IDIAP RESEARCH REPORT

EIGHT YEARS OF FACE RECOGNITION RESEARCH: REPRODUCIBILITY, ACHIEVEMENTS AND OPEN ISSUES

Tiago de Freitas Pereira Dominic Schmidli
Yu Linghu Xinyi Zhang Sébastien Marcel
Manuel Günther

Idiap-RR-09-2022

AUGUST 2022

Centre du Parc, Centre du Parc, Rue Marconi 19, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

Eight Years of Face Recognition Research: Reproducibility, Achievements and Open Issues

Tiago de Freitas Pereira,[1] Dominic Schmidli,[2] Yu Linghu,[2]
Xinyi Zhang,[2] Sébastien Marcel,[1,3] and Manuel Günther[2]

[1]Idiap Research Institute [2]University of Zurich [3]University of Lausanne

Abstract

Automatic face recognition is a research area with high popularity. Many different face recognition algorithms have been proposed in the last thirty years of intensive research in the field. With the popularity of deep learning and its capability to solve a huge variety of different problems, face recognition researchers have concentrated effort on creating better models under this paradigm. From the year 2015, state-of-the-art face recognition has been rooted in deep learning models. Despite the availability of large-scale and diverse datasets for evaluating the performance of face recognition algorithms, many of the modern datasets just combine different factors that influence face recognition, such as face pose, occlusion, illumination, facial expression and image quality. When algorithms produce errors on these datasets, it is not clear which of the factors has caused this error and, hence, there is no guidance in which direction more research is required. This work is a followup from our previous works developed in 2014 and eventually published in 2016, showing the impact of various facial aspects on face recognition algorithms. By comparing the current state-of-the-art with the best systems from the past, we demonstrate that faces under strong occlusions, some types of illumination, and strong expressions are problems mastered by deep learning algorithms, whereas recognition with low-resolution images, extreme pose variations, and open-set recognition is still an open problem. To show this, we run a sequence of experiments using six different datasets and five different face recognition algorithms in an open-source and reproducible manner. We provide the source code to run all of our experiments, which is easily extensible so that utilizing your own deep network in our evaluation is just a few minutes away.

1 Introduction

Biometric recognition has attracted much attention in the past decades. Commonly used examples of biometric recognition include methods of recognizing one's face, iris, voice, ear, palm print, gait, or signature [3]. Face recognition is one of the most popular forms of biometric recognition and its development has made great progress in the last decades, mainly influenced by the availability of different open-source methods for face processing, including face and facial landmark detection and face recognition [4]. Furthermore, its field of application is very versatile, as almost every mobile device, including laptops and smartphones, nowadays offers the possibility to unlock its screen through face recognition. Another popular application is video surveillance and through security cameras [5] where face recognition can help to identify criminals or find missing persons. In these and many other fields, the need for robust facial recognition systems has increased year over year [6] and already in 2007 automatic face recognition has superseded human performance in controlled and constrained environments [7]. Thus, in security-relevant applications such as automatic border control, frontal faces, neutral expressions, and good illumination are enforced [8]. However, such an environment can not always be found. Especially in outdoor surveillance situations, illumination from the sun is often not ideal for recognizing faces and people may show different expressions and will likely not look into the camera [9]. Furthermore, subjects may wear hats or glasses, faces might be partially occluded, and the quality and size of the image can vary greatly [10]. All these conditions can seriously interfere with the performance of face recognition. Before the era of deep learning, face recognition employed hand-crafted features and traditional algorithms. When conditions for capturing faces are not optimal, such as when there are different

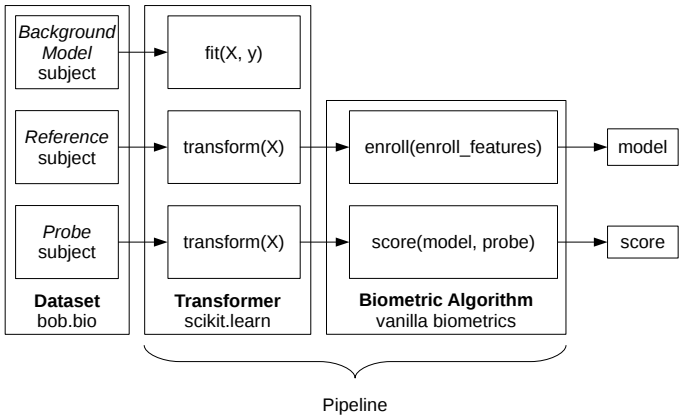


Figure 1: Biometric Recognition Pipeline in Bob. *The dataset implements the evaluation protocol, i.e. which samples to use for enrollment and probing. Each subject's data goes through the Transformer (or series of Transformers) to extract features, which are then given to enroll models or to compute similarity scores between model and probe features.*

facial expressions, lighting conditions or face poses, the performance of these traditional algorithms drop significantly [1]. The biggest issue of hand-crafted features is that they fail to capture important information from the faces, and traditional algorithms are not well-suited for all different aspects of face recognition, particularly, for changes in illumination and pose.

This changed with the development of convolutional neural networks, the Compute Unified Device Architecture (CUDA), which allows data-intensive training of deep networks [3], and libraries that easily put all these elements together, such as Theano, Caffe, MxNet, Tensorflow, PyTorch, PaddlePaddle and so on. Since then, face recognition has been dominated by neural networks, and performance has steadily improved since deep networks automatically learn the best suited features to extract from the images, as well as provide algorithms to learn the different aspects of face recognition in an integrated way. While researchers first focused on creating deeper networks with the support of advanced network architectures, the attention has shifted toward creating more powerful loss functions that are better adapted to face recognition requirements, i.e., to increase between-class and decrease within-class variability of the deep features extracted from the networks. One of the methods that has demonstrated a state-of-the-art performance is ArcFace [11], which uses an additive angular margin loss function to separate the features. Latest inventions in face recognition include MagFace [12], which include metrics for image quality into training deep networks.

Another main area of research is the creation of datasets, which are not only important for training, but also essential for performance comparison. Many papers and several surveys [3–6, 10, 13] show the performance of different deep neural networks in challenging environments that achieve impressive results on the LFW [14] and IJB-C [15] datasets. Unfortunately, these datasets include only mixtures of different aspects of face recognition, but it is impossible to evaluate which of these aspects can be considered to be solved and which aspects require more research focus. Only in rare cases, aspects such as ethnicity [5] or aging [13] are evaluated separately. Also, even though the title of the latest review [4] suggests to review open-source face recognition **frameworks**, the authors only investigate different **parts of the face recognition pipeline** (different datasets, face and facial landmark detection methods, deep learning models and evaluation techniques), but they do not test combinations of different parts since no **framework** for face recognition is available and maintained – except for the Bob framework [16–18] utilized and advertised in this work. It is important to note that those surveys only duplicate the results promised by the reviewed papers **without having a chance to reproduce these numbers or to change evaluation criteria**.

The present work aims to close this gap by comparing the performance of state-of-the-art deep neural networks in different challenging face recognition environments in a reproducible, comparable and extensible manner. To achieve this, four pre-trained networks from the state of the art are examined. Experiments are performed on six datasets, including AR face [19], Multi-PIE [20], SCface [21], GBU [22], IJB-C [15], and MOBIO [23], each of which represents a different aspect of face recognition. Our implementations of evaluation protocols [1, 2] allow the isolated consideration of different aspects of face variations, such as different types of occlusion, facial expressions, or poses. The performed experiments make use of the open-source biometrics recognition pipeline [18] of Bob [16, 17]. Together with this study, an open-source implementation to re-run (or at least re-evaluate) the experiments is provided.[1]

This work extends our previous work that was executed in 2014 and finally published in 2016 [1] and 2017 [2], in which we evaluated the performance of several traditional algorithms on different datasets and, thereby, considered individual challenging face recognition conditions in isolation. In that study, we found that strong occlusion has a significant impact on recognition rates. For

[1] https://gitlab.idiap.ch/bob/bob.paper.8years

face recognition across pose, especially with angles deviating beyond ±45° from frontal, traditional algorithms almost completely fail. Also, facial image comparison including low-resolution probe images poses a major challenge to those algorithms.

