Neurocomputing 334 (2019) 114–124

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Muti-stage learning for gender and age prediction

Jie Fang[a,b], Yuan Yuan[c], Xiaoqiang Lu[a,*], Yachuang Feng[a]

[a] Key Laboratory of Spectral Imaging Technology CAS, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119 Shaanxi, China
[b] University of Chinese Academy of Sciences, Beijing 100049, China
[c] Center for Optical Imagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xian 710072, China

ARTICLE INFO

Article history:
Received 6 September 2018
Revised 18 November 2018
Accepted 31 December 2018
Available online 11 January 2019

Communicated by Prof. Junwei Han

Keywords:
Gender and age prediction
Muti-stage learning
Segmentation network

ABSTRACT

Automatic gender and age prediction has become relevant to an increasing amount of applications, particularly under the rise of social platforms and social media. However, the performances of existing methods on real-world images are still not satisfactory as we expected, especially when compared to that of face recognition. The reason is that, facial images for gender and age prediction have inherent small inter-class and big intra-class differences, i.e., two images with different skin colors and same age category label have big intra-class difference. However, most existing methods have not constructed discriminative representations for digging out these inherent characteristics very well. In this paper, a method based on muti-stage learning is proposed: The first stage is marking the object regions with an encoder-decoder based segmentation network. Specifically, the segmentation network can classify each pixel into two classes, "people" and others, and only the "people" regions are used for the subsequent processing. The second stage is precisely predicting the gender and age information with the proposed prediction network, which encodes global information, local region information and the interactions among different local regions into the final representation, and then finalizes the prediction. Additionally, we evaluate our method on three public and challenging datasets, and the experimental results verify the effectiveness of our proposed method.

1. Introduction

Recently, gender and age have played fundamental roles in social interactions [1,2], particularly since the rapid development of social platforms and social media. Languages reserve different grammar rules and salutations for male and female, in addition, different vocabularies are often used when describing elders compared to young people. Despite the basic roles these attributes play in our daily lives, the ability to automatically predict (see Fig. 1) them accurately and reliably from face images is still far from meeting the needs of the increasing applications [3,4], such as in social interaction, electronic commerce, laws and medical treatment. Though people can easily distinguish faces in different age ranges, it is still challenging for a computer or machine to complete that [5,6]. There are many factors that contribute to the difficulty of automatic gender and age prediction. Gender and age prediction is sensitive to some intrinsic factors, such as identity, ethnicity and so on, as well as extrinsic factors, for example, pose, illumination and expression. These factors lead to the small inter-class and big intra-class differences for facial images, i.e., two images with different gender labels have small inter-class difference if they are photographed on the same scene. As a result, the small inter-class and big intra-class differences increase the difficulty to model the gender and age patterns.

As one of the most powerful tools of machine learning, convolutional neural networks (CNNs) [7–9] have developed rapidly in recent years. Lots of complex problems such as human pose estimation [10], face parsing, have been solved successfully by this architecture because of its strong capacity for visual feature extraction. Recent approaches to classifying or estimating gender and age from images have relied on differences in facial features or tailored face descriptors. Most have employed CNN schemes [6,11] designed particularly for the specific task. Levi and Hassncer [6] consider gender and age prediction as two classification tasks, and they utilize CNN in an end-to-end fashion to directly predict the gender and age information. Similarly, Rothe et al. [11] also use CNN to depict the feature of the image, in order to obtain the precise representation, they use the off-of-the-shelf face detectors to obtain the location and size of the face in data preprocessing procedure. Even though these CNN-based methods have achieved significant performances, they still have some common limitations: *The first*

* Corresponding author.
E-mail address: luxq666666@gmail.com (X. Lu).

https://doi.org/10.1016/j.neucom.2018.12.073
0925-2312/

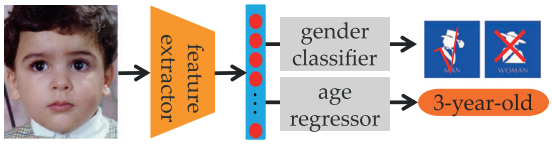


Fig. 1. The gender and age prediction system. Including three components, feature extractor, gender classifier and age regressor.

limitation is that, even though CNNs can depict the image very well, it is difficult to train a traditional big scale CNN sufficiently with little dataset [12]. However, a big scale of images with gender and age labels is difficult to obtain because of some private causes. In the gender and age prediction problem, the quantities of training data are usually very limited and the face image of same people only cover a narrow range of ages. Such insufficient data is hard to exploiting the most general features of the age so that the models tend to suffer from overfitting. *The second limitation* is that, the existing methods accept the whole image as an input of the network, the complex background information interferes the feature extraction process to a large extent, and affects the performance of prediction furthermore. For example, an old man and a boy are photographed in the same scene, the similar backgrounds decrease the intra-distance of two images, so the age prediction will be interfered if the feature extractor accepts the original image as the input. *The third limitation* is that, most existing CNN-based methods only use the output from the last fully connected layer of the network to depict the image, which consider the global information well but ignore the relationships among different local regions. Specific to the facial images, their global information are similar since they have similar objects such as “mouse” and “nose”. In other words, only using global features can not precisely reflect the real differences among images.

To address the aforementioned limitations, a muti-stage learning based method for gender and age prediction task is proposed in this paper. Firstly, an encoder-decoder saliency detection network is proposed to extract the interest area, which is used to avoid the interference of the complex background. Specifically, the saliency detection network is modified from the deconvolution semantic segmentation network, which can classify each pixel into two classes, “people” and others, and we only use the “people” regions to finalize the prediction. Secondly, the VGG19-Net [9] based prediction network is proposed to obtain the final gender and age information of the image. The details of the prediction network are introduced as follows: (1) We replace the fully connected layers with average-pooling layer to depict the global information of the image, this operation decreases the parameters of the network and avoids the overfitting problem to a certain extent. The reason is that, most parameters of CNNs are distributed in the fully connected layers. (2) A combination layer is used to encode the outputs of the last convolutional layer as the local feature of the image, then a local-region-interaction (LRI) layer followed the combination layer is used to enhance the local features. The LRI layer encodes the relationships among different regions in the feature, and it makes the representation more effective. In addition, the global feature and enhanced local feature are fused as the final representation of the image. (3) A classification sub-branch is used to predict the gender information, while the age information is obtained from a regression sub-branch. Additionally, the proposed method achieves state-of-the-art experiment results on three public and challenging datasets: FG-Net [13], Adience [14] and CACD [15]. In summary, the contributions of our work are listed as follows:

1. We propose an encoder-decoder saliency detection network to mark the object regions of the images, only the interest regions are used to predict the gender and age information. This avoids the noises and interferences of the complex background, and further ensures the accuracy of the prediction.
2. We propose a novel prediction network to estimate the gender and age category labels of the age images. The global information, local region information and relationships among different local regions are encoded in the final representation simultaneously, which makes the representation more effective for the gender and age prediction task.
3. We propose a combination layer and local-region-interaction (LRI) layer, which is used to compress the local features and encode the interaction of each local-region-pair, respectively. Additionally, we replace the fully connected layers with average pooling layer to avoid the overfitting problem.
4. We consider the age prediction as a regression task but not the traditional classification one, which enhances the generalization capacity.

The rest of our paper is organized as follows: In Section 2, we introduce the related works to gender and age prediction. Section 3 describes the proposed method. We report the experiments in Section 4 and conclude the paper in Section 5.

2. Related works

This section reviews the existing methods for gender and age prediction task. Additionally, because the proposed method is based on deep neural networks (DNNs), some advantages and disadvantages of DNNs for facial images are introduced as well.

