A Comprehensive Experimental and Reproducible Study on Selfie Biometrics in Multistream and Heterogeneous Settings

Guillaume Heusch, Tiago de Freitas Pereira, *Member, IEEE*, and Sébastien Marcel, *Senior Member, IEEE*

***Abstract*—This contribution presents a new database to address current challenges in face recognition. It contains face video sequences of 75 individuals acquired either through a laptop webcam or when mimicking the front-facing camera of a smartphone. Sequences have been acquired with a device allowing to record visual, near-infrared, and depth data at the same time. Recordings have been made across three sessions with different, challenging illumination conditions and variations in pose. Together with the database, several experimental protocols are provided and correspond to real world scenarios, when a mismatch in conditions between enrollment and probe images occurs. A comprehensive set of baseline experiments using publicly available baseline algorithms show that extreme illumination conditions and pose variations are remaining issues. However, the usage of different data domains—and their fusion—allows to mitigate such variation. Finally, experiments on heterogeneous face recognition are also presented using a state-of-the-art model based on deep neural networks, and showed better performance. When applied to other tasks, this model proved to surpass all existing baselines as well. The data, as well as the code to reproduce all experiments are made publicly available to help foster research in selfie biometrics using latest imaging devices.**

***Index Terms*—Face recognition, database, multistream, illumination, pose, heterogeneous, reproducible research.**

I. Introduction

FACE recognition is a popular topic in computer vision, with a wide range of applications such as physical access control, remote authentication or smart device unlocking, to name a few. After more than 30 years of active research, face recognition in *controlled* environments can be considered mature, as evidenced by numerous applications deployed at scale, like automatic gates at border crossing in airports. However, there are challenges remaining. Large variations in pose and in illumination conditions have been consistently identified as the main limitations of face recognition systems [1], [2], [3]. Thanks to the availability of very large face datasets together with the recent and rapid progress of deep neural networks, noticeable progress have been made and computer-based face recognition is now close to human performance. However, most of these recent approaches have been made on faces retrieved from Web-based color images, and although such data exhibit some variations in terms of pose and illumination conditions, they remain quite modest. Furthermore, recent studies show that state-of-the-art face recognition systems are prone to presentation attacks (or spoofing) [4]. Hence, for a typical application, i.e., authentication on a smartphone, to be successful and widely accepted, such threats should be taken into account. Consequently, both academia and the industry are investigating new sensors for secure face recognition (e.g., the Apple iPhone X).

The usage of 3D data can greatly help to alleviate all of the aforementioned challenges. Indeed, it can cope with pose and illumination variations [5], [6] and can help the detection of presentation attacks [7]. However, most 3D approaches require high-resolution scanners, which are both bulky and extremely costly. The recent development of low-cost RGB-D devices allowed researchers to investigate face-related tasks with such data. As a consequence, several RGB-D face databases, mostly collected with the Microsoft Kinect device, have been proposed in recent years [8], [9]. Examples relevant to our work include the CurtinFace database [10], the Eurecom database [11], the KaspAROV database [12] and the Lock3DFace database [13], that are all dedicated to face recognition.

While current approaches using RGB-D data can, to some extent, cope with pose and illumination variations [3], [10] near-infrared (NIR) imaging has also been shown to be robust to illumination variations, even in extreme conditions [2]. As a consequence, this modality has also been extensively studied in the context of face recognition, whether in a single [14], [15] or multimodal setting [2]. Therefore, there also exists face database containing both RGB and near-infrared (NIR) images: examples include the CASIA NIR-VIS 2.0 [16] and the Near-Infrared and Visible-Light (NIVL) from the University of Notre Dame [17]. However, these databases do not contain 3D data, and are usually composed of frontal faces acquired in a controlled environment.

Although there exist datasets providing all three modalities, significant differences with the proposed dataset should be noted. For instance, the Multi-Dim database [18] contains, for each subject, high-quality 3D face scans, 2D high-quality still

Manuscript received November 28, 2018; revised March 12, 2019; accepted May 5, 2019. Date of publication July 11, 2019; date of current version October 30, 2019. This work was supported by the Swiss Commission for Technology and Innovation (CTI) through the project entitled "Convenient and Secure 3D Face Recognition based on RGB-D Cameras" under Project 18713.1. This paper was recommended for publication by Associate Editor Z. Sun upon evaluation of the reviewers' comments. *(Corresponding author: Guillaume Heusch.)*

The authors are with the Biometrics Security and Privacy Group, Idiap Research Institute, 1920 Martigny, Switzerland (e-mail: guillaume.heusch@idiap.ch; tiago.pereira@idiap.ch; sebastien.marcel@idiap.ch).

Digital Object Identifier 10.1109/TBIOM.2019.2927692

TABLE I
OVERVIEW OF MULTIMODAL FACE DATABASES DEDICATED TO FACE RECOGNITION

Name	# of subjects	NIR	depth	light variation	pose variation	selfie mode	annotations
Eurecom	52	✗	✓	no	yaw	✗	6 points
IIIT-D RGB-D	106	✗	✓	no	moderate	✗	none
CurtinFaces	52	✗	✓	yes	yaw + pitch	✗	none
CASIA NIR-VIS 2.0	725	✓	✗	moderate	moderate	✗	eyes
NIVL	402	✓	✗	no	no	✗	eyes
CASIA HFB	100	✓	✓	no	no	✗	eyes
KaspAROV	108	✓(52 subj.)	✓	yes	yaw + pitch	✗	none
Lock3DFace	509	✓	✓	no	yaw + pitch	✗	4 points
Multi-Dim	124	✓	✓	yes	moderate	✗	unknown
FARGO	75	✓	✓	yes	yaw + pitch	✓	16 points

color images and surveillance video clips in both the visual spectrum and the near-infrared spectrum. In this case, images are extracted from different devices and are not aligned (RGB and NIR devices were not set in the same location) and consequently do not contain the same conditions in terms of both pose and illumination. The CASIA-HFB database [19] also contains RGB, NIR and depth information. It was recorded in a typical laboratory setting, where the subject quietly sits in front of the camera. While there are some variations in expression, the pose is always frontal and illumination conditions do not vary. Finally, the Lock3DFace database [13] is maybe the closest to ours, since it was recorded using a Kinect v2 device. It hence allowed to record RGB, NIR and depth data at the same time, but unfortunately, most of the subjects were recorded during one session only, preventing the dataset from having variations in illumination conditions. Besides, it is worth mentioning that none of the databases presented above contain recordings carried out in *outdoor settings*, and hence do not provide a wide-range of realistic conditions for probe images. Table I summarizes relevant multimodal face databases (including ours) with their respective features.

In order to help the research community tackle current limitations in face recognition, we present the FARGO database: a face dataset containing RGB, depth and near infrared (NIR) video sequences. No less than 75 subjects have been recorded with an Intel RealSense Camera (model SR300), and constitute, to the best of our knowledge, the first public face dataset with *synchronized* RGB, NIR and depth sequences acquired in various conditions, including outdoors.

Equipped with this new database, we also provide various baseline experiments in several face verification scenarios. Frontal face verification experiments are conducted in each of the modalities, and on a variety of mismatched, challenging illumination conditions. Then, fusion of different algorithms, but also of different modalities are presented. Experiments on pose-varying and heterogeneous face recognition are also reported. Note that all experiments presented in this contribution are reproducible and that the data,[1] as well as the code[2] are both publicly and freely available. Several algorithms are compared for each of the different tasks: one is based on Gabor local features [20] and another relies on a advanced statistical model, Inter-session Variability [21]. Recent approaches using deep convolutional networks, comprising the classical VGG model [22], the computationally efficient LightCNN [23], and models especially dedicated to pose-invariant face recognition [24] or heterogeneous face recognition [25] have also been also considered. Although the main purpose of this dataset is face recognition, we believe that it could benefit other face-related tasks, such as facial feature localization, head pose estimation, and so on.

The rest of this contribution is organized as follows: Section II presents the FARGO database and the face verification protocols in more details. Section III describes the algorithms used as baselines. Section IV is dedicated to the experimental evaluation: it presents results on the various face verification tasks and also provides a discussion and qualitative results. Finally, Section V summarizes our contribution and concludes the paper.

