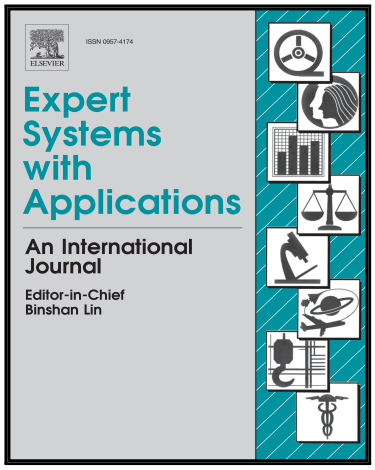
Journal Pre-proof



Learning Local Representations for Scalable RGB-D Face Recognition

Nesrine Grati, Achraf Ben-Hamadou, Mohamed Hammami

PII: S0957-4174(20)30144-5
DOI: https://doi.org/10.1016/j.eswa.2020.113319
Reference: ESWA 113319

To appear in: *Expert Systems With Applications*

Received date: 11 September 2019
Revised date: 16 February 2020
Accepted date: 16 February 2020

Please cite this article as: Nesrine Grati, Achraf Ben-Hamadou, Mohamed Hammami, Learning Local Representations for Scalable RGB-D Face Recognition, *Expert Systems With Applications* (2020), doi: https://doi.org/10.1016/j.eswa.2020.113319

Highlights

- Efficient learning of local representations for RGB-D face recognition.
- Efficient high-level decision fusion scheme based on a sparse representation.
- Dynamic dictionary selection for a scalable RGB-D face recognition.

Learning Local Representations for Scalable RGB-D Face Recognition

Nesrine Grati[a,*], Achraf Ben-Hamadou[b], Mohamed Hammami[a]

[a] *MIRACL-FS, Sfax University, Road Sokra Km 3 BP 802, 3018 Sfax, Tunisia*
[b] *Centre de Recherche en Numérique de Sfax, 3021, Sfax, Tunisie*

Abstract

In this article we present a novel RGB-D learned local representations for face recognition based on facial patch description and matching. The major contribution of the proposed approach is an efficient learning and combination of data-driven descriptors to characterize local patches extracted around image reference points. We explored the complementarity between both of deep learning and statistical image features as data-driven descriptors. In addition, we proposed an efficient high-level fusion scheme based on a sparse representation algorithm to leverage the complementarity between image and depth modalities and also the used data-driven features. Our approach was extensively evaluated on four well-known benchmarks to prove its robustness against known challenges in the case of face recognition. The obtained experimental results are competitive with the state-of-the-art methods while providing a scalable and adaptive RGB-D face recognition method.

Keywords: face recognition, SRC, data-driven descriptors, convolutional neural networks, BSIF, RGB-D sensors, deep learning.

*Corresponding author
Email addresses: grati.nesrine@gmail.com (Nesrine Grati), achraf.benhamadou@crns.rnrt.tn (Achraf Ben-Hamadou), mohamed.hammami@fss.usf.tn (Mohamed Hammami)

1. Introduction

Face recognition for an automated person identification has received great attention over the years as it offers the most user-friendly and non-invasive modality. Face recognition based on standard two dimensional (2-D) images was extensively studied but it still suffers from problems related to imaging conditions and face pose variations. Thanks to the progress in three-dimensional (3-D) technology, recent research has shifted from 2-D to 3-D (Abbad et al., 2018). Indeed, 3-D face representation ensures a reliable surface shape description and adds geometric shape information to the face characterization. Most recently, some researchers proposed to use image and depth data captured from cost-effective RGB-D sensors like MS Kinect or Intel RealSense instead of bulky and expensive 3-D scanners. In addition to color images, RGB-D sensors provide depth maps describing the scene 3-D shape by active vision or an alternative technology. Driven by the emergence of this type of sensors and the latest advances in deep learning techniques, RGB-D face recognition is now becoming at the heart of several recent research studies. Indeed, it is nowadays crystal clear that data-driven feature extraction, using Convolutional Neural Networks (CNNs) for example, outperforms traditional hand-crafted features for many computer vision tasks like object detection (Szegedy et al., 2013), image classification (Krizhevsky et al., 2012), *etc.* When it comes to the RGB-D face recognition, the observed challenges basically deal with face pose variations, partial occlusions, imaging conditions, and discriminant feature extraction.

In this article, we proposed multimodal data-driven representation for RGB-D face classification in a scalable manner to deal with typical issues like illumination variations, head pose variations, and disguise in controlled environments. Our contribution is many-fold. First, the proposed pipeline does not require any prior knowledge on the face pose nor does it rely on a semantic analysis of the face before performing the recognition. Typically, image interest points (like SURF or SIFT) can be detected offering a repeatable, efficient and stable results across different viewing conditions. These interest points are extracted

from facial images then RGB-D patches are simply extracted around to get local facial regions rather than an entire face. Second, in contrast to hand-crafted features, we proposed to learn discriminant local data-driven features based on deep learning techniques and statistical binary features for an optimal face patches representation. The effectiveness of combining these representations on both image and depth modalities was proved. Third, we suggested a patch matching algorithm based on a Sparse Representation Classification (SRC) method in a scalable way. Preliminary, the SRC algorithm had a dictionary that is filled with the entire samples in the gallery; but this would slow down the matching process of each patch to all patches in the gallery. A dynamic patch dictionary selection was then performed to pick only the closest patches from the gallery to drastically speed-up patch matching process. Finally, we proposed a late-fusion strategy leveraging the complementarity between image and depth data representations.

The remaining sections of this article are structured as follows. First, Section 2 gives an overview of the related work. Then, we detail the proposed RGB-D face recognition approach in Section 3. Section 4 summarizes the performed experiments and the obtained results to validate our approach. Finally, we conclude this study in Section 5 with some observations and perspectives for future work.

2. Related Work

In this section, we provide an overview on the RGB-D face recognition methods closely related to our work. This discussion can be intuitively driven under three categories. The first includes the initial efforts made to develop pose-invariant solutions where the main contributions generally focused on the pre-processing part. This can be explained by the poor quality of the depth data acquired with the first low-cost RGB-D sensors. In the second category, other solutions explored the adaptation of standard hand-crafted image descriptors to characterize RGB-D face data. Finally, as a last trend, RGB-D face recog-

nition techniques recently shifted from using hand-crafted features to applying learned-features, grounding on deep learning techniques.

The method of (Li et al., 2013) is among the first methods proposed for RGB-D face recognition. The pre-processing includes face data cropping from the 3-D scan by centering a sphere on the nose tip which is manually selected as the closest point to the sensor. Then, all the cropped facial scans are aligned with a generic face model using an Iterative Closest Point (ICP) algorithm to generate a canonical frontal view for both image and depth data. A symmetric filling process is then applied on the missing depth data caused by self-occlusion in non-frontal poses. For image data, the Discriminant Color Space (DCS) operator is used as a feature extractor. Then, the pre-processed depth map and the 2-D DCS features are classified separately by applying an SRC algorithm before performing a late fusion to obtain the final identity of a given probe.

(Hsu et al., 2014) fit a 3-D face model to the face data to build a 3-D textured face model for each person in the gallery. For a new probe, the face pose is estimated based on facial landmarks detection (Zhu & Ramanan, 2012) to be able to apply it on all the 3-D textured models in the gallery. This allows generating 2-D images corresponding to the probe facial pose by plan projection. Then, a Local Binary Pattern (LBP) descriptor is applied on all the projected 2-D images to perform the classification using an SRC algorithm.

Similarly, (Sang et al., 2016) estimate the face pose from the probe samples by aligning a template face model to the depth data using an ICP algorithm then the image data in the gallery can be rendered to the same view as the probe. For feature extraction, the well-known Histograms of Oriented Gradients (HOG) operator is applied on both image and depth data, then a Joint Bayesian Classifier is used and the final decision is made by a weighted sum of similarity scores obtained from the image and depth classification.

It can be clearly observed that these previous approaches focused on pre-processing especially when dealing with pose variation by aligning the probe to the gallery samples. Although this kind of sequential processing may lead to error propagation from pose estimation to the classification, it yielded promising

results (Hsu et al., 2014). Alternatively, to deal with pose variation, (Ciaccio et al., 2013) propose to complete the gallery by generating a set of new images, from a single RGB-D data, corresponding to a large range of predefined face orientations. This is achieved in the following way. Firstly, all faces are cropped then aligned based on a facial landmarks detector (Zhu & Ramanan, 2012). Then, each face is rotated around the Y axis every 5 degrees to render new images. Now, for all gallery samples, including both original and generated ones, each face image is represented by a set of densely sampled patches of 10×10 pixels using a step size of 5 pixels. The patches corresponding to self-occluded parts are identified and then discarded based on the estimated pose. The remaining patches are described by the LBP and co-variance descriptor computed from pixel locations, intensity derivatives, and edge orientations. Then, the matching part is performed using only the filtered patches of each gallery face image based on the Euclidean distance in the feature space. The similarity measures over the selected patches are integrated together and normalized to obtain the final similarity score between the probe and the gallery set. Finally, a combination with probabilistic integration of the resulting scores were made and a Bayesian decision was performed.

In the second category of methods, this overview focused on those methods mainly interested in the feature extraction part from RGB-D face data. In (Dai et al., 2015), an Enhanced Local Mixed Derivative Pattern descriptor is separately applied on 2-D Gabor features extracted from image and depth data. This descriptor is a mixed feature descriptor of different orders of local derivative patterns and local binary patterns. To attribute an the identity of a given probe, a nearest neighbor search algorithm is applied separately for each modality, and the final similarity score is produced by combining the scores computed for both image and depth modalities. In (Goswami et al., 2014), faces are represented by a set of texture features and geometric attributes computed from both image and depth data. For texture features, the HOG operator is applied on saliency and entropy maps obtained from both image and depth data. The set of geometric attributes are computed based on Euclidean distances between facial landmarks

located on the depth maps. Finally, a random forest classifier is used for the classification part. (Boutellaa et al., 2015) explored more feature combinations, a bunch of hand-crafted features (*e.g.,* LBP, Local Phase Quantization (LPQ), and HOG) was applied respectively on RGB and depth face crops, and finally a Support Vector Machine (SVM) classifier was used for the classification.

