CAAI Transactions on Intelligence Technology

Research Article

Three-stage network for age estimation

ISSN 2468-2322
Received on 18th March 2019
Revised on 28th April 2019
Accepted on 30th April 2019
doi: 10.1049/trit.2019.0017
www.ietdl.org

Yu Tingting, Wang Junqian, Wu Lintai, Xu Yong ✉

School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen), Shenzhen, People's Republic of China
✉ *E-mail: laterfall@hit.edu.cn*

Abstract: Age estimation on the basis of the face has been widely used in the field of human–computer interaction and intelligent surveillance. Many existing methods extract deeper global features from the facial image and achieve significant improvement on age estimation. However, local features and their relationship are important for age estimation. In this study, the authors propose a model to use local features for age estimation. The proposed model consists of three stages, preliminary abstraction stage for extracting deeper features, local feature encoding stage to model the relationship between local features and recall stage for the combination of temporary local impressions. Extensive experiments show that their proposed method outperforms previous state-of-the-art methods.

1 Introduction

Facial age estimation is widely concerned and plays an important role in the cases where the age of the users should be estimated. In addition, facial age estimation has a huge application market in social and business [1, 2] such as human–computer interaction and accurate advertising. Generally, age estimation falls into two main categories: the accurate age estimation and the age group estimation. The accurate age estimation predicts the exact age of a person directly via a model. The age group estimation mainly divides the continuous ages into disjoint intervals and predicts which internal the age of a person belongs to. Since the age group estimation is easier than the accurate age estimation, the estimation result of it is better. This paper focuses on the accurate age estimation.

In the last few decades, many types of researches have worked for facial age estimation and most of them focused on feature extraction. The anthropometric model, active appearance model [3] and biometric inspired feature [4] are well known models to extract features for age estimation. Recently, deep learning has become the mainstream feature extraction method and has been widely used in the field of computer vision. It has made breakthroughs in many tasks such as image classification, object detection and image segmentation. Many facial age estimation methods apply convolutional neural network (CNN) to learn deep facial age features from large-scale facial data and achieve significant improvement on age estimation. For example, ordinal regression with CNN (ORCNN) [5] and deep random forest (DRFs) [6] use VGG [7], ResNet [8] and other structures for deep feature extraction and focus on improving the decision-making schemes. However, few works focus on using a local feature which has great significance in age estimation. We observe that age information presented in the face may be affected by some local features such as crow's feet. In addition, the relationship between local features may contribute to the estimation result. For example, the presence of crow's feet shows that a man is not youthful and the white beard on his chin can support this speculation, and thus the prediction result may be more accurate. Therefore, such a relationship may help a fine estimation.

To address these issues, we propose a novel model, which is inspired by human observation process. As shown in Fig. 1, the observation process of human consists of three stages. First, they extract deep features to gain a global impression of images. Second, for fine prediction, they exploit the feature from top to bottom carefully to gain a local impression. Finally, memory recall is conducted to ensure more exact answers. Mimicking this procedure, our model is also composed of three stages. We attain deep feature in the first stage. Next, we slice this feature to get local features and put them into a recurrent neural network (RNN) to get local impressions. Additionally, a previous local impression will be supplied as context information and transferred to the latter encoding process to capture the relationship between local features. Finally, age estimation is conducted based on these local impressions. In addition, we carry out experiments on multiple standard age prediction datasets like the craniofacial longitudinal morphological face database (MORPH) [9], MegaAge [10] and MegaAge-Asian10 and achieve good results, which proves the validity of the proposed model and the feasibility of simulating human vision behaviour on age estimation task.

The main contributions of this paper consist of the following three points: (i) inspired by human visual observation habits, we propose a three-stage model that simulates preliminary abstraction, local feature encoding and recall processing. (ii) By introducing the local feature encoding module, we capture local features and their relationship information to get a fine prediction. (iii) On the public dataset MORPH, MegaAge and MegaAge-Asian, the proposed network structure achieves good results.

The remaining parts of this paper have the following organisation. Section 2 describes the related work of the proposed method. Section 3 presents the proposed method. Section 4 offers a performance analysis of the network. Section 5 provides the conclusion of this paper.