II. THE FARGO DATABASE

The FARGO database has been recorded across a time period of 5 months on three different sites. 75 subjects have been recorded, among which 20 are females and 55 males. At the time of recording, most of the subjects were aged between 20 and 30 years old - the exact age is available as metadata. The recordings have been made using an Intel RealSense SR-300 device, allowing to capture classical RGB, Near-Infrared (NIR) and depth maps video sequences at the same time. Each subject was recorded during three sessions. The first session took place in an indoor environment with controlled lighting, ensuring the face to be well lit. The second session has been recorded in a very dark room, and the third one has been recorded outdoor, and hence contains arbitrary illumination conditions. Figure 2 provides example images for each session. In each session and for each subject, four video sequences were recorded: two where the device was mounted as a webcam on a laptop, and two where the device was mimicking the frontal camera of a mobile phone (see Figure 1).

During each recording, the subject has been asked to remain still for the first five seconds, and then to move his head to the left, to the right, to the top and to the bottom, *while still looking at the device*. This has been done for two reasons: the movements in yaw will allow to address the challenge

[1] https://www.idiap.ch/dataset/fargo

[2] https://gitlab.idiap.ch/bob/bob.paper.fargo_tbiom_2019

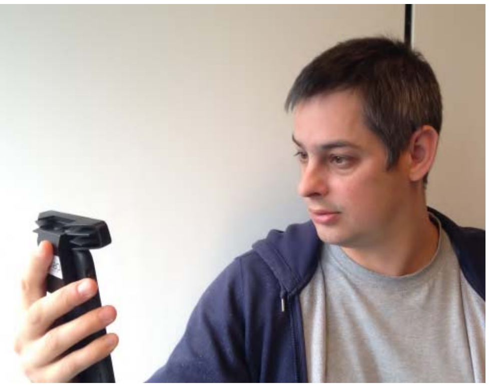


Fig. 1. Illustration of the capture process when the device is mimicking the front-facing camera of a smartphone.

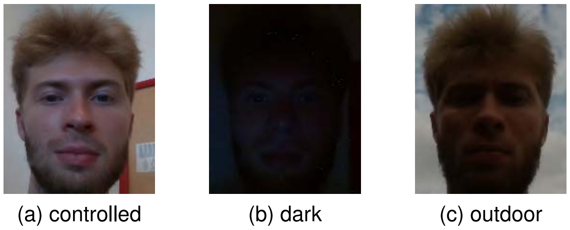


Fig. 2. Example of images acquired in each session.

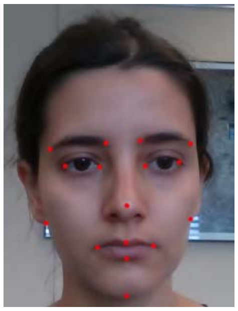


Fig. 3. The 16 annotated fiducial points.

of face recognition across pose and the movements in pitch are trying to mimic the typical pose variations one can observe when using a front-facing smartphone camera. For all recorded face video sequences, 13 specific frames have been manually annotated. Roughly, these frames correspond to a frontal view of the face, to the extreme positions attained when the subject moves her/his head (left, right, top and bottom), plus two frames in between the extreme position and the frontal view. Selected frames have been annotated with 16 keypoints corresponding to salient facial features, as depicted in Figure 3.

A. Protocols

In our different scenarios, the specific task of face verification is addressed: a client is claiming her/his identity and supports this claim by providing an image of her/his face (usually called a *probe*). The goal here is to decide whether the claim is legit by comparing the probe image with a previously enrolled model. In our framework, face verification is made on still images. As a consequence, various frames from different video sequences have been extracted, depending on the task at hand. Experimental protocols are provided for three different tasks corresponding to current challenges in face recognition:

1) Frontal face verification across illumination conditions: images for enrollment are in clean conditions, whereas probe images contains low or arbitrary (outdoor) illumination conditions.
2) Pose-varying face verification: images for enrollment contains frontal faces, but probe images contain faces at a different pitch or yaw.
3) Heterogeneous face verification: images for enrollment comes from one domain (i.e., RGB), probe images are taken from another domain (i.e., NIR).

For each protocol, and as it is a standard practice in biometrics, the dataset has been divided into 3 distinct subsets, each containing 25 identities: training, development and evaluation. Special care has been taken to balance these subsets with respect to gender and recording location. The training set is used to build our prior knowledge of the problem and is not involved in verification experiments *per se* (*i.e.,* no identities present in this set will be used as either clients or zero-effort impostors). The various parameters of the different algorithms and the decision threshold are tuned on the development set, and the final assessment is made on the evaluation set. This framework ensures an unbiased assessment and is widely adopted in biometric verification experiments [26], [27].

1) Frontal Face Verification: One aim of this dataset is to assess the performance of the different modalities in unmatched, difficult lighting conditions. As a result, three face verification protocols (for each of the modalities) have been devised. In particular, experiments reflect the fact that usually, images for enrollment are recorded under controlled conditions whereas probe images could be acquired in arbitrary conditions. The three protocols use the same clean images for both training and enrollment, but differ in probe images:

1) *Matched Controlled (MC):* Probes are taken in controlled conditions.
2) *Unmatched Dark (UD):* Probes are taken in dark conditions.
3) *Unmatched Outdoor (UO):* Probes are taken in outdoor conditions.

The first scenario aims at establishing a baseline for face verification performance, and the two unmatched conditions are used to assess the behaviour of the different modalities - and of the different algorithms - in more realistic conditions. A summary of these protocols is given in Table II. For each subject, 10 images containing a frontal face have been extracted in each recorded video sequence. Since two recordings with two different mounting of the device have been acquired in each session, there is a total of 40 images per subject in each illumination condition. This yield a training set of 1000 images (25 subjects in clean conditions). Note that the number of probe images is halved for the MC protocol, since half of the images have been used for enrollment.

2) Pose-Varying Face Verification: To address the problem of pose-varying face verification, images at different poses have been extracted and clustered to build additional probe

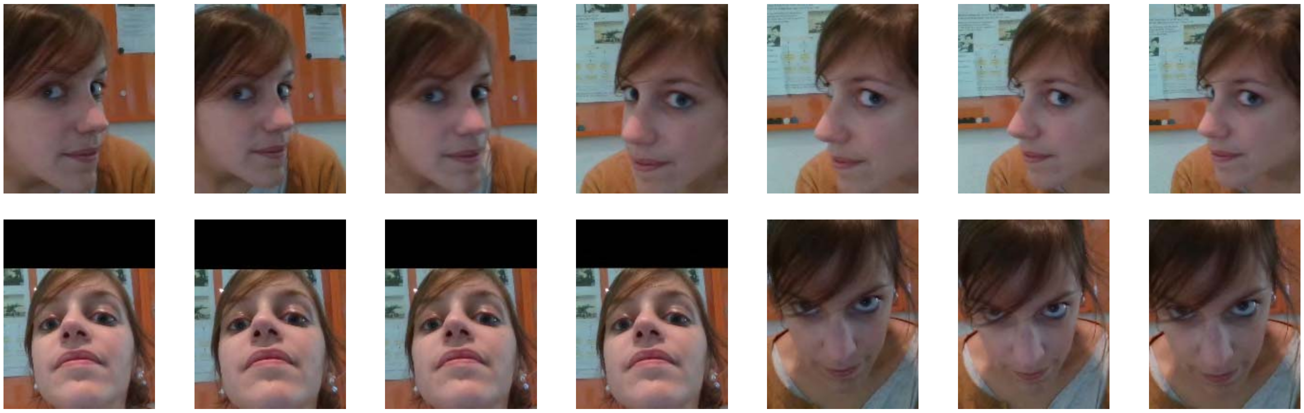


Fig. 4. Example of probe images for pose-varying face verification. First row contains variation in yaw, whereas second row shows variations in pitch.

TABLE II

SUMMARY OF THE DIFFERENT PROTOCOLS FOR FRONTAL FACE VERIFICATION: c STANDS FOR CONTROLLED, **d** FOR DARK AND **o** FOR OUTDOOR. THE NUMBER OF IMAGES PER SUBJECT IS GIVEN IN PARENTHESIS

	Training	**Dev**		**Eval**	
		Enroll	Probe	Enroll	Probe
MC	c (40)	c (20)	c (20)	c (20)	c (20)
UD	c (40)	c (20)	**d** (40)	c (20)	**d** (40)
UO	c (40)	c (20)	**o** (40)	c (20)	**o** (40)

images sets. For training and enrollment images however, the same set as in frontal face verification experiments is used: it contains frontal face images only. Note also that protocols with mismatched probes in terms of pose are performed under controlled conditions.

The different probe sets have been built by taking advantage of annotated frames. In particular, the annotated frames at the end of each movement and the one between the fully frontal position have been used. This allow to extract a set of images in each directions, and also to ensure that the face is not too frontal. Then, images corresponding to rotation in yaw and in pitch have been clustered together, yielding two additional probe sets. An example of probe images with variation in poses extracted from one sequence is given in Figure 4. Note that since not all the subjects were moving their head at the same speed, the number of extracted images may vary in each sequence. Consequently, probe sets with variations in yaw and pitch contain an average of 22.5 and 26.8 images per subject respectively.

3) Heterogeneous Face Verification: To address the problem of heterogeneous face recognition, two major protocols were designed. The first one addresses the task of matching RGB images to NIR images and the second one addresses the task of matching RGB images to depth maps. The same protocols as in Section II-A1 are used here. The difference lies in the modalities present in each set. The training set contains images from both *source* and *target* domains: this is required in the training procedure of the Domain Specific Units algorithm (see [25] for details). Enrollment is then done using images coming from the *source* domain (i.e., RGB) and probe images come from the *target* domain (i.e., either NIR or depth).

III. Face Recognition Algorithms

This section presents face recognition algorithms used to establish baselines for the various face verification scenarios of the FARGO database. The different approaches have been selected since their implementation are publicly available, and they have been shown to achieve high recognition rates under different scenarios [23], [28], [29], while being quite different from each other. Two other approaches based on deep neural networks have also been investigated, since they have specifically been designed to handle particular tasks. The DR-GAN [24] was made to handle pose-varying face recognition and Domain Specific Units [25] addresses heterogeneous face recognition.

A. Gabor Grid Graphs

The Gabor Grid Graph (GGG) algorithm has been proposed by Günther *et al.* [20]. As its name suggests, it relies on the Gabor wavelet transform and borrows ideas from the Elastic Bunch Graph Matching algorithm [30]. Rather than having an elastic graph, this algorithm computes the response of 40 Gabor wavelets (8 orientations and 5 scales) on points located on a regular grid across the image. A set of Gabor Jets, which can be viewed as local texture features, is then formed by concatenating the 40 responses at each image location defined by the grid. Two Gabor jets, at respective locations $\mathcal{J}$ and $\mathcal{J}'$, are compared using a similarity measure given by [20]:

$$S\big(\mathcal{J}, \mathcal{J}'\big) = \sum_j \left[\frac{a_j - a'_j}{a_j + a'_j} + \cos\Big(\phi_j - \phi'_j - \vec{k}_j^T \vec{d}\Big) \right] \quad (1)$$

Face verification is achieved by comparing Gabor Jets extracted from a gallery of templates and the given probe. The similarities at each grid location are then averaged to reach a final score. Note finally that this algorithm does not require any training phase.

B. Inter-Session Variability

The aim of Inter-Session variability (ISV) is to learn a generative model to describe the distribution of local features extracted from the face image [21]. Local features are obtained by considering overlapping blocks within the image.

Each block is mean and variance-normalized before extracting the D lowest-frequency components of the 2D Discrete Cosine Transform. As a result, the face image is represented with a set of K feature vectors of dimension D: $O = \{o_1, o_2, \ldots, o_K\}$. In order to build models for each client, a Universal Background Model (UBM) is first trained using a training set comprising different identities. A *mean super-vector* $\boldsymbol{m}$ is then built by simply concatenating the GMM component means. To derive a specific client model $\boldsymbol{c}_i$, the UBM super-vector is *adapted* in the following way:

$$\boldsymbol{c}_i = \boldsymbol{m} + \boldsymbol{d}_i \tag{2}$$

where $\boldsymbol{d}_i$ is an offset characterizing client i. Since client images may contain significant variations (such as facial expressions for instance), this within-class variability is also embedded in $\boldsymbol{c}_i$ and may result in a degradation of verification performance. The aim of ISV is hence to model and suppress this within-class variability. The super-vector corresponding to the enrollment image j of client i is in this case written as:

$$\boldsymbol{\mu}_{i,j} = \boldsymbol{m} + \boldsymbol{u}_{i,j} + \boldsymbol{d}_i, \tag{3}$$

where the extra term $\boldsymbol{u}_{i,j}$ represents the particular condition of image j, and $\boldsymbol{d}_i$ is now closer to the *true* client offset, from which external variability has been removed. ISV assumes that within-client variation is contained in a linear subspace of the GMM mean super-vector:

$$\boldsymbol{u}_{i,j} = \boldsymbol{U}\boldsymbol{x}_{i,j}, \tag{4}$$

where $\boldsymbol{U}$ is the low-dimensional subspace containing within-client variation, and $\boldsymbol{x}_{i,j} \sim \mathcal{N}(0, \mathcal{I})$. Similarly, the client-dependent offset, $\boldsymbol{d}_i$, can be expressed as:

$$\boldsymbol{d}_i = \boldsymbol{D}\boldsymbol{z}_i, \tag{5}$$

where $\boldsymbol{D}$ is a diagonal matrix derived from the diagonal variances of the UBM, and $z_i \sim \mathcal{N}(0, \mathcal{I})$. Latent variables $\boldsymbol{x}_{i,j}$ and z_i are estimated during enrollment and the client specific super-vector (*i.e.,* the client model) is finally given by:

$$\boldsymbol{c}_i = \boldsymbol{m} + \boldsymbol{d}_i \tag{6}$$

At verification time, this model is compared to the super-vector extracted from the probe image to generate a score.

C. Deep Convolutional Neural Networks

Thanks to the explosion of available data and the constant advancement in computing power, computer vision tasks can nowadays fully benefit of deep learning approaches. Within this framework, detection and recognition tasks are typically performed using (deep) Convolutional Neural Networks (CNN) [31], and such models have become the *de facto* standard. State-of-the-art systems in face recognition are no exceptions, and among the huge amount of CNN-based approaches addressing this problem, one can mention DeepFace [32], FaceNet [33], and VGG [29]. All these models relies on massive datasets to be trained (up to 200 millions images for [33]) and training them from scratch is not a trivial task. In our contribution, deep convolutional networks are used as feature extractors to represent the identity of a client. In particular, the output of a specific layer in the network (an *embedding*) is used as a feature vector. Classification is then performed using a nearest-neighbour classifier with the cosine distance:

$$d(x, y) = 1 - \frac{x \cdot y}{||x|| \cdot ||y||} \tag{7}$$

where x is a feature extracted from an enrollment image and y a feature extracted from a probe image. To derive the feature vector x, four different architectures are investigated in this work: they are described in more details in the next subsections.

1) VGG: The architecture of the VGG network originates from a very deep network designed in the first place for object recognition [22]. It is very deep in the sense that it contains no less than 13 convolutional layers, followed by 3 fully-connected layers. It has a substantial amount of free parameters (138 millions) and was hence trained on a dataset containing around 2.6 millions images. Fortunately, a pre-trained VGG model has been made available to the research community.[3] In our work, it is used both as-is and when fine-tuned on the FARGO training data. The output of the second fully-connected layer (FC7) is used as a feature to represent a face image.

2) Light CNN: The Light CNN [23] is a lightweight Convolutional Neural Network introducing the so-called Max-Feature-Map (MFM) operation. The benefits of MFM is twofold: it allows to obtain a more compact representation while increasing the robustness of the model by performing filter feature selection. Furthermore, the authors trained the proposed architecture using a semantic bootstrapping method to mitigate the effect of noisy labels present in large databases. In this work, the architecture containing 29 layers has been used, and the released pre-trained model[4] is again used both as-is and with fine-tuning. Finally, it is worth mentioning that this architecture has 10 times less parameters than VGG and is 5 times faster, making it a more practical for a deployment on mobile devices.

3) DR-GAN: The recently proposed DR-GAN model [24] is especially dedicated to face recognition across pose, and relies on Generative Adversarial Networks (GAN) [34]. In this framework two networks compete against each other: a generator tries to synthetize "fake" images to fool the discriminator, and makes it believe that they are actually real. The DR-GAN model proposes an encoder-decoder architecture for the generator, which is able to synthetize a face image of a *given identity, and at a given pose*. For this purpose, an input face image x is first forwarded through a CNN (G_{enc}) to encode the identity $f(x)$. Then the encoded identity is concatenated with a conditional pose c and some input noise z. Finally, this concatenated vector is given as input to a neural network (G_{dec}) consisting in several layers of deconvolution operations, which result in an image of the same dimension of the input. Its architecture is depicted in Figure 5.

