Proceedings of Seminar and Project

TITLE

SEMESTER

Oliver Wasenmüller and Prof. Didier Stricker
Department Augmented Vision
University of Kaiserslautern and DFKI GmbH

Introduction

The seminar and project TITLE (INF-XX-XX-S-X, INF-XX-XX-L-X) are continuative courses based on and applying the knowledge taught in the lectures 3D Computer Vision (INF-73-51-V-7) and Computer Vision: Object and People Tracking (INF-73-52-V-7). The goal of the project is to research, design, implement and evaluate algorithms and methods for tackling computer vision problems. The seminar is more theoretical. Its educational objective is to train the ability to become acquainted with a specific research topic, review scientific articles and give a comprehensive presentation supported by media.

In the XXX semester XXX, XXX projects addressing XXX were developed. Moreover, XXX seminar works addressed XXX. The results are documented in these proceedings.

Organisers and supervisors

The courses are organised by the Department Augmented Vision (`http://ags.cs.uni-kl.de`), more specifically by:

Oliver Wasenmüller
Prof. Dr. Didier Stricker

In the XXX semester XXX, the projects were supervised by the following department members:

NAME

MONTH YEAR

Apparent/Real Age Estimation using Deep Learning

Basavaraj Hampiholi[1] and Mohamed Selim[2]

[1] basavaraj.hampiholi@dfki.uni-kl.de
[2] mohamed.selim@dfki.de

Abstract. This study describes the estimation of real/apparent age estimation in still face images using deep convolutional neural networks(CNNs). The CNNs achieved the better results compared to bio inspired features. In this case, a special CNN architecture VGG-16 is used for training. Age estimation is regression problem, but can be considered as classification by using discrete classes for each year. The larger datasets like IMDB-WIKI[11],MORPH-II,LAP,FG-NET helped to consider age as a classification problem as more samples per each class are available. A pre-trained VGG-16 on Imagenet is used as a classifier to train the IMDB-WIKI dataset The resulting model is fine-tuned using LAP dataset to find the apparent age. IMDB-WIKI dataset is imbalanced and contains less images of children compared to adults. Hence a separate children specific CNNs are introduced to train the children images between 0-12 years old. In this,first face detection is applied on test images and then CNN estimates the age from an ensemble of 20 networks.

Keywords: DEX, IMDB-WIKI, LAP, CNN, VGG-16

1 Introduction

Face analysis is one of the most important and rapidly growing area of research in Computer Vision and Pattern Recognition community. Automatic age estimation from facial images is also one of the most challenging topics because of following reasons: aging process is uncontrollable, aging patterns are personalized as age depends upon food, race, gender etc and variation among faces of the same age. Age estimation has many applications like customer profiling, search optimization in large databases, assistance of bio-metrics systems,video surveillance, Demographic statistics collection.

Majority of researches focus on real age estimation since from [8], but with less significant results compared to CNNs. This field regained interest with the availability of large databases FG-NET,MORPH-NET. With larger number of samples, the discretization error between each class is low and hence people started estimating age with classification rather than regression. In contrast, the estimation of apparent age, that is the age perceived by the others is also progressing rapidly since the introduction of the ChaLearn's LAP dataset for apparent age estimation.

Main motive of this study is to estimate real and apparent age using deep convolutional neural networks(CNNs). Although there were many researches in this field, introduction of the larger public datasets like IMDB-WIKI(real age) by [11] and LAP(apparent age) has really promoted the research in this area. IMDB-WIKI is the largest available dataset for real age estimation and LAP dataset is the first State of the art database for apparent age.

The rest of the paper is organized as follows: in section 2, related works on real/apparent age estimation using different state of the art methods are presented. In section 3, the datasets like IMDB-WIKI and LAP are described. Section 4 discusses about different approaches for real/apparent age estimation, mainly DEX and children specific DEX. In section 5, evaluation protocol for age estimation, results and different experiments on validation set with different processing steps are discussed. Finally section 6, summarizes the conclusions and future work.

