Unsupervised Facial Biometric Data Filtering for Age and Gender Estimation

Krešimir Bešenić[1], Jörgen Ahlberg[2] and Igor S. Pandžić[1]
[1]*Faculty of Electrical Engineering and Computing, University of Zagreb, Unska 3, 10000 Zagreb, Croatia*
[2]*Computer Vision Laboratory, Linköping University, 58183 Linköping, Sweden*

Keywords: Filtering, Unsupervised, Biometric, Web-Scraping, Age, Gender.

Abstract: Availability 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, datasets with age and gender annotations are scarce and usually limited in terms of size and sample diversity. Web-scraping approaches for automatic data collection can produce large amounts weakly labeled noisy data. The unsupervised facial biometric data filtering method presented in this paper greatly reduces label noise levels in web-scraped facial biometric data. Experiments on two large state-of-the-art web-scraped facial datasets demonstrate the effectiveness of the proposed method, with respect to training and validation scores, training convergence, and generalization capabilities of trained age and gender estimators.

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 processing power enabled researchers to design methods that surpassed human performance on difficult tasks such as image classification (He et al., 2015) and face recognition (Sun et al., 2015). 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 race estimation. For more than a decade, facial age estimation research was based 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. 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.

This paper presents an efficient unsupervised method for biometric data filtering that can significantly reduce label noise in facial image datasets. To best of our knowledge, this is the first completely automatic and parameter-free method for facial dataset filtering that does not require training of dataset-specific systems and utilizes only general purpose, off-the-shelf algorithms. The method was experimentally applied to two large-scale biometric facial datasets, and the results were evaluated on three different biometric estimation tasks (e.g. real age estimation, apparent age estimation, and gender classification).

The rest of the paper is organized as follows. Section 2 reviews important manually collected datasets for age and gender estimation, as well as most relevant automatic web-scraping and dataset filtering methods. Further, Section 3 describes the proposed method for unsupervised biometric data filtering, Section 4 experimentally validates the method's effectiveness on the age and gender estimation tasks and Section 5 briefly concludes the findings of this work.

Bešenić, K., Ahlberg, J. and Pandžić, I.
Unsupervised Facial Biometric Data Filtering for Age and Gender Estimation.
DOI: 10.5220/0007257202090217
In *Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2019)*, pages 209-217
ISBN: 978-989-758-354-4

Table 1: Publicly available age and gender datasets.

Dataset	Images	Subjects	Images/Subject	Labels	Age Range
FG-NET (Panis and Lanitis, 2014)	1,002	82	12.2	age	0-69
MORPH (Ricanek and Tesafaye, 2006)	55,134	13,618	4.1	age, gender	16-77
CACD (Chen et al., 2014)	163,446	2,000	81.7	age	14-62
IMDB-WIKI (Rothe et al., 2016)	523,051	20,284	25.8	age, gender	0-100

2 RELATED WORK

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

2.1 Manually Collected Datasets

Early research on the automatic facial age and gender estimation was conducted on small manually collected datasets often having less than 100 samples (Golomb et al., 1990; **?**).

One of the first publicly available datasets for age and gender estimation was The Face and Gesture Recognition Research Network (FG-NET) dataset. It is a cross-age dataset, consisting of 1002 images from 82 subjects. To collect the dataset, subjects were asked to scan photos of themselves, ranging from photos from their childhoods to their adult lives. Although small in size, this manually collected dataset was a difficult challenge and a stepping stone for early age estimation research, as reviewed by (Panis and Lanitis, 2014).

Another important milestone for the facial age and gender estimation research was the introduction of The Craniofacial Longitudinal Morphological Face Database (MORPH)(Ricanek and Tesafaye, 2006). MORPH is a mug-shot dataset consisting of more than 55,000 images, taken in a correctional facility over a period of 4 years, with annotations for age, gender, and race. Even though the images were collected in highly controlled environment and the dataset has 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 that field.

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 age and gender data, while Table 1 summarises the basic properties of the mentioned 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 by (Jia and Cristianini, 2015). By querying search engines with a list of gender-specific names, they 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, (Hu et al., 2017) proposed a method for web-based collection of samples with age difference labels. To build their dataset, they used Flickr[1] to crawl large amounts of images by the query names from the LFW dataset (Huang et al., 2008) along with descriptions containing dates of image-taking. Although they did not collect the actual age information, pre-training their network for age-difference estimation improved their final real age estimation results.

