A Deep CNN Model For Age Estimation

Beichen Zhang
Tokyo City University
1-28-1 Tamazutsumi
Setagaya-ku
158-8557, Tokyo, Japan
bzhang@tcu.ac.jp

Yue Bao
Tokyo City University
1-28-1 Tamazutsumi
Setagaya-ku
158-8557, Tokyo, Japan
bao@g.tcu.ac.jp

ABSTRACT

Age estimation from human faces is an important yet challenging task in computer vision because of the large differences between physical age and apparent age. Although many inspiring works have focused on the age estimation of a single human face through deep learning, the existing methods still have lower performance when dealing with faces in videos because of the differences in head pose between frames. In this paper, a combined system of age estimation and head pose estimation is proposed to improve the performance of age estimation from faces in videos. We use deep regression forests (DRFs) to estimate the age of facial images, while a multi-loss convolutional neural network is also utilized to estimate the head pose. Accordingly, we estimate the age of faces only for head poses within a set degree threshold to enable value refinement. First, we divided the images in the Cross-Age Celebrity Dataset (CACD) and the Asian Face Age Dataset (AFAD) according to the estimated head pose degrees and generated separate age estimates for images with different poses. The experimental results showed that the accuracy of age estimation from frontal facial images was better than that for faces at different angles. Further experiments were conducted on several videos to estimate the age of the same person with his or her face at different angles, and the results show that our proposed combined system can provide more precise and reliable age estimates than a system without head pose estimation.

Keywords

age estimation; deep learning; CNN; head pose estimation

1 INTRODUCTION

Age estimation from a facial image has become an important yet challenging problem in many applications, such as human-computer interaction[1], identification[2], security[5], and precision advertising[3].

In recent years, deep learning has made impressive works on various computer vision tasks[4], including age estimation[8, 7]. However, all these works have used datasets including only frontal facial images, which cannot adequately reflect the conditions of real-life applications. Different from most facial images in datasets, the head pose may vary greatly in videos or webcam streams, leading to intolerable errors in the estimated age.

In this work, a combined system of age estimation and head pose estimation is proposed to solve the problem of age estimation from faces in videos or webcam streams. First, we use deep regression forests (DRFs) [7] to estimate the age of facial images, which can achieve high precision for frontal facial images. Meanwhile, a multiloss convolutional neural network (CNN) is also utilized to estimate the head pose [9]. Then, we can use the trained system to estimate age and head pose from several videos frame by frame. When using the trained mapping between age and head pose, we set a degree threshold for the head pose and perform age estimation only for frames where the head pose is within this threshold to enable value refinement of the age estimated from the video.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

2 RELATED WORK

For age estimation from faces in videos, the most closely related work is the Deep Age Estimation Model [11], in which Ji et al. use a CNN with an attention mechanism; Facial features are extracted by CNN then aggregated from features vectors to a single feature by an attention block. They trained the model using a new loss function leading to better precision and stability across every frames for age estimation. However, to guarantee stability, this model used continuous frames as input data rather than using a single image. In

addition, to train this model, a new dataset must be collected with labels annotation.

Another work for age estimaion that static and dynamic features can be learned from expressions of face simultaneously in videos called the Spatially-Indexed Attention Model (SIAM) [10]. In this model, Ji Pei et al. employ CNNs to extract the latent appearance features from each frame and then uses recurrent networks to process all the features to simulate time dynamics. However, this method has limitations in terms of which types of facial expression images it can consider; specifically, only smile and disgust databases were used in experiments.

3 PROPOSED METHOD

In this section, each step of the system flow will be explained in detail.

3.1 Age estimation

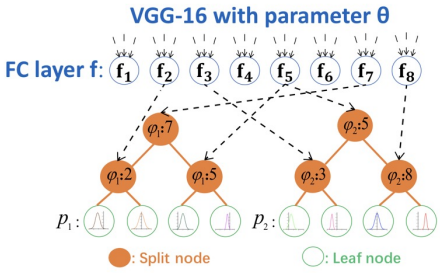

Figure 1: Illustration of a DRF.

Fig. 1 shows a diagram of a DRF [7].

A CNN combined with deep regression forests is introduced in this work and estimate the real age from facial image. The model is trained on facial image datasets with known ages and face landmarks as labels. The training process in this paper begins with the pretrained weights from the ImageNet dataset, as with the same model used in [12]. Then, the CNN is fine-tuned on the two target datasets using for age estimation. The fine-tuning process make the CNN to obtain the features, distribution, and bias of each dataset and optimizes the performance.

The upper blue circles represent the output neurons from the CNN defined by the function $\mathbf{f}$ with parameter Θ. All these neurons come from the last fully-connected (FC) layer of VGG-16. The middle orange circles represent the split nodes and the bottom green circles represent the leaf nodes of deep regression forests. φ_1 and φ_2 represent the index functions of each tree. The black dashed arrows point out the correspondence from the split nodes of each tree to the neurons of VGG-16 FC layer. Each neuron may correspond to the split nodes of different trees. Each tree has its own distribution π for its leaf node(represented by the distribution curves on the leaf nodes). The final output for the whole forest can be calculated as the mix of the predictions of the individual trees. The parameter $f(\cdot;\Theta)$ and π will be trained simultaneously end- to-end.

3.2 Head pose estimation

We adopted Ruiz's method [9], in which deep multiloss CNNs are trained for head pose estimation with satisfactory accuracy. The ResNet50 networks [13] was introduced for headpose estimation and three losses are used for three angles separately. There are two parts of each loss: the mean squared error regressed directly and the cross-entropy loss from classification of pose. There are three FC layers being used for three angles and shared the previous parts of the network. By adopting additional cross-entropy losses from classification, we constructed three signals to be backpropagated to improve the learning process. The predictions of three output angles was computed as the final head pose results. The details of the architecture are shown in Fig. 2.

4 EXPERIMENTS

4.1 Testing on AFAD and CACD

In this section, the performance of DRFs for age estimation based on frontal and nonfrontal facial images is presented. The frequently used AFAD and CACD datasets, representing Asians and Europeans, respectively, were used in this experiment. We used the trained multiloss CNN to estimate the head poses in both datasets. For each facial image, three rotational angles were estimated, one on each axis. We set 30 degrees as the threshold for the sum of the three angles, and images with head pose angle estimates summing to more than 30 degrees were defined as nonfrontal images. Fig. 3 depicts exemplar images of nonfrontal facial images from the datasets.

On AFAD

Based on the estimated angles, AFAD was divided into frontal and nonfrontal subsets consisting of 53,983 and 5,361 images, respectively. Both subsets were randomly split into training/test (85%/15%) sets, and the training process was repeated 5 times with different random separation and the final outcome is the average of 5 times outputs. The quantitative results are summarized in Table 1. The results show that the accuracy of age estimation from frontal facial images is significantly better than that for nonfrontal images.

Subset	MAE
Frontal	3.73
Nonfrontal	4.97

Table 1: Performance (MAE) comparison on the frontal and nonfrontal subsets of AFAD [14]

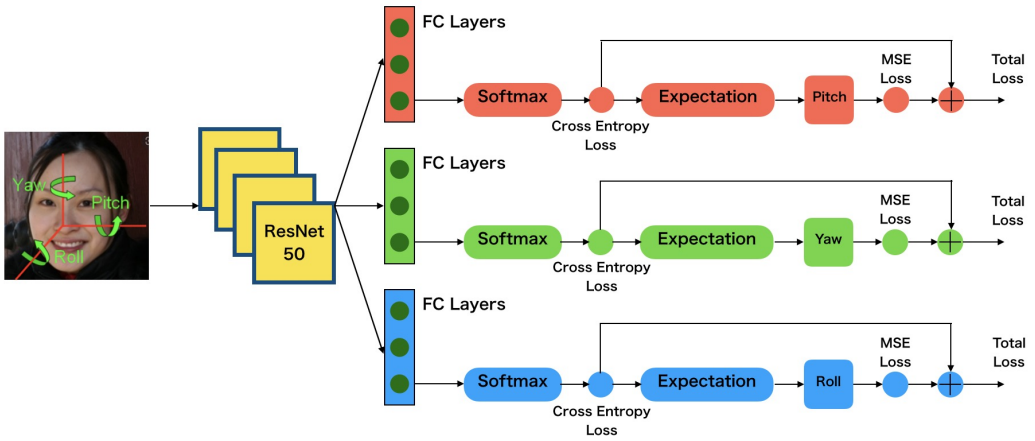

Figure 2: CNN with combined mean squared error and cross-entropy losses.

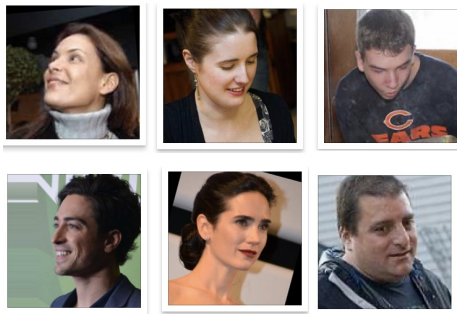

Figure 3: Examples of nonfrontal facial images.

On CACD

Based on the estimated angles, CACD was divided into frontal and nonfrontal subsets consisting of 15,145 and 3,026 images, respectively. Both subsets were randomly split into training/test (85%/15%) sets, and the training process was repeated 5 times with different random separation and the final outcome is the average of 5 times outputs. The quantitative results are summarized in Table 2. The results show that the accuracy of age estimation from frontal facial images is significantly better than that for nonfrontal images.

Subset	MAE
Frontal	4.59
Nonfrontal	5.65

Table 2: Performance (MAE) comparison on the frontal and nonfrontal subsets of CACD [6]

4.2 Testing on facial video datasets

Two new facial video datasets were constructed to evaluate our model in terms of age estimation performance. We collected 18,282 and 18,944 frames from two twelve-minute facial videos of Asian and European subjects, respectively. It should be noted that each facial video dataset was collected from the same person, and these datasets were used only for evaluating the age estimation models; currently, there is no facial video dataset available to be used for training the whole model. We first trained DRFs on AFAD and CACD, representing Asians and Europeans, respectively. Then, we tested the two trained models on the facial video datasets with simultaneous head pose estimation. Examples of the test images are shown in Fig. 4. We performed age estimation only for faces with head poses within 30 degrees, and we compared the results with the results for all images without head pose restrictions. Several other models were also trained on AFAD and CACD and then tested on the facial video datasets for more comprehensive comparisons.

Figure 4: Examples from the facial video datasets with age and head pose estimates. The numbers represent the predicted age. Green and red colors indicate that the sum of the head pose rotational angles is less than and greater than 30 degrees, respectively.

On the Asian facial video dataset We trained a DRF on AFAD and tested the model on the Asian video dataset with head pose restrictions. We compared the results of our method with those of other outstanding age estimation models, and the quantitative results are summarized in Table 3. All models were trained on AFAD with the same training strategy to ensure fair comparisons. On the task of facial video estimation, our method achieves the best MAE of 5.12, and the variance is reduced by 0.62 compared to the best existing method.

Method	MAE	Variance
AlexNet [15]	6.19	6.92
DEX [16]	6.72	8.65
DRF [7]	5.96	4.12
Our method	**5.12**	**3.50**

Table 3: Accuracy (MAE) and variance results for comparison with state-of-the-art methods on the Asian facial video dataset

On the European facial video dataset We trained a DRF on CACD and tested the model on the European video dataset with head pose restrictions. We compared the results of our method with those of other outstanding age estimation models, and the quantitative results are summarized in Table 4. All models were trained on CACD with the same training strategy to ensure fair comparisons. On the task of facial video estimation, our method achieves the best MAE of 5.56, and the variance is reduced by 1.53 compared to the best existing method.

Method	MAE	Variance
AlexNet [15]	6.93	7.15
DEX [16]	7.17	8.22
DRF [7]	6.39	5.84
Our method	**5.56**	**4.31**

Table 4: Accuracy (MAE) and variance results for comparison with state-of-the-art methods on the European facial video dataset

5 CONCLUSIONS

In this paper, a combined system of age estimation and head pose estimation is proposed to solve the problem of age estimation based on faces in videos or webcam streams, where different head poses may lead to intolerable errors on the estimated ages. Experimental results show that with a head pose restriction such that age estimation is performed only for facial images with head poses within a specified degree threshold to ensure value refinement, our method achieves promising improvements in accuracy and stability for age estimation from video.