2.1. Gender prediction

Mkinen and Raisamo made a detailed survey of gender prediction methods in [3] to promote the development of this task, and Reid et al. made a more recent survey in [16]. One of the early methods for gender recognition [17] used a neural network trained on a near-frontal face image dataset. The 3D structure of the head was used for gender recognition in [18]. Support vector machine (SVM) classifier was used to image intensities in [19]. Similar to [19], Baluja and Rowley used adaboost for gender prediction in [20], which is applied to image intensities. Finally, viewpoint-invariant age and gender recognition was proposed in [21]. More recently, Ullah et al. used the webers local texture descriptor [22] for gender prediction [23]. Most of these methods discussed above used FERET benchmark [24] both to develop the proposed systems and to evaluate their performances. Images in the FERET dataset were taken under highly controlled condition. Therefore, they are much less challenging than in-the-wild face images. Moreover, the results obtained on this benchmark demonstrate that it is saturated and not challenging for modern competitive methods. In this case, it is difficult to verify the actual relative benefit of these techniques. As a consequence, Shan proposed a method based on LBP features and adaboost classifier in [25], and the authors experimented on the popular labeled faces in the wild (LFW) [2] benchmark, which is primarily used for face recognition task.

2.2. Age prediction

Besides the gender prediction task, the problem of automatically obtaining age information from facial images has also received increasing attention recently, and many methods have been proposed for it. Fu et al. made a detailed survey of age estimation approaches in [26], and Hu et al. made a more recent one [4]. In general, the methods for age prediction can be divided into two branches, measurement-based methods and representation-based methods.

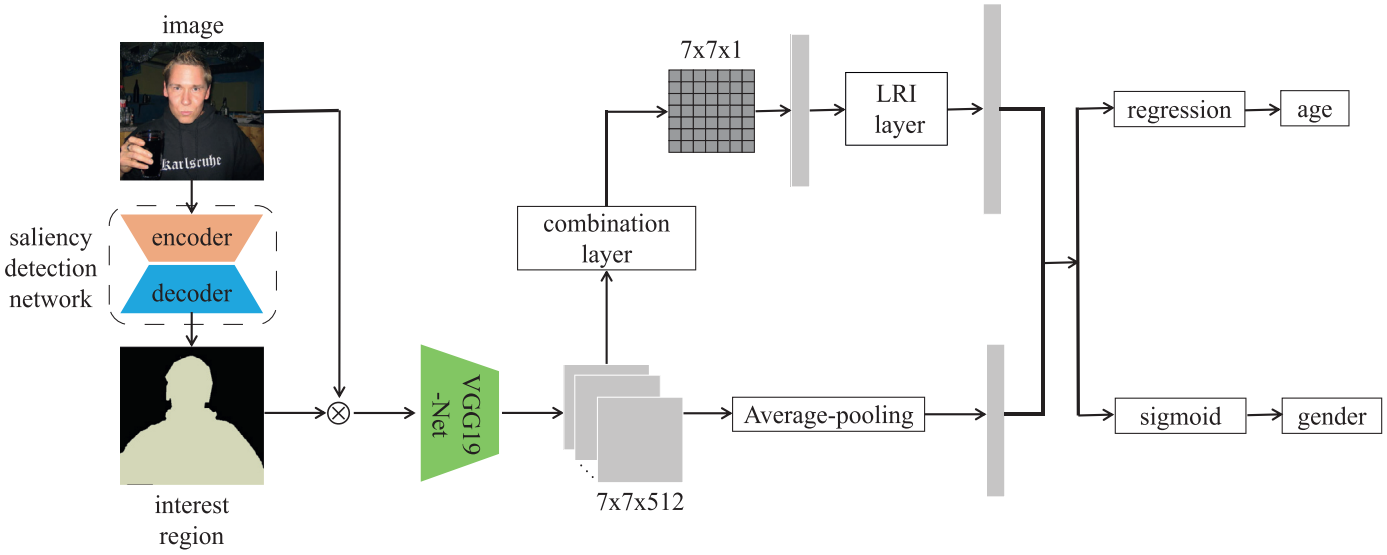


Fig. 2. The overview of the proposed method, which consists two important components, saliency detection network and prediction network. The saliency detection network is used to mark the "person" regions in the image, which aims to avoid the interferences and the noises of the complex background. The prediction network is used to finalize the gender and age prediction procedure, which considers the global information, local information and the relationships among different local regions of the image adequately.

Measurement-based methods. Once facial features are localized and their distances are measured, ratios between them are calculated and used for classifying the face into different age categories. Most early methods for age estimation task are based on calculating the metrics of facial features [27,28]. More recently, Ramanathan and Chellappa [29] proposed a craniofacial growth model, which characterizes growth related shape variations observed in human faces during formative years, to predict the age information in a measurement way. Even though these methods have achieved competitive performances, they are ill-suited for unconstrained images. The reason is that, all of these methods require accurate localization of facial features, which is a challenging problem by itself.

Representation-based methods. Different from the aforementioned methods that use local features for measuring the distances among different images, in [30], Gaussian mixture models (GMMs) [31] were used to represent the distribution of facial patches. Also in [32], GMMs were used for representing the distribution of local facial measurements, but robust descriptors were used instead of pixel patches. Instead of GMM, hidden-markov-model (HMM) and super-vectors [33] were used for representing face patch distributions [34]. More recently, Yi et al. deployed a multi-scale CNN [35] to learn the features of facial images. Wang *et al.* incorporated a manifold learning algorithm in traditional CNN to improve the performance of age prediction [36]. Rothe et al. went deeper with CNNs and SVR [37] for accurate real age estimation on the top of the CNN learned featured. Besides, a novel CNN architecture [6] with shared convolution weights is proposed to predict the gender and age information simultaneously, which has achieved satisfactory performance because the powerful feature extraction capability of deep CNN. Similarly, in [11], the authors used a CNN to predict the age from a single input face image. The difference is that, the off-of-the-shelf face detector is used to obtain the location and size of the face in each input image and the detected face images are rotated to improve the robustness of the method.

2.3. Deep neural network

Currently, with the increment of computational power, the data acquisition ability and the techniques for large scale training of deep neural networks (DNNs) [38], the DNN based methods have achieved impressive results on many classification and recognition tasks. Particularly, the successful use of convolutional neural network on complex computer vision applications, such as human pose estimation [10], face parsing [39], facial keypoint detection [8], object detection [40], object segmentation [41], speech recognition [42] and action classification [43] are notable.

Despite of the advantages of the deep neural networks, there are some disadvantages when they are used for some specific tasks. One of the limitation is that the big scale network models need big storage space to store, which makes them can not apply to mobile devices conveniently. Another limitation is that, it usually needs many annotated data when training deep neural networks because they are equipped with many parameters. However, for some specific tasks, obtaining a big scale of high quality annotated data is not easy, hence the network model cannot be trained sufficiently and the overfitting problem usually leads to the unsatisfactory performances. Therefore, designing dedicated small scale networks for specific tasks is a huge demand.

3. Proposed method

This section describes the multi-stage learning method for gender and age prediction, which contains saliency detection learning stage and prediction learning stage. The overall flowchart is shown in Fig. 2, and the procedures of the proposed method are described as follows:

1. Saliency detection learning stage. An encoder-decoder based segmentation network is proposed to mark the object regions. Specifically, the segmentation network can classify each pixel into two classes, "people" and others, and we only use the "people" regions to finalize the prediction. This stage avoids the noises and interferences of the complex background.
2. Prediction learning stage. A VGG19-Net based prediction network is proposed to predict the precise gender and age information. Specifically, besides the overfitting problem, the prediction network considers the global information, local information and the relationships among different local regions effectively. This stage enhances the representation of the image and makes the prediction results more precise.