2 Related work

People use the words 'baby face' or 'obsolete' to describe the gap between the real and apparent ages. Specifically, in the real world, real and apparent ages of people are inconsistent. It is not easy to judge the true age from a photograph. To address this problem, some scholars utilise neural networks to design lots of methods. These methods can be roughly divided into three kinds: classification, regression and ranking method.

2.1 Classification methods for age estimation

In this kind of method, different ages are regarded as different categories. There are many examples of this kind of method.

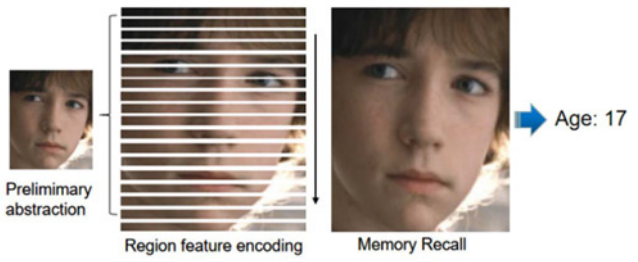

Fig. 1 *Three stages of the human observation process for age estimation*

For example, in 2016, Rothe *et al.* [11] simply used the VGG16 model to extract facial images and calculated the age expectation via the obtained probability from the model output. Their success was driven by the largest scale age estimation dataset IMDBWIKI proposed in their work either. Later, based on [11], Uřičář *et al.* [12] used an ensemble of ten multi-class structured output support vector machine (SO-SVM) classifiers as the final classifier rather than just a single fully connected (FC) layer to get the more general result and achieved a relative improvement. Noting that these approaches utilised age expectation as a final result while ignoring the age distribution. Pan *et al.* [13] proposed a mean-variance loss to drive estimated distribution close to real distribution and achieved state of the art in several datasets.

2.2 Regression methods for age estimation

Age growth is an orderly and constantly changing process, so age estimation can also be seen as a regression issue. To overcome the heterogeneous and discontinuity of data, Huang *et al.* [14] proposed soft margin mixture of regression which simultaneously finds homogeneous partitions in the joint input–output space using max-margin classification and learns a local regressor for each partition. However, their regression model cannot be jointly trained with other CNN model. Preventing these defect, Shen *et al.* proposed an end-to-end model, DRFs, which connects the split nodes to a FC layer and jointly learns input-dependent data partitions to deal with heterogeneous data in age estimation [8].

2.3 Ranking methods for age estimation

Ordinal information between ages is also very important. Treating age labels as ranking order, Chang *et al.* [15] put forward the innovative algorithm ordinal hyperplanes ranker to capture ordering relation information, while the ordinary regression training process is individual and sub-optimised. Inspired by Chang *et al.* [15], Niu *et al.* [7] transformed ordinal regression as a combination of multiple binary classification models based on CNN which can be trained end to end easily. Similarly, Chen *et al.* [16] exploits a face image through multiple binary category models to make age decision and the number of models depends on the age range; however, it is complex and computational.

3 Proposed method

In this section, we describe the proposed method for age estimation. Fig. 2 shows the architecture of the proposed model, which consists of three parts: preliminary extraction module, local feature encoding module and recall module. Furthermore, we will present the proposed three modules.

3.1 Preliminary abstraction module

At the first stage, we use the CNN network to extract features to mimic the initial visual abstraction of images of humans. Specifically, this module has two main challenges as follows: (i) extracting specific visual features from face images, which could lead to a higher estimation accuracy. (ii) Compressing the image size without loss of important information. To address these two problems, we use a face recognition model, sphere-net [17], to carry out initial visual feature extraction task due to the two following metrics: (i) it is obvious that the features extracted from the face recognition model can better express the original face images. (ii) Features extracted by the face recognition model contains rich information about both identity and age; this is proved by Wang [18]. On the basis of these two reasons, we believe that it is reasonable to utilise the face recognition model as the basic feature extraction procedure.