The main contributions of this paper are as follows: (1) We are the first to couple age estimation and head pose estimation for age estimation in videos. (2) Our method shows significantly improved performance in age estimation on facial video datasets compared to other state-of-the-art methods in terms of both accuracy and variance.

DATA AVAILABILITY

Links to datasets used in this paper.

CACD: `https://bcsiriuschen.github.io/CARC/`

AFAD: `https://afad-dataset.github.io/`

6 REFERENCES

[1] A. Lanitis, C. Draganova, and C. Christodoulou. Comparing different classifiers for automatic age estimation. *IEEE Transaction on Systems, Man, and Cybernetics, Part B: Cybernetics*, 34(1):621-628, 2004.

DOI: 10.1109/TSMCB.2003.817091

[2] A. Lanitis, C. J. Taylor, and T. F. Cootes. Toward automatic simulation of aging effects on face images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(4):442-455, 2002.

DOI: 10.1109/34.993553

[3] C. Shan, F. Porikli, T. Xiang, and S. Gong, editors. Video Analytics for Business Intelligence. *Studies in Computational Intelligence. Springer*, 2012.

[4] M. Stefan Ìczyk, and T. Bochen Ìski. Mixing deep learning with classical vision for object recognition. *Journal of WSCG*, 28(1-2): 147-154, 2020.

DOI: 10.24132/JWSCG.2020.28.18

[5] Z. Song, B. Ni, D. Guo, T. Sim, and S.Yan. Learning universal multi-view age estimator using video context. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 241-248, Nov 2011. 1 DOI: 10.1109/ICCV.2011.6126248

[6] B. Chen, C. Chen, and W. H. Hsu. Face recognition and retrieval using cross-age reference coding with cross-age celebrity dataset. *IEEE Trans. Multimedia*, 17(6):804-815, 2015. DOI: 10.1109/TMM.2015.2420374

[7] W. Shen, Y. Guo, Y. Wang, K. Zhao, B. Wang, and A. Yuille. Deep Regression Forests for Age Estimation. In *IEEE CVPR*, pages 2304-2313, 2018.

DOI: 10.48550/arXiv.1712.07195

[8] S. Chen, C. Zhang, M. Dong, J. Le, and M. Rao. Using ranking-CNN for age estimation. In *IEEE ICCV*, pages 5183-5192, 2017.

DOI: 10.1109/CVPR.2017.86

[9] Ruiz. N, Chong. E, Rehg. J.M. Fine-grained head pose estimation without key-points. In *CVPR workshops*, pp. 2074-2083, 2018.

DOI: 10.48550/arXiv.1710.00925

[10] W. Pei, H. DibeklioÄlu, T. BaltruÅ¡aitis and D. Tax. Attended End-to-End Architecture for Age Estimation From Facial Expression Videos. In *IEEE Transactions on Image Processing*, volume 29, pages 1972-1984, 2019.

DOI: 10.1109/TIP.2019.2948288

[11] Z. Ji, C. Lang, K. Li and J. Xing. Deep Age

Estimation Model Stabilization from Images to Videos. In *International Conference on Pattern Recognition*, 2018.

DOI: 10.1109/ICPR.2018.8545283

[12] Simonyan. K, Zisserman. A. Very deep convolutional networks for large-scale image recognition. CoRR abs/1409.1556, 2014.

DOI: 10.48550/arXiv.1409.1556

[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

DOI: 10.48550/arXiv.1512.03385

[14] Z. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua. Ordinal regression with multiple output cnn for age estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4920-4928, 2016.

DOI: 10.1109/CVPR.2016.532

[15] K. Chang, C. Chen, and Y. Hung. A ranking approach for human age estimation based on face images. In *ICPR*, 2010.

DOI: 10.1109/ICPR.2010.829

[16] K. Chang, C. Chen, and Y. Hung. Ordinal hyperplanes ranker with cost sensitivities for age estimation. In *CVPR*, pages 585-592, 2011.

DOI: 10.1109/CVPR.2011.5995437