IET Biometrics

Research Article

IET Journals
The Institution of Engineering and Technology

Automated cleaning of identity label noise in a large face dataset with quality control

ISSN 2047-4938
Received on 21st April 2019
Revised 12th July 2019
Accepted on 29th August 2019
E-First on 11th December 2019
doi: 10.1049/iet-bmt.2019.0081
www.ietdl.org

Mohamad Al Jazaery[1], Guodong Guo[1] ✉

[1]*Department of CSEE, West Virginia University, Morgantown, WV 26506, USA*

✉ *E-mail: Guodong.Guo@mail.wvu.edu*

Abstract: For face recognition, some very large-scale datasets are publicly available in recent years, which are usually collected from the Internet using search engines, and thus have many faces with wrong identity (ID) labels (outliers). Additionally, the face images in these datasets have different qualities because of uncontrolled situations. The authors propose a novel approach for cleaning the ID label error, handling face images in different qualities. The face ID labels cleaned by their method can train better models for low-quality face recognition since more low-quality images are correctly labelled for training a deep model. In their low-to-high-quality face verification experiments, the deep model trained on their cleaning results of MS-Celeb-1M.v1 face dataset outperforms the same model trained on the same dataset cleaned by the semantic bootstrapping method. They also apply their ID label cleaning method on a subset of the cross-age celebrity dataset (CACD) face dataset, in which their quality-based cleaning can deliver higher precision and recall than a previous method on detecting the ID label errors.

1 Introduction

Owing to recent advances in deep learning techniques for face recognition, the need for large face datasets with accurate identity (ID) labels has increased dramatically. To build large datasets, researchers typically collect a large number of face images from the Internet. However, these kinds of datasets usually contain ID label ambiguity. Also, the fact of being large scale makes them almost impossible to be cleaned from ID label errors by just taking a manual approach.

Furthermore, these large face datasets are not only filled with wrong ID labels but also have different levels of quality. Low-quality face images with low resolutions in addition to uncontrolled poses and illumination conditions are hard to identify. In the current automated ID label cleaning methods, low-quality faces are likely to be removed because they are less similar to face of higher qualities. Therefore, developing an automated ID label cleaning method, which keeps more inlier low-quality face images, helps in training better face models by providing a diversity of face images. The problem of low-quality face matching and recognition happens very often in real life, where face images are usually captured by surveillance cameras in unconstrained conditions and matched to passport-style high-quality face images in the gallery.

Additionally, many research studies tried to tackle the low-quality face recognition problem by introducing new data-processing methods [1], loss functions [2, 3], robust local face features [4] and model structures [5]. In our work, we believe that improving the process of collecting and cleaning face datasets to include more accurate low-quality face images will improve the low-quality face recognition models. Therefore, we propose a novel approach for reducing ID label errors in a large face dataset with more considerations to the face image quality issue. Our contributions include:

- Developing a novel method to detect ID label errors in a large face dataset using a face image quality assessment (FIQA) which can preserve low-quality face images as inliers.
- An evaluation of the proposed approach in comparison with other representative approaches: indirect comparison to the semantic bootstrapping cleaning approach [5] shows that our method can produce better training data for low-to-high-quality deep face matching. Also, a comparison with human annotations shows that our ID label cleaning approach achieves higher recall and precision on a larger face dataset than Ng and Winkler's [5] dataset.

The remaining of this paper is organised as follows: Section 2 introduces prior face dataset cleaning works and approaches. In Section 3, our novel ID label cleaning for a large face dataset using FIQA is presented with all steps in details. In Sections 4 and 5, we compare our method indirectly to semantic bootstrapping [5] by conducting two low-to-high-quality deep face matching experiments. Also, we compare our ID label cleaning output to human annotations as a direct way to evaluate our method. Finally, concluding remarks and future directions are given in Section 6.

2 Related work

Recently, a number of large face datasets consisting of unconstrained face images have been constructed. During the construction, different methods were applied for correct ID annotation and noise label removal. Ng and Winkler [5] proposed a method to identify the outliers by formulating it as a quadratic programming problem that combines the outputs of an outlier detection classifier and a gender classifier, enforcing visual dissimilarity between the outliers and inliers, while at the same time constrains to at most one face per image to be an inlier. Their results on the FaceScrub dataset show that the method can effectively clean the raw data. To clean the VGG-Face dataset, Parkhi *et al.* [6] first used human annotators to select the identities with over 90% pure images, then removed erroneous faces in each set automatically using the support vector machine (SVM) trained for each ID with the Fisher vector faces descriptor [7]. After that, they removed the near duplicates by clustering the vector of locally aggregated descriptors [8] of the images. Finally, they used human annotators after ranking images within each ID set to decrease the likelihood of being an outlier. Yi *et al.* [9] built a large-scale dataset, which includes about 10,000 identities and 500,000 images, called CASIA-WebFace. They crawled the Internet Movie Database, a well-structured website containing rich information of celebrities, to collect the images. Then, all images were processed by a multi-view face detector. After that, they used a tag-similarity clustering method to clean the dataset. Later on, to illustrate the quality of CASIA-WebFace, they trained a deep convolutional