In addition, for subsequent processing and further compression of the model, we have made the following changes:

(i) To prevent loss of spatial information caused by stretching of two-dimensional features into a vector, we removed a FC layer in the original model.
(ii) Following origin model settings, feature maps are scared from 28×24 to 14×12 in B3. To save more detail information, the stride of the first convolution layer in B3 was changed to 1; the feature map size is still 28×24 without decrease. For the same reason, B4 in the origin model is removed either.
(iii) Finally, using 1×1 convolution layers to compress the feature map of 256 channels into a single channel, which is used as subsequent input in the next stage.

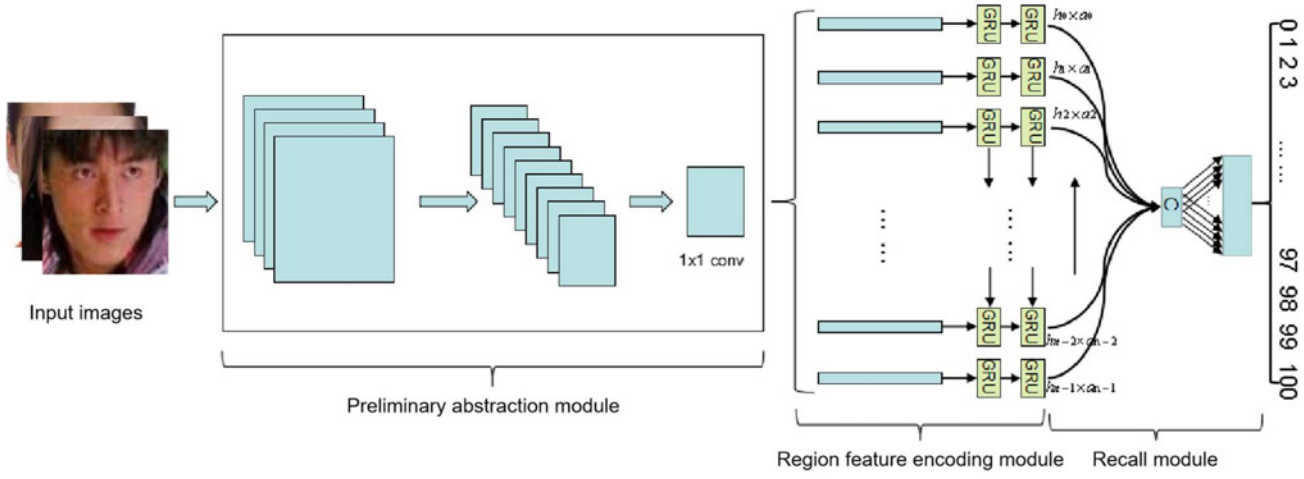

Fig. 2 *Architecture of the proposed method. There are three modules: preliminary abstraction module, region feature encoding module and recall module. Matching to the description in Fig. 1*

Sphere-net and modified module is illustrated in Fig. 3. Red boxes indicate a 3×3 convolution layer with a stride of 2, while black boxes are a 3×3 convolution layer with a stride of 1. The blue box in Fig. 3*b* represents a 1×1 convolution layer with a stride of 1 and the number of output channels is 1.

3.2 Local feature encoding module

A visual feature with the size 28×24 is gained after preliminary abstraction in the first stage, which is a general operation. We note that human have different predictions when they view different facial regions. Similarly, we hope that the model can have different feedbacks with different visual regions, which is essential to making decisions. Thus, we segment the feature to simulate different visual region and put these vectors into gated recurrent unit (GRU)-RNN [19] to generate different impressions in different time steps.

The processing of this module is shown in Fig. 4, the feature defined as X is split into many vectors denoted as $x_0, x_1, x_2, \ldots, x_{n-1}$ and then the vectors are input to GRU unit orderly. Owing to the properties of GRU-RNN, information of adjacent positions can still remain after the artificial division. For example, when x_2 is input into this module directly, x_1 and x_0 are input as context information through the hidden unit to have an impact on the result of x_2.

In different time t, this module has a different visual region. Since the visual region is increased as the encoding proceeds, we can get a series of different local features and the global impression is obtained in the last time step.