The purpose of the current study is to extend and update our previous open-source package[2] in order to see how large the progress of face recognition for various aspects of face recognition has been in the last eight years. Therefore, we compare the best traditional open-source systems, i.e., the inter-session variability (ISV) modeling [1,24] with current state-of-the-art open-source algorithms found today.

The contributions of this work are:

- We evaluate current state-of-the-art deep-learning face recognition algorithms on *various aspects of face recognition separately* and compare them with the best traditional face recognition algorithm.
- In comparison to our old evaluation, we change to a better-suited evaluation procedure that better *accounts for requirements of real-world applications.*
- We show that even extreme facial expressions and large occlusions of the faces are handled well by deep learning algorithms, but *research should focus more on comparing faces across pose, on low-resolution images* and on open-set identification.
- We run all experiments in an *open-source and reproducible manner*, and provide tools to *easily extend this research* to novel future findings and developments.

2 Related Work

This section provides an overview of the current state of research regarding face recognition and deep learning. First, available face datasets are described. Other than a few popular datasets, the focus is put on the datasets utilized in our experiments. More datasets and algorithms can, for example, be found in [4]. The next section gives a brief summary of our previous evaluation [1,2], on which this study is based, and presents the current state of the art in deep learning for face recognition.

2.1 Datasets

A major research interest in the area of deep learning lies in the development of new datasets. There are a large number of facial image datasets that differ greatly in the number of images and identities, as well as in the diversity of the images. While older datasets have often been split into parts for training and for evaluation of algorithms, data-hungry deep learning methods require more data to train and, therefore, training and evaluation datasets have been disentangled. This section gives an overview of some commonly used datasets for training and evaluating face recognition models.

[2] https://gitlab.idiap.ch/bob/bob.chapter.frice

Table 1: PREPROCESSING. *Summary of the employed eye and mouth positions for the experiments. All parameters are given in Bob's `(y,x)` order. Right and left eye coordinates were used for frontal faces, eye and mouth for profile faces.*

	Facenet	**ArcFace-100**, **Zoo-AttentionNet** and **Idiap-Resnet50**	ISV
Resolution	160 × 160	112 × 112	80 × 64
Right Eye	(32, 39)	(52, 38)	(16, 15)
Left Eye	(32, 120)	(52, 74)	(16, 48)
Eye	(32, 64)	(52, 56)	(16, 25)
Mouth	(106, 64)	(91, 56)	(52, 25)

2.1.1 Training Datasets

The Visual Geometry Group (VGG) in Oxford developed a five-step guide to compile a large dataset, and applied these instructions to images of the celebrities in the Internet Movie Database (IMDB). The VGGFace dataset [25] consists of over 2.6 million images from 2'622 different celebrities, with about five percent of these images showing profile faces, the others being mostly frontal. The extended VGGFace2 [26] dataset consists about three million images of 9'131 identities with facial images varying in pose, background, age and illumination, yet all images are of relatively high resolution.

CASIA-WebFace [27] features about half a million images from 10'000 identities. It is also often used for face verification and face identification. The images were collected from celebrities of various years of birth. MS-Celeb-1M (MS1M) [28] is a dataset with ten million images of celebrities collected from the Internet representing a variety of nationalities and professions such as politicians, actors, writers, and singers. It consists of 100'000 identities in total with about 100 images per identity, with over three quarters of the subjects being female. Several researchers released extensions of this dataset, most of them handling some mislabeling issues that a dataset of such size always contains [11,29]. More recently, the WebFace260M dataset was released [30].

This dataset is currently the largest public face recognition dataset. It is composed of noisy 260M faces of 4M identities and a cleaned version composed of 42M faces and 2M identities.

2.1.2 Evaluation Datasets

While there is a plethora of old and small-scale face datasets for evaluation, we here focus only on the ones that suit our purpose best. More datasets can, for example, be found online.[3]

With about 3312 images taken of 76 males and 60 females, AR face [19] is a relatively small dataset, but it is still in use today due to its unique face variations. The images vary in facial expressions, illumination, and occlusion in the form of scarves and sunglasses. Multi-PIE [20] contains about 755'370 images shot in four sessions from 337 different subjects, covering 15 different camera view points, 19 different lighting conditions, and 7 distinct facial expressions. SCface [21] contains 4160 images from 130 subjects taken by five video surveillance cameras of different qualities that were installed slightly above the head position. Pictures of the participants were taken from three different distances where the smallest faces were just about 20 pixels in height. The MOBIO dataset [23] includes facial videos, images and speech recordings of 152 people taken with mobile devices over 12 different recordings. This dataset is of particular interest since the view-point and background seen in the recordings are different from the default forward-facing images, and it provides two gender-dependent evaluation protocols.

Besides these datasets that allow to investigate different aspects of face recognition (occlusion, expression, and pose), other larger datasets are often used for evaluation. The Good, the Bad & the Ugly (GBU) dataset [22] consists of 8'638 frontal images from 782 different identities. It provides three protocols that mainly evaluate different illumination conditions called Good, Bad, and Ugly, where Ugly is the most difficult protocol, while Good is the easiest. The IARPA Janus Benchmark C (IJB-C) [15] is currently the most widely used benchmark for face recognition. IJB-C has the highest diversity in occlusion, occupation, and geographic origin, and image quality to better represent as much of the world's population as possible. The dataset consists of a total of 31'334 images and 11'779 videos of 3'531 identities.

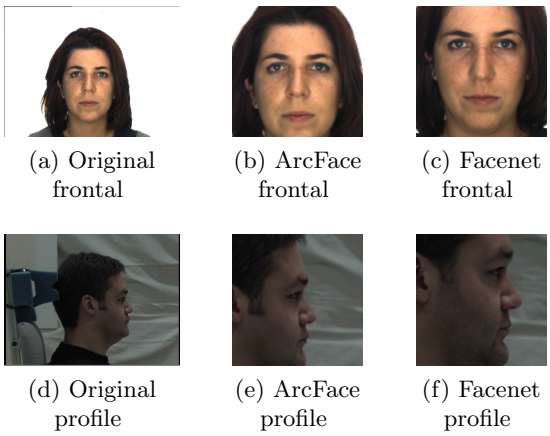
[3] http://face-rec.org/databases

2.2 Algorithms

Before deep learning, face recognition was accomplished through traditional face recognition algorithms. In [1,2], we surveyed and evaluated the performance of several traditional face recognition methods, where we used the open-source software Bob [16] and also published our code.[2] We showed that most of the traditional algorithms worked relatively well in good conditions but failed strongly when differences in facial expressions, illuminations, occlusions, and poses were evaluated. The algorithm with the highest stability against most of these factors was found to be Inter-Session Variability (ISV) modeling [24].

In recent years, deep learning has dominated and revolutionized the field of face recognition so that current face recognition surveys and reviews are full of deep learning methods. These algorithms have advanced face recognition to a level that traditional methods can no longer reach [10]. There are two main research directions in the academic community that have tried to improve the performance of neural networks, especially in unconstrained face recognition environments: the engineering of new network topologies and the definition of new loss functions. Early versions of deep face recognition systems were using only a few convolutional layers, e.g., the VGGFace network [25] used 13 such layers. One of the most relevant contributions in terms of network topologies is the Residual Network [31], which introduced residual connections between layers that allow training much deeper network structures than it was able before, the most common topologies have 18 to 152 layers. Another architecture is the Squeeze and Excitation network [32], which integrates a special block into current network architectures that allows for automatic weighting of individual convolution channels. Additionally, so-called lightweight network architectures have recently been developed. One of these is the MobileNet [33], which uses depth-wise separable convolutions and neural architectural search to lead to a considerable reduction of parameters and, therewith, reduced the computing requirements compared to other networks with similar depth.

The most common loss function employed in classification tasks is the categorical cross-entropy loss used in combination with softmax activation, which is often called SoftMax loss. The basic hypothesis of this loss is that the final embeddings (aka. deep features) that result from this closed-set end-to-end training are sufficiently discriminative for open-set problems, i.e., when subjects from the test datasets differ from the people used dur-



(a) Original frontal (b) ArcFace frontal (c) Facenet frontal

(d) Original profile (e) ArcFace profile (f) Facenet profile

Figure 2: PREPROCESSING EXAMPLES. *This figure shows for each network some preprocessed example images from the AR face and Multi-PIE datasets.*

ing training. Several extensions on top of this basic hypothesis were created over the years, including Center loss [34], which works in conjunction with the SoftMax loss to minimize the within-class distance of embeddings by learning a center for each class and penalizing the distance of features to the corresponding center [35].

Large-margin loss [36], also known as L-SoftMax, was one of the first loss functions that extended SoftMax with an angular margin, which was finally employed for face recognition by SphereFace [37]. Later, CosFace [38] extended this loss to use cosine similarity instead of angular losses and, finally, ArcFace [11] introduced an additive angular margin to both maximize intra-class similarity and inter-class diversity. The big advantage of this margin is that it allows some similarity between faces of different people and does not force all of them to be as dissimilar as possible. Latest loss functions include AdaCos [39], PS2Grad [40], ring loss [41], and MagFace [12]. Also, loss functions that explicitly tackle the issue of open-set face recognition have been proposed [42], but those techniques will not be discussed further in our study. Other losses that are worth mentioning are Triplet [43] and Contrastive Loss [44], which are metric learning approaches that work directly on the embedding space by explicitly minimizing within-class and maximizing between-class variability.

3 Evaluation procedure

The experiments described in the present work all rely on the software Bob [16, 17], an open-source signal processing and machine learning toolbox. Particularly, we make extensive use of its Biometric Recognition Pipelines [18], which are *easily extensible to use new face recognition algorithms based on deep learning* and allow an easy way of reproducing[4] experiments [17]. In our current evaluation we will make use of the newly added interfaces for running experiments with deep networks [45] and the work of [46].

Fig. 1 illustrates the three different steps in the biometric recognition process of Bob.[5] The biometric **Dataset** stores all information required to run a biometric recognition process, which are the original images and their identity labels, facial landmarks used for alignment, and the evaluation protocol that defines which images should be compared. The **Transformer** is essentially a scikit-learn Pipeline[6] containing a sequence of steps to process a sample. Such a pipeline can assume a different sequence of steps depending on the biometric algorithm. In the case of face recognition it is usually composed of a face and facial landmark detector, face alignment, and a feature extractor. Since in this work we are evaluating face recognition algorithms and not facial landmark detectors, we replace the face detector by using hand-annotated landmarks for the alignment step in our experiments. Finally, the **Biometric Algorithm** has functions to enroll a client (create a biometric template) and compute a similarity score between a given template and probe sample. When several images are used for enrollment, the simple average of the embeddings is computed.

3.1 Evaluated Algorithms

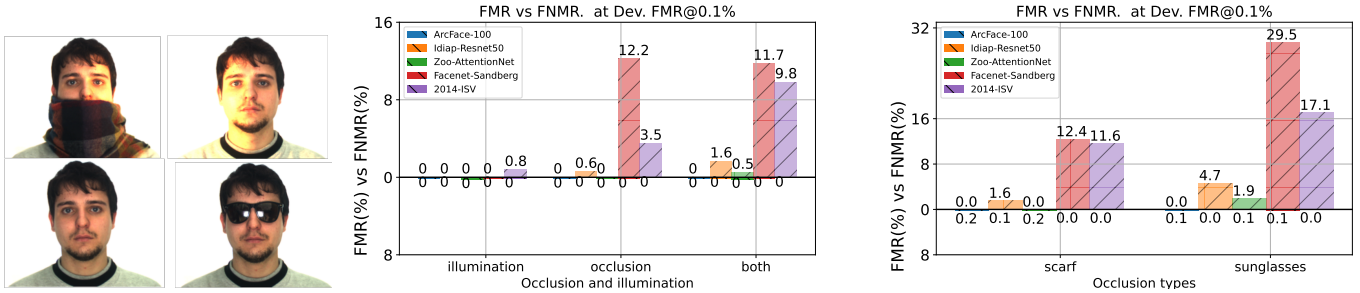
Bob's face recognition package[7] has more than 30 different face recognition systems available (including traditional methods based on hand-crafted features, as well as many modern deep-learning algorithms and pre-trained networks) ready to be used. Because of page limits, in this work we will make use of four different deep-learning-based face recognition systems available in Bob. However, scores from all the available systems will be available.[1] The four systems are the following – in chronological order of publication: The first system is the **Facenet-**

[4] We are aware that some of the employed datasets are no longer publicly available – and we are not allowed to share the data ourselves – which limits the reproducibility of some of our experiments. We provide the resulting recognition scores for further evaluation.

[5] https://www.idiap.ch/software/bob/docs/bob/bob.bio.base/v5.0.0/biometrics_intro.html

[6] https://scikit-learn.org/stable/modules/generated/sklearn.pipeline.Pipeline.html

[7] http://gitlab.idiap.ch/bob/bob.bio.face



(a) neutral, occlusion, illumination

(b) Effect of illumination and occlusion

(c) Effect of different occlusion types

Figure 3: Partial Occlusion. *Image examples of the AR face dataset and the effect of partial occlusion of the face on the tested algorithms. False Match Rates (FMR%) and False Non-Match Rates (FNMR%) on the* eval *set are computed based on the threshold at FMR=0.1% on the* dev *set.*

Sandberg[8] trained using a pruned version of the MS-Celeb-1M dataset using the Inception-ResNet-v1 backbone [43]. The second network is the ArcFace model [11] from InsightFace (**ArcFace-100**) that is based on ResNet-101 architecture and trained using the ArcFace loss. The third network is taken from the FaceX-Zoo models [47], which contain several face recognition systems which are all integrated in Bob; in this work we have used the **Zoo-AttentionNet** backbone. The forth method is based on ResNet-50 architecture trained using the ArcFace loss (**Idiap-Resnet50**) on a pruned version of the MSCeleb-1M dataset. Finally, to show the improvement made over the last eight years, we also provide results of the top-performing method of our old evaluation, i.e., the Inter-Session Variability (ISV) modeling of Discrete Cosine Transform (DCT) features [24].

3.2 Image Preprocessing

Preprocessing plays a crucial role in the face recognition process as it can affect the performance of feature extraction networks [48]. For all datasets used in this work – except for IJB-C – hand-labeled facial landmarks are available, which can be used to align faces directly according to the desired size and the location of these landmarks in the target images. Unfortunately, many research papers lack detailed information on how preprocessing is done, making it particularly difficult for others to reproduce experiments. For ArcFace, only the required input image dimension of 112×112 is provided [11]. There were also some scripts to align faces based on landmarks detected with a MTCNN [49], but our hand-labeled landmarks do not correspond to those extracted by MTCNN and, thus, it was not entirely clear how to achieve alignment. Six pictures, which indicate some kind of alignment, could be taken from their GitHub[9] repository, which we used to manually estimate average landmark locations in the target images. In a similar manner, face images are cropped to 160×160 for Facenet.

The alignment for almost all experiments in this work has been done based on eye landmarks.[10] The experiments on the Multi-PIE dataset required an additional alignment point since the images do not always provide two visible eyes [20]. In these cases, the visible eye and the respective corner of the mouth served as a reference point. Defining these landmarks was even more cumbersome and required trail-and-error executions of algorithms [48]. While our experiments indicate that these landmarks work approximately well, there is no guarantee that these are the best landmark locations to be used. Examples of preprocessed images can be found in Fig. 2, while exact landmark locations in the aligned images are given in Tab. 1.

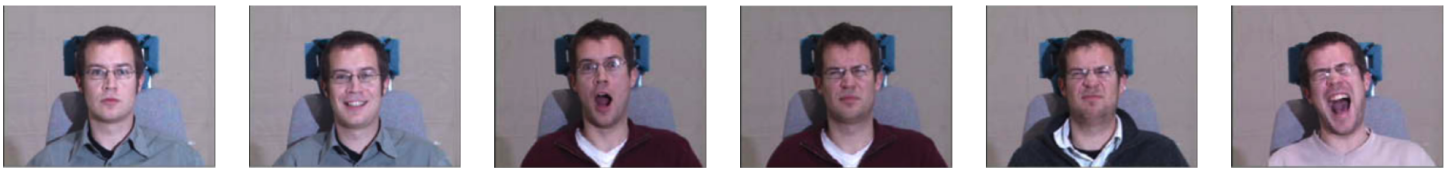
3.3 Evaluation Criteria

With the computed similarity scores from our face recognition pipeline, several evaluations are possible. To evaluate verification protocols, usually the False Match Rate (FMR) and the False Non-Match Rate (FNMR) are computed based on a certain similarity score threshold. By varying this threshold, Receiver Operating Characteristics (ROC) can be plotted and compared. Unfortunately, however, ROCs have the issue that they are computed over the test set and any threshold selected on an unseen

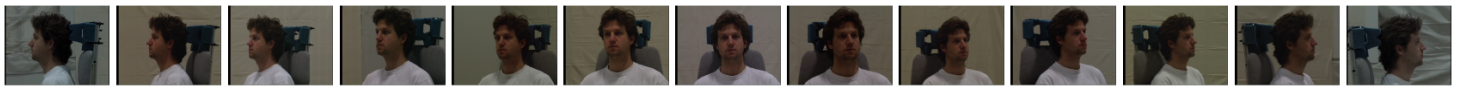
[8] https://github.com/davidsandberg/facenet

[9] https://github.com/deepinsight/insightface

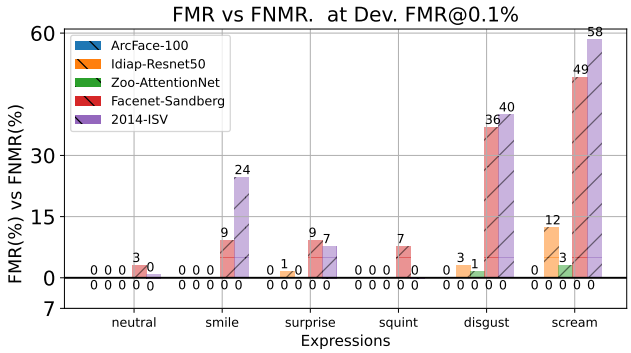
[10] See our face alignment guide



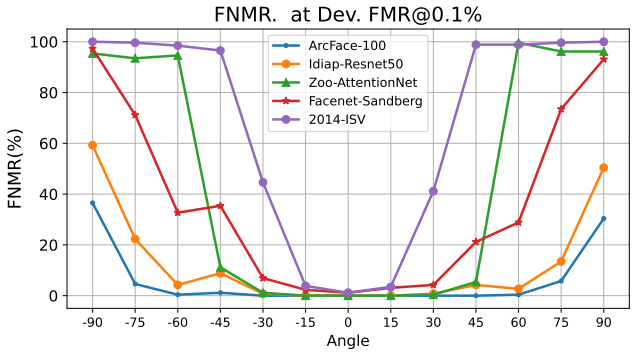
(a) Facial expressions examples: neutral, smile, surprise, squint, disgust, scream



(b) Poses examples from -90 to 90 in steps of 15 degrees



(c) Effect of different expressions



(d) Effect of different poses

Figure 4: Expression and Pose. *This figure shows the effect of different expressions and poses of the face on the tested neural networks. Example images for facial expressions and poses are displayed in subfigure (a) and subfigure (b).*

set of images is not guaranteed to perform equally well in practice.

For this reason, unless the datasets provide different standard evaluation procedures, we have split all of our evaluation protocols into two groups of non-overlapping identities. The first group *dev* is used to compute a threshold based on some criteria, and this threshold is then applied to the second group *eval* to compute the final performance metrics. In our previous study [1, 2], we mainly computed the threshold based on the Equal Error Rate (EER) on *dev* and reported Half Total Error Rates (HTER) on *eval*. Since most security-relevant applications of face recognition want to assure a very small risk of imposters being recognized as genuine, more reasonable thresholds are rather computed for low FMR values. Therefore, in our current evaluation we have switched to compute the threshold based on an FMR of 0.001 (or 0.1 %) on the *dev* set. Notably, when running an evaluation with several sub-protocols (Sec. 4) we compute a **single** threshold for the **combined** scores over all protocols on the *dev* set. Finally, we report both FMR and FNRM on the *eval* set for each sub-protocol separately.

Open-set identification systems are evaluated using the Detection and Identification Rate Curve [50], which is also called the Open-Set ROC and is the standard metric in NIST evaluations.[11] For consistency, we will use the NIST terms and evaluate the False Positive Identification Rate (FPIR) and True Positive Identification Rate (TPIR) at rank 1 based on a certain similarity threshold. By varying this threshold, the TPIR can be plotted over the FPIR. Note that the closed-set identification performance can be obtained at FPIR=1, i.e., on the right-hand side of the plot. While we are aware that this measure has the same deficits as the verification ROC discussed above, we leave the development of better-suited open-set evaluation metrics for future work.

4 Experiments

In the following, we present the results of all face recognition experiments performed using the four networks **Facenet-Sandberg**, **ArcFace-100**, **Zoo-EfficientNet**, and **Idiap-Resnet50**. Furthermore, we have included experiments from the best overall baseline [24] from our previous publications [1, 2], which we

[11] https://www.nist.gov/programs-projects/face-recognition-vendor-test-frvt-ongoing

have marked as **2014-ISV**. The scores from other systems evaluated in our old study are also available for comparison.[12] In this way, we have a perspective on which fronts current state-of-the-art systems improved.

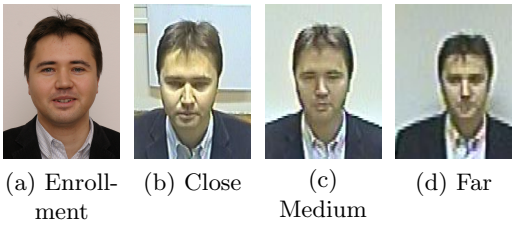
4.1 Face Variations

First, the algorithms are tested against three types of face variations, more precisely partial occlusion, different expressions and poses. In all evaluated datasets in this section, gallery templates are enrolled from neutral faces, i.e., images with neutral frontal illumination showing a face in frontal pose and with neutral facial expression. On the other hand, probe images are equipped with one of the above-mentioned variations. This assures that recognition does not happen because gallery and probe share the same variation, but only because the algorithm is able to ignore the variation and still can recognize the person.

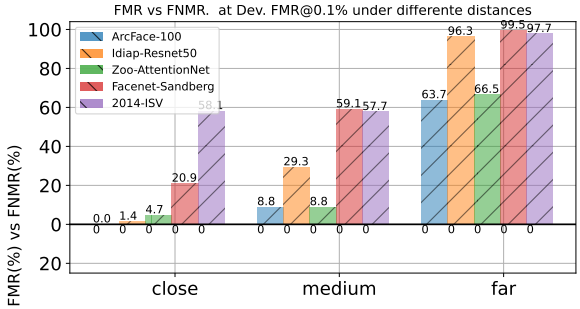
4.1.1 Partial Occlusion

Partial occlusion is a common issue in unconstrained face recognition environments, which makes the recognition of identities harder. Especially during the COVID-19 pandemic, when this work was written, many people wore masks that covered their faces from chin to nose, which has been shown to influence face recognition [51]. The AR face dataset [19] is used to evaluate the performance of the algorithms with respect to different partial occlusions. It consists of four protocols expression, occlusion, illumination, and occlusion_and_illumination, Fig. 3(a) displays some example images from the used protocols. For all experiments on this dataset, only images with neutral facial expressions were used to observe the influence of occlusion and illumination as isolated as possible. The identities were split up into 24 males and 19 females for each *dev* and *eval* sets.

In Fig. 3(b) we present the False Non-Match Rate (FNMR) and False Match Rate (FMR) on the *eval* set using the score threshold at FMR at 0.1% in the *dev* set. As can be seen, most of the networks are not severely affected by occlusion or illumination. The more recent **ArcFace-100**, **Zoo-AttentionNet**, and **Idiap-Resnet50** present both FNMR around 1% and FMR around the operational threshold of 0.1%. The slightly older **Facenet-Sandberg** presents some difficulties with



(a) Enrollment (b) Close (c) Medium (d) Far



(e) Results

Figure 5: SCface. *This figure displays some example images of the SCface dataset and the performance of the tested neural networks.*

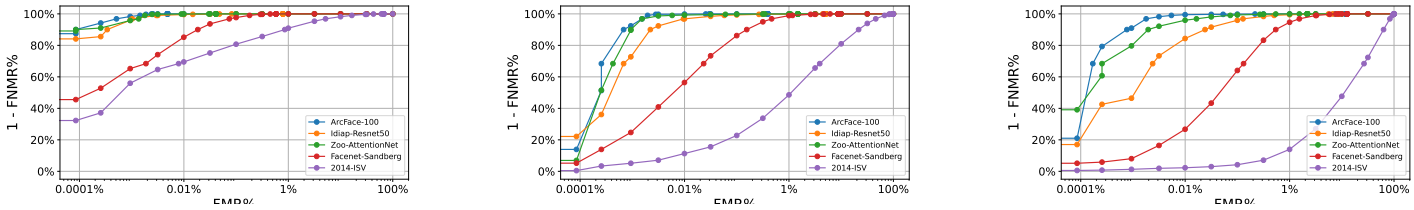
different types of occlusion that are isolated in Fig. 3(c), e.g., an FNMR of 29% is reported for sunglasses. Our selected top three algorithms present significant improvements in terms of FMR and FNMR compared with our **2014-ISV** system from eight years ago, which is severely impacted by both occlusion and illumination.

4.1.2 Facial Expressions

Humans are emotional beings and tend to show their emotions intensely through facial expressions, which has a significant visual impact on facial features [6]. Therefore, modern face recognition algorithms must be able to handle a wide range of facial expressions. The Multi-PIE [20] dataset with its protocol E is used to test the algorithms against a variety of expressions seen in Fig. 4(a). 64 identities are used in the *dev* set, and the *eval* set is composed of 65 identities. Five faces per identity with neutral expressions were considered for gallery template enrollment.

Fig. 4(c) shows the FNMR and FMR for different expressions by setting the decision threshold at FMR at 0.1% in the development set. The plot reveals that most recent networks can handle facial expressions well. All systems present an FMR around the operation threshold, which is an indication of homogeneity of both development set and evaluation sets. **ArcFace-100** and **Zoo-AttentionNet** show FNMR around 1% for all ex-

[12] See the FRICE 2016 section in https://www.idiap.ch/webarchives/sites/www.idiap.ch/resource/biometric



(a) Good (b) Bad (c) Ugly

Figure 6: The Good, the Bad & the Ugly. *ROC curves for the protocols Good, Bad and Ugly of the GBU dataset.*

pressions. The **Idiap-Resnet-50** presents an FNRM of around 1% for all expressions except for "scream" whose FNMR drastically increases to 12.3%. Both **Facenet-Sandberg** and **2014-ISV** present high FNRM for almost all expressions, reaching the highest values of 49% for and 59% for the "scream" expression.

4.1.3 Face Poses

Another aspect that challenges face recognition is the presence of different face poses. It is known that the performance of neural networks significantly drops when faces are no longer frontal [52]. Protocol P from the Multi-PIE dataset is used to observe the performance of neural networks on pose variations. This protocol provides faces rotated from left to right in steps of 15 degrees of yaw angle, examples can be seen in Fig. 4(b). The facial expressions are neutral, without any type of occlusion or strong illumination. When both eyes are visible, i.e. for ±45°, the hand-labeled eye positions are used for alignment. For the poses with deviations of more than 45° from frontal, only one eye is visible and, therefore, the alignment used the visible eye and the corner of the mouth. 64 identities are used for the *dev* set and 65 for the *eval* set, and five frontal images per identity are enrolled in a gallery template.

Fig. 4(d) shows the FNMR for different pose angles. For this experiment, we observed a similar trend in terms of FMR as in the previous one (all FMR are around the operational threshold). For that reason, we are plotting only the values of FNMR. It can be observed that all systems are able to handle well frontal poses with an angle of less than ±15°. FNMR starts to drastically increase for the **2014-ISV** for angles larger than ±30°, while all modern systems present similar FNMR for this particular set of angles. For 45° angles, the FNMR of **Facenet-Sandberg** starts to increase to around 35 %, while **Arcface-100** still presents an FNMR 0 % for this angle and **Idiap-Resnet50** and **Zoo-AttentionNet** increase to an FNMR of around 6%. For angles above ±45° the FNMR of **Zoo-AttentionNet** drastically increases to around 100%, while the **Idiap-Resnet50** and **Arcface-100** slowly increase to around 60% and 35% for angles of ±90°.

4.1.4 Face Sizes

The surveillance camera face (SCface) dataset [21] contains images taken by different low-resolution video surveillance cameras at three different distances. The three protocols close, medium, and far are used to evaluate the performance on different camera distances. For each protocol, images of 44 identities are used for the *dev* set, and 43 for the *eval* set. One frontal image taken in passport quality as shown in Fig. 5(a) is used for model enrollment, which differs dramatically in quality from the probe images, e.g., the far probe face as shown in Fig. 5(d) has only about 20 pixels of height.

Fig. 5(e) shows the FMR and FNMR on the *eval* set, indicating FMR values close to the estimated operational threshold. In terms of FNMR we could observe that for short distances the **ArcFace-100** is the best system, presenting 0%, followed by **Idiap-Resnet50** and **Zoo-AttentionNet** with 1.4% and 4.7% respectively. With around 20.9%, **Facenet-Sandberg** presents a very high FNMR. Once the distance between the probe subject and the camera increases, decreasing the image resolution, the FNRM also increases. At long distances (far), the **ArcFace-100** and the **Zoo-AttentionNet** present an FNMR of ≈ 65%. For the **Facenet-Sandberg** and **Idiap-Resnet50** FNMR reaches above 90%. Low-resolution probe samples seem to have a substantial impact on the performance of the algorithms.