To perform recognition, this model can be used either to synthetize frontal face images as a pre-processing step to any frontal face verification algorithm, or by directly using

[3] http://www.robots.ox.ac.uk/∼vgg/software/vgg_face/

[4] https://github.com/AlfredXiangWu/LightCNN

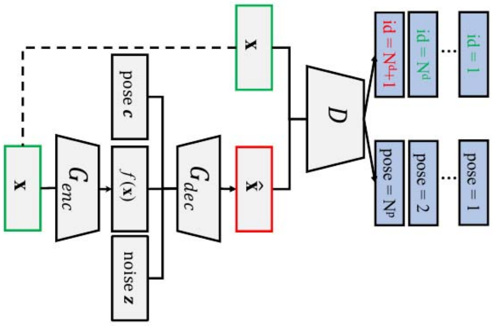


Fig. 5. Encoder-decoder architecture of the DR-GAN.

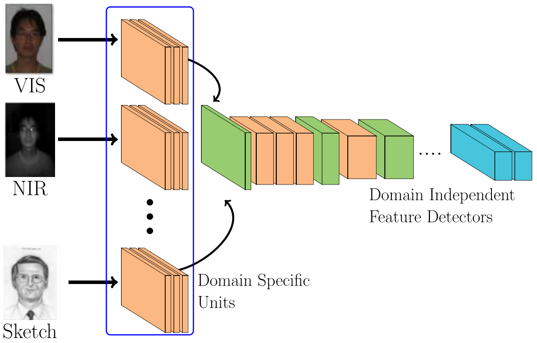


Fig. 6. Domain Specific Units - General Schematic.

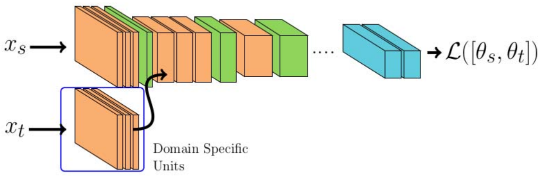


Fig. 7. Siamese Neural Networks training: pairs of images from both the source (x_s) and the target (x_t) domains are forwarded. Only the first layers (blue box), specific to the target modality, are udapted after error backpropagation.

pose-independent identity vectors $f(x)$. In our work, the latter approach has been used. The original DR-GAN model can be freely downloaded,[5] and the results presented in this contribution have been obtained with this pre-trained model.

4) Domain Specific Units: The Domain Specific Units (DSU) algorithm has been proposed in [25] for the task of Heterogeneous Face Recognition. It hypothesizes that Deep Convolutional Neural Networks high level features trained with RGB images are domain independent. With that established, the task of face recognition between different image domains (RGB to NIR, RGB to thermograms, etc.) can be carried out by adapting the low level features of the target domain only. High level features (Domain Independent Feature Detectors) are hence shared between all image modalities and the low level ones are trained specifically for each image domain (see Figure 6).

Using two different base architectures and two different methods to train such networks, the authors showed recognition rates improvements in three different image domains (RGB images to NIR, RGB to sketches and RGB to Thermal). In this work, the best setup from [25] is used. The selected architecture is the Inception ResNet v2 [33], which has been pre-trained with grayscale images of the MSCeleb-1M dataset [35]. The pre-trained model has been made available for download.[6] Using this model as a prior, the Domain Specific Units (i.e., the first 5 layers) are trained using the Siamese Neural Network method with the training set of the FARGO database with both source and target modalities. In this case, pairs of face images are processed: the image from the source domain is forwarded to the main network, and the image form the target domain is first forwarded through its own few first layers, and then to the remaining layers of the main network (which is shared among different image modalities). The error $\mathcal{L}(\theta_s, \theta_t)$ is then backpropagated, and only the modality specific layers are updated. Figure 7 shows the DSU training procedure.

IV. Experiments & Results

In this section, experimental results are presented for a variety of scenarios and tasks. First, performance measures are defined before proceeding with experiments on frontal face verification in different illumination conditions. Different fusion experiments - on both algorithms and modalities - are then shown to improve verification performance. Baselines for face verification across pose are then presented, and the section ends with experiments on heterogeneous face verification.

A. Performance Measures

As required by ISO standard [36], the performance of a biometric verification system should be reported in terms of *False Match Rate* (FMR) and *False Non-Match Rate* (FNMR). The FMR is defined as the expected probability that a zero-effort impostor sample will be falsely declared to match a enrolled client template. Mathematically, the FMR is computed as:

$$FMR = \frac{\text{\# of accepted impostor accesses}}{\text{\# of impostor accesses}} \tag{8}$$

Conversely, the FMNR is defined as the expected probability that a client sample will be falsely declared not to match its enrolled template. Mathematically, the FNMR is computed as:

$$FNMR = \frac{\text{\# of rejected client accesses}}{\text{\# of client accesses}} \tag{9}$$

Since both the FMR and the FNMR depends on a threshold τ, they are strongly related to each other: increasing the FMR will reduce the FNMR and vice-versa. For this reason, verification results are often presented using either Receiver Operating Characteristic (ROC) or Detection-Error Tradeoff (DET) curves, which basically plot the FMR versus the FNMR for different thresholds [37]. In this contribution, and to provide a single measure, the FNMR at FMR = 1% is reported. Note however that in following Tables, both FNMR and FMR

[5] http://cvlab.cse.msu.edu/project-dr-gan.html

[6] https://gitlab.idiap.ch/bob/bob.bio.face_ongoing

TABLE III
FMR AND FMNR [%] ON THE EVALUATION SETS IN THE VISUAL (RGB) SPECTRUM

	MC		UD		UO	
	FMR	FNMR	FMR	FNMR	FMR	FNMR
GGG	1.6	5.2	2.5	70.0	0.8	69.3
ISV	0.6	2.4	**2.1**	**57.7**	0.8	59.2
VGG	0.1	5.2	0.8	72.0	0.8	45.7
VGG (ft)	5.2	5.0	1.7	88.7	0.9	85.2
LCNN	0.9	66.4	0.0	100.0	0.6	96.3
LCNN (ft)	**0.3**	**0.4**	6.3	80.4	**1.7**	**38.0**

TABLE IV
FMR AND FMNR [%] ON THE EVALUATION SETS IN THE NEAR-INFRARED (NIR) SPECTRUM

	MC		UD		UO	
	FMR	FNMR	FMR	FNMR	FMR	FNMR
GGG	1.4	14.0	1.8	15.1	0.9	80.9
ISV	1.0	1.2	0.8	3.0	0.4	62.1
VGG	1.2	29.4	1.6	24.5	1.2	62.3
VGG (ft)	9.8	13.2	0.5	69.1	0.7	90.4
LCNN	0.6	92.2	0.6	96.4	0.2	99.0
LCNN (ft)	**0.5**	**0.2**	**0.5**	**1.6**	**0.5**	**26.0**

are reported *on the evaluation set*: the threshold reaching a FMR of 1% is selected *a priori* on the development set. As a consequence, applying the same threshold on the evaluation set usually leads to a FMR close to 1%, but with some variations.

B. Face Verification Under Difficult Illumination

In this section, experiments on frontal face verification with strong mismatch in terms of illumination conditions are presented. Experiments have been performed for all the protocols defined in Section II-A1 and on each modality. This allows to (i) assess baseline results for each of the algorithms and (ii) quantify their performance in mismatched conditions. Furthermore, since images and protocols are the same across the different data streams, the differences across modalities can be easily compared and discussed.

1) Preprocessing: In all experiments, the face is first located using a detector based on a boosted cascade of LBP features [38]. For both the Gabor Grid Graph and ISV algorithms, registered faces are cropped, converted to grayscale and resized to 84x60 pixels. Face images are photometrically preprocessed using LBP normalization [39] in the case of GGG, or with the Tan-Triggs algorithm [40] in the case of ISV. For VGG, face images have been cropped and resized according to the requirements of the model (224x224). In the case of LightCNN, face images were cropped, resized to 128x128 and converted to grayscale. Other than that, all algorithms are used with their default parameters. Note that the ISV background model has been trained using FARGO data only, in contrast to the pre-trained CNN models (VGG and LightCNN), which have been pre-trained on various large-scale databases. Note finally that GGG does not require any training phase.

2) Evaluation in the Visual Spectrum: Table III presents results obtained on RGB images. Note that for approaches relying on deep neural networks (i.e., VGG and LightCNN), results are presented with both the released pre-trained model and also with the model fine-tuned using FARGO training data, containing clean conditions only. As expected, all algorithms in the Matched Controlled (MC) protocol present low error rates. Indeed, in this case, images used for enrollment and for probes are very similar to each other. A notable exception is the pre-trained LightCNN: this suggests that the databases used for training this model may be very different from ours.

When considering the more challenging UD protocol, where probe images are acquired in a very dark room, one can observe a dramatic drop in performance of all algorithms, as expected. An interesting result though, is that ISV is performing better than all deep learning approaches, despite the fact that it has been trained on clean FARGO data only. This may be explained by the ability of the ISV algorithm to properly capture - and suppress to some extent - the within-class variability. However, performance in this case remains far from being acceptable: with a FNMR of 57.7%, a legitimate client has a higher chance to be rejected than to be accepted. Although such conditions may seem extreme (see Figure 2), it may happen in a real-life scenario: think of a person willing to unlock his smartphone in a basement for instance. On the UO protocol, the fine-tuned LightCNN model achieves the best performance with a FMNR of 38%. This can be explained by the fact that this model has been pre-trained on a very large database, which likely contains outdoor conditions. While results on the UO protocol are generally better than in dark conditions, they could not be considered as satisfactory either.

It is also worth noting the different behavior in both CNN models: VGG usually reaches better performance when it is *not* fine-tuned while the opposite is observed for LightCNN. This may be due to the large number of parameters in VGG, in conjunction with the small amount of training data to perform the fine-tuning. This may lead to an overfitting to the (clean) training conditions of the FARGO database. On the other hand, since LightCNN contains fewer parameters, this phenomenon is not observed and results are significantly better when the model is fine-tuned. Note that this observation contradicts the claim in [23], where authors declare that “Light CNN achieves state-of-the-art results on five face benchmarks without supervised fine-tuning”. This also suggests that the proposed database contains conditions that are not present in traditionally used large face databases, typically collected from the Web. Overall, this first experiment shows that face recognition in unconstrained environment is still a current issue, at least with conventional imaging methods.

3) Evaluation in the Near-Infrared Spectrum: In this section, the same experiments as before, but on the Near-Infrared (NIR) spectrum are presented. Note that here, the FARGO training data contains NIR images. As a consequence, the fine-tuning of the CNNs has been made using this modality.

Table IV shows the performance using near-infrared images. As in the previous case, the fine-tuned LightCNN model reaches the best performance on the MC protocol, and the

TABLE V
FMR AND FMNR [%] ON THE EVALUATION SETS ON DEPTH MAPS

	MC		UD		UO	
	FMR	FNMR	FMR	FNMR	FMR	FNMR
GGG	0.9	35.2	1.2	45.2	0.2	92.7
ISV	**0.4**	**30.0**	**0.6**	**53.8**	**0.9**	**92.0**
VGG	0.2	94.0	0.1	96.6	1.2	96.9
VGG (ft)	15.3	36.4	0.4	92.9	2.2	95.1
LCNN	0.5	97.4	0.4	98.1	0.3	99.4
LCNN (ft)	0.3	79.8	0.2	76.0	0.4	98.1

FNMR is even lower (0.2%) than with color images, again suggesting than fine-tuning is effective for this model. An interesting, although expected, observation is that NIR images drastically improves the performance in the case of very low-light condition, and this is the case for all algorithms. However, when considering images taken outdoors, all algorithms but the fine-tuned LightCNN perform worse than on RGB images. Indeed, NIR imaging has issue to properly deal with such conditions, and does not provide good quality images. This may be due to the stronger near-infrared components contained in sunlight [41].

4) Evaluation on Depth Maps: Since methods devised for traditional two-dimensional face verification have already been applied directly to depth maps (see [42] for instance), and for the sake of completeness, results of our baseline algorithms applied on depth images are provided here as well. Note that in this case, no special preprocessing has been applied to the face image: photometric normalization is not applicable on depth maps. For registration, we used the detection results obtained from the face detector on NIR images, since these streams are both temporally and spatially aligned.

Table V presents results obtained using depth data. While good performance was not expected using only such data, the UD protocol provides an interesting insight for GGG and ISV. In the case of dark conditions, and despite the coarseness of depth data, better performance is observed using this modality as compared to RGB images. This could be explained by the fact that both of these algorithm are acting on depth data only - as opposed to CNN approaches, where pre-trained models on RGB images have been used. In outdoor conditions however, depth data proves to be useless: good quality depth information is hard to obtain and hence contain many holes, for the same interference reasons as with NIR.

5) Fusion: Since different algorithms and different modalities are available, the usage of score fusion is presented here. Fusion can help face verification performance, and in our case, it makes sense to investigate this strategy from two aspects:

1) From the *algorithms* point of view: since the algorithms used in this work are very different from each other, they might be complementary. Some algorithms can fail in certain conditions where others would be successful.
2) From the *modality* point of view: different modalities are available and as a consequence, one should consider fusing verification results obtained with each of them. Indeed, NIR images in dark conditions are obviously a useful complement to RGB images acquired in the same condition for instance.

Investigating all possible combinations between algorithms and modalities fusion is beyond the scope of this contribution. Experiments have hence been performed using three algorithms in the same modality, and the same algorithm was used across different modalities. In particular, we did not consider fusing different algorithms in different modalities (i.e., VGG in RGB fused with ISV in NIR for instance). In our work, scores were fused using an approach based on logistic regression; scores from different systems are combined into a new feature vector, which is then classified as either coming from the true client, or being an impostor. This approach has already been successfully applied in biometric verification [43].

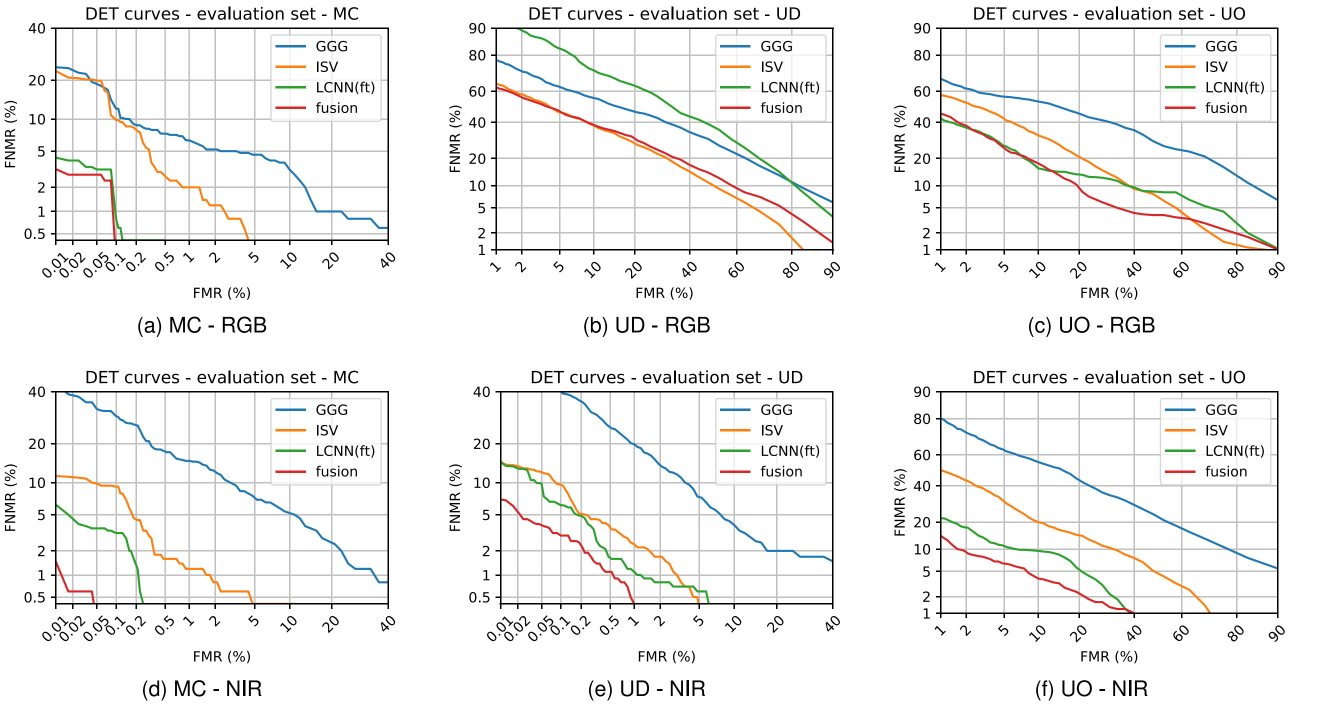
Algorithm Fusion: Figure 8 shows DET curves on the evaluation sets for the different protocols. Here we report result when fusing GGG, ISV and LCNN(ft). These algorithms have been selected since they are very different from each other, hence having a better chance to be complementary. Also, LCNN(ft) has been chosen since it the best performing CNN. The first row of Figure 8 shows results of algorithms fusion in RGB and the second row in NIR.

As can be observed on the curves in Figure 8, fusing the algorithms generally yields better performance than any of the algorithms alone *for each condition and in each modality*. This shows that these three algorithms are complementary to each other: one may fail on certain probe images, where others will be successful. However, the increase in performance improvement is not significant in the visual spectrum. On the other hand, fusing the different algorithms in NIR is more effective and performance is improved in all three protocols, with FNMRs going from 0.2% to 0.0%, from 1.6% to 1.0%, and from 26.0% to 14.4% for the MC, UD and UO protocol respectively.

Modalities Fusion: Figure 9 shows DET curves on the evaluation sets when fusing modalities. Note that here we only provide results for the LCNN(ft) algorithm, since it has been applied on the three modalities, and presents the lowest error rates on average. Note in this case that performance obtained with fusion does not necessarily improve the result, and when it does, the increase in performance is quite marginal. There are, in our opinion, two main explanations for this behaviour: results obtained with depth data may be more confusing than useful, and actually penalise the fusion. Second, it is clear that NIR imaging can cope with low lighting conditions, and again, other modalities are not helping in this particular case. However in the UO protocol, fusing the different modalities helps at achieving better performance, suggesting that RGB and NIR modalities carry complementary information.

C. *Face Verification Across Pose*

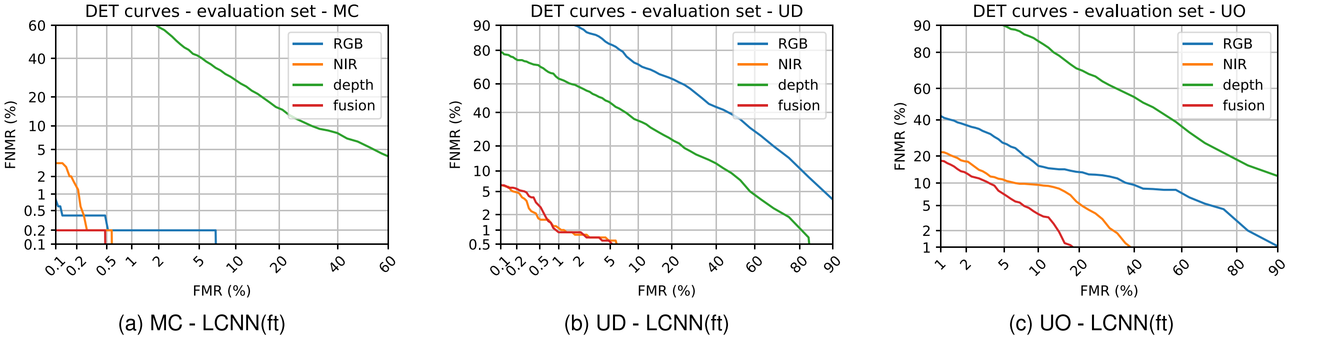
In this section, experiments for face verification in different pose conditions are presented. Experiments are performed for two different pose variations, namely pitch and yaw. Basically, the same settings have been used here as in Section IV-B on frontal face verification (i.e., preprocessing and default parameters). Additionally, results are also provided for a



(a) MC - RGB (b) UD - RGB (c) UO - RGB

(d) MC - NIR (e) UD - NIR (f) UO - NIR

Fig. 8. DET curves of different algorithms and their fusion on the evaluation sets for frontal face verification protocols. The first row presents results obtained in the visual spectrum, and the second row shows results on the near-infrared spectrum.



(a) MC - LCNN(ft) (b) UD - LCNN(ft) (c) UO - LCNN(ft)

Fig. 9. DET curves of different modalities and their fusion on the evaluation sets for frontal face verification protocols. DET curves have been obtained with the fine-tuned LightCNN algorithm.

TABLE VI
FMR AND FMNR [%] ON THE EVALUATION SETS FOR POSE-VARYING FACE VERIFICATION IN THE VISUAL SPECTRUM

	Yaw		Pitch	
	FMR	FNMR	FMR	FNMR
GGG	0.6	80.4	0.8	51.4
ISV	1.1	44.7	1.0	37.4
VGG	0.4	47.0	0.8	39.6
VGG (ft)	0.6	82.8	1.4	64.2
LCNN	1.1	83.7	1.1	84.0
LCNN (ft)	**1.1**	**21.0**	**1.0**	**18.5**
DR-GAN	0.2	57.0	0.2	55.3

state-of-the-art model especially dedicated to face recognition across pose: the so-called DR-GAN [24]. Experiments have been performed using RGB data only and error rates are presented in Table VI.

Generally, the error rates of the investigated baseline algorithms are not so low, with best FNMRs of 21% and 18.5% for yaw and pitch respectively. From these results, it can be seen that rotations in yaw are more difficult to address than rotations in pitch. This is probably due to the self-occlusion caused by large rotation in yaw, which are not as important with pitch. A surprising result is the poor performance obtained by the DR-GAN model, which is especially tailored for pose-varying face recognition, and has been trained on two databases containing face images with various rotations: CASIA WebFace [44] and AFLW [45]. This is in contrast to ISV, where *frontal images only* have been seen during both training and enrollment, and for which performance is significantly better.

D. Heterogeneous Face Verification

In this section, experiments on heterogeneous face verification (HFR) are presented. Experiments have been performed

TABLE VII
FMR AND FMNR [%] ON THE EVALUATION SETS USING RGB IMAGES FOR ENROLLMENT AND NIR IMAGES FOR PROBES

	MC		UD		UO	
	FMR	FNMR	FMR	FNMR	FMR	FNMR
No adapt.	1.4	9.8	1.9	22.7	0.8	25.7
DSU	**0.8**	**1.0**	**1.0**	**4.7**	**1.1**	**8.5**

TABLE VIII
FMR AND FMNR [%] ON THE EVALUATION SETS USING RGB IMAGES FOR ENROLLMENT AND DEPTH MAPS FOR PROBES

	MC		UD		UO	
	FMR	FNMR	FMR	FNMR	FMR	FNMR
No adapt.	8.1	91.2	7.4	93.8	5.8	92.8
DSU	**1.2**	**95.8**	**3.4**	**90.5**	**3.2**	**97.3**

on the MC, UD and UO protocols. In this case, unlike experiments with single modalities of Section IV-B, the training set now contains images from both the source and target domain, to be used in the Siamese training approach (see Figure 7). Besides, images used for enrollment are coming from the source domain (i.e., RGB), whereas probes come from the target domain (i.e., NIR or depth). Here, we are interested in (i) assessing the behaviour of the selected CNN in transfer learning between image domains, and (ii) in quantifying the benefits of using Domain Specific Units adaptation.

1) RGB to NIR: Table VII presents FMR and FNMR obtained on the evaluation set for the protocols defined in Section II-A1. One can see that without adaptation (first line), performance is generally worse than the one obtained when considering NIR images for both enrollment and probe presented in Section IV-B3 (see Table IV). However, when the CNN is adapted to account for the target modality, performance is significantly improved. In the challenging UO protocol, the FNMR goes down to 8.5%, despite the fact that enrollment and probe images come from different modalities. This suggests that (i) Domain Specific Units adaptation is of great help, and that (ii) the selected architecture (Inception ResNet v2) in conjunction with the Siamese training approach is very effective: indeed, this is the best performance achieved so far in outdoor conditions.

2) RGB to Depth: Table VIII presents the two possible error rates (FMR and FNMR) using the RGB to depth data in the evaluation set. One can notice that performance using depth information as probes is quite low, whether or not Domain Specific Units adaptation is performed. These results are similar to those presented in Table V: it shows that applying CNNs pre-trained with images in the visual spectrum are not able to properly deal with depth data. In this case, using Domain Specific Units adaptation does not guarantee an improvement in performance - on the contrary to the RGB to NIR case. Again, it stresses that the task of recognizing people using low-quality depth maps as probes (with a gallery of RGB images) is a very challenging task: none of the models (without and with DSU adaption) presents FNMR lower than 90%.

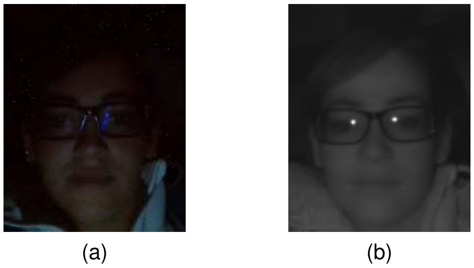


Fig. 10. Examples where RGB fails (a), but NIR succeeds (b).

E. Discussion

In this section, a discussion of obtained results is made. It highlights successes, but also examines current challenges by providing qualitative results: misclassified face images in various scenarios are provided to illustrate limitations.

