Int J Comput Vis
DOI 10.1007/s11263-016-0940-3

Deep Expectation of Real and Apparent Age from a Single Image Without Facial Landmarks

Rasmus Rothe[1] · **Radu Timofte**[1] · **Luc Van Gool**[1,2]

Received: 15 February 2016 / Accepted: 20 July 2016

Abstract In this paper we propose a deep learning solution to age estimation from a single face image without the use of facial landmarks and introduce the IMDB-WIKI dataset, the largest public dataset of face images with age and gender labels. If the real age estimation research spans over decades, the study of apparent age estimation or the age as perceived by other humans from a face image is a recent endeavor. We tackle both tasks with our convolutional neural networks (CNNs) of VGG-16 architecture which are pre-trained on ImageNet for image classification. We pose the age estimation problem as a deep classification problem followed by a softmax expected value refinement. The key factors of our solution are: deep learned models from large data, robust face alignment, and expected value formulation for age regression. We validate our methods on standard benchmarks and achieve state-of-the-art results for both real and apparent age estimation.

Keywords Age estimation · Deep learning · CNN · Regression

1 Introduction

Age estimation from a single face image (see Fig. 1) is an important task in human and computer vision which has many applications such as in forensics or social media. It is closely related to the prediction of other biometrics and facial attributes tasks such as gender, ethnicity, hair color and expressions. A large amount of research has been devoted to age estimation from a face image under its most known form—the real, biological, age estimation. This research spans decades as summarized in large studies (Panis et al. 2016; Chen et al. 2015; Eidinger et al. 2014; Han et al. 2013; Guo 2012). Several public standard datasets (Chen et al. 2015; Panis et al. 2016; Ricanek and Tesafaye 2006) for real age estimation permit public performance comparison of the proposed methods. In contrast, the study of apparent age, that is the age as perceived by other humans, is at the beginning. The ChaLearn Looking At People ICCV 2015 challenge (Escalera et al. 2015) provided the largest dataset known to date of images with apparent age annotations, here

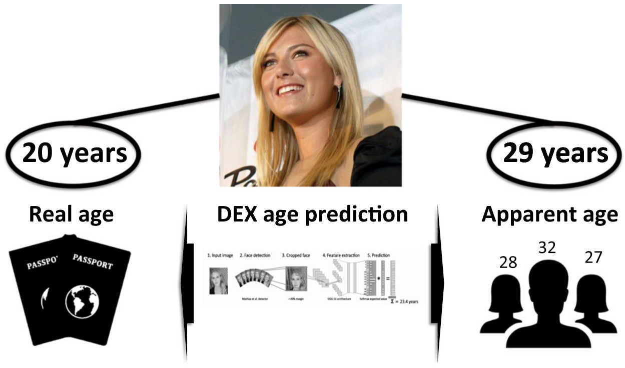

Fig. 1 Predicting the real and apparent age of a person

Communicated by Cordelia Schmid, Sergio Escalera, Jordi Gonzàlez, Xavier Barò, Isabelle Guyon and Hugo Jair Escalante.

✉ Rasmus Rothe
rrothe@vision.ee.ethz.ch

Radu Timofte
timofter@vision.ee.ethz.ch

Luc Van Gool
vangool@vision.ee.ethz.ch

[1] Computer Vision Lab, D-ITET, ETH Zurich, Switzerland

[2] VISICS/iMinds, ESAT, KU Leuven, Belgium

Published online: 10 August 2016

called the LAP dataset, and 115 registered teams proposed novel solutions to the problem.

With the recent rapid emergence of the intelligent applications there is a growing demand for automatic extraction of biometric information from face images or videos. Applications where age estimation can play an important role include: (1) access control, e.g., restricting the access of minors to sensible products like alcohol from vending machines or to events with adult content; (2) human–computer interaction (HCI), e.g., by a smart agent estimating the age of a nearby person or an advertisement board adapting its offer for young, adult, or elderly people, accordingly; (3) law enforcement, e.g., automatic scanning of video records for suspects with an age estimation can help during investigations; (4) surveillance, e.g., automatic detection of unattended children at unusual hours and places; (5) perceived age, e.g., there is a large interest of the general public in the perceived age (c.f. https://howhot.io/), also relevant when assessing plastic surgery, facial beauty product development, theater and movie role casting, or human resources help for public age specific role employment.

One should note that the intelligent applications need to tackle age estimation under unconstrained settings, that is, the face is not aligned and under known, unchanged, light and background conditions. Therefore, in the wild, a face needs first to be detected, then aligned, and, finally, used as input for an age estimator. It is particularly this setup we target in our paper with our system. Despite the recent progress (Panis et al. 2016; Rothe et al. 2016; Escalera et al. 2015) the handling of faces in the wild and the accurate prediction of age remains a challenging problem.

1.1 Proposed Method

Our approach—called Deep EXpectation (DEX)—to age estimation is motivated by the recent advances in fields such as image classification (Ciregan et al. 2012; Krizhevsky et al. 2012; Russakovsky et al. 2015) or object detection (Girshick et al. 2014) fueled by deep learning. From the deep learning literature we learn four key ideas that we apply to our solution: (1) the deeper the neural networks (by sheer increase of parameters / model complexity) are the better the capacity to model highly non-linear transformations—with some optimal depth on current architectures as (He et al. 2015) suggests; (2) the larger and more diverse the datasets used for training are the better the network learns to generalize and the more robust it becomes to over-fitting; (3) the alignment of the object in the input image impacts the overall performance; (4) when the training data is small the best is to fine-tune a network pre-trained for comparable inputs and goals and thus to benefit from the transferred knowledge.

We always start by first rotating the input image at different angles to then pick the face detection (Mathias et al. 2014) with the highest score. We align the face using the angle and crop it for the subsequent steps. This is a simple but robust procedure which does not involve facial landmark detection. For our convolutional neural networks (CNNs) we use the deep VGG-16 architecture (Simonyan and Zisserman 2014). We always start from pre-trained CNNs on the large ImageNet (Russakovsky et al. 2015) dataset for image classification such that (1) to benefit from the representation learned to discriminate 1000 object categories in images, and (2) to have a meaningful representation and a warm start for further re-training or fine-tuning on relatively small(er) face datasets. Fine-tuning the CNNs on face images with age annotations is a necessary step for superior performance, as the CNN adapts to best fit to the particular data distribution and target of age estimation. Due to the scarcity of face images with (apparent) age annotation, we explore the benefit of fine-tuning over crawled Internet face images with available (biological, real) age. We crawl 523,051 face images from the IMDb and Wikipedia websites to form IMDB-WIKI - our new dataset which we make publicly available. Figure 4 shows some images. It is the largest public dataset with gender and real age annotations. While age estimation is a regression problem, we go further and cast the age estimation as a multi-class classification of age bins followed by a softmax expected value refinement.

Our main contributions are as follows:

1. The IMDB-WIKI dataset, the largest dataset with real age and gender annotations;
2. A novel regression formulation through a deep classification followed by expected value refinement;
3. The DEX system, winner of the LAP 2015 challenge (Escalera et al. 2015) on apparent age estimation.

This work is an extended and detailed version of our previous LAP challenge report paper Rothe et al. (2015). We now officially introduce our IMDB-WIKI dataset for apparent age estimation, provide a more in depth analysis of the proposed DEX system, and apply the method and report results also on standard real age estimation datasets.

The remainder of the paper is organized as follows. Section 2 briefly reviews related age estimation literature. Section 3 introduces our proposed method (DEX). Section 4 introduces publicly our new IMDB-WIKI dataset with faces in the wild and age and gender labels, then describes the experimental setups and discusses the achieved results. Section 5 concludes the paper.

2 Related Work

While almost all literature prior the LAP 2015 challenge focuses on real (biological) age estimation from a face image, Han et al. (2015) provide a study on demographic estimation in relation to human perception and machine performance. In the next, we briefly review the age estimation literature and describe a couple of methods that most relate with our proposed method. We refer to Panis et al. (2016), Guo (2012), Fu et al. (2010), Han et al. (2015), Chen et al. (2015) and Eidinger et al. (2014) for broader literature reviews.