(Kaashki et al., 2018) also explored the usage of similar feature extractors like HOG, LBP, and 3DLBP. However, these descriptors were applied locally on patches around located facial landmarks. An SVM classifier was also used for the classification. In (Hayat et al., 2016) an image set classification is proposed for RGB-D face recognition. For a given set of images, the face regions and the head poses are, firstly, detected using (Fanelli et al., 2011) then clustered into multiple subsets according to the estimated pose. A block based covariance matrix representation from the LBP features is applied to model each subsets on the Riemannian manifold space. As classifier, an SVM is used for each subset for both modalities, and a final decision with a majority vote rule is made.

Unlike hand-crafted features, feature learning has started to draw increasing interest for face recognition, initially, in 2-D image-based approaches and currently on RGB-D data-based approaches. Recently, and for image set classification (Hayat et al., 2014, 2015) proposed a deep learning approach based on Auto-Encoder (AE) to learn a class-specific model called Deep Reconstruction Model (DRM) for each set of images. In the offline phase, Template Deep Reconstruction Model (TDRM) weights are, firstly, initialized using Gaussian Restricted Boltzmann Machines (GRBMs) and then fine-tuned for each class of the training image sets.

In the test phase, given a new probe, the face image and depth data are encoded and decoded using all the learned class-specific models separately. The underlying idea is to perform the classification based on the evaluation of the residual error between the original face data and the reconstructed ones (*i.e.,* output of the decoder network) for all the learned models. The major drawback of this approach is the lack of scalability as it requires to learn a class-specific model for any new person to add to the gallery, and for the test phase, all the

models should be run on the input set of data to evaluate the reconstruction errors. This means that the run-time is linearly dependent on the number of persons in the gallery.

(Lee et al., 2016) proposed to learn deep features from both image and depth data. The CNN model is, firstly, trained on color and gray-scale facial images. Then, the obtained model is fine-tuned on depth face data for transfer learning. A step of depth enhancement is performed to recover the facial depth image by projecting the depth pixels onto the 3-D space and rendering back onto images again after a series of processing steps like noise reduction, depth fusion, hole filling, pose estimation, frontalization, *etc.* For classification, an SVM is applied with probability estimation taking into consideration deep representation similarities, head pose and database similarity standard deviation to estimate a confidence score and make the final decision.

(Zhang et al., 2018) introduced a novel method which processes the multi-modal and the cross-modal matching allowing measuring the similarity between image and depth data. In more details, a set of complimentary and common features are learned from image and depth data. On the one hand, the authors started with learning two modality-specific feature networks based on Inception-v2 (Ioffe & Szegedy, 2015), then they introduced a joint loss architecture taking activation from both networks to enforce complementary feature learning. On the other hand, for learning heterogeneous feature from image and depth data, the modality-specific features are used again to obtain RGB-to-RGB and RGB-to-depth matching scores. Finally, the resulting similarity scores are combined with a weighted sum rule.

(Neto et al., 2019) proposed a depth-based face recognition approach by learning from 3D-LBP images. Firstly, two 3D-LBP variants are computed from the depth image. Then, a shallow CNN is designed for classification. The resulting scores in the last softmax layer for each descriptor image are combined with weighted sum rules to get the final decision.

The previous overview makes it clear that all of these approaches require a pre-processing step to precisely localize the face, estimate its pose, or even

accurately detect face landmarks which could be prone to error propagation in a sequential processing and add further dependencies to the approach. For example, (Hsu et al., 2014; Li et al., 2013; Ciaccio et al., 2013; Sang et al., 2016) aimed mainly to overcome pose variations either through pose correction or gallery completion by generating new images in different views. For data representation, the aforementioned works (Dai et al., 2015; Goswami et al., 2014; Boutellaa et al., 2015; Kaashki et al., 2018) settle for the adaptation of the classic hand-crafted descriptors (*i.e.,* HOG, LBP, *etc.*) while the extraction of more appropriate features could be obtained by data-driven learning techniques. Later and with the arrival of the new era of deep learning techniques, the alternative approaches like those of (Lee et al., 2016; Hayat et al., 2015; Zhang et al., 2018) started to take benefit from learning more appropriate features and boosted their RGB-D face recognition performances. Opposite to (Hayat et al., 2015) who focused on intra-class compactness and did not consider the relationship between classes with maximizing the inter-class separability, (Zhang et al., 2018; Lee et al., 2016) proposed to learn discriminant features for a multi-modal recognition taking the whole face as input. While only a few techniques apply local learned features for a RGB-D face recognition, our approach highlighted how to learn a discriminant representation of local regions in the face data, and showed an undeniable ability to compete with standard hand-crafted features in the case of RGB-D face recognition. Compared to the global description of the whole face images, local features have proven to be robust against many variations especially occlusion (Tan et al., 2006). Generally for local feature extraction, feature detectors are applied to extract the face distinctive information from local regions. These regions are cropped with sampling the input image with fixed stride in grid or alternatively around detected landmarks. In our approach, we simply considered locating a set of image interest points on the face to get rid of facial landmarks detection or further facial analysis. A given face is represented by a set of patches around salient detected image interest points. Each of these patches is transformed using a selection of learned descriptors, namely CNNs and the Binarized Statistical Image Features (BSIF) (Kannala

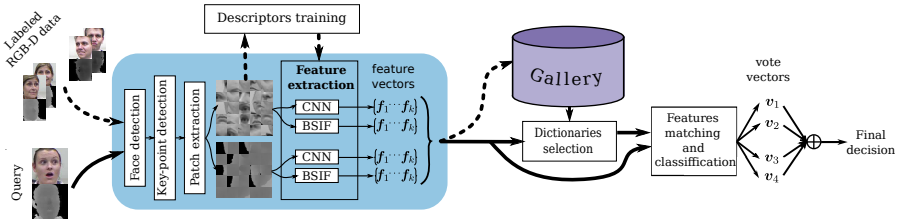


Figure 1: Flowchart of the proposed RGB-D face recognition approach. The pipeline involves online and offline stages sharing some processings grouped under the highlighted block like face detection, patches extraction and CNNS and BSIF features computing. The offline phase is outlined with dashed arrows to indicate the gallery construction and the descriptors training. In the online phase, a dynamic dictionary selection is performed whenever a given feature vector is matched to the gallery. The process of feature matching and classification generates four vote vectors corresponding to the four applied descriptors. These vote vectors are then combined to generate the final decision.

& Rahtu, 2012), before feeding the classification part. For CNNs, we propose an effective training algorithm leading to a discriminant space for face patches representation. Additionally, although the majority of deep learning approaches often rely on the deep architecture and on the availability of huge amounts of data in order to achieve state-of-the art performance, we believe that building a shallow deep model with fewer parameters can efficiently learn discriminant local features from small data and achieve a concurrent performance. BSIF is a popular statistical descriptor used for several computer vision tasks. (Boutellaa et al., 2015) proved the usefulness of low-resolution depth data in different face analysis tasks compared to hand-crafted features and achieved prominent classification rates. Here, we demonstrated the effectiveness of combining statistical and CNNs features to properly describe facial local patches. Finally, the correspondence between patches was performed based on the SRC algorithm with a dynamic patch dictionary selection. The final classification decision is obtained by a majority vote rule.

3. Proposed RGB-D face recognition approach

The general pipeline of the proposed approach is outlined in Figure 1. It involves online and offline phases sharing some processing blocks like raw data pre-processing (*i.e.,* median and bilateral filtering), face localization, patch ex-

traction and feature vector computing. The offline phase is mainly dedicated to train or update the data-driven descriptors and construct the gallery. Whereas the online phase is dedicated to the identity recognition given a face query. This online phase follows the following steps. Firstly, the face is localized in the image. It is then represented by a set of patches cropped around the image interest points extracted on the face. We considered two data-driven descriptors, namely CNN and BSIF, applied on both input and gallery patches. The input patches are matched to those of the gallery based on their feature vector and using a sparse representation algorithm. The application of this algorithm on each patch separately yields a set of votes. These are later combined to obtain the final identity of the input face.

The remainder of this section detailed the main modules involved in our proposed approach.

3.1. Face pre-processing and patch extraction

The face pre-processing shared between the offline and online phases of our system includes median and bilateral filtering for the depth maps and face localization (Zhu & Ramanan, 2012)[1]. The face detection is performed on the texture image then the obtained bounding box is mapped on the depth image. The cropped face region is resized to 96×96 pixels to ensure a normalized face spatial resolution. To get rid of face landmarks localization, we only consider the image interest points without any further semantic analysis and without loss of generality. In other words, we do not try to catch specific facial landmarks. Although the use of facial landmarks may appear to be intuitive, we believe that, overall the whole pipeline the use of image interest points is more suitable and robust against all variations and challenges related to the facial recognition in our context. Fundamentally, we believe that a precise facial landmarks localization is still challenging under head pose variations, facial expressions and is not robust enough under facial occlusions. Also, facial landmarks detection

[1]We used only the face localization, facial landmarks were not used.

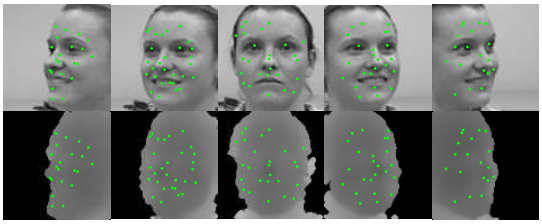


Figure 2: Illustration of extracted image interest points from image and depth data under different angles of view.

is an independent research field, and we would like to get rid of an additional dependency in our pipeline. This can be justified by the fact that imprecise landmark detection errors could propagate to the rest of the pipeline and affect the recognition performance. Instead, the image interest point detection does not need any advanced analysis of the face region. It is rather stable, straightforward, and fast.

We used SURF detector (Bay et al., 2006) to extract image interest points on the cropped and resized face images. SURF presents an efficient scale and rotation invariant detector and descriptor which outperforms all other feature detection techniques in terms of repeatability, distinctiveness, robustness and speed using the concept of integral image and Hessian approximation (Bay et al., 2006, 2008). The coordinates of the detected SURF interest points are mapped from image to depth data using the sensor calibration (see Figure 2 for an example). Around each interest point, we extract two patches of 21×21 pixels from both image and depth data. Again, the mapping between image and depth data is ensured by the RGB-D sensor calibration.

3.2. RGB-D data-driven descriptor

This section details the data-driven descriptors used for an optimal facial patch description. We first introduce our CNN architecture to learn discriminant patch features. Secondly, we present the BSIF descriptor and how it was adapted for local face patch representation.

3.2.1. CNN based features

Features based on deep learning are widely applied on 2-D face recognition and has achieved promising results in many works like DeepFace (Taigman et al., 2014), DeepID2 (Sun et al., 2014), FaceNet (Schroff et al., 2015), VGG-DeepFace (Parkhi et al., 2015), Center Loss (Wen et al., 2016), SphereFace (Liu et al., 2017), Cosface (Wang et al., 2018), ArcFace (Deng et al., 2018). The success of these methods is attributed to the ability of the CNN to learn rich features from the whole face. In our case, the face is represented by a set of small patches (21×21 pixels), pre-trained networks, like VGG, for feature extraction can not be applied. We rather designed a relatively shallow CNN architecture detailed in Figure 3. We chose standard processing blocks instead of complex ones like skip connections, residual, *etc.* due to the small input patches and the shallowness of the network. It consists of four types of processing layers including 2-D Convolution, Normalization, Pooling, and Fully Connected Layer (FCL). This architecture produces at the end a feature vector of 128 dimensions for a given input patch. We separately trained the designed CNN for both image and depth modalities. Many loss functions have been proposed to train the CNNs such as contrastive and cross-entropy losses. In our case, we opted for the triplet loss (Schroff et al., 2015). It has the advantage of enforcing the inter-class separability and in a same way the intra-class compactness. In other words, triplet loss ensures a discriminant patch representation by enforcing the closest patches to emanate from a same person. It takes a triplet of patches as input in the form $\{\mathcal{A}, \mathcal{P}, \mathcal{N}\}$, where $\mathcal{A}$ is the anchor patch, $\mathcal{P}$ is the positive patch, which is a different sample from the same class person as $\mathcal{A}$, and $\mathcal{N}$ standing for negative patch is a sample belonging to a different class. The objective of the optimization process, is to update the network parameters in such way that the patches $\mathcal{A}$ and $\mathcal{P}$ become closer in the embedded feature space, and $\mathcal{A}$ and $\mathcal{N}$ are further apart in terms of their Euclidean distances as shown in Figure 4.

The triplet loss formula is given in Equation (1) where f function stands for

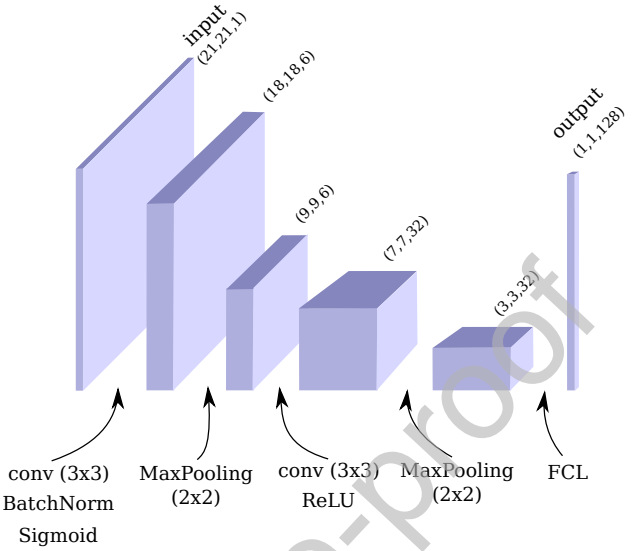


Figure 3: The proposed CNN architecture to be trained separately on image and depth data.

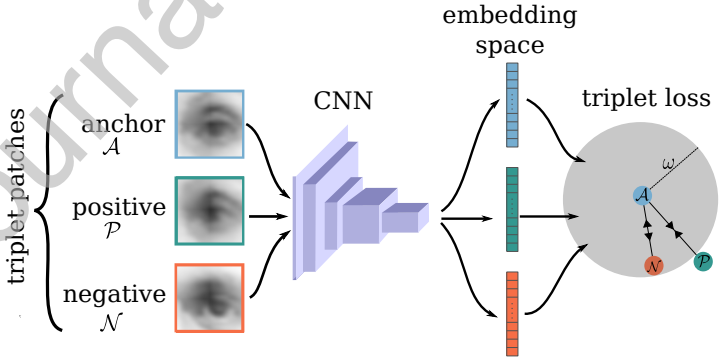


Figure 4: Illustration of the triplet loss with facial patches images. (left) input patches $(\mathcal{A}, \mathcal{P}, \mathcal{N})$, transformed by the CNN network into the embedding space (middle). In this space, the objective of the triplet loss is to pull the representation of $\mathcal{P}$ inside the ω-radius hypersphere centered on the representation of $\mathcal{A}$ and push the representation of $\mathcal{N}$ out of the same hypersphere.

the application of the CNN on a given input patch ($\mathcal{A}$,$\mathcal{P}$, or $\mathcal{N}$) to generate a feature vector. The triplet loss equation involves an additional parameter called the margin ω [2] defining how far away the dissimilarities should be.

$$\mathcal{L} = \sum_{a,p,n \in \{\mathcal{A},\mathcal{P},\mathcal{N}\}} \left(||f(a) - f(p)||_2^2 - ||f(a) - f(n)||_2^2 + \omega \right) \quad (1)$$

Minimizing $\mathcal{L}$ enforces the maximization of the Euclidean distance between patches from different classes which should be greater than the distance between anchor and positive features. For an efficient training, only the triplet patches that independently verify the constraint $\mathcal{L} > 0$ are online selected during the optimization iteration as valid triplets. In practice, a single patch triplet is obtained following these 3 steps:

1. Randomly select one anchor patch from the pool of patches related to a given person c.
2. Randomly select the positive patch from the remaining patches in the same pool.
3. Randomly select a negative patch from the patch pool related to other persons ($\neq c$)

3.2.2. BSIF based features

The BSIF descriptor was previously applied in (Boutellaa et al., 2015) to prove the usefulness of low-resolution depth data in different face analysis tasks. It achieved the best classification rates when compared to the use of hand-crafted features for the same purpose. The BSIF is a data-driven image descriptor that aims to compute a binary code for each pixel in an input image to represent its local structure. The value of each bit within the BSIF descriptor is computed by quantizing the response of a linear filter. Each bit in the binary code is associated to a specific filter, and the desired length of the bit string determines the number of filters used. In the original BSIF method, the authors used a pre-

[2] In our case ω is experimentally set to 0.2.

defined number of convolution filters are learned from a set of training image patches selected to maximize the statistical independence between the responses of the convolutions of each individual filter and the given image patches (Kannala & Rahtu, 2012). A BSIF code of length M bits is computed for a given image patch $\boldsymbol{X}$ through the following equation:

$$BSIF_M = \sum_{m=1}^{M} b_m \times 2^{m-1} \tag{2}$$

$$\text{where } b_m = \begin{cases} 1 & \text{if } \boldsymbol{W}_m * \boldsymbol{X} > 0 \\ 0 & \text{otherwise} \end{cases}$$

In Equation (2), $*$ is the convolution operator and $\boldsymbol{W}_m$ is the m-th filter of the same size as the image patch $\boldsymbol{X}$. The response of the application of a filter ($\boldsymbol{W}_m * \boldsymbol{X}$) is thresholded at zero to obtain the m-th bit of the BSIF code corresponding to $\boldsymbol{X}$. On a full image or a larger region, Equation (2) is applied in exactly the same way at each position in the input image yielding a BSIF code for each pixel of the input image.

Originally, the BSIF filters were learned for texture analysis purposes on a bunch of patches extracted randomly from a set of natural images (*e.g.,* landscapes, grass, *etc.*). However, it could be trained for different applications and contexts. In our case, we trained the BSIF filters on image and depth patches separately. We consider different filter sizes $\{7, 9, 11, 13, 15, 17\}$ combined to different bit lengths $\{5, 6, 7, 8, 9, 10, 11, 12\}$ resulting in a total of 48 filters to be trained for each modality. The BSIF filters training consists mainly of three steps: 1) mean subtraction of each patch, 2) dimensionality reduction using Principle Component Analysis, and 3) estimation of statistically independent filters (or basis) using Independent Component Analysis. Figure 5 displays an illustration of the learned BSIF filters applied to an image patch.

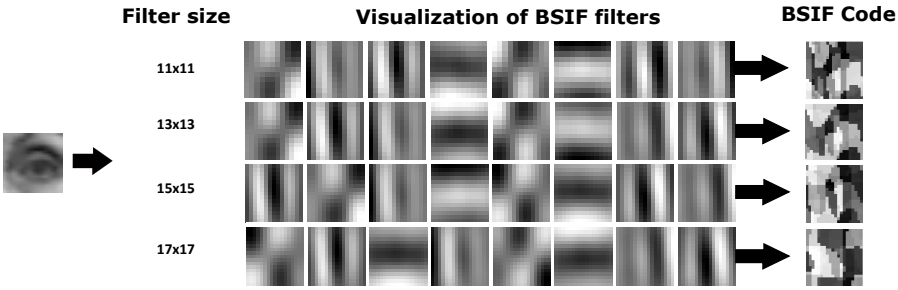


Figure 5: Example of extracted BSIF features. Input patch (on the left) is transformed using the learned BSIF filters (visualized in the middle) with different sizes $11 \times 11, 13 \times 13, 15 \times 15, 17 \times 17$ to obtain their corresponding BSIF codes (on the right). Please note that the filters are resized only for the purpose of illustration.

3.3. Patch matching and classification

Given a new RGB-D probe face, a set of K patches is extracted from both image and depth data, and then transformed using their corresponding CNN and BSIF descriptors to obtain a set of 2 feature vectors per extracted patch and per modality, which yields a total of $4 \times K$ feature vectors representing the input face. Let us define $d \in [1, .., 4]$ as the index of descriptors that can take one of the 4 options; either CNN or BSIF descriptors applied on either image or depth patches. We propose to match each feature vector separately to the gallery before taking the final decision. The matching algorithm we propose is based on a sparse representation technique where the main objective is to approximate an input feature vector by a sparse linear regression of a dictionary constructed from the feature vectors computed on dynamically selected gallery patches. As described in Algorithm 1, the matching process is outlined in three main steps repeated separately for each feature vector $\boldsymbol{y}_d^k$: 1) adaptive and dynamic dictionary selection; 2) sparse representation of $\boldsymbol{y}_d^k$; and 3) prediction of the person identity corresponding to the given feature vector. In more details and for a given feature vector $\boldsymbol{y}_k^d$, an adaptive dictionary $\hat{\boldsymbol{D}}_k^d$ is selected gathering the nearest atoms in the putative full dictionary $\boldsymbol{D}_k^d$ corresponding to the descriptor d (see Equation (3)). We used KD-Tree search algorithm following our previous work (Grati et al., 2016). This allows us to substantially speed-up the overall matching process and maintain a scalable classification with respect

to the gallery size. Afterward, we approximate $\boldsymbol{y}_k^d$ by a sparse linear regression of the selected dictionary $\hat{\boldsymbol{D}}_k^d$ (see Equation (4)). In Equation (4), $\boldsymbol{x}_k^d \in \mathbb{R}^{\hat{N}}$ corresponds to the resulting sparse coefficient vector whose non-zero values are related to the atoms in $\hat{\boldsymbol{D}}_k^d$ contributed to the reconstruction of $\boldsymbol{y}_k^d$, and $\hat{N}$ is experimentally fixed to 20. We used (Mairal et al., 2010) to solve the LASSO ℓ_1 minimization problem in equation (4). The obtained sparse representation $\hat{\boldsymbol{x}}_k^d$ is then used in Equation (5) to compute $\boldsymbol{r}_k^d$ vector gathering the reconstruction error for each class in the gallery. In Equation (5), $\boldsymbol{M}_{k,i}^d$ is a diagonal matrix whose values take 1 for all atoms in $\hat{\boldsymbol{D}}_k^d$ emanating from the i-th class and 0 otherwise. Finally, the identity attributed to the current feature vector $\boldsymbol{y}_k^d$ corresponds to the lowest reconstruction error in $\boldsymbol{r}_k^d$ (see Equation (6)) which allows incrementing the associated vote in $\boldsymbol{v}^d$ as shown in Equation (7).

Algorithm 1: Sparse Representation based Matching.

Input:

- A set of query feature vectors $\{\boldsymbol{y}_k^d\}_{k=1,..,K}$ of the descriptor d computed on K patches of a query face.
- Putative dictionary $\boldsymbol{D}^d$ corresponding to the descriptor d and consisting of the set of feature vectors computed for the descriptor d on all the N patches in the gallery.
- C is the number of classes.

Output: Vote vector $\boldsymbol{v}^d$ corresponding to the descriptor d.

- Initialize $\boldsymbol{v}^d$ values to zero.

for *$k = 1$ to K* **do**

Step 1:

- Select the dynamic dictionary $\hat{\boldsymbol{D}}_k^d$ from $\boldsymbol{D}^d$ gathering the nearest $\hat{N}$ atoms to $\boldsymbol{y}_k^d$ using KD-Tree algorithm, where $\hat{N} \ll N$:

$$\hat{\boldsymbol{D}}_k^d = \text{KD-Tree}(\boldsymbol{y}_k^d, \boldsymbol{D}^d) \tag{3}$$

Step 2:

- Approximate $\boldsymbol{y}_k^d$ by a sparse linear regression of the corresponding dictionary $\hat{D}_k^d$ (*i.e.*, $\boldsymbol{y}_k^d \approx \hat{\boldsymbol{D}}_k^d\ \hat{\boldsymbol{x}}_k^d$) by solving the LASSO ℓ_1 minimization problem:

$$\hat{\boldsymbol{x}}_k^d = \arg\min_{\boldsymbol{x}_k^d} \|\hat{\boldsymbol{D}}_k^d \boldsymbol{x}_k^d - \boldsymbol{y}_k^d\|_2 + \lambda \|\boldsymbol{x}_k^d\|_1 \tag{4}$$

Step 3:

- Calculate the reconstruction residual as:

for *each class $i = 1$ to C* **do**

$$\boldsymbol{r}_k^d(i) = \|\boldsymbol{y}_k^d - \hat{\boldsymbol{D}}_k^d\ \boldsymbol{M}_{k,i}^d\ \hat{\boldsymbol{x}}_k^d)\|_2 \tag{5}$$

end

- Select the identity to attribute to the current feature vector $\boldsymbol{y}_k^d$ corresponding to the lowest reconstruction error:

$$\tilde{c} = \arg\min_i\ \boldsymbol{r}_k^d(i) \tag{6}$$

- Increment the number of votes associated to the selected class $\tilde{c}$:

$$\boldsymbol{v}^d(\tilde{c}) += 1 \tag{7}$$

end

3.4. Votes fusion and final decision

After the application of the matching algorithm on all the set of feature vectors, we end up with a set of four vote vectors $\{\boldsymbol{v}^d\}_{d=1,..,4}$. These should be combined into a single vote vector to deduce the final identity to attribute to the face query. To this end, we simply sum all the resulting votes and apply a majority vote rule to obtain the final decision, which corresponds to the index of the highest vote over the elements of $\hat{\boldsymbol{v}}$ (see Equation (8)).

$$\hat{\boldsymbol{v}} = \sum_{d=1,..,4} \boldsymbol{v}^d \tag{8}$$

It is worth noticing that two other fusion schemes were tested at two different levels of our pipeline:

- raw data-level fusion: the corresponding image and depth patches are concatenated at the very beginning and fed as two-channels input tensors to a single CNN network. In this case, the CNN leverages the complementarity between image and depth patches to output a single feature vector.
- feature-level fusion: two CNNs are applied separately on each modality and combined via attention module to learn the most informative and discriminant components of a pair of representations yielding a single feature vector.

The reader can find more details on these fusion schemes in the technical report given in (Grati et al., 2020). Although the majority vote rule looks simple, it outperforms the other elaborated fusion solutions and allows the approach to be versatile and adaptive in case other local descriptors should be added or in the case of missing modality (*i.e.,* image or depth).

4. Experimental results

We conducted a set of experiments to assess the performance of our proposed RGB-D face recognition approach against known challenges in face recognition

applications, including face pose variations, partial occlusions, and imaging conditions. We keep the same experimental protocols followed by the state-of-the-art works using four well-known benchmark databases. The obtained results are separately presented for each database to simplify the comparison with the state-of-the-art results. In addition, we pushed further our experiments to deal with large-scale and heterogeneous databases by combining all the four benchmark databases into a single, larger and more challenging database. Finally, we compared our patch-based approach to state-of-the-arts for CNN-based face representation and recognition. All the reported results in all the conducted experiments were obtained by averaging five runs with five separately trained CNN models and BSIF filters.

The remaining of this section is organized as follows: we first start by introducing the four considered benchmark databases, then we detail the settings of our descriptors training, and finally we provide and discussed the obtained experimental results.

4.1. Evaluation databases

Four RGB-D face recognition benchmark databases were used for the validation of our approach: CurtinFaces (Li et al., 2013), Eurecom (Min et al., 2014), BIWI (Fanelli et al., 2011), and VAP (Hg et al., 2012).

- CurtinFaces database consists of 52 subjects, 10 females and 42 males. Each subject has 97 images taken under different conditions: combinations of 7 facial expressions, 7 poses, 5 illumination configurations, and 2 occlusion situations. This database is considered as the most challenging database as it offers different combinations in terms of variations of poses, and expression and illumination.

- Eurecom database is also made up of 52 subjects, 14 females and 38 males. Each person has a set of 9 images taken at two different time sessions (separated about half a month). Each session covers 9 settings: neutral,

smiling, open mouth, illumination variation, left end right profile, occlusion on the eyes, occlusion on the mouth, and finally occlusion with a white paper-sheet.

- BIWI Kinect Head Pose database contains 24 sequence of 20 different persons (6 females and 14 males). We refered to it here as the BIWI database for simplicity. This database is actually the least used database for face recognition validation since it was originally released for head pose estimation tasks.

  This database brings more challenges to deal with face pose variations. Indeed, each person rolls his face at different orientations within $\pm 75^{\circ}$ in yaw, $\pm 60^{\circ}$ in pitch, and $\pm 50^{\circ}$ in roll.

- VAP face database was initially proposed for testing a face detection algorithm. It contains 1581 images taken for 31 persons (1 female and 30 males). The dataset has 51 images for each person captured under 17 different variations in poses and expressions repeated three times. Thirteen points on a wall behind a Kinect sensor were chosen and each person looked at these points sequentially to achieve roughly the same angles for each person.

4.2. Settings of the descriptors training

The BSIF filters training is relatively straightforward. In practice, we followed the same training procedure described in (Hyvärinen et al., 2009) and we used the original implementation[3]. Basically, there are only the filter sizes, the bit lengths, and the input training data to be set for this training. We randomly selected a set of 20 image and depth facial data. We set the patch size to the desired filter size, then each image is randomly sampled yielding a large set of about 50000 extracted patches for each modality.