The Cross-Age Celebrity Dataset (CACD) was the first public large-scale web-scraped facial dataset with age annotations, initially collected for cross-age face recognition by(Chen et al., 2014). Their 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[2]) to find the

[1] www.flickr.com
[2] www.imdb.com

50 most popular celebrities for each birth year from 1951 to 1990, resulting in a list containing 2000 subjects. After the list was created, they used Google Image Search[3] 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 by (Rothe et al., 2016). They 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 metadata. 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(Chen et al., 2014), the authors note that they can not 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 by (Ni et al., 2009). 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 all the images with a large reconstruction error.

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 removed.

Even though the benefits of pre-training on the large and noisy IMDB-WIKI dataset were clearly demonstrated by (Rothe et al., 2016), a cleaned version could further improve their age estimation results. In order to create a cleaned version of the dataset, (Antipov et al., 2016) combined automatic and manual processing steps. 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 average annotator estimates age with high mean absolute error (MAE) of 4.7 - 7.2 years (Han et al., 2013), indicating that even this seemingly trivial step can produce additional noisy outcomes.

Compared to the limited work presented on the 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 (Ng and Winkler, 2014). To identify the faces to be removed from their dataset, they exploited the ob-

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

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

servations 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. Assuming 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 (Cao et al., 2018) adopted and improved a multi-step semi-automatic approach from the original VGGFace paper (Parkhi et al., 2015). To achieve their goal of a 96% pure dataset, their effort 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 the 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.

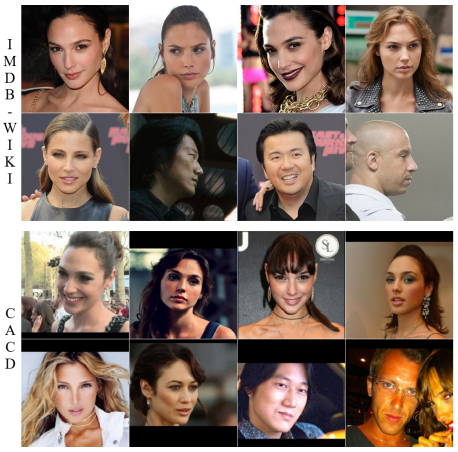

Figure 1: Web-scraping noise in IMDB-WIKI and CACD data for subject Gal Gadot; for each dataset, top row shows valid samples while bottom row shows image samples that have been wrongly paired with Gal Gadot's meta-data.

3 PROPOSED APPROACH

To design an efficient filtering method, we firstly analyse the common sources of label noise in the current state-of-the-art biometric facial datasets.

Due to the nature of commonly used web-scraping approaches described in Section 2.2, there are two main sources of label noise. The first problem is the unreliability of the automatic age annotation process itself. Although the dates of birth are mostly correct, the year of photo-taking can be inaccurate or misleading. As mentioned in(Rothe et al., 2016), a large number of images are actually movie screenshots annotated with the year of the movie release, and some movies have production times spanning over several years. While this problem usually causes relatively small age annotation errors, the following problem can cause large annotation discrepancies for both age and gender labels.

In case of multi-person images, face detector failures or simply bad image search results, collected meta-data can be paired with a face detection of the wrong person. 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 a have wrongly assigned gender and a high age annotation error (e.g. 30 years). Figure 1 shows examples of correctly paired and mismatched images for one subject from CACD and IMDB-WIKI datasets.

In order to reduce the number of samples with large label errors caused by mismatched identities and to mitigate this main source of label noise, we propose a filtering method described in the next section.

3.1 Unsupervised Filtering Method

The main idea of our unsupervised filtering method is to automatically group samples with the same biometric meta-data (images collected for the same person) into clusters of samples with matching identity, and only to keep the samples from the largest cluster, while all other samples get discarded. There are two prerequisites that need to be satisfied for this approach to function:

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

There are a number of clustering algorithms that can perform the required clustering efficiently, but regardless of the type of the automatic clustering algorithm, the clustering performance will greatly depend on the way the samples are 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 favourable option for that task are the facial recognition algorithms. Facial recognition algorithms are specifically designed to project high-dimensional facial image data to highly discriminative low-dimensional feature vectors (i.e. face descriptors) that encode subject's identities. However, our method is not restricted to work on descriptors obtained by facial recognition algorithms - any image descriptor could be used.

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.