2.1 Real Age Estimation

Most of the prior literature assumes a normalized (frontal) view of the face in the input image or employ a face preprocessing step such that the face is localized and an alignment of the face is determined for the subsequent processing steps. Generally, the age estimators work on a number of extracted features, feature representations and learn models from training data such that to minimize the age estimation error on a validation data. The whole process assumes that the train, validation, and test data have the same distribution and are captured under the same conditions.

FG-NET (Panis et al. 2016) and MORPH (Ricanek and Tesafaye 2006) datasets with face images and (real) age labels are the most used datasets allowing for comparison of methods and performance reporting under the same benchmarking conditions. We refer to Panis et al. (2016) for an overview of research (365+ indexed papers) on facial aging with results reported on FG-NET dataset.

A large number of face models has been proposed. We follow the taxonomy from Guo (2012) and mention: wrinkle models (Kwon and Vitoria 1999), anthropometric models (Farkas and Schendel 1995; Kwon and Vitoria 1999; Ramanathan and Chellappa 2006), active appearance models (AAM) (Cootes et al. 2001), aging pattern subspace (Geng et al. 2007), age manifold (Fu and Huang 2008; Guo et al. 2008; Guo and Mu 2011), biologically-inspired models [including biologically-inspired features (BIF) (Guo et al. 2009)], compositional and dynamic models (Xu et al. 2008; Suo et al. 2010), local spatially fexible patches (Yan et al. 2008), and methods using fast Fourier transform (FFT) and genetic algorithm (GA) for feature extraction and selection (Fukai et al. 2007), local binary patterns (LBP) (Yang and Ai 2007), Gabor filters (Gao and Ai 2009). Recently, the convolutional neural networks (CNN) (LeCun et al. 1998), biologically inspired, were successfully deployed for face modeling and age estimation (Wang et al. 2015; Levi and Hassner 2015; Uricar et al. 2016).

The age estimation problem can be seen as a regression (Fu and Huang 2008) or as a classification problem up to a quantization error (Lanitis et al. 2004; Geng et al. 2007). Among the most popular regression techniques we mention Support Vector Regression (SVR) (Drucker et al. 1997), Partial Least Squares (PLS) (Geladi and Kowalski 1986), Canonical Correlation Analysis (CCA) (Hardoon et al. 2004), while for classification the traditional nearest neighbor (NN) and Support Vector Matchines (SVMs) (Cortes and Vapnik 1995).

In the next we select a couple of the representative (real) age estimation methods. Yan et al. (2007) employ a regressor learning from uncertain labels, Guo et al. (2008) learn a manifold and local SVRs, Han et al. (2015) apply age group classification and within group regression (DIF), Geng et al. (2007) introduce AGES (AGing pattErn Subspace), Zhang and Yeung (2010) propose a multi-task warped gaussian process (MTWGP), Chen et al. (2013) derive CA-SVR with a cumulative attribute space and SVR, Chang et al. (2011) rank hyperplanes for age estimation (OHRank), Huerta et al. (2014) fuse texture and local appearance descriptors, Luu et al. (2009) use AAM and SVR, while Guo and Mu (2014) use CCA and PLS.

Recently, Yi et al. (2014) deployed a multiscale CNN, Wang et al. (2015) used deep learned features (DLA) in a CNN way, while Rothe et al. (2016) went deeper with CNNs and SVR for accurate real age estimation on top of the CNN learned features.

2.2 Apparent Age Estimation

Our DEX (Rothe et al. 2015) method (CVL_ETHZ team, 1st place in LAP challenge) was initially introduced for apparent age estimation at the ChaLearn LAP 2015 challenge (Escalera et al. 2015). This work is an extension, releasing the IMDB-WIKI age estimation dataset with some in-depth analysis. Furthermore, this paper shows that the model presented in Rothe et al. (2015) achieves state-of-the-art also on real age estimation. Some more detailed qualitative and quantitative evaluations in this paper confirm the robustness and good performance of the DEX model. We review several runner-up methods that relate the most to our work and refer to Escalera et al. (2015) and Sect. 4.2.2 for more details on the LAP challenge. These methods are representative since LAP is the largest dataset to date on apparent age estimation and the methods employ deep learning and are the best out of 115 registered participants. A note is due: all the following apparent age estimation techniques are either pre-trained for real age estimation or can easily be adapted to it.

Liu et al. (2015) [ICT-VIPL team, 2nd place in LAP challenge (Escalera et al. 2015), see Table 3] proposes the following approach based on general to specific deep transfer learning and GoogleNet architecture (Szegedy et al. 2015) for 22-layers CNN. (1) Pre-train CNN for multiclass face classification using the CASIA-WebFace database and softmax loss; (2) fine-tune CNN for age estimation on large extra age dataset with two losses: Euclidean for age encoding and cross-entropy loss of label distribution learning based age

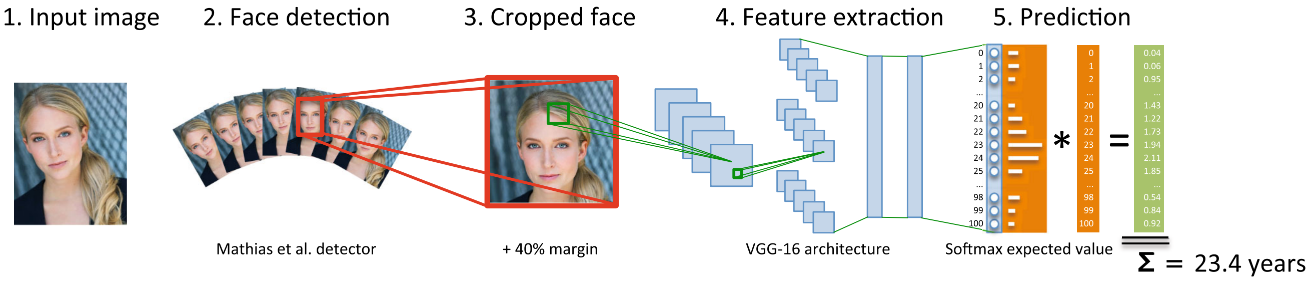

Fig. 2 Pipeline of DEX method for age estimation

encoding; (3) fine-tune CNN on the LAP apparent age data; (4) ensemble Learning and fusion of 10 CNNs.

Zhu et al. (2015) (WVU_CVL team, 3rd place in LAP challenge) employ GoogleNet (Szegedy et al. 2015) deep CNN networks trained on thousands of public facial images with real age labels. These are then fine-tuned on LAP apparent age data and then the CNN features are extracted. Random Forest and SVR are learned on each of ten age groups for age estimation and then their results are fused at test time.

Yang et al. (2015) (SEU-NJU team, 4th place in LAP challenge) use face and landmark detection for face alignment and the VGG-16 architecture (Szegedy et al. 2015) for modeling. Private and MORPH 2 data are used for training of multiple networks with different setups, aligned and non-aligned faces, different color spaces, filters, objective losses. The final prediction is a fusion.

UMD team (5th place in LAP challenge) employs face and landmark detection (Kazemi and Sullivan 2014), a CNN model (Chen et al. 2016), Adience (Eidinger et al. 2014) and MORPH datasets. A classification in three age groups is followed by age regression.

Enjuto team (6th place in LAP challenge) use face detection (Mathias et al. 2014) and face landmark detection (Kazemi and Sullivan 2014) and six CNNs for classification in three age groups and for local (part face) and global (whole face) prediction of age. The results are fused.

3 Proposed Method (DEX)

The proposed method, Deep EXpectation (DEX) follows the pipeline in Fig. 2. In this section each step of the pipeline is explained in detail.

3.1 Face Alignment

As many datasets used in this work do not show centered frontal faces but rather faces in the wild (cf. Figs. 2(1), 4), we detect and align the faces for both training and testing.

An ideal input face image should be of the same or comparable size, centered, and aligned to a normalized position and with minimum background. We choose the off-the-shelf (Mathias et al. 2014) face detector to obtain the location and size (scale) of the face in each image. This state-of-the-art face detector uses the Deformable Parts Model (DPM) (Felzenszwalb et al. 2010) and inherits robust performance. As expected, by cropping the detected face for the following age estimation processing instead of using the entire image we obtain massive increases in performance.