2 Related Works

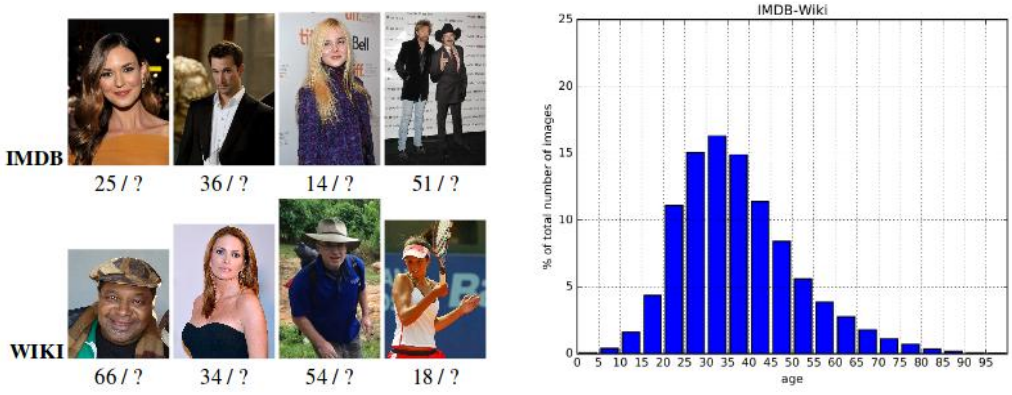
Age Estimation Automatic age estimation is an important and challenging problem in facial analysis for computer vision and pattern recognition community. Real age estimation has made significant progress with impressive accuracies after decades of research due to the availability of large public face databases. The relevant research and literature is rich [4],[5],[8],[15]. However apparent age estimation is new area to explore and significant advances have been also made in this direction too and ChaLearns Look at People (LAP) challenge has really enhanced the research in this field. The recent studies are [1],[2], [11]

Deep Learning for Facial Analysis Recently, Deep Learning has inspired many computer vision domains like face detection, face recognition, facial gender recognition, and facial emotion recognition. However several issues of face analysis are still open like larger face detection and recognition, emotion recognition in which the community keeps improving with excellent results. Facial age estimation is among one of them and its making slower progress compared to other face analysis problems as there are less available public face datasets with age information.

Learning with CNNs Inspired by the animal visual cortex, convolutional neural networks have been impressive in solving computer vision and pattern recognition problems. Alex Krizhevsky et al.[7], trained a large deep convolutional neural network to classify the 1.2 million high-resolution images in the ILSVRC-2012 and achieved a winning top error rate of 15.4% compared to next best result of 26.2%. This shows CNNs work better compared to any other SoA with image data. Since our problem is estimating the age by looking at face images, the CNNs can be used as solution.Followed by AlexNet, many CNN architectures were introduced by several researchers and major of them are GoogleNet[13], VGG Net[12] and Microsoft ResNet[6]. Among them, VGG-16 is a simple and deep architecture with significant less number of parameters compared to AlexNet.

3 Datasets

There are many datasets available for real age estimation, like IMDB-WIKI, MORPH-II, FG-NET, Adience. Some of them contains the samples for age group instead of single age value, like Adience. The datasets for apparent age estimation are very less and the only available public dataset is LAP. In this SoA both IMDB-WIKI and LAP dataset are used for real and apparent age estimation respectively. Hence we discuss the process of dataset preparation for IMDB-WIKI and LAP in below sections.



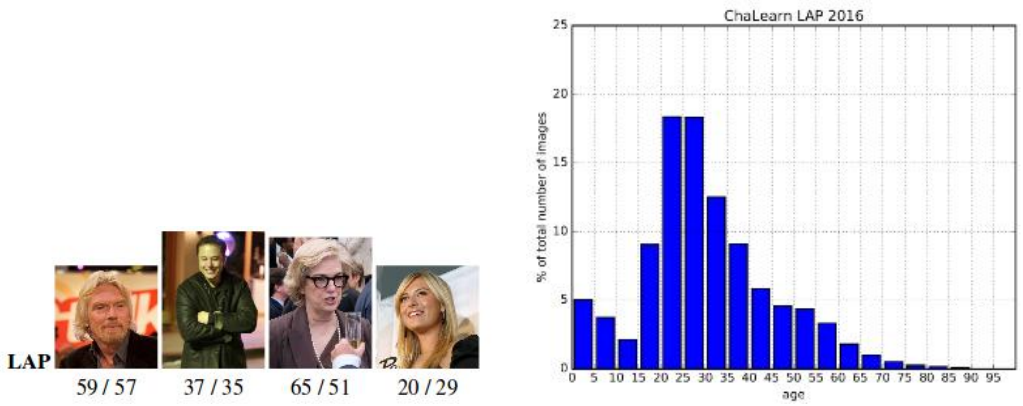
(a) (b)

Fig. 1. IMDB-WIKI dataset with sample images and distribution [11][2]

3.1 IMDB-Wiki Dataset

This is largest dataset available for real age estimation problem. Most popular 100,000 actors were listed as per the IMDB ranking and automatically crawled from their profiles birth dates, images, and annotations. Difference between the date of birth and photo taken year is labeled as real age. Images without photo taken year and with multiple high scored face detections are removed. But the accuracy of the dataset is not guaranteed as many images are stills from movies and have wrong time stamps. In total, 461,871 face images of celebrities were obtained.

Wikipedia profile pictures were crawled and filtered as per the same criteria applied to IMBD images and collected 62,359 images. Finally, total of 524,230 images were taken from both the sources with age information. In case of images with several faces, images with second strongest face detection score below a threshold value were taken into consideration. Age distribution was equalized by randomly dropping some of the images of the most frequent ages. In Fig.1.(a) shows some sample images from LAP annotated with real age and Fig.1.(b) shows the dataset distribution. From the dataset distribution it can be seen that there are less number of samples for children and senior citizens compared to adults.



(a) (b)

Fig. 2. LAP dataset with sample images and distribution [11][2]

3.2 LAP Dataset

The first state of the art database for Apparent Age Recognition rather than Real Age recognition. To the date, ChaLearn LAP dataset V2 contains around 8000 images, which are labeled by multiple individuals(at least 10) using a collaborative Facebook implementation and Amazon Mechanical Turk. Hence the dataset is labeled with mean and standard deviation. The votes variance is used as a measure of the error for the predictions. The dataset is split into 4113 images for training, 1500 for validation and rest of the images for testing. The age distribution is the same in all the three sets of the LAP dataset.In Fig.2.(a) shows some sample images from LAP annotated with apparent age and Fig.2.(b) shows the dataset distribution. From the dataset distribution it can be seen that there are less number of samples for children and senior citizens compared to adults.

4 Approaches

There are many SoA approaches for age estimation are available using CNNs itself. One among them is DEX: Deep expectation of apparent age from single still faces[11] which attained excellent results and won LAP-2015 challenge for apparent age estimation. Another SoA is children specialized DEX[2] based on DEX which is also able achieve better results than DEX itself and won the LAP-2016 challenge. We discuss both the (SoA)s in below sections.

4.1 DEX: Deep Expectation of Apparent Age

Age estimation using CNNs follows the process of face detection, face alignment and training. DEX uses VGG-16 CNN architecture for training. Fig.3. shows the pipeline in detail and the same is explained below.

Face detection Mathias et al., face detector is used to obtain the location of the face. Face detector is run on the original image as well as on all rotated versions between -60 to 60 degrees in 5 degree steps and also on -90,90 and 180 degrees for upside down photos. The face with strongest detection score is taken and rotated it accordingly to up-frontal position. In case face detector failed (¡0.2) to find the faces, entire image is taken. Then face size is extended by adding 40 margin to left, right, above and below.

Training Pipeline The pipeline for training is shown in Figure 3. One of the most simple and deep CNN architecture VGG-16 is used for training. A pre-trained model of VGG-16 on ImageNet, is fine-tuned on IMDB-WIKI dataset. While training the classifier the output layer is changed to 101(0-100) output neurons and for regression only one neuron at the output layer. For real age estimation, the IMDB-WIKI dataset is divided into training, testing splits and pre-trained VGG-16 is trained on it. For apparent age estimation, the LAP dataset is divided into 20 splits with equal age distribution in each split. In each split, 90 of the data is used for training and 10 for testing. The data augmentation is performed on LAP dataset before training. Finally all the split data is trained on ensemble of 20 networks and the prediction is average of all. For improving the accuracy of model softmax expected value E is computed as follows:

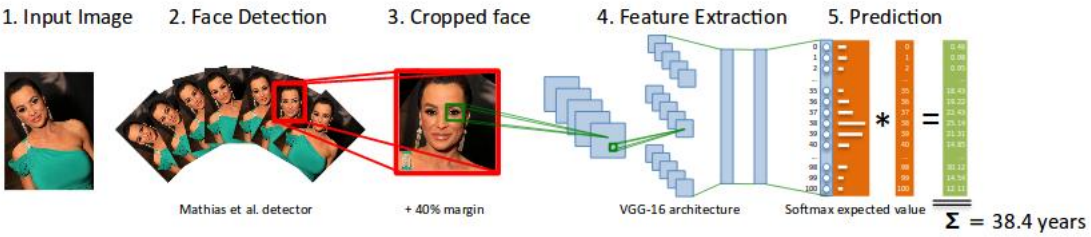


Fig. 3. Training pipeline of DEX method for apparent age estimation[11]

4.2 Children Specialized DEX

This SoA is based upon the DEX approach itself with more improvements. In this Mathias face detector is applied for face detection same like DEX but aligned the faces using facial landmark detection which was missing in original DEX. As it can be observed from the IMDB-WIKI dataset, the children images are very less and hence a private children dataset is collected. A separate children specific CNN is trained using this children dataset. One more improvement is using the age label encoding which is missing in DEX. All these are discussed in detail in following sections.

Face Detection The open source face detector Head Hunter is used for face detection and each image is rotated at all angles in the range [-90 90] with step of 50 . Finally images with strongest face detection scores are taken for face alignment step. If no faces found in images 2 upscaling operations are done to find at least one face. As described in 4.1, the face area is extended by 40 margin on height, width, right and left.

Face Alignment The state of the art for face alignment is based upon the multi-view facial landmark detection by [14]. There are 5 landmark detection models: a frontal, 2 profile, 2 half-profile

models. Model with highest confidence score is selected and then perform an affine transformation from the detected landmarks.

Age Labels Encoding

Real Number Encoding: The age labels are the real numbers and its for regression approach

0/1 Classification Encoding: In this, the age labels are encoded as binary vectors containing a single non-zero value corresponding to the class. This is pure classification approach.

Label Distribution Encoding: Here, there are predefined number of classes as well as the labels are encoded with real-valued vectors which represents probability distributions

Children Dataset There are very less images of children(0-12) in IMDB-WIKI dataset. Hence there are high chances of age estimation model failure for children images. As per the annotations of LAP dataset, it can be observed that standard deviation of human votes for children images is about 1 and for others it is about 5. This shows that humans estimate an age of the child almost 5 times accurately than age of an adult. Therefore, a private dataset of 5723 children(ages between 0-12) images were manually collected and used for training.

Network Architecture In this approach, there are two different CNNs are used-General and Children specific CNNs both are of VGG-16 architecture. The number of output neurons are different for each of them. In General CNNs the output layer contains 100 neurons(0-99) with Sigmoid activation function. In case of Children specific CNNs, the output layer contains 13 neurons(0-12) with global softmax activation function. Both the CNNs are optimized using gradient descent with momentum of 0.9 with mini-batches of 32 images. The deep learning framework used for this task is Caffe and the GPU is Tesla K40c.

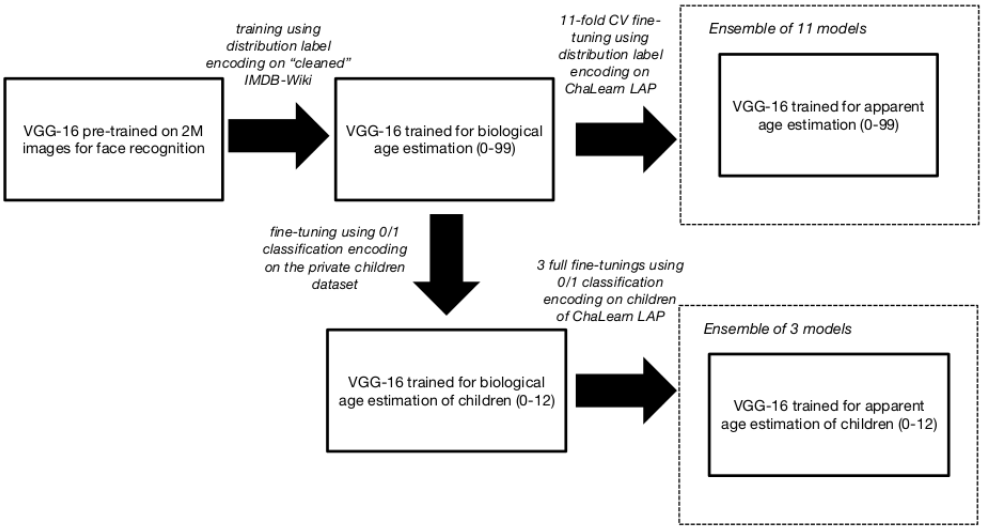


Fig. 4. Training pipeline for children specialized DEX [2]

Training Pipeline The pipeline for testing is shown in Figure 4. In this state of the art, the pretrained VGGNet from [10] is used to train the general CNN for biological age estimation for all ages between 0 to 99 on IMDB-WIKI dataset using distributed labeling strategy. The children CNN for biological age estimation between 0 to 12 years is then fine-tuned on the obtained network using 0/1 encoding strategy. Then the two CNNs are fine-tuned for apparent age estimation. For general CNNs both training and validation images LAP dataset are combined and fine-tuned 11 general CNNs for apparent age estimation using 11-fold cross validation. In case of children CNNs, children

images from both training and validation sets are combined and fine-tuned CNNs for apparent age estimation without any validation by just saving the weights at 3 predefined points which have been chosen by experimentation on validation set. Hence there are three children CNNs.

Testing Pipeline In this phase, the images are pre-processed using head hunter face detector and aligned using facial landmarks detection same as training set. The image is resized to 224*224 as an input to VGG-16. The resulting image is then modified to seven different versions like, the ones rotated at +-5, the ones shifted by 5 perc on left or right, the ones scaled in/out by 5 perc.In total there are 8 images including the original image. After all these pre-processing steps, all the obtained images are processed using 11 general CNNs having 100 output neurons. The output of each CNN is averaged and normalized to sum upto 1. Hence the final genaral age prediction is calculated as an expected value of the softmax probabilities:*general age*= $\sum_{i=0}^{99} i * p_i$. If the predicted age is greater than 12, then it is considered as final apparent age. In case, if it is below 12, the same 8 images are processed for 3 children CNNs which are having 13 output neurons(0-12). Similar to general CNNs, the average of all 3 children CNNs are taken and normalized them to 1. Again the final children age prediction with 13(0-12) values representing probabilities are calculated as an expected value: *children age*= $\sum_{i=0}^{12} i * p_i$

5 Experiments and Results

The results are evaluated by using the standard mean absolute error (MAE) and Gaussian error(ϵ). MAE is computed as the average of absolute errors between the estimated age(ex) and the ground truth ages (x). The Gaussian error fits the normal distribution with mean μ and standard deviation σ of the votes for each image.

$$MAE=\frac{1}{N}\sum_{i=1}^{N}|ex_i - x| \quad \text{and} \quad \epsilon = 1 - e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

5.1 Results

DEX: The network is experimented to train for both training regression and learning a regression on top of CNN features from convolutional layers. But the softmax expected value on the network trained for classification worked better than both of these methods. Table.1 reports the MAE and e-error for different setups. This shows training LAP without IMDB WIKI dataset results in large error rate compared to training with IMDB WIKI dataset. Also the network directly trained on regression, classification has more error than the classification+Expected value.The softmax expected value on LAP validation set resulted in less error rate with MAE of 3.221 and e-error of 0.278.

pretrain	finetune	Learning	MAE	ϵ-error
ImageNet	LAP	Regression	5.007	0.431
		Classification	7.216	0.549
		Classification+Expected Value	6.082	0.508
ImageNet and IMDB-WIKI	LAP	Regression	3.531	0.301
		Classification	3.349	0.291
		Classification+Expected Value	3.221	0.278

Table 1. Performance of DEX on validation set of ChaLearn LAP 2015[11]

However this method fails in some cases. The causes are, failure in face detection phase, dark images, images with glasses and old photographs. Fig.5 shows the face images for which DEX fails to predict the age.

Children Specialized DEX: The experimental results for this SoA are tabulated in Table 2. The table contains ϵ-error for different setups like using different face detectors, only biological age

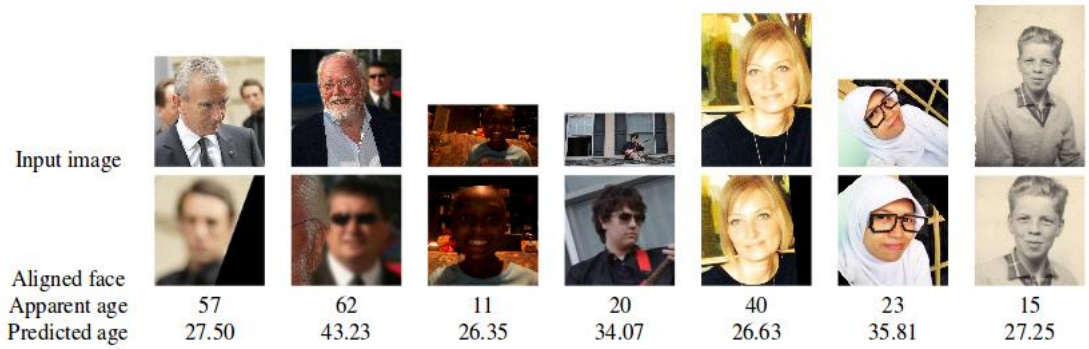


Fig. 5. Images for which DEX fails[11]

estimation, data augmentation and using children model. First, the model is only trained on IMDB WIKI dataset to estimate real age and ϵ-error was 0.3927. To estimate the apparent age the model was first fine-tuned on IMDB WIKI dataset and then trained on LAP dataset which resulted ϵ-error of 0.2986.This clearly shows the gap of almost 0.1 ϵ-error score indicating the importance of fine-tuning on the competition dataset as well as the difference between the real age and apparent age estimation. Also the face detection and alignment using [14] and [9] helped to reduce the ϵ-error by 0.01. The data augmentation during the fine-tuning for apparent age estimation has further reduced the ϵ-error. Finally ϵ-error for apparent age estimation using children model on validation set has reduced to 0.2609. In the LAP challenge the final score on test data was 0.2411.

Biological age training	Apparent age fine-tuning	Image preprocessing(face detection+face alignment)	Data augmentation during apparent age fine-tuning	Data augmentation during testing	Children model	ϵ-score
Yes	No	[9]+[14]	No	No	No	0.3927
Yes	Yes	[16]+[3]	No	No	No	0.3086
Yes	Yes	[9]+[14]	No	No	No	0.2986
Yes	Yes	[9]+[14]	Yes	No	No	0.2825
Yes	Yes	[9]+[14]	Yes	Yes	No	0.2782
Yes	Yes	[9]+[14]	Yes	Yes	Yes	0.2609

Table 2. Performance of children specialized DEX on validation set of ChaLearn LAP 2016[2]

5.2 Discussion and Comparison

There are many researches on Real age estimation using different (SoA)s since [8]. The ChaLearn's LAP challenge has really motivated the research on apparent age estimation too and there are many recent researches progressing in this field. One such break-through in this direction was by [11] as they introduced new dataset IMDB-WIKI for age estimation. This SoA method produced excellent results compared to the previous which led them to win ChaLearn's 2015 challenge. It's also discussed that the results would be better by applying facial landmark detection. One more area of improvement is the distribution of the IMDB-WIKI dataset. The dataset is imbalanced and majority of the sample images are of adults. To balance the dataset and improve accuracy [2] introduced children specialized CNNs. There are about 5700 images of children(ages between 0-12) were collected manually and trained on separate CNN. There are two CNNs trained- General

and Children specific. In this SoA, facial landmark detection is used for face alignment which was missing in DEX. All these resulted in better validation accuracy of 0.2609 compared to DEX 0.278. Final ϵ-score on LAP test dataset was 0.2411 compared to DEX score 0.2649 with a significant difference of 0.02.

6 Conclusions and Future Work

In this work, the estimation of both real age and apparent age in still face images using CNNs is presented. The results of this study show that CNNs yield better accuracy compared other (SoA)s like[5, 4]. The accuracy can further be improved by using proper image preprocessing and data augmentation techniques. Also it is evident from children specialized DEX that, a pretrained model on face images[10] helps to increase the accuracy than a model pretrained on ImageNet. Since age is considered as classification problem in this study, it is better to have balanced dataset for all classes(0-99), but IMDB-WIKI is highly imbalanced and concentrates on adults. Hence, collecting and training more images in the category of children and senior citizens can further improve accuracy of the model.

References

1. E. Agustsson, R. Timofte, S. Escalera, X. Baro, I. Guyon, and R. Rothe. *Apparent and Real Age Estimation in Still Images with Deep Residual Regressors on Appa-Real Database*. May 2017.
2. Grigory Antipov, Moez Baccouche, Sid-Ahmed Berrani, and Jean-Luc Dugelay. *Apparent Age Estimation from Face Images Combining General and Children-Specialized Deep Learning Models*. Las Vegas, USA, 2016.
3. P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar. *Localizing Parts of Faces Using a Consensus of Exemplars*. CVPR '11. IEEE Computer Society, Washington, DC, USA, 2011.
4. X. Geng, Z. H. Zhou, and K. Smith-Miles. *Automatic Age Estimation Based on Facial Aging Patterns*, volume 29. Dec 2007.
5. G. Guo, Guowang Mu, Y. Fu, and T. S. Huang. *Human age estimation using bio-inspired features*. June 2009.
6. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. *Deep Residual Learning for Image Recognition*, volume abs/1512.03385. 2015.
7. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. *Imagenet classification with deep convolutional neural networks*. 2012.
8. Young Ho Kwon and N. da Vitoria Lobo. *Age classification from facial images*. Jun 1994.
9. Markus Mathias, Rodrigo Benenson, Marco Pedersoli, and Luc Van Gool. *Face Detection without Bells and Whistles*. Springer International Publishing, Cham, 2014.
10. O. M. Parkhi, A. Vedaldi, and A. Zisserman. *Deep Face Recognition*. 2015.
11. Rasmus Rothe, Radu Timofte, and Luc Van Gool. *DEX: Deep EXpectation of apparent age from a single image*. December 2015.
12. K. Simonyan and A. Zisserman. *Very Deep Convolutional Networks for Large-Scale Image Recognition*, volume abs/1409.1556. 2014.
13. Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. *Going Deeper with Convolutions*. 2015.
14. M. Ui, V. Franc, D. Thomas, A. Sugimoto, and V. Hlav. *Real-time multi-view facial landmark detector learned by the structured output SVM*, volume 02. May 2015.
15. Dong Yi, Zhen Lei, and Stan Z. Li. *Age Estimation by Multi-scale Convolutional Network*. Springer International Publishing, Cham, 2015.
16. Lun Zhang, Rufeng Chu, Shiming Xiang, Shengcai Liao, and Stan Z. Li. *Face Detection Based on Multi-Block LBP Representation*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007.