For the approach to be completely parameter-less and unsupervised, the descriptor grouping should be done with a clustering method that is capable of discovering the number of underlying groups (identities) automatically. For this purpose, a number of clustering algorithms, such as Chinese Whispers Clustering (Biemann, 2006), Affinity Propagation Clustering (Frey and Dueck, 2007) or Mean Shift Clustering (Cheng, 1995), can be used.

3.2 Implementation

Based on the method described in Section 3.1, we implement a filtering pipeline consisting of several steps. First, the datasets are reorganized into subject-based galleries containing all the images collected for each of the dataset subjects.

The second step amounts to detecting all the faces in the image. Although both datasets provide facial bounding box information, given bounding boxes lack consistency with respect to bounding box scale and positioning. To ensure more consistent inputs to the feature 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 (Bulat and Tzimiropoulos, 2017) to precisely determine bounding box position and scale based on the detected facial landmark points. The bounding box information provided by the datasets' authors is used only in the rare cases of a face detection failure, and even then it is corrected by the face alignment step.

The next step is the extraction of face descriptors for all detected faces. We utilized the dlib's powerful facial recognition model based on ResNet architecture (He et al., 2016) to extract compact 512-dimensional identity descriptors. By calculating distances between the extracted feature vectors, the probability of two descriptors representing the same subject can be efficiently estimated, and by using dlib's default descriptor similarity threshold, a reliable identity matching can be obtained.

The final step is the sample clustering based on the extracted facial descriptors. Since the number of identities in each gallery is unknown, we utilize the Chinese Whispers clustering; an efficient graph-based parameter-free clustering algorithm introduced by (Biemann, 2006) which discovers the number of clusters in a simple iterative process.

3.3 Filtering Results

We apply the proposed filtering pipeline to the two largest publicly available facial biometric estimation datasets; the CACD dataset and the IMDB-WIKI dataset. As we can see from Table 1, the average number of images per subject is 81.7 for the CACD, and 25.8 for the IMDB-WIKI dataset, indicating that the method's first prerequisite from Section 3.1 will be satisfied for the majority of subjects.

The probability of a well-defined image search producing more bad than good results is very low. The probability of a subject not being most frequently

[5] http://dlib.net

appearing person on his/her 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.

Images that were damaged, or had an extremely low resolution or biologically impossible labels were removed in the initial preprocessing step and were not considered in any of the experiments. Additionally, only the IMDB part of the IMDB-WIKI dataset was used considering that only one image per subject was collected for the WIKI part[6].

Prior to the label filtering, the subsets of the CACD and the IMDB-WIKI datasets used in this work had 150,383 and 452,261 samples, respectively. After filtering, 130,571 samples were retained in the CACD dataset (13.2% reduction), and only 216,939 samples remained in the IMDB dataset (52.0% reduction). As we can see in Figure 2, the sample distributions of the unfiltered and filtered subsets of the datasets remained similar, while the number of samples was greatly reduced.

To examine the filtering results more closely, outputs for several galleries were manually inspected and showed consistent results. Figure 3 shows the results of a statistical analysis of filtering outputs for one of the subjects from the IMDB dataset. The figure contains a histogram 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 subject's most popular associates. The emphasised part of the Figure 3

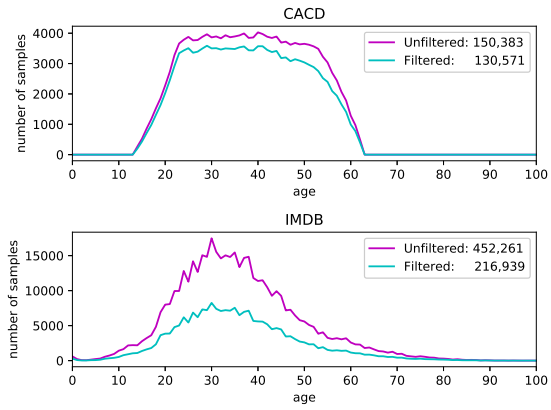

Figure 2: Distribution of samples across ages for CACD and IMDB datasets before and after filtering.

[6]IMDB and WIKI galleries could have been merged by IMDb IDs

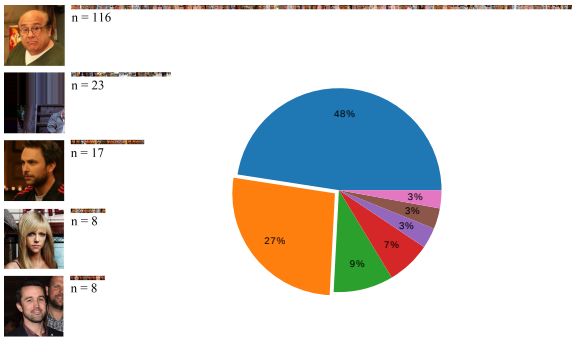

Figure 3: Clustering results for the IMDB subject Danny DeVito; images on the left show representative images for top 5 clusters, histogram bars contain all cluster samples, and chart on the right shows cluster size distribution with emphasised part representing all clusters containing 1 to 3 samples.

chart represents the clusters with only 1-3 samples (1-3 occurrences per identity) therefore grouping the less frequently appearing outliers.

4 EXPERIMENTS

The proposed filtering pipeline described in Section 3.2 resulted with strong sample count reduction, as presented in Section 3.3. 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 dataset's applicability to the facial biometric tasks they were designed for, we performed a set of age and gender estimation experiments.

To compare results before and after the proposed filtering method was applied, we trained age and gender estimation algorithms on both versions of the dataset under identical conditions, and analysed differences in terms of convergence rate, training error, validation error, and testing error. All the networks were trained for 100 epochs on an 80% split of the dataset, with the remaining 20% of samples being used for validation. The training set was augmented with random bounding box perturbations and horizontal flipping.

Good generalization capabilities, crucial for real world in-the-wild applications, often directly depend on the training set sample count and diversity. To validate that our quite aggressive sample reduction does not impair the generalization capacity of the trained estimation systems, we perform testing on separate in-the-wild benchmarks.

To further show that the proposed filtering method is beneficial even in case of highly specialized transfer learning, we performed additional fine-tuning on the training parts of the benchmark datasets.

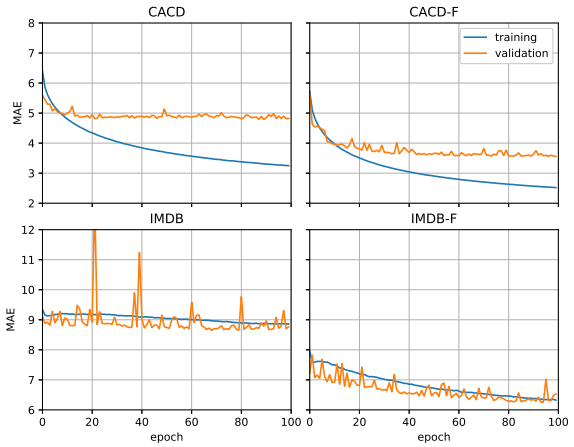

Figure 4: Training and validation MAE for the first 100 epochs of training on the unfiltered and filtered (F) versions of the CACD and IMDB datasets.

The deep learning estimation algorithms used in the experiments rely on a simple 9-layer CNN model based on the open-source architecture Tiny DarkNet[7]. This minimalistic 1M-parameter architecture was pre-trained for the task of facial recognition and further modified to take a low-resolution $3 \times 96 \times 96$ RGB input, allowing efficient training and single-core CPU inference in real time, even on mobile devices.

4.1 Age Estimation Experiments

To train the age estimators, the network architecture was modified to output a single soft value ranging from 0 to 100. To this extent, we added a fully connected layer, applied a sigmoid function to its sole output and multiplied it by 100. To train the network, we used the Mean Square Error (MSE) Loss and the widely adopted Adadelta optimization algorithm (Zeiler, 2012) with the learning rate set to 10^{-1}. To calculate the estimation errors, we adopted the standard mean absolute error (MAE) measure. The trainings were performed on CACD and IMDB datasets separately.

Figure 4 shows training and validation MAEs over 100 epochs for the 4 experiments. It is notable that the trainings based on the filtered data (denoted as CACD-F and IMDB-F) resulted in lower training and validation errors, as well as better general convergence.

Table 2: Appa-Real real age estimation testing (MAE).

Dataset	Fine-tuned	Unfiltered	Filtered
CACD	no	13.73	11.68
IMDB	no	8.94	6.67
CACD	yes	7.30	6.87
IMDB	yes	6.85	6.26

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

Table 3: Appa-Real apparent age estimation testing (MAE).

Dataset	Fine-tuned	Unfiltered	Filtered
CACD	no	12.40	10.40
IMDB	no	7.55	5.51
CACD	yes	5.71	5.32
IMDB	yes	5.57	4.92

The testing was performed on the Appa-Real age estimation benchmark, introduced in (Agustsson et al., 2017). The Appa-Real dataset is a small in-the-wild dataset consisting of 7,591 samples with highly reliable real and apparent age annotations, allowing us to perform testing on the two separate tasks. Real age estimation is the task of estimation of 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. The fine-tunings were performed on the two separate tasks for 500 epochs with SGD optimization and a relatively low learning rate (10^{-5}).