Many approaches employ rather complex alignment procedures involving accurate facial landmark detectors and image warping (Taigman et al. 2014; Yang et al. 2015). In our preliminary experiments we observed that the failure of the landmark detectors is difficult to predict and harms the performance as it leads to wrong face alignments. Since we target faces in the wild, use a robust face detector, and our CNNs can tolerate small alignment errors, we build our alignment procedure as follows.

We explicitly handle rotation by running the detector not only on the original image but on images rotated with steps of 5° [cf. Fig. 2(2)]. Due to limited computational resources we check only angles between −60° and 60°. Additionally we run the detector at −90°, 90° and 180° to cope with flipped or rotated images. At the end the face with the highest detection score across all rotations is picked and then rotated to up-frontal position.

For very few face images, the detector is unable to detect a face. In those cases the entire image is taken as the face. We notice that performance increases when considering also the context around the face. Therefore we extend the detected face by taking additional 40 % of the width and height of the face on all sides [cf. Fig. 2(3)]. If the face is too large so that there is no context on some of the sides, the last pixel at the border is just repeated. This ensures that the face is always placed in the same location in the image. As the aspect ratio of the resulting image might differ, it is squeezed to 256 × 256 pixels. This forms the input for the deep convolutional neural network.

Our alignment procedure is rather unorthodox but initial experiments showed that this works much better than using facial landmark detection for alignment. Our method employs a rough and robust alignment that in very rare cases fails to upright align the face by rotation. In contrast, the facial landmark detectors can provide accurate landmarks in a majority of cases which is useful for accurate alignments (especially when the face is already upright). However, the landmark detectors fail much more often (∼5 % of the images in the LAP dataset) than our robust approach (∼1 % of the images in the LAP dataset) which leads to wrong alignments. It is difficult to determine these failure cases and to recover from them to ensure that the age estimation part is still successful. Therefore we decided against alignment with landmarks.

We refer to our face alignment procedure as "robust", since there are very few cases where it fails completely and gives always a rough alignment. Though our procedure does not provide very precise pixel-wise alignments, our CNN copes well with such level of precision.

3.2 Age Estimation

We employ a convolutional neural network (CNN) to predict the age of a person starting from a single input face image. This takes an aligned face with context as input and returns a prediction for the age. The CNN is trained on face images with known age.

3.2.1 *CNN Architecture*

Our method uses a CNN with the VGG-16 (Simonyan and Zisserman 2014) architecture [cf. Fig. 2(4)]. Our choice is motivated (1) by the deep but manageable architecture, (2) by the impressive results achieved using VGG-16 on the ImageNet challenge (Russakovsky et al. 2015), (3) by the fact that as in our case the VGG-16 architecture starts from an input image of medium resolution (256 × 256), (4) and that pre-trained models for classification are publicly available allowing warm starts for training.

The VGG-16 architecture is much deeper than previous architectures such as the AlexNet (Krizhevsky et al. 2012) with 16 layers in total, 13 convolutional and 3 fully connected layers. It can be characterized by its small convolutional filters of 3 × 3 pixels with a stride of 1. AlexNet in comparison employs much larger filters with a size of up to 11 × 11 at a stride of 4. Thereby each filter in VGG-16 captures simpler geometrical structures but in comparison allows more complex reasoning through its increased depth. For all our experiments we start with the convolutional neural network pre-trained on the ImageNet images, the same models used in Simonyan and Zisserman (2014). Unless otherwise noted, we fine-tune the CNN on the images from the newly introduced IMDB-WIKI dataset to adapt to face image contents and age estimation. Finally, we tune the network on the training part of each actual dataset on which we evaluate. The fine-tuning allows the CNN to pick up the particularities, the distribution, and the bias of each dataset and thus to maximize the performance.

3.3 Evaluation Protocol

For quantitative evaluation in our experiments we use two different measures.

MAE. For all experiments we report the mean absolute error (MAE) in years. This is the average of the absolute error between the predicted age and the ground truth age. MAE is the most used measure in the literature, a *de facto* standard for age estimation.

ϵ-error. The LAP challenge proposes the ϵ-error as a quantitative measure. ϵ-error applies for datasets where there is no ground truth age but instead a group of people guessing the ground truth. It takes into account the standard deviation σ of the age voted by the people who labelled the images. Thus if the labelled age for an image varies significantly among the votes, a wrong prediction is penalized less. By assuming that those votes are following a normal distribution with mean age μ and standard deviation σ the error is then measured as

$$\epsilon = 1 - e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \tag{1}$$

The final ϵ-error is the average over all images. Its value ranges from 0 (perfect predictions) to 1 (completely wrong).

3.4 Output Layer and Expected Value

The pre-trained CNN (with VGG-16 architecture) for the ImageNet classification task has an output layer of 1000 softmax-normalized neurons, one for each of the object classes. In contrast, age estimation is a regression and not a classification problem, as age is continuous rather than a set of discrete classes.

For regression we replace the last layer with only 1 output neuron and employ an Euclidean loss function. Unfortunately training a CNN directly for regression is relatively unstable as outliers cause a large error term. This results in very large gradients which makes it difficult for the network to converge and leads to unstable predictions.

Instead, we phrase the prediction problem as a classification problem where the age values are discretized into $|Y|$ ranges of age. Each age range Y_i covers a range of ages from $Y_i^{\min}$ to $Y_i^{\max}$ and votes for the mean of all training samples in this age range, y_i. In our experiments we consider: (a) uniform ranges where each age range covers the same number of years and (b) balanced ranges such that each age range covers approximately the same number of training samples and, thus, fit the data distribution. The number of age ranges depends on the training set size, i.e., each age range needs sufficiently many training samples and thus finer discretization requires more samples. In this way, we train our CNN for classification and at test time we compute the expected value over the softmax-normalized output probabilities of the $|Y|$ neurons

$$E(O) = \sum_{i=1}^{|Y|} y_i \cdot o_i, \tag{2}$$

where $O = \{1, 2, ..., |Y|\}$ is the $|Y|$-dimensional output layer and $o_i \in O$ is the softmax-normalized output probability of neuron i. Experimental results show that this formulation increases robustness during training and accuracy during testing. Additionally it allows some interpretation of the output probability distribution to estimate the confidence of the prediction, which is not possible when training directly for regression.

3.5 Implementation Details

Depending on the experiment, the CNN is trained for regression or classification. In the case of classification we report both the performance when testing for classification, i.e., the predicted age is the age of the neuron with the highest probability, and the expected value over the softmax normalized output probabilities.

When training the CNN for classification instead of regression, the age ranges are formed in two different ways: (a) uniform ranges such that each age range covers the same number of years and (b) balanced ranges where each age range covers approximately the same number of training samples.

For all experiments the CNN is initialized with the weights from training on ImageNet. This model is then further pretrained on the IMDB-WIKI dataset for classification with 101 output neurons and uniform age ranges. Finally the CNN is trained on the dataset to test on.

We split the training set into 90 % for learning the weights and 10 % for validation during the training phase. The training is terminated when then network begins to over-fit on the validation set. All experiments start with the pre-trained ImageNet weights from Simonyan and Zisserman (2014). For any fine-tuning the learning rate for all layers except the last layer is set to 0.0001. As we change the number of output neurons, the weights of the last layer are initialized randomly. To allow quick adjustment of those new weights, we set the learning rate for the output layer to 0.001. We train with a momentum of 0.9 and a weight decay of 0.0005. The learning rate is reduced every 10 passes through the entire data by a factor of 10.

The models are trained using the Caffe framework (Jia et al. 2014) on Nvidia Titan X GPUs. Training on the IMDB-WIKI and CACD datasets took several days whereas fine-tuning on the smaller datasets took only a couple of hours.

3.6 Parameters for Output Layer

Both Table 1 and Fig. 3 show how varying the number of output neurons and the prediction of ranges of age affects the performance. For all the settings we use LAP train data for training and report on the LAP validation data. Note that for the case where the settings are kept identical with the IMDB-WIKI pre-training which was done with 101 output neurons and uniform balancing, we additionally report performance for the case when the last layer is reinitialized when training on the LAP dataset. There seems to be a sweet spot, *i.e.* too many neurons result into too little training data per neuron and at the same time too few neurons lack a fine-grained ranges of the ages and thus make prediction less precise. Surprisingly, with 10 output neurons the performance is still very good despite the large distance in age between the neurons. Balanced ranges seems to perform slightly better than uniform ranges, especially when combined with few neurons.

