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 pretrained 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.

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 [42, 2, 9, 25, 20]. Several public standard datasets [2, 42, 44] 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 [10] provided the largest dataset known to date of images with apparent age annotations, here 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: (i) access control, e.g., restricting the access of minors to sensible products like alcohol from vending machines or to events with adult content; (ii) 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; (iii) law enforcement, e.g., automatic scanning of video records for suspects with an age estimation can help during investigations; (iv) surveillance, e.g., automatic detection of unattended children at unusual hours and places; (v) perceived age, e.g., there is a large interest of the general public in the perceived age (c.f. 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 [42, 46, 10] the handling of faces in the wild and the accurate prediction of age remains a challenging problem.

Related work

While almost all literature prior the LAP 2015 challenge focuses on real (biological) age estimation from a face image, Han et al. [26] 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 [42, 20, 14, 26, 2, 9] for broader literature reviews. 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 [42] and MORPH [44] 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 [42] for an overview of research (365+ indexed papers) on facial aging with results reported on FGNET dataset.

A large number of face models has been proposed. We follow the taxonomy from [20] and mention: wrinkle models [33], anthropometric models [11, 33, 43], active appearance models (AAM) [6], aging pattern subspace [18], age manifold [13, 23, 21], biologically-inspired models (including biologically-inspired features (BIF) [24]), compositional and dynamic models [54, 49], local spatially fexible patches [56], and methods using fast Fourier transform (FFT) and genetic algorithm (GA) for feature extraction and selection [15], local binary patterns (LBP) [58], Gabor filters [16]. Recently, the convolutional neural networks (CNN) [35], biologically inspired, were successfully deployed for face modeling and age estimation [53, 36, 52].

The age estimation problem can be seen as a regression [13] or as a classification problem up to a quantization error [34, 18]. Among the most popular regression techniques we mention Support Vector Regression (SVR) [8], Partial Least Squares (PLS) [17], Canonical Correlation Analysis (CCA) [27], while for classification the traditional nearest neighbor (NN) and Support Vector Matchines (SVMs) [7].

In the next we select a couple of the representative (real) age estimation methods. Yan et al. [55] employ a regressor learning from uncertain labels, Guo et al. [23] learn a manifold and local SVRs, Han et al. [26] apply age group classification and within group regression

(DIF), Geng et al. [18] introduce AGES (AGing pattErn Subspace), Zhang et al. [61] propose a multi-task warped gaussian process (MTWGP), Chen et al. [4] derive CA-SVR with a cumulative attribute space and SVR, Chang et al. [1] rank hyperplanes for age estimation (OHRank), Huerta et al. [29] fuse texture and local appearance descriptors, Luu et al. [38] use AAM and SVR, while Guo and Mu [22] use CCA and PLS. Recently, Yi et al. [59] deployed a multiscale CNN, Wang et al. [53] used deep learned features (DLA) in a CNN way, while Rothe et al. [46] went deeper with CNNs and SVR for accurate real age estimation on top of the CNN learned features.

Proposed method

**Datasets**

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 estimation 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 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 pre-trained 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.

**Our approach**

Deep EXpectation (DEX) – to age estimation is motivated by the recent advances in fields such as image classification [5, 32, 47] or object detection [19] fueled by deep learning. From the deep learning literature we learn four key ideas that we apply to our solution: (i) 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 [28] suggests; (ii) 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; (iii) the alignment of the object in the input image impacts the overall performance; (iv) 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 [41] 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 [48]. We always start from pre-trained CNNs on the large ImageNet [47] dataset for image classification such that (i) to benefit from the representation learned to discriminate 1000 object categories in images, and (ii) 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. Fig. 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 [10] on apparent age estimation.

This work is an extended and detailed version of our previous LAP challenge report paper [45]. 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.

**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.

**CNN architecture**

Our method uses a CNN with the VGG-16 [48] architecture (cf. Fig. 2 (4)). Our choice is motivated (i) by the deep but manageable architecture, (ii) by the impressive results achieved using VGG-16 on the ImageNet challenge [47], (iii) by the fact that as in our case the VGG-16 architecture starts from an input image of medium resolution (256 × 256), (iv) and that pretrained 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 [32] with 16 layers in total, 13 convolutional and 3 fully connected layers. It can be characterized by its small convolutional filters of 3x3 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 [48]. 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.

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.

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.

References

1. Chang KY, Chen CS, Hung YP (2011) Ordinal hyperplanes ranker with cost sensitivities for age estimation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR)

2. Chen BC, Chen CS, Hsu WH (2015) Face recognition and retrieval using cross-age reference coding with crossage celebrity dataset. IEEE Transactions on Multimedia 17(6):804–815

3. Chen JC, Patel VM, Chellappa R (2016) Unconstrained face verification using deep CNN features. In: IEEE Winter Conference on Applications of Computer Vision (WACV)

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

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

6. Cootes TF, Edwards GJ, Taylor CJ (2001) Active appearance models. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 23(6):681–685

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

8. Drucker H, Burges CJC, Kaufman L, Smola AJ, Vapnik V (1997) Support vector regression machines. In: Advances in Neural Information Processing Systems 9, pp 155–161

9. 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

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

11. Farkas LG, Schendel SA (1995) Anthropometry of the head and face. American Journal of Orthodontics and Dentofacial Orthopedics 107(1):112–112

12. Felzenszwalb PF, Girshick RB, 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

13. Fu Y, Huang TS (2008) Human age estimation with regression on discriminative aging manifold. IEEE Transactions on Multimedia 10(4):578–584