Tables 2 and 3 present the real and apparent age estimation errors on the Appa-Real test-set, respectively. The first two rows show errors for models that are directly trained on the unfiltered and filtered versions of the CACD and IMDB datasets, while the last two rows show errors for models that were fine-tuned on the Appa-Real training set. Note that the high CACD testing errors were caused by the lack of young and old people in the CACD dataset, as shown by Figure 2, and were greatly reduced after fine-tuning on the Appa-Real training set.

For both datasets and both types of age estimation tasks, the models trained on the filtered versions of the data resulted with consistently reduced mean absolute age estimation error by more than 2 years. Even after the specialized fine-tunings, versions pre-trained on the filtered data consistently yielded ~ 0.5 years lower MAE compared to the models based on the unfiltered data, regardless of the type of the age estimation task.

4.2 Gender Classification Experiments

The gender estimator used in the experiments was designed as a simple binary classifier. The base architecture was extended with one fully connected layer producing two softmax outputs, along with an additi-

Table 4: LFW gender classification accuracy testing.

Dataset	Fine-tuned	Unfiltered	Filtered
IMDB	no	96.06	96.71
IMDB	yes	96.36	96.84

onal dropout layer to reduce possible overfitting problems. The networks were trained by employing the Cross-Entropy Loss and Stochastic Gradient Descent (SGD) optimization with learning rate set to $2 \cdot 10^{-2}$. Due to the lack of gender labels in the CACD dataset, the experiments were performed only on the IMDB dataset.

Figure 5 shows training measurements for the unfiltered and filtered (F) versions of the IMDB data. The training and validation classification accuracies for the training based on the filtered data were consistently higher by a large margin of $\sim 17\%$ over all 100 epochs, indicating very high amounts of gender label noise in the unfiltered version of the dataset.

The testing was performed on a version of the LFW datasets aligned by (Huang et al., 2012) with manually verified gender labels provided by (Afifi and Abdelhamed, 2017). The dataset consists of 13,233 in-the-wild images initially collected for the facial recognition testing. All tests were performed on the images of the official test-set subjects to prevent images of the same person appearing in both training and test sets. Similar to the procedure from the Section 4.1, the fine-tunings were performed on the training part of the LFW benchmark dataset for 500 epochs with SGD optimization and learning rate set to 10^{-5}.

Table 4 presents the gender classification accuracy scores on the LFW test-set. The testing accuracy obtained with the gender classifier trained on the unfiltered version of the IMDB dataset was almost 19% higher than the highest validation accuracy reached during the training (77.09%). This interesting result further indicates that the cause of low training and validation accuracies during the training on the unfiltered data was gender label noise since the trained classifier demonstrated good performance on the clean, manually verified LFW test-set.

In the case of simple tasks, such as binary classification, modern deep learning methods can achieve good generalization capabilities despite large amounts of label noise in the training set. However, even in case of one of the simplest facial analysis tasks (i.e. gender classification), the testing accuracy obtained by the classifier trained on the filtered version of the data was notably higher. Even after performing fine-tuning on the manually cleaned LFW training set, the pre-training on the cleaned version of the dataset was shown to be beneficial.

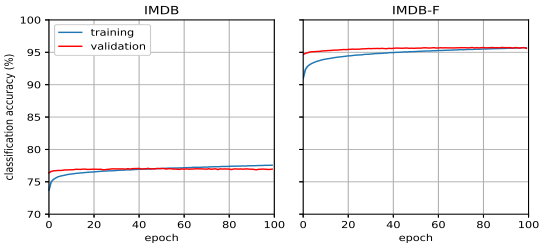

Figure 5: Training and validation gender classification accuracies for the first 100 epochs of training on the unfiltered and filtered (F) versions of the IMDB datasets.

5 CONCLUSIONS

Compared to the manually collected facial datasets for biometric estimation, datasets collected with automatic web-scraping methods can be far superior with respect to the sample count and variety, but share a common downside in terms of label noise. The filtering methods for label noise reduction often require dataset-specific trainings and manual intervention.

The proposed method for unsupervised biometric data filtering, build upon parameterless identity-based clustering, can automatically reduce the number of noisy samples in facial web-scraped datasets by combining only several general-purpose algorithms.

The implemented filtering pipeline resulted with strong sample count reduction (up to 52%) on two state-of-the-art web-scraped facial datasets (i.e. CACD and IMDB). The filtering results were validated by training separate age and gender estimators on unfiltered and filtered data under identical setup. Models based on filtered data demonstrated better convergence rates and better training and validation scores, indicating lower amounts of label noise and improved label consistency, with an additional benefit of shorter training times.

The generalization capabilities of the models trained on the filtered data were shown to be considerably improved by performing testing on separate in-the-wild age and gender benchmarks. In case of age estimation, Appa-Real testing MAE was consistently lowered by more than 2 years for both datasets and two separate age estimation tasks (i.e. real and apparent age estimation). The gender classification accuracy on the LFW test-set was improved by 0.65%, and the large testing-validation accuracy gap ($\sim 19\%$) for the model trained on the unfiltered data further indicated very high amounts of label noise, compared to a gap of only 0.96% in case of the filtered data.

The proposed filtering method was additionally shown to consistently improve results for all 3 biometric tasks even in case of specialized fine-tuning on manually cleaned benchmark train-sets.

REFERENCES

Afifi, M. and Abdelhamed, A. (2017). Afif4: Deep gender classification based on adaboost-based fusion of isolated facial features and foggy faces. *arXiv preprint arXiv:1706.04277*.

Agustsson, E., Timofte, R., Escalera, S., Baro, X., Guyon, I., and Rothe, R. (2017). Apparent and real age estimation in still images with deep residual regressors on appa-real database. In *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*, pages 87–94. IEEE.

Antipov, G., Baccouche, M., Berrani, S.-A., and Dugelay, J.-L. (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*, pages 96–104.

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*, pages 73–80. Association for Computational Linguistics.

Bulat, A. and 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 *International Conference on Computer Vision*.

Cao, Q., Shen, L., Xie, W., Parkhi, O. M., and 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*, pages 67–74. IEEE.

Chen, B.-C., Chen, C.-S., and Hsu, W. H. (2014). Cross-age reference coding for age-invariant face recognition and retrieval. In *European conference on computer vision*, pages 768–783. Springer.

Cheng, Y. (1995). Mean shift, mode seeking, and clustering. *IEEE transactions on pattern analysis and machine intelligence*, 17(8):790–799.

Frey, B. J. and Dueck, D. (2007). Clustering by passing messages between data points. *science*, 315(5814):972–976.

Golomb, B. A., Lawrence, D. T., and Sejnowski, T. J. (1990). Sexnet: A neural network identifies sex from human faces. In *NIPS*, volume 1, page 2.

Han, H., Otto, C., Jain, A. K., et al. (2013). Age estimation from face images: Human vs. machine performance. *ICB*, 13:1–8.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Hu, Z., Wen, Y., Wang, J., Wang, M., Hong, R., and Yan, S. (2017). Facial age estimation with age difference. *IEEE Transactions on Image Processing*, 26(7):3087–3097.

Huang, G., Mattar, M., Lee, H., and Learned-Miller, E. G. (2012). Learning to align from scratch. In *Advances in neural information processing systems*, pages 764–772.

Huang, G. B., Mattar, M., Berg, T., and 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*.

Jia, S. and Cristianini, N. (2015). Learning to classify gender from four million images. *Pattern recognition letters*, 58:35–41.

Ng, H.-W. and Winkler, S. (2014). A data-driven approach to cleaning large face datasets. In *Image Processing (ICIP), 2014 IEEE International Conference on*, pages 343–347. IEEE.

Ni, B., Song, Z., and Yan, S. (2009). Web image mining towards universal age estimator. In *Proceedings of the 17th ACM international conference on Multimedia*, pages 85–94. ACM.

Panis, G. and Lanitis, A. (2014). An overview of research activities in facial age estimation using the fg-net aging database. In *European Conference on Computer Vision*, pages 737–750. Springer.

Parkhi, O. M., Vedaldi, A., Zisserman, A., et al. (2015). Deep face recognition. In *BMVC*, volume 1, page 6.

Ricanek, K. and 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*, pages 341–345. IEEE.

Rothe, R., Timofte, R., and Van Gool, L. (2016). Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision*, pages 1–14.

Sun, Y., Wang, X., and Tang, X. (2015). Deeply learned face representations are sparse, selective, and robust. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2892–2900.

Zeiler, M. D. (2012). Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.