Experiments conducted on frontal face verification provide several insights. First, it should be noted that when images used for enrollment and for probes are acquired in clean illumination conditions, performance of current approaches are satisfactory, and such a problem can be considered as solved. However, when illumination conditions differs, the performance is not acceptable for a real-world, day to day usage. However, Table IV shows that near-infrared (NIR) imaging can mitigate the case of dark environmental conditions. Figure 10 provides an example where the best approach in the visual domain for this protocol (ISV) fails, but successfully authenticate the subject in NIR. As a consequence, using the near-infrared spectrum on dark images reaches performance *on par* with clean conditions. However, images acquired outdoor with natural lighting remain a challenge, both in the visual spectrum and using NIR information. Figure 11 presents examples where the fine-tuned LightCNN model fails at recognizing people in outdoor condition. The difference between the first row (enrollment image) and the second row (probe) is certainly noticeable, but not so severe - for instance, a human may not be fooled by such variations. It should also be noted that face detection has an influence on the overall performance. Indeed, our face detector returns the best face candidate in each image, and hence may contain approximative detections. Note finally that the captured depth maps are not of sufficient quality to be reliably used for face verification. This is most likely because subjects were not constrained to remain still at an optimal distance. However, it may provide valuable information when used in conjunction with other modalities.

Score fusion, whether between algorithms or modalities, can help to some extent, and improves performance in all investigated scenarios. But the gap considering the MC protocol as a reference is still substantial. Nevertheless, score fusion has not been thoroughly addressed in this study and further investigation may reveal significant improvement. Also, using different sources of information, for frontalizing faces or for fusing different features are directions worth exploring, see [46] and [47] for examples.

Regarding face recognition across pose, conducted experiments showed that all investigated algorithms still have trouble

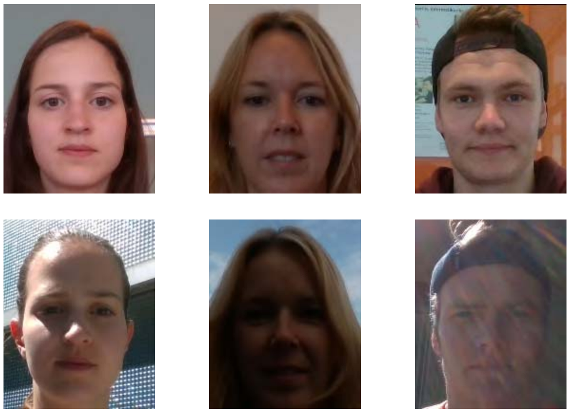


Fig. 11. Example of enrollment images (first row) and probes leading to a false non match (second row).

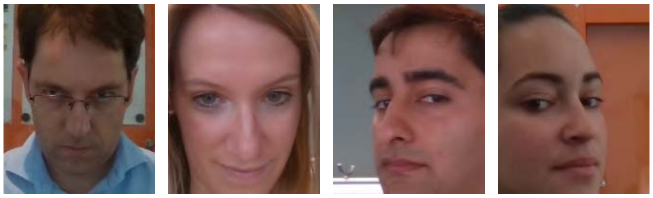


Fig. 12. Examples of probe images leading to a False Non Match.

to reach low error rates. Figure 12 shows examples of False Non Match declared by the fine-tuned LightCNN: it should be noted that these images show faces with little pose distortion, where all facial features are visible. This shows that even moderate variations can be source of confusion.

Finally, experiments performed in the context of heterogeneous face verification obtained the best results on the difficult outdoor conditions. This is certainly due to the usage of a state-of-the-art deep neural networks with a particularly well-suited training mechanism (i.e., Inception ResNet v2 with Siamese training). Indeed, when applying this architecture directly to NIR images without adaptation, performance on the UO protocol is already *on par* with LCNN(ft) in the same settings. Further, Domain Specific Units adaptation leads to even better results, with an FNMR of 8.5%.

As a consequence, and to fully assess this promising model, further experiments have been conducted using the Inception ResNet v2 with images in the visual spectrum, on the different scenarios defined in Sections II-A1 and II-A2. In this case, the fine-tuning was done using the Siamese network approach and the training set of the FARGO database, containing clean and frontal RGB face images only (i.e., the model was not adapted for other modalities as done in Section IV-D). Obtained results for face verification under difficult illumination conditions and for varying poses are presented in Table IX and Table X: they should be compared to Table III and Table VI respectively.

TABLE IX
FMR AND FMNR [%] FOR FACE VERIFICATION UNDER DIFFICULT ILLUMINATION, USING THE INCEPTION RESNET V2 MODEL

	MC		UD		UO	
	FMR	FNMR	FMR	FNMR	FMR	FNMR
ResNet v2	1.6	0.0	2.9	17.3	1.4	6.5

TABLE X
FMR AND FMNR [%] FOR POSE-VARYING FACE VERIFICATION, USING THE INCEPTION RESNET V2 MODEL

	Yaw		Pitch	
	FMR	FNMR	FMR	FNMR
ResNet v2	1.6	4.2	1.5	2.7

This state-of-the-art model obtain substantially lower error rates, as compared to the baseline algorithms used earlier. Indeed, for comparable FMR, the FNMR is reduced from 57.7% to 17.3% for the UD protocol, and from 45.7% to 6.5% for the UO protocol. An even better improvement is observed for pose-varying face verification, going from an FNMR of 47.0% to 4.2% and from 39.6% to 2.7% for variations in yaw and pitch respectively.

V. CONCLUSION

This contribution introduced a new face database, consisting in video sequences captured using latest imaging devices and comprising several modalities: RGB, NIR and depth maps. Data have been acquired in various, difficult illumination conditions and with varying pose as well, allowing to address current challenges in face recognition. Several experimental protocols have been proposed, and baselines results using publicly available algorithms have been presented. Our experiments show that strong illumination mismatch and pose variation are still challenging issues for existing face recognition approaches, including fine-tuned deep neural networks. As expected, dark illumination conditions can be mitigated by using data captured in the Near-Infrared spectrum. Also, score fusion, whether across different modalities or across different face recognition approaches, has been shown to improve performance. Finally, the task of heterogeneous face verification has been addressed using a state-of-the-art model based on the Inception ResNet v2 and significant improvements have been attained. This is especially true for the task of RGB to NIR. On the contrary, this model was not able to perform well in the RGB to depth settings. Considering the good performance obtained by this model, it has also been applied to the aforementioned tasks (face verification under difficult illumination and with pose variation) and significantly surpassed all the baselines.

It is our hope that this new dataset will help the research community to address the identified challenges in face recognition, but also enable various research endeavours in face recognition and related tasks. Indeed, in complement to face recognition, the proposed dataset allows to explore the combination of several modalities in a wide variety of tasks such as facial feature localization and tracking, facial animation, or head pose estimation for instance. Furthermore, the recent introduction of advanced imaging capabilities in latest consumer-grade devices (i.e., the iPhone X) makes this dataset particularly interesting for research purposes. Note finally that the data and the code to reproduce presented experiments has been made freely available, and could be easily extended.

Acknowledgment

The authors would like to thank ams AG and more specifically its Martigny Lab for its valuable help in building the dataset.

References

[1] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM Comput. Surveys*, vol. 35, no. 4, pp. 399–458, 2003.

[2] H. Zhou, A. Mian, L. Wei, D. Creighton, M. Hossny, and S. Nahavandi, "Recent advances on singlemodal and multimodal face recognition: A survey," *IEEE Trans. Human–Mach. Syst.*, vol. 44, no. 6, pp. 701–716, Dec. 2014.

[3] C. Ding and D. Tao, "A comprehensive survey on pose-invariant face recognition," *ACM Trans. Intell. Syst. Technol.*, vol. 7, no. 3, pp. 1–42, 2016.

[4] S. Bhattacharjee, A. Mohammadi, and S. Marcel, "Spoofing deep face recognition with custom silicone masks," in *Proc. IEEE Int. Conf. Biometrics Theory Appl. Syst. (BTAS)*, 2018, pp. 1–7.

[5] S. Romdhani, J. Ho, T. Vetter, and D. J. Kriegman, "Face recognition using 3-D models: Pose and illumination," *Proc. IEEE*, vol. 94, no. 11, pp. 1977–1999, Nov. 2006.

[6] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter, "A 3D face model for pose and illumination invariant face recognition," in *Proc. IEEE Int Conf. Adv. Video Signal Based Surveillance*, 2009, pp. 296–301.

[7] Y. Atoum, Y. Liu, A. Jourabloo, and X. Liu, "Face anti-spoofing using patch and depth-based CNNs," in *Proc. Int Joint Conf. Biometrics*, 2017, pp. 319–328.

