Picking out the bad apples: unsupervised biometric data filtering for refined age estimation

Krešimir Bešenić, Jörgen Ahlberg and Igor S. Pandžić

The Visual Computer manuscript No.
(will be inserted by the editor)

Picking out the bad apples - Unsupervised biometric data filtering for refined age estimation

Krešimir Bešenić · Jörgen Ahlberg · Igor S. Pandžić

Received: date / Accepted: date

Abstract Introduction of large training datasets was essential for the recent advancement and success of deep learning methods. Due to the difficulties related to biometric data collection, facial image datasets with biometric trait labels are scarce and usually limited in terms of size and sample diversity. Web-scraping approaches for automatic data collection can produce large amounts of weakly labeled and noisy data. This work is focused on *picking out the bad apples* from web-scraped facial datasets by automatically removing erroneous samples that impair their usability. The unsupervised facial biometric data filtering method presented in this work greatly reduces label noise levels in web-scraped facial biometric data. Experiments on two large state-of-the-art web-scraped datasets demonstrate the effectiveness of the proposed method with respect to real and apparent age estimation based on five different age estimation methods. Furthermore, we apply the proposed method, together with a newly devised strategy for merging multiple datasets, to data collected from three major web-based data sources (i.e. IMDb, Wikipedia, Google), and derive the new Biometrically Filtered Famous Figure Dataset or B3FD. The proposed dataset, which is made publicly available, enables considerable performance gains for all tested age estimation methods and age estimation tasks. This work highlights the importance of training data quality compared to data quantity and selection of the estimation method.

Keywords Filtering · Biometric · Unsupervised · Web scraping · Age estimation · Dataset design

Krešimir Bešenić
University of Zagreb, Faculty of Electrical Engineering and Computing, Unska 3, 10000 Zagreb, Croatia
E-mail: kresimir.besenic@fer.hr
ORCID: 0000-0002-5861-7076

Jörgen Ahlberg
Linköping University, Computer Vision Laboratory, 58183 Linköping, Sweden
E-mail: jorgen.ahlberg@liu.se

Igor S. Pandžić
University of Zagreb, Faculty of Electrical Engineering and Computing, Unska 3, 10000 Zagreb, Croatia
E-mail: igor.pandzic@fer.hr

1 Introduction

In recent years, algorithms based on deep learning became a prominent technique for solving complex computer vision tasks. Advancements in training algorithms and model architectures along with large amounts of available data and computing infrastructure enabled researchers to design methods that surpassed human performance on difficult tasks such as image classification [20] and face recognition [42]. The main remaining barrier for solving many similar tasks is the lack of sufficient amounts of labeled data. While techniques like transfer learning are frequently being utilized to mitigate this problem and achieve state-of-the-art results, training with small numbers of task-specific samples can result with domain overfitting, questionable generalization capabilities, and unsatisfying performance in unconstrained environments.

As biometric data collection becomes an increasingly sensitive issue, the research community struggles with collection of large amounts of reliable data for biometric tasks such as gender, age, and ethnicity estimation. For more than a decade, image-based face analysis research relied on small manually collected datasets, ranging from 1,000 to 50,000 samples. More recently, several research groups successfully utilized automatic web-scraping methods to collect large amounts of noisy data and improve the state-of-the-art facial analysis algorithms [25, 23, 41, 33]. Although a low amount of noise in the training data is not considered to be a problem for modern deep learning algorithms and can, in some cases,

even help to reduce overfitting problems, large amounts of noise can reduce the smoothness of the cost function hyperplane, lower the convergence rate, and impair the final performance.

The goal of our extended work from [3] is to automatically reduce the level of label noise in web-scraped facial datasets by filtering out the wrongly labeled or otherwise faulty samples in order to improve the resulting face analysis algorithms. We focus on the age estimation task, as it is one of the most difficult face analysis problems [13]. The contributions of our work are summarized as follows:

- We present an efficient unsupervised method for biometric data filtering that can significantly reduce label noise in facial image datasets. To the best of our knowledge, this is the first completely automatic and parameter-free method for facial dataset filtering that does not require supervised training of dataset-specific systems but utilizes only general purpose, off-the-shelf algorithms.
- We apply the proposed filtering method to two state-of-the-art web-scraped datasets and demonstrate its benefits to 5 different age estimation methods and with respect to generalization capabilities in unconstrained conditions.
- We propose a biometric filtering strategy to reinforce and refine the merging process of multiple facial datasets and derive the new Biometrically Filtered Famous Figure Dataset (B3FD). We demonstrate B3FD's superiority over existing state-of-the-art age estimation datasets with respect to both real and apparent age estimation and make the dataset publicly available.
- We highlight the importance of training data quality compared to the training data quantity and demonstrate that the proposed refinements of the training data result in larger margin of improvement than utilization of more advanced age estimation methods.

The rest of the paper is organized as follows. Section 2 reviews important manually collected datasets for age estimation, most relevant automatic web-scraping and dataset filtering methods, as well as recent age estimation methods. Further, Section 3 describes the proposed method for unsupervised biometric data filtering and provides experimental validation of the method's effectiveness. Section 4 describes the design strategy of the new famous figure age estimation dataset and provides comparison with the state-of-the-art. Section 5 briefly concludes the findings of this work.

2 Related work

This section presents a review of the most relevant work on manually collected facial age datasets, automatic web-scraping methods for biometric data collection, large-scale facial dataset filtering, as well as relevant recent age estimation methods.

2.1 Manually collected datasets

Early research on automatic facial biometric trait estimation was conducted on small manually collected datasets often having less than 100 samples [27, 18]. Whereas early research was mostly focused on real age estimation, in recent years apparent age estimation also started to gain traction. Real age estimation is the task of estimating the subject's biological age, while apparent age estimation refers to the estimation of age as humans perceive it, based on the subject's physical appearance. Early work was also mostly focused on age estimation in controlled environment, whereas in recent years focus shifted towards more difficult task of age estimation in unconstrained in-the-wild conditions, meaning that environmental conditions such as background, lighting, camera position, and occlusions are not controlled.

One of the first publicly available datasets for facial age estimation was The Face and Gesture Recognition Research Network (FG-NET) dataset. It is a cross-age dataset, consisting of 1,002 images from 82 subjects. To collect the dataset, subjects were asked to scan their personal photos from childhood and adulthood. Although small in size, this manually collected dataset was a difficult challenge and a stepping-stone for early age estimation research, as reviewed in [37].

Another important milestone for facial age estimation research was the introduction of The Craniofacial Longitudinal Morphological Face Database [40]. MORPH is a mugshot dataset consisting of more than 55,000 images, taken in a correctional facility over a period of 4 years. It provides annotations for age, gender, and race. Even though the images were collected in a highly controlled environment and the dataset has an unbalanced distribution of samples across gender (85% male), age (80% between 20 and 50 years, no children and old people), and race (77% African American), it increased the number of publicly available samples for age estimation research by a factor of 55 and made a great impact in the field.

The Appa-Real dataset [1] is a more recent manually collected dataset, based on data from CLAP 2015 [12] and CLAP 2016 [13] challenges. The authors designed a data collection and labeling web application and utilized the Facebook API and the Amazon Mechanical Turk platform to get diversified data. The Appa-Real dataset consists of 7,591 unconstrained in-the-wild samples with age range from 0 to 95. Whereas the ChaLearn LAP datasets provided only apparent age labels, the Appa-Real dataset is extended with real age labels, making it the only dataset that provides highly reliable labels for both real and apparent age. The authors utilized their platform to collect almost 300,000 apparent age votes, which amounts to approximately 38 votes per image.

The AgeDB dataset is the most recent and largest manually collected in-the-wild dataset with real age labels. It consists of 16,488 images with real age, gender, and iden-

Table 1 Frequently used publicly available facial image datasets with age labels.

	General			Labels		Collection		
Dataset	Images	Subjects	Im/Sub	Type	Range	Images	Labels	Environment
FG-NET [37]	1,002	82	12.22	real age	0-69	manual	manual	semi-controlled
CLAP 2015 [12]	4,699	-	-	apparent age	3-85	manual	manual	uncontrolled
CLAP 2016 [13]	7,592	-	-	apparent age	0-95	manual	manual	uncontrolled
Appa-Real [1]	7,591	≈7,000	≈1.08	real age, apparent age	0-95	manual	manual	uncontrolled
MORPH [40]	55,134	13,618	4.05	real age	16-77	manual	manual	controlled
AgeDB [31]	16,488	568	29,03	real age	1-101	manual	manual	uncontrolled
GROUPS [15]	5,080	28,231	1.00	7 apparent age group	0-66+	web scraping	manual	semi-controlled
Adience [11]	26,580	2,284	11.64	8 apparent age groups	0-60+	web scraping	manual	uncontrolled
MegaAge [47]	41,941	-	-	apparent age	0-70	web scraping	manual	uncontrolled
CACD [7]	163,446	2,000	81.72	real age	14-62	web scraping	web scraping	uncontrolled
WIKI [41]	62,328	62,328	1.00	real age	0-100+	web scraping	web scraping	uncontrolled
IMDB [41]	460,723	20,284	22.71	real age	0-100+	web scraping	web scraping	uncontrolled
IMDB-WIKI [41][1]	523,051	82,612	6.33	real age	0-100+	web scraping	web scraping	uncontrolled

[1] Separate entries for the IMDB and WIKI subsets were added to highlight the differences in their properties.

tity labels. The dataset was collected by manually searching for images of famous people via the Google Image Search[1] platform and keeping only images for which the exact age of the depicted subject was explicitly mentioned in the metadata.

Small amounts of samples, lack of sample diversity, and biased sample distributions are some of the recurrent obstacles for the development of systems with good generalization capabilities and robustness to in-the-wild conditions. The next section reviews work on automatic web-based collection of large amounts of diverse facial biometric data, while Table 1 summarizes the basic properties of the described public datasets.