3.3 Memory recall module

A global scanning impression vector is attained in the last time step and we can make an age determination based on it without loss of spatial information. However, the information storage capacity of the GRU-RNN [19] is finite. As the observation time increases, the previous impression and local information will be blurred and lost which results in a large deviation between prediction value and ground truth.

A recall process is used to prevent loss of local information and get a more precise result. Accordingly, we calculate the weights of all temporary impressions and final global impression. To model the recall process, a new vector is generated by combining those impressions with weights. The calculation process is as follows:

$$u_t = \boldsymbol{V}^{\mathrm{T}} \tan h(\boldsymbol{W}_1 h_t + \boldsymbol{W}_2 s) \tag{1}$$

$$\alpha_t = \text{soft max}\,(u_t) \tag{2}$$

$$c = \sum_t \alpha_t h_t \tag{3}$$

In the above formulations, S is the state of the hidden unit at the final time; $\boldsymbol{V}^{\mathrm{T}}$, $\boldsymbol{W}_1$, $\boldsymbol{W}_2$ are weight matrices which are learned; α_t is the weight of vector at the time t; and $\boldsymbol{c}$ is the new vector. Classification decision can be directly made based on this vector without loss of context information.

3.4 Training settings

The implementation of the above model is based on the open resource of the deep learning framework, PyTorch, and is performed on a computer with an Intel i7 CPU and two NVIDIA GTX1080Ti. In the training process of the model, we use the Adam optimisation algorithm to perform back propagation calculations. In the preliminary abstraction module, we use a model pre-trained on the VGGFace2 dataset, while the parameters of subsequent modules are randomly initialised. The learning rate is $1{\times}10^{-4}$ at first and gradually decreases at 20th, 30th and 40th epochs. The attenuation rate is 5 with 512 batch sizes and the loss function is cross-entropy.

Complex models and low data volume can lead to overfitting, so we use a variety of image augmentations to reduce the possibility of overfitting including random horizontal flip, random rotation (angle between -10 and 10), randomly change image contrast, saturation and brightness transitions, and all these transformations can be easily achieved by torchvision tools. Besides, in order to avoid the influence of complex background, we use the multi-task convolutional neural network (MTCNN) face detection algorithm proposed by Zhang *et al.* [20] in 2016 to detect the face position and crop a 160×160 face image with margin 10. In addition,

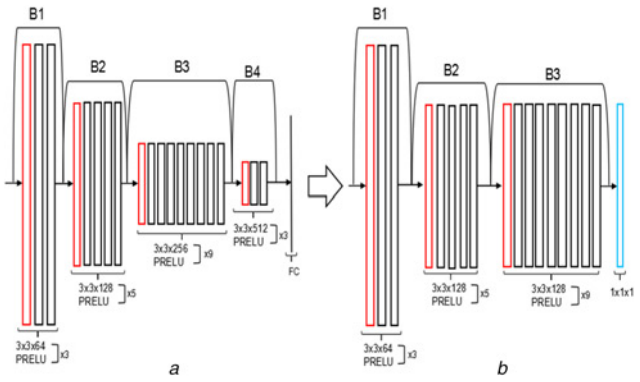

Fig. 3 *Sphere-net and modified module*

a Structure of the sphere-net
b Modified structure of the preliminary abstraction module

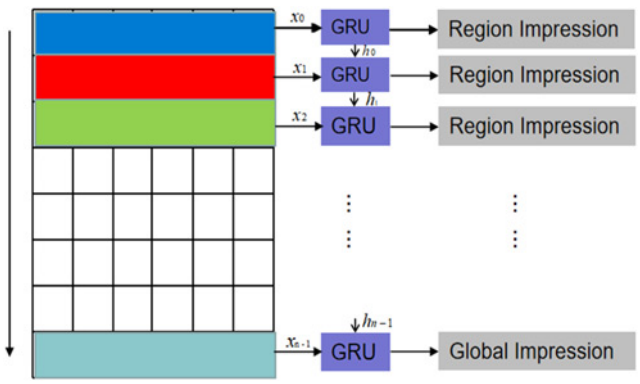

Fig. 4 *Process of the second module in the feature map*

Table 1 Experiment results in MORPH dataset with setting I

Methods	MAE
DEX	3.25
ARN	3.00
DRFs	2.91
model 1(w/o modules 2, 3)	3.30
model 2(w/o module 3)	3.06
full model	**2.85**