For reference in Table 1 we report the performance when employing standard Support Vector Regression (SVR) with RBF kernel and ϵ-insensitive loss function on deep features extracted from the last pooling layer (conv5_3), last (fc7) and penultimate (fc6) fully connected layer of our deep architecture without and with pre-training on IMDB-WIKI dataset. As expected the specialized layers lead to better performance than the more generic pooling layer when the network is adapted to the age estimation task, otherwise the more generic pooling layer provides better features for SVR. With IMDB-WIKI pretraining, SVR on fc7 is slightly below the direct application of the network learned for apparent age regression.

4 Experiments

In this section we present the experimental results. We first introduce the datasets used. In the following we present both quantitative as well as qualitative results. We conclude the section with a discussion about the results.

4.1 Datasets

In this paper we use five different datasets for real (biological) and apparent age. Figure 4 depicts exemplar images for each dataset. Table 2 shows the size of each dataset and the corresponding splits for training and testing.

IMDB-WIKI. We introduce a new dataset for age estimation which we name IMDB-WIKI. To the best of our knowledge this is the largest publicly available dataset for age esti-

Table 1 Performance on validation set of ChaLearn LAP 2015 apparent age estimation challenge

Learning strategy	Number output neurons	w/o IMDB-WIKI pre-training ranges				w/ IMDB-WIKI pre-training ranges			
		Uniform		Balanced		Uniform		Balanced	
		MAE	ϵ-error	MAE	ϵ-error	MAE	ϵ-error	MAE	ϵ-error
SVR on conv5_3		8.472	0.647			4.570	0.411		
SVR on fc6		15.086	0.787			3.690	0.329		
SVR on fc7		12.083	0.720			3.670	0.321		
SVR on conv5_3†		7.150	0.560			4.020	0.356		
SVR on fc6†		9.695	0.663			3.406	0.297		
SVR on fc7†		9.069	0.664			3.323	0.288		
Regression	1	5.586	0.475			3.650	0.310		
Classification	5	6.953	0.563	6.275	0.501	5.944	0.529	4.394	0.369
Classification	10	6.404	0.511	6.352	0.516	4.243	0.388	3.912	0.337
Classification	25	6.474	0.521	6.507	0.516	3.563	0.309	3.676	0.322
Classification	50	6.424	0.510	7.044	0.555	3.463	0.298	3.517	0.306
Classification	101	7.083	0.548			3.640	0.310		
Classification*	101					3.521	0.305		
Expected value	5	6.306	0.535	5.589	0.464	5.226	0.481	3.955	0.329
Expected value	10	5.586	0.470	**5.369**	**0.456**	3.553	0.315	3.505	0.296
Expected value	25	**5.580**	**0.469**	5.522	0.468	3.306	0.289	3.353	0.290
Expected value	50	5.653	0.473	6.042	0.509	3.349	0.291	**3.318**	**0.289**
Expected value	101	5.965	0.493			3.444	0.299		
Expected value*	101					**3.252**	**0.282**		

Bold values indicate the highest performing setting

Varying number of output neurons [* last layer initialized with weights from IMDB-WIKI pre-training, † fine-tuned on LAP (expected value* 101 setup) before training SVR]. conv5_3 (100,352 dim) is the last convolutional layer. fc6 (4096 dim) and fc7 (4096 dim) are the penultimate and last fully connected layers, respectively

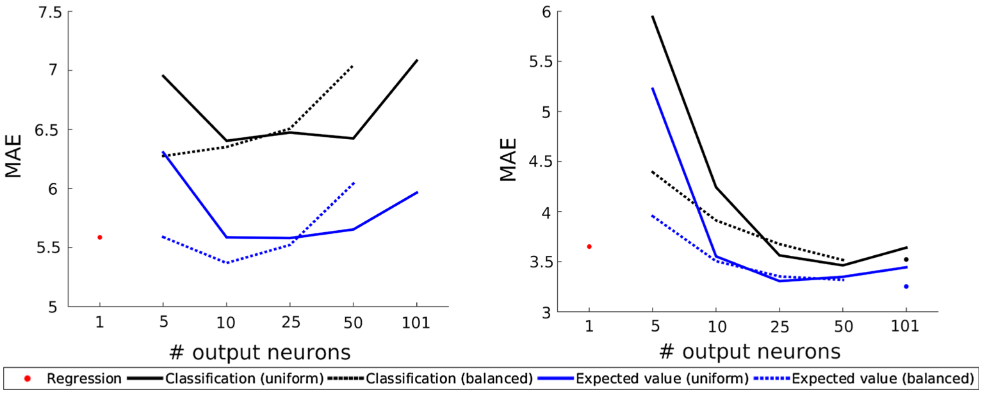

Fig. 3 Impact of the number of output neurons and the age ranges on the MAE performance

mation of people in the wild containing more than half a million labelled images. Most face datasets which are currently in use (1) are either small (i.e., tens of thousands of images) (2) contain only frontal aligned faces or (3) miss age labels. As the amount of training data strongly affects the accuracy of the trained models, especially those employing deep learning, there is a clear need for large datasets.

For our IMDB-WIKI dataset we crawl images of celebrities from IMDb[1] and Wikipedia[2]. For this, we use the list of the 100,000 most popular actors as listed on the IMDb website and automatically crawl from their profiles date of

[1] www.imdb.com.

[2] https://en.wikipedia.org/.

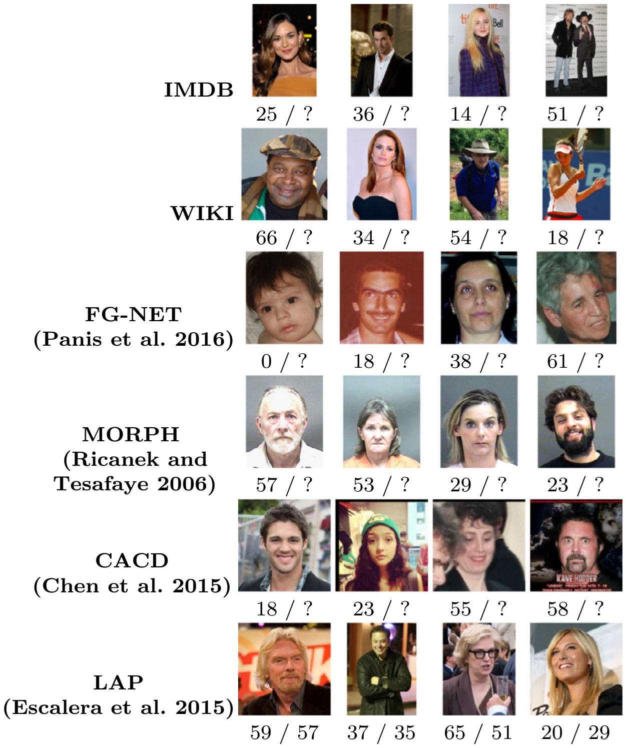

Fig. 4 Real/apparent age of exemplar images for each dataset

Table 2 The proposed method is evaluated on five datasets

Dataset	Number of images
IMDB-WIKI	**523,051**
IMDB	460,723
Wikipedia	62,328
IMDB-WIKI train	260,282
FG-NET (Panis et al. 2016)	1002
Train	990 (average)
Test	12 (average)
MORPH (Ricanek and Tesafaye 2006)	55134
Train	4380
Test	1095
CACD (Chen et al. 2015)	163,446
Train	145275 (1800 celebs)
Val	7600 (80 celebs)
Test	10571 (120 celebs)
LAP (Escalera et al. 2015)	4691
Train	2476
Val	1136
Test	1079

This table shows the number of images per dataset and the corresponding training and testing split

birth, name, gender and all the images related to that person. Additionally, we crawl all profile images from pages of people from Wikipedia with the same meta information.