2.2 Automatic web scraping

A very simple, yet effective, method for automatic collection of a large web-scraped gender dataset was presented in [25]. By querying search engines with a list of gender-specific names, the authors collected 4 million weakly labeled samples and demonstrated the importance of large-scale datasets for in-the-wild gender estimation. The dataset was unfortunately not made publicly available.

To avoid the need for large-scale public datasets with exact age annotations, a method for web-based collection of samples with age difference labels was proposed in [23]. To build their dataset, the authors used Flickr[2] to crawl large amounts of images by the query names from the LFW dataset [24] along with descriptions containing dates of image acquisition. Although they did not collect the actual age information, pre-training their network for age-difference estimation improved their final real age estimation results.

[1] https://images.google.com/

[2] www.flickr.com

Automatic image scraping from the Flickr platform was also used by authors of the GROUPS [15] and Adience [11] datasets. Both datasets provide age group and gender labels, while the Adience dataset also provides identity labels. The GROUPS dataset was collected for analysis of contextual features of groups of people. The Adience dataset was collected to provide a challenging in-the-wild dataset for age group and gender classification. While images from both datasets were collected via automatic web scraping, the labels were manually estimated by the authors. The age labels are estimates of the approximate age group (e.g. 8-13 or 25-32), meaning that they represent apparent age groups rather than the real age.

The MegaAge dataset [47] is another example of a dataset consisting of images automatically scraped from the Flickr platform. The dataset is made of samples randomly selected from the large in-the-wild MegaFace face recognition dataset [26] and the YFCC100M dataset [34]. Both MegaFace and YFCC100M are based on images from the Flickr platform. The dataset consists of 41,941 images with age posterior distribution labels. The labels were obtained by manually comparing images to multiple annotated images from the FG-NET dataset. Annotators were asked to estimate if the person in the image appears younger or older than the person in the reference image. The posterior reflects the apparent age with an estimated uncertainty range.

The Cross-Age Celebrity Dataset (CACD) was the first public large-scale web-scraped facial dataset with real age annotations, initially introduced for cross-age face recognition in [7]. The goal was to create a large-scale dataset with good sample variety with respect to the subject's age. The list of subjects was created based on two main criteria: (1) the subjects on the list should have varying ages, and (2) they must have large numbers of images available on the Internet.

To satisfy the latter term, they decided to collect images of celebrities. To deal with the former term, they decided to collect images of celebrities born in a 40-year period. They used the popular online movie database (IMDb[3]) to find the 50 most popular celebrities for each birth year from 1951 to 1990, resulting in a list containing 2,000 subjects. After the list was created, they used Google Image Search to collect images. In order to collect samples across different ages, they used combinations of celebrity names and years as search phrases. After removing duplicate images with a simple duplicate-detection algorithm and dismissing images without detected faces, they ended up with more than 160,000 images. The years from the search phrases were used in combination with the birth years collected from the IMDb to automatically produce the age labels. Although the authors admit this simple approach produces a lot of noisy labels, the collected dataset was far superior to the existing ones in terms of size and sample variety.

Another similar famous people image-crawling-based approach was presented in [41]. The authors managed to collect more than 500,000 images with age and gender annotations from IMDb and Wikipedia[4]. The dataset was named IMDB-WIKI and is the single largest public dataset for age and gender estimation to date. The authors used the IMDb to obtain a list of 100,000 most popular actors and crawled images directly from their IMDb profiles, along with gender and birth date information. Additionally, they collected Wikipedia profile pictures with the same meta-data. After removing all the images that do not list the year in which they were taken, they used the listed years and the date of birth from the subject's profile to automatically obtain age labels. In case of images with multiple face detections, they decided to keep only the images where all secondary face detection confidences were under a certain threshold. Similar to [7], the authors note that they cannot vouch for the accuracy of the assigned age and gender information.

2.3 Facial dataset filtering

Web-scraped datasets such as CACD and IMDB-WIKI are shown to be superior to the manually collected datasets in terms of size and sample variety, but their overall quality is undermined by the high amounts of label noise. This section reviews efforts made toward cleaning noisy web-scraped facial datasets.

An early example of an automatic facial dataset filtering method was presented in [33]. In an attempt of designing a robust and universal age estimator, the authors used image search engines and a set of age-related queries to collect a large facial dataset with weak age labels. In order to reduce the label noise levels, they designed a simple two-step filtering approach. In the first step, they used parallel face detection based on multiple state-of-the-art face detectors. To remove non-facial images and dismiss misaligned detections, they only retained samples with multiple detections overlapping more than 90%. To further reduce the number of false positive detections and to reduce the number of faces not correctly corresponding to the search-query age, they applied the Principal Component Analysis (PCA) to all images collected for a certain age and dismissed images with large reconstruction errors.

The age-specific PCA filtering step was intended to remove age-category outliers based on their apparent age, but the largest reconstruction errors were caused by face occlusions and non-frontal head poses, thus removing samples crucial for training a robust age estimator. Furthermore, due to the strict criterion of multiple face detection overlap, an additional large number of valuable difficult samples was discarded.

Even though the benefits of pre-training on the large and noisy IMDB-WIKI dataset were clearly demonstrated in [41], a cleaned version could further improve their age estimation results. In order to create a cleaned version of the dataset, combined automatic and manual processing steps were used in [2]. In the first step, all the images with multiple face detections were removed to increase the probability of the detected face corresponding to the provided age label. In the second step, a subset of the remaining multi-face images was manually filtered via a crowdsourcing annotation process.

The authors state that the first step ensures the correctness of the age labels, but both false positive and false negative detections induce considerable amounts of label noise even in the single-detection images. In the manual step, the annotators were asked to pair the provided annotation with one of the faces in the image. A study on human performance showed that the average annotator estimates age with high mean absolute error (MAE) of 4.7 - 7.2 years [19], indicating that even this seemingly trivial step can produce additional noisy outcomes.

Compared to the limited work presented on age and gender data filtering, several more advanced approaches for facial dataset filtering were proposed in the facial recognition field, as it has become one of the most data-hungry image analysis fields in general.

A data-driven approach for cleaning large face datasets was presented by authors of the FaceScrub dataset [32]. To identify the faces to be removed from their dataset, they exploited the observations that the same person should appear at most once per image, have the same gender, and look similar. The task of outlier detection was formulated as a query-specific quadratic programming (QP) problem based on a combination of terms related to those observations. Assum-

[3] www.imdb.com
[4] https://en.wikipedia.org/

ing that falsely detected faces form only a small portion of the detected set, they were able to train a one-class SVM and use the output of its decision function as a score for a false positive term. To enforce a gender term, they trained a two-class linear SVM for gender classification with query-based gender labels. Similar to the false detection term, the outputs of its decision function were used as gender scores. A similarity term was encouraged by graph regularization based on the normalized graph Laplacian, and an additional prior term was used to encode the assumption that most faces are correct.

By manually annotating a part of their dataset, the authors assessed their algorithm and demonstrated that their QP formulation outperforms the naive approach where the classifiers were used separately. However, the discussed benefit of manual workload reduction was somewhat impaired by the need for the dataset-specific classifier trainings.

The latest large-scale web-scraped facial recognition dataset, named VGGFace2 [6], adopted and improved a multi-step semi-automatic approach from the original VGGFace paper [38]. To achieve their goal of a 96% pure dataset, their efforts included more than 3 months of manual annotations. The majority of that time was spent on the initial name list filtering. The annotation team reduced the initial list from 500,000 to only 9,244 names by dismissing all the subjects for whom the top 100 Google Image Search results were not at least 90% pure. After applying a relatively strict face detection step, a set of 1-vs-rest classifiers was trained to discriminate between the 9,244 subjects. The threshold was selected by manually checking results for 500 subjects and all samples with a score below the selected threshold were dismissed. The next step, designed to remove near-duplicate images, used VLAD descriptor clustering and retained only one image per cluster. To detect overlapping subjects (names referring to the same person), an additional classifier was trained to generate a confusion matrix and remove classes mostly confused with others. The final, partially manual step consisted of iterative retraining of the 1-vs-rest classifiers with an annotator team manually filtering only part of the samples based on the classification scores.

To reach their target in terms of data purity, the authors of the VGGFace2 trained several versions of more than 9,000 1-vs-rest classifiers, trained an additional classifier for overlap detection, performed manual threshold search and substantial amounts of manual filtering. This impressive data filtering effort resulted in a state-of-the-art face recognition dataset.

2.4 Age estimation

Although the age estimation based on facial images is a prominent research field with more than two decades of publishing history, in this section we focus on recent CNN-based methods that brought us closer to closing the gap to human performance. Most existing age estimation methods fall into one of the four main categories; regression, classification, ranking, and label distribution learning methods.

The regression-based method introduced in [45] used a multi-scale CNN to extract features from multiple facial regions and utilized a rudimental square loss function. A CNN-based age and gender classification was introduced in [29]. A framework for deep learned aging pattern extraction (DLA) based on feature maps obtained from different CNN layers was proposed in [43] and applied to age regression and classification. CNN-based regression and classification approaches were also evaluated in [41] and outperformed by introducing the Deep Expectation formulation that calculates expected age value as a weighted average of the Softmax outputs. Whereas only classification-based loss was used in [41] to train the model, Mean-Variance loss formulation introduced in [36] added regression-based and distribution-oriented components to the loss function to produce state-of-the-art results.

Proponents of ranking-based methods [35, 8] argue that the basic classification approaches disregard the ordinality of the age estimation task and propose to formulate ordinal learning as a series of binary classification sub-problems. A Ranking-CNN framework introduced in [8] estimates age as an aggregation of binary results of multiple CNN networks trained for ordinal age estimation. Multi-output CNN for ordinal regression proposed in [35] collectively solves the classification sub-problems in an end-to-end learning fashion.

Label distribution learning methods represent age labels as distributions and utilize the Kullback-Leibler Divergence Loss to train the model. In [44] and [16] the ground truth age label distribution is approximated by a Gaussian function. The Deep Label Distribution Learning (DLDL) approach from [16] was extended by introduction of an expectation regression module to alleviate inconsistency between the training objectives and evaluation metrics in [17]. In [30], the authors argue that fixed-form age distribution is not suitable to represent complicated facial image domains. To mitigate this problem, the label distribution is constructed by learning the cross-age correlation between context-neighboring samples in [22], while [30] proposed a label distribution refinery to adaptively learn the continuous age distribution.

3 Unsupervised biometric data filtering

To design an efficient filtering method, overview of which is given in Figure 1, we analyzed the common sources of label noise in the current state-of-the-art biometric facial datasets.

Due to the nature of the commonly used web-scraping approaches described in Section 2.2, there are two main sources of label noise. The first problem is the unreliability

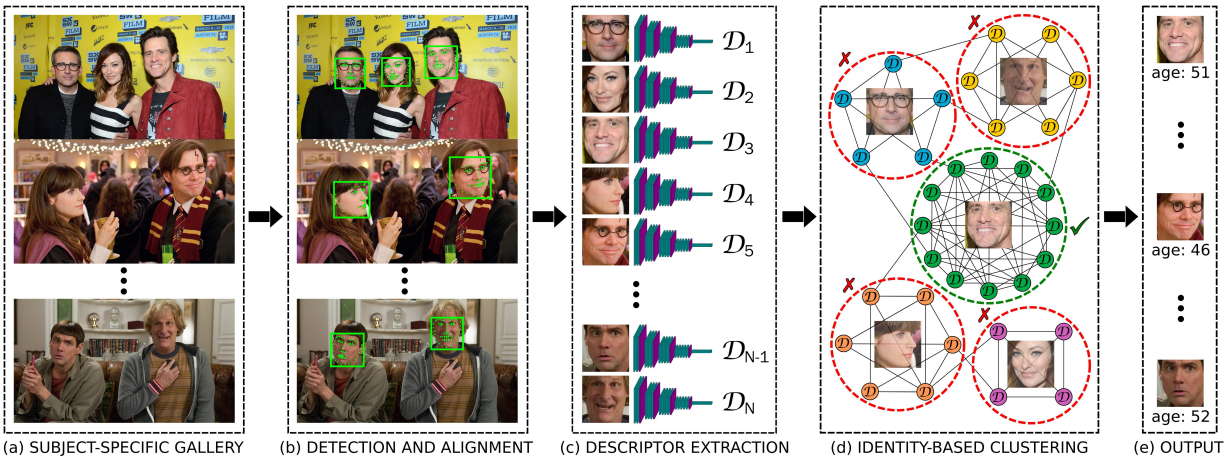

Fig. 1 The proposed unsupervised biometric filtering pipeline for the IMDB subject Jim Carrey. (a) Samples are grouped into subject-specific galleries based on the provided identity meta-data. Every image in the subject-specific gallery contains gallery's owner (e.g. Jim Carrey), but many also contain other subjects that can be mismatched with the gallery owner's biometric labels (e.g. age and gender). (b) Face detection and alignment of faces in the image gallery. Most of the detected faces belong to the gallery's owner, while others are sources of label noise. (c) CNN-based face recognition descriptor extraction from all detected faces. (d) Graph-based clustering of the extracted facial descriptors. Clusters are formed based on identity matching. The gallery owner's cluster is the largest because its face is most frequently appearing in the image gallery. (e) Samples from the largest cluster are retained, while samples from all other clusters are discarded as noise.

of the automatic age annotation process itself. Although the dates of birth are mostly correct, the year of image acquisition can be inaccurate or misleading. As mentioned in [41], a large number of images are actually movie screenshots annotated with the year of the movie release, and some movies have production time spanning over several years. This problem usually causes only minor age annotation errors.

A much more serious issue, causing large discrepancies for age and other biometrics labels, are mismatched identities. In case of multi-person images, face detector failures or bad image search results, collected meta-data can be paired with a face detection of a wrong subject. For example, if the image is a photo of a female actress and her son, and the son's face gets detected as the primary face, the image will have wrongly assigned gender and high age annotation error (i.e. up to several decades).

Fig. 2 Web-scraping noise in IMDB-WIKI and CACD data for subject Gal Gadot. For each dataset, the top row shows valid samples while the bottom row shows image samples that have been wrongly paired with Gal Gadot's meta-data.

Figure 2 shows examples of correctly paired and mismatched images for one subject appearing in both CACD and IMDB-WIKI datasets. Compared to typically constrained manually collected data, the top-row (i.e. correct) samples exhibit superior variation with respect to many important aspects such as head pose, facial expression, lighting, and background. On the other hand, the bottom-row (i.e. mismatched) samples greatly impair the overall quality and usability of the data.

In order to reduce the number of labeled samples with erroneously matched facial images and to mitigate this most detrimental source of label noise, we propose a filtering method described in the next section.

3.1 Proposed filtering method

The main idea of the unsupervised filtering method is to automatically group samples from subject-specific image galleries (i.e. images with the same identity meta-data) into clusters of samples with matching biometric descriptors, and to keep only the samples from the largest cluster, while all other samples get discarded. This way the samples of the most frequently appearing subject in the image gallery can be retained, while samples erroneously matched with that subject's meta-data are removed. This fundamental idea is expanded as follows.

Prevalent approach for automatic grouping of samples based on a certain property is utilization of clustering algorithms. There are a number of clustering algorithms that can perform the required grouping efficiently and without supervision, but regardless of the type of the clustering algorithm, the clustering performance will greatly depend on the way the sample's grouping property is represented numerically.

For a good performance, numerical representations should be compact and highly descriptive of the property of interest (e.g. the subject's identity). In case of facial image data, a favorable option for the task is the facial recognition algorithms. Facial recognition algorithms are specifically designed to project high-dimensional facial image data to highly discriminative low-dimensional biometric feature vectors (i.e. face descriptors) that encode subject's identities. However, the proposed method is not restricted to work on descriptors obtained by facial recognition algorithms as other biometric or image descriptors could be utilized.

To reduce the undesired effects of feature extraction from misaligned and inconsistent detections, we propose to employ a two-step detection procedure consisting of regular object (face) detection followed by key-point detection that allows precise calculation of bounding box position and scale, as well as in-plane image alignment.

For the approach to be completely parameterless and unsupervised, the descriptor grouping should be done with a clustering method that is capable of discovering the number of underlying groups (i.e. identities) automatically. For this purpose, a number of clustering algorithms, such as Chinese Whispers Clustering [4], Affinity Propagation Clustering [14] or Mean Shift Clustering [10], can be used.

The aforementioned outlines the basic concepts of the proposed unsupervised biometric filtering method, while the following section gives the implementation details of the designed filtering pipeline.

3.1.1 Implementation details

Based on the previously described concepts, we implemented a filtering system consisting of five main consecutive steps. A graphical summary of the proposed filtering pipeline is given in Figure 1, depicting the five steps for one specific subject. The implementation details of the presented steps are as follows.

Subject-specific gallery. Firstly, according to the provided identity meta-data (e.g. subject name or ID), the datasets are reorganized into series of subject-specific galleries. Each gallery contains all the images collected for one specific subject. While the gallery's owner should be present in all the images in the gallery, many other subjects may appear as well, but less frequently.

Detection and alignment. The second step amounts to detection and alignment of all faces in the subject-specific image galleries. Although the bounding box information is usually provided, given bounding boxes lack consistency with respect to bounding box scale and positioning. To ensure more consistent inputs to the following descriptor extraction step, we first utilize a face detection algorithm based on dlib's[5] CNN face detector to re-detect faces, and then use a facial alignment algorithm robust to bounding-box imprecisions [5] to precisely determine bounding box position and scale based on the detected facial landmark points. Moreover, the facial landmark points are used to perform in-plane image alignment. The bounding box information provided by the datasets' authors is used only in the rare cases of face detection failure, and even then it is corrected by the face alignment step.

Descriptor extraction. The third step is the extraction of face descriptors for all the faces detected in the previous step. We utilized the dlib's powerful facial recognition model based on ResNet architecture [21] to extract compact 512-dimensional identity descriptors. By calculating distances between the extracted face descriptors, the probability of two descriptors representing the same subject (i.e. identity) can be efficiently estimated, and by using dlib's default descriptor similarity threshold, a reliable identity matching can be achieved. Well performing face recognition systems offer reliable identity matching regardless of facial expression, orientation, and even occlusion to some extent, as well as environmental factors such as lighting and background.

Identity-based clustering. The fourth step is designed to identify samples belonging to the gallery's owner and to separate them from the other faces that are coincidentally present in the gallery and wrongly paired with the gallery's owner meta-data. This step is based on applying a clustering algorithm to facial descriptors extracted from all the detected faces. This enables us to group facial images based on their biometric features (i.e. identity). Since the number of identities in the gallery is unknown, we utilize the Chinese Whispers clustering; an efficient graph-based parameter-free clustering algorithm introduced in [4] which discovers the number of clusters in a simple iterative process.

Output. For each subject-specific gallery, the pipeline is finalized by retaining facial samples and associated labels only from the largest identity cluster. All other samples, belonging to clusters associated with subjects appearing less frequently than the gallery's owner, are discarded.

The designed filtering pipeline is completely parameterless and utilizes only generic off-the-shelf models and algorithms. This approach does not require supervised training of dataset-specific models or any type of manual effort.

[5] http://dlib.net

3.2 Dataset filtering

The proposed filtering pipeline is applied to the two largest publicly available facial age estimation datasets; the CACD dataset and the IMDB-WIKI dataset. In this section we discuss method's prerequisites, data preprocessing, and results of the dataset filtering based on the proposed method.

3.2.1 Prerequisites

There are two prerequisites that need to be satisfied for the proposed filtering method to be applicable:

1. There must be multiple images of every subject.
2. For each subject-specific image gallery, the number of appearances of the gallery owner's face must exceed the number of appearances of any other subject.

As we can see from Table 1, the average number of images per subject is 81.72 for the CACD, and 22.71 for the IMDB dataset, indicating that the method's first prerequisite will be satisfied for the majority of subjects. The WIKI subset of the IMDB-WIKI dataset has only one image per subject, therefore it will not be filtered with this method.

The probability of a well-defined image search producing more bad than good results is very low. The probability of a subject not being the most frequently appearing person on its IMDb/Wikipedia profile photos is even lower. Therefore, the method's second prerequisite is satisfied intrinsically for the majority of samples from the CACD and IMDB-WIKI datasets.

3.2.2 Preprocessing

Although the training data can in some cases be used in its raw form for the supervised training of machine learning models, in most cases a set of initial processing steps is applied to the data. Initial processing of the image data typically consists of image cropping, alignment, and filtering according to attainable image properties. Based on the label data, samples can also be filtered to remove invalid labels, improve sample distribution, etc. The goal of this section is to distinguish the three versions of data used in our experiments; raw, processed, and filtered data.

Raw data (R). Both CACD and IMDB-WIKI datasets provide pre-cropped facial images with associated age labels. As face detection and cropping are the minimal preprocessing steps for the typical face analysis systems, we consider this to be the raw data. This data consists of loose unaligned face crops of varying quality and resolution produced based on face detections obtained by the datasets' authors. The associated labels are most likely the direct product of the automatic labeling procedures, so invalid labels can be present in the data. To make this raw data compatible with our training framework, we apply center-cropping to the non-square images, resize all images to the same resolution and limit the age labels to the $[0,100]$ range by applying function $age = min(max(age,0),100)$. None of the samples are discarded. This results in CACD-R, IMDB-R, and WIKI-R dataset variants.

Processed data (P). The processed data refers to the outcome of the typically applied automatic preprocessing steps that are designed to improve the raw data quality and remove some of the obvious outliers that impact the training of the models. Whereas the originally provided face detections are used in the raw data (R), the processed data (P) consists of re-detected, aligned, and re-cropped samples obtained by the detection and alignment approach described in the section 3.1.1. Images that were damaged, had very low resolution or in any other way caused the described detection, alignment, and cropping pipeline to fail were discarded. Additionally, a small number of samples that had age labels with biologically impossible (i.e. negative) or highly improbable (i.e. greater than 100) values were also discarded since they are most probably result of failed automatic labeling procedures.

Filtered data (F). The filtered data is the data obtained by applying the proposed filtering method described in the Section 3.1 to the previously described processed versions of the CACD and IMDB datasets (i.e. CACD-P and IMDB-P). This way the filtered data (i.e. CACD-F and IMDB-F) is directly comparable to the processed data and the impact of the proposed filtering method is unambiguously observable.

Data augmentations, such as random cropping and image flipping, are not considered to be part of the initial processing as the same augmentation techniques are applied to all three versions of the data (i.e. R, P, and F) during training.

3.2.3 Filtering results

The raw versions of the CACD and IMDB datasets have 163,446 and 460,723 samples, respectively. By applying initial processing, described in the previous section, we produced the CACD-P dataset with 150,383 samples and the IMDB-P dataset with 451,571 samples. After the proposed filtering method was applied to the processed data, 130,571 samples were retained from the CACD-P dataset (13.2% reduction), and only 216,595 samples from the IMDB-P dataset (52.0% reduction), giving us the final CACD-F and IMDB-F dataset versions. As we can see in Figure 3, the sample distributions of the filtered subsets of the CACD and IMDB datasets remained similar to the raw and processed versions, while the number of samples was greatly reduced.

To examine the filtering results more closely, outputs for several subject-specific galleries were manually inspected and showed consistent results. Figure 4 shows the results of a statistical analysis of filtering outputs for one of the subjects from the IMDB dataset. The figure contains a his-

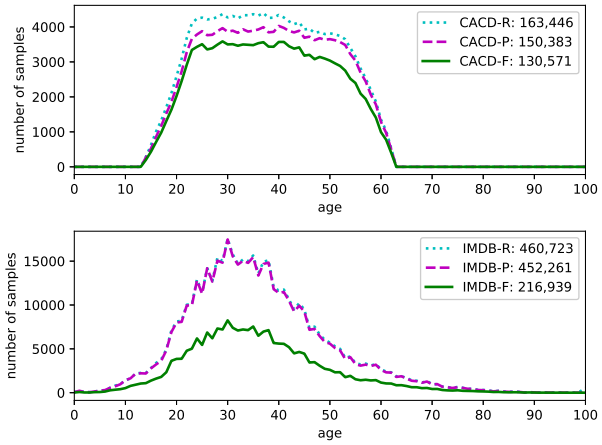

Fig. 3 Age label distributions for raw (R), processed (P), and filtered (F) versions of the CACD and IMDB datasets.

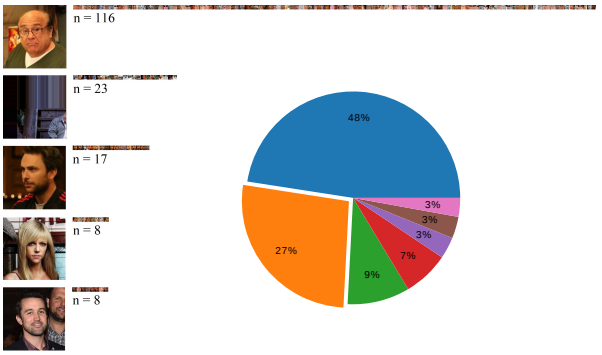

Fig. 4 Clustering results for the IMDB subject Danny DeVito. Images on the left show representative samples for the top 5 clusters, horizontal bars contain all cluster samples (best viewed in electronic version with extreme zoom), and the chart on the right shows cluster size distribution with emphasized part representing all clusters containing 1 to 3 samples [3].

togram for the top five sample clusters and a chart representing the cluster sizes. The 48% of the samples that were grouped into the largest cluster were kept while 52% of the samples were filtered-out. The analysis showed that the second largest cluster (9%) grouped primarily non-facial images caused by false-positive detections, and the subsequent clusters contained facial images of actors with whom the subject is most frequently associated. The emphasized part of the chart in Figure 4 represents all clusters with only 1 to 3 samples (1 to 3 occurrences per identity) therefore grouping the less frequently appearing outliers.

3.3 Experimental evaluation

The proposed filtering pipeline described in Section 3.1.1 resulted in strong sample count reduction, as presented in Section 3.2.3. To validate that the resulting subsets of the original datasets have higher percentages of valid data and that the proposed automatic filtering approach is beneficial to the datasets' applicability to the facial biometric task they were designed for, we perform an extensive set of age estimation experiments.

Good generalization capabilities, crucial for real world in-the-wild applications, often directly depend on the training data sample count and diversity. To validate that our aggressive sample reduction does not impair the generalization capabilities of the trained models, we performed cross-dataset testing on the unconstrained manually collected Appa-Real benchmark.

To validate if the benefits are method-invariant, we perform evaluation based on two frequently used and three advanced age estimation methods, described in the following section.

3.3.1 Age estimation methods

To perform unbiased evaluation of different age estimation methods, we used an identical CNN feature extraction model for all methods, thus supporting each of the methods with the same number of learnable parameters. The models were designed to output 101-dimensional feature vectors intended to facilitate estimation of 101 age values ranging from 0 to 100. Based on this setup, we implemented five relevant age estimation methods, described after the following concise explanation of the mathematical notation.

Formally, in a dataset with N samples, let $x_i \in \mathbb{R}^{h\times w\times c}$ denote i-th image sample where h, w, and c are the image height, width, and number of channels, respectively. Assuming that $y_i \in \{0,1,\ldots,K\}$ is the corresponding age label with range from 0 to K, $z_i \in \mathbb{R}^K$ represents the CNN model output feature vector. Vector z_i is obtained as $z_i = f(x_i;\theta)$, where θ denotes CNN model parameters.

Softmax. One of the simplest and most frequently used age estimation approaches is to interpret age estimation as a classification problem with number of classes corresponding to the number of different age values. We adopt the most common approach based on the Softmax function and Cross Entropy Loss, commonly referred to as the Softmax method. First, we obtain estimated probability distribution $\hat{p}_i \in \mathbb{R}^K$ based on the CNN output z_i by applying the Softmax function

$$\hat{p}_{i,j} = \frac{e^{z_{i,j}}}{\sum_{k=1}^{K} e^{z_{i,k}}}, \tag{1}$$

for $j \in \{1,2,\ldots,K\}$. To train the model parameters θ we utilize Cross Entropy Loss defined by Eq. 2. At inference time, age estimation is calculated as $\hat{y}_i = \underset{j}{\operatorname{argmax}}(\hat{p}_{i,j})$.

$$L_S = \frac{1}{N}\sum_{i=1}^{N} -y_i \log \hat{p}_{i,y_i} \tag{2}$$

Euclidean. The second frequently used age estimation approach is to treat age estimation as a regression task. To accommodate this approach, the used CNN model needs to output a single value $\hat{y}_i \in \mathbb{R}^1$. For this purpose, we calculate the output $\hat{y}_i$ based on the CNN model feature vector z_i as

$$\hat{y}_i = \sum_{j=1}^{K} j * \hat{p}_{i,j}, \tag{3}$$

where $\hat{p}_{i,j}$ once again denotes the estimated Softmax distribution from Eq. 1. To train the model parameters θ we adopt the Mean Square Error Loss (MSE), commonly referred to as Euclidean loss, defined by Eq. 4.

$$L_E = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{4}$$

DEX. Similar to the Softmax method, the Deep Expectation approach [41] uses Softmax function from Eq. 1 and Cross Entropy Loss defined by Eq. 2 to train the CNN model parameters θ. However, contrary to the Softmax method, estimated probability distributions are used at inference time to calculate a continuous value for age estimation as a weighted average defined by Eq. 3. This improved method uses Softmax-based learning and regression-based inference-time age calculation.

Mean-Variance. The fourth implemented method extends the idea introduced in [41] by enhancing the basic Softmax Loss with two additional components; Mean Loss and Variance Loss. The Mean-Variance Loss [36] is defined as

$$L_{MV} = \frac{1}{N}\sum_{i=1}^{N}\left(-y_i \log \hat{p}_{i,y_i} + \frac{\lambda_1}{2}(y_i - \hat{y}_i)^2 + \lambda_2 \sum_{j=1}^{K} p_{i,j} * (j - \hat{y}_i)^2\right), \tag{5}$$

where the factor λ_1 controls the contribution of the Mean Loss component, while λ_2 controls the contribution of the Variance Loss. Following the setup from [36], we set λ_1 and λ_2 to 0.2 and 0.05, respectively. This advanced method combines classification and regression-based components with an additional distribution-oriented variance component at training time. At inference time, the estimated age is calculated according to Eq. 3.

DLDL-v2. The final age estimation method used for evaluation is DLDL-v2 [17]. This label-distribution-based method trains the CNN model parameters θ by utilizing a loss function defined as

$$L_{LD} = \frac{1}{N}\sum_{i=1}^{N}\left(\sum_{j=1}^{K} -p_{i,j} \log \hat{p}_{i.j} + \lambda \mid y_i - \hat{y}_i \mid\right), \tag{6}$$

where p_i denotes approximated ground truth label distribution. The aforementioned distribution is calculated based on the ground truth age label y_i as

$$p_{i,j} = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(j - y_i)^2}{2\sigma^2}\right) \tag{7}$$

for $j \in \{1, 2, \ldots, K\}$. This advanced method combines the distribution-based Kullback-Leibler Divergence Loss and regression-based L1 Loss. The parameter λ, which controls the balance between the loss components, was set to 1 in all experiments, while σ was set to 2. At inference time, the estimated age is once again calculated according to Eq. 3.

3.3.2 Experimental setup

To compare results before and after the proposed filtering method was applied, we train the models on the processed (P) and the resulting filtered (F) versions of the datasets under identical conditions. Additionally, we train the baseline models on the raw data (R) under the same setup.

A set of image data augmentations, consisting of randomized horizontal flipping (image mirroring), bounding box perturbations, and in-plane rotations, was applied to training subsets of all three versions of the datasets. Random in-plane rotations were limited to $[-10°, 10°]$ range. The bounding box perturbations were achieved by resizing the image to 110×110 pixels and performing random cropping to 96×96 dimensions expected by the used model, effectively obtaining perturbations of up to 15% of bounding box scale.

The model selected for this extensive evaluation was a simple 9-layer CNN model based on the open-source architecture Tiny DarkNet[6]. This minimalistic 1M-parameter architecture, proposed for real-time performance in [39], was pre-trained for the task of face recognition and further modified to take low-resolution $3 \times 96 \times 96$ RGB inputs and produce 101 outputs, corresponding to age values from 0 to 100.

All models were trained for 100 epochs with batch size of 64 and optimized based on widely adopted Adadelta optimization algorithm [46] with learning rate set to 10^{-1}. To calculate the estimation errors, we adopted the common mean absolute error measure (MAE).

5-fold validation. The first set of experiments was based on 5-fold validation and the most common age estimation method; the Euclidean method (i.e. regression). The datasets were randomly divided into 5 equally-sized parts and for each of the 5 folds, a different part was used for validation, while the rest of the data was used for trainings. This way, 5 different 80-20 training-validation splits were created and used to achieve robust verification based on mean

[6] https://pjreddie.com/darknet/tiny-darknet/

Table 2 Results of the 5-fold validation of the raw (R), processed (P), and filtered (F) versions of the CACD and IMDB datasets. The testing and fine-tuning were performed on the Appa-Real benchmark. R-MAE and A-MAE denote mean absolute errors for real and apparent age estimation, respectively.

		Validation	Testing		Fine-tuning	
Dataset	Images	R-MAE	R-MAE	A-MAE	R-MAE	A-MAE
CACD-R	163,446	5.56 ± 0.03	14.60 ± 0.58	13.15 ± 0.67	9.11 ± 0.59	7.39 ± 0.72
CACD-P	150,383	4.66 ± 0.01	13.61 ± 0.13	12.30 ± 0.15	7.64 ± 0.17	5.78 ± 0.13
CACD-F	130,571	$\mathbf{3.46 \pm 0.01}$	$\mathbf{11.83 \pm 0.26}$	$\mathbf{10.60 \pm 0.26}$	$\mathbf{7.17 \pm 0.13}$	$\mathbf{5.46 \pm 0.15}$
IMDB-R	460,723	8.11 ± 0.03	11.14 ± 0.11	9.51 ± 0.08	9.84 ± 0.37	8.12 ± 0.37
IMDB-P	451,571	7.70 ± 0.02	8.51 ± 0.18	7.20 ± 0.14	6.73 ± 0.20	5.20 ± 0.34
IMDB-F	216,595	$\mathbf{5.07 \pm 0.01}$	$\mathbf{6.83 \pm 0.13}$	$\mathbf{5.63 \pm 0.14}$	$\mathbf{6.31 \pm 0.06}$	$\mathbf{4.72 \pm 0.10}$

and standard deviation of the 5 result variations. To further show that the proposed filtering method is beneficial even in case of highly specialized transfer learning, we performed additional fine-tunings on the training part of a separate benchmark dataset. The fine-tunings were performed for 500 epochs with SGD optimization [28] and a relatively low learning rate (10^{-5}). Separate models were trained based on real and apparent age labels.

5-method validation. The second set of experiments was designed to evaluate if consistent performance gains are obtainable for each of the 5 age estimation methods described in the previous section, thus indicating that the benefits of the proposed filtering are method-invariant. As mentioned in the previous section, identical CNN backbone model was used for all methods. 80% of the data from the evaluated dataset was used for training, 20% for validation, while testing was performed on a separate unconstrained benchmark dataset.

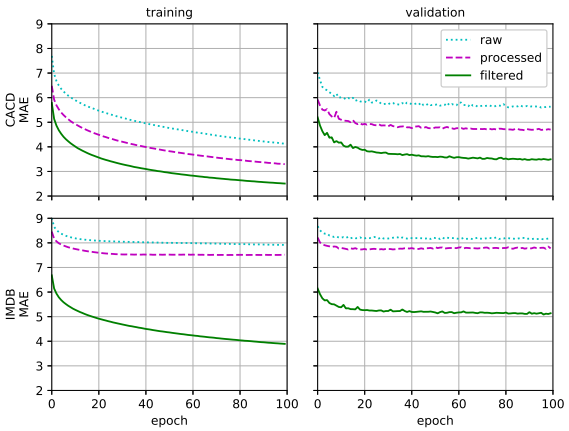

Fig. 5 Average training and validation MAEs for the first 100 epochs of 5-fold trainings on the raw, processed, and filtered versions of the CACD and IMDB datasets. The first row presents results for the CACD dataset variants (CACD-R, CACD-P, and CACD-F), while the second row presents results for the IMDB dataset variants (IMDB-R, IMDB-P, and IMDB-F).

3.3.3 Evaluation results

The results of the 5-fold evaluation are presented in Figure 5 and Table 2. Figure 5 shows the average training and validation MAEs over 100 training epochs for the 6 different dataset variations. In addition to the benefits of the standard data processing, the graphs show that the proposed data filtering introduces further large reduction in training and validation errors. The results presented in Table 2 further support the proposed filtering. Compared to the processed CACD and IMDB data, the respective filtered data 5-fold results were improved by 1.20 and 2.63 years for the validation MAE, 1.78 and 1.68 years for the real age testing MAE, and 1.70 and 1.57 years for the apparent age testing MAE. Even after the specialized fine-tunings, versions pre-trained on the filtered data consistently yielded improved results of up to 0.48 years, regardless of the type of age estimation task. Note that the high CACD testing errors were caused by the lack of young and old people in the CACD dataset, as shown in Figure 3, and were greatly reduced after the Appa-Real finetunings. Compared to the baseline raw data, the performance was improved with up to 4.31 years.

The results of the 5-method evaluation are presented in Table 3. The results demonstrate considerable performance gains regardless of the age estimation method. Compared to the processed version, the filtered versions of the datasets improved the results by between 1.50 and 2.21 years and between 1.49 and 1.99 years for the CACD and IMDB datasets, respectively. Compared to the baseline raw data, the performance was improved with up to 4.39 years.

For both datasets and both types of age estimation task, robust 5-fold and diverse 5-method validations demonstrated the advantages of using versions of the datasets filtered by the proposed method. Expectedly, the raw versions of the dataset resulted in worst performing age estimation models. The standard simple data processing techniques resulted in significant performance gains. Nonetheless, the direct subsets of the processed data, obtained by the proposed filter-

Table 3 Results of the 5-method validation of the raw (R), processed (P), and proposed filtered (F) versions of the CACD and IMDB datasets. The testing was performed on the Appa-Real benchmark. The values represent real age mean absolute error (MAE).

Dataset	Images	Softmax	Euclidean	DEX [41]	MV [36]	DLDL-v2 [17]
CACD-R	163,446	15.787	14.730	14.513	14.164	14.886
CACD-P	150,383	14.045	13.605	13.420	12.988	13.338
CACD-F	130,571	**12.545**	**11.390**	**11.854**	**11.453**	**11.589**
IMDB-R	460,723	11.547	11.109	11.650	11.160	10.938
IMDB-P	451,571	9.667	8.568	9.338	8.503	8.491
IMDB-F	216,595	**7.807**	**7.080**	**7.349**	**6.771**	**6.787**

ing method, resulted in additional substantial performance gains, thus demonstrating the benefits of the proposed method and highlighting that the data quality is more important than the data quantity. Interestingly, the results also demonstrate that the training data processing and filtering can result in larger margin of improvement than utilization of more advanced age estimation methods.

4 Biometrically Filtered Famous Figure Dataset

This section presents the designed strategy undertaken to derive a new age estimation dataset, extensive evaluation results, and comparison with the state-of-the-art.

In section 3.3 we demonstrated that by applying the unsupervised biometric filtering method introduced in section 3.1 we produced superior versions of the IMDB and CACD datasets, respectively denoted as IMDB-F and CACD-F. To create a new famous figures dataset with improved properties, we explore options for combining and further filtering the publicly available web-scraped data from IMDB-F, CACD-F, and WIKI datasets.

4.1 Dataset design

As the first step in dataset design, we analyze interrelation between the IMDB-F, CACD-F, and WIKI datasets. Since all three datasets were collected by scraping images of famous figures and they all provide subject's names in the meta-data, we decided to analyze identity overlaps of the 3 datasets.

For the purpose of identity overlap analysis, we convert the originally provided names from plain Unicode versions to a representation that contains only lower-case ASCII alphabetic symbols. This step helps to mitigate name matching failures by eliminating potentially ambiguous white-space, diacritic, and other special symbols. Figure 6 visualizes obtained identity overlaps between the 3 datasets.

By combining information from Table 1 and Figure 6, we observe that the WIKI dataset, although having the lowest number of samples, has the highest number of unique

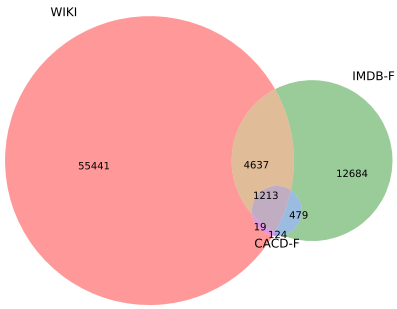

Fig. 6 Identity overlaps for WIKI, IMDB-F and CACD-F datasets.

identities. In fact, the WIKI dataset contains only one image per sample. Compared to the WIKI dataset, IMDB and CACD datasets have lower numbers of unique identities, but have large average numbers of samples per subject, thus enabling us to apply the proposed filtering method from section 3.1. While IMDB-F and CACD-F datasets can contribute with large amounts of filtered samples and substantial dataset depth, adding samples from the WIKI dataset can potentially increase the overall number of unique identities by 280%, hence greatly improving the dataset breadth.

We apply the initial processing described in section 3.2.2 to the raw WIKI data to produce the WIKI-P dataset version. To further enhance the WIKI-P dataset, we take advantage of the observation from section 3.2.3 and figure 4; non-facial images (i.e. false positive detections) have tendency to form clusters in the face recognition descriptor feature space. We exploit this finding to design a simple false positive detection filtering method.

Finally, leveraging the interrelation between the 3 datasets, we design an additional majority voting filtering approach that takes advantage of the dataset identity overlaps.

4.1.1 Biometric false positive detection filtering

Once again, we *pick out the bad apples* by designing a biometric clustering approach similar to the method proposed in section 3.1, but applicable to the WIKI data. This simplified version of the aforementioned method does not require

Fig. 7 Representative WIKI dataset samples from 4 clusters (C1, C2, C3, and C4) obtained by K-Means clustering with K set to 64. While clusters C1, C3, and C4 contain similar samples w.r.t. face recognition descriptors, cluster C2 groups mostly non-facial samples.

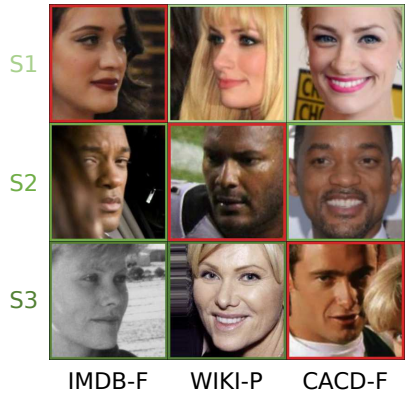

Fig. 8 Outliers detected by majority vote filtering for subjects S1, S2, and S3 from datasets IMDB-F, WIKI-P, and CACD-F. Outliers are placed on the main diagonal and marked with red frames (best viewed in color).

multiple images of the same subject as the formed clusters are not subject-based. The method consists of 3 main steps; feature extraction, sample clustering, and elimination of the *bad* clusters containing false positive detections.

For the feature extraction step we retain the face recognition descriptor extraction step that provides compact and highly descriptive numerical representation of facial images with tendency for clustering of non-facial samples, as was observed in Section 3.2.3.

Whereas we chose to use a graph-based clustering approach to implement the filtering method proposed in section 3.1, here we decide to use the distance-based K-Means clustering method as it also provides cluster centers useful for the final cluster elimination step. To form candidate clusters, we apply K-Means clustering with an arbitrary large K. Figure 7 shows 4 representative cluster samples obtained by clustering the WIKI-P data with K set to 64.

To identify clusters with false positive detections, we choose a single seed false positive detection sample and compare its face recognition descriptor with the central descriptor of each of the K-Means clusters. The second row in Figure 7 shows representative samples from a WIKI-P false positive detection cluster identified by this automatic approach.

This method can be applied to detect non-facial outliers in any type of facial image dataset. We apply it to CACD-F and IMDB-F datasets as well to achieve an additional level of data refinement.

4.1.2 Majority vote filtering

To avoid conflicting labels that could emerge in the process of merging the IMDB-F, CACD-F, and WIKI-P data, we leverage the identity overlaps between the three datasets shown in Figure 6 to implement a majority vote filtering method and further refine the data.

Once again, we base our filtering method on face recognition descriptors and name-based identifiers formed as described at the beginning of section 4.1. Algorithm 1 describes our majority vote filtering approach in detail, while Figure 8 shows representative samples of outliers identified by the described algorithm.

Assuming that in in Algorithm 1 $A,B \in [I,C,W]$ denotes a dataset from a set including IMDB-F, CACD-F and WIKI-P datasets, I_A represents set of identities in the dataset A, S_A^{id} represents set of samples from the dataset A corresponding to the specific identity id, and D_A^{id} represents the mean descriptor for all samples in S_A^{id}. Furthermore, $D_A^{id} \approx D_B^{id}$ denotes that descriptors for identity id from datasets A and B are similar, while $mean(D_A^{id}, D_B^{id})$ represents the mean descriptor for descriptors D_A^{id} and D_B^{id}.

Mean descriptors are calculated by simple averaging. Cosine similarity is used as the similarity measure with the similarity threshold set to 0.25. This relatively loose threshold, along with comparison based on mean descriptors, reduces chances of important difficult facial samples to be removed.

4.2 Dataset properties

By combining proposed filtering methods from sections 3.1, 4.1.1, and 4.1.2, and merging the refined IMDB, CACD, and WIKI data, we produce a new derived facial age estimation dataset dubbed Biometrically Filtered Famous Figure Dataset or B3FD in short. The dataset is made publicly available[7].

B3FD contains 375,592 facial image samples with corresponding age labels. It has 53,759 unique subjects, which amounts to 6.99 samples per subject on average. The age labels are ranging from 0 to 100. Distribution of age la-

[7] https://github.com/kbesenic/B3FD

Algorithm 1: Majority vote filtering

Input: $I_I, I_C, I_W, D_I, D_C, D_W, S_I, S_C, S_W$
for $id \in (I_I \cup I_C \cup I_W)$ **do**
  if $id \in (I_I \cap I_C \cap I_W)$ **then**
    if $D_I^{id} \approx D_C^{id}$ ***and*** $D_W^{id} \not\approx mean(D_I^{id}, D_C^{id})$ **then**
      remove S_W^{id};
    end
    if $D_C^{id} \approx D_W^{id}$ ***and*** $D_I^{id} \not\approx mean(D_C^{id}, D_W^{id})$ **then**
      remove S_I^{id};
    end
    if $D_I^{id} \approx D_W^{id}$ ***and*** $D_C^{id} \not\approx mean(D_I^{id}, D_W^{id})$ **then**
      remove S_C^{id};
    end
  else if $id \in (I_I \cap I_C)$ **then**
    if $D_I^{id} \not\approx D_C^{id}$ **then**
      if $|S_I^{id}| < |S_C^{id}|$ **then**
        remove S_I^{id};
      else
        remove S_C^{id};
      end
    end
  else if $id \in (I_W \cap I_C)$ **then**
    if $D_W^{id} \not\approx D_C^{id}$ **then**
      if $|S_W^{id}| < |S_C^{id}|$ **then**
        remove S_W^{id};
      else
        remove S_C^{id};
      end
    end
  else if $id \in (I_W \cap I_I)$ **then**
    if $D_W^{id} \not\approx D_I^{id}$ **then**
      if $|S_W^{id}| < |S_I^{id}|$ **then**
        remove S_W^{id};
      else
        remove S_I^{id};
      end
    end
  end
end

bels for B3FD dataset and its main components (i.e. IMDB-F, CACD-F, and WIKI-P) is shown in Figure 9. Our combined filtering efforts resulted in removal of 310,905 samples, which makes 45.29% of the originally provided data.

The B3FD datasets is composed of two main subsets determined by the data origin; the IMDB-WIKI subset (i.e. B3FD-IWS) and the CACD subset (i.e. B3FD-CS). B3FD-IWS consists of 245,204 processed samples from the IMDB-WIKI dataset with 53,568 unique subjects, which amounts to 4.58 samples per subject on average. B3FD-CS consists of 130,388 processed samples from the CACD dataset with 1,831 unique subjects, which amounts to 71.21 samples per subject. These subsets can be useful in case of data-origin-based constraints (e.g. only IMDB-WIKI data can be used).

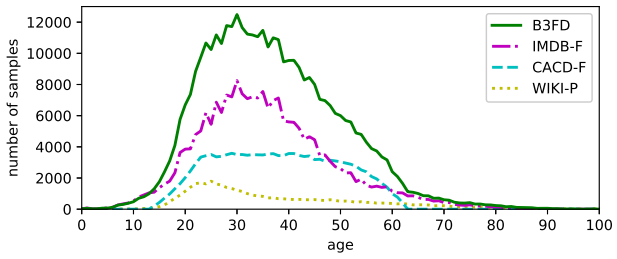

Fig. 9 Age label distributions for the proposed B3FD dataset and its components (i.e. IMDB-F, CACD-F, and WIKI-P datasets).

4.3 Comparison with the state of the art

This section provides an extensive evaluation of the B3FD dataset and its subsets presented in section 4.2. The evaluation is performed by comparing performance of models trained on the B3FD dataset variations and the previous state-of-the-art age estimation datasets with respect to real and apparent age estimation accuracy. To demonstrate generalization capabilities of the models trained on evaluated datasets, the comparison is once again facilitated by a separate unconstrained in-the-wild age estimation benchmark with real and apparent age labels. Furthermore, to show that performance gain is method-indifferent, we perform evaluation based on 5 different age estimation methods from Section 3.3.1.

4.3.1 Experimental setup

To facilitate this extensive evaluation, we chose to use a CNN architecture from MobileFaceNet family [9] designed specifically for efficient face analysis. Table 5 describes the used model architecture in detail. The standard MobileFaceNet architecture was slightly modified to use 128×128 RGB inputs, output 101 values corresponding to ages from 0 to 100, and pre-trained for the face recognition task. The model has 1M parameters with inference-time computational cost of 0.99 GFLOPs.

The conventional data processing, explained in Section 3.2.2 was applied to all evaluated datasets. A set of image data augmentations, consisting of randomized horizontal flipping, bounding box perturbations, and in-plane rotations, described in Section 3.3.2, was applied to all training sets, with only difference being the resulting image resolution (i.e. 128×128).

All evaluated models were trained for 50 epochs with batch size of 64 and SGD optimization [28]. We set the weight decay parameter to 10^{-4} and momentum to 0.9. A learning rate scheduler was used to reduce the learning rate by factor of 10 every 15 epochs. Initial learning rate was set to 10^{-3}.

A random 90-10 training-validation split was used for all datasets except Appa-Real where the predefined validation set was used and the MegaAge where the predefined test set was used as the validation set. The models used for dataset evaluation were chosen based on the validation set MAE.

Table 4 Performance of age estimation models trained on different datasets with Softmax and Euclidean age estimation methods and evaluated on Appa-Real test set w.r.t. mean absolute error for real and apparent age, along with vote-distribution-based ε-error. R-MAE and A-MAE denote mean absolute errors for real and apparent age estimation, respectively.

		Softmax			Euclidean		
Dataset	Images	R-MAE	A-MAE	ε-error	R-MAE	A-MAE	ε-error
Appa-Real [1]	7,591	7.252	6.295	0.453	8.287	6.984	0.517
AgeDB [31]	16,488	9.851	9.259	0.545	10.847	9.988	0.563
MegaAge [47]	41,941	9.588	8.434	0.587	10.404	9.364	0.633
MORPH [40]	55,134	13.787	12.478	0.667	13.467	11.698	0.642
CACD [7]	163,446[1]	13.369	12.448	0.631	12.667	11.707	0.636
IMDB-WIKI [41]	523,051[2]	6.509	6.393	0.434	7.294	6.323	0.432
IMDB-WIKI + CACD [41, 7]	686,497[3]	6.986	6.981	0.449	7.819	6.921	0.449
B3FD-CS [ours]	130,388	11.143	10.164	0.592	11.327	10.239	0.598
B3FD-IWS [ours]	245,204	**5.423**	**5.275**	**0.408**	5.828	5.238	**0.408**
B3FD [ours]	375,592	5.547	5.532	0.415	**5.707**	**5.186**	0.409

[1] The processed subset of the data (P) with 150,383 samples was used as it was shown to be superior to the raw data in Section 3.3.
[2] The processed subset of the data (P) with 494,158 samples was used as it was shown to be superior to the raw data in Section 3.3
[3] The sum of used samples from IMDB-WIKI and CACD is 644,541 for reasons explained in preceding notes.

Table 5 Architectural details of the used MobileFaceNet-based model. Following MobileFaceNet notation from [9], rows describe a sequence of operators repeated n times with c output channels. Stride of first layer in each sequence is denoted by s while t represents the expansion factor applied to the input size.

Input	Operator	t	c	n	s
$\mathbf{128^2 \times 3}$	conv3x3	–	64	1	2
$64^2 \times 64$	depthwise conv3x3	–	64	1	1
$64^2 \times 64$	bottleneck	2	64	5	2
$32^2 \times 64$	bottleneck	4	128	1	2
$16^2 \times 128$	bottleneck	2	128	6	1
$16^2 \times 128$	bottleneck	4	128	1	2
$8^2 \times 128$	bottleneck	2	128	2	1
$8^2 \times 128$	conv1x1	–	512	1	1
$8^2 \times 512$	linear GDConv8x8	–	512	1	1
$1^2 \times 512$	linear conv1x1	–	**101**	1	1

All models were evaluated on the Appa-Real test set, as it is the only age estimation dataset providing real and apparent age labels for unconstrained in-the-wild facial images.

We evaluate if performance gains are indifferent to the type of age estimation task by providing MAE values for both real and apparent age. To additionally evaluate if the performance gains are indifferent to the type of evaluation metric, we also report ε-error [12] based on distribution of votes for apparent age. The ε-error is defined as

$$\varepsilon = 1 - e^{-\frac{(\hat{y}-\mu)^2}{2\sigma^2}}, \tag{8}$$

where μ and σ are the mean and standard deviation of the apparent age votes, while $\hat{y}$ represents the age prediction.

4.3.2 Evaluation results

The results of the experimental evaluation of the proposed B3FD dataset are presented in tables 4 and 6. Table 4 presents evaluation results for two most frequently used age estimation methods; classification-based Softmax method and regression-based Euclidean method. Table 6 presents results for 3 additional advanced age estimation methods described in section 3.3.1; DEX, Mean-Variance, and DLDL-v2.

To validate if the proposed B3FD data outperforms its main components, we evaluated CACD and IMDB-WIKI datasets under identical conditions. Furthermore, to validate the proposed merging strategy from section 4.1, we also evaluated performance of the naively merged IMDB-WIKI + CACD dataset. To validate if the proposed B3FD data provides better generalization capabilities than manually collected data, we evaluated performance of the largest manually collected dataset (i.e. MORPH) and the more recent in-the-wild AgeDB dataset. To validate if the proposed B3FD data outperforms the largest in-the-wild dataset for apparent age estimation, we evaluated performance of the MegaAge dataset. Finally, to validate if the proposed dataset outperforms manually collected domain-specific data, we evaluated models trained on the training part of the evaluation Appa-Real benchmark.

The results show that the models based on the proposed B3FD data outperformed models trained on the other evaluated datasets by a notable margin, without exception. The experiments demonstrated that the performance gains are obtained regardless of the used age estimation method, type of age estimation task, or metric. Compared to the naive combination of all evaluated web-scraped data, our derived B3FD dataset provided MAE reduction between 1.44 and

Table 6 Performance of age estimation models trained on different datasets with 3 advanced age estimation methods and evaluated on Appa-Real test set w.r.t. mean absolute error for real and apparent age, along with vote-distribution-based ε-error. R-MAE and A-MAE denote mean absolute errors for real and apparent age estimation, respectively.

		DEX [41]			Mean-Variance [36]			DLDL-v2 [17]		
Dataset	Images	R-MAE	A-MAE	ε-error	R-MAE	A-MAE	ε-error	R-MAE	A-MAE	ε-error
Appa-Real [1]	7,591	7.511	6.382	0.478	7.462	6.355	0.477	7.369	6.099	0.457
AgeDB [31]	16,488	10.023	9.303	0.536	10.378	9.625	0.548	9.851	9.251	0.549
MegaAge [47]	41,941	10.666	9.685	0.643	9.941	8.938	0.623	10.021	9.073	0.633
MORPH [40]	55,134	13.608	12.171	0.652	13.434	11.768	0.642	12.871	11.211	0.614
CACD [7]	163,446[1]	13.538	12.454	0.637	12.555	11.409	0.613	12.079	11.005	0.605
IMDB-WIKI [41]	523,051[2]	7.155	6.673	0.433	7.077	6.259	0.431	6.568	6.101	0.418
IMDB-WIKI + CACD [41, 7]	686,497[3]	7.658	7.174	0.447	7.606	6.831	0.449	7.106	6.704	0.417
B3FD-CS [ours]	130,388	10.914	9.816	0.583	11.033	9.992	0.587	10.844	9.869	0.590
B3FD-IWS [ours]	245,204	5.340	**5.072**	**0.388**	**5.394**	**4.808**	**0.386**	5.158	**4.684**	**0.383**
B3FD [ours]	375,592	**5.282**	5.118	0.393	5.441	5.025	0.403	**5.077**	5.064	0.408

[1] The processed subset of the data (P) with 150,383 samples was used as it was shown to be superior to the raw data in Section 3.3.
[2] The processed subset of the data (P) with 494,158 samples was used as it was shown to be superior to the raw data in Section 3.3
[3] The sum of used samples from IMDB-WIKI and CACD is 644,541 for reasons explained in preceding notes.

2.38 years for real age and between 1.45 and 2.06 years for apparent age. Compared to the MORPH dataset, the MAE reductions range from 6.15 to more than 8 years, most probably due to the constraints and biases associated with manual data collection. The most recent in-the-wild manually collected AgeDB dataset outperforms the larger manually collected MORPH dataset, but still falls behind the proposed B3FD data by between 4.30 and 5.14 years for real age and 3.73 and 4.82 years for the apparent age. The proposed B3FD data also outperformed the MegaAge dataset by a similar margin for both real and more importantly apparent age, despite MegaAge providing apparent age labels, while the proposed B3FD data provides real age labels. Even in case of the domain-specific manually collected in-the-wild data from Appa-Real, the performance was improved by a considerable margin of up to 2.58 years. B3FD-CS outperformed its corresponding CACD superset by up to 2.62 years for real and 2.64 for apparent age, while B3FD-IWS outperformed its corresponding IMDB-WIKI superset by up to 1.81 years for real and 1.60 years for apparent age. B3FD-IWS even partially outperformed its B3FD superset, presumably due to the age distribution and identity number constraints of the CACD data.

Once again, the results highlighted that the training data quality is more important than the training data quantity and that the performance gains obtained by applying more advanced age estimation methods can be fairly less significant than the performance gains obtained by utilization of more relevant and filtered training data. This becomes more obvious in our experimental setup, where all age estimation methods are based on the same feature extraction backbone model. For example, the difference between the worst and the best performing age estimation methods on the combined IMDB-WIKI and CACD data is 0.83 years, while using the corresponding filtered B3FD dataset reduces the real age MAE by up to 2.38 years.

5 Conclusions

Compared to the manually collected facial datasets for biometric trait estimation, datasets collected by automatic web-scraping methods are far superior with respect to the sample count and variety but have a significant downside in terms of label noise. The filtering methods for label noise reduction often require dataset-specific trainings and manual efforts.

The proposed method for unsupervised biometric data filtering can automatically reduce the number of erroneous samples in facial web-scraped datasets by combining only a few general-purpose algorithms. The implemented filtering pipeline resulted in strong sample count reduction on two state-of-the-art web-scraped facial datasets as up to 52% of the samples were discarded. The robust 5-fold and diverse 5-method validations with cross-dataset testing both demonstrated that the models based on the filtered data outperform the models based on raw and conventionally processed data by a considerable margin, indicating lower amounts of faulty samples and improved label consistency, with an additional benefit of reduced training time. The testing of generalization capabilities on the in-the-wild data also indicates that the data diversity is not impaired by the proposed filtering method, despite the strong sample count reduction.

The results obtained by the proposed filtering method were extended by an additional biometric filtering strategy devised to reinforce and refine the merging process of the publicly available web-scraped datasets. The proposed merging process of the 3 different web-scraped data sources re-

sulted in a new derived in-the-wild age estimation dataset. The introduced Biometrically Filtered Famous Figure Dataset (B3FD) was experimentally evaluated and compared with both manually collected and web-scraped state-of-the-art datasets. The B3FD data consistently outperformed all evaluated datasets with respect to both real and apparent in-the-wild age estimation. B3FD is made publicly available.

The extensive experimental evaluation also highlighted the importance of training data quality and label consistency, as the results of models trained on the dataset subsets produced by the proposed filtering methods were superior to the results of models trained on larger datasets, as well as to models trained with more advanced age estimation methods.

Declarations

Funding: The author K. Bešenić receives Ph.D. scholarship from the company Visage Technologies.
Conflicts of interest: The author K. Bešenić is employed by Visage Technologies. Authors J. Ahlberg and I. S. Pandžić are members of Visage Technologies' board of directors. All three authors own stock in the company.
Availability of data and material: The proposed dataset is available at https://github.com/kbesenic/B3FD.

References

1. Agustsson E, Timofte R, Escalera S, Baro X, Guyon I, Rothe R (2017) Apparent and real age estimation in still images with deep residual regressors on appa-real database. In: 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), IEEE, pp 87–94
2. Antipov G, Baccouche M, Berrani SA, Dugelay JL (2016) Apparent age estimation from face images combining general and children-specialized deep learning models. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp 96–104
3. Bešenić K, Ahlberg J, Pandžić IS (2019) Unsupervised facial biometric data filtering for age and gender estimation. In: Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 5: VISAPP, pp 209–217
4. Biemann C (2006) Chinese whispers: an efficient graph clustering algorithm and its application to natural language processing problems. In: Proceedings of the first workshop on graph based methods for natural language processing, Association for Computational Linguistics, pp 73–80
5. Bulat A, Tzimiropoulos G (2017) How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks). In: Proceedings of the IEEE International Conference on Computer Vision, pp 1021–1030
6. Cao Q, Shen L, Xie W, Parkhi OM, Zisserman A (2018) Vggface2: A dataset for recognising faces across pose and age. In: Automatic Face & Gesture Recognition (FG 2018), 2018 13th IEEE International Conference on, IEEE, pp 67–74
7. Chen BC, Chen CS, Hsu WH (2014) Cross-age reference coding for age-invariant face recognition and retrieval. In: European conference on computer vision, Springer, pp 768–783
8. Chen S, Zhang C, Dong M, Le J, Rao M (2017) Using ranking-cnn for age estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 5183–5192
9. Chen S, Liu Y, Gao X, Han Z (2018) Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices. In: Chinese Conference on Biometric Recognition, Springer, pp 428–438
10. Cheng Y (1995) Mean shift, mode seeking, and clustering. IEEE transactions on pattern analysis and machine intelligence 17(8):790–799
11. Eidinger E, Enbar R, Hassner T (2014) Age and gender estimation of unfiltered faces. IEEE Transactions on In-

formation Forensics and Security 9(12):2170–2179
12. Escalera S, Fabian J, Pardo P, Baró 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: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp 1–9
13. Escalera S, Torres Torres M, Martinez B, Baró X, Jair Escalante H, Guyon I, Tzimiropoulos G, Corneou C, Oliu M, Ali Bagheri M, et al. (2016) Chalearn looking at people and faces of the world: Face analysis workshop and challenge 2016. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp 1–8
14. Frey BJ, Dueck D (2007) Clustering by passing messages between data points. science 315(5814):972–976
15. Gallagher AC, Chen T (2009) Understanding images of groups of people. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp 256–263
16. Gao BB, Xing C, Xie CW, Wu J, Geng X (2017) Deep label distribution learning with label ambiguity. IEEE Transactions on Image Processing 26(6):2825–2838
17. Gao BB, Zhou HY, Wu J, Geng X (2018) Age estimation using expectation of label distribution learning. In: IJCAI, pp 712–718
18. Golomb BA, Lawrence DT, Sejnowski TJ (1990) Sexnet: A neural network identifies sex from human faces. In: NIPS, vol 1, p 2
19. Han H, Otto C, Jain AK, et al. (2013) Age estimation from face images: Human vs. machine performance. ICB 13:1–8
20. He K, Zhang X, Ren S, Sun J (2015) Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE international conference on computer vision, pp 1026–1034
21. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
22. He Z, Li X, Zhang Z, Wu F, Geng X, Zhang Y, Yang MH, Zhuang Y (2017) Data-dependent label distribution learning for age estimation. IEEE Transactions on Image processing 26(8):3846–3858
23. Hu Z, Wen Y, Wang J, Wang M, Hong R, Yan S (2017) Facial age estimation with age difference. IEEE Transactions on Image Processing 26(7):3087–3097
24. Huang GB, Mattar M, Berg T, Learned-Miller E (2008) Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In: Workshop on faces in'Real-Life'Images: detection, alignment, and recognition
25. Jia S, Cristianini N (2015) Learning to classify gender from four million images. Pattern recognition letters 58:35–41
26. Kemelmacher-Shlizerman I, Seitz SM, Miller D, Brossard E (2016) The megaface benchmark: 1 million faces for recognition at scale. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4873–4882
27. Kwon YH, et al. (1994) Age classification from facial images. In: Computer Vision and Pattern Recognition, 1994. Proceedings CVPR'94., 1994 IEEE Computer Society Conference on, IEEE, pp 762–767
28. LeCun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Proceedings of the IEEE 86(11):2278–2324
29. Levi G, Hassner T (2015) Age and gender classification using convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp 34–42
30. Li P, Hu Y, Wu X, He R, Sun Z (2020) Deep label refinement for age estimation. Pattern Recognition 100:107178
31. Moschoglou S, Papaioannou A, Sagonas C, Deng J, Kotsia I, Zafeiriou S (2017) Agedb: the first manually collected, in-the-wild age database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp 51–59
32. Ng HW, Winkler S (2014) A data-driven approach to cleaning large face datasets. In: Image Processing (ICIP), 2014 IEEE International Conference on, IEEE, pp 343–347
33. Ni B, Song Z, Yan S (2009) Web image mining towards universal age estimator. In: Proceedings of the 17th ACM international conference on Multimedia, ACM, pp 85–94
34. Ni K, Pearce R, Boakye K, Van Essen B, Borth D, Chen B, Wang E (2015) Large-scale deep learning on the yfcc100m dataset. arXiv preprint arXiv:150203409
35. Niu Z, Zhou M, Wang L, Gao X, Hua G (2016) Ordinal regression with multiple output cnn for age estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4920–4928
36. Pan H, Han H, Shan S, Chen X (2018) Mean-variance loss for deep age estimation from a face. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 5285–5294
37. Panis G, Lanitis A (2014) An overview of research activities in facial age estimation using the fg-net aging database. In: European Conference on Computer Vision, Springer, pp 737–750
38. Parkhi OM, Vedaldi A, Zisserman A, et al. (2015) Deep face recognition. In: BMVC, vol 1, p 6

39. Redmon J, Farhadi A (2017) Yolo9000: better, faster, stronger. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7263–7271
40. Ricanek K, Tesafaye T (2006) Morph: A longitudinal image database of normal adult age-progression. In: Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on, IEEE, pp 341–345
41. Rothe R, Timofte R, Van Gool L (2016) Deep expectation of real and apparent age from a single image without facial landmarks. International Journal of Computer Vision pp 1–14
42. Sun Y, Wang X, Tang X (2015) Deeply learned face representations are sparse, selective, and robust. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2892–2900
43. Wang X, Guo R, Kambhamettu C (2015) Deeply-learned feature for age estimation. In: 2015 IEEE Winter Conference on Applications of Computer Vision, IEEE, pp 534–541
44. Yang X, Gao BB, Xing C, Huo ZW, Wei XS, Zhou Y, Wu J, Geng X (2015) Deep label distribution learning for apparent age estimation. In: Proceedings of the IEEE international conference on computer vision workshops, pp 102–108
45. Yi D, Lei Z, Li SZ (2014) Age estimation by multi-scale convolutional network. In: Asian conference on computer vision, Springer, pp 144–158
46. Zeiler MD (2012) Adadelta: an adaptive learning rate method. arXiv preprint arXiv:12125701
47. Zhang Y, Liu L, Li C, et al. (2017) Quantifying facial age by posterior of age comparisons. arXiv preprint arXiv:170809687