Table 1 Large-scale face datasets

Dataset	Type	Identities	Images	Cleaning
FaceScrub	public	695	141,130	automated
VGG-Face	public	2622	~2.6M	hybrid
WebFace	public	10,575	494,414	automated
MS-1M-2R	public	79,077	5049,824	automated

In the cleaning method column, automated means there was no human involvement in the cleaning process. The hybrid method means it used a combination of automated and human processing.

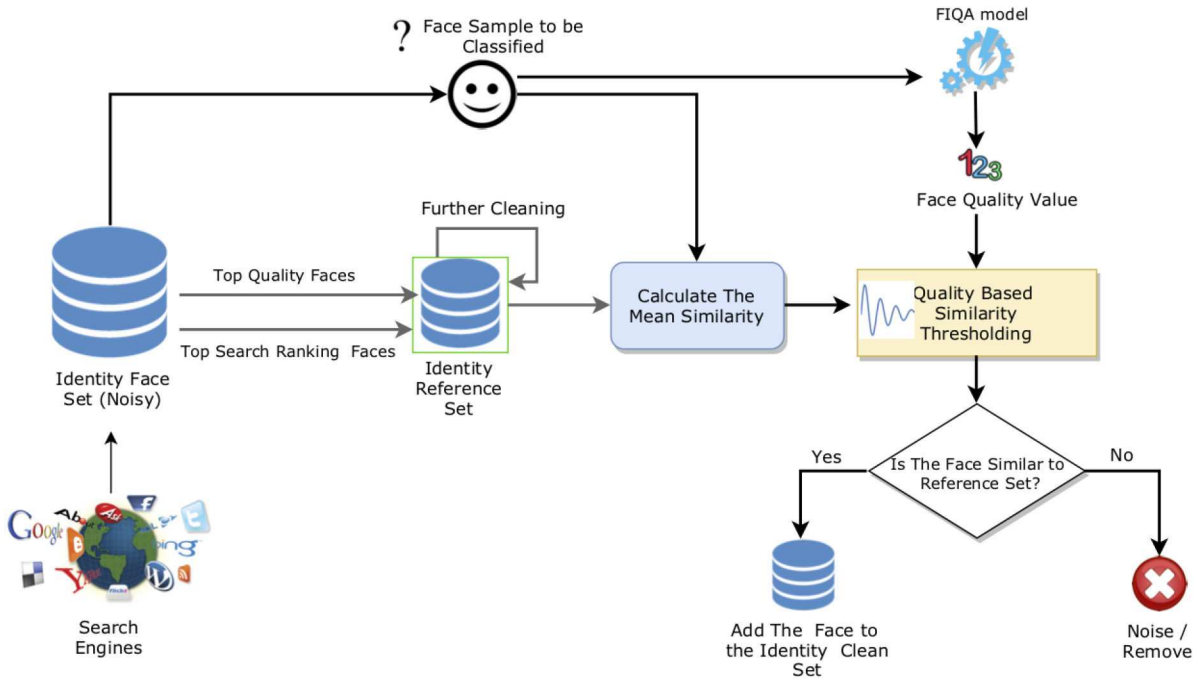

Fig. 1 *Flowchart of the proposed method for cleaning the face images with incorrect ID labels, using the FIQA. First, the method defines a reference set out of the collected noisy face set. Then, using quality-based similarity threshold, decides whether the face image is similar to the reference set. If not, the system considers the face image as a noise; otherwise, the image will be added to the output ID cleaned face set*

neural network (CNN) on it. Wu *et al.* [5] used semantic bootstrapping to clean the MS-Celeb-1M.v1 [10]. First, they trained a light CNN [11] model on the original noisy label dataset. Second, the trained model was utilised to predict the labels. Finally, using a threshold, they decided to whether accept or reject the prediction according to a conditional probability. They called their cleaned result set as MS-1M-2R. Table 1 gives a comparative view of the different cleaned face datasets; some are with manual works.