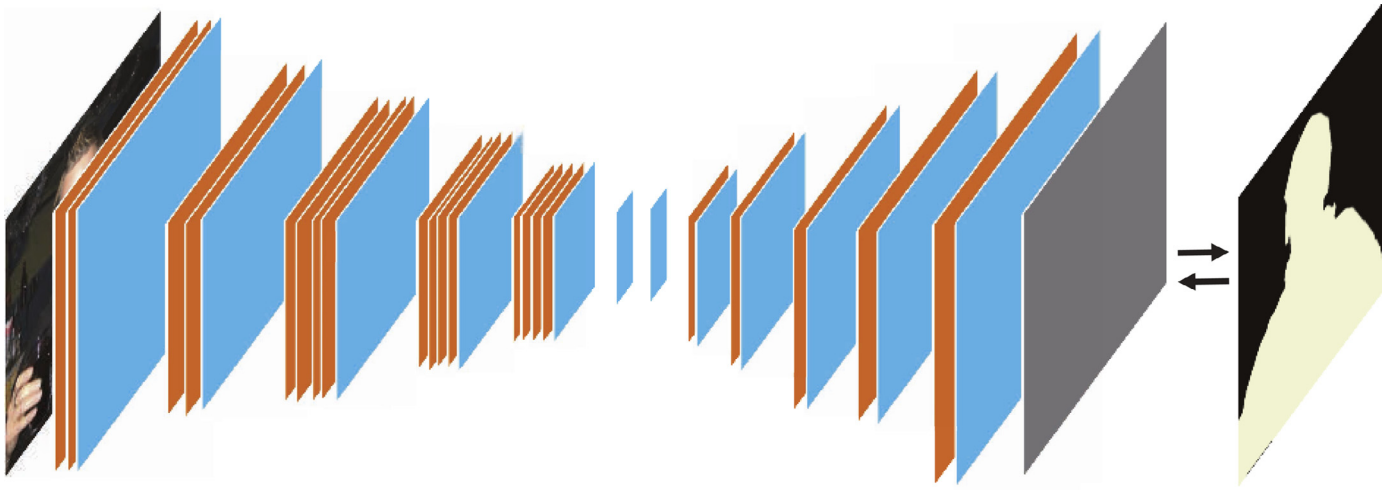


Fig. 3. The structure of the proposed saliency detection network. Including encoder part and decoder part.

Table 1
The inner architecture of saliency detection network.

layer	input	kernel / stride	output
data	/	/	$448 \times 448 \times 3$
VGG16-Net(1 ~ 14 layers)	$448 \times 448 \times 3$	/	$14 \times 14 \times 512$
conv	$14 \times 14 \times 512$	$1 \times 1 \times 512$	$14 \times 14 \times 512$
relu	$14 \times 14 \times 512$	/	$14 \times 14 \times 512$
unpooling	$14 \times 14 \times 512$	2	$28 \times 28 \times 512$
conv	$28 \times 28 \times 512$	$5 \times 5 \times 512$	$28 \times 28 \times 512$
relu	$28 \times 28 \times 512$	/	$28 \times 28 \times 512$
unpooling	$28 \times 28 \times 512$	2	$56 \times 56 \times 512$
conv	$56 \times 56 \times 512$	$5 \times 5 \times 256$	$56 \times 56 \times 256$
relu	$56 \times 56 \times 256$	/	$56 \times 56 \times 256$
unpooling	$56 \times 56 \times 256$	2	$112 \times 112 \times 256$
conv	$112 \times 112 \times 256$	$5 \times 5 \times 128$	$112 \times 112 \times 128$
relu	$112 \times 112 \times 128$	/	$112 \times 112 \times 128$
unpooling	$112 \times 112 \times 128$	2	$224 \times 224 \times 128$
conv	$224 \times 224 \times 128$	$5 \times 5 \times 64$	$224 \times 224 \times 64$
relu	$224 \times 224 \times 64$	/	$224 \times 224 \times 64$
unpooling	$224 \times 224 \times 64$	2	$448 \times 448 \times 64$
conv	$448 \times 448 \times 64$	$5 \times 5 \times 64$	$448 \times 448 \times 64$
softmax	$448 \times 448 \times 64$	/	$448 \times 448 \times 2$

The details of these two subnetworks are detailed in Sections 3.1 and 3.2, respectively.

3.1. Saliency detection network

The first subpart of our proposed method is a deep encoder-decoder based segmentation network, which has achieved success in many other computer vision tasks such as boundary prediction [44], hole filling [45] and semantic segmentation [46]. Compared to the existing unsupervised or weakly supervised learning methods such as [47,48], the proposed encoder-decoder saliency detection network can extract more accurate foreground information of the image. The reasons mainly include two aspects: (1) the strong feature representation capability of CNNs, which can depict the image information very well; (2) based on transfer learning techniques, the proposed saliency detection network is actually a pixel-level classification problem, and the model is trained in a supervised way. Compared to the unsupervised or weakly supervised learning methods, it is of more directivity.

Network architecture: The saliency detection network consists an encoder network and a decoder network, as is shown in Fig. 3. The input to the encoder network is transformed into down-sampled feature maps by subsequent convolutional layers and max-pooling layers. The decoder network in turn uses subsequent unpooling layers which reverse the max-pooling operation and convolutional layers to unsample the feature maps and have the desired output, saliency detection result of people in this paper. Specifically, the proposed encoder network has 14 convolutional layers and 5 max-pooling layers. To simplify the network, the decoder is designed with 6 convolutional layers, 5 unpooling layers followed by a final saliency prediction layer. The structure of the saliency detection network is shown in Fig. 3 and the inner architecture is shown in Table 1.

Training dataset: We consider the saliency detection as a semantic segmentation task which has only two semantic categories, person and others. Unfortunately, there is no existing dataset can train our network directly. In other words, human images with gender, age and pixel-level saliency ground truth information are not easy to obtain. In order to train our saliency detection network, we modified the ground truth images of PASCAL VOC 2012 semantic segmentation benchmark. Firstly, we find out the images with “person” and the corresponding ground truths. Secondly, for a specific ground truth image, we keep the person category labels and assign background to other regions. Some samples is shown in Fig. 4.

An early concern is whether the network trained on the modified PASCAL VOC 2012 dataset can be transformed to the gender and age prediction set well. Actually, PASCAL VOC 2012 dataset contains images with 21 semantic categories, which is more diverse one compared with the facial image dataset. In other words,

Fig. 4. From semantic segmentation to saliency detection.

the model trained on PASCAL VOC 2012 can be generalized to the facial image dataset and marks the "people" regions well.

Loss function: Our network leverages the traditional sigmoid-based cross-entropy loss, which is to predict the affiliation of each pixel, foreground or background. The loss is described as the following equation,

$$L_s = -\frac{1}{n^2}\left[\sum_{i=1}^{n^2} y_s^i \log\left(h_\theta\left(x^i\right)\right) + \left(1 - y_i^s\right)\log\left(1 - h_\theta\left(x^i\right)\right)\right]. \quad (1)$$

Where θ represents the parameters of the network. n is the size of the image. y_s^i is the label of pixel i, it equals to 1 if the pixel belongs to foreground and equals to 0 otherwise. Additionally, $h_\theta(z) = g(\theta^T z)$, and $g(t)$ is the sigmoid function, which is defined as the following equation,

$$g(t) = \frac{1}{1+e^{-t}}. \quad (2)$$

Implementation: To avoid overfitting as well as to leverage the training data more effectively, we use several training strategies. The encoder portion of the network is initialized with the first 14 convolutional layers of VGG16-Net (the 14th layer is fully connected layer "fc6" which can be transformed to a convolutional layer). All the decoder parameters are initialized with Xavier random variables.

3.2. Gender and age prediction network

The gender and age prediction network is based on deep convolutional neural network, which contains two brances: classifier-based gender prediction branch and regressor based age prediction branch. Additionally, except the classifier layer and regressor layer themselves, two branches of the network share the weights. Finally, in order to depict the relationships among different local regions of the facial images precisely, a novel LRI layer is incorporated in the network.

Network architecture: The network is based on the VGG19-Net, but we replace its fully connected layers with our region-relationship based layer, and two extra separated layers are introduced to predict the gender and age information respectively. Specifically, a series of operations are adopted: (1) we use an extra 1×1 convolutional layer to encode the $14 \times 14 \times 512$ feature maps from the last convolutional layer of VGG19-Net as a $14 \times 14 \times 1$ one, then a maxpooling layer is used to encode the feature map to a $7 \times 7 \times 1$ one. These layers actually complete the linear combination of the 512 feature maps, besides enhancing the feature representation, it also decreases the number of parameters of the feature effectively. (2) The output of the proposed convolutional layer is reshaped as a 49×1 vector, each element in the vector represents the feature of corresponding 32×32 region in the input image. Consider gender and age prediction task furthermore, besides the region themselves, the relationships among different regions are very important for estimation. The reason is that, each facial image has almost similar "object", such as "nose", "mouse" and "eye". In this case, not these "objects" themselves but their relationships are sensitive to the gender and age. Additionally, according to the causes referred above, we utilize LRI operation for the 49×1 feature vector to depict the features of each region and the relationships among them effectively, and we obtain a final feature vector sized 2058×1 (because of ignoring the interactions among local regions in the same row compared to the traditional BP, the vector size becomes $7^3 \times (7-1) = 2058$). (3) Following the LRI layer, two separated branches based on classification and regression structure are used to estimate the gender and age information of the facial image. The inner architecture is shown in Table 2.

Table 2
The inner architecture of prediction network.

layer	input	kernel	output
data	/	/	$448 \times 448 \times 3$
VGG19-Net($1 \sim 16$ layers)	$448 \times 448 \times 3$	/	$14 \times 14 \times 512$
conv	$14 \times 14 \times 512$	$1 \times 1 \times 1$	$14 \times 14 \times 1$
maxpooling	$14 \times 14 \times 1$	2	$7 \times 7 \times 1$
reshape	$7 \times 7 \times 1$	/	49×1
LRI	49×1	/	2058×1
softmax / regression	2570×1	/	2 / 1

Local-region-interaction operation: Recently, bilinear pooling models have achieved competitive performances on fine-grained visual classification task [49,50]. The reason is that, BP model considers the 2-order interaction information of different local regions

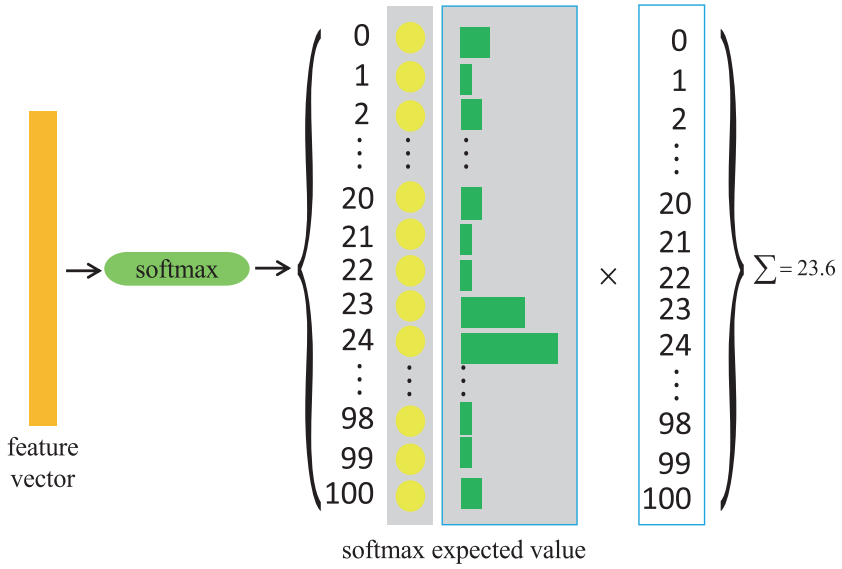


Fig. 5. Mid-mode age prediction. The product of each age category and the corresponding score predicted by the SoftMax classifier is seen as the final age result. Actually, the final predicted result of mid-mode method is a mathematical expectation of the cla-mode method.

in the image, and construct a more robust and discriminative representation. However, the computational complexity of BP is high because of calculating the interaction of each local-region-pair in the image. Particularly, for the gender and age datasets, the extracted interesting regions, e.g., human regions, are generally symmetrical. In this case, based on BP, we propose a local-region-interaction (LRI) operation. Specifically, LRI operation ignores the interactions among local regions in the same row, which enhances the representation of the feature by eliminating the redundant information. The formulation of LRI is defined as the following equation,

$$\mathbf{o}^{N^3(N-1)} = LRI\left(\mathbf{F}^{N\times N}\right) \tag{3}$$

Where **o** is the result of the LRI operation. **F** is the feature map from the combination layer of the prediction network. N is the size of **F**. Additionally, the result **o** is calculated through the following equation,

$$o_k = F_{i,j} \cdot F_{p,q},\ if\ i \neq p \tag{4}$$

Where o_k is the kth element of **o**. $1 \leq i, j, p, q \leq N$ are four indexes and $F_{i,j}$ is the element of **F**. From Eqs. (3) and (4), we can see that, besides enhancing the representation, LRI reduces multiplication operations compared to the traditional BP operation. Additionally, because of the application of combination layer, LRI fuses the feature maps of different channels, compared to BP.

Loss function: The gender and age prediction branches are trained separately, and the loss function of the gender prediction branch is shown as the below equation.

$$L_s = -\frac{1}{N}\left[\sum_{i=1}^{N} y_g^{(i)} \log\left(h_\theta\left(x^{(i)}\right)\right) + \left(1 - y_g^{(i)}\right) \log\left(1 - h_\theta\left(x^{(i)}\right)\right)\right]. \tag{5}$$

where θ represents the parameters of the sigmoid layer. N is the number of images in the dataset. $x^{(i)}$ is the feature vector of the ith facial image. $y_g^{(i)}$ is the gender label, which belongs to {0, 1}, represents male and female respectively. Additionally, the loss function of the age prediction branch is shown as the below equation.

$$\ell_a = \frac{1}{N}\sum_{i=1}^{n} \left(y_a^i - f(x^{(i)})\right)^2 + \lambda \cdot |w|_F^2 \tag{6}$$

where N is the number of images in the dataset. $x^{(i)}$ is the feature vector of the i_{th} facial image. $f(\cdot)$ represents the operation of the regression layer and $f(x^{(i)})$ is the predicted age of the ith facial image. y_a^i is the real age of the i_{th} facial image. w represents the parameters of the regression layer. In testing step, we use $\lfloor p_a + 0.5 \rfloor$ to obtain the integer age label. Where $\lfloor \cdot \rfloor$ is the symbol of the integer operation, and p_a is the result variable from the age prediction network.

Implementation: Even though the prediction network is divided into two branches, they have several common layers. To train the whole network adequately and effectively, we train the gender prediction branch first and then train the age prediction branch. The reason is that, gender prediction is a simpler task compared to age prediction. Additionally, age information is often influenced by the the gender category. Specifically, the age sensitivities of facial images for male and female are different, and usually females' facial characteristics are more sensitive for the age growth. The first 16 layers of the network are initialized with the first 16 convolutional layers of VGG19-Net, and other parameters of the network are initialized with Xavier random variables [51].

Differ from other classification tasks, age is actually a continuous variate, samples from different category are not independent. Using a classification mode directly to predict the age information is coarse. To verify whether the regression mode is more appropriate for the subtask of age prediction, inspired by the method in [52], we design a mode that between classification mode and regression mode. Based on the classification mode method, the product of each category and the corresponding score is seen as the final age predicted result. This mode method considers the details between different age categories. This mode is named as "mid-mode", and its flowchart is shown as Fig. 5.

4. Experiments

This section details the experiments, including datasets, experiment settings, evaluation metrics and experiment results and analysis.

4.1. Datasets

There are many datasets for age and face recognition across age, the proposed method is tested on three public and challenging datasets: FG-Net [13], Adience [14] and CACD [15], these datasets

Fig. 6. Samples of the three datasets used in this paper. The images in the first row are samples from FG-Net. The images in the second row are samples from Adience dataset. The images in the third row are samples from the CACD dataset.

are the most widely used ones. Some samples of the three datasets are shown in Fig. 6.

FG-Net. The Face and Gesture Recognition Research Network (FG-Net) aging dataset contains 1002 color and gray images, which was taken in a totally uncontrolled environment. On average, there are 12 images for each of the 82 subjects, whose age ranges from 0 to 69.

Adience. The Adience dataset was published in 2014, it contains 26,580 images of 2284 subjects, with age and gender labels.

CACD. The Cross-Age Celebrity Dataset (CACD) contains 163,446 images from 20,000 celebrities collected from the internet. The images are collected from search engines using celebrity name and year information. The age is estimated using the query year and the known data of birth.

4.2. Experiment settings

In this paper, we choose 80% images of the dataset to train the network, and use the others to be the testing set. We train the network using mini-batch SGD with patch size 224×224 and batch size 10. The initial learning rate is set to 2.5×10^{-4}, weight decay is set to 5×10^{-4}, momentum is 0.9 and the network is trained for 200 epoches.

4.3. Evaluation metrics

For gender prediction, gender classification accuracy (g.cacc, Eq. (7)) is used to measure the methods. For age prediction, age absolute difference (a.ad, Eq. (8)) is used to measure the error. Besides, even though the age absolute difference (a.ad) can reflect the error between predicted value and label value, it is uniformly distributed and hence does not fit with the human perception sometimes. For instance, the a.rd between predicted value (1) and the corresponding label value (2) is 1, and the a.rd between predicted value (55) and the corresponding label value (56) is also 1. However, from the human perspective, the latter prediction is more accurate because of its bigger base. To address the aforementioned perception gap, the age relative difference (a.rd) is proposed, which is defined as Eq. (9).

$$g.cacc = \frac{n_m^c + n_f^c}{n_m + n_f} \times 100\%, \tag{7}$$

$$a.ad = \frac{1}{n}\sum_{i=1}^{n} \left| y_i^p - y_i^l \right|, \tag{8}$$

$$a.rd = \frac{1}{n}\sum_{i=1}^{n} \frac{\left| y_i^p - y_i^l \right|}{y_i^l} \times 100\%, \tag{9}$$

where n is the total number of the dataset. n_m and n_f represent the number of male and female images respectively. n_m^c and n_f^c are the number of male and female images which gender category are predicted correctly. y_i^p and y_i^l are the predicted and real age label of the ith facial image respectively.

4.4. Contrasting approaches

Here we demonstrate seven contrasting approaches which have achieved satisfactory performance on age estimate task.

Hierarchical age estimation (HAE) [4]: Each facial component is first classified into one of four disjoint age groups using a decision tree based on SVM (SVM-DBT). Within each age group, a separate SVM age regressor is trained to predict the final age.

Deep expectation (DEX) [11]: The off-of-the-shelf face detector is used to obtain the location and size of the face in each input image. Additionally, the authors rotated the detected face images to improve the robust of the method, they used a convolutional neural network to predicted the age of a person starting from a single input face image.

Deep convolutional neural network (DCNN) [6]: Through using the deep convolutional neural network to obtain the representation of the image, and two classifiers are used to predict the gender and age information of the image respectively.

Multi-agent (MA) [53]: Different techniques for acquisition, preprocessing and processing of images are integrated for the gender and age classification task.

Metric regression with CNN (MR-CNN) [54]: MR-CNN considers the age estimation problem as a simple metric regression problem, and addresses this problem with an end-to-end CNN learning algorithm.

Ordinal Regression with CNN (OR-CNN) [54]: OR-CNN considers the age estimation problem as an ordinal regression problem, and addresses this problem with an end-to-end CNN learning algorithm.

Ranking CNN [55]: Ranking-CNN contains a series of basic CNNs, each of which is trained ordinal age labels. Then their binaty outputs are aggregated for the final prediction.

Table 3
The influence of saliency detection (%).

Methods	*g.cacc* (%)	*a.ad*	*a.rd*(%)
None	94.81	4.25	12.14
ScaleFace network [56]	95.06	3.88	10.49
Our segmented network	95.31	3.67	9.66

Table 4
The influence of LRI operation (%).

Methods	*g.cacc* (%)	*a.ad*	*a.rd*(%)
Without LRI product	96.31	3.68	9.51
With LRI product	97.80	3.06	8.74

4.5. Experiment results and analysis

This section details the experiment results on three public and challenging datasets: FG-Net, MORPH and CACD. Before the real experiment, we test the effectiveness of the proposed saliency detection and LRI operations on FG-Net.

The influence of saliency detection. To verify the effectiveness of the saliency detection, both gender and age information are predicted by "pre-trained VGG19-Net+SVM". Additionally, the SVM parameters are trained with feature vectors from the pretrained VGG19-Net and the corresponding gender or age labels of the images. The comparison experiment settings are similar, the only difference is that which saliency detection method is used to extract the interesting regions. Three contrasting methods are without saliency detection operation, with ScaleFace network [56], and with our saliency-based segmentation network respectively. Specifically, the 4096×1 feature vector from the VGG19-Net is used to classify the gender and age information of the corresponding facial image. The experiment result is shown as Table 3.

As is shown in Table 3, the ScaleFace network improves the performances of gender and age prediction, because it avoids the noises and inferences of the complex background. Additionally, the proposed saliency-based segmentation network improves the performance of the prediction, compared to the ScaleFace network. For example, it obtains 0.25% improvement in terms of *g.cacc*. The main reason is that, compared to ScaleFace network, the interesting regions extracted with the proposed saliency-based segmented network not only contain face regions, but also contain the texture information such as hair and clothes, which are vital for gender and age prediction task.

The influence of LRI operation. To verify the effectiveness of the LRI, both gender and age information are predicted by the proposed prediction network, and the saliency detection network is used in this step. The comparison experiment settings are similar, the only difference is that whether the LRI layer is introduced in the network. Specifically, 49×1 or $(49 \times (49 - 7)) \times 1$ feature vector is input to the prediction layer. In other words, the feature vector is input to a sigmoid branch to predict the gender category label, while it is input to a regression branch layer to predict the age information. The experiment results is shown as Table 4.

As can be seen from Table 4, the LRI layer dramatically improves the experiment results. For instance, it gains 0.62 decrement in terms of *a.ad*. The reason is that, through the LRI layer, the relationships among different regions are dug out, it enhances the representation of the image. Additionally, differ from the original bilinear pooling operation, specific to the facial image, the proposed LRI operation ignores the relationships of the regions in the same row. This avoids the interferences of the asymmetric local regions in facial image, meanwhile, it improves the inimitable symbol of different regions and the independence of the representations.

Table 5
The influence of average pooling feature (%).

Methods	*g.cacc* (%)	*a.ad*	*a.rd*(%)
Without average pooling-feature	97.81	3.05	8.71
With average pooling-feature	98.55	2.68	7.44

Table 6
Results on FG-Net (%).

Methods	*g.cacc* (%)	*a.ad*	*a.rd*(%)
HAE [4]	–	4.60	13.14
DEX [11]	95.53	3.09	8.27
DCNN [6]	90.70	5.84	15.63
MA [53]	87.71	6.13	16.41
OR-CNN [54]	–	3.18	8.75
MR-CNN [54]	–	3.11	8.52
Ranking-CNN [55]	–	2.94	8.36
Ours (age cla-mode)	98.80 (S)	2.97	8.49
Ours (age mid-mode)	–	2.84	8.11
Ours (age reg-mode)	–	2.69	7.68

The influence of average-pooling feature. To verify the effectiveness of the average-pooling feature vector, both gender and age information are predicted by the prediction network. The comparison experiment settings are similar, the only difference is that whether the average-pooling layers are use in the prediction process. Additionally, both the saliency detection network and the LRI layer are used in this step. The experiment result is shown in Table 5.

From Table 5, when the average pooling-feature is considered as the final representation of the image, the prediction network obtains more satisfactory performance. For example, it obtains 0.74% decrement in terms of *a.rd* when the average pooling-feature is fused in the final representation of the image. The reason is that, despite of the local information and the relationships among the local relationships, the global information is also very important for the representation. The average pooling-feature can depict the global feature of the image well, and it is an effective supplement of the local features from the LRI layer.

In the following, we will report the specific experiment results on three public and challenging datasets. Because the method in [11] is designed only for age prediction, to verify the superiority of the representation capability of the proposed method, we use the final feature+SVM to predict the gender category label. In order to ensure the fairness of the comparison, we use a separated approach to predict the gender information when the proposed method is tested: firstly, the feature vector is extracted with the proposed network. And then, a SVM classifier is trained to predict the gender category label.

4.5.1. Results on FG-Net

The result on FG-Net is shown as Table 6, where 'S' represents the separated approach. Some samples with wrong predicted gender category label are shown in Fig. 7, from which we can see that, the factors such as hairstyle and face expression may affect the result of gender prediction. Additionally, as the first picture shown in Fig. 7, the gender of child is very difficult to predict even for the humans, this demonstrate the challenge of the gender prediction task again.

From Table 6, we can see that the Ranking-CNN method in [55] achieves 2.90 decrement in terms of *a.ad* compare to the DCNN method [6], because Ranking-CNN thoroughly considers the age-related ordinal information when predict the age label. Additionally, the proposed method achieves better performance than the method in [54]. For example, the (reg-mode) method achieves 0.25 decrement in terms of *a.ad* compared with the method in

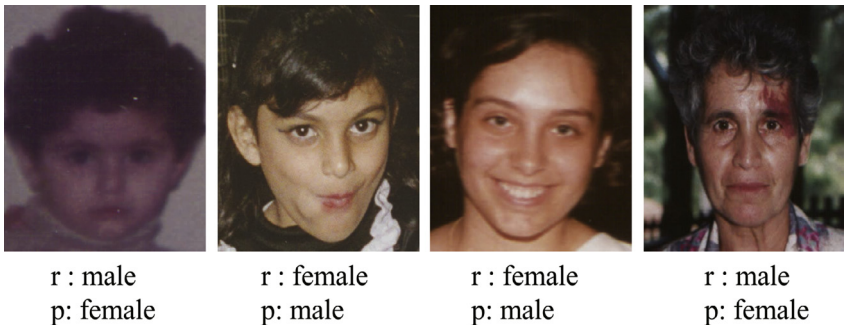


Fig. 7. Samples with wrong predicted gender category in FG-Net. '*r*' represents the real gender category label while '*p*' represents the predicted gender category label.

Table 7
Results on Adience.

Methods	*g.cacc* (%)	*a.ad*	*a.rd* (%)
HAE [4]	–	3.03	8.99
DEX [11]	90.16	2.55	7.56
DCNN [6]	86.85	4.48	13.27
MA [53]	82.43	5.69	17.78
OR-CNN [54]	–	2.67	8.34
MR-CNN [54]	–	2.60	8.12
Ranking-CNN [55]	–	2.24	7.25
Ours (age cla-mode)	93.52	2.23	6.61
Ours (age mid-mode)	–	2.08	6.17
Ours (age reg-mode)	–	1.84	5.02

Table 8
Results on CACD.

Methods	*g.cacc* (%)	*a.ad*	*a.rd* (%)
HAE [4]	–	7.04	18.05
DEX [11]	93.48 (S)	6.56	16.82
DCNN [6]	89.23	9.44	24.20
MA [53]	86.74	10.28	26.35
OR-CNN [54]	–	6.49	16.63
MR-CNN [54]	–	6.61	17.06
Ranking-CNN [55]	–	6.52	16.82
Ours (age cla-mode)	95.01	6.33	16.23
Ours (age mid-mode)	–	5.94	15.45
Ours (age reg-mode)	–	5.38	13.77

[54]. The reasons mainly include two aspects: (1) the saliency detection network extracts the interest regions in the image and only these regions are used to complete the prediction tasks. This avoids the interference information of the complex background. (2) besides the global information of the image, the local region information and their relationships are considered in the final representation through the LRI layer. The local region and detailed information are more important for the facial images, because their global information are similar to each other and can not depict the differences of different images very well. Finally, the (age reg-mode) version of the proposed method achieves the best age prediction result. For example, the (age reg-mode) method achieves 0.43% *a.rd* decrement than (age mid-mode) method. This is because that age is a continuous variable, it will break the inner distribution and dependence of the data if a classification based system is used to deal with the task.

4.5.2. Result on Adience

The result on Adience is shown in Table 7.

From Table 7, we can see that DCNN and DEX methods achieve better performance than that of MA. For example, they achieve 4.42% and 7.73% increments in terms of *g.cacc* than that of MA respectively. This is because that deep convolutional neural networks have powerful capability in visual feature extraction. In addition, all versions of the proposed method achieve more satisfactory than that of OR-CNN and MR-CNN. For instance, the regression mode of the proposed method achieves 0.76 decrement in terms of *a.rd* than that of MR-CNN. The reason is that, the proposed method construct a more discriminative and robust feature. It not only considers the global information of the image, but also considers the local region information and the relationships among different local regions, which enhances the image representation and improves the experiment performances furthermore.

4.5.3. Results on CACD

The result on CACD is shown as Table 8. Some samples with large age predicted error is shown in Fig. 8, from which we can see that skin color is an important effect factor for age prediction. Specifically, the system tends to predict the white younger while predict the black older, this may result from the unbalanced distribution of the training set.

From Table 8, we can see that the proposed method achieves more satisfactory performance. For example, it gains 1.53% improvements in terms of *g.cacc* compared to DEX [11]. This indicates the final feature extracted from our proposed network owns more stronger representation capacity, compared to the feature extracted from the last fully connected layer of the DEX network. Additionally, the regression mode of our proposed method achieves the best performance for age prediction. For instance, it gains 0.28 decrement in terms of a.ad compared to the classification mode of our proposed method. The reasons mainly concludes two aspects: (1) differ from most discrete variables with specific category label, age is a continuous one, methods based on classification architecture may lose detailed information and lead to larger prediction error. (2) when the classification-based method are used to complete the age prediction task, the training process of classifier is very important. For example, if the distribution of the training samples is uneven, the trained classifier will be in an ill state, it is sensitive to some samples but not to others, this is likely to lead to the decrement of the prediction results. However, the regression-based methods can avoids these two limitations to a certain extent. The reason is that, regression-based methods do not have the limits of ill classifier problem result from unbalanced data distribution.

5. Conclusion and future work

In this paper, a multi-stage based learning approach is proposed for gender and age prediction. We use the saliency detection network and transfer learning strategy to find out the local regions we are interested in, which avoids the interferences and noises of the complex background. And then, we use the region relationship-based network to predict the precise gender and age information of the facial image, which considers the global, local region, and the relationships among local region information at the same time.

Fig. 8. Samples with large age predicted error in CACD. '*r_a*' represents the real age label while '*p_a*' represents the predicted age category label.

Generally speaking, according to the strong representation capacity of our extracted features, the proposed method can be generalized to the related works which refer to facial images conveniently.

The proposed method in this paper only considers the gender and age prediction for images which contain and only contain one person, but images with no people or with more than one person may exist in the real application. To address these problems, the preliminary idea is that, incorporate some existing image recognition and object detection methods into our prediction architecture, to improve its flexibility and generalization capability. We will do further researches to cope with these problems in future works.

Additionally, as is aforementioned, facial images have inherent small inter-class and big intra-class differences, and which increases the difficulty of the gender and age prediction. In this case, inspired by the satisfactory performances of existing metric learning methods [57–59] in this condition, we will try utilizing metric learning strategy for constructing more effective feature representation in our future work.

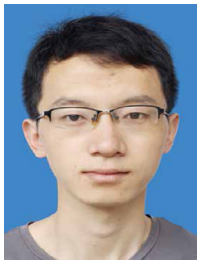
Acknowledgments

This work was supported in part by the National Key R&D Program of China (Grant No. 2017YFB0502900), the National Natural Science Foundation of China (Grant Nos. 61632081, 61772510, and 61702498), the Young Top-Notch Talent Program of Chinese Academy of Sciences (Grant No. QYZDB-SSW-JSC015), and the CAS Light of West China Program (Grant No. XAB2017B15).

References

[1] K.Y. Chang, C.S. Chen, Y.P. Hung, Ordinal hyperplanes ranker with cost sensitivities for age estimation, in: Proceedings of the Computer Vision and Pattern Recognition, 2011, pp. 585–592.
[2] G.B. Huang, M. Mattar, T. Berg, E. Learned-Miller, Labeled faces in the wild: a database for studying face recognition in unconstrained environments, Month (2007).
[3] E. Mkinen, R. Raisamo, Evaluation of gender classification methods with automatically detected and aligned faces, IEEE Trans. Pattern Anal. Mach. Intell. 30 (3) (2008) 541.
[4] H. Hu, C. Otto, A.K. Jain, Age estimation from face images: human vs. machine performance, in: Proceedings of the International Conference on Biometrics, 2013, pp. 1–8.
[5] W.L. Chao, J.Z. Liu, J.J. Ding, Facial age estimation based on label-sensitive learning and age-oriented regression, Pattern Recognit. 46 (3) (2013) 628–641.
[6] G. Levi, T. Hassncer, Age and gender classification using convolutional neural networks, in: Proceedings of the Computer Vision and Pattern Recognition Workshops, 2015, pp. 34–42.
[7] Y. Lecun, B. Boser, J.S. Denker, D. Henderson, R.E. Howard, W. Hubbard, L.D. Jackel, Backpropagation applied to handwritten zip code recognition, Neural Comput. 1 (4) (2014) 541–551.
[8] Y. Sun, X. Wang, X. Tang, Deep convolutional network cascade for facial point detection, in: Proceedings of the Computer Vision and Pattern Recognition, 2013, pp. 3476–3483.
[9] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, Comput. Sci. (2014).
[10] A. Toshev, C. Szegedy, DeepPose: human pose estimation via deep neural networks, in: Proceedings of the Computer Vision and Pattern Recognition, 2014, pp. 1653–1660.
[11] R. Rothe, R. Timofte, L.V. Gool, Deep expectation of real and apparent age from a single image without facial landmarks, Int. J. Comput. Vis. (2016) 1–14.
[12] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Commun. ACM 60 (2) (2017) 2012.
[13] G. Panis, A. Lanitis, N. Tsapatsoulis, T.F. Cootes, Overview of research on facial ageing using the FG-NET ageing database, IET Biom. 5 (2) (2016) 37–46.
[14] E. Eidinger, R. Enbar, T. Hassner, Age and gender estimation of unfiltered faces, IEEE Trans. Inf. Forensics Secur. 9 (12) (2014) 2170–2179.
[15] B.C. Chen, C.S. Chen, W.H. Hsu, Face recognition and retrieval using cross-age reference coding with cross-age celebrity dataset, IEEE Trans. Multimed. 17 (6) (2015) 804–815.
[16] D.A. Reid, S. Samangooei, C. Chen, M.S. Nixon, A. Ross, Soft Biometrics for Surveillance: An Overview, Elsevier Science and Technology, 2013.
[17] B.A. Golomb, D.T. Lawrence, T.J. Sejnowski, Sexnet: A neural network identifies sex from human faces, in: Proceedings of the Advances in Neural Information Processing Systems, 1991, pp. 572–579.
[18] A.J. O'Toole, T. Vetter, N.F. Troje, H.H. Blthoff, Sex classification is better with three-dimensional head structure than with image intensity information, Perception 26 (1) (1997) 75–84.
[19] B. Moghaddam, M.H. Yang, Learning gender with support faces, IEEE Trans. Pattern Anal. Mach. Intell. 24 (5) (2002) 707–711.
[20] S. Baluja, H.A. Rowley, Boosting sex identification performance, Int. J. Comput. Vis. 71 (1) (2007) 111–119.
[21] M. Toews, T. Arbel, Detection, localization, and sex classification of faces from arbitrary viewpoints and under occlusion, IEEE Trans. Pattern Anal. Mach. Intell. 31 (9) (2009) 1567–1581.
[22] J. Chen, S. Shan, C. He, G. Zhao, M. Pietikainen, X. Chen, W. Gao, Wld: A robust local image descriptor, IEEE Trans. Pattern Anal. Mach. Intell. 32 (9) (2010) 1705–1720.
[23] I. Ullah, M. Hussain, G. Muhammad, H. Aboalsamh, Gender recognition from face images with local WLD descriptor, in: Proceedings of the International Conference on Systems, Signals and Image Processing, 2012, pp. 417–420.
[24] P.J. Phillips, H. Wechsler, J. Huang, P.J. Rauss, The feret database and evaluation procedure for face-recognition algorithms, Image Vis. Comput. 16 (5) (1998) 295–306.
[25] C. Shan, Learning local binary patterns for gender classification on real-world face images, Pattern Recognit. Lett. 33 (4) (2012) 431–437.
[26] Y. Fu, G. Guo, T.S. Huang, Age synthesis and estimation via faces: a survey, IEEE Trans. Pattern Anal. Mach. Intell. 32 (11) (2010) 1955.
[27] X. Geng, Z.H. Zhou, K. Smithmiles, Automatic age estimation based on facial aging patterns., IEEE Trans. Pattern Anal. Mach. Intell. 29 (12) (2007) 2234.
[28] Y.H. Kwon, N.D. Vitoria Lobo, Age classification from facial images, Comput. Vis. Image Underst. 74 (1) (1999) 1–21.
[29] N. Ramanathan, R. Chellappa, Modeling age progression in young faces, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2006, pp. 387–394.
[30] S. Yan, X. Zhou, M. Liu, M. Hasegawa-Johnson, T.S. Huang, Regression from patch-kernel (2008) 1–8.
[31] A.R. Webb, K.D. Copsey, Introduction to Statistical Pattern Recognition, John Wiley and Sons, Ltd, 1990.
[32] S. Yan, M. Liu, T.S. Huang, Extracting age information from local spatially flexible patches, in: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2008, pp. 737–740.
[33] L.R. Rabiner, B.H. Juang, An introduction to hidden Markov models, Curr. Protoc. Bioinform. 3 (2007). Appendix 3.
[34] X. Zhuang, X. Zhou, M. Hasegawa-Johnson, T. Huang, Face age estimation using patch-based hidden Markov model supervectors, in: Proceedings of the International Conference on Pattern Recognition, 2009, pp. 1–4.
[35] D. Yi, Z. Lei, S.Z. Li, Age estimation by multi-scale convolutional network, in: Proceedings of the Asian Conference on Computer Vision, 2014, pp. 144–158.

[36] X. Wang, R. Guo, C. Kambhamettu, Deeply-learned feature for age estimation, in: Proceedings of the Applications of Computer Vision, 2015, pp. 534–541.
[37] R. Rothe, R. Timofte, L.V. Gool, Some like it hot ł visual guidance for preference prediction, in: Proceedings of the Computer Vision and Pattern Recognition, 2016, pp. 5553–5561.
[38] X. Peng, J. Feng, S. Xiao, W. Yau, J.T. Zhou, S. Yang, Structured autoencoders for subspace clustering, IEEE Trans. Image Process. 27 (10) (2018) 5076–5086, doi:10.1109/TIP.2018.2848470.
[39] P. Luo, X. Wang, X. Tang, Hierarchical face parsing via deep learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 2480–2487.
[40] J. Han, D. Zhang, G. Cheng, N. Liu, D. Xu, Advanced deep-learning techniques for salient and category-specific object detection: a survey, IEEE Signal Process. Mag. 35 (1) (2018a) 84–100, doi:10.1109/MSP.2017.2749125.
[41] J. Han, R. Quan, D. Zhang, F. Nie, Robust object co-segmentation using background prior, IEEE Trans. Image Process. 27 (4) (2018b) 1639–1651, doi:10.1109/TIP.2017.2781424.
[42] A. Graves, A.R. Mohamed, G. Hinton, Speech recognition with deep recurrent neural networks, in: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2013, pp. 6645–6649.
[43] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, F.F. Li, Large-scale video classification with convolutional neural networks, in: Proceedings of the Computer Vision and Pattern Recognition, 2014, pp. 1725–1732.
[44] J. Yang, B. Price, S. Cohen, H. Lee, M.H. Yang, Object contour detection with a fully convolutional encoder-decoder network (2016) 193–202.
[45] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, A.A. Efros, Context Encoders: Feature Learning by Inpainting (2016) 2536–2544.
[46] V. Badrinarayanan, A. Kendall, R. Cipolla, Segnet: A deep convolutional encoder-decoder architecture for scene segmentation, IEEE Tran. Pattern Anal. Mach. Intell. 39 (12) (2017) 2481–2495.
[47] J. Han, K.N. Ngan, M. Li, H. Zhang, Unsupervised extraction of visual attention objects in color images, IEEE Trans. Circuits Syst. Video Techn. 16 (1) (2006) 141–145, doi:10.1109/TCSVT.2005.859028.
[48] J. Han, D. Zhang, G. Cheng, L. Guo, J. Ren, Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning, IEEE Trans. Geosci. Remote Sens. 53 (6) (2015) 3325–3337, doi:10.1109/TGRS.2014.2374218.
[49] T. Lin, A. Roy Chowdhury, S. Maji, Bilinear CNN models for fine-grained visual recognition, in: Proceedings of the IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7–13, 2015, 2015, pp. 1449–1457, doi:10.1109/ICCV.2015.170.
[50] Y. Gao, O. Beijbom, N. Zhang, T. Darrell, Compact bilinear pooling, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2016, pp. 317–326.
[51] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, CoRR (2014). http://arxiv.org/abs/1409.1556.
[52] H. Pan, H. Hu, S. Shan, X. Chen, Mean-variance loss for deep age estimation from a face, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.
[53] A. Gonzlez-Briones, G. Villarrubia, J.F.D. Paz, J.M. Corchado, A multi-agent system for the classification of gender and age from images, Comput. Vis. Image Underst. (2018) 98–106.
[54] Z. Niu, M. Zhou, L. Wang, X. Gao, G. Hua, Ordinal regression with multiple output CNN for age estimation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27–30, 2016, 2016, pp. 4920–4928, doi:10.1109/CVPR.2016.532.
[55] S. Chen, C. Zhang, M. Dong, J. Le, M. Rao, Using ranking-CNN for age estimation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21–26, 2017, 2017, pp. 742–751, doi:10.1109/CVPR.2017.86.
[56] S. Yang, Y. Xiong, C.C. Loy, X. Tang, Face detection through scale-friendly deep convolutional networks, CoRR (2017). http://arxiv.org/abs/1706.02863
[57] G. Cheng, P. Zhou, J. Han, Duplex metric learning for image set classification, IEEE Trans. Image Process. 27 (1) (2018) 281–292, doi:10.1109/TIP.2017.2760512.
[58] J. Han, G. Cheng, Z. Li, D. Zhang, A unified metric learning-based framework for co-saliency detection, IEEE Trans. Circuits Syst. Video Techn. 28 (10) (2018) 2473–2483, doi:10.1109/TCSVT.2017.2706264.
[59] G. Cheng, C. Yang, X. Yao, L. Guo, J. Han, When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative cnns, IEEE Trans. Geosci. Remote Sens. 56 (5) (2018) 2811–2821, doi:10.1109/TGRS.2017.2783902.



Jie Fang received B. S. degree in school of electronic engineering from XiDian University, Xi'an, Shaanxi, P. R. China in 2015. He is currently pursing the Ph. D degree in signal and information processing techniques with the Key Laboratory of Spectral Imaging Technology CAS, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, China and with the University of Chinese Academy of Sciences, Beijing 100049, China. His research interests include Artificial Intelligence, Machine Learning and Image Understanding.

Yuan Yuan (M'05-SM'09) is currently a Full Professor with the Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an 710072, China. Her current research interests include Visual Information Processing and Image/Video Content Analysis.

Xiaoqiang Lu (M'14-SM'15) is a Full Professor with the Key Laboratory of Spectral Imaging Technology CAS, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, China. His current research interests include Pattern Recognition, Machine Learning, Hyperspectral Image Analysis, Cellular Automata, and Medical Imaging.

Yachuang Feng is currently an assitant researcher with the Key Laboratory of Spectral Imaging Technology CAS, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, China. His current research interests include Image Processing and Video Content Analysis.