4.2 Unconstrained Evaluations

This section provides the results for some common datasets, some of which were also evaluated in [2], in-

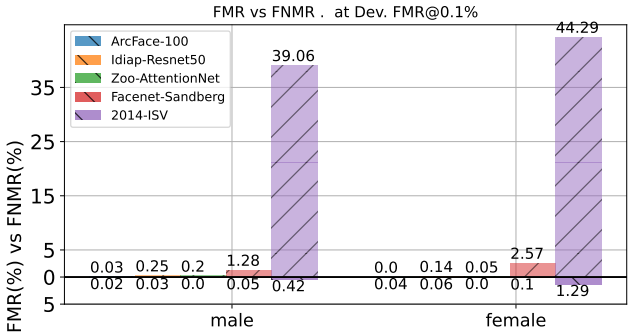


Figure 7: MOBIO. *This figure shows the performance of the tested neural networks on the MOBIO dataset.*

cluding MOBIO, GBU, and IJB-C.

4.2.1 MOBIO

The MOBIO dataset consists of video frames taken via mobile phone or laptop from male and female subjects. The *dev* set consists of 18 females and 24 males with 1'890 and 3'520 images, respectively, while the *eval* set contains 20 females and 38 males with 2'100 and 2'990 images. Five images per person are used for model enrollment.

Fig. 7 shows the FMR and FNMR on the *eval* set. **ArcFace-100**, **Idiap-Resnet50**, and **Zoo-AttentionNet** present an FNMR below 0.2% for all genders. The **Facenet-Sandberg** presents a slightly higher FNMR compared with the others, around 2% for both genders. These FNMRs present a substantial improvement compared with the **2014-ISV**, which reaches around 40%, and for which also the FMR raises substantially.

4.2.2 The Good, the Bad, & the Ugly

The experiments on the GBU [22] dataset are performed on all three protocols Good, Bad, and Ugly. Since the default evaluation protocols are not split into *dev* and *eval* sets, we are only able to report the ROC curve. The model enrollment uses only one image per identity, but there are several models per identity defined by the protocol.

All networks performed well on protocol Good, with the exception of **Facenet-Sandberg**. The best performance was achieved with the **ArcFace-100**, which provides the best FNMR for all FMR operational points. For the protocol Bad, **ArcFace-100** performed the best, while the **Idiap-Resnet50** had a slight decrease for FMR below 0.01%. The same trends could be followed by the protocols Ugly. Interestingly, for very low FMR values, results on the Ugly protocol are even better than on the Bad protocol, which indicates that the definition of Bad and Ugly has shifted since the development of this dataset.

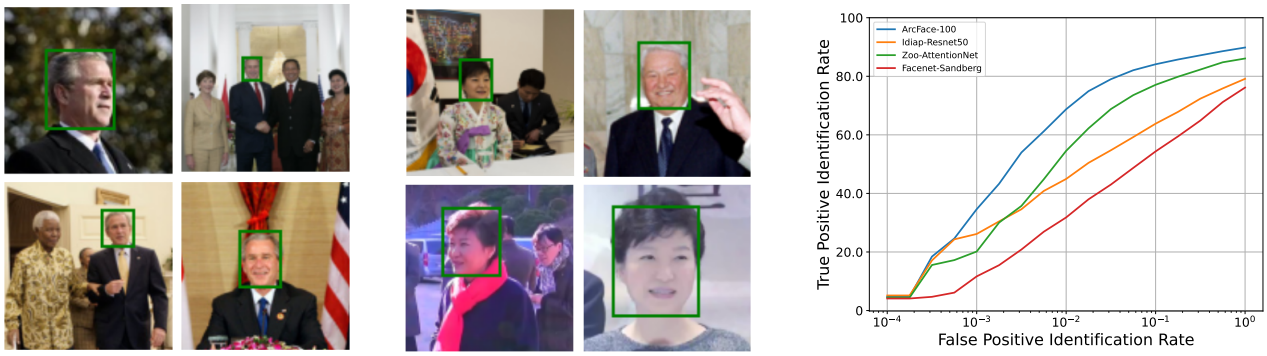
4.2.3 IJB-C

The IARPA Janus Benchmark C (IJB-C) is one of the most challenging evaluation datasets in face recognition research. This dataset contains evaluation protocols for face detection, face clustering, face verification, and open-set face identification. In this work we focused on open-set evaluation and as such, we will use the protocol test4-g1. This protocol contains a gallery of 1'170 subjects and a set of 1'759 "unknown" probes. Since IJB-C does not provide eye locations for face alignment, we first cropped the faces according to a slightly enlarged ground-truth face bounding box, detected the facial landmarks via MTCNN, and used the detected eye locations for alignment according to Tab. 1.

It is possible to observe that in this setup, **ArcFace-100** presents the bests TPIR, followed by the **Zoo-AttentionNet**. For closed-set results, i.e. TPIR=1 at the right-hand side of the plot, identification rates up to 90 % can be reached by the best-performing network. However, the TPIR rapidly decreases with decreasing FPIR. For FPIR=0.001 (0.1 %) all systems operate with a TPIR well-below 40 %. This means, to have a level of false positives of one in a thousand, we should expect an identification rate of 40 %. This is far from being practical in any surveillance camera application.

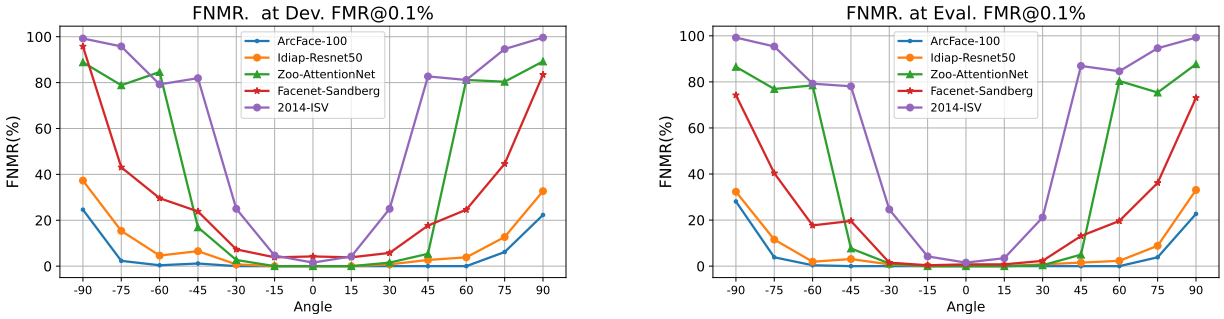
5 Discussion

The goal of this work is threefold. The first aim of this paper is to show the increase of face recognition performance in the last eight years, i.e., with the advent of the deep learning technology. In our experiments we have utilized four different pre-trained deep networks that were developed at various times during the eight years, and compared their performance with the best-performing technique from our previous study [1, 2]. In most of our experiments, we find that the performance decreased when the age of the algorithm increased. For example, **2014-ISV** generally performs worst, followed by the **FaceNet-Sandberg** model developed in 2015 and the **Idiap-Resnet-50** trained in 2020 ranging third. The second-best of the evaluated methods is the **Zoo-AttentionNet** and our winner **ArcFace-100** shows the



(a) Example of enrollment samples (b) Example of probe samples (c) Open-Set ROC Curve

Figure 8: IJB-C. *This figure displays some example images of the IJB-C dataset and the Open-Set ROC curve.*



(a) Protocol-Specific Threshold on *dev* (b) Protocol-Specific Threshold on *eval*

Figure 9: Different Ways to Select a Threshold. *We depict the effect of two different wrong ways to select a score threshold. In (a), separate score thresholds are estimated on the* dev *set for each sub-protocol. In (b), separate score thresholds are selected directly on the* eval *set. A comparison to our proposed evaluation procedure in Fig. 4(d) reveals that both ways result in an unwarrented decrease in error rates.*

most stable performance across all of our experiments. While our selection covers many modern deep learning algorithms, it cannot be excluded that other inventions perform better in one or the other task. Fortunately, our open-source, reproducible and easily extensible framework[1] allows to incorporate novel algorithms quickly and produces fair evaluations with respect to our tested models.

The second goal of this work is to assess what are the next steps that research should put a focus on. We evaluate various different aspects of face recognition. We are surprised by our finding that – when evaluated separately – partial occlusions and facial expressions are practically solved by our winner network **ArcFace-100** since nothing in the training procedure particularly focused on solving occlusion or expressions. Only the datasets used for training these models include some images with expressions and mild occlusions. For the aspect of face pose, the algorithms have improved drastically over the years. While **2014-ISV** reaches more than 40 % FNMR already with angles of ±30°, the evaluated networks can work with angles of ±45° or even up to ±75° for **Idiap-Resnet50** reasonably well; only for full-profile images, error rates are beyond expectation. The most critical evaluation is on the SCface dataset, where none of the algorithms is able to work with the lowest-resolution faces. Additionally, we show that open-set recognition is far from being solved and, hence, the research needs to focus more on this aspect so that the technology can be utilized for the very important task of identifying offenders in surveillance cameras.

The final goal of this paper is to show that the evaluation procedure must be adapted from how it is currently performed. First, many evaluation protocols require to

provide separate results for each protocol, and we have done the same mistake in our previous evaluation [1, 2]. For example, the Celebrities in Frontal-Profile (CFP) dataset [52] provides two protocols, a frontal-frontal and a frontal-profile protocol, and more than 94 % accuracy is reported in [10]. When evaluating both protocols separately, a different score threshold is selected for both protocols. This is, however, not how face recognition works in practice where a single threshold is used independently of the type of image (frontal or profile) at hand. To highlight the difference, we repeat the evaluation of the different poses from Fig. 4(d) by selecting a separate threshold per sub-protocol (one threshold for each face pose) on the *dev* set and plotting the FNRM on the *eval* set in Fig. 9(a). Clearly, the error rates drop drastically when a separate threshold is computed. Second, in most datasets and evaluation protocols, ROC curves are plotted that show the performance of the evaluated system only on the *eval* set (aka. the test set). How well a threshold selected on this test set translates to previously unseen subjects is not clear, but from comparing Fig. 9(b) with Fig. 9(a) we can see that there is a trend to reduce error rates when selecting the threshold on the *eval* set directly. We believe, splitting the protocols into *dev* and *eval* is critical to evaluate the algorithm on data that has not been seen at any stage of the process.

Finally, we want to highlight the utmost importance of reproducible research [17] and the requirement of providing all required details both in the paper and in code. For example, the alignment of faces is an important step for the ArcFace network, but neither the paper nor the source code clearly shows how to do a good alignment. Especially the alignment procedure required for handling profile images is nowhere to be found and, consequently, we had to come up with our own alignment procedure, cf. Tab. 1, that seemed to have provided good results [46, 48]. Only for the networks for which we know the exact alignment of profile faces (**Idiap-Resnet-50** and **FaceNet-Sandberg**), results do not abruptly degrade between ±45° and ±60°.

6 Conclusion

This work provides an overview on the challenges that still remains in face recognition research by running a similar set of evaluations that we carried out eight years ago in our previous work [1] in a reproducible and open-source manner. Our evaluation protocols allow an isolated examination of single aspects of face recognition (e.g., pose, occlusion, illumination, low resolution, unconstrained open-set identification), as well as a more application-oriented evaluation. Below follow the problems that are solved well in face recognition research and the remaining challenges.

6.1 Problems Mastered in Eight Years

In general, we could observe that certain types of occlusion are handled well by the state-of-the-art networks using the AR face dataset as a proxy. Two networks (**ArcFace-100**, **Zoo-AttentionNet**) presented a FNMR of 0% at FMR 0.1%, which is a substantial improvement from the best system we have in our previous work. A similar trend is observed with face expressions and face recognition in mobile phones using the MultiPIE and the MOBIO datasets as respective proxies. Illumination from different directions is no longer an obstacle, but different illumination types still constitute a gap for further research [46]. Nevertheless, for face recognition on mobile phones, we were able to decrease FNMR from 44% to 0%.

6.2 Problems Remaining to be Solved

Despite substantial improvement, we could observe that recognition under strong pose variations is still a problem in face recognition. Recognition under angles until 60° is very well handled by most networks. However, once this angle increases, the number of false non-matches substantially increases as we could observe using Multi-PIE as proxy. Recognition at a distance or with low-resolution or low-quality probe images is also an open problem in face recognition. We could observe extremely high FNMRs using the "Ugly" protocol on GBU and the "far" protocol from SCface. For instance, the state-of-the-art **ArcFace-100** presented an FNMR of 69% on SCface, which is impractical in the real world. Another open problem is open-set face recognition. Using IJB-C as a proxy, we could observe a closed-set recognition rate of 90% (for FPIR=100%) using the state-of-the-art **ArcFace-100**. This figure of merit goes down to around 35% for a more realistic value of FPIR=0.1%. Finally, the reproducibility of research still is a problem. For example, the developers of ArcFace decided to change their alignment procedure and their popular previously trained networks – one of which we have used in our experiments – are no longer to be found online. Also, the alignment procedure required for profile faces in ArcFace and AttentionNet is not clear. While we spent some effort to find optimal alignments for ArcFace both in their

code and empirically [46], the results of AttentionNet should be able to be improved with better alignment.[13] We are utilizing the open-source and reproducible face recognition framework of Bob [18] and providing all relevant details of all our experiments.[1] This makes our research distinct from other reviews that can only rely on results reported in the literature since they cannot re-run experiments or change evaluation metrics.

While some problems still remain to be solved, we could observe great progress in face recognition research in the last eight years. It is worth noting that none of the tested algorithms were carefully crafted to handle the above-mentioned aspects. The availability of large amounts of data definitely plays an important role in the recent state-of-the-art networks. Furthermore, the three best systems we presented use different variations of the ArcFace loss [11], which definitely played an important role as well.

We are aware that we only used a small subset of available deep networks for face recognition, and we are sorry if we missed your particular network. Furthermore, this work did not consider security aspects in face recognition systems, such as morphing or presentation attacks [53, 54]. Possible extensibility to cover these aspects would require new work with a new experimental setup that we were not able to cover in this one. Fortunately, the source code for this study is publicly available,[1] and new implementations [45] in Bob's biometric recognition framework [16–18] allow for a very easy extension to include your network.

References

[1] M. Günther, L. El Shafey, and S. Marcel, *Face recognition across the imaging spectrum*. Springer, 2016, ch. Face recognition in challenging environments: An experimental and reproducible research survey.

[2] ——, "2D face recognition: An experimental and reproducible research survey," Idiap, Tech. Rep., 2017.

[3] S. Minaee, A. Abdolrashidi, H. Su, M. Bennamoun, and D. Zhang, "Biometrics recognition using deep learning: A survey," *arXiv*, 2021.

[4] D. Wanyonyi and T. Celik, "Open-source face recognition frameworks: A review of the landscape," *IEEE Access*, vol. 10, 2022.

[5] I. Masi, Y. Wu, T. Hassner, and P. Natarajan, "Deep face recognition: A survey," in *Conference on Graphics, Patterns and Images (SIBGRAPI)*, 2018.

[6] G. Guo and N. Zhang, "A survey on deep learning based face recognition," *Computer Vision and Image Understanding*, vol. 189, 2019.

[7] A. J. O'Toole, P. J. Phillips, F. Jiang, J. H. Ayyad, N. Penard, and H. Abdi, "Face recognition algorithms surpass humans matching faces over changes in illumination," *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, no. 9, 2007.