3 Quality-based cleaning method

The input is a set of ID face images with possible ID label noise. Here, noise means the set includes outlier face images, which do not belong to the assigned ID. The output is the input face images excluding the outliers. The method starts with defining a clean subset from the input set using some preliminary assumptions. We call this set the 'reference set'. Then, a further cleaning is done on this reference set. After that, a 'quality adaptive similarity (QAS) threshold' is applied to decide whether a sample face image is similar to the reference set (inlier) or not. The QAS threshold means using adaptive threshold values for the face images based on their qualities. Since the low-quality inlier faces are likely to achieve a lower similarity score to the reference than the high-quality inlier faces, using a QAS threshold to classify a sample face image may save many low-quality face images from being falsely classified as the noise. Fig. 1 is an illustration of our framework, showing the major steps and components.

We can divide our cleaning method into four steps: (i) constructing an initial reference set based on some preliminary measures, (ii) processing this reference set to become cleaned, (iii) applying the QAS thresholds, and (iv) building the final cleaned set. In the following, we provide explanations of our quality-based cleaning method in details. See Algorithm 1 (see Fig. 2) for a procedural description of the proposed method.

3.1 Defining an initial reference set

Building a reference set, which includes images with a high probability to be inliers helps in classifying any sample from the original noisy ID set S, such that, samples which are similar to the reference set are considered inliers. Thus, we build an initial reference set R_{init} consisting of images that have a high probability of being inliers. First, because the images are collected via a search engine, there is usually a ranking of images based on the search result, and the images with high rankings have a relatively high probability of being inlier images of the ID. On the basis of that, the top three searches ranked images related to the ID are added to R_{init}. Second, because the majority of the high-quality face images are potential inliers, all the images above the mean quality value $\overline{Q}$ are considered high quality and added to the initial reference set R_{init}. The mean quality $\overline{Q}$ is the average of the quality values for all the face images in the dataset. At this point, the ID initial reference set R_{init} contains the top high-quality images, and the top three searches ranked images of that ID.

3.2 Fixing the reference set

To avoid any noise in the initial reference set R_{init}, we estimated a similarity threshold T_{init}, such that, any image in R_{init} that has a similarity measure less than the threshold T_{init} to the remaining set of face images in R_{init} is not added to the final reference set. By excluding any potential outliers, we create the reference R from R_{init}. In other words, the reference set R is a subset of R_{init}, where the image $I \in R_{\text{init}}$ is considered as a good reference only if its

Input : S The identity noisy face image set.
Output: C The identity cleaned face image set.

begin
  // Building the initial reference set.
  $R_{init} \longleftarrow \{top.search.ranked.images(S)\}$
  for $I \in S$ **do**
    /* Add the high quality images only. */
    if $Q(I) > \overline{Q}$ **then**
      $R_{init} \longleftarrow R_{init} \cup \{I\}$
    end
  end
  // Building the final reference set.
  for $I \in R_{init}$ **do**
    /* Exclude possible outliers. */
    if $mean.sim(I, R_{init}) > T_{init}$ **then**
      $R \longleftarrow R \cup \{I\}$
    end
  end
  // Building the final clean set.
  for $I \in S$ **do**
    /* Classify using the Quality Adaptive Similarity Threshold. */
    if $mean.sim(I, R) > T_{QAS}(I)$ **then**
      $C \longleftarrow C \cup \{I\}$
    end
  end
end

Fig. 2 *Algorithm 1: quality-based face ID label cleaning*

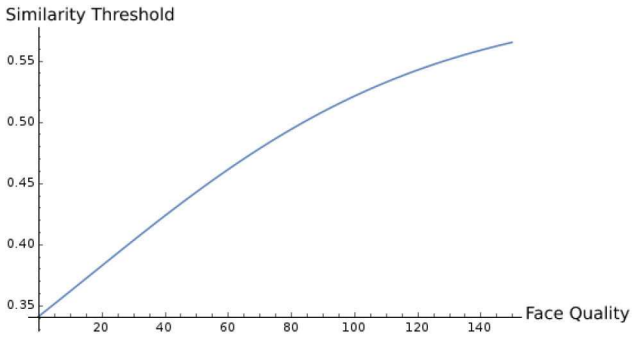

Fig. 3 *QAS threshold function $T_{QAS}(\cdot)$ which is designed for in the MS-Celeb-1M.v1 cleaning experiment. The similarity threshold values increase as the quality of the face images increases*

average similarity to the images in the set R_{init} is above a similarity threshold T_{init}. The ID reference set R is defined as follows:

$$R = \{I | \text{sim}(I, R_{\text{init}}) > T_{\text{init}}, I \in R_{\text{init}}\}, \quad (1)$$

where $\text{sim}(\cdot)$ is the function that calculates the mean similarity between image I and the images in the set R_{init} using a cosine distance measure and T_{init} is an estimated similarity threshold.