[3]http://www.naturalimagestatistics.net/

Our CNN trainings, however, required parameters to tune. All the trainings have been performed with a batch size of 64, a decay and momentum values of 0.0005 and 0.09 respectively, and an initial learning rate set to 0.001. We used PyTorch[4] framework to implement and train the CNNs. For each database, the training data were obtained from a classical train/validation split, and the pool of patch triplets as needed for our loss calculation were generated from all the persons equally and updated each succession of 10 optimization epochs. We kept the same training settings for all the databases and experiments. The CNN model were trained exclusively on patches extracted from the gallery sets. Firstly, the CurtinFaces database was used to obtain the first models; then, they were fine-tuned separately on the other databases (*i.e.,* EURECOM, BIWI and VAP). The training patches were selected the same way for the four databases. It consists in selecting all the patches in the gallery set which corresponds to 32 patches in average per modality and per sample.

4.3. Experiments on CurtinFaces database

We followed the same protocol originally proposed in (Li et al., 2013) and adopted by most of the works in the literature. It defines a bunch of evaluation experiments to assess the face recognition methods performance in dealing with many challenges like head pose, illumination and facial expression variations and occlusion. Here the gallery for all the experiments contains 18 captures per person. Each of these captures involved only one kind of variation, namely, illumination, pose or expression variations. CurtinFaces benchmark is considered as the most challenging and representative database. We took advantage of the experiments performed on this database to evaluate the complementarity between the image and depth modalities, and also between the CNN and BSIF features. To this end, we derived four additional baselines from our method, named Our$_{image}$, Our$_{depth}$, Our$_{CNN}$, and Our$_{BSIF}$. Our$_{image}$ and Our$_{depth}$corresponded to our method and used either image or depth data as

[4]https://pytorch.org/

Figure 6: Sample of probe images of a smiling person under various poses. The frontal view is at the center, on the left and right sides images corresponding to face yaw variations with ±30°, ±60°, and ±90°, and finally, the top and bottom images correspond to ±60°pitch variation.

input while Our$_{CNN}$ and Our$_{BSIF}$corresponded to our method and considered either CNNs or BSIF features for the patch description.

4.3.1. Experiments on CurtinFaces database under pose and facial expression variations

These experiments were performed on the CurtinFaces section including non-occluded faces with frontal, left and right profile view added to 4 yaw and 2 pitch distinct poses coupled to 6 different facial expressions. A total of 39 probes per subject were considered for this experiment (see Figure 6 for a sample).

Table 1 summarizes and compares the obtained recognition rates to those of the state-of-the-art methods. From a first glance, one could observe that all the results are very competitive for the frontal view, yaw pose angles ±30°and ±60°, and pitch pose angles ±60°. Our approach outperforms those of (Ciaccio et al., 2013; Li et al., 2013; Kaashki et al., 2018).

For extreme yaw profile poses (±90°), (Hsu et al., 2014; Sang et al., 2016) stands out as they rely on a specific processing to handle pose variation. They achieved 93.5% and 95.1%, respectively, while we obtained 86.55% without any additional processing to deal with pose variation. In general, our approach shows promising results compared to the state-of-the-art and the only drop in the recognition rates is observed for yaw pose angles ±90°. This can be explained by the fact that CurtinFaces provides only one capture per person for this pose,

Table 1: Comparison of recognition rates on CurtinFaces database under pose and facial expression variation.

Pose	Ciaccio et al. (2013)	(Hsu et al., 2014)	(Li et al., 2013)	(Sang et al., 2016)	(Kaashki et al., 2018)	Proposed method		
						Our_{image}	Our_{depth}	Our
Frontal	N/A	100%	100%	100%	100%	100%	100%	**100**%
Yaw±30°	94.2%	99.4%	99.4%	99.5%	90.3%	99.35%	99.35%	**99.84**%
Yaw±60°	84.6%	98.2%	98.2%	98.4%	58.6%	97.92%	95.67%	**98.45**%
Yaw±90°	75.0%	93.5%	84.6%	**95.1**%	N/A	84.61%	78.85%	86.55%
Pitch±60°	N/A	N/A	92.8%	96.7%	97.6%	96.8%	97.2%	**98.1**%

which is that of the probe, and consequently, there are no patches corresponding to this extreme profile pose in the gallery.

The second observation we can make from Table 1 is that the two baselines Our_{image} and Our_{depth} achieved similar and relatively high recognition rates, although they are still lower than those of our full pipeline. This highlights the complementarity between both image and depth modalities for our RGB-D face recognition approach. Under this CurtinFaces section, we also evaluated and compared the pipelines Our_{CNN} and Our_{BSIF} in order to study the combination of the considered CNN and BSIF data-driven features. Figure 7 shows the cumulative match curve (CMC) of Our_{CNN} and Our_{BSIF} in addition to our full pipeline. Roughly speaking, the curves of Our_{CNN} and Our_{BSIF} seem to be similar, however the curve of the full pipeline is significantly higher, especially for the first ranks (*e.g.,* 99.73% against 98.1% and 97.52% at rank 1). These results plead in favor of the combination of CNN and BSIF features for facial RGB-D patches representation.

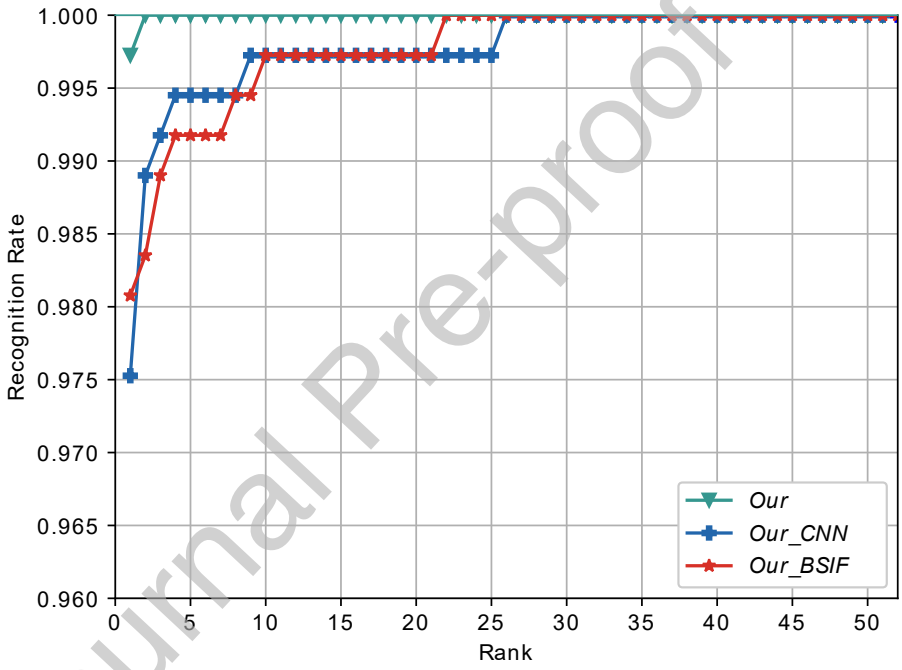


Figure 7: The CMC of our proposed method in comparison with the pipelines Our_{CNN} and Our_{BSIF}.

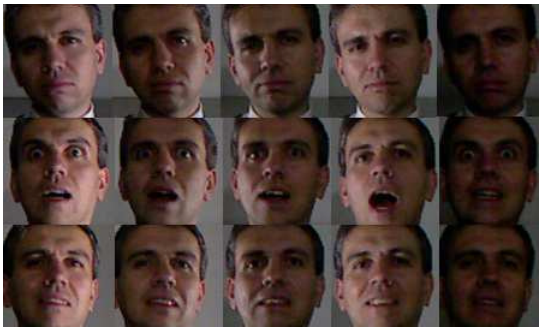


Figure 8: Examples from CurtinFaces of 15 probes images for the same person under simultaneous variation in lighting conditions (columns) and facial expressions (rows).

4.3.2. Experiments on CurtinFaces database under illumination condition variations

The actual experiments are conducted to evaluate the effectiveness of our proposed approach under simultaneous variation in lighting condition and facial expression. The validation protocol of CurtinFaces related to this experiment defines a probe set including 30 captures per subject covering five illumination conditions combined with six different facial expressions (see the examples in Figure 8).

Table 2 summarizes the obtained recognition rates for our approach and compare them to the state-of-the-art methods with respect to each proposed illumination condition. We achieved an average recognition rate of 99.6% which outperforms other results like those of (Li et al., 2013; Dai et al., 2015) who obtained 98.4%, 97.4% respectively. In addition, our results was similar to that of (Kaashki et al., 2018) in this challenging situation, who achieved an average rate of 99.5%. Nevertheless, we have already outperformed this method in all pose and expression situations where no results are reported on the profile pose. These results reflect the robustness of our approach against several illumination condition changes.

Table 2: Face recognition rates under simultaneous variation in lighting condition and facial expression from brighter to darker degrees.

Situations	(Li et al., 2013)	(Kaashki et al., 2018)	(Dai et al., 2015)	Proposed method		
				Our_{image}	Our_{depth}	Our
Front	98.9%	99.5%	N/A	99.3%	98.2%	**99.7**%
Back	98.6%	99.3%	N/A	99.1%	97.8%	**99.45**%
Low Ambient	97.1%	99.7%	N/A	99.2%	98.2%	**99.65**%
Average	98.4%	99.5%	97.4%	99.2%	98.1%	**99.6** %

Figure 9: Examples of eye and mouth occlusion situations under various pose and illumination variations in CurtinFaces database.

4.3.3. Experiments on CurtinFaces database with face occlusions

CurtinFaces database handles two separate occlusion situations; eyes occlusion with sunglasses and the mouth covered with a hand. These occlusion situations include additionally both pose and illumination variations. Here the original validation protocol of CurtinFaces provides a probe set containing 5 captures per subject and per occlusion situation. Figure 9 shows a few examples.

The obtained recognition rates for our approach are detailed in Table 3. For the eyes occlusion in the frontal view, we achieved a significant improvement compared to (Li et al., 2013) (+6%) with a rate of 94.2%, while achieving similar performances under pose variation. However, under illumination variations, we

Table 3: Face recognition rates obtained for our approach with eyes and hand occlusion situations.

Situations	Eyes			Mouth		
	Our$_{image}$	Our$_{depth}$	Our	Our$_{image}$	Our$_{depth}$	Our
Frontal view	100%	98.1%	**100%**	100%	96.2%	**100%**
Illumination	84.61%	83.7%	**91.35%**	94.3%	87.50%	**96.2%**
Pose	82.7%	76.95%	**85.57%**	88.5%	83.65%	**92.3%**
Average	86.92%	83.88%	**90.77**	93.1%	87.7%	**95.4%**

can observe that the face recognition rates under hand occlusion situations are higher than those obtained for eyes occlusion. Indeed, they decrease on average from 96.2% to 91.35%. Also, we can observe that pose variation combined to either eyes or mouth occlusion situations represents the most challenging combination as it drops the performances down to 85.57% and 92.3%. Moreover, our approach reached an overall average of 93.11% for all the combinations and occlusion situations, which outperforms (Dai et al., 2015) with only 88.9%.

These results prove the robustness of the proposed approach since the local face description has an obvious advantage over the global ones in occlusion situations. Although, the training set of CNN and BSIF descriptors did not include any occluded images, we reported a promising RGB-D face recognition performance.

4.3.4. Image interest points vs facial landmarks

We conducted a set of experiments to highlight the benefits of using image interest points over facial landmarks in our pipeline. We considered the CurtinFaces database as it involves the most representative challenges (*i.e.,* pose and facial expression variations and occlusion situations). The obtained results shown in Figure 4 prove that the image interest points are experimentally a better choice than facial landmarks especially in the case of head pose variations

Table 4: Comparison of the obtained recognition rates on CurtinFaces database under occlusion situations and also joint pose and facial expression variations for our pipeline considering image interest points and facial landmarks.

Situations	Image interest points			Facial landmarks		
	Our_{image}	Our_{depth}	Our	Our_{image}	Our_{depth}	Our
Pose+Expression	97.45%	96.62%	**98.25%**	83.81%	81.60%	83.90%
Eyes Occlusion	86.92%	83.88%	**90.77**	75.35%	75.21%	79.25%
Mouth Occlusion	93.1%	87.7%	**95.4%**	77.95%	80.72%	82.32%

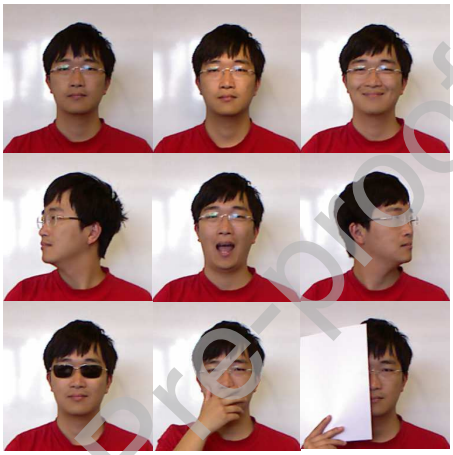


Figure 10: A sample from Eurecom database with 9 conditions going as follows: neutral, light on, smile, left profile, open mouth, right profile, eyes, hand and paper occlusions.

and occlusion situations.

4.4. Experiments on Eurecom Database with pose, illumination and occlusion conditions.

In this experiment we evaluated our approach using the Eurecom database with the proposed 9 challenging conditions as shown in Figure 10. We followed the original protocol consisting in taking the first session data-set as a gallery while the second session as the probe set. Each subject in the data-set has 9 captures in the gallery set and 9 other captures in the probe set.

The recognition performances of our RGB-D face recognition approach based on local learned features were compared to (Hsu et al., 2014) and (Sang et al., 2016) methods and the reported results are displayed in Figure 11. We can observe that our approach achieves a considerably higher performance in the

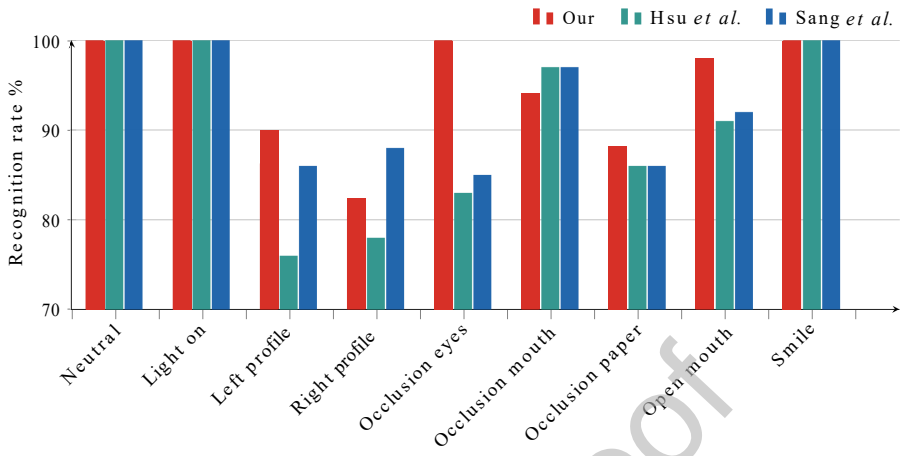


Figure 11: Comparison of the obtained recognition rates on the Eurecom database under the 9 different settings.

situations of open mouth, eyes, and paper occlusions. It is also obvious that we achieved similar results as both referenced approaches, in neutral, light-on, smile and left profile situations.

The only situations where the results obtained by (Hsu et al., 2014) and (Sang et al., 2016) are better than ours where for the the mouth occlusion situation and the right profile view. In the profile view, Sang *et al.* results are better than ours and those of Hsu *et al.* This can be explained by the frontalization and the symmetric filling pre-process applied for the non-frontal faces in order to overcome the self-occlusion caused by the large head rotation.

In the following, we compared our approach with the state-of-the-arts works based on learned features for face representation. For a fair comparison, we followed the proposed protocol in (Lee et al., 2016; Neto et al., 2019) where the profile images in the gallery set are discarded and the test set is differently selected on each experiment as bellow:

- Experiment 1: the gallery and the probe set contain seven variations from session 1 and 2, respectively: neutral, smile, illumination, paper occlusion, mouth occlusion, eyes occlusion, and open mouth.
- Experiment 2: the gallery set contains seven variations from session 1 and

the probe set contains non occluded images (neutral, smile, illumination) from session 2.

- Experiment 3: the gallery and the probe set contain only the neutral faces from both sessions.

Table 5: The obtained recognition rates on the Eurecom database in comparison with deep learning-based approaches.

Experiments	(Lee et al., 2016)			(Neto et al., 2019)	Proposed method		
	image	depth	fusion		Our_{image}	Our_{depth}	Our
1	99%	80.8%	99%	90.75%	96.1%	93.6%	**100**%
2	97%	78.8%	97.6%	98%	100%	99.35%	**100**%
3	N/A	N/A	N/A	91.1%	94.2%	90.2%	**95.8**%

Table 5 reports the obtained results in comparison with the aforementioned works. For all the experiment settings, our recognition rates with the proposed baselines outperform those achieved in Neto and in Lee works. Although they focus on generating and improving the quality of new depth images, Our_{depth} performances are better than the those of the two methods. Besides, the obtained results also proved that our designed shallow architectures are able to learn discriminant features instead of using a deep 12-layer architecture as in (Lee et al., 2016). On the other hand, while we achieved an overall recognition rate of 95.5% (9 images vs 9 images), the recorded performance in (Zhang et al., 2018) is still slightly better than ours with an average of 96,6% better. This can be explained by the importance of learning multi-modal facial data in some cases at an early stage to attain a more compact fusion strategy.

4.5. Experiment on BIWI Database

The BIWI database was originally released for the 3-D head pose estimation. Unfortunately, there is not an explicitly defined gallery/probe split for face recognition purposes. In our experiments, the gallery contains 20 persons and for each person, we randomly selected 30 captures covering a large range of

Figure 12: Example of RGB-D images from the BIWI database under different yaw angles of view.

face poses and the remaining captures as the probe set. Figure 12 shows few examples of image and depth captures for the same person under different facial poses. Following (Hsu et al., 2014; Sang et al., 2016), the testing data consist of a set of other 15 captures for each person and pose cluster (*i.e.,* L60, L45, L30, L15, frontal, R15, R30, R45, R60).

Figure 13 shows the obtained recognition rates for our approach alongside the results of (Hsu et al., 2014) and (Sang et al., 2016). Although their testing protocol is different, we still provide their results as a reference. As expected, the recognition rates for all the three approaches drop systematically when the face shifts from a frontal pose towards R60 or L60. Indeed, our approach takes advantage of a large number of captures to construct the gallery. In addition, it does not rely at all on facial landmarks detection which could be a source of recognition errors especially in the case of large facial pose variations.

4.6. Experiment on VAP Database

For these experiments, we used one face image per pose in the gallery (17 images), and the remaining face images (34 images) as probe set. We achieved an overall recognition rate of 99.71% for the image data, 99.05% for depth data and 99.80% after fusion, which outperforms the recognition rates obtained in (Zhang et al., 2018) with an average of 90.8%. From Table 6 we proved once again the complementarity between the modalities.

4.7. Experiments on a large database

So far, the performed experiments have been conducted on separate benchmarks covering different challenging situations. Now, we would like to evaluate

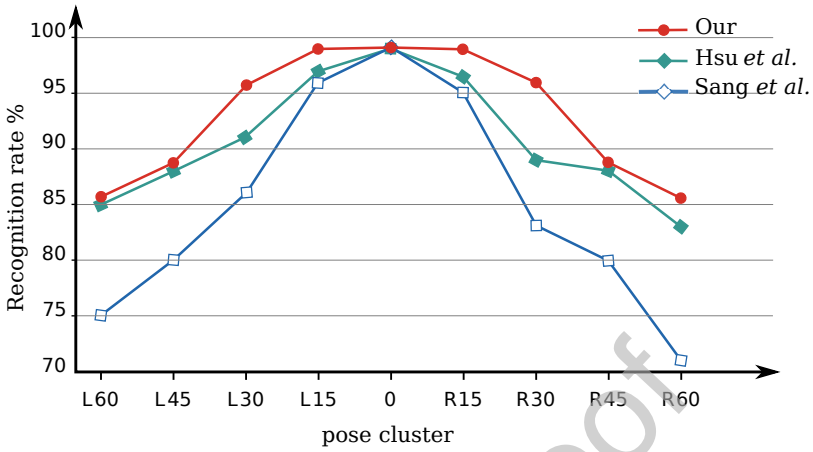


Figure 13: Obtained recognition rates on the BIWI database under head pose variation from left 60°to right 60°.Despite the difference of the testing protocol over Hsu *et al.* and Sang *et al.* methods, we still plot their results as a reference.

Table 6: Face recognition rates on VAP database under pose and facial expression variation.

Situations	(Zhang et al., 2018)	Proposed method		
		Our_{image}	Our_{depth}	Our
Pose	N/A	99.87%	99.87%	**100**%
Expression	N/A	99.2%	96.35%	**99.2** %
Average	90.8%	99.71%	99.05%	**99.80**%

our approach on a larger and more challenging benchmark by concatenating the four databases (*i.e.,* CurtinFaces, Eurecom, BIWI, and VAP) into a single one. The aim of this evaluation was twofold. The first is to access the performance of our approach in case of a larger and heterogeneous database and compare it to the results already obtained for the four benchmarks separately. The second is that we would like to investigate the effectiveness of training CNN and BSIF descriptors in case of a gradually increasing size database. In particular, we would like to check if there is a need to train or fine-tune our data-driven descriptors, especially the CNN, whenever a new subject is integrated in the database. This was achieved in the following way: all the data from the four benchmarks were concatenated and shuffled at the level of persons yielding a single database con-

sisting of 156 persons. Here, the gallery set is simply constructed by combining the gallery sets of each of the four involved benchmarks. The probe set is also constructed by combining all the probe sets from the involved benchmarks.

Table 7: Average recognition rates of our pipeline for CNN and BSIF descriptors training data corresponding to different proportion values.

Portions	Proposed method		
	Our$_{image}$	Our$_{depth}$	Our
20%	93.26% ±0.38	91.43% ±0.48	**94.30**% ± 0.27
40%	94.40% ± 0.24	93.78% ± 0.26	**95.33**% ± 0.21
60%	96.15% ±0.21	95.06% ±0.22	**96.72**% ±0.19
80%	96.15% ±0.21	95.06% ±0.23	**96.72**% ±0.20
100%	96.15% ±0.19	95.06% ±0.21	**96.72**% ±0.19

We considered five scenarios depending on the proportion of persons randomly selected for the training patches of CNN and BSIF descriptors. The defined five data proportions represent 20%, 40%, 60%, 80%, and 100% of the 156 persons in the database. For each scenario, we train our CNN and BSIF descriptors as previously detailed, then we apply our face recognition pipeline on all the probe set. As an example, the experiment related to the proportion 20% actually corresponds to the scenario of an extreme situation where we use only 20% of the gallery for training our CNN models. This is different from the other methods based on data-driven features, which require systematically an update of the network architecture and/or at least some iterations of fine-tuning (Hayat et al., 2015; Lee et al., 2016) when a new person is added to the database.

The obtained performances for the five scenarios are shown in Table 7. In addition to the average recognition rates we reported also the related standard variation values since these particular experiments involved the largest amount of data and the various training subsets (proportions) were randomly selected from the gallery at each run. The low standard deviation values (less than 1%) proved the stability of our model training procedure. In general, the obtained recognition rate (96.72%) for the proportion 100% setting is comparable to the

results obtained for the smaller benchmarks. Also, as expected the highest recognition rate is obtained for the proportion 100% setting and the lowest one corresponds to proportion 20%, which means, the drop is relatively unsignificant (*i.e.,* from 94.30% to 96.72% respectively). Moreover, we may observe that the recognition rates seems to be asymptotic starting from proportion 60%. In other words, it is not mandatory to update our data-driven descriptors when a new person is inserted or inversely removed from the gallery.

4.8. Patch based vs whole face based

To highlight the benefits of considering patches instead of the whole face for our pipeline especially in facial occlusion situations, we basically compared our pipeline to two other baselines using the whole face as input:

1. Our$_{VGG}$: is a modified version of our method with a VGG architecture[5] for CNN feature extraction representing the whole face region. Here the facial region is cropped and resized to 224 × 224 pixels. We basically used our CNN training mechanism to fine-tuned the pre-trained VGG-Face model (Parkhi et al., 2015) to extract 128 features representing the whole face before performing the SRC algorithm yielding 4 vote vectors (*i.e.,* 2 vectors for CNNs and 2 others for BSIF descriptors), and the final decision is made the same way as in our pipeline. The BSIF descriptor is also trained following our method except of using face cropped images instead of local patches.
2. FaceNet: is a reference method for RGB face recognition (Schroff et al., 2015) with its original implementation[6] taking as input the RGB face cropped images and respecting the original training protocol. This method takes as input facial crops resized to 96 × 96 pixels and outputs a 128 dimensional face representation fed to classification with linear SVM.

We can observe from the obtained recognition rates summarized in Table

[5]Only the last two layers were modified to output 128 features instead of 4096.
[6]https://github.com/davidsandberg/facenet hash:096ed77

Table 8: Comparison of recognition rates between our patch-based approach and a state-of-the-art whole-face based method.

Variations	FaceNet	Our$_{VGG}$	Our
Pose+Expression	95.7%	88.3%	**97.1**%
Illumination	94.6%	87.1	**98.3**%
Occlusion	85.2%	70.2%	**92.1**%
Average	94.5%	86.2%	**96.72**%

8 that our method provides a promising results and outperforms the FaceNet and Our$_{VGG}$ baselines on different face challenges including pose, expression, illumination variations and occlusion. These whole-face-based methods have difficulty in handling occlusion situations. They achieved 85.2% and 70.2% respectively for Facenet and Our$_{VGG}$ while our method achieved 92.1%. In general, considering the patches in our pipeline allows us to implicitly handle partial occlusions since the intermediate classification results of most of the patches located on visible facial regions will converge towards a same identity, meanwhile other patches located on occluded parts will almost be scattered on different identities. At the end, the majority vote rule outputs the final decision with the maximum of votes.

4.9. Run-time complexity

We evaluated our pipeline on an Intel Core i7 3 GHZ CPU with 16GB RAM and an 512GB SSD. We used a python binding of the SPAMS library (Mairal, 2014) for the sparse representation classification. We computed the average run-time for the main processing blocks in our pipeline on CurtinFaces database to identify one query RGB-D face image from 52 persons in the gallery (936 images). Also, we highlighted the benefit of the Dynamic Dictionary selection module in the run-time efficiency by carrying out an ablation study for this module. Although run-times strictly depended on the implementation details and the used machine or platform, we compared our method to the methods that already shared their run-time performances (Li et al., 2013; Sang et al., 2016).

Table 9: The average running time (in seconds) of our proposed approach in comparison with the state-of-the-art approaches based on the CurtinFaces database.

Steps	SRC with DD*	SRC without DD	Li et al. (2013)	Sang et al. (2016)
Face Cropping	0.2380	0.2380	0.061	0.127
Specific pre-processing	-	-	4.933	8.426
Points detection	0.0042	0.0042	-	-
Patch extraction	0.0092	0.0092	-	-
Feature extraction	0.2655	0.2655	-	-
Dynamic dictionary	1.2464	-	-	-
Image classification	0.0148	17.9931	0.084	0.057
Depth classification	0.0156	17.5947	0.026	0.036
Fusion	0.0002	0.0002	0.017	0.024
Total	**3.0403**	36.1049	5.114	8.748

* DD:Dynamic Dictionary.

From Table 9, we can see that our proposed method provides an acceptable running time, around 3 seconds to recognize a single RGB-D query image compared to 5 seconds and 9 seconds respectively for (Li et al., 2013) and (Sang et al., 2016). In fact, the time consuming parts in their method seem to lay in the pre-processing like symmetry filling, face registration, face rendering, *etc.* We can also notice the advantage of the Dynamic Dictionary selection in accelerating the matching process (*i.e.,* $\approx 1,26$ second instead of ≈ 18 seconds).

4.10. General observations

Several important observations can be made while considering the presented experiments and tests. The first is related to the usefulness of learning local data-driven representations from image and depth data to recognize RGB-D face recognition. Regarding fusion, we have not only proved the complementarity between image and depth modalities, but also between deep learning and statistical features (*i.e.,* CNN and BSIF respectively) for facial patch representation.

Besides, we have demonstrated the robustness of the proposed pipeline under different expression, illumination and pose variations. Meanwhile, compared to

the frontal and half-profile poses, the extreme profile pose is still a challenging condition for our pipeline whose results are outperformed by such solutions as (Hsu et al., 2014; Sang et al., 2016) including a pre-processing to handle face pose variations. We observed that if the gallery includes a large range of poses, as in the BIWI experiments, our pipeline allows handling implicitly the facial pose variations even if the facial pose related to the query is not exactly represented in the gallery. In this case, the gallery recording protocol, especially in a controlled environment, is very important to reduce the effect of facial pose variations. Otherwise, and as a perspective, our pipeline could be extended with a facial modelization as a pre-processing step for all faces in the gallery so that they could be rendered in any point of view as in (Hsu et al., 2014). On another side, these methods can also apply our data-driven representation even for the whole face instead of local patches.

Furthermore, the reported results on occlusion situations have proved the robustness of the proposed approach especially with local face description which is an obvious advantage over global face description in case of occlusion situations. This is also proved by comparing our approach to two other global state-of-the-art methods.

5. Conclusion

This research study detailed a new RGB-D face recognition approach based on two data-driven representations for face patches description. The proposed pipeline does not require any prior on the face pose or rely on a semantic analysis of the face. The RGB-D patches are simply extracted around some detected image interest points. Each of these patches is transformed using a selected learned descriptors, namely CNN and BSIF, before feeding the classification part. For the CNN, we propose an effective training approach leading to a discriminant embedding space for face patches representation. The learned local features fulfill the intra-class compactness and inter-class separability constraints with applying a triplet-loss as learning metric. Finally, the correspondence between patches is performed based on an SRC algorithm with a dynamic patch dic-

tionary selection. The final classification decision is obtained by a score-level fusion scheme. The experimental results performed on well-known RGB-D face benchmark databases support our claim of the effectiveness and robustness of the proposed approach. Also, we obtained competitive results relative to the state-of-the-art. As a future perspective, we would like to extend our pipeline with learning a multi-modal representation to combine both image and depth data within a more elaborated CNN architecture to leverage the complementarity between both modalities. Also, we would like to explore dynamic patch weighting to give more attention to representative and discriminant patches. Finally, we would like to extend our pipeline with the 3D facial modelization of all faces in the gallery to handle facial pose variations with less constraints on the gallery recording.

References

Abbad, A., Abbad, K., & Tairi, H. (2018). 3D face recognition: Multi-scale strategy based on geometric and local descriptors. *Computers & Electrical Engineering*, *70*, 525–537. doi:10.1016/j.compeleceng.2017.08.017.

Bay, H., Ess, A., & Tuytelaars (2008). Speeded-up robust features. *Computer vision and image understanding*, *110*, 346–359. doi:"10.1016/j.cviu.2007.09.014".

Bay, H., Tuytelaars, T., & Van Gool, L. (2006). Surf: Speeded up robust features. In *European conference on computer vision* (pp. 404–417). Graz, Austria. doi:10.1007/11744023_32.

Boutellaa, E., Hadid, A., Bengherabi, M., & Ait-Aoudia, S. (2015). On the use of kinect depth data for identity, gender and ethnicity classification from facial images. *Pattern Recognition Letters*, *68*, 270–277. doi:10.1016/j.patrec.2015.06.027.

Ciaccio, C., Wen, L., & Guo, G. (2013). Face recognition robust to head pose changes based on the RGB-D sensor. In *IEEE Sixth International Conference*

on Biometrics: Theory, Applications and Systems (pp. 1–6). Arlington, VA, USA. doi:10.1109/BTAS.2013.6712718.

Dai, X., Yin, S., Ouyang, P., Liu, L., & Wei, S. (2015). A multi-modal 2D+3D face recognition method with a novel local feature descriptor. In *IEEE Winter Conference on Applications of Computer Vision* (pp. 657–662). Waikoloa, HI, USA. doi:10.1109/WACV.2015.93.

Deng, J., Guo, J., Xue, N., & Zafeiriou, S. (2018). Arcface: Additive angular margin loss for deep face recognition. *arXiv preprint arXiv:1801.07698*, .

Fanelli, G., Gall, J., & Van Gool, L. (2011). Real time head pose estimation with random regression forests. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 617–624). Colorado Springs, CO, USA. doi:10.1109/CVPR.2011.5995458.

Goswami, G., Vatsa, M., & Singh, R. (2014). RGB-D face recognition with texture and attribute features. *IEEE Transactions on Information Forensics and Security*, *9*, 1629–1640. doi:10.1109/TIFS.2014.2343913.

Grati, N., Ben-Hamadou, A., & Hammami, M. (2016). A scalable patch-based approach for RGB-D face recognition. In *International Conference on Neural Information Processing* (pp. 286–293). Kyoto, Japan. doi:10.1007/978-3-319-46681-1_35.

Hayat, M., Bennamoun, M., & An, S. (2014). Learning non-linear reconstruction models for image set classification. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1907–1914). Columbus, OH, USA.

Hayat, M., Bennamoun, M., & An, S. (2015). Deep reconstruction models for image set classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *37*, 713–727.

Hayat, M., Bennamoun, M., & El-Sallam, A. A. (2016). An RGB-D based image set classification for robust face recognition from kinect data. *Neurocomputing*, *171*, 889–900. doi:10.1016/j.neucom.2015.07.027.

Hg, R., Jasek, P., Rofidal, C., Nasrollahi, K., Moeslund, T. B., & Tranchet, G. (2012). An rgb-d database using microsoft's kinect for windows for face detection. In *2012 Eighth International Conference on Signal Image Technology and Internet Based Systems* (pp. 42–46). Sorrento, Italy.

Hsu, G.-S. J., Liu, Y.-L., Peng, H.-C., & Wu, P.-X. (2014). RGB-D-based face reconstruction and recognition. *IEEE Transactions on Information Forensics and Security*, *9*, 2110–2118. doi:10.1109/TIFS.2014.2361028.

Hyvärinen, A., Hurri, J., & Hoyer, P. O. (2009). *Natural image statistics: A probabilistic approach to early computational vision.* volume 39. Springer Science & Business Media. doi:10.1007/978-1-84882-491-1.

Ioffe, S., & Szegedy, C. (2015). Batch normalization: accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning* (pp. 448–456). Lille, France.

Kaashki, N. N., & Safabakhsh, R. (2018). RGB-D face recognition under various conditions via 3D constrained local model. *Journal of Visual Communication and Image Representation*, *52*, 66–85. doi:10.1016/j.jvcir.2018.02.003.

Kannala, J., & Rahtu, E. (2012). BSIF: Binarized statistical image features. In *Proceedings of the 21st International Conference on Pattern Recognition* (pp. 1363–1366). Providence, RI, USA.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems* (pp. 1097–1105). Lake Tahoe, Nevada, USA. doi:10.1145/3065386.

Lee, Y.-C., Chen, J., Tseng, C. W., & Lai, S.-H. (2016). Accurate and robust face recognition from RGB-D images with a deep learning approach. In *Proceedings of the British Machine Vision Conference* (pp. 123.1–123.14). York, UK. doi:10.5244/C.30.123.

Li, B. Y., Mian, A., Liu, W., & Krishna, A. (2013). Using kinect for face recognition under varying poses, expressions, illumination and disguise. In *IEEE Workshop on Applications of Computer Vision* (pp. 186–192). Tampa, FL, USA. doi:10.1109/WACV.2013.6475017.

Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., & Song, L. (2017). Sphereface: Deep hypersphere embedding for face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 212–220). Honolulu, HI, USA. doi:10.1109/CVPR.2018.00552.

Mairal, J., Bach, F., Ponce, J., & Sapiro, G. (2010). Online learning for matrix factorization and sparse coding. *The Journal of Machine Learning Research*, *11*, 19–60. doi:10.1145/1756006.1756008.

Min, R., Kose, N., & Dugelay, J.-L. (2014). Kinectfacedb: A kinect database for face recognition. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, *44*, 1534–1548. doi:10.1109/TSMC.2014.2331215.

Neto, J. B. C., Marana, A. N., Ferrari, C., Berretti, S., & Bimbo, A. D. (2019). Depth-Based Face Recognition by Learning from 3D-LBP Images. In *Eurographics Workshop on 3D Object Retrieval*. Genoa, Italy.

Parkhi, O. M., Vedaldi, A., Zisserman, A. et al. (2015). Deep face recognition. In *Proceedings of the British Machine Vision Conference* (p. 6). Swansea, UK. doi:10.5244/C.29.41.

Sang, G., Li, J., & Zhao, Q. (2016). Pose-invariant face recognition via RGB-D images. *Computational intelligence and neuroscience*, *2016*, 13. doi:10.1155/2016/3563758.

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 815–823). Boston, Massachusetts, USA. doi:10.1109/CVPR.2015.7298682.

Sun, Y., Chen, Y., Wang, X., & Tang, X. (2014). Deep learning face representation by joint identification-verification. In *Advances in neural information processing systems* (pp. 1988–1996). Montreal, Canada.

Szegedy, C., Toshev, A., & Erhan, D. (2013). Deep neural networks for object detection. In *Proceedings of the 26th International Conference on Neural Information Processing Systems* (pp. 2553–2561). Lake Tahoe, Nevada, USA.

Taigman, Y., Yang, M., Ranzato, M., & Wolf, L. (2014). Deepface: Closing the gap to human-level performance in face verification. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1701–1708). Columbus, OH, USA. doi:10.1109/CVPR.2014.246.

Tan, X., Chen, S., Zhou, Z.-H., & Zhang, F. (2006). Face recognition from a single image per person: A survey. *Pattern recognition*, *39*, 1725–1745. doi:10.1016/j.patcog.2006.03.013.

Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., & Liu, W. (2018). Cosface: Large margin cosine loss for deep face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 5265–5274). Salt Lake City, UT, USA.

Wen, Y., Zhang, K., Li, Z., & Qiao, Y. (2016). A discriminative feature learning approach for deep face recognition. In *European conference on computer vision* (pp. 499–515). Amsterdam, The Netherlands. doi:10.1007/978-3-319-46478-7_31.

Zhang, H., Han, H., Cui, J., Shan, S., & Chen, X. (2018). RGB-D face recognition via deep complementary and common feature learning. In *13th IEEE International Conference on Automatic Face & Gesture Recognition* (pp. 8–15). Xi'an, China. doi:10.1109/FG.2018.00012.

Zhu, X., & Ramanan, D. (2012). Face detection, pose estimation, and landmark localization in the wild. In *IEEE Conference on Computer Vision and Pattern*

Recognition (pp. 2879–2886). Providence, RI, USA. doi:10.1109/CVPR.2012.6248014.

Mairal, J. (2014). SPAMS a sparse modeling software, v2. 5. URL: http://spams-devel.gforge.inria.fr.

Grati, N., Ben-Hamadou, A., & Hammami, M.(2020). Comparative study on RGB-D local representations fusion schemes for face recognition. *arXiv preprint arXiv (to appear)*, .

Declaration of interests

☒The authors declare that they have no known competing financialinterestsor personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: