PROCEEDINGS OF SCIENCE

Video Age Estimation with Multiple Stacked CNN Models

Zhiqin Zhang[1][2]
Wuhan Donghu University, Wuhan 430000,China
*E-mail:*120991228@qq.com

Automatic age classification has become relevant to an increasing amount of applications, particularly after the occurence of many social platforms and social medias where the video age recognition is important for the improvement of user experience; however, performance of the existing methods on real-world video continuous images is in great shortage, especially when it is compared to the “super-human” improvement of recognition precision reported for the related task of object and face recognition. In this paper, we proposed a new cnn structure by combining several stacked deep-convolutional neural networks (CNN), which consists of an improved alexnet and an improved grouped googlenet. The stacked models can be used to estimate the apparent age of the people from coarse to fine. Experiments showed that a significant improvement in performance can be obtained on the video tasks. We evaluated our method on the recent benchmark for video apparent age estimation and showed it to outperform current state-of-the-art methods.

CENet2017
22-23 July 2017
Shanghai, China

[1] Speaker
[2] This study is supported by the Foundation for Young Scholars in Wuhan Donghu University in 2016(2016dhzk05)

1.Introduction

Age recognition is becoming more and more important in social interactions. Different vocabularies are used when addressing elders compared to young people and the likeness of them are different too. To meet the needs of commercial applications in our day-to-day lives, the ability to automatically estimate their age accurately and reliably from face images is strongly required. This is particularly perplexing when considering recent claims to super-human capabilities in the related task of face recognition. Thereforethe super-improvement accuracy of age estimation is necessary and imminent, especially for video.

In the past decades, approaches to distinguishing these features from face images wererelied on searching in different facial feature spaces [1] or “manual selected” face descriptors [2, 3, 4]). Most have employed classification schemes designed particularly for age classification tasks. Few of these past methods were designed to handle the big challenges of unconstrained video conditions [2]. Moreover, the past deep learning methods employed by these systems did not use the relation of different deep network structures, and the more important thing for improving classification accuracy is to exploit the massive numbers of video image examples available through internet.

In this paper, we attempt to find the reason why automatic face recognition capabilities and those of age classification methods have big gap on accuracy and furthermore to improve it with a stacked CNN structure. In the past several years,the face recognition achieved tremendous improvement by using deep convolutional neural networks (CNN). Following the road of face recognition improvement, we demonstrated similar gains with a mixed stacked CNN network architecture, designed by considering the feature of video age examples and the continuity of video images.

We tested our stacked network on the newly released benchmark for age classification of unconstrained video. We showed that despite of the very challenging nature of the video images, by designing the fit stacked CNN network, the proposed method could outperform existing state of the art method. Although these methods provided the same deep-learning-based structure, a lot of improvements couldn be achieved with more elaborate model structure designs, and we believe that the problem of accurately estimating age in the unconstrained video settings can be solved applicably in the near future.

2.Related Work

Before describing the proposed method we briefly reviewed related methods for age classification and provided an overview of convolutional neural networks.

One of the first applications of convolutional neural networks (CNN) is perhaps the LeNet-5 network for optical character recognition. Compared to modern deep CNN, their network was relatively modest due to limited computational resources of the time and the algorithmic challenges of training bigger networks. In practice, the very large-scale deep convolutional neural network has yielded quite impressive performance in image recognition problem[5] .The proposed combination of the two improved VGG versions models achieved 7.1% top-5 error on ILSVRC-2012-val, and 7.0% top-5 error on ILSVRC-2012-test[6]. The author Kaiming He proposed the ResNet for COCO2015 competitions[7], which won the 1st places in: ImageNet classification, ImageNet detection, ImageNet localization, COCO detection, and COCO segmentation. From above we can see that the proposed CNN models are more complex and the recognition results are more precise.

The traditional methods are subspace mapping which map the aging image into a subspace [8] or a manifold [9], but the recognition precision was influenced by the quality of images which should be near-frontal and well-aligned. As a result, such methods do not suit for unconstrained videos or images. An age and gender classification method was proposed [10], which used convolutional neural networks to classify the age and gender with the Adience benchmark, but the proposed network was single and simple,and the continuous feature of video images and the relation between multiple models were not considered.

In the following sections, we showed a mixed multiple stacked convolutional net architecture that could be used to estimate the apparent age of the people on the videos with high accuracy. Finally, we evaluated our method on the recent benchmark for video apparent age estimation and showed it to outperform current state-of-the-art methods.

3. Stacked CNN Model Structure

In order to achieve the improved performance, a stacked two CNN models structure was proposed which stacked a simple and coarse cnn model with a complex and accurate cnn model. The input image samples were resized into 128×128, and we randomly cropped 112×112 image as the input of our stacked cnn model in training time. The batch size was set to 256. The whole sample set included different gender and color of skin people data for the sake of generality.We selected 15 percent of the whole sample set as the validate set. The figure1 showed the whole sample set year distribution. The stacked structure was showed in Figure 2

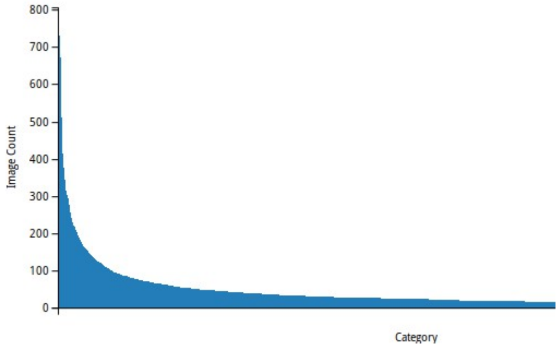


Figure 1: The whole sample set distribution: the category represents the year class index and image count represents every class samples count

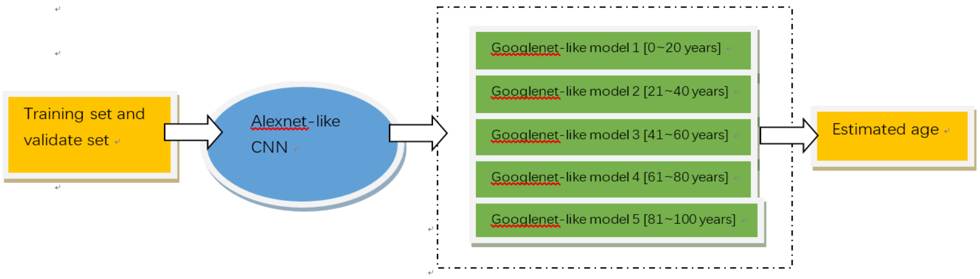

Training set and validate set
Alexnet-like CNN
Googlenet-like model 1 [0~20 years]
Googlenet-like model 2 [21~40 years]
Googlenet-like model 3 [41~60 years]
Googlenet-like model 4 [61~80 years]
Googlenet-like model 5 [81~100 years]
Estimated age

Figure 2: Stacked CNN model structure

3.1Simple and Coarse Alexnet-like Cnn Model

At this stage, the input samples were grouped into five subsets, one grouped sample subset corresponded to one age group and one age group corresponded to 20 years. Therefore,the whole samples set included all of training face age images which consist of all the training images for the 100 year classes. An alexnet-like cnn model [figure.2] was selected to classify the whole age duration into five coarse age parts, the model consisted of five convolution layers, three maxpool layers and one full connection layer, the last layer was softmax layer. Every convolution layer and full connection layer followed by a relu activation layer. In the model we cut the norm layer and the last dropout layer and addedthe batch-norm layer[11].This change could accelerate the convergence rate and improve the accuracy. And we found the accuracy could be improved three percent by this change and the training time could be saved by 30%.

3.2Complex and accurate googlenet-like cnn model

After the first coarse classification, the face age was classified into five age parts, then the first stage classification result was sent to the second classification stage which we chose the googlenet-like cnn model which consisted of thirteen convolution layers, three maxpooling layers and one mean pooling layer. In the net structure, the stage consisted of five googlenet models and every model could be used to classify the inputted face age image into fine age group. We simplified the googlenet model, and added two resnet layers [12]and one batchnorm layer in the net. And finally, a mixed loss layer was selected to classify the face age of the first stage result into 20 subclass, the details of the loss layer were showed in next section. The final age result wasthe fusion of the five maximized probability class of softmax output, and the computation details was showed in formula3.1:

$$R_{est}=Mean(Top5(softmax(prob))) \tag{3.1}$$

Where $softmax(prob)$ represents the output of softmax layer, function $Top5$ means the top five maximized probability year classes, function $WeightedMean()$ means how to compute the final concrete face age using the top five probability year classes output, formula is as follows:

$$WeightedMean(x_{1..5})=w_1x_1+w_2x_2+w_3x_3+w_4x_4+w_5x_5 \tag{3.2}$$

Where x_i represents the corresponding face age of the ith largest probability output of softmax layer, w_i decreases exponentially and should be normalized, we chose the suitable w by grid search training.

3.3Improved Loss Ensemble

It is difficult to train the stacked cnn models. Alternatively we can finetune the model from the pre-trained face recognition model., However, it may not be helpful for all the models and we found directly finetuning may fall into local extreme. To solve the training difficulty problem, we adopted the mixed loss[13],the formula is as follows

$$L=Ls+\lambda Lc=-\sum_{(i=1)}^{m}\log\left(e^{\left(W_{(y_i)}^T x_i+b_{(y_i)}\right)}/\sum_{(j=1)}^{n}e^{\left(W_j^T x_i+b_j\right)}\right)+\lambda/2\sum_{(i=1)}^{m}\|x_i-c_{(y_i)}\|_2^2 \tag{3.3}$$

Where $L_{softmax}$ is softmax loss, L_{cl} is the center loss[14] and $\varnothing$ is the weight which was set to 0.004 in all our experiments.

The improved loss can decrease inter class variance and increase extra class variance which is helpful to promote the validate accuracy. We found the accuracy can be increased by10 percent with the loss ensemble.

3.4Video Face Tracking

When the single frame face age was estimated, a mixed tracker consisting of RGB feature template and colored histogram feature template was used to track the face stably.

$$H_{mixed}=\alpha\times h_{rgb}+\beta\times h_{ch} \qquad (3.4)$$

Where H_{mixed} is the final tracking similarity result , h_{rgb} is the computed RGB feature template similarity value and h_{ch} is the colored histogram feature template similarity value, α and β are the weights set to 0.7 and 0.3 respectively in all of our experiments. We selected the template match as the tracking algorithm.

4.Experiments

4.1Data Set Selection

The ICCV2016 Apparent Age Estimation Challenge aimed to investigate the performance of estimation methods on apparent age rather than real age. A dataset of 4699 images was provided. Each image was labeled a real number from 0 to 100 indicating the apparent age. The images were collected from two web-bases application and labeled by at least 10 different users. We divided the dataset into five coarse age parts manually for the first alexnet-like model, and the original 4699 images were used for the training of the four googlenet-like models of the second stacked cnn models group. The ground truth wascalculated with all the users' votes, where the mean age from all the votes was considered as the apparent age label.

In the final test phase, both the training set and validation set can be used for model training. The performance wasmeasured by mean normalized error calculated as:

$$\varepsilon=\frac{1}{N}\sum_{i=1}^{N}1-e^{\frac{-(x-\mu)^2}{2\sigma^2}} \qquad (4.1)$$

where x is the estimated age , μ is the mean age, ε is the mean normalized error, N is the number of test samples, and σ is the standard deviation.

4.2Implementation Details

Firstly, the inputted RGB 124*124 images were cropped to 112*112 around the image center.

Secondly, we began to pre-train our two submodels using CASIA dataset which contained about 350,000 face images for face recognition, and then, fine-tune was performed on the combined face age datase., We formed the whole training age dataset using audience、FGNET、Morph subsets.

We trained our stacked model structure using the Caffe framework [15]. The base learning rate was set to 0.04 and multiplied by 0.3 when the validate loss stop decreased . The momentum was set as 0.95 and the weight decay was set as 0.0002. All the experiments was performed on an GTX-TitanX GPU machine with 11GB memory.

4.3Results

Before the training, we augmented the training data set., Three sample augment strategies were selected including:the three color channel pixel value added an offset randomly, respectively;50% probability mirror; producing occlusion at a random image small patch.

In Table 1, a comparison results were showed with and without data augmentation and an obvious total improved 5% gap could be achieved from data augmentation.

Data set augmentation strategy	accuracy
No augmentation	85.1%
With Only Color augmentation	87.5%
With Only mirror augmentation	86.3%
With Only patch occlusion augmentation	85.4%
With all the three augmentation strategies	90.8%

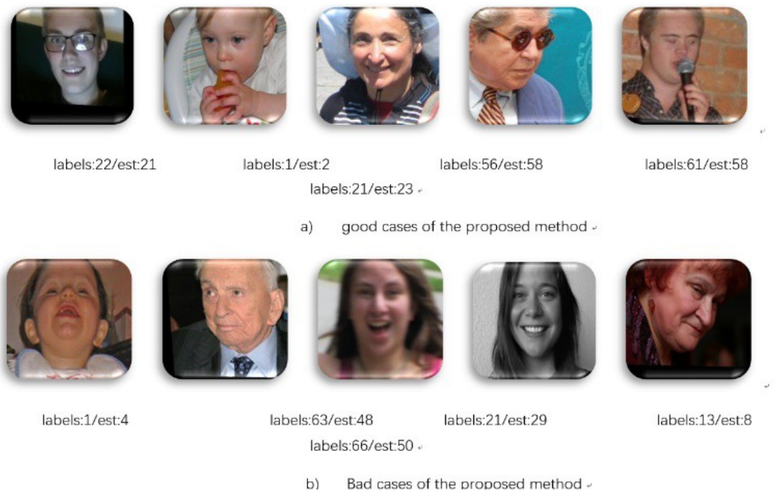
Table1: Data Augmentation Comparison

Table 2 presented our results for video age classification. In the table, the proposed method was compared with the methods proposed by Eidinger and Gil Levi. Evidently, the proposed method outperformed the reported state-of-the-art on the three tasks with considerable gaps. We thought that the improvements were achieved by the stacked cnn structure which can learn more discriminative features and reduce the overfitting.

Method	Accuracy
Best from Eidinger's	74.5±1.1
Best from Gil Levi's	82.3±1.5
Our method	90.8±1.2

Table 2: Three Methods Comparison

The partial experiment results were showed in Figure 3, where both good and bad cases of apparent age estimation results were presented. The results showed that our approach was robust to variations in pose, lighting, glass, occlusion. However, our approach did not work very well for face blur, mis-alignment, huge pose transformation and gray images ,which may caused by the lack of corresponded training data.



a) good cases of the proposed method

b) Bad cases of the proposed method

Figure 3: Examples of Age Estimation Results of the Proposed Method

5.Conclusion

In this paper, a robust video face age estimation was proposed byusing a stacked cnn model structure to estimated age. To reduce the risk of over-fitting, we firstly trained a alexnet-like model to classify the face age coarsely, then an accurate four combined googlenet-like models were used to estimate the final face age.Finally a multiple model tracker was used to track the face stably. The experiments showed that the propose method can achieve a good gap over the state of the art method.In the future, we will explore how to extend the data to solve the problems of low precision caused by gray model and huge pose transformation images.

References

[1]Y. H. Kwon and N. da Vitoria Lobo, *Age classification from facial images*[C]. Proceedings of the CVPR '94.IEEE,New Jersey. pp, 762–767(1994).

[2]E. Eidinger, R. Enbar, and T. Hassner. *Age and gender estimation of unfiltered faces*[J]. Trans. on Inform.Forensics and Security,9(12):2170-2179(2014).

[3]F. Gao and H. Ai, *Face age classification on consumer images with gabor feature and fuzzy lda method*[C], In Advances in biometrics.Springer, Berlin.pp, 132–141(2009).

[4]C. Liu and H. Wechsler,*Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition*[J]. Trans. Image Processing, 11(4):467–476(2002).

[5]K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. *Return of the devil in the details: Delving deep into convolutional nets*. arXiv preprint arXiv:1405.3531(2014).

[6]K. Simonyan, A. Zisserman .*Very Deep Convolutional Networks for Large-Scale Image Recognition.* arXiv:1409.1556(2014).

[7]Kaiming He and Xiangyu Zhang and Shaoqing Ren and Jian Sun,*Deep Residual Learning for Image Recognition.* arXiv preprint arXiv:1512.03385(2015).

[8]X. Geng, Z.-H. Zhou, and K. Smith-Miles. *Automatic age estimation based on facial aging patterns*[J]. Trans. Pattern Anal. Mach. Intell., 29(12):2234–2240(2007).

[9]G. Guo, Y. Fu, C. R. Dyer, and T. S. Huang. *Imagebased human age estimation by manifold learning and locally adjusted robust regression*[J]. Trans. Image Processing,17(7):1178–1188(2008).

[10]Gil Levi and Tal Hassner .*Age and Gender Classification using Convolutional Neural Networks*[C] .2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW).IEEE,New Jersey. pp, 34-42(2015)

[11]Sergey Ioffe, Christian Szegedy. *Batch Normalization: accelerating deep network training by reducing internal covariate shift*. arXiv:1502.03167v3(2015)

[12]Kaiming He, Xiangyu Zhang. *Deep residual learning for image recognition.* arXiv:1502.03385v1(2015)

[13]Wen,Y.,Li,Z.,Qiao,Y. *Latent factor guided convolutional neural networks for age-invariant face recognition*[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.IEEE,New Jersey.pp,4893–4901(2016).

[14]Yandong Wen, Kaipeng Zhang, Zhifeng Li, Yu Qiao .*A Discriminative Feature Learning Approach for Deep Face Recognition*[C] European Conference on Computer Vision. Springer, Berlin.pp, 499-515(2016).

[15]Y.Jia, E.Shelhamer, J.Donahue, S.Karayev, J.Long, R.Girshick, S. Guadarrama, and T. Darrell. *Caffe: Convolutional architecture for fast feature embedding*. arXiv preprint arXiv:1408.5093(2014).