3.3 QAS threshold

The QAS depends on the FIQA to determine the best similarity threshold for the sample. The method developed by Chen *et al.* [12] is used to estimate the quality for each face image. It is learning to rank-based FIQA method, which uses the ranking SVMs trained on a rank-ordered set of face images. At first, the rank SVMs learn rank weights for five different image features (histogram of oriented gradients (HoG), Gabor, Gist, local binary pattern (LBP) and CNN features), then the features are fused into a single feature set using a polynomial kernel mapping and another weight vector is learned for the fusion feature. To get the predicted score for an image $\boldsymbol{I}$, the five image features are extracted and multiplied by their corresponding weight vectors, then fused into a second-level feature, and finally multiplied with the fusion feature weight. The quality score is then normalised to a value within the range of 0–100. If $f(\cdot)$ is the function that extracts the feature vector from an image, the quality assessment function $Q(\cdot)$ can be defined as

$$Q(\boldsymbol{I}) = P(\boldsymbol{\omega}^{\text{T}} f(I))\boldsymbol{\omega}' \quad (2)$$

where $\boldsymbol{I}$ is an image, $\boldsymbol{\omega}$ is the learned weight vector for first-level features, $P(\cdot)$ is the polynomial kernel mapping function and ω' is the learned weight for the fused features.

As mentioned earlier, to decide whether a sample is an inlier or not, it should be compared with the ID reference set. To do that, a similarity threshold is needed. Since R has mostly high-quality images, it is highly possible that the low-quality inlier samples will have small similarity scores than the high-quality ones when comparing them to the reference set R. If we try to classify the samples using a strictly high similarity threshold, we could falsely classify the inlier low-quality faces as non-match. On the other hand, if we use a low threshold, it could lead to many outliers to be falsely included.

To solve this dichotomy, an adaptive similarity threshold is proposed, where the threshold goes lower when the image quality is lower. However, the relation between the quality and the similarity threshold is not strictly linear because the threshold is highly affected by the quality of the image in the low-quality range, whereas the range of middle to-high-quality images has less influence by the similarity threshold. On the basis of these facts, we define our QAS threshold function $T_{\text{QAS}}(\cdot)$ as follows:

$$T_{\text{QAS}}(I) = T_{\max} - \frac{(T_{\max} - T_{\min})}{e^{(Q(I)/2\overline{Q})}} \quad (3)$$

where I is a face image sample, $T_{\max}$, $T_{\min}$ are the maximum and minimum allowed similarity thresholds, $\overline{Q}$ is the average quality and $Q(I)$ is the function provided in (2). T_{QAS} threshold changes faster in the low-to-mid-quality range but smoother for the range above the average quality ($\overline{Q}$). Fig. 3 shows one example of how the function $T_{\text{QAS}}(\cdot)$ values change for different quality values.

3.4 Building the final clean set

After calculating the QAS threshold at step 3, the images from the original noisy set that achieves a mean similarity to the reference set R above threshold T_{QAS} are considered as inliers. Therefore, the final ID clean set C is defined as follows:

$$C = \{I | \text{sim}(I, R) > T_{\text{QAS}}), I \in S\}, \quad (4)$$

where S is the noisy ID set, $\text{sim}(\cdot)$ is the function that calculates the mean similarity between the image I and the images in the set R using the cosine distance similarity measure and T_{QAS} is QAS similarity threshold, varying with the face image qualities.

By repeating the method for each ID set in the noisy dataset, we obtain the final cleaned sets for all the identities.

4 Experimental settings

In this section, we will discuss our experimental settings in details before we present the experimental results.

4.1 Dataset description

To evaluate the proposed method, we have chosen the MS-Celeb-1M.v1 dataset [10]. It contains about 100 K identities of 10 million face images. This is a subset of the one million celebrity list collected from a knowledge graph called freebase. Public search engines were used to collect ∼100 images for each celebrity, resulting in about 10M web images. The one million celebrities list includes people with more than 2000 different professions from more than 200 distinct countries/regions. It also covers all major ethnic groups of the world and has a large age range. Some sample images from the dataset are shown in Fig. 4. The outlier faces are

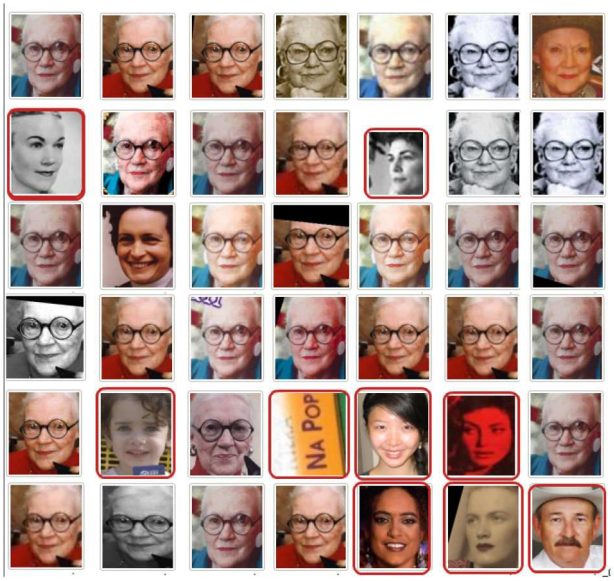

Fig. 4 *Sample images of ID in the MS-Celeb-1M.v1 dataset. Noise images are highlighted by red boxes*

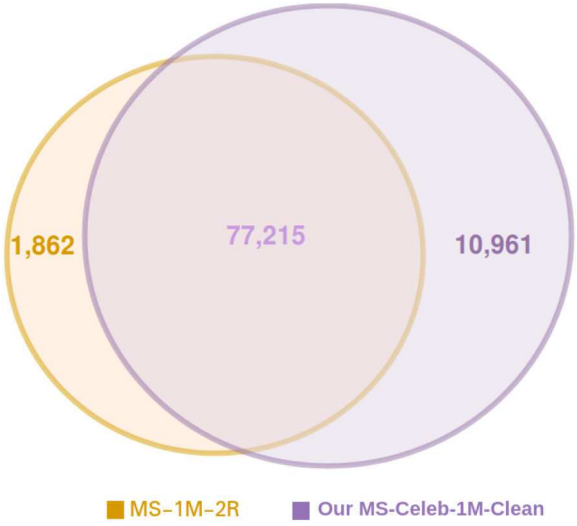

Fig. 5 *Our MS-Celeb-1M-Clean and MS-M1-2R (semantic bootstrapping) ID sets of the MS-Celeb-1M cleaning results (77,215 overlapped identities, 10,961 identities are only in our MS-Celeb-1M-Clean and 1862 identities are only in MS-1M-2R)*

highlighted. We can also see that the collected faces have different qualities in this dataset.

As mentioned in [10], Guo *et al.* do not manually remove noise labels in the dataset because its size is beyond the scope of manual cleaning. In some cases, we found that the outliers for one ID can contain up to five different identities.

4.2 Face detection and alignment settings

All face images are detected, aligned, converted to grey-scale images and normalised into a size of 144×144 for the training data, and 140×140 for the testing data. We use the OpenFace [13, 14] library to detect facial landmarks. The mouth, ears and eyes from detected landmarks are used in the normalisation and alignment process.

4.3 Similarity measure settings

Since our automated cleaning method uses a face matching, we trained a light CNN on CASIA-WebFace dataset using the same settings as in [5]. The momentum is set to 0.9, and the weight decay is set to 5×10^{-4} and the learning rate is set to 1×10^{-3}. The fully connected layer 'eltwise_fc1', which has 256 dimensions, is used to extract deep features. The similarity measure is based on the cosine distance computation.

4.4 MS-Celeb-1M.v1 dataset cleaning settings

Before starting the actual cleaning process, we need to estimate the values of the cleaning parameters which are appropriate to clean MS-Celeb-1M.v1, e.g. the mean quality value ($\overline{Q}$), the initial reference similarity threshold (T_{init}) and both the minimum and maximum thresholds in the quality-based similarity function (T_{min}, T_{max}). The mean quality value ($\overline{Q}$) is the mean quality value of all the images in the dataset and is equal to 54. To find the best values for the other parameters, we defined a validation set of 40 identities. Our validation experiments show that the best clean result is obtained when $T_{\text{init}} = 0.25$, $T_{\text{min}} = 0.34$, $T_{\text{max}} = 0.63$. Fig. 3 shows the function $T_{\text{QAS}}(\cdot)$ with the mentioned settings.

5 Experiments and results

In this section, we discuss the results of the two conducted experiments and show how effective our method is in cleaning the data and preserving the low-quality face images.

5.1 Comparison to the state of the art

Wu *et al.* [5] used a semantic bootstrapping to clean the MS-Celeb-1M.v1. In the following, we compare our MS-Celeb-1M.v1 cleaning results with their published cleaning results.

5.1.1 MS-Celeb-1M.v1 cleaning results: Our final quality-based cleaned set contains 88,176 identities and 4,517,039 face images. The average number of images per ID is 49. From here on, we denote the cleaned version of 'MS-Celeb-1M.v1' as 'MS-Celeb-1M-Clean' dataset. Compared to the semantic bootstrapping cleaning results sets (MS-1M-1R and MS-1M-2R), our method can keep around 10K more identities. Fig. 5 shows the comparison between the number of identities of our MS-Celeb-1M-Clean dataset and the semantic bootstrapping dataset MS-1M-2R. Our MS-Celeb-1M-Clean dataset has 10,961 identities that do not exist in MS-1M-2R. However, 1862 identities in MS-1M-2R could be falsely removed from ours, which means that our method could be improved further to include more identities. Since we limited our method to use light CNN deep architecture and WebFace dataset for a fair comparison with the semantic bootstrapping, using better deep models and more data to train could overcome what looks like some limitations.

Additionally, to show the effectiveness of our method on correctly classifying the low-quality images, we visually compare with the semantic bootstrapping method. Fig. 6 shows examples of identities mainly with low-quality face images, are correctly kept by our method but are falsely considered as noise by the semantic bootstrapping method. We see that there are a number of low-quality faces with blurry, low resolution, large pose change and partially covered faces in these examples.

5.1.2 Low-to-high-quality face verification comparison: To evaluate the usefulness of our cleaned data, especially keeping low-quality face images, we chose the deep network proposed in [5] using the original settings but trained with our cleaned data. Using the original settings helps to make a fair comparison between our model and [5] released model, which is trained with the MS-Celeb-1M.v1 cleaned by the bootstrapping method (bootstrapping model). Since the main goal of our method is preserving inlier low-quality face images, we designed two low-to-high-quality face verification experiments using two different face datasets. In the following sections, we present the experimental details and results on the IJB-A [15] and FaceScrub [5] datasets.

IJB-A experiment: Using the protocol in [16], IJB-A dataset is divided into two subsets based on the face image quality: (i) 10,089 high-quality images and (ii) 362 low-quality images. To perform the low-to-high-quality match experiments, we chose 6676 positive pairs and 3,645,542 negative pairs. Each pair contains one low- and one high-quality images. Our model can obtain 6% higher

Fig. 6 *Some examples from our MS-Celeb-1M-Clean dataset which were falsely classified by MS-M1-2R (semantic bootstrapping) as noise*

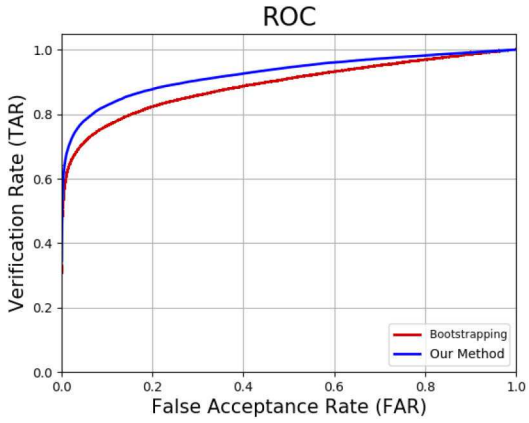

Fig. 7 *ROC comparison on IJB-A low-to-high-quality face verification experiments*

verification rate (VR) at false acceptance rate (FAR) equal to 10^{-3} than the bootstrapping model, where our model achieves 50% VR@FAR = 10^{-3}, while the bootstrapping model achieves 44% VR@FAR = 10^{-3}. The ROC curve comparison is shown in Fig. 7. Additionally, the verification accuracy comparison is given in Table 2. On the basis of the results, our model achieves better face VRs for various FARs on the IJB-A dataset, compared with the bootstrapping model.

FaceScrub experiment: Similarly and using the protocol in [16], the FaceScrub dataset is divided into two subsets based on the quality: (i) 1543 high-quality images, and (ii) 6196 low-quality images. To perform the low-to-high-quality match experiments, there are 18,978 positive pairs and 9541,450 negative pairs. Each pair contains one low- and high-quality images. Similar to the

Table 2 Performance comparison of IJB-A low-to-high-quality face verification experiments

	Bootstrapping [5]	Our model
FAR = 0.001	0.44	**0.50**
FAR = 0.01	0.60	**0.67**
FAR = 0.1	0.76	**0.82**
equal error rate (EER)	0.18	**0.14**
area under the curve (AUC)	0.88	**0.92**

Bold values indicate that the best performance.

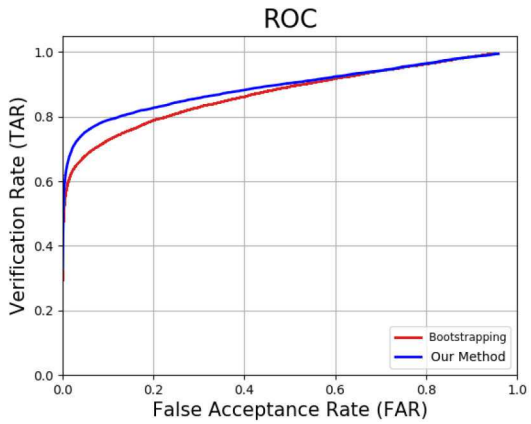

Fig. 8 *ROC comparison on FaceScrub low-to-high-quality face verification experiments*

Table 3 Performance comparison on FaceScrub low-to-high-quality face verification experiments

	Bootstrapping [5]	Our model
FAR = 0.001	0.42	**0.48**
FAR = 0.01	0.58	**0.64**
FAR = 0.1	0.72	**0.78**
EER	0.20	**0.18**
AUC	0.87	**0.89**

Bold values indicate that the best performance.

recognition results on IJB-A dataset, our model can obtain 6% higher VR than the bootstrapping model at FAR equal to 10^{-3}, where our model achieves 48% VR@FAR = 10^{-3}, while the bootstrapping model [5] achieves 42% VR@FAR = 10^{-3}. The ROC curve performance comparison is shown in Fig. 8. Additionally, the verification accuracy comparisons are given in Table 3. Again, our model outperforms the bootstrapping model by achieving better face VRs for various FARs on the FaceScrub dataset.

Since the low-to-high-quality face verification is very challenging compared with other face verification problem, the marginal accuracy improvement achieved by our method is still very valuable. Therefore, we can say that compared with[5], our clean version of MS-Celeb-1M.v1 contains more face variations in terms of the face quality and better training data for low-to-high face matching tasks.

5.2 Comparison with human annotation

Up to now, human labelling is still considered as the best possible annotation and cleaning method, even though it is time-consuming and error-prone to some extent. For this reason, we decided to evaluate our cleaning method using a manually cleaned dataset. Then, we compare the result with another state-of-the-art cleaning method proposed by Ng and Winkler's method [5] that has also been compared with a manually labelled dataset.

Note that, it was not possible for us to perform a direct comparison with Ng and Winkler's cleaning method [5] since their code and the original datasets are not publicly available. So we perform an *ad hoc* study by applying our cleaning method on a manually annotated noisy subset of CACD [17] dataset and

Table 4 Data size and cleaning results for the comparison with the human annotations

Dataset	Number of identities	Number of images	Number of outliers
Ng and Winkler [5]	20	5791	794
ours	500	40,757	6967

Method	Recall	Precision	F1 score
Ng and Winkler [5]	0.72	0.52	0.60
ours	0.76	0.58	0.66

measure the recall and precision. Then, we compare the precision–recall results with their published result. Our argument is that, even though we do not perform our comparative analysis using the same dataset, if we use a much larger dataset with much more noisy labels and still get a better precision–recall curve than theirs, this can indirectly indicate that our method could be better than theirs.

CACD [17] is a large dataset collected for cross-age face recognition in 2014, which includes 2000 identities of 162,815 face images. We manually cleaned a subset of the CACD dataset for our experiment. We chose 500 random identities of 40,757 face images and manually annotated the faces as inliers or outliers. Our manual cleaning found out 6967 outliers in this chosen subset of the CACD.

To clean those 500 identities with our proposed method, we used the same settings as we used to clean MS-Celeb-1M.v1, except we set the maximum and minimum similarity thresholds to higher values, where $T_{\min} = 0.36$ and $T_{\max} = 0.66$. Higher similarity threshold values give better cleaning results since there are higher-quality faces overall in the CACD dataset compared with MS-Celeb-1M.v1.

Our method successfully detected 76% of the outliers (true positive (TP) rate) but removed 11% of the inliers (false positive (FP) rate). Comparing to Ng and Winkler's [5] method, our cleaning method outperforms their reported results in terms of both the recall and precision. Our cleaning results have a recall of 0.76 and precision of 0.58, whereas their method reported 0.72 recall and 0.52 precision. Note that, our test dataset is much larger compared with [5], since their test set contains 5791 face images from 20 people, with 794 of them being outliers. Compared to theirs, our test dataset has 25 times more identities of 40,757 face images with 6967 outliers (Table 4).