[8] J. S. del Rio, D. Moctezuma, C. Conde, I. M. de Diego, and E. Cabello, "Automated border control e-gates and facial recognition systems," *Computers & Security*, vol. 62, 2016.

[9] M. Günther, P. Hu, C. Herrmann, C.-H. Chan, M. Jiang, S. Yang, A. R. Dhamija, D. Ramanan, J. Beyerer, J. Kittler *et al.*, "Unconstrained face detection and open-set face recognition challenge," in *International Joint Conference on Biometrics (IJCB)*. IEEE, 2017.

[10] M. Wang and W. Deng, "Deep face recognition: A survey," *Neurocomputing*, vol. 429, 2021.

[11] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[12] Q. Meng, S. Zhao, Z. Huang, and F. Zhou, "MagFace: A universal representation for face recognition and quality assessment," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[13] M. M. Sawant and K. M. Bhurchandi, "Age invariant face recognition: a survey on facial aging databases, techniques and effect of aging," *Artificial Intelligence Review*, vol. 52, no. 2, 2019.

[14] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," University of Massachusetts, Amherst, Tech. Rep., 2007.

[13] Some code found in the AttentionNet repository suggests that ArcFace and AttentionNet use the same image alignment, but it is unclear if this code was actually used to align their images.

[15] B. Maze, J. Adams, J. A. Duncan, N. Kalka, T. Miller, C. Otto, A. K. Jain, W. T. Niggel, J. Anderson, J. Cheney, and P. Grother, "IARPA Janus Benchmark - C: Face dataset and protocol," in *International Conference on Biometrics (ICB)*, 2018.

[16] A. Anjos, L. El-Shafey, R. Wallace, M. Günther, C. McCool, and S. Marcel, "Bob: a free signal processing and machine learning toolbox for researchers," in *ACM Conference on Multimedia Systems (ACMMM)*, 2012.

[17] A. Anjos, M. Günther, T. de Freitas Pereira, P. Korshunov, A. Mohammadi, and S. Marcel, "Continuously reproducing toolchains in pattern recognition and machine learning experiments," in *International Conference on Machine Learning (ICML)*, 2017.

[18] M. Günther, R. Wallace, and S. Marcel, "An open source framework for standardized comparisons of face recognition algorithms," in *European Conference on Computer Vision (ECCV) Workshops and Demonstrations*. Springer, 2012.

[19] A. Martínez and R. Benavente, "The AR face database," Computer Vision Center, Tech. Rep. 24, 1998.

[20] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-PIE," *Image and Vision Computing*, vol. 28, no. 5, 2010.

[21] M. Grgic, K. Delac, and S. Grgic, "SCface – surveillance cameras face database," *Multimedia Tools and Applications*, vol. 51, no. 3, 2011.

[22] P. J. Phillips, J. R. Beveridge, B. A. Draper, G. Givens, A. J. O'Toole, D. S. Bolme, J. Dunlop, Y. M. Lui, H. Sahibzada, and S. Weimer, "An introduction to the good, the bad, & the ugly face recognition challenge problem," in *International Conference on Automatic Face Gesture Recognition (FG)*, 2011.

[23] C. McCool, S. Marcel, A. Hadid, M. Pietikainen, P. Matejka, J. Cernocky, N. Poh, J. Kittler, A. Larcher, C. Levy, D. Matrouf, J.-F. Bonastre, P. Tresadern, and T. Cootes, "Bi-modal person recognition on a mobile phone: Using mobile phone data," in *International Conference on Multimedia and Expo Workshops*, 2012.

[24] R. Wallace, M. McLaren, C. McCool, and S. Marcel, "Inter-session variability modelling and joint factor analysis for face authentication," in *International Joint Conference on Biometrics (IJCB)*, 2011.

[25] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *British Machine Vision Conference (BMVC)*, 2015.

[26] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VGGFace2: A dataset for recognising faces across pose and age," in *International Conference on Automatic Face Gesture Recognition (FG)*, 2018.

[27] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," *arXiv*, 2014.

[28] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "MS-Celeb-1M: A dataset and benchmark for large-scale face recognition," in *European Conference on Computer Vision (ECCV)*. Springer, 2016.

[29] T. de Freitas Pereira, A. Anjos, and S. Marcel, "Heterogeneous face recognition using domain specific units," *Transactions on Information Forensics and Security (TIFS)*, vol. 14, no. 7, 2018.

[30] Z. Zhu, G. Huang, J. Deng, Y. Ye, J. Huang, X. Chen, J. Zhu, T. Yang, J. Lu, D. Du *et al.*, "WebFace260M: A benchmark unveiling the power of million-scale deep face recognition," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[32] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[33] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," *arXiv*, 2017.

[34] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *European Conference on Computer Vision (ECCV)*, 2016.

[35] G.-S. J. Hsu, H.-Y. Wu, and M. H. Yap, "A comprehensive study on loss functions for cross-factor face recognition," in *Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2020.

[36] W. Liu, Y. Wen, Z. Yu, and M. Yang, "Large-margin softmax loss for convolutional neural networks," in *International Conference on Machine Learning (ICML)*, 2016.

[37] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "SphereFace: Deep hypersphere embedding for face recognition," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[38] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "CosFace: Large margin cosine loss for deep face recognition," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[39] X. Zhang, R. Zhao, Y. Qiao, X. Wang, and H. Li, "AdaCos: Adaptively scaling cosine logits for effectively learning deep face representations," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[40] X. Zhang, R. Zhao, J. Yan, M. Gao, Y. Qiao, X. Wang, and H. Li, "P2SGrad: Refined gradients for optimizing deep face models," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[41] Y. Zheng, D. K. Pal, and M. Savvides, "Ring loss: Convex feature normalization for face recognition," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[42] M. Günther, A. R. Dhamija, and T. E. Boult, "Watchlist adaptation: Protecting the innocent," in *International Conference of the Biometrics Special Interest Group (BIOSIG)*, 2020.

[43] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[44] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.

[45] Y. Linghu and X. Zhang, "Open-source package for generic deep-network-based face detection and recognition in bob," Master's Project, University of Zurich, 2021.

[46] D. Schmidli, "Face recognition aspects with DNNs: An experimental and reproducible research survey," Bachelor's Thesis, University of Zurich, 2021.

[47] J. Wang, Y. Liu, Y. Hu, H. Shi, and T. Mei, "FaceX-Zoo: A pytorch toolbox for face recognition," in *ACM International Conference on Multimedia (ACMMM)*, 2021.

[48] T. Wartmann, "Frontal to profile face recognition with rank lists," Bachelor's Thesis, University of Zurich, 2021.

[49] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *Signal Processing Letters*, vol. 23, no. 10, 2016.

[50] P. J. Phillips, P. Grother, and R. Micheals, *Handbook of face recognition.* Springer, 2011, vol. 2, ch. Evaluation Metrics in Face Recognition.

[51] N. Damer, J. H. Grebe, C. Chen, F. Boutros, F. Kirchbuchner, and A. Kuijper, "The effect of wearing a mask on face recognition performance," in *International Conference of the Biometrics Special Interest Group (BIOSIG)*, 2020.

[52] S. Sengupta, J.-C. Chen, C. Castillo, V. M. Patel, R. Chellappa, and D. W. Jacobs, "Frontal to profile face verification in the wild," in *Winter Conference on Applications of Computer Vision (WACV)*, 2016.

[53] U. Scherhag, C. Rathgeb, J. Merkle, R. Breithaupt, and C. Busch, "Face recognition systems under morphing attacks: A survey," *IEEE Access*, vol. 7, 2019.

[54] S. Marcel, M. S. Nixon, J. Fierrez, and N. Evans, *Handbook of biometric anti-spoofing: Presentation attack detection.* Springer, 2019, vol. 2.