[8] G. Fanelli, T. Weise, J. Gall, and L. Van Gool, "Real time head pose estimation from consumer depth cameras," in *Proc. Joint Pattern Recognit. Symp.*, 2011, pp. 101–110.

[9] R. I. Hg, P. Jasek, C. Rofidal, K. Nasrollahi, T. B. Moeslund, and G. Tranchet, "An RGB-D database using Microsoft's kinect for windows for face detection," in *Proc. Int Conf. Signal Image Technol. Internet Based Syst. (SITIS)*, 2012, pp. 42–46.

[10] B. Y. Li, A. S. Mian, W. Liu, and A. Krishna, "Using kinect for face recognition under varying poses, expressions, illumination and disguise," in *Proc. IEEE Workshop Appl. Comput. Vis. (WACV)*, 2013, pp. 186–192.

[11] R. Min, N. Kose, and J.-L. Dugelay, "KinectFaceDB: A kinect database for face recognition," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 44, no. 11, pp. 1534–1548, Nov. 2014.

[12] A. Chowdhury, S. Ghosh, R. Singh, and M. Vatsa, "RGB-D face recognition via learning-based reconstruction," in *Proc. IEEE Int. Conf. Biometrics Theory Appl. Syst. (BTAS)*, 2016, pp. 1–7.

[13] J. Zhang, D. Huang, Y. Wang, and J. Sun, "Lock3DFace: A large-scale database of low-cost kinect 3D faces," in *Proc. Int. Conf. Biometrics (ICB)*, 2016, pp. 1–8.

[14] R. S. Ghiass, O. Arandjelović, A. Bendada, and X. Maldague, "Infrared face recognition: A comprehensive review of methodologies and databases," *Pattern Recognit.*, vol. 47, no. 9, pp. 2807–2824, 2014.

[15] S. Farokhi, J. Flusser, and U. U. Sheikh, "Near infrared face recognition: A literature survey," *Comput. Sci. Rev.*, vol. 21, pp. 1–17, Aug. 2016.

[16] S. Z. Li, D. Yi, Z. Lei, and S. Liao, "The CASIA NIR-VIS 2.0 face database," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2013, pp. 348–353.

[17] J. Bernhard, J. Barr, K. W. Bowyer, and P. Flynn, "Near-IR to visible light face matching: Effectiveness of pre-processing options for commercial matchers," in *Proc. IEEE Int. Conf. Biometrics Theory Appl. Syst. (BTAS)*, 2015, pp. 1–8.

[18] F. Liu, J. Hu, J. Sun, Y. Wang, and Q. Zhao, "Multi-dim: A multidimensional face database towards the application of 3D technology in real-world scenarios," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, 2017, pp. 342–351.

[19] S. Z. Li, Z. Lei, and M. Ao, "The HFB face database for heterogeneous face biometrics research," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2009, pp. 1–8.

[20] M. Günther, D. Haufe, and R. P. Würtz, "Face recognition with disparity corrected Gabor phase differences," in *Proc. Int Conf. Artif. Neural Netw. (ICANN)*, vol. 7552, Sep. 2012, pp. 411–418.

[21] R. Wallace, M. McLaren, C. McCool, and S. Marcel, "Inter-session variability modelling and joint factor analysis for face authentication," in *Proc. Int. Joint Conf. Biometrics (IJCB)*, 2011, pp. 1–8.

[22] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015.

[23] X. Wu, R. He, Z. Sun, and T. Tan, "A light CNN for deep face representation with noisy labels," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 11, pp. 2884–2896, Nov. 2018.

[24] L. Tran, X. Yin, and X. Liu, "Disentangled representation learning GAN for pose-invariant face recognition," in *Proc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 1283–1292.

[25] T. D. F. Pereira, A. Anjos, and S. Marcel, "Heterogeneous face recognition using domain specific units," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 7, pp. 1803–1816, Jul. 2019.

[26] E. Bailly-Bailliére *et al.*, "The BANCA database and evaluation protocol," in *Proc. Int Conf. Audio Video Based Biometric Pers. Authentication (AVBPA)*, 2003, pp. 625–638.

[27] C. McCool *et al.*, "Bi-modal person recognition on a mobile phone: Using mobile phone data," in *Proc. IEEE ICME Workshop Hot Topics Mobile Multimedia*, Jul. 2012, pp. 635–640.

[28] M. Günther, L. El Shafey, and S. Marcel, "Face recognition in challenging environments: An experimental and reproducible research survey," in *Face Recognition Across the Imaging Spectrum*, 1st ed., T. Bourlai, Ed. Cham, Switzerland: Springer, Feb. 2016.

[29] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2015, pp. 1–12.

[30] L. Wiskott, J.-M. Fellous, N. Krüger, and C. Von Der Malsburg, "Face recognition by elastic bunch graph matching," in *Intelligent Biometric Techniques in Fingerprint and Face Recognition*, L. C. Jain, U. Halici, I. Hayashi, and S. B. Lee, Eds. Boca Raton, FL, USA: CRC Press, 1999, ch. 11, pp. 355–396.

[31] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[32] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "DeepFace: Closing the gap to human-level performance in face verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2014, pp. 1701–1708.

[33] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 815–823.

[34] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2014, pp. 2672–2680.

[35] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "MS-Celeb-1M: A dataset and benchmark for large-scale face recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 1–17.

[36] *Information Technology—Biometric Performance Testing and Reporting—Part 1: Principles and Framework*, Standard ISO/IEC 19795-1:2006, 2006.

[37] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," in *Proc. Eurospeech*, 1997, pp. 1895–1898.

[38] C. Atanasoaei, "Multivariate boosting with look-up tables for face processing," Ph.D. dissertation, Eng. Sci. Technol., Ecole Polytechnique Federale de Lausanne, Lausanne, Switzerland, 2012.

[39] G. Heusch, Y. Rodriguez, and S. Marcel, "Local binary patterns as an image preprocessing for face authentication," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit. (AFGR)*, 2006, pp. 9–14.

[40] X. Tan and B. Triggs, "Enhanced local texture feature sets for face recognition under difficult lighting conditions," in *Proc. IEEE Int. Workshop Anal. Model. Faces Gestures (AMFG)*, 2007, pp. 168–182.

[41] D. Yi, R. Liu, R. Chu, R. Wang, D. Liu, and S. Z. Li, "Outdoor face recognition using enhanced near infrared imaging," in *Proc. Int. Conf. Biometrics*, 2007, pp. 415–423.

[42] C. McCool, J. Sanchez-Riera, and S. Marcel, "Feature distribution modelling techniques for 3D face verification," *Pattern Recognit. Lett.*, vol. 31, no. 11, pp. 1324–1330, 2010.

[43] I. Chingovska, A. Anjos, and S. Marcel, "Biometrics evaluation under spoofing attacks," *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 12, pp. 2264–2276, Dec. 2014.

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

[45] M. Köstinger, P. Wohlhart, P. M. Roth, and H. Bischof, “Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization,” in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, 2011, pp. 2144–2151.

[46] G. Goswami, M. Vatsa, and R. Singh, “RGB-D face recognition with texture and attribute features,” *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 10, pp. 1629–1640, Oct. 2014.

[47] Y. Lee, J. Chen, C.-W. Tseng, and S.-H. Lai, “Accurate and robust face recognition from RGB-D images with a deep learning approach,” in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2016, pp. 1–14.

Guillaume Heusch received the M.Sc. degree in communication systems and the Ph.D. degree in electrical engineering from the Ecole Polytechnique Fédérale de Lausanne in 2005 and 2010, respectively. Then, he spent several years working as a Computer Vision Research Engineer in various industries. He is currently a Research Associate with Idiap Research Institute. His current research interests are computer vision, machine learning and, on a broader perspective, the extraction of meaningful information from raw data.

Tiago de Freitas Pereira (M’18) received the B.S. degree in computer science from the University of São Paulo in 2010 and the M.S. degree in electrical engineering from the University of Campinas in 2013. He is currently pursuing the Ph.D. degree from the Ecole Polytechnique Fédérale de Lausanne. He is currently a Research Assistant with Idiap Research Institute. His current interests include face recognition, pattern recognition, image processing, and machine learning.

Sébastien Marcel (SM’17) received the Ph.D. degree in signal processing from the Université de Rennes I, France, in 2000, at CNET, the research center of France Telecom (currently, Orange Labs). He is the Head of the Biometrics and Security Group and a Senior Research Scientist with Idiap Research Institute. He leads a team conducting research on multimodal biometrics (face, speaker, and vein) and presentation attack detection. He is currently interested in pattern recognition and machine learning with a focus on multimodal biometric person recognition. He serves as an Associate Editor for the IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY.