$. It jointly considers both global and local information, which helps the network to capture complementary information.

Each sub-network in the global or local branches owns three residual blocks, while each sub-network in the global-local branch owns two. In each sub-network, the last residual block $C^{B_4}_{c_k}$ or $C^{B_4}_{c_k}$ generates the high dimensional features of 512 channels. It increases the parameters in the later layers especially in the Fully Connected (FC) layer and the LSTM unit, which will lead to overfitting. Therefore, we employ a convolutional layer with only 128 kernels, a Global Average Pooling (GAP) and a FC layer with only 128 neurons followed by the last residual block in each sub-network. Then, each sub-network extracts a feature vector with only 128 dimensions. For convenience, the extracted features for $I_i$ in the $k^{th}$ sub-network of the global, local and global-local branches are denoted as $z^{g_k}_i$, $z^{l_k}_i$ and $z^{g_l}_i$, respectively.

2.4 Recurrent Fusion

In each branch, we adopt a Long Short-Term Memory (LSTM) network and a sum operation to explore the correlation among different region features. Taking the global branch as an example, the encoded hidden state can be represented as:

$$h^{g_k}_i = LSTM(z^{g_k}_i, h^{g_k-1}_i)$$

where $LSTM$ denotes an LSTM unit, which encodes the hidden state $h^{g_k}_i$ by using both the features $z^{g_k}_i$ and the previous hidden state $h^{g_k-1}_i$ as the input. Then, we sum all hidden features to obtain the final global features $h^{g}_i$:

$$h^{g}_i = \sum_k h^{g_k}_i$$

(3)

Recurrent fusion of LSTM is used to explore the spatial dependency and correlation between different regions. With the learned topological contextual information among regions, we can extract more discriminative features for facial age estimation. Similarly, we can obtain the final local and global-local features, i.e., $h^{l}_i$ and $h^{g_l}_i$, respectively. The fusion features $h^{g}_i$, $h^{l}_i$ and $h^{g_l}_i$ encode different facial information for facial age estimation, e.g., $h^{g}_i$ mainly contains the global structure and $h^{l}_i$ mainly encodes the local fine-grained details, and $h^{g_l}_i$ mainly reserves the integrated configuration of global and local information. Consequently, we concatenate these three features to form discriminative and effective features $x_i = concat (h^{g}_i, h^{l}_i, h^{g_l}_i)$ for final age estimation. To our best knowledge, such recurrent feature fusion is the first attempt in the facial age estimation community.

2.5 The Loss Function

Most previous works [Niu et al., 2016; Rothe et al., 2018; Tan et al., 2019] train a network by only employing the loss at the final level, e.g., DEX [Rothe et al., 2018] trains the network with only a softmax loss followed by the final layer. However, it may hardly learn independently in different regions when extracting the region features. Then, we employ multiple objective loss functions to train our DHAA. We enforce a separate objective loss function with the same label constraint in each sub-network, which facilitates each sub-network to learn their own independent discriminative features. Besides, a cross entropy classification loss is followed by the final layer to guide the whole network to extract the effective and discriminative features for final age categorization. Formally speaking, we use the posterior probability $p_j(x_i)$ as the probability of assigning the features $x_i$ to the $j^{th}$ class: $p_j(x_i) = \frac{\exp(w^j_i x_i)}{\sum_j \exp(w^j_i x_i)}$, where $w_j$ denotes the parameters of $j$ class in the softmax classifier. Similarly, we can define $p_j(z^{g_k}_i)$, $p_j(z^{l_k}_i)$, $p_j(z^{g_l}_i)$ for the features $z^{g_k}_i$, $z^{l_k}_i$ and $z^{g_l}_i$ of the $j^{th}$ class, respectively. Thus, the whole training losses can be represented as:

$$L = -\frac{1}{n} \sum_{i=1}^{n} \log(p_{y_i}(x_i)) - \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{3} 3 \log(p_{y_i}(z^{g_k}_i))$$

(4)

where $n$ is the total number of training images, $y_i$ denotes the age label of image $I_i$, and $\phi = \{g, l, c\}$ refers to a set of global, local and global-local branches. In the end, an expectation refinement followed by the final layer is employed to predict the exact age. It can be denoted as: $y'_i = \sum_j j \cdot p_j(x_i)$. Note that the classifiers followed by the sub-network are only employed to assist the network training, and in test we only use the last classifier for final age prediction.

3 Web-FaceAge Dataset

3.1 Data Collection and Processing

Image Crawler and Face Collection. We collect face images from popular search engines, such as Google, Bing, Baidu and Yahoo. Only one image is kept when images have the same url. A face detection method based on faster r-cnn [Ren et al., 2015] is applied to all of the downloaded images, and each face is saved as a single image.

Duplicate Image Removal. We use the pHash method to generate a binary vector representation with 64-bits for each
image. Images compute their similarities by using the hammering distance. In our experiments, if the distance between any two images is smaller than 6, we treat them as the same image and remove one of them randomly.

Age Labeling. At first, the coarse age labels are obtained by some Face APIs, such as the Face++ API\(^1\), AuthenMetric\(^2\). We use the mean values of the above APIs as the coarse labels. Then, we ask more than 20 students and academic staff to refine those coarse labels. It takes about two months to get about 124K face images with apparent age labels.

3.2 Data Statistics

Our collected database comprises 124,917 face images that are accurately labeled by human annotators. Some image samples and labeled ages are given in the Fig. 3. There are large variability in subjects, ethnicity, head poses, scenes, camera angles, lights and so on. The collected Web-FaceAge has a large age range with $[0, 100]$. It especially contains many images of children, which is scarce in the existing datasets. For evaluation purposes, we split the dataset into two sets: 20,000 images are randomly selected for test, and the left (104,917 images) are used for training. Fig. 4 shows the age distributions in training and test sets.

4 Experiments

4.1 Datasets

We evaluate the proposed method on six benchmark datasets: Web-FaceAge, Morph II [Ricanek and Tesafaye, 2006], CACD [Chen et al., 2014], FG-NET\(^3\) and Challearn LAP 2015 [Escalera et al., 2015]. The Web-FaceAge has already been introduced in Section 3, and its description is omitted here. Besides, we also introduce the IMDB-WIKI dataset [Rothe et al., 2018] used for pretraining.

Morph II: This dataset contains 55,134 face images in total. In our experiments, three typical protocols are employed for evaluation: (1) **80-20 protocol**: as in the works [Niu et al., 2016; Gao et al., 2018], the dataset is randomly divided to two parts, where 80% for training and 20% for test. (2) **Partial 80-20 protocol**: Following the works [Rothe et al., 2018; Tan et al., 2018], a subset of 5493 face images from Caucasian descent are used, which reduces the cross-race influence for facial age estimation. Then, we randomly split the subset into two parts: 80% images for training and the left for test. (3) **S1-S2-S3 protocol**: the dataset is split into three non-overlapped subsets S1, S2 and S3 followed by the work [Tan et al., 2018]. All experiments are repeated twice and the averaging results are used for evaluation: 1. training with S1 and test with S2+S3; 2. training with S2 and test with S1+S3.

FG-NET: This dataset contains 1,002 images of 82 subjects. We follow the previous works [Rothe et al., 2018] to take leave-one person-out (LOPO) strategy for evaluation.

CACD: CACD dataset contains more than 160 thousand images of 2,000 celebrities. According to the previous work [Tan et al., 2018], we employ the subset of 1,800 celebrities with less precise labeling for training, and 80 and 120 cleaned celebrities for validation and test, respectively.

Challearn LAP 2015: This dataset is a competition dataset for apparent age estimation. It contains 4691 images, and each image is labeled by an apparent age and a standard deviation. This dataset includes training, validation and test subsets with 2476, 1136 and 1079 images, respectively. We adopts the experimental settings of [Rothe et al., 2018; Tan et al., 2018] for evaluation.

IMDB-WIKI: This dataset is usually used as a pre-training dataset due to its noises. According to the work [Tan et al., 2018; Rothe et al., 2018], we remove non-face images and part of images with multiple faces. To the end, about 300 thousand images are reserved for pretraining.

4.2 Experimental Settings

Pre-processing and Augmentation: The images are aligned with the eyes center and the upper lip two landmarks according to the work [Tan et al., 2018]. To obtain the features with large size for ARP, we adopt the images with the size of $256 \times 256$ as the inputs. Following the settings in [Gao et al., 2018], we augment the face images with random horizontal flipping, scaling, rotation and translation in the training stage. In the test stage, both the test image and its flipped copy are fed into the network, and then the averaging prediction is used as the final prediction. Note that only random horizontal flipping is employed when evaluating on CACD and Web-FaceAge datasets, because those two datasets have a large number of training images.

Training Details: All networks are first pretrained on ImageNet and optimized by SGD with Nesterov momentum. The weight decay and the momentum are set to 0.0005 and 0.9. The initial learning rate is set to 0.01 and reduced by a factor of 10 with number of iteration increases. All models are implemented with PyTorch on GTX 1080Ti GPU.

Evaluation Metrics: According to previous works [Tan et al., 2018; Rothe et al., 2018], Mean Absolute Error
Table 1: The analysis of global, local and global-local features learning (denoted by G, L and G-L, respectively) on Morph II and Web-FaceAge datasets.

<table>
<thead>
<tr>
<th>Method</th>
<th>Morph II</th>
<th>Web-FaceAge</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>G</td>
<td>L</td>
</tr>
<tr>
<td>Res18+EX w/o L-S</td>
<td>3.549</td>
<td>3.295</td>
</tr>
<tr>
<td>(ours)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ThinAgeNet+FC</td>
<td>3.739</td>
<td>3.295</td>
</tr>
<tr>
<td>ThinAgeNet+L-S</td>
<td>3.739</td>
<td>3.295</td>
</tr>
<tr>
<td>ThinAgeNet w/o L-S</td>
<td>3.739</td>
<td>3.295</td>
</tr>
<tr>
<td>ThinAgeNet w/o FC</td>
<td>3.739</td>
<td>3.295</td>
</tr>
<tr>
<td>ThinAgeNet w/o L-S+FC</td>
<td>3.739</td>
<td>3.295</td>
</tr>
</tbody>
</table>

Figure 5: MAE results on Morph II dataset. ROI means replacing the ARP with the ROI pooling. FC Fus. denotes replacing the recurrent fusion with the FC Fusion in DHAA. w/o L-S indicates training the network without using the losses on sub-networks.

(MAE) and ε-error are adopted for evaluation.

4.3 Ablation Studies

In this subsection, we select the classical Morph II and the proposed Web-FaceAge dataset to validate the effectiveness on four components: global and local features learning, ARP, recurrent feature fusion and the losses on sub-networks.

Global and Local Features Learning: The results of analyzing global, local and global-local features learning are shown in Table 1. The first row means the baseline method, where the plain ResNet-18 model with EXPectation [Rothe et al., 2018] (Res18+EX) is used. For global, local and global-local three branches, they all outperform the baseline method, which demonstrate the effectiveness of the proposed three branches with multiple sub-networks and recurrent fusions. Besides, the global-local branch can achieve lowest MAE than other two branches. It shows that the integrated features learning (both global and local) can extract more effective features than using global or local features learning alone. The joint learning of global and local branches can achieve better performance than the global-local branch. When all the three kinds of features learning are employed, the performance can be further improved. It demonstrates the effectiveness of the joint learning of global context, local details, and the integrated features.

Aligned Region Pooling (ARP): We also evaluate the performance by replacing ARP with ROI pooling. When ROI pooling is used, the nose tip is used as the center of generating the global regions while we use the center of the left eye center, the middle between the nose tip and the middle of two eyes, the center of mouth as the centers to generate three local regions. From Fig. 5, we can find that when ROI pooling is used as the substitute of ARP, the performance drops on both Morph II and Web-FaceAge datasets, which shows ARP is a better choice than ROI to extract more effective features.

Recent Feature Fusion: To investigate the effectiveness of the recurrent fusion, we also show the results of FC fusion (fusing features using a fully connected layer in each branch) which replaces the recurrent fusion. As shown in Fig. 5, the performance drops evidently on both Morph II and Web-FaceAge datasets, which shows that the recurrent fusion is more effective than FC fusion in capturing the spatial dependency and correlation in different regions.

Losses on Sub-networks: As in Fig. 5, the performance drops a lot when removing the losses on sub-networks. This may be caused by the following: 1) when learning the features of multiple regions, the losses help each network to extract independent features and the whole network to extract the diverse and discriminative features; 2) when the training data is insufficient (especially on partial 80-20 protocol), the losses on sub-networks can help to avoid the over-fitting.

4.4 Comparisons with Prior Arts

Results on Web-FaceAge: We compare the proposed D-HAA with the baseline Res18+EX and the re-implemented AgeED [Tan et al., 2018] with parameters $n = 9$ and $p = 2$. As shown in Table 2, our DHAA achieves the best performance with an MAE of 3.549. Compared with the baseline Res18+EX, the proposed method reduces the MAE by 0.190, which is a promising improvement.

Results on Morph II: In Table 3, we compare the proposed DHAA with the baseline Res18+EX, ThinAgeNet [Gao et al., 2018], DRFs [Shen et al., 2018], M-V Loss [Pan et al., 2018], SSR-Net [Yang et al., 2018], AgeED [Tan et al., 2018], DEX [Rothe et al., 2018] and so on. The proposed method outperforms the previous state-of-the-art methods on all three protocols, with achieving the MAE of 1.908, 2.448, 2.489 on 80-20, partial 80-20 and S1-S2-S3 protocols, respectively. Note that we do not employ any additional face dataset for pretraining. Even so, our D-HAA can still outperform those methods with using additional face dataset for pretraining. To our best knowledge, it is the first work which reduces the MAE to under 2 years on 80-20 protocol without finetuning on the additional face dataset.

Table 2: The MAE comparisons on Web-FaceAge dataset.

<table>
<thead>
<tr>
<th>Method</th>
<th>80-20</th>
<th>Partial 80-20</th>
<th>S1-S2-S3</th>
</tr>
</thead>
<tbody>
<tr>
<td>AgeED [Tan et al., 2018]</td>
<td>3.928</td>
<td>3.739</td>
<td>3.549</td>
</tr>
<tr>
<td>Res18+EX(ours)</td>
<td>3.739</td>
<td>3.549</td>
<td></td>
</tr>
<tr>
<td>DHAA(ours)</td>
<td>3.739</td>
<td>3.549</td>
<td></td>
</tr>
</tbody>
</table>

Table 3: The MAE comparisons on Morph II dataset.

<table>
<thead>
<tr>
<th>Method</th>
<th>80-20</th>
<th>Partial 80-20</th>
<th>S1-S2-S3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Soft softmax [Tan et al., 2016]</td>
<td>–</td>
<td>–</td>
<td>3.14/3.03*</td>
</tr>
<tr>
<td>OR-CNN [Niu et al., 2016]</td>
<td>3.34</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>CasCNN [Wan et al., 2018]</td>
<td>3.30</td>
<td>2.93</td>
<td></td>
</tr>
<tr>
<td>DEX [Rothe et al., 2018]</td>
<td>3.00</td>
<td>3.25/2.68*</td>
<td>–</td>
</tr>
<tr>
<td>ARN [Agustsson et al., 2017]</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>AgeED [Tan et al., 2018]</td>
<td>2.93/2.52*</td>
<td>2.86/2.70*</td>
<td></td>
</tr>
<tr>
<td>SSR-Net [Yang et al., 2018]</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>M-V Loss [Pan et al., 2018]</td>
<td>2.41/2.16*</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>DRFs [Shen et al., 2018]</td>
<td>2.17</td>
<td>2.91</td>
<td></td>
</tr>
<tr>
<td>ThinAgeNet [Gao et al., 2018]</td>
<td>1.969*</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Res18+EX(ours)</td>
<td>2.331</td>
<td>2.718</td>
<td>2.695</td>
</tr>
<tr>
<td>DHAA(ours)</td>
<td>1.908</td>
<td>2.487</td>
<td>2.489</td>
</tr>
</tbody>
</table>

pretraining: * IMDB-WIKI, † MS-Celeb-1M [Gao et al., 2016].
The IMDB-WIKI dataset is used for pretraining. As shown in Table 4, the top three methods in the competition. In our experiments, our DHAA achieves the lowest MAE and $\epsilon$-error results on validation and test sets. On validation set, our DHAA reduces the MAE to 2.595 and 3.721 respectively) than Res18+EX. Based on our DHAA, the extracted features with the same age stay closer and thus being more reliable for facial age estimation.

### 5 Conclusion

We have presented a unified network for jointly learning the global context, local details and integrated features, called Deep Hybrid-Aligned Architecture (DHAA), for facial age estimation. Moreover, the DHAA integrates novel components, such as Aligned Region Pooling (ARP), recurrent fusion, and employing separate losses on sub-networks to capture reliable features. A new age dataset has been assembled with a large number of samples and instances, diverse scenes, and a large age range, which is also useful for aging analysis. Extensive experiments have shown that the proposed method achieves better results than the state-of-the-art on multiple aging datasets.

### Acknowledgment

This work was supported by the National Key Research and Development Plan (Grant No.2016YFC0801002), the Chinese National Natural Science Foundation Projects.
References