The bold values in Tables 1–3 are used to highlight the best results

Table 2 Experiment result in MORPH dataset with setting II

Methods	MAE
OHRank	6.07[a]
DIF	3.00
ranking-CNN	2.96[a]
mean-variance loss	**2.80**
model 1 (w/o modules 2, 3)	3.55
model 2 (w/o module 3)	2.96
full model	2.94

[a]For experiment on protocol a

CAAI Trans. Intell. Technol., 2019, Vol. 4, Iss. 2, pp. 122–126

Table 3 Performance comparison on MegaAge and MegaAge-Asian

Method	MegaAge-Aisan			MegaAge		
	CA3	CA5	CA7	CA3	CA5	CA7
CA	63.19	80.43	90.57	35.17	52.60	66.80
Zhang *et al.* [10] w/o L_{hyper}	60.94	77.57	88.24	35.62	52.52	66.30
Zhang *et al.* [10] w/o L_{KL}	64.08	80.43	90.42	38.69	57.90	**73.15**
Zhang *et al.* [10] full model	64.23	82.15	90.80	41.17	58.37	72.31
model 1 (w/o modules 2 and 3)	63.58	80.30	89.30	39.20	57.20	70.50
model 2 (w/o module 3)	64.05	81.36	90.04	40.02	58.30	72.04
full model (B3 stride = 2)	60.38	76.26	87.63	34.21	51.41	64.78
full model	**64.8**	**83.20**	**91.40**	**42.19**	**60.00**	72.70

following the rules of most face recognition models, we scale the input image to 112×96 finally.

4 Experimental results

In this section, we analyse the performance of the proposed algorithm on two standard public datasets MORPH: MegaAge and MegaAge-Asian. First, we introduce the datasets followed by related assessment criteria. Finally, the experiments are performed to prove the effect of the proposed model.

4.1 Dataset

We conducted experiments on two public datasets: MORPH [9] and MegaAge [10].

The Craniofacial Longitudinal Morphological Face Database (MORPH) is a longitudinal face database. 55,000 images are collected from 13,000 individuals between the ages of 16 and 77.

The MegaAge dataset contains 41,941 images and the MegaAge-Asian contains 40,000 images. They are encompassing ages from 0 to 70, while the former consists of human faces of different regions and the latter are only composed of Asian faces.

4.2 Evaluation criteria

We employ cumulative accuracy (CA) and mean absolute error (MAE) as our evaluation criteria. MAE is defined as: $\mathrm{MAE} = |y' - y|$, where y' is the predicted value and y is the true value. CA is defined as: $\mathrm{CA}(n) = K_n / K \times 100$, where K is a total number of all images in dataset and K_n is the number of test images, whose absolute error with ground truth is lower than n.

4.3 Experiment on MORPH

MORPH dataset is the most popular dataset in the field of age estimation. For evaluation, there exist two mainstream experimental schemes: (i) setting I [8, 11, 21] requires the experiment to be performed on a random subset of all images. This subset is composed of about 5000 images which 80% of them are used for training and others for testing. (ii) Setting II performs five-fold cross-validation on the whole dataset and takes an average of the five experimental results as the final result. Nevertheless, there are two split protocols: (a) five-fold random split [15, 16]; (b) five-fold subject-exclusive protocol [13, 22]. Protocol b means that a subject should only exist in one fold, thus it is more difficult. Our evaluation will perform on setting I and setting II with protocol b.

As shown in Table 1, our proposed model consisting of three modules achieves the lowest MAE of 2.85 in setting I. We compare our model with recent methods including deep expectation (DEX) of apparent age [11], anchored regression network [21] and DRFs [8]. Among them, DEX and DRFs are using VGG16 as the basic feature extraction model. The results show that our method has an obvious improvement.

For setting II, as illustrated in Table 2, the lowest MAE our model gets is 2.94, which is better than some state-of-the-art approaches such as ordinal hyperplane ranking (OHRank) [15] ORCNN [7], DIF [22] and ranking-CNN [16]. However, it is slightly worst than the lowest MAE 2.80 got by mean-variance loss proposed by Pan *et al.* [13].

In addition, we also summarise ablation studies to prove the validity of the last two modules: Model 1 is the model with only the first module and Model 2 is the model with the first and second module. The full model is the model with all three modules. Results in Tables 1 and 2 suggest that the performance of the model without the second and the third module has a great drop and prove that these two modules are essential and important.

4.4 Experiments on MegaAge and MegaAge-Asian

Table 3 reports the performance of the proposed model and a set of state-of-the-art network models [10, 23] for age estimation on MegaAge and MegaAge-Asian. Following the experiment settings of Zhang [10], we also use three criteria CA(3), CA(5) and CA(7) to measure the performance of models in these datasets. The method of Zhang uses VGG, pre-trained on VGGFace dataset, as basic feature extraction model and achieve 60.94 CA(3) on MegaAge-Asian, whereas 35.62 CA(3) on MegaAge. Comparing with their work, our basic feature extraction model has huge advance and improve about 2% on MegaAge-Asian and have 4% improvement on MegaAge, which proves the effectiveness of face recognition model in this task. Furthermore, we can transform the existing pre-trained face recognition model directly without retraining the new model with a large dataset. As declaimed in Section 3, we modify the original face recognition model to get a larger feature map. The output size of B3 in the model with stride 2 is 14×12. To prevent the decreasing of output size, we remove B4 and modify the stride to 1, so that the output size turns to 28×24. We conduct an experiment on different basic models. In Table 3, the result of the full model with stride 2 in B3 is worst than that of the model with stride 1 in B3, which proves that a larger feature map is better for the following processing to achieve higher accuracy.

We also provide ablation experiment on MegaAge and MegaAge-Asian. Among the three mechanisms, our full model performs outstandingly in the three evaluation criteria. Compared our model with state-of-the-art methods cumulative attribute [23] and method of Zhang [10]. It can be directly concluded from Table 3 that our proposed method significantly outperforms than the above methods in different evaluation criteria. Contrasting the result, all these methods achieve the performance on MegaAge-Asian significantly outperforms on MegaAge, which proves that human races have a huge impact on age estimation. Cross-population age estimation is still waiting for addressing.

5 Conclusion

Inspired by the works of predecessors and the habit of human observation, this paper proposes a novel method to simulate three stages of human visual abstraction, local feature re-abstraction and

memory recall processes. In this paper, We also analyse the effectiveness of each module by ablation experiment and demonstrate the validity of the algorithm on standard age estimation datasets. In the future, we will continue the research by trying to use domain adaption to diminish the influence of race.

6 References

[1] Han, H., Otto, C., Jain, A.K.: 'Age estimation from face images: human vs. machine performance'. Int. Conf. Biometrics IEEE, Madrid, Spain, 2013

[2] Guo, G.: 'Video analytics for business intelligence' (Springer, Berlin Heidelberg, Germany, 2012)

[3] Cootes, T.F., Taylor, C.J., Cooper, D.H., *et al*.: 'Active shape models – their training and application', *Comput. Vis. Image Underst.*, 1995, **61**, (1), pp. 38–59

[4] Guo, G.G.G., Mu, G.M.G., Fu, Y.F.Y., *et al*.: 'Human age estimation using bio-inspired features'. IEEE Conf. Computer Vision & Pattern Recognition IEEE, Miami, Florida, USA, 2009

[5] Niu, Z., Zhou, M., Wang, L., *et al*.: 'Ordinal regression with multiple output CNN for age estimation'. 2016 IEEE Conf. Computer Vision and Pattern Recognition (CVPR), Las Vegas, Nevada, USA, 2016, pp. 4920–4928

[6] Shen, W., Guo, Y., Wang, Y., *et al*.: 'Deep regression forests for age estimation'. Proc. IEEE Conf. Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018, pp. 2304–2313

[7] Simonyan, K., Zisserman, A.: 'Very deep convolutional networks for large-scale image recognition', *Comput. Sci.*, ICLR, San Diego, USA, 2014,

[8] He, K., Zhang, X., Ren, S., *et al*.: 'Deep residual learning for image recognition'. Proceedings of the IEEE conference on computer vision and pattern recognition, Las Vegas, NV, USA, 2016, pp. 770–778

[9] Ricanek, K., Tesafaye, T.: 'MORPH: a longitudinal image database of normal adult age-progression'. Seventh Int. Conf. Automatic Face and Gesture Recognition (FGR06), Southampton, UK, 2006, pp. 341–345

[10] Zhang, Y., Liu, L., Li, C., *et al*.: 'Quantifying facial age by posterior of age comparisons', 2017

[11] Rothe, R., Timofte, R., Van Gool, L.: 'Deep expectation of real and apparent age from a single image without facial landmarks', *Int. J. Comput. Vis.*, 2018, **126**, (2–4), pp. 144–157

[12] Uricár, M., Timofte, R., Rothe, R., *et al*.: 'Structured output SVM prediction of apparent age, gender and smile from deep features'. Proc. IEEE Conf. Computer Vision and Pattern Recognition Workshops, Las Vegas, Nevada, USA, 2016, pp. 25–33

[13] Pan, H., Han, H., Shan, S., *et al*.: 'Mean-variance loss for deep age estimation from a face'. Proc. IEEE Conf. Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018, pp. 5285–5294

[14] Huang, D., Han, L., De la Torre, F.: 'Soft-margin mixture of regressions'. Proc. IEEE Conf. Computer Vision and Pattern Recognition, Honolulu, Hawaii, USA, 2017, pp. 6532–6540

[15] Chang, K.Y., Chen, C.S., Hung, Y.P.: 'Ordinal hyperplanes ranker with cost sensitivities for age estimation'. Computer Vision and Pattern Recognition, Colorado Springs, Colorado, USA, 2011, pp. 585–592

[16] Chen, S., Zhang, C., Dong, M., *et al*.: 'Using ranking-CNN for age estimation'. Proc. IEEE Conf. Computer Vision and Pattern Recognition, Honolulu, Hawaii, USA, 2017, pp. 5183–5192

[17] Liu, W., Wen, Y., Yu, Z., *et al*.: 'SphereFace: deep hypersphere embedding for face recognition'. Proc. IEEE Conf. Computer Vision and Pattern Recognition, Honolulu, Hawaii, USA, 2017, pp. 212–220

[18] Wang, Y., Gong, D., Zhou, Z., *et al*.: 'Orthogonal deep features decomposition for age-invariant face recognition'. Proc. European Conf. Computer Vision (ECCV), Munich, Germany, 2018, pp. 738–753

[19] Cho, K., Van Merriënboer, B., Gulcehre, C., *et al*.: 'Learning phrase representations using RNN encoder–decoder for statistical machine translation', *Proceedings of the Empiricial Methods in Natural Language Processing*, Doha, Qatar, 2014, pp. 1724–1734.

[20] Zhang, K., Zhang, Z., Li, Z., *et al*.: 'Joint face detection and alignment using multitask cascaded convolutional networks', *IEEE Signal Process. Lett.*, 2016, **23**, (10), pp. 1499–1503

[21] Agustsson, E., Timofte, R., Van Gool, L.: 'Anchored regression networks applied to age estimation and super resolution'. Proc. IEEE Int. Conf. Computer Vision, Venice, Italy, 2017, pp. 1643–1652

[22] Han, H., Jain, A.K., Wang, F., *et al*.: 'Heterogeneous face attribute estimation: a deep multi-task learning approach', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2018, **40**, (11), pp. 2597–2609

[23] Chen, K., Gong, S., Xiang, T., *et al*.: 'Cumulative attribute space for age and crowd density estimation'. Proc. IEEE Conf. Computer Vision and Pattern Recognition, Portland, Oregon, 2013, pp. 2467–24

CAAI Trans. Intell. Technol., 2019, Vol. 4, Iss. 2, pp. 122–126