A comparative analysis of deep and shallow features for multimodal face recognition in a novel RGB-D-IR dataset

Tiago Freitas[1,2⋆], Pedro G. Alves[2], Cristiana Carpinteiro[2], Joana Rodrigues[2], Margarida Fernandes[2], Marina Castro[2], João C. Monteiro[1,2], Jaime S. Cardoso[1,2]

[1]INESC-TEC, Porto, Portugal;
[2]Faculty of Engineering University of Porto, Portugal;

Abstract. With new trends like 3D and deep learning alternatives for face recognition becoming more popular, it becomes essential to establish a complete benchmark for the evaluation of such algorithms, in a wide variety of data sources and non-ideal scenarios. We propose a new RGB-depth-infrared (RGB-D-IR) dataset, RealFace, acquired with the novel Intel® RealSense™ collection of sensors, and characterized by multiple variations in pose, lighting and disguise. As baseline for future works, we assess the performance of multiple deep and "shallow" feature descriptors. We conclude that our dataset presents some relevant challenges and that deep feature descriptors present both higher robustness in RGB images, as well as an interesting margin for improvement in alternative sources, such as depth and IR.

1 Introduction

Over the past few years, the issue of face recognition has been on the spotlight of many research works in pattern recognition, due to its wide array of real-world applications. The *face* is a natural, easily acquirable, trait with a high degree of uniqueness, representing one of the main sources of information during human interaction. These marked advantages, however, fall short when images of limited quality, acquired under unconstrained environments, are presented to the system. The fact that humans perform and rely on face recognition routinely and effortlessly throughout their daily lives leads to an increased interest in replicating this process in an automated way, even when above limitations are known to frequently occur.

Whereas technological improvements in image capturing and transmitting equipment managed to attenuate most noise factors, partial face occlusions, severe illumination changes and extreme pose variations still represent genuine challenges to automated face recognition [1,2]. Approaching these issues will, therefore, be a matter of either exploring new sources of data, to compensate

⋆ Mailing author: vilab.biometrics@gmail.com

the more traditional alternatives in less ideal scenarios, or designing more robust algorithms, capable of encompassing such limitations.

Recently, a new trend has been observed in face recognition works, with information from the three-dimensional structure of the face being incorporated in recognition frameworks, in an attempt to grant higher robustness in scenarios such as critically low illumination, where the extraction of color information is severely limited, or extreme pose variations. In conjunction with the more traditional color images, 3D data can be used to develop more robust multimodal approaches [3].

Research in automated face recognition has also found an interesting new alternative in methodologies based on Deep Learning, such as deep Convolutional Neural Networks (CNN). These approaches have shown increased performance in a multiplicity of image recognition tasks, due to their capacity to learn abstract and invariant high-level features when compared to the more traditional application-tailored features [4].

The development of biometric recognition systems is generally limited by the shortage of large public databases acquired under real unconstrained working conditions. Database collection represents a complicated process, in which a high degree of cooperation from a large number of participants is needed. For that reason, nowadays, the number of existing public databases that can be used to evaluate the performance of biometric recognition systems in real-life acquisition conditions and making use of multiple sources of information is quite limited.

Motivated by this need and the growing interest of the research community in both 3D and deep learning strategies for face recognition, we present a new database, named RealFace, acquired using the novel Intel® RealSense™ collection of sensors. In addition to this new dataset, we also establish an experimental setup and a performance baseline using a set of more traditional tailored feature descriptors as well as some deep learning alternatives. We aim to assess the accuracy, robustness and generalization capability of such features with regards to both color and 3D information, as well as establishing a solid baseline for further research in the biometrics scientific community. Finally, we present a new CNN, trained specifically for face recognition using depth representations of the 3D structure of the face, validated both on the state-of-the-art EURECOM dataset, as well as our proposed RealFace dataset.

2 Related Work

Following the good results obtained in object recognition by Krizhevsky et al. [5] using deep CNNs, their use has shown promising performance in many computer vision related tasks, as they can achieve more correct assumptions about the image's local pixel dependencies. This approach showed an absolute decrease in the error rate of about 10% when compared with densely-sampled SIFT keypoint descriptors applied to the same tasks [5]. In the field of face recognition, DeepFace [4] has achieved 97.35% accuracy and FaceNet [6] from Google has achieved 99.63% recognition accuracy on the benchmark Labeled Faces in the

Wild dataset. These results surpass those achieved with "shallow" methods like Local Binary Patterns (95.15% accuracy) [7] and SIFT [8] (93.03% accuracy), due to the ability of deep neural networks to better handle the large amounts of data present in such datasets, extracting and learning more high dimensional, invariant features. This translates in increased robustness of the system to variations in occlusion, pose and illumination.

Simpler networks have also been proposed, achieving results comparable to the state-of-the art, such as the VGG-Face network by Parkhi et al. [9]. This network has the advantage of being one of the few pre-trained, publicly available, deep CNNs. To the extent of our knowledge only a few approaches have made use of this network [10,11], as of the writing of this paper. All these works conclude that the VGG-Face can outperform more hand-crafted feature extraction methods, as it extracts highly discriminative and invariant face descriptors. This was further noted in the recent International Challenge on Biometric Recognition in the Wild (ICB-RW), where all the top-ranked algorithms made use of face descriptors extracted by the VGG-Face. A more thorough exploration of deep learning approaches in this field is deemed necessary, especially when new data sources, such as 3D and infrared data, are being made more easily accessible.

Recent technological advances have made it feasible to deploy low-cost alternatives to the more traditional high-cost 3D scanners, such as Minolta, Inspeck, CyberWare and 3dMD [12]. The appearance of Microsoft Kinect has opened a wide array of opportunities to include three-dimensional information in computer vision solutions that were, otherwise, limited by the wider availability of color images. These sensors provide two types of data: depth images and 3D models, that can be point clouds (PC) or meshes. 3D models consist in a representation that retains all geometric information of the head. On the other hand, depth images, or 2.5D, are bi-dimensional representations of a set of 3D points, in which each pixel in the XY plane stores the depth z value. While the use of multiple sources of information has been shown to improve performance in a vast number of biometric recognition works [**?**], the potential of using a single sensor to acquire multiple representations of the same data makes it worth to invest on such alternatives.

A number of datasets have already been built using RGB-D sensors. Some of these databases include the Aalborg University RGB-D Face Database, the Florence Superface Dataset, CurtinFaces, FaceWarehouse, EURECOM Kinect Face database, IIIT-D face database and the Labeled Infrared-Depth Face. A more detailed state-of-the-art-review regarding these datasets is presented in our previous work [**?**]. While the aforementioned datasets present a wide variety of conditions, there is still not enough available relevant public data that uses the more recent sensors like the Kinect v2 or the Intel® RealSense™ models. Data acquired with these sensors could present useful alternatives for the face biometrics research community. The scientific relevance of the Intel® RealSense™ sensor is even higher considering that, to the extent of our knowledge, no publicly available dataset was built on this novel sensor. Intel® provides two models, the SR300 (previously named F200) for short range applications, and the R200 for

long range acquisitions. Both sensors, similarly to Kinect v2, also provide IR images. Both models are based on the same technology, consisting in 3 streams that provide RGB images, stereoscopic IR and its resulting depth-map representations of 3D shape. The prospect of designing a dataset that comprises all these modalities in conjunction with simulated real-world acquisition environments would certainly result in a strong contribution to the field. With this prospect in mind, the next section will serve as a detailed presentation of our proposed RealFace multimodal face dataset, with regards to both the acquisition setup as well as its final composition.

3 RealFace Dataset Description

The RealFace dataset was acquired from a set of 42 volunteers, with different ethnicities, ages and genders. [1] Ages ranged from 18 to 40 years, gender distribution was 22 male and 20 female, while regarding nationality 41 were Portuguese and 1 was Venezuelan. After signing an agreement for the sole use of the images for scientific research purposes, each of these individuals carried out the acquisition protocol detailed below.

The acquisition protocol followed in the present work was designed so that the environmental conditions presented to the sensor would closely simulate a realistic set of real-world unconstrained conditions. With this in mind variations in pose (frontal, left/right profile and left/right $\pm 45°$), facial expression (neutral and open mouth), occlusions (handkerchief and glasses) and illumination (natural, artificial and darkness) were considered in the acquisition setup. All combinations of occlusion (2) and facial expression (2) were replicated for every illumination (3) and pose (5, plus an extra neutral) conditions, and acquisition was made in a sequential way so as not to render the whole process too long and tedious for the volunteers. Due to the different optimal operating ranges, the whole process was repeated for each sensor, with the distance to the sensor being varied from 0.5 m, for the SR300 model, to 1.3 m, for the R200. The full acquisition setup is depicted in Fig. 1 and the whole acquisition process took approximately 12 minutes per subject, resulting in a total of 72 conditions. Some representative examples are depicted in Fig. 2.

To take advantage of all data streams made available by the Intel® RealSense™ sensors, for each of the aforementioned conditions, we acquired an RGB image and its respective Point Cloud, as well as the IR images provided by the integrated sensors (two for the R200 and one for the SR300) and the corresponding depth maps. This wide array of modalities and conditions confers our dataset a high versatility regarding its possible uses within the biometrics research community. A representative example of the data obtained with the SR300 model for a single acquisition is depicted in Fig. 3.

Additionally to the multiple data sources, for each RGB/Point Cloud/IR group, a set of facial keypoints were also manually annotated, to both facilitate

[1] All volunteers were gathered from the students and staff community of the Faculty of Engineering of the University of Porto, Portugal).

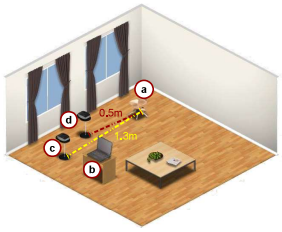

Fig. 1: RealFace dataset acquisition setup: (a) subject; (b) acquisition control software; (c) SR300 model and (d) R200 model.

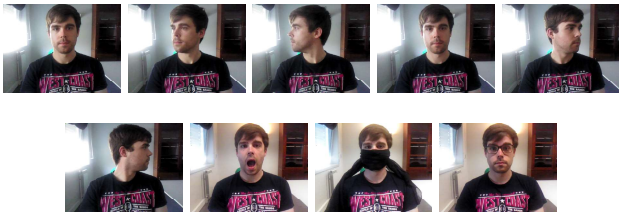

Fig. 2: Representative examples of the poses, occlusion, illumination and expression variations considered for each subject during the RealFace acquisition process.

the region-of-interest (ROI) segmentation, as well as allowing the dataset to be used as a benchmark for keypoint detection in any of the presented modalities. The amount and nature of the annotated keypoints depended on the variations that characterized each image. In frontal images, both eye centers, the nose tip and both mouth corners were annotated, except for the handkerchief occlusion scenario, where both nose tip and mouth corners were not considered. In profile and rotated pictures, the closest visible eye center, the nose tip and the closest visible mouth corner were considered, as well as the visible ear lobe. The handkerchief occlusion limitations were also verified here, with no nose tip and mouth corners being considered. When hair occlusion resulted in no visible ear lobe this point was also left out from the annotation. A visual example of the manual annotation in each of these scenarios is depicted in Fig. 4.

Fig. 3: Multimodal data from a single acquisition in the RealFace dataset: (a) RGB; (b) Depth-map; (c) Point-Cloud and (d) Infrared.

Fig. 4: Manual annotation on the RealFace dataset: (a) Frontal neutral; (b) Frontal occlusion; (c) Profile neutral and (d) Profile occlusion.

The whole RealFace dataset will be made publicly available for research purposes. For more information regarding its availability contact the mailing author.

4 Experimental setup for the RealSense dataset

In this section we set an experimental setup for performance assessment in the RealFace dataset. The baseline results obtained following the setup will be presented afterwards in this section. We will cover data partitioning and region-of-interest segmentation, as well as the shallow and deep feature representations chosen for the baseline performance assessment. We start this analysis by presenting some global considerations regarding some cases we chose to leave out of the present work, but could be the focus of future endeavors on this dataset.

Global considerations: We chose to work solely on frontal poses, leaving both the $\pm 45^{\circ}$ and the profile images out of our baseline analysis. We felt that including this kind of images would dilute the focus of the paper. The R200 model was also left out of this work due to the fact that we confirmed what it had been previously reported in literature: the quality of the depth images acquired with this model is still very low and unfit for object recognition problems [13]. Nevertheless, we propose a region-of-interest segmentation strategy for non-frontal images, and the whole setup is easily extrapolated for future work with the R200 images.

Pre-processing: The manually annotated keypoints were used to crop a region-of-interest (ROI) around the nose. In frontal images, a square ROI was considered, with the center corresponding to the nose tip and the side set to twice the distance between the eye centers. For $\pm 45^{\circ}$ variations, on the other hand, we considered a square box centered on the horizontal line passing through the nose tip, and with side corresponding from $1.5\times$ the distance between the nose tip and the ear lobe. For the profile images a similar strategy was followed, but the side corresponded to $1\times$ the aforementioned distance.

Data partitioning: We chose images characterized by both neutral expression and artificial illumination to serve as training data for each individual. This decision is based on the fact that we want the most controlled scenarios to be used during training, and the most complicated ones be left for testing. It is intuitive to understand that training an algorithm to encompass all possible acquisition scenarios is unfeasible, when real-world applications are considered.

Thus, by using the more stable images during training we aim to assess the capability of algorithms of presenting robust behaviour when more complicated challenges are presented to them. All other combinations of conditions were assessed individually during testing: illumination (natural - N; artificial - A; darkness - D), occlusion (S - scarf; G - glasses) and expression (N - neutral; OpM - open mouth).

Shallow features: From an extensive list of state-of-the-art feature extraction methods for face recognition using RGB images, the top performances were observed for PHOW, TPLBP and FHOG. A similar analysis was carried out for depth images, with both PHOW and FHOG presenting consistently better performances, while 3D-LBP finished the top-ranked descriptors. TPLBP (Three-Patch LBP) and 3DLBP are variants of traditional LBP. Presented in [14], 3DLBP was proposed as a variation of traditional LBP, for depth images, where depth differences are encoded in the final descriptor. In [15] TPBLP was proposed as an upgrade of traditional LBP descriptor for face identification. Here, three patches are considered to produce a single bit value for each pixel. The Felzenszwalb's HOG (FHOG) descriptor has been described in [16] as a variant of traditional HOG for object detection, where a feature pyramid is calculated for a finite number of scales, using repeated smoothing and sub-sampling. PHOW, presented in [17] consists in a variation of dense-SIFT which is applied at multiple scales and combined with VLAD (Vector of Linearly Aggregated Descriptors) enconding.

Deep features: Using the pre-trained model provided by [18], we tested the VGG-Face CNN for all modalities, by extracting features from the $fc7$ layer and using them to train a logistic regression classifier, as described in the following paragraph. For depth images, we also decided to train a new CNN from scratch. To serve as training we used data from 195 subjects obtained from all the available datasets presented in Section 2 (except EURECOM and RealFace, which were left for performance assessment). To avoid the class unbalance caused by the high degree of heterogeneity in the original number of samples per individual, we chose to generate synthetic depth maps, by flipping and rotating the original point clouds, until a total of 1000 samples per subject were obtained. The tested architecture consisted in 5 conv-relu-conv-relu-pool blocks followed by 2 fully connected layers. All conv layers include 3×3 filters, whereas the number of filters for each block is $8 - 16 - 32 - 64$, respectively. Finally the two fully connected layers consist of 256 units each. A batch size of 256 and a logarithmically decaying learning rate from 10^{-1} to 10^{-6} were considered, for a total of $50k$ iterations.

Classification: A set of logistic regression models was trained for classification, using each of the aforementioned shallow and deep feature descriptors. The model choice was motivated by its simplicity, leaving a considerable margin for improvement for future works on the dataset, as well as the good performance that it revealed when compared to other alternatives, such as SVM and k-nearest neighbors. The fact that class probabilities can be easily obtained was also a ruling factor of this choice, as it facilitated the multimodal fusion process described

in the next section. Decision for a single feature representation and modality is carried out by maximum probability, with regards to all possible IDs.

Multimodal fusion: As referred earlier, the joint use of multiple data sources to solve the biometric recognition problem has shown improved performance in a multiplicity of recent works. To evaluate such effect in the RealFace dataset, we combine the individual logistic regression probabilities for a given ID from the RGB, depth and IR representations of a single test sample, $p_{mod}(ID|x_i)$, using a weighted-sum rule, $p(ID|x_i) = w_{RGB} \cdot p_{RGB}(ID|x_i) + w_d \cdot p_d(ID|x_i) + w_{IR} \cdot p_{IR}(ID|x_i)$, with $\sum_{mod} w_{mod} = 1$ and w_{mod} optimized by grid search. Decision is then carried out by maximizing the fusion probability $p(ID|x_i)$ with regards to all possible IDs. To overcome the loss of performance in the case of RGB in darkness conditions, a new method is proposed to deal with severely low illumination conditions. For all test images, the individual mean intensity of gray-scale converted RGB image, μ_i, is calculated and, depending on this value, a corrected weight for RGB-modality, w^*_{RGB}, is calculated, using a logistic function, $w^*_{RGB} = \frac{1}{1+e^{(-0.5(-\theta+\mu_i)}} \cdot w_{RGB}$, where θ was empirically set to 20 as it was observed to be the mean transition intensity between fair and poor illumination conditions. This adaptation allows the algorithm to self-adapt its performance, by adjusting the RGB weight to be higher in better illumination, and lower in less ideal low illumination conditions. The weight loss $w^*_{RGB} - w_{RGB}$ is then divided equally between the other modalities.

Following the experimental setup described throughout this section the baseline performance for the RealFace dataset will be presented next. Furthermore, some experiments were also carried out on the RGB-D EURECOM dataset, so as to better understand the challenges of our proposed dataset when compared with a state-of-the-art alternative.

5 Results and Discussion

In this section we start by giving some insight into the EURECOM dataset and the experimental setup used for performance assessment in this alternative. We then proceed with the discussion of the results obtained for each tested dataset, with regards to the specific challenges that each one poses.

5.1 EURECOM Dataset

Composition: The EURECOM dataset, acquired with the Microsfot Kinect v1 sensor, is composed by a set of well-aligned 2D, 2.5D, 3D and video data. It includes scans from 52 subjects (38 males and 14 females) from two sessions interleaved from 5 to 14 days. Each session has nine types of scans that include: neutral face (N), open mouth (OpM), smile (S), strong illumination (LO), occlusion with sunglasses (OE), occlusion by hand (OM), occlusion by paper (OP), right face profile and left face profile. The acquisition environment is controlled in terms of luminosity, with the individuals always in a range from 0.7 to 0.9 meters to the sensor. A blank background was chosen to make the processing of

the data easier. An example of the 2D and 2.5D images from a single individual is presented in Fig. 5.

Fig. 5: Example RGB (a)-(g) and depth (h)-(o) images from the multiple subsets of the EURECOM dataset.

Experimental setup: We chose to follow a setup similar to the one we proposed for the RealFace dataset. The neutral images from both sessions were, therefore, chosen for training and all other subsets were considered individually for testing. ROI segmentation was carried out using the keypoints provided by the dataset, using a methodology analogous to the one described in Section 4 for both RGB and depth images. Resizing was carried out to 96×96 and 224×224, for depth and RGB data respectively. These dimensions were chosen to correspond to the inputs expected by the CNNs used in this work.

5.2 Performance analysis

The results for the EURECOM dataset are summarized in Table 1. In RGB images, shallow and deep features presented similar high performance for all tested conditions. PHOW with VLAD encoding was the shallow descriptor with better overall performance, and showed great versatility in image description by achieving the highest overall performance also for depth images. For such images, the overall performance drop comparatively to their RGB counterparts is clear. In this case, shallow features outperform deep features by a considerable margin. This drop in performance can be understood by the fact that only RGB images were considered during the training of the VGG-Face CNN. As there is no trivial visual similarity between the two types of images, it is logic to conclude that the filters learnt for the RGB problem are not directly applicable to depth inputs. This observation is also corroborated by the fair results presented by the proposed pre-trained CNN (PT_{CNN}) evaluated in the non-occlusion cases. Clearly, the learnt filters are able to achieve some discrimination, unlike the VGG-face alternative, but fail to adapt to non-ideal cases related to occlusions. When both modalities are combined, as referred in the state-of-the-art, the global performance is slightly increased, although not statistically relevant due to the already very high performances obtained by the RGB modality alone.

Table 2 presents the main results obtained for the RealFace dataset. As expected from being a more challenging dataset than EURECOM, the overall performance drop is evident. In RGB images, deep features clearly present a

Table 1: Performance comparison of shallow (S) and deep (D) features on the RGB and depth modalities of the EURECOM dataset.

		RGB							Depth						
	Feat/SS	LO	OE	OM	OP	OpM	Sm	G	LO	OE	OM	OP	OpM	Sm	G
S	FHOG	100	96.2	84.6	63.5	88.5	98.1	88.5	89.4	89.4	23.1	6.7	87.5	100	66.0
	PHOW	99.0	96.2	100	97.1	100	100	98.7	98.1	92.3	52.9	26.0	89.4	99.1	76.3
	LBP	100	95.2	95.2	88.5	95.2	98.1	95.7	–	–	–	–	–	–	–
	3D-LBP	–	–	–	–	–	–	–	90.4	93.3	12.5	4.8	89.4	99.1	64.9
D	VGG-F_O	100	98.1	95.2	96.2	100	100	98.2	34.6	12.5	18.3	13.5	16.4	38.46	19.1
	PT_{CNN}	–	–	–	–	–	–	–	85.6	82.7	6.7	2.9	49.0	84.6	51.9
MM	VGG+PHOW	100	98.1	96.2	96.2	100	100	98.4							
	PHOW+PHOW	100	99.1	100	97.1	100	100	99.4							

more robust behaviour, when presented to more variable illumination and occlusion conditions. The PHOW shallow descriptor, however, keeps the highest performance for depth images, proving to be an interesting alternative for object description in this type of data. The same observations regarding the VGG-Face and our proposed CNN for depth images can be made for this dataset, with the occlusion scenarios severely compromising global performance. In the IR modality some interesting observations can also be made. First, both PHOW and the deep descriptors from VGG-Face achieve the best overall performances. While the obtained performance is still significantly lower than the observed for RGB images, it is interesting to note how the filters learnt by VGG-Face still carry some of the discriminative power to this new modality. As referred above, for depth images, the visual similarity between RGB and IR images might translate into similar responses to the pre-trained filters, thus justifying the similar observed behaviour. The improvement caused by multimodal fusion in this dataset is more clearly noted than in EURECOM. It should be considered that no darkness conditions were evaluated for the RGB modality alone and, therefore, direct comparison of multimodal performance can only be carried out with the remaining modalities.

6 Conclusion and Future Work

The growing interest in 3D information for face recognition, as well as the emergence of new low-cost sensors, such as the Intel RealSense, has motivated the creation of the RealFace dataset, a multimodal set of images acquired under a wide array of non-ideal conditions to be used for performance assessment in a multiplicity of applications. Even though we only assessed its use in biometric recognition, we acknowledge that its usability can extend to fields such as face alignment, gender and age prediction as well as face detection in depth and IR images. The manually annotated keypoints, as well as the defined ROI segmentation methodologies, make the performed experiments easily replicable and confer the presented performance baseline a strong starting point for further research in the community. However, the number of enrolled subjects is still not as high as desirable, and an extended version of the dataset would be an interesting line of research in the future. The inclusion of more intermediate non-frontal poses

Table 2: Performance comparison of shallow (S) and deep (D) features on the RGB, depth and IR modalities of the RealFace dataset.

		NN	NOpM	NS	NG	AOpM	AS	AG	DN	DOpM	DS	DG	G
	RGB												
	FHOG	54.8	39.3	26.2	66.7	96.4	42.9	90.5	–	–	–	–	59.5
	PHOW	42.9	33.3	14.3	42.9	100	54.8	97.6	–	–	–	–	55.1
	LBP	64.3	42.9	27.4	65.5	96.4	56.0	95.2	–	–	–	–	64.0
	VGG-F_o	98.8	96.4	81.0	89.3	100	88.1	95.2	–	–	–	–	92.7
	Depth												
	Feat/SS	NN	NOpM	NS	NG	AOpM	AS	AG	DN	DOpM	DS	DG	G
S	FHOG	71.4	46.4	28.6	79.8	78.6	33.3	76.2	90.5	66.7	32.1	70.2	61.3
S	PHOW	76.2	66.7	46.4	66.7	88.1	48.8	65.5	88.1	77.4	36.9	54.8	65.0
S	3D-LBP	54.8	47.6	26.2	66.7	76.2	21.4	61.9	83.3	60.7	16.7	57.1	52.1
D	VGG-F_o	17.9	13.1	6.0	15.5	11.9	9.5	14.3	19.1	9.5	9.5	15.5	12.9
D	PT_{CNN}	48.8	44.1	15.5	66.7	72.6	16.7	56.0	58.3	51.2	13.1	53.6	45.1
	IR												
	Feat/SS	NN	NOpM	NS	NG	AOpM	AS	AG	DN	DOpM	DS	DG	G
	FHOG	88.1	75.0	45.2	89.3	94.1	53.6	89.3	95.2	79.8	44.1	82.1	76.0
	PHOW	97.6	94.1	60.7	90.5	100	66.7	96.4	98.8	91.7	48.8	88.1	84.9
	LBP	84.5	76.2	42.9	85.7	94.1	45.4	86.9	92.9	77.4	33.3	81.0	72.8
	VGG-F_o	100	96.4	76.2	78.6	96.4	75.0	73.8	98.8	95.2	61.9	71.4	84.0
	Multimodal												
	Feat/SS	NN	NOpM	NS	NG	AOpM	AS	AG	DN	DOpM	DS	DG	G
	VGG+PHOW+VGG	100	100	89.3	91.7	98.8	90.5	95.2	98.8	95.2	63.1	76.2	90.8
	VGG+PHOW+PHOW	98.8	96.4	81.0	89.3	100	88.1	95.2	100	90.5	58.3	90.5	89.8

would further extend the usability of the dataset for alternative applications such as pose quantification.

Regarding the comparative analysis between deep and shallow features we observed that a few challenges are still unsolved. While the publicly available VGG-face network showed excellent performance in the RGB modality for all tested scenarios in both datasets, surpassing all alternative shallow feature alternatives, performance dropped considerably for depth images. The pre-trained CNN that we presented showed increased performance in some scenarios, but still stays below the results obtained with specific tailored features such as PHOW and FHOG. The fact that the amount of data used to train VGG-Face is considerably higher than the amount of depth data used to train our CNN may account for these observations. With the appearance of more datasets based on depth representations of faces, and the consequent growth in the amount of available data, an improved version of the proposed CNN could also be easily obtained, namely by augmenting the training dataset to better deal with the presence of occlusions.

Acknowledgments

This work was funded by the Project NanoSTIMA: MacrotoNano Human Sensing: Towards Integrated Multimodal Health Monitoring and Analytics/NORTE-01-0145-FEDER-000016 financed by the North Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement, and through the European Regional Development Fund (ERDF), and also by Fundação para a Ciência e Tecnologia (FCT) within PhD grant number SFRH/BD/87392/2012.

References

1. Nech, A., Kemelmacher-Shlizerman, I.: Megaface 2: 672,000 identities for face recognition. (2016)
2. Monteiro, J.C., Cardoso, J.S.: A cognitively-motivated framework for partial face recognition in unconstrained scenarios. Sensors **15** (2015) 1903–1924
3. Monteiro, J.C., Freitas, T., Cardoso, J.S.: Multimodal hierarchical face recognition using information from 2.5 d images. U. Porto Journal of Engineering **2** (2016) 39–54
4. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: Deepface: Closing the gap to human-level performance in face verification. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2014)
5. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Advances in neural information processing systems. (2012) 1097–1105
6. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 815–823
7. Chen, D., Cao, X., Wen, F., Sun, J.: Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2013)
8. Simonyan, K., Parkhi, O.M., Vedaldi, A., Zisserman, A.: Fisher Vector Faces in the Wild. In: British Machine Vision Conference. (2013)
9. Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition. In: British Machine Vision Conference. (2015)
10. Crosswhite, N., Byrne, J., Parkhi, O.M., Stauffer, C., Cao, Q., Zisserman, A.: Template adaptation for face verification and identification. CoRR **abs/1603.03958** (2016)
11. El Khiyari, H., Wechsler, H., et al.: Face recognition across time lapse using convolutional neural networks. Journal of Information Security **7** (2016) 141
12. Min, R., Kose, N., Dugelay, J.L.: Kinectfacedb: A kinect database for face recognition. IEEE Transactions on Systems, Man, and Cybernetics: Systems **44** (2014) 1534–1548
13. Song, S., Lichtenberg, S.P., Xiao, J.: Sun rgb-d: A rgb-d scene understanding benchmark suite. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 567–576
14. Huang, Y., Wang, Y., Tan, T.: Combining statistics of geometrical and correlative features for 3d face recognition. In: BMVC, Citeseer (2006) 879–888
15. Wolf, L., Hassner, T., Taigman, Y.: Descriptor based methods in the wild. In: Workshop on Faces in'Real-Life'Images: Detection, Alignment, and Recognition. (2008)
16. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE transactions on pattern analysis and machine intelligence **32** (2010) 1627–1645
17. Bosch, A., Zisserman, A., Munoz, X.: Image classification using random forests and ferns. In: Proceedings of the International Conference on Computer Vision (ICCV). (2007)
18. Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition. Proceedings of the British Machine Vision **1** (2015) 6