Face Verification from Depth using Privileged Information

Guido Borghi
guido.borghi@unimore.it
Stefano Pini
stefano.pini@unimore.it
Filippo Grazioli
filippo.grazioli@unimore.it
Roberto Vezzani
roberto.vezzani@unimore.it
Rita Cucchiara
rita.cucchiara@unimore.it

Department of Engineering
"Enzo Ferrari"
University of Modena and Reggio Emilia
Modena, Italy

Abstract

In this paper, a deep Siamese architecture for depth-based face verification is presented. The proposed approach efficiently verifies if two face images belong to the same person while handling a great variety of head poses and occlusions. The architecture, namely *JanusNet*, consists in a combination of a depth, a RGB and a hybrid Siamese network. During the training phase, the hybrid network learns to extract complementary mid-level convolutional features which mimic the features of the RGB network, simultaneously leveraging on the light invariance of depth images. At testing time, the model, relying only on depth data, achieves state-of-art results and real time performance, despite the lack of deep-oriented depth-based datasets.

1 Introduction

The computer vision community has broadly addressed the face recognition problem in both the RGB and the depth domain. Traditionally, this problem is categorized in two tasks: *face identification* and *face verification*. The former consists in the comparison of an unknown subject's face with a set of faces (*one-to-many*), while the latter consists in comparing two faces in order to determine whether they belong to the same person or not (*one-to-one*).
The majority of existing face recognition algorithms is based on the processing of RGB images, while only a minority of methods investigates the use of other image types, like *depth maps* or *thermal images* [27, 28]. Recent works [29, 34, 35] employ very deep convolutional networks for the embedding of face images in a d-dimensional hyperspace. Unfortunately, these very deep architectures used for face recognition tasks typically rely upon very large scale datasets which only contain RGB or intensity images, such as *Labeled Faces in the Wild* (LFW) [13], *YouTube Faces Database* (YTF) [43] and *MS-Celeb-1M* [11].

The main goal of this work is to present a framework, namely *JanusNet* and depicted in Figure 1, that tackles the face verification task analysing depth images only. In particular,

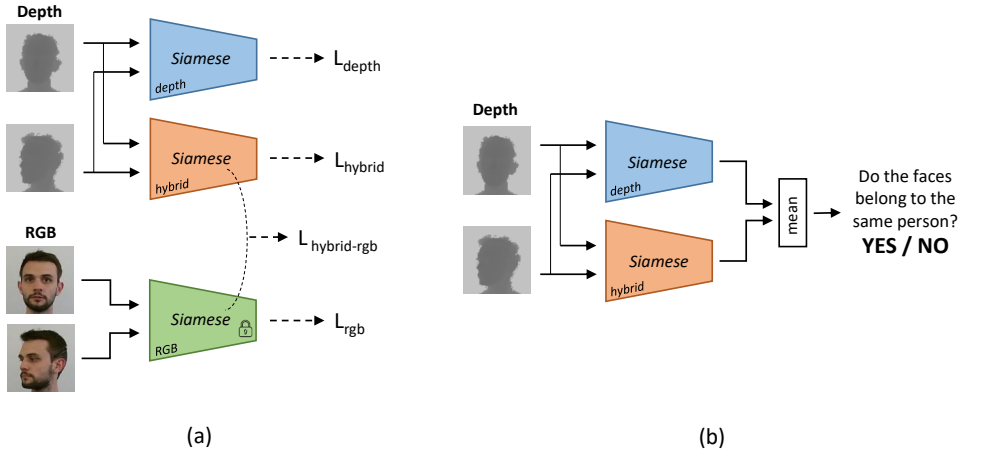

Figure 1: Overview of the *JanusNet* architecture. During the training phase (a), the model is composed of the depth and the hybrid network, which analyse depth face images, and the RGB network, which analyses the RGB ones. Each Siamese network predicts the similarity between an input pair of images. At testing time (b), only the depth and the hybrid networks are employed for the face verification task.

the use of shallow deep architectures is investigated in order to obtain real time performance and to deal with the small scale of the existing depth-based face datasets, like [4, 24]. In fact, despite the recent introduction of deep-learning oriented depth-based datasets and cheap commercial depth sensors, the usual size of depth datasets is not big enough to train very deep neural models [3, 14, 22].
Furthermore, we aim to directly detect the identity of a person without strong a priori hypotheses, like facial landmark or nose tip localisation, which could compromise the whole following pipeline. Under the hypothesis that intensity information improves the face verification task, RGB side information is incorporated during the training phase.

In this paper, a combination of Siamese models, composed by a depth, a hybrid and a RGB network, is proposed, taking partial inspiration from [12]. The Siamese networks, exploiting the architecture depicted in Figure 2, are meant to predict whether two images belong to the same person or not. During the training phase, the hybrid Siamese network is conditioned by a specific loss that forces its feature maps to mimic the mid-level features maps of the RGB network. At testing time, the RGB network is not employed, while the depth and the hybrid Siamese network are fed with the same pair of depth images and jointly predict if they belong to the same person.

2 Related Works

Face Verification. Traditional RGB-based face verification approaches tend to be sensitive to variations in illumination, pose and expression changes [6, 26]. In the last decades, a vast body of literature has exploited algorithms based on the classification of hand-crafted features [1, 2, 16, 17, 18, 40, 46]. Recently, very deep neural networks have achieved best performance on RGB images [38, 39]. Taigman *et al.* [39] presented *DeepFace*, a deep

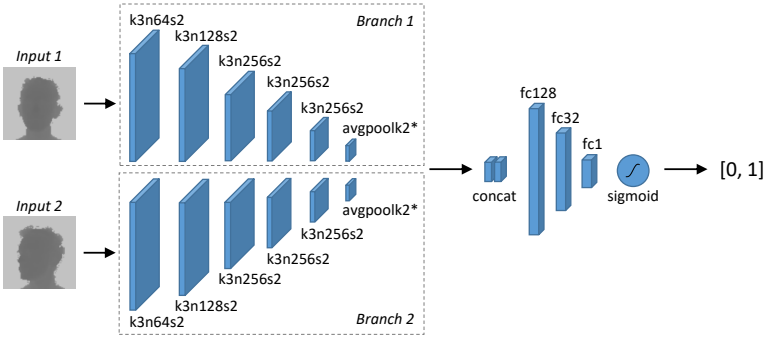

Figure 2: Architecture of each Siamese module, composed of two branches. The loss functions $L_{hybrid\text{-}rgb_{1,2}}$ are computed between each branch of the RGB and the hybrid Siamese networks, on the last convolutional layers (marked with *). k, n, s, fc correspond to kernel size, number of feature maps, stride and units. The network outputs a continuous similarity score.

convolutional network designed for the verification task. In particular, a Siamese architecture is exploited in conjunction with pre-processing steps, such as face alignment and face frontalization. In [34], a triplet loss and a face embedding space were proposed. A deep convolutional network, called *FaceNet*, is trained to directly optimise a common embedding space, achieving state-of-art results at the time of publication. Kumar *et al*. [20] proposed a SVM classifier that learns the similarity of faces by recognizing the presence of visual attributes, *e.g*. gender, age and ethnicity.
Several other works tackled the face verification task exploiting Siamese architectures and RGB images in order to learn a similarity metric directly from the data [5, 7, 10, 19]. Usually, these methods require a huge amount of training images and a threshold value on the learned similarity metric to define identities.

The recent introduction of cheap and accurate depth sensors, like *Microsoft Kinect* or *Intel RealSense* series, has increased the relevance on methods based on 2.5D (*depth maps* or *depth images*) and 3D (*point clouds*) data, despite the lack of big-scale annotated datasets [47]. Depth devices can be divided in respect to the technology used to retrieve depth information, *i.e*. *Structured Light*, *Time-of-Flight* (ToF) or *Stereo Cameras*. Each technology has its pros and cons [33]. Generally, the acquired depth data presents light invariance, since it is based on infrared light. Point cloud-based methods usually reconstruct 3D face models exploiting an ICP-like registration algorithm [6]. This process is extremely sensitive to the quality of the 3D input data and it often requires an expensive range camera. 2.5D images are acquired with cheap and high-quality sensors that often provide RGB data as well.

Mantecon *et al*. [24] used a *Pegasos SVM* [35] with a substantial modification of the popular Local Binary Patterns (LBP) algorithm, namely *DLQP* feature, to solve the *one-vs-all* face identification task. Moreover, a new dataset, called *High-Resolution Range-based Face Database* (HRRFaceD), was proposed. Then, the method was improved in [25] through a novel and highly discriminative face image descriptor, referred as *Bag-D3P*. Both methods assume the knowledge of all subjects during the training phase.
Furthermore, several methods [21, 22, 23, 32] combine the use of both RGB and depth data, assuming the presence of both type of data at training and testing time, to compensate the

relativity low resolution of depth maps. However, depth data are frequently exploited only in a complementary way with respect to RGB images or point clouds [8, 9, 31]. In addition, most of these methods are based on facial landmark detection to perform face alignment and frontalisation and on a supplementary classifier to perform a joint classification of multi-modal features.

Privileged Information. Also called *Side Information*, *Privileged Information* was introduced in [41]. The main idea is to add knowledge at training time in order to improve the performance of the system at testing time. Several works [36, 42, 44] successfully proved the strength of this approach when multi-modal data are available.
In [45], authors proposed to use depth images in order to improve the learning of a distance metric for the face verification and person re-identification task on RGB images. Recently, this concept was exploited by Hoffman *et al*. in [17]. They trained a convolutional architecture for RGB-based object recognition incorporating depth information during the training phase. The hallucination branch, trained on RGB images, is able to mimic mid-level features of the depth branch, improving the final accuracy score.
Differently from this work, we investigate the use of raw depth images as input in conjuction with Siamese architectures. Moreover, a simpler training procedure is adopted in order to reduce the number of loss functions and to address the different final task, *i.e*. the face verification.

3 JanusNet

An overview of the proposed framework, called *JanusNet*, is depicted in Figure 1.
The goal of the architecture is the face verification task, *i.e*. estimating whether two face images belong to the same person or not. The model is composed of three Siamese modules, called depth, hybrid and RGB Siamese network, that share the same architecture, detailed in Section 3.1. Depth image pairs are the input data of both the depth and the hybrid Siamese network, while RGB images are the input of the RGB Siamese module. During the training phase, a *binary cross-entropy* loss function is applied separately to the output of each module, corresponding to the terms L_{depth}, L_{hybrid} and L_{rgb} reported in Figure 1.
At the same time, two loss functions ($L_{hybrid\text{-}rgb_1}$ and $L_{hybrid\text{-}rgb_2}$), defined as follows, are applied between the last convolutional layer of each Siamese branch of the RGB and the hybrid module:

$$L_{hybrid\text{-}rgb_{1,2}} = \frac{1}{N}\sum_{n}^{N}\left(\mathbf{y}_n^{hybrid} - \mathbf{y}_n^{rgb}\right)^2 \tag{1}$$

where N is the number of feature maps of the last convolutional layer of the Siamese modules, while $\mathbf{y}_n^{hybrid}$ and $\mathbf{y}_n^{rgb}$ are the n-th feature maps of the hybrid and the RGB network, respectively.
Since the input of these modules is composed of corresponding depth and RGB images, the hybrid module is forced to learn visual features that are characteristic of RGB images from depth maps. Specifically, the loss functions $L_{hybrid\text{-}rgb_{1,2}}$ force the mid-level feature maps of the hybrid network to mimic those of the RGB network.
Experimental results, reported in Section 4.3, confirm that these features are complementary to the features extracted from the depth Siamese module, improving the overall performance

(a) (b)

Figure 3: Depth and RGB sample frames taken from *Pandora* dataset. Part (a) contains frames taken from sequences S_1, S_2, S_3 while part (b) from sequences S_4, S_5, in which garments are introduced. Subset details are reported in Section 4.1.

of the proposed method. The final loss function for the proposed models is:

$$L = \alpha \left(L_{hybrid\text{-}rgb_1} + L_{hybrid\text{-}rgb_2} \right) + \beta \left(L_{depth} + L_{hybrid} + L_{rgb} \right) \tag{2}$$

where α and β are the loss weights. With regard to the weight initialisation, the depth and the RGB network are randomly initialised and pre-trained on depth and RGB image pairs, respectively, while the hybrid network is initialised with the pre-trained weights of the RGB module. The weights of the RGB network are not updated during the training of the hybrid network in order to avoid that RGB mid-level features mimic the depth ones.
During the testing phase, the *JanusNet* architecture relies upon two modules: only depth image pairs are processed by both the depth and the hybrid network (Fig. 1). In this phase, the RGB module does not play a role and is therefore discarded.

3.1 Siamese Architecture

A shallow architecture has been adopted for the Siamese modules in order to deal with the relatively small size of existent depth datasets and to achieve real time performance. The input of each Siamese module is a pair of two images of size 100×100. Each Siamese branch presents 5 convolutional layers with kernel size 3×3, stride 2×2 and an increasing number of feature maps.
In details, the convolutional layers have 64 and 128 kernels for the first two layers and 256 for the following ones. Convolutional layers are followed by an average pooling layer with kernel size 2×2, whose output is a tensor of shape $256 \times 1 \times 1$.
The output tensors of the two branches are flattened and concatenated, obtaining a 512-*d* feature vector, and followed by 3 fully connected layers with 128, 32 and 1 units, respectively. The output of the last layer of the Siamese network is the predicted similarity between the input depth faces, expressed through a continuous value in the $[0,1]$ range. *Rectified Linear Unit* (ReLU) activation function is exploited in every layer, except the last one that applies a *Sigmoid* activation function. *Stochastic Gradient Descent* (SGD) is used for the training in conjunction with *dropout* [37] and *batch normalization* [15] which are employed for regularisation purposes. An overview of the Siamese module is presented in Figure 2.

Model	Data type	Accuracy
FaceNet [34]	RGB	0.8232
Hybrid network	Depth	0.7553
RGB network	RGB	0.7631
Depth network	Depth	0.7950
JanusNet	**Depth**	**0.8142**

Table 1: Accuracy comparison for the face verification task on the fixed test set of the *Pandora* dataset. Results are reported on depth and RGB data for the Siamese modules, on depth data for the proposed architecture, namely *JanusNet*, and on RGB data for the *FaceNet*-like architecture.

3.2 Head Crop

The head localisation task, which is out of the scope of this paper, is performed exploiting depth information and Pandora dataset annotations. In particular, we accurately crop the head including the relevant part of the foreground. Given the head center coordinates (x_H, y_H) in the frame, a dynamic size cropping provides the head bounding box that includes a small portion of the background. The head bounding box has the barycentre located in (x_H, y_H). Width w_H and height h_H are computed as follows:

$$w_H = \frac{f_x \cdot R_x}{D} \qquad h_H = \frac{f_y \cdot R_y}{D} \tag{3}$$

where D is the distance between the acquisition device and the head centre, $f_{x,y}$ are the horizontal and vertical focal lengths (expressed in pixel) and $R_{x,y}$ represent the average width and height of a generic face. In our experiments, $f_{x,y} = 365.337$ and $R_{x,y} = 320$. Final head crops are then resized to 100×100 pixels. Some example of the extracted head bounding boxes are reported in Figure 3.

4 Experiments

We analyse the performance of the proposed architecture using two public datasets, both containing the RGB and the corresponding depth face images. Firstly, *JanusNet* and its single modules are trained and tested on the Pandora dataset. Then, the whole architecture is trained, tested and compared with depth based-only state-of-art methods [24, 25].
In addition, we conduct an ablation study on the whole system, to understand the contribution of each siamese network that forms the framework – *i.e.* the depth, the hybrid and the RGB network (see Figure 1) – to the final result. Furthermore, we investigate how head poses and occlusions influence the overall performance.
For full reproducibility, we release the source code of the proposed method along with the list of the testing couples[1] used to collect experimental results.

4.1 Datasets

Pandora. Borghi *et al.* [4] introduced the *Pandora* dataset, which was specifically created for the head and shoulder pose estimation tasks. The acquisition device (*i.e. Microsoft Kinect*

[1] http://imagelab.ing.unimore.it/janusnet

	Pegasos SVM				JanusNet		
	LBP	SIFT	DLQP	Bag-D3P	*max*	*avg*	*voting*
Accuracy	0.5917	0.7194	0.7347	0.9430	0.9756	**0.9877**	0.9804
Improvement	-	+12.7	+14.3	+35.1	+38.4	**+39.6**	+38.9

Table 2: Accuracy comparison for the face recognition task on the *HRRFaceD* dataset. Comparison results are taken from [24, 25], in which different features are tested. See Equation 5a, 5b and 5c for details about *max*, *avg* and *voting*.

for Windows v2, also called *Microsoft Kinect One*) acquires only the upper-body part of the subjects (10 males and 12 females). Even though it was not designed for the face recognition task, this dataset presents remarkable challenges due to the significant head pose variance and numerous facial occlusions, both with objects (*e.g.* smartphones, tablets) and garments. In particular, head pose angles range within the following intervals: *roll*: $\pm 70°$, *pitch*: $\pm 100°$, *yaw*: $\pm 125°$. As reported in the original work, we create a subject-independent test set using subjects number 10, 14, 16 and 20. Furthermore, we create an additional subject-independent validation set composed by subjects 9, 18, 21 and 22.

Each recorded subject presents 5 different sequences of frames. We split the sequences into two sets. In the first set (here referred as sequences S_1, S_2, S_3), actions are performed with constrained movements, *i.e.* yaw, pitch and roll angles vary one at a time, both for head and shoulders. The second set (sequences S_4, S_5) contains both simple and complex movements, as well as challenging camouflage and occlusions, which even seriously affects the face appearance in both the RGB and depth images. Experiments are performed on both these two sets in order to investigate the effects of the mentioned differences.

Moreover, we split the dataset taking head pose angles into account. Given the angles yaw, pitch and roll as ρ, θ and σ, respectively, for each sample $s_{\rho\theta\sigma}$, we create three head pose-based subsets, defined as follows:

$$A_1 = \left\{ s_{\rho\theta\sigma} \,|\, \forall \gamma \in \{\rho, \theta, \sigma\} : -10° \leq \gamma \leq 10° \right\} \tag{4a}$$

$$A_2 = \left\{ s_{\rho\theta\sigma} \,|\, \exists \gamma \in \{\rho, \theta, \sigma\} : \gamma < -10° \vee \gamma > 10° \right\} \tag{4b}$$

$$A_3 = \left\{ s_{\rho\theta\sigma} \,|\, \forall \gamma \in \{\rho, \theta, \sigma\} : \gamma < -10° \vee \gamma > 10° \right\} \tag{4c}$$

A_1 consists of only frontal face images, while non-frontal face images are included in A_2. A_3 contains extreme head poses. As reported above, Siamese modules take image pairs as input. Since a dataset with N images contains a huge number of possible pairs (*i.e.* $\binom{N}{2}$ unique pairs), we created two fixed set of image couples, a validation and a test set, in order to allow repeatable and comparable experiments. Reported results are obtained evaluating the model on the fixed test set in correspondence to the lower loss achieved on the fixed validation set.

HRRFaceDatabase. Mantecon *et al.* [24] introduced the *High-Resolution Range-based Face Database*. It is composed of about 20*k* images of 18 different male and female subjects recorded from different perspectives. Like Pandora, it was collected by using the *Kinect One* device and placing users at a distance of about 50cm from the sensor. All subjects extensively rotated their head during the dataset acquisition. Training and testing sequences are obtained by sampling from the same recording for each subject. We exploit the train-test split reported in the reference work. *HRRFaceD* dataset contains already-cropped face images.

Train / Test	$\{S_1,S_2,S_3\}$	$\{S_4,S_5\}$	$\{S_1,S_2,S_3,S_4,S_5\}$
$\{S_1,S_2,S_3\}$	0.8442	0.7464	0.7734
$\{S_4,S_5\}$	0.7921	0.7127	0.7426
$\{S_1,S_2,S_3,S_4,S_5\}$	0.8049	0.7323	0.7620

Table 3: Accuracy comparison for the face verification task as a function of different data subsets. Results are reported for the proposed Siamese model on depth data only.

4.2 JanusNet

The results of the Siamese modules and the *JanusNet* architecture on *Pandora* dataset are reported in Table 1. It is worth to notice that the Siamese network reaches similar results on both the RGB and depth data, confirming that depth images provide as much discriminant information as RGB images. The models are trained on the A_2 subset (see Sect. 4.4) since it assures better results. Notwithstanding the hybrid network obtains similar performance to the RGB network, the *JanusNet* architecture, corresponding to the fusion of the depth and the hybrid network, further improves the accuracy, confirming that learned features from the depth and the hybrid network are complementary and jointly participate to the final prediction. The network is trained with $\alpha = \beta = 1$, learning rate set to 0.002, momentum to 0.9 and a batch size of 64. Furthermore, we report results of a deep model[2], based on the *FaceNet* architecture and pre-trained on a subset of [1]. Even if it has not been fine-tuned on *Pandora* dataset, we use it as a comparison to verify that *JanusNet* achieves similar performance with respect to the RGB state-of-art method, trained on a big-scale dataset. For a fair comparison, we adopt a slightly different crop procedure to generate appropriate RGB input images for *FaceNet*. Finally, we note that the *Facenet* network has about 140M parameters, while only 3.2M parameters are exploited by *JanusNet* during the testing phase.
From our experiments, the whole framework is able to run at more than 200 fps (4.5 ms per frame) on a machine with a *i7-6850k* (3.60 GHz) processor and a *Nvidia GTX 1080Ti* with extreme low memory hardware requirements (less than 1 GB). The architecture is implemented using the popular framework *PyTorch* [30].

4.3 External Comparison

We compare the proposed framework with state-of-art methods based on depth data only. In particular, we focus on depth images acquired with recent *ToF* sensors. Competitors [24, 25] implement methods for the *Face Identification* task, testing them on *HRRFaceD* dataset. Hence, we adapt *JanusNet*, which is designed for the *Face Verification* task, to tackle the *Face Identification* one. In particular, to deal with the *one-to-many* comparison, *JanusNet* is exploited to obtain a similarity score between all possible face couples contained in *HRRFaceD* dataset. Thus, a variety of functions to determine the final identity can be used. Given s as the testing sample whose identity needs to be determined, $J(s,s')$ as the final score of the *JanusNet* given a pair of face images (s,s') and S_i as the set of images that belong to

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

Train \ Test	A_1	A_2	A_3	$\{A_1,A_2\}$
A_1	0.8016	0.6603	0.6179	0.6888
A_2	0.8337	0.7859	0.7664	0.7950
A_3	0.5054	0.5028	0.5044	0.5002
$\{A_1,A_2\}$	0.7984	0.7505	0.7273	0.7620

Table 4: Accuracy comparison for the face verification task as a function of different angle subsets. Results are reported for the proposed Siamese model on depth data only. The description of head angle subsets is reported in Section 4.3.

the i-th subject, the following functions can be defined:

$$y = \arg\max_i J\left(s,s'\right),\ \forall s' \in S_i \tag{5a}$$

$$y = \arg\max_i \operatorname*{avg}_{s' \in S_i} J\left(s,s'\right) \tag{5b}$$

$$y = \arg\max_i \#\{S_i \mid J(s,s') \geq t\},\ \forall s' \in S_i \tag{5c}$$

In Equation 5a, the final identity corresponds to the couple with the max score, while the identity with the highest mean value is chosen in 5b. In Equation 5c, a voting procedure, in which every couple with a score greater than a threshold t is counted, is exploited to find the final identity. We found that best results are obtained using the mean scoring function, as reported in Table 2. Regarding the training procedure, we train *JanusNet* on *Pandora* dataset. Then, we finetune the depth network on *HRRFaceD* dataset. As shown in Table 2, our model outperforms previous state-of-art methods with a clear margin.

4.4 Internal Comparison

In the following, we test the depth Siamese module as a function of different head poses and subject sequences, defined in Section 4.1. *Pandora* dataset is used in every experiment. Table 3 presents results obtained training and testing the model on different sequence types. We note that including sequences $\{S_4,S_5\}$ in the training data reduces the performance on the whole dataset, since strong occlusions could compromise the learning of the network. Angle subset experiments, reported in Table 4, show that the face verification task becomes very challenging when head poses present extreme angles (A_3), since only small portions of faces are visible. Differently from A_3, training on A_2 provides higher accuracy than the other subsets and the whole dataset, since it provides a more representative distribution of the exploited dataset, without outliers (extreme angles) nor too easy samples. Last line of Table 4 contains performance of the model trained on the whole dataset defined as $\{A_1,A_2\}$.

5 Conclusion

In this paper, we present a framework, namely *JanusNet*, that tackles the face verification task using only depth maps at testing time. Moreover, the training procedure boosts accuracy performance by exploiting RGB images through a hybrid procedure. Shallow deep convolutional architectures are used in order to deal with the limited size of existing depth

datasets and to obtain a real time processing, with low hardware requirements. The proposed system achieves state-of-art results on two public datasets. This confirms the feasibility of the system in real-world scenarios, in which stable light sources cannot be guaranteed. The system is developed in a modular way: if RGB images are not available during the training phase, a single branch of JanusNet can be used, reaching satisfactory results.

(a) (b)

Figure 4: Samples of the output of *JanusNet* taken from the *Pandora* test set. Both correct (a) and wrong (b) predictions are depicted. The proposed system is able to handle face occlusions, wide poses and garments. Here, contrast is increased for a better visualisation.

References

[1] Timo Ahonen, Esa Rahtu, Ville Ojansivu, and Janne Heikkila. Recognition of blurred faces using local phase quantization. In *International Conference on Pattern Recognition*. IEEE, 2008.

[2] S Anith, D Vaithiyanathan, and R Seshasayanan. Face recognition system based on feature extration. In *IEEE International Conference on Information Communication and Embedded Systems*. IEEE, 2013.

[3] Stefano Berretti, Alberto Del Bimbo, and Pietro Pala. Superfaces: A super-resolution model for 3d faces. In *European Conference on Computer Vision*. Springer, 2012.

[4] Guido Borghi, Marco Venturelli, Roberto Vezzani, and Rita Cucchiara. Poseidon: Face-from-depth for driver pose estimation. In *IEEE International Conference on Computer Vision and Pattern Recognition*. IEEE, 2017.

[5] Jun-Cheng Chen, Vishal M Patel, and Rama Chellappa. Unconstrained face verification using deep cnn features. In *IEEE Winter Conference on Applications of Computer Vision*. IEEE, 2016.

[6] Jongmoo Choi, Ayush Sharma, and Gérard Medioni. Comparing strategies for 3d face recognition from a 3d sensor. In *IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 2013.

[7] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *IEEE International Conference on Computer Vision and Pattern Recognition*. IEEE, 2005.

[8] Liu Ding, Xiaoqing Ding, and Chi Fang. Continuous pose normalization for pose-robust face recognition. *IEEE Signal Processing Letters*, 2012.

[9] Tao Gao, XL Feng, He Lu, and JH Zhai. A novel face feature descriptor using adaptively weighted extended lbp pyramid. *Optik-International Journal for Light and Electron Optics*, 2013.

[10] Hosseinali Ghiassirad and Mohammad Teshnehlab. Similarity measurement in convolutional space. In *IEEE International Conference on Intelligent Systems*. IEEE, 2012.

[11] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. MS-Celeb-1M: A dataset and benchmark for large scale face recognition. In *European Conference on Computer Vision*. Springer, 2016.

[12] Judy Hoffman, Saurabh Gupta, and Trevor Darrell. Learning with side information through modality hallucination. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2016.

[13] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, 2007.

[14] Tri Huynh, Rui Min, and Jean-Luc Dugelay. An efficient lbp-based descriptor for facial depth images applied to gender recognition using rgb-d face data. In *Asian Conference on Computer Vision*. Springer, 2012.

[15] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *icml*, 2015.

[16] Bongjin Jun, Inho Choi, and Daijin Kim. Local transform features and hybridization for accurate face and human detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013.

[17] Juho Kannala and Esa Rahtu. Bsif: Binarized statistical image features. In *International Conference on Pattern Recognition*. IEEE, 2012.

[18] Güney Kayım, Cihan Sarı, and Ceyhun Burak Akgül. Facial feature selection for gender recognition based on random decision forests. In *21st Signal Processing and Communications Applications Conference*. IEEE, 2013.

[19] Mohamed Khalil-Hani and Liew Shan Sung. A convolutional neural network approach for face verification. In *International Conference on High Performance Computing & Simulation*. IEEE, 2014.

[20] Neeraj Kumar, Alexander C Berg, Peter N Belhumeur, and Shree K Nayar. Attribute and simile classifiers for face verification. In *IEEE International Conference on Computer Vision*. IEEE, 2009.

[21] Yuan-Cheng Lee, Jiancong Chen, Ching Wei Tseng, and Shang-Hong Lai. Accurate and robust face recognition from rgb-d images with a deep learning approach. In *British Machine Vision Conference*, 2016.

[22] Billy YL Li, Ajmal S Mian, Wanquan Liu, and Aneesh Krishna. Using kinect for face recognition under varying poses, expressions, illumination and disguise. In *IEEE Workshop on Applications of Computer Vision*. IEEE, 2013.

[23] Han Liu, Feixiang He, Qijun Zhao, and Xiangdong Fei. Matching depth to rgb for boosting face verification. In *Chinese Conference on Biometric Recognition*. Springer, 2017.

[24] Tomas Mantecon, Carlos R del Bianco, Fernando Jaureguizar, and Narciso García. Depth-based face recognition using local quantized patterns adapted for range data. In *IEEE International Conference on Image Processing*. IEEE, 2014.

[25] Tomás Mantecón, Carlos R del Blanco, Fernando Jaureguizar, and Narciso García. Visual face recognition using bag of dense derivative depth patterns. *IEEE Signal Processing Letters*, 2016.

[26] Gérard Medioni, Jongmoo Choi, Cheng-Hao Kuo, and Douglas Fidaleo. Identifying noncooperative subjects at a distance using face images and inferred three-dimensional face models. *IEEE Transactions on Systems, Man and Cybernetics*, 2009.

[27] Andreas Mogelmose, Chris Bahnsen, Thomas Moeslund, Albert Clapes, and Sergio Escalera. Tri-modal person re-identification with rgb, depth and thermal features. In *IEEE International Conference on Computer Vision and Pattern Recognition Workshops*, 2013.

[28] Olegs Nikisins, Kamal Nasrollahi, Modris Greitans, and Thomas B Moeslund. Rgb-dt based face recognition. In *International Conference on Pattern Recognition*. IEEE, 2014.

[29] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, et al. Deep face recognition. In *British Machine Vision Conference*, 2015.

[30] Adam Paszke, Sam Gross, Soumith Chintala, and Gregory Chanan. Pytorch, 2017.

[31] Marcelo Romero, Cesar Flores, Vianney Muñoz, and Luis Carlos Altamirano. Face recognition using eigensurface on kinect depth-maps. In *International Conference on Image Processing, Computer Vision, and Pattern Recognition*, 2016.

[32] Gaoli Sang, Jing Li, and Qijun Zhao. Pose-invariant face recognition via rgb-d images. *Computational Intelligence and Neuroscience*, 2016.

[33] Hamed Sarbolandi, Damien Lefloch, and Andreas Kolb. Kinect range sensing: Structured-light versus time-of-flight kinect. *Computer Vision and Image Understanding*, 2015.

[34] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2015.

[35] Shai Shalev-Shwartz, Yoram Singer, Nathan Srebro, and Andrew Cotter. Pegasos: Primal estimated sub-gradient solver for svm. *Mathematical Programming*, 2011.

[36] Viktoriia Sharmanska, Novi Quadrianto, and Christoph H Lampert. Learning to rank using privileged information. In *IEEE International Conference on Computer Vision*. IEEE, 2013.

[37] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 2014.

[38] Yi Sun, Ding Liang, Xiaogang Wang, and Xiaoou Tang. Deepid3: Face recognition with very deep neural networks. *arXiv preprint arXiv:1502.00873*, 2015.

[39] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2014.

[40] Matthew Turk and Alex Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 1991.

[41] Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural Networks*, 2009.

[42] Vladimir Vapnik, Akshay Vashist, and Natalya Pavlovitch. Learning using hidden information (learning with teacher). In *International Joint Conference on Neural Networks*. IEEE, 2009.

[43] Lior Wolf, Tal Hassner, and Itay Maoz. Face recognition in unconstrained videos with matched background similarity. In *IEEE International Conference on Computer Vision and Pattern Recognition*. IEEE, 2011.

[44] Eric P Xing, Michael I Jordan, Stuart J Russell, and Andrew Y Ng. Distance metric learning with application to clustering with side-information. In *Advances in Neural Information Processing Systems*, 2003.

[45] Xinxing Xu, Wen Li, and Dong Xu. Distance metric learning using privileged information for face verification and person re-identification. *IEEE Transactions on Neural Networks and Learning Systems*, 2015.

[46] Wenchao Zhang, Shiguang Shan, Wen Gao, Xilin Chen, and Hongming Zhang. Local gabor binary pattern histogram sequence (lgbphs): a novel non-statistical model for face representation and recognition. In *IEEE International Conference on Computer Vision*. IEEE, 2005.

[47] Zhengyou Zhang. Microsoft kinect sensor and its effect. *IEEE MultiMedia*, 2012.