For both data sources we remove the images that do not list the year when it was taken in the caption. Assuming that the images with single faces are likely to show the celebrity and that the year when it was taken and date of birth are correct, we are able to assign to each such image the biological (real) age. Especially the images from IMDb often contain several people. To ensure that we always use the face of the correct celebrity, we only use the photos where the second strongest face detection is below a threshold. Note that we can not vouch for the accuracy of the assigned age information. Besides incorrect captions, many images are stills from movies—movies that can have extended production times. Nonetheless for the majority of the images the age labels are correct. In total IMDB-WIKI dataset contains 523,051 face images: 460,723 face images from 20,284 celebrities from IMDb and 62,328 from Wikipedia. Only 5 % of the celebrities have more than 100 photos, and on average each celebrity has around 23 images.

We make the dataset publicly available at http://data.vision.ee.ethz.ch/cvl/rrothe/imdb-wiki/. We also release pretrained models. Note that this dataset can also be used for gender classification. We provide the entire image, the location of the face, its score and the score of the second most confident face detection.

FG-NET. The Face and Gesture Recognition Research Network (FG-NET) (Panis et al. 2016) aging database consists of 1002 color and greyscale images which were 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. For evaluation we adopt the setup of Yan et al. (2007), Guo et al. (2008), Chang et al. (2011) and Chen et al. (2013). They use leave-one person-out (LOPO) cross validation and report the average performance over the 82 splits.

MORPH. The Craniofacial Longitudinal Morphological Face Database (MORPH) (Ricanek and Tesafaye 2006) is the largest publicly available longitudinal face database containing more than fifty thousand mug shots. For our experiments we adopt the a setup often used in the literature Chang et al. (2011), Chen et al. (2013), Guo et al. (2008), Wang et al. (2015) and Rothe et al. (2016), where a subset of 5475 photos is used whose age ranges from 16 to 77. For evaluation, the dataset is randomly divided into 80 % for training and 20 % for testing. Some works Guo and Mu (2014), Huerta et al. (2014) use different splits. We still report them, however they are not directly comparable.

CACD. The Cross-Age Celebrity Dataset (CACD) (Chen et al. 2015) contains 163,446 images from 2000 celebrities collected from the Internet. The images are collected from search engines using celebrity name and year (2004–2013) as keywords. The age is estimated using the query year and the known date of birth. The dataset splits into three parts, 1800 celebrities are used for training, 80 for validation and 120 for testing. The validation and test sets are cleaned whereas the

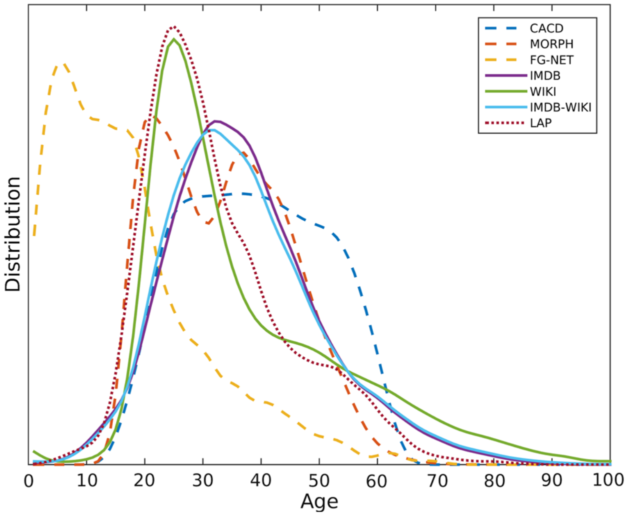

Fig. 5 Age distribution of people for all five datasets

training set is noisy. In our experiments we report results on the test set.

LAP. The ChaLearn LAP dataset (Escalera et al. 2015) contains 4699 images collectively age labeled using two web-based applications. According to the organizers of the LAP challenge this is the largest dataset on apparent age estimation. Each age label is the averaged opinion of at least ten independent users. Additionally, the standard deviation σ is also provided for each age label. The LAP dataset is split into 2476 images for training, 1136 images for validation and 1087 images for testing. The age distribution is very similar in all the three sets of the LAP dataset. Regarding the distribution of ages, the LAP datasets covers the 20–40 years interval the best. For the [0, 15] and [65, 100] age intervals it suffers from a small number of images per year.

As Fig. 5 depicts, the distribution of age between the datasets differs greatly. FG-NET contains images with by far the youngest people. MORPH has 2 peeks, one around early 20s and one at 40, suggesting that the images come from two data sources. CACD has few images from people below 20 or above 60 but is very balanced between those ages. The majority of face images on Wikipedia seem to show people slightly younger than on IMDb. In contrast Wikipedia has a long tail for old ages. The combined IMDB-WIKI dataset then follows a very similar distribution to the IMDb dataset as the ratio between IMDB and WIKI images is about 8 to 1. LAP and WIKI datasets have similar distributions.

4.2 Quantitative Results

In this section we report quantitative results of our proposed DEX method for biological and apparent age estimation. Additionally the results from the ChaLearn Looking at People (LAP) 2015 challenge (Escalera et al. 2015) on apparent age estimation are presented.

4.2.1 Apparent Age Estimation

We report performance of our DEX method for apparent age estimation. Table 1 summarizes the results when testing on the validation set of the LAP dataset.

The best performance for pre-training on the IMDB-WIKI dataset and taking the expected value reaches 0.282 ϵ-error (MAE 3.252) compared to 0.456 ϵ-error (MAE 5.369) when training directly on the LAP dataset. Training for regression instead performs worse at 0.475 ϵ-error (MAE 5.586) and 0.310 ϵ-error (MAE 3.650), respectively.

4.2.2 Looking at People (LAP) Challenge

Our DEX method is the winner of the ChaLearn Looking at People (LAP) 2015 challenge (Escalera et al. 2015) on apparent age estimation with 115 registered teams, significantly outperforming the human reference. The challenge had two phases: development and test.

Development Phase. In this phase the training and validation images of the LAP dataset are accessible. For the training set the apparent age labels are known, whereas for the validation set they are not released. The teams are able to submit their predictions for the validation images to a server to get the overall performance on those images. A public score board shows the latest performance of each team. As the previous score of each team is overwritten we build a crawler to check the score board every couple of seconds. Figure 6 shows the scores over the last month of the development phase. It can

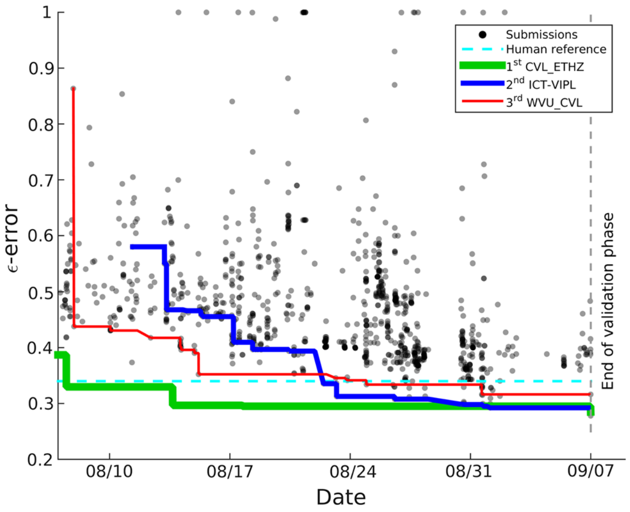

Fig. 6 One month validation entries for LAP challenge. For the top three teams we plot the best scores curves. **CVL_ETHZ is ours.**

Table 3 ChaLearn LAP 2015 (Escalera et al. 2015) final ranking on the test set

Rank	Team	ϵ error
1	CVL_ETHZ (ours) (Rothe et al. 2015)	0.264975
2	ICT-VIPL (Liu et al. 2015)	0.270685
3	WVU_CVL (Zhu et al. 2015)	0.294835
4	SEU-NJU (Yang et al. 2015)	0.305763
	Human reference	*0.34*
5	UMD	0.373352
6	Enjuto	0.374390
7	Sungbin Choi	0.420554
8	Lab219A	0.499181
9	Bogazici	0.524055
10	Notts CVLab	0.594248

115 registered participants. AgeSeer does not provide codes. The human reference is the one reported by the organizers

Table 4 Comparison results (MAE) for real (biological) age estimation

Method	MORPH 2	FG-NET
Human workers (Han et al. 2015)	6.30	4.70
DIF (Han et al. 2015)	3.80*	4.80
AGES (Geng et al. 2007)	8.83	6.77
MTWGP (Zhang and Yeung 2010)	6.28	4.83
CA-SVR (Chen et al. 2013)	5.88	4.67
SVR (Guo et al. 2008)	5.77	5.66
OHRank (Chang et al. 2011)	5.69	4.85
DLA (Wang et al. 2015)	4.77	4.26
Huerta et al. (2014)	4.25*	N/A
Guo and Mu (2011)	4.18*	N/A
Guo and Mu (2014)	3.92*	N/A
Luu et al. (2009)	N/A	4.37*
Luu et al. (2011)	N/A	4.12**
Yi et al. (2014)	3.63*	N/A
Rothe et al. (2016)	3.45	5.01
DEX	3.25	4.63
DEX (IMDB-WIKI)	**2.68**	**3.09**

Bold values indicate the highest performing setting

Our DEX method achieves the state-of-the-art performance on the MORPH and FG-NET standard datasets (* different split, ** landmark pre-training)

be clearly seen that as the end of the phase approaches the teams steadily improve their performance.

Test Phase. This is the final phase of the competition. The organizers of the challenge release the apparent age labels for the validation set and the images for the final test set. Now the algorithm is re-trained on both training and validation images to then predict the apparent age on the final test set. Our final results are obtained by training a full ensemble of 20 CNNs with 101 age bins on the training and validation images and then averaging the 20 predictions for each of the test images. Note that for all other results in this paper we report the performance of a single CNN.

Final Results. Figure 3 shows the final ranking of the competition. The best four methods achieve performance above the human reference of an ϵ-error of 0.34, as reported by the organizers. Our method is the only method within the top six methods which does not employ facial landmarks.

4.2.3 Real Age Estimation

In this section we present the performance of our proposed method for estimating the real (biological) age. In recent years, both the FG-NET and MORPH dataset have become the standard benchmark for the existing methods.

On the MORPH Dataset (Ricanek and Tesafaye 2006) our DEX method achieves a mean average error (MAE) of 3.25 when just fine-tuning the CNN on the training MORPH data. This improves over previous state-of-the-art reported in Rothe et al. (2016) by 0.2 years (see Table 4). Additional fine-tuning on our novel IMDB-WIKI dataset before fine-tuning on the MORPH dataset leads to a MAE of 2.68 years. To the best of our knowledge this is the first work reporting an error below 3 years on this common evaluation setup for MORPH, improving over the state-of-the-art by nearly 0.8 years.

On the FG-NET Dataset (Panis et al. 2016) without fine-tuning on IMDB-WIKI we achieve 4.63 years. Note that the larger error is due to the fact that FG-NET is a very small dataset (1000 images) and thus training a CNN on it is difficult. However, training on the IMDB-WIKI dataset before fine-tuning on FG-NET leads to a MAE of 3.09 years. This improves over DLA (Wang et al. 2015) by more than 1 year in average error. The results are summarized in Table 4.

On the CACD Dataset (Chen et al. 2015) we run additional experiments. The results are shown in Table 5. In comparison to MORPH and FG-NET the CACD dataset is much larger but not manually annotated. When training on the 145,275 training images we achieve a MAE of 4.785 years. When only training on the manually cleaned validation set with 7600 images the performance drops to a MAE of 6.521. This suggests that having a large training set with slightly imprecise labels results in better performance than a carefully annotated dataset of much smaller size.

4.2.4 Age Group Estimation

Besides real age estimation, we also evaluate our approach for predicting age groups. This is a somewhat simpler task as

Table 5 DEX results (MAE) on CACD dataset

Training on	CACD (Chen et al. 2015)
Train	4.785
Val	6.521

Table 6 Age group estimation results [mean accuracy(%) ± standard deviation] on Adience benchmark (Eidinger et al. 2014)

Method	Exact	1-off
DEX w/ IMDB-WIKI pretrain	**64.0 ± 4.2**	**96.6 ± 0.9**
DEX w/o IMDB-WIKI pretrain	55.6 ± 6.1	89.7 ± 1.8
Best from Levi and Hassner (2015)	50.7 ± 5.1	84.7 ± 2.2
Best from Eidinger et al. (2014)	45.1 ± 2.6	79.5 ± 1.4

Bold values indicate the highest performing setting

the goal is to predict whether a person's age falls within some range instead of predicting the precise biological age. We evaluate the performance on the Adience dataset (Eidinger et al. 2014) which consists of 26,580 images from 2284 subjects from Flickr. The dataset has 8 age groups (0–2, 4–6, 8–13, 15–20, 25–32, 38–43, 48–53, 60- years) and we report the results on the 5-fold cross validation proposed by the authors of the dataset. For this task we train our network for classification with 8 classes and report the exact accuracy (correct age group predicted) and 1-off accuracy (correct or adjacent age group predicted). We report results with and without pre-training on IMDB-WIKI. As it can be seen in Table 6 we achieve an accuracy of 64.0 % compared to the previous state-of-the-art of 50.7 %. When predicting the 1-off accuracy we achieve 96.6 %, i.e., our model is nearly always able to predict at least the adjacent age group.

4.3 Insight Experiments

In the following we present various insight experiments. These experiments are both quantitative and qualitative and give a deeper understanding of the method.

Visual Assessment. Figure 7 shows examples of face images in the wild (from LAP dataset) with good age estimation by our DEX with a single CNN. We observe that in these cases also the faces are aligned very well. Failure cases are also shown in Fig. 7. The failures are mostly caused by a failure in the detection stage (i.e., wrong or no face detected) or difficult conditions due to glasses, other forms of occlusions, or bad lightning.

Dataset Bias. In Fig. 8 we reveal the existence of a dataset bias. By testing the trained models on a dataset other than it was trained for (trained on LAP and tested on MORPH, and vice versa) we show the biases which come with each dataset. In Fig. 8a we show the distribution of predicted ages on LAP dataset for two models trained on MORPH dataset and LAP dataset, resp., and of the LAP dataset. The LAP model follows the distribution of the dataset and has the better MAE. In contrast the MORPH model exhibits a bi-modal distribution which is more similar to the MORPH dataset (cf. Fig. 8b). A similar behavior is observed when testing both

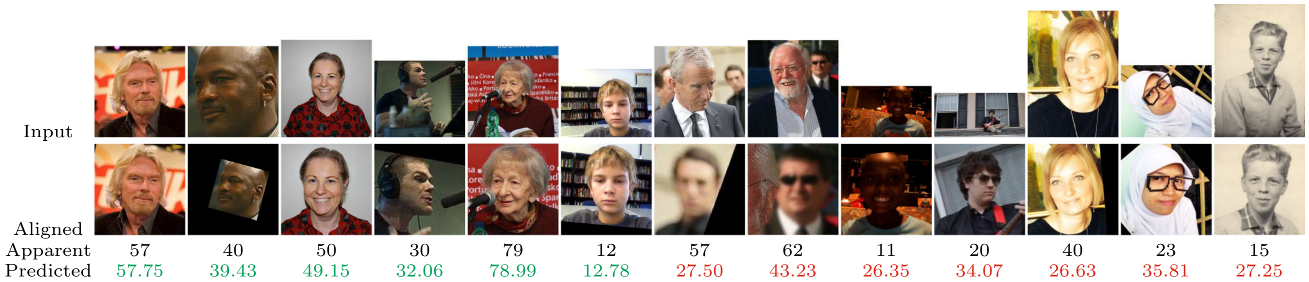

Fig. 7 Examples of face images with good and bad age estimation by DEX

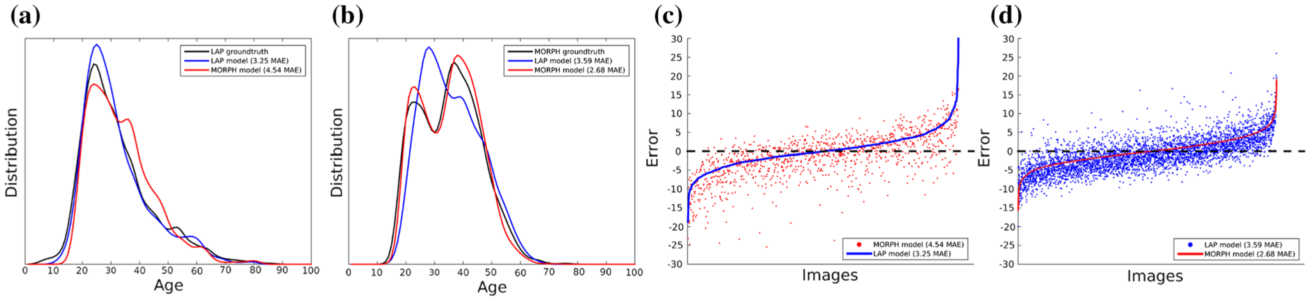

Fig. 8 Dataset Bias of LAP and MORPH

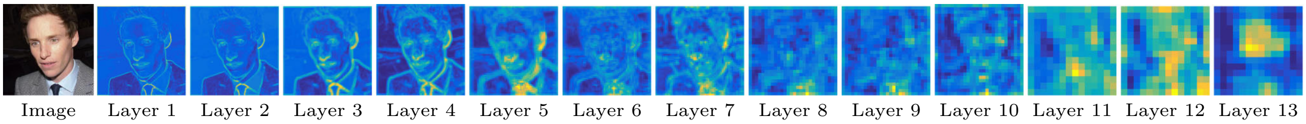

Fig. 9 Activation across CNN for a test image. The *color* indicates the maximum activation for any feature map for a particular layer

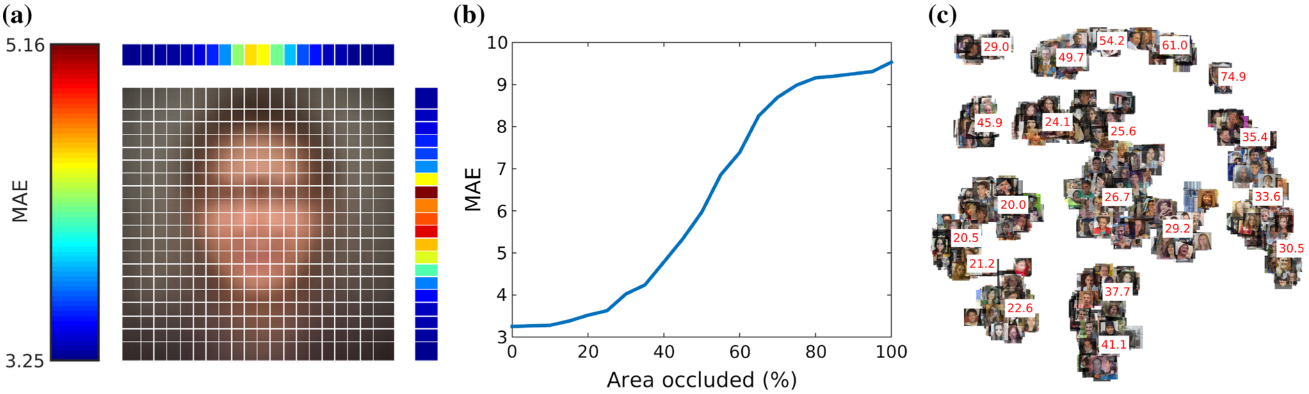

Fig. 10 In depth experiments and visualization of CNN models

models on the MORPH dataset (see Fig. 8b). In Fig. 8c, d the individual errors for each test image are plotted. The images are sorted according to the original dataset, i.e., in Fig. 8c when testing on LAP they are sorted according to the error of the model trained on LAP. On LAP dataset, in Fig. 8c, it can be seen that even though the error of the MORPH model is bigger overall, its predictions follow the curve of the LAP trained model and thereby both models similarly over- or underestimate the age of a person. A similar reasoning applies to the plots in Fig. 8d.

Important Face Regions. In order to determine which parts of a face image correlate and contribute the most to the overall age estimation accuracy we devise the following experiment. We systematically occlude a vertical or horizontal strip of the image by setting it to the mean image, as in Zeiler and Fergus (2014). Each of the 20 strips has a width of 10 % of the input face image. In Fig. 10a we report the MAE on the LAP dataset (validation images) for each of the vertical or horizontal strip occlusions. The results are intuitive, occlusions in the face area from the eyes to the chin and between ears affect the most the estimation accuracy. The results show that occluding the eyes with a horizontal strip increases the MAE the most, suggesting that the eyes are the most important indicator for age in the human face. The eyes are seconded by the horizontal strip region passing the upper lip and bottom of the nose. At the same time the horizontal strip occlusions lead to larger MAE than the vertical ones. A reason for this is that the face has horizontal symmetry and therefore for vertical occlusions except the strip that passes through the center of the face, there is always a corresponding symmetrical strip that is not occluded providing important information to the CNN model.

Robustness to Block Occlusions. To determine the robustness of our solution to occlusion we apply a block occlusion mask at random locations in the input face image. We report the MAE over the LAP dataset as the size of the occluded area is increased in Fig. 10b. When less than 20 % of the image is occluded the MAE is still low, i.e., the trained CNN is robust to those fairly small occlusions. Above 40 % occlusion the MAE performance rapidly deteriorates.

CNN Model Visualization. Figure 10c shows a t-Distributed Stochastic Neighbor Embedding (t-SNE) (van der Maaten and Hinton 2008) of the last fully connected layer of the model trained on the LAP dataset for the validation images. The feature vector of dimensionality 4096 is preprocessed using PCA to a dimensionality of 50. The visualization shows the test images for a perplexity of 10. We further cluster the embedded data into 20 clusters and report the average age of each cluster. The separation of images by age suggests that the features learned are discriminative for age prediction.

CNN Activations. Figure 9 shows the activation across our CNN trained on LAP for a test image using a color heatmap. The color indicates the maximum activation energy for any feature map for a particular layer. In the first couple of layers the face of the person can still be recognized and we can generally have the intuition that the neurons corresponding to the face region and the face edges activate the most. However, as we go deeper into the CNN the representation becomes more abstract and difficult to interpret.

4.4 Discussion

The proposed DEX method shows state-of-the-art results on MORPH and FG-NET for biological age and LAP for apparent age. Training the CNN for classification instead of regression not only improves performance but also stabilizes the training process. Without relying on landmarks and by robustly handling small occlusions the proposed method confirms its applicability for age estimation in the wild. Pretraining on the IMDB-WIKI dataset results in a large boost in performance suggesting that the lack of a larger dataset for age estimation was overdue for a long time.

In future work the training dataset could be further enlarged. Fine-tuning the face detector on the target dataset can reduce the failure rate of the face detection step. Using a very robust landmark detector can lead to better alignment. The recently introduced Residual Nets by He et al. (2015) with more than 150 layers show that an even deeper architecture than VGG-16 might help to improve performance if sufficient training data is available. Though at the same time the work suggests that there is an optimal depth, as the network with 1000 layers performs worse.

Ultimately the proposed DEX pipeline can be used for other prediction tasks of facial features including gender, ethnicity, attractiveness or attributes (i.e., does the person have glasses, a beard, blond hair).

5 Conclusions

In this paper we proposed a solution for real and apparent age estimation. Our Deep EXpectation (DEX) formulation builds upon a robust face alignment, the VGG-16 deep architecture and a classification followed by a expected value formulation of the age estimation problem. Another contribution is IMDB-WIKI, the largest public face images dataset to date with age and gender annotations. We validate our solution on standard benchmarks and achieve state-of-the-art results.

Acknowledgments This work was supported by the KTI-SUPSI (#2-69650-14) project and by an NVidia GPU grant.

References

Chang, K.Y., Chen, C.S., & Hung, Y.P. (2011). Ordinal hyperplanes ranker with cost sensitivities for age estimation. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*

Chen, B. C., Chen, C. S., & Hsu, W. H. (2015). Face recognition and retrieval using cross-age reference coding with cross-age celebrity dataset. *IEEE Transactions on Multimedia*, *17*(6), 804–815.

Chen, J.C., Patel, V.M., & Chellappa, R. (2016). Unconstrained face verification using deep CNN features. *IEEE Winter Conference on Applications of Computer Vision (WACV)*

Chen, K., Gong, S., Xiang, T., & Change Loy, C. (2013). Cumulative attribute space for age and crowd density estimation. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*

Ciregan, D., Meier, U., & Schmidhuber, J. (2012). Multi-column deep neural networks for image classification. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*

Cootes, T. F., Edwards, G. J., & Taylor, C. J. (2001). Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, *23*(6), 681–685.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273–297.

Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A. J., & Vapnik, V. (1997). Support vector regression machines. *Advances in Neural Information Processing Systems*, *9*, 155–161.

Eidinger, E., Enbar, R., & Hassner, T. (2014). Age and gender estimation of unfiltered faces. *IEEE Transactions on Information Forensics and Security*, *9*(12), 2170–2179.

Escalera, S., Fabian, J., Pardo, P., Baro, X., Gonzalez, J., Escalante, H.J., Misevic, D., Steiner, U., & Guyon, I. (2015). Chalearn looking at people 2015: apparent age and cultural event recognition datasets and results. *IEEE International Conference on Computer Vision (ICCV) Workshops*

Farkas, L. G., & Schendel, S. A. (1995). Anthropometry of the head and face. *American Journal of Orthodontics and Dentofacial Orthopedics*, *107*(1), 112–112.

Felzenszwalb, P. F., Girshick, R. B., McAllester, D., & Ramanan, D. (2010). Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, *32*(9), 1627–1645.

Fu, Y., & Huang, T. S. (2008). Human age estimation with regression on discriminative aging manifold. *IEEE Transactions on Multimedia*, *10*(4), 578–584.

Fu, Y., Guo, G., & Huang, T. S. (2010). Age synthesis and estimation via faces: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, *32*(11), 1955–1976.

Fukai, H., Takimoto, H., Mitsukura, Y., & Fukumi, M. (2007). Apparent age estimation system based on age perception. *SICE Annual Conference*

Gao, F., & Ai, H. (2009). Face age classification on consumer images with gabor feature and fuzzy lda method. *International Conference on Biometrics (ICB)*, pp 132–141

Geladi, P., & Kowalski, B. R. (1986). Partial least-squares regression: a tutorial. *Analytica Chimica Acta*, *185*, 1–17.

Geng, X., Zhou, Z. H., & Smith-Miles, K. (2007). Automatic age estimation based on facial aging patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, *29*(12), 2234–2240.

Girshick, R.B., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*

Guo, G. (2012). Human age estimation and sex classification. *Video Analytics for Business Intelligence*, pp 101–131

Guo, G., & Mu, G. (2011). Simultaneous dimensionality reduction and human age estimation via kernel partial least squares regression. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*

Guo, G., & Mu, G. (2014). A framework for joint estimation of age, gender and ethnicity on a large database. *Image and Vision Computing*, *32*(10), 761–770.

Guo, G., Fu, Y., Dyer, C. R., & Huang, T. S. (2008). Image-based human age estimation by manifold learning and locally adjusted robust regression. *IEEE Transactions on Image Processing*, *17*(7), 1178–1188.

Guo, G., Mu, G., Fu, Y., & Huang, T. (2009). Human age estimation using bio-inspired features. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*

Han, H., Otto, C., & Jain, A.K. (2013). Age estimation from face images: Human vs. machine performance. *International Conference on Biometrics (ICB)*

Han, H., Otto, C., Liu, X.,& Jain, A.K. (2015). Demographic estimation from face images: Human vs. machine performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)37*(6):1148–1161

Hardoon, D. R., Szedmak, S., & Shawe-Taylor, J. (2004). Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, *16*(12), 2639–2664.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep residual learning for image recognition. CoRR abs/1512.03385

Huerta, I., Fernández, C., & Prati, A. (2014). Facial age estimation through the fusion of texture and local appearance descriptors. *IEEE European conference on computer vision (ECCV)*

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., & Darrell, T. (2014). Caffe: Convolutional architecture for fast feature embedding. *International Conference on Multimedia*

Kazemi, V., & Sullivan, J. (2014). One millisecond face alignment with an ensemble of regression trees. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*

Krizhevsky, A., Sutskever, I., & Hinton, G.E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems (NIPS)*

Kwon, Y. H., & da Vitoria, Lobo N. (1999). Age classification from facial images. *Computer Vision and Image Understanding (CVIU)*, *74*(1), 1–21.

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

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324.

Levi, G., & Hassner, T. (2015). Age and gender classification using convolutional neural networks. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp 34–42

Liu, X., Li, S., Kan, M., Zhang, J., Wu, S., Liu, W., Han, H., Shan, S., & Chen, X. (2015). Agenet: Deeply learned regressor and classifier for robust apparent age estimation. *IEEE International Conference on Computer Vision (ICCV) Workshops*

Luu, K., Ricanek, K., Bui, T. D., & Suen, C. Y. (2009). Age estimation using active appearance models and support vector machine regression. *IEEE International Conference on Biometrics: Theory, Applications, and Systems* (ed).

Luu, K., Seshadri, K., Savvides, M., Bui, T.D., & Suen, C.Y. (2011). Contourlet appearance model for facial age estimation. *International Joint Conference on Biometrics (IJCB)*

Mathias, M., Benenson, R., Pedersoli, M., & Van Gool, L. (2014). Face detection without bells and whistles. *IEEE European Conference on Computer Vision (ECCV)*

Panis, G., Lanitis, A., Tsapatsoulis, N., & Cootes, T. F. (2016). Overview of research on facial ageing using the fg-net ageing database. *IET Biometrics*, *5*(2), 37–46.

Ramanathan, N., & Chellappa, R. (2006). Modeling age progression in young faces. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*

Ricanek, K., & Tesafaye, T. (2006). Morph: a longitudinal image database of normal adult age-progression. *Automatic Face and Gesture Recognition (FGR)*

Rothe, R., Timofte, R., & Van Gool, L. (2015). Dex: Deep expectation of apparent age from a single image. *IEEE International Conference on Computer Vision (ICCV) Workshops*

Rothe, R., Timofte, R., & Van Gool, L. (2016). Some like it hot-visual guidance for preference prediction. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., et al. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, *115*(3), 211–252.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. CoRR abs/1409.1556

Suo, J., Zhu, S. C., Shan, S., & Chen, X. (2010). A compositional and dynamic model for face aging. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, *32*(3), 385–401.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. *TIEEE Conference on Computer Vision and Pattern Recognition (CVPR)*

Taigman, Y., Yang, M., Ranzato, M., & Wolf, L. (2014). Deepface: Closing the gap to human-level performance in face verification. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*

Uricar, M., Timofte, R., Rothe, R., Matas, J., & Van Gool, L. (2016). Structured output svm prediction of apparent age, gender and smile from deep features. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*

van der Maaten, L., & Hinton, G. E. (2008). Visualizing high-dimensional data using t-SNE. *Journal of Machine Learning Research (JMLR)*, *9*, 2579–2605.

Wang, X., Guo, R., & Kambhamettu, C. (2015). Deeply-learned feature for age estimation. *IEEE Winter Conference on Applications of Computer Vision (WACV)*

Xu, Z., Chen, H., Zhu, S. C., & Luo, J. (2008). A hierarchical compositional model for face representation and sketching. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, *30*(6), 955–969.

Yan, S., Wang, H., Tang, X., & Huang, T.S. (2007). Learning auto-structured regressor from uncertain nonnegative labels. *IEEE International Conference on Computer Vision (ICCV)*

Yan, S., Zhou, X., Liu, M., Hasegawa-Johnson, M., & Huang, T.S. (2008). Regression from patch-kernel. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*

Yang, X., Gao, B.B., Xing, C., Huo, Z.W., Wei, X.S., Zhou, Y., Wu, J., & Geng, X. (2015). Deep label distribution learning for apparent age estimation. *IEEE International Conference on Computer Vision (ICCV) Workshops*

Yang, Z., & Ai, H. (2007). Demographic classification with local binary patterns. *International Conference on Biometrics (ICB)*

Yi, D., Lei, Z., & Li, S.Z. (2014). Age estimation by multi-scale convolutional network. *Asian Conference on Computer Vision (ACCV)*

Zeiler, M.D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. *IEEE European Conference on Computer Vision (ECCV)*

Zhang, Y., & Yeung, D.Y. (2010). Multi-task warped gaussian process for personalized age estimation. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*

Zhu, Y., Li, Y., Mu, G., & Guo, G. (2015). A study on apparent age estimation. *IEEE International Conference on Computer Vision (ICCV) Workshops*