6 Conclusion

Cleaning large-scale face datasets have become a major challenge recently. We have presented a novel method for cleaning very large-scale face image datasets using an FIQA scheme. Our method has shown that it can more efficiently solve the ID label noise problem in a large face dataset. Our high-to-low-quality face verification experiments on FaceScrub and IJB-A datasets have shown the effectiveness of our method in face data cleaning by keeping more low-quality face images. Our cleaned version of MS-Celeb-1M.v1 has 10,000 more identities than the bootstrapping-based cleaned version [5] and contains more low-quality faces. Our MS-Celeb-1M-Clean dataset can be released to other researchers. Also, a benchmark analysis of our cleaning method has been performed using a human-annotated test set. Our cleaning results have produced higher recall and precision than a previous method, even working on a much larger dataset.

7 Acknowledgments

This work was partly supported by an NSF-CITeR grant and a WV-HEPC grant. The authors thank Mohammad Iqbal Nouyed for help on face quality assessment.

8 References

[1] Zhao, J., Cheng, Y., Xu, Y., *et al.*: 'Towards pose invariant face recognition in the wild'. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), Utah, USA, 2018

[2] Deng, J., Guo, J., Zafeiriou, S.: 'ArcFace: additive angular margin loss for deep face recognition', *CoRR*, 2018, **abs/1801.07698**, pp. 4690–4699

[3] Wang, H., Wang, Y., Zhou, Z., *et al.*: 'CosFace: large margin cosine loss for deep face recognition'. 2018 IEEE/CVF Conf. Computer Vision and Pattern Recognition, Utah, USA, 2018, pp. 5265–5274

[4] Chen, J., Patel, V.M., Liu, L., *et al.*: 'Robust local features for remote face recognition', *Image Vis. Comput.*, 2017, **64**, pp. 34–46

[5] Ng, H.W., Winkler, S.: 'A data-driven approach to cleaning large face datasets'. 2014 IEEE Int. Conf. Image Processing (ICIP), Paris, France, 2014, pp. 343–347

[6] Parkhi, O.M., Vedaldi, A., Zisserman, A.: 'Deep face recognition'. British Machine Vision Conf., Swansea, UK, 2015, vol. **1**, p.6

[7] Parkhi, O.M., Simonyan, K., Vedaldi, A, *et al.*: 'A compact and discriminative face track descriptor'. 2014 IEEE Conf. Computer Vision and Pattern Recognition, Ohio, USA, 2014, pp. 1693–1700

[8] Arandjelovic, R., Zisserman, A.: 'All about VLAD'. 2013 IEEE Conf. Computer Vision and Pattern Recognition, Oregon, USA, 2013, pp. 1578–1585

[9] Yi, D., Lei, Z., Liao, S, *et al.*: 'Learning face representation from scratch', *CoRR*, 2014, **abs/1411.7923**

[10] Guo, Y., Zhang, L., Hu, Y., *et al.*: '*MS-Celeb-1M: a dataset and benchmark for large-scale face recognition*' (Springer International Publishing, Cham, 2016), pp. 87–102

[11] Wu, X., He, R., Sun, Z, *et al.*: 'A light CNN for deep face representation with noisy labels', *IEEE Trans. Inf. Forensics Sec.*, 2018, **13**, (11), pp. 2884–2896

[12] Chen, J., Deng, Y., Bai, G, *et al.*: 'Face image quality assessment based on learning to rank', *IEEE Signal Process. Lett.*, 2015, **22**, (1), pp. 90–94

[13] Baltrusaitis, T., Robinson, P., Morency, L.P.: 'Constrained local neural fields for robust facial landmark detection in the wild'. 2013 IEEE Int. Conf. Computer Vision Workshops, Sydney, Australia, 2013, pp. 354–361

[14] Baltrušaitis, T., Robinson, P., Morency, L. P.: 'OpenFace: an open source facial behavior analysis toolkit'. 2016 IEEE Winter Conf. Applications of Computer Vision (WACV), New York, USA, 2016, pp. 1–10

[15] Klare, B.F., Klein, B., Taborsky, E., *et al.*: 'Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus Benchmark A'. 2015 IEEE Conf. Computer Vision and Pattern Recognition (CVPR), Boston, USA, 2015, pp. 1931–1939

[16] Guo, G., Zhang, N.: 'What is the challenge for deep learning in unconstrained face recognition?'. 2018 13th IEEE Int. Conf. Automatic Face Gesture Recognition (FG 2018), 2018, pp. 436–442

[17] Chen, B.-C., Chen, C.-S., Hsu, W.H.: '*Cross-age reference coding for age-invariant face recognition and retrieval*' (Springer International Publishing, Cham, 2014), pp. 768–783