Discriminant Patch Representation for RGB-D Face Recognition using Convolutional Neural Networks

Nesrine Grati[1], Achraf Ben-Hamadou[2] and Mohamed Hammami[1]
[1]*MIRACL-FS, Sfax University, Road Sokra Km 3 BP 802, 3018 Sfax, Tunisia*
[2]*Digital Research Center of Sfax, Technopark of Sfax, PO Box 275, 3021 Sfax, Tunisia*

Keywords: RGB-D Face Recognition, Convolution Neural Networks, Data-Driven Representation, Discriminant Representation, Consumer RGB-D Sensors.

Abstract: This paper focuses on designing data-driven models to learn a discriminant representation space for face recognition using RGB-D data. Unlike hand-crafted representations, learned models can extract and organize the discriminant information from the data, and can automatically adapt to build new compute vision applications faster. We proposed an effective way to train Convolutional Neural Networks to learn face patch discriminant features. The proposed solution was tested and validated on state-of-the-art RGB-D datasets and showed competitive and promising results relatively to standard hand-crafted feature extractors.

1 INTRODUCTION

With the increase demand of robust security systems in real-life applications, several automated biometrics systems for person identity recognition are developed, where the most user-friendly and non-invasive modality is the face. Face recognition using 2D images was well treated but still affected by imaging conditions. Thanks to the 3D technology progress, the recent research has shifted from 2D to 3D (Bowyer et al., 2006; Abbad et al., 2018). Indeed, three-dimensional face representation ensures a reliable surface shape description and add geometric shape information to the face appearance. Most recently, some researchers used image and depth data capture from low-cost RGB-D sensors like MS Kinect or Asus Xtion instead of bulky and expensive 3D scanners. In addition to color images, RGB-D sensors provide depth maps describing the scene 3D shape by active vision. With the availability of cost-effective RGB-D sensors, many researchers proposed and adapted feature extraction operators to the raw data for different computer vision applications like gait analysis (Wu et al., 2012), lips movement analysis (Rekik et al., 2016; Rekik et al., 2015b; Rekik et al., 2015a), and gender recognition (Huynh et al., 2012). Hand-crafted or engineered feature extractors such as LBP, Local Phase Quantization(LPQ), HOG were mainly used to deal with RGB-D data for face recognition. The main benefits of these feature extractors is that they are relatively simple and efficient to compute. Alternatively, learned features, for example with Convolution Neural Networks (CNNs), achieve a very prominent performance in many computer vision tasks (*e.g.,* object detection (Szegedy et al., 2013), image classification (Krizhevsky et al., 2012) *etc.*). The basic idea behind is to learn data-driven models that transform the raw data to an optimal representation space leading to appropriate features without manual intervention.

In this context, this paper focuses on the feature extraction part in our face recognition pipeline. A given face is represented by a set of patches extracted from image and depth data. We propose to learn discriminant local features using data-driven representation to describe the face patches before feeding a Sparse Representation Classification (SRC) algorithm to attribute the person identity.

The rest of this paper is organized as follows. First, an overview on the most prominent RGB-D face recognition systems is given in section 2. Then, we detail our proposed system in section 3. Section 4 summarizes the performed experiments and the obtained results to validate our proposed system. Finally, we conclude this study in section 5 with some observations and perspectives for future work.

Grati, N., Ben-Hamadou, A. and Hammami, M.
Discriminant Patch Representation for RGB-D Face Recognition using Convolutional Neural Networks.
DOI: 10.5220/0007403305100516
In *Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2019)*, pages 510-516
ISBN: 978-989-758-354-4

2 RELATED WORKS

In this overview, we focus mainly on the feature extractors for face recognition using RGB-D sensors. Actually, many other aspects can be discussed like the pre-processing techniques, or the overall classification schemes. In (Li et al., 2013), the nose tip is manually detected and the facial scans are aligned with a generic face model using the Iterative Closest Point (ICP) algorithm to normalize the head orientation and generate a canonical frontal view for both image and depth data. A symmetric filling process is applied on the missing depth data specifically for the non frontal view. For image data, Discriminant Color Space (DCS) operator is used as feature extractor. Then, obtained depth frontal view and DCS features are classified separately using SRC before late fusion to get the person identity.

(Hsu et al., 2014) fits a 3D face model to the face data to reconstruct a single 3D textured face model for each person in the gallery. The approach requires to estimate the pose for any new probe to be able to apply it to all 3D textured models in the gallery. This allows to generate 2D images by plan projection and then compute the LBP descriptor on the whole projected 2D images to perform the classification using an SRC algorithm. Likewise, (Sang et al., 2016) used the depth data for pose correction based on ICP algorithm to render the gallery view as the probe one. However, contrary to (Hsu et al., 2014), the authors applied Joint Bayesian Classifier on RGB and depth HOG descriptors extracted from the both data and the final decision is made via weighted sum of their similarity scores.

From the discussion above, the most focus to preprocessing especially dealing with pose variation by aligning the query data to the gallery samples. Although this kind of sequential processing may lead to error propagation from pose estimation to the classification, it gave a very good results (Hsu et al., 2014). Alternatively, to deal with pose variation, (Ciaccio et al., 2013) used a large number of image sets in the gallery under different poses angles from a single RGB-D data. Also, the face pose is estimated via the detection and alignment of standard facial landmarks in the images (Zhu and Ramanan, 2012). Each face is then represented using a set of extracted patches centered on the detected landmarks and described by a set of LBP descriptor, co-variance of edge orientation, and pixel location and intensity derivative. The classification is then performed by computing distances between patch descriptors, inferring probabilities, and lately performing a Bayesian decision.

The following works, gave more attention to feature extraction from face RGB-D data than preprocessing and dealing with head pose variation. In (Dai et al., 2015), a single ELMDP (Enhanced Local Mixed Derivative Pattern) descriptors are extracted and Nearest Neighbor algorithm is used for the combined features with confidence weights. In (Goswami et al., 2014), a combination of HOG applied on saliency and entropy features, and geometric attributes computed from the Euclidean distances between face landmarks are used as face signature. The random forest classifier is then used for the identity classification. In (Boutellaa et al., 2015), a series of handcrafted feature extractors (*i.e.,* LBP, LPQ, and HOG) are applied respectively on texture and depth crops and finally SVM classifier is carried out for face identification. The only use of the carefully-engineered representation was with feature Binarized Statistical Image Features detector (BSIF).

In (Kaashki and Safabakhsh, 2018) three-dimensional constrained local model (CLM-Z) is applied for face-modeling and landmarks points localization. Local features HOG, LBP and 3DLBP around landmarks points are extracted then SVM classifier is used for the classification.

Indeed, (Hayat et al., 2016) proposed the first RGB-D image set classification for face recognition. For a given set of images (which can captured frames with Kinect sensor), the face regions and the head poses are firstly detected with (Fanelli et al., 2011) algorithm's than clustered into multiple subsets according to the estimated pose. A block based covariance matrix representation with LBP features is applied to model each subsets on the Riemannian manifold space and SVM classification is performed on all subsets for the both modality. The final decision is made with a majority vote fusion. The proposed technique has been evaluated on a combination of three RGB-D data sets and achieved an identification rate of 94.73% which concurrent the single image based classification accuracy's.

Observations. For the classification part, we observe that SRC is used in the most popular RGB-D face recognition systems (Li et al., 2013; Hsu et al., 2014). Indeed, after its successful application in (Wright et al., 2009) for face recognition, SRC has attracted the attention in many other computer vision tasks. Also, SRC is a good choice when the number of classes in the dataset is variable and in a constant increase, which is the case of face recognition applications.

We note that the most prominent methods (Hsu et al., 2014) (Li et al., 2013) aim principally to overcome the issues of pose variation either with pose cor-

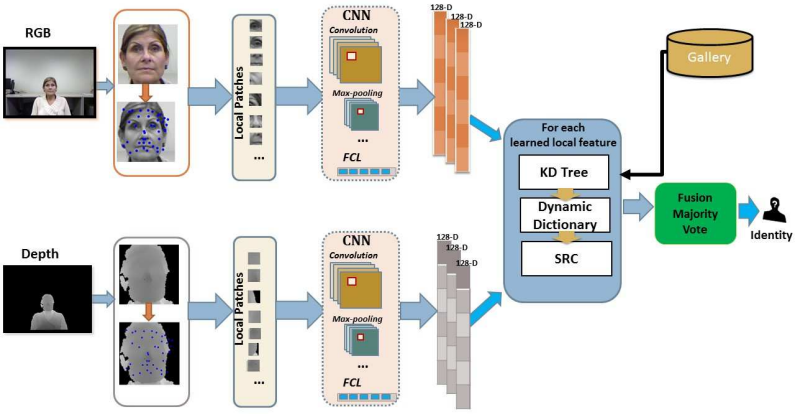

Figure 1: An overview of the online process of the proposed method. First, the facial regions are detected from image and depth data. Local patches are then extracted from both of the modalities. Afterward, two CNNs are used to transform the extracted patches to obtain feature vectors to feed the SRC algorithm and finally obtain the person ID.

rection or with augmenting the gallery through generating new images in different view or even capturing multi-view data for each face in the gallery. Data representation and appropriate feature extraction remains underestimated in the aforementioned works and they settle for hand-crafted features combinations and fusion. In this new era of deep learning techniques, we believe that RGB-D face recognition systems can take benefits of CNNs to learn appropriate features and boost their performances.

Contribution. The main objective of this paper is to highlight how CNNs can contribute to learn a discriminant representation of local regions in face image, and how it competes with standard hand-crafted feature extractors in the case of RGB-D face recognition. A given face is represented by a set of patches around detected salient key-points. Each of these patches is assigned an ID by SRC technique that associate the patch to one of the most similar patches in the database. The raw patches data (image and depth) are transformed using a CNN before feeding the classification part. We propose an effective approach to learn our CNN weights leading to a discriminant space for face patches representation.

3 PROPOSED FACE RECOGNITION SYSTEM

The proposed approach involves online and offline phases sharing some processing blocks. The offline phase is to train our CNNs while the online phase is dedicated to predict the person identity given a face query. Figure 1 sketches the online phase. It goes along the following steps. Firstly, the face is localized in the image. It is then represented by a set of patches cropped around key-points extracted on the face. Afterward, two trained CNNs are applied to transform these patches to get a feature vector for each one, and an SRC algorithm is used to attribute an ID for each feature vector before making the late fusion leading to the predicted identity. The remaining of the section gives details about the different processing blocks just introduced including the training of the CNNs.

3.1 Face Pre-processing and Patch Extraction

The face pre-processing shared between the offline and online phase of our system includes median and bilateral filtering for the depth maps and face localization (Zhu and Ramanan, 2012)[1]. The face region is cropped and resized to 96×96 pixels to ensure a normalized face spatial resolution. To get rid of face landmarks localization, we only consider salient image key-points without any further semantic analysis and without loss of generality. In other words, we do not try to catch specific anatomical reference points on the face. That said, the repeatability of image feature points for face analysis was proven. We used the SURF operator (Bay et al., 2006) to extract interest key-points on the cropped face images. The number of extracted key-points is variable and depends on face textures and also the position of the person in the frustum of the RGB-D camera. The key-points coordinates are mapped on the depth crop using the sensor calibration parameters. Around each key-point, we crop from both image and depth data two patches of 20×20 pixels. Again, the mapping

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

between image and depth map can be ensured by the RGB-D sensor geometric calibration (Ben-Hamadou et al., 2013).

3.2 CNN Architecture and Training

Since the size of the CNN input patches is small (20×20 pixels), we designed a relatively shallow CNN architecture as described in Table 1. It is worth to notice that at this level of the study we did not try complex architectures or fancy connectivity (skip connections, residual, *etc.*.) but it could be explored later in future works. We train separate models with the same architecture for each modality (*i.e.,* image and depth patches).

Research on face analysis usually focuses on finding an improvement in faithful face characterization with discriminant and robust representation. Learning descriptors with neural networks is entirely a data-driven approach. The objective of the discriminant descriptors learning is to find a transformation from raw space to an another space in which features from same classes are closer than features from different classes. Metric learning using a triplet network was introduced by Google's FaceNet (Schroff et al., 2015), where a triplet-loss is used to train an embedding space for face image using online triplet mining which outperforms a Siamese networks in manifolds clustering. Good face embedding satisfy similarity's constraint by the way faces from the same person should be close together while those of different faces are far away from each other. The intra-class distance should to be smaller than the inter-class distance and form well separated clusters.

In this paper, the triplet loss takes a triplet of patches as input in the form a, p, n , where a is the anchor patch, p is the positive patch, which is a different sample of the same class as a, and n standing for negative patch is a sample belonging to a different class. The objective of the optimization process, is to update the parameters of the network in such way that

Table 1: Our CNN architecture for small RGB-D Patches. *odim* stands for number of channels in the output tensors, similarly *idim* is the number of channels in the input tensors, and *ks* is the kernel size.

Layer	Parameters	Output tensor
Convolution	odim: 6, *ks*: 3	(6,18,18)
BatchNorm		(6,18,18)
Sigmoid		(6,18,18)
Max Pooling	*ks*: 2	(6,9,9)
Convolution	*odim*: 32, *ks*: 3	(32,7,7)
ReLU		(32,7,7)
Max Pooling	*ks*: 2	(32,3,3)
FCL	*idim*: 288, *odim*: 128	128

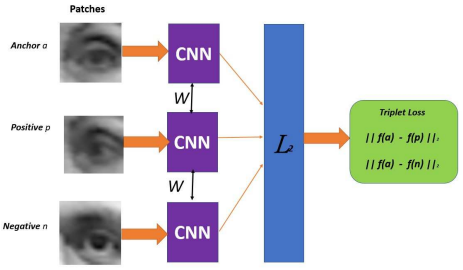

Figure 2: Local feature descriptor training pipeline with triplet loss.

patches a and p become closer in the embedded feature space, and a and n are further apart in terms of their Euclidean distances as presented in Figure 2. The triplet loss formula is given in Equation 1. $f(x)$ stands for the application of CNN on a given input x to generate a feature vector (embedding). Another hyper parameter is added to the loss equation called the margin, it defines how far away the dissimilarities should be. Minimizing L_{tr} enforces to maximize the Euclidean distance between patches from different classes which should be greater than the distance between anchor and positive features distance. For efficient training, only the triplets patches that verify the constraint $L_{tr} > 0$ are online selected as a valid triplet to improve the training.

$$L_{tr} = \sum_{a,n,p} \left(||f(a) - f(p)||_2^2 - ||f(a) - f(n)||_2^2 + margin \right) \quad (1)$$

3.3 Patch Classification

We followed (Grati et al., 2016) for the adaptive and dynamic dictionary selection in the SRC process. The objective the SRC is to reconstruct an input signal by a linear combination of atoms in a selected dictionary. We denote by $\boldsymbol{y}_k \in \mathbb{R}^M$ the input feature vector obtained from the application of the trained CNN on the k-th extracted patch and M is the its dimension. Also, we note by $\tilde{\boldsymbol{D}}_k \in \mathbb{R}^{M \times \tilde{N}}$ the dictionary. It consists of the closest $\tilde{N}$ gallery patches (atoms). Equation 2 gives the linear regression leading to the reconstructed feature vector $\tilde{\boldsymbol{y}}_k$. $\boldsymbol{x}_k \in \mathbb{R}^{\tilde{N}}$ is the sparse coefficient vector whose nonzero values are related to the atoms in $\tilde{D}_k$ used for the reconstruction of $\boldsymbol{y}_k$, ε_k captures noise, and $\tilde{N}$ is experimentally fixed.

$$\tilde{\boldsymbol{y}}_k = \tilde{\boldsymbol{D}}_k \, \boldsymbol{x}_k + \varepsilon_k \quad (2)$$

The estimation of the sparse coefficients $\boldsymbol{x}_k$ is formulated by a LASSO problem with an ℓ_1 minimization using (Mairal et al., 2010):

$$\arg min(\| \tilde{\boldsymbol{D}}_k \; x_k - \boldsymbol{y}_k \|_2 + \lambda \| \boldsymbol{x}_k \|_1) \quad (3)$$

Finally, the identity associated to the k-th patch is classically the class generating the less reconstruction error. Once the sparse representation of all local patches in the query image is obtained, a score level fusion strategy is applied then a majority vote rule predicts the final person identity.

4 EXPERIMENTS AND RESULTS

4.1 Training Details

Our CNNs has been trained with a batch size of 64, a decay value of 0.0005 a momentum value of 0.09 and an initial learning rate set to 0.001. We used PyTorch [2] framework to implement and train the CNNs. Firstly, the set of patches is classically split to training and testing sets. The pool of patch triplets needed for the CNNs training are generated and updated every ten epochs of the training by gathering the triplets from all the persons equally. A single patch triplet is obtained following these 3 steps:

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

4.2 Datasets

Our approach is validated on two publicly RGB-D face databases: Eurecom (Huynh et al., 2012), and Curtin faces (Li et al., 2013).

- Eurecom dataset is composed by 52 subjects, 14 females and 38 males. Each person has a set of 9 images in two different sessions. Each session contains 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 papersheet.
- CurtinFaces dataset consists of 52 subjects. Each subject has 97 images captured under different variations: combinations of 7 facial expression, 7 poses, 5 illuminations, and 2 occlusions. CurtinFaces database with low quality face models is more challenging in terms of variations of poses, and expression and illumination face models.

[2]https://pytorch.org/

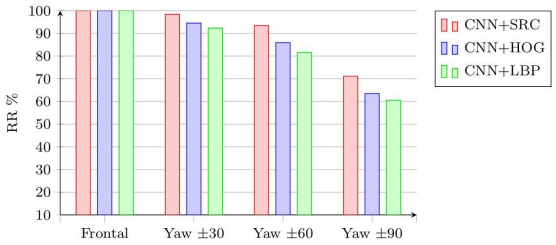

Figure 3: Performance comparison between our CNN features and the standard hand-crafted features HOG and LBP.

4.3 Validation and Results

We performed two sets of experiments to evaluate our approach. The first set is dedicated to compare with state-of-the-art systems, and the second set is to evaluate the proposed learned features comparing to hand-crafted features (*i.e.,* HOG and LBP) taking our face recognition system as baseline.

For the first set of experiments, and on CurtinFaces dataset, we report our results as well as those of (Hsu et al., 2014) (Li et al., 2013), (Ciaccio et al., 2013), (Kaashki and Safabakhsh, 2018), and (Grati et al., 2016). To make a fair comparison, we use the same protocol as (Li et al., 2013). 18 images containing only one kind of variations in illumination, pose or expression are selected for the training and testing images include the rest of non-occluded faces, there are a total of 6 different yaw poses with 6 different expressions added to the neutral frontal views. The reported results of our system on RGB, depth and fusion scheme in comparison with (Li et al., 2013) are summarized in table 2.

Table 2: Face recognition performance under yaw pose and facial expression variations.

Pose	Modality	Our Work	DSC+SRC
Frontal	RGB	100%	100%
	Depth	100%	100%
	Fusion	100%	100%
Yaw ± 30	RGB	99.03 %	99.8%
	Depth	94.55%	88.3 %
	Fusion	98.40%	99.4 %
Yaw ± 60	RGB	93.75 %	97.4%
	Depth	86.05%	87.0%
	Fusion	93.43%	98.2%
Yaw ± 90	RGB	70.2%	83.7%
	Depth	63.45%	74%
	Fusion	71.15%	84.6 %
Average	RGB	94.6%	96.70%
	Depth	88.67%	86.65%
	Fusion	94.23%	96.98%

The presented results demonstrates that our method can works equally to the aforementioned work

Table 3: CurtinFaces Database performances on differently approach.

Pose	Our	Cov+LBP	LBP+SRC	DCS+SRC	BSIF+SRC	HLF+SVM
Frontal	100%	N/A	100%	100%	100%	100%
Yaw $\pm$30	98.40%	94.2%	99.4 %	99.8%	99.04%	90.3%
Yaw $\pm$60	93.43%	84.6%	98.2 %	97.4	95.51%	58.6%
Yaw $\pm$90	71.15%	75.0%	93.5%	83.7%	60.58%	N/A

which is based on expensive pre-processing stage to frontalize and to fill symmetrically the self-occluded part in the face due to head rotation. An overall of 94.6%, 88.67% and 94.23% are achieved with our system respectively for RGB, depth and multimodal data. It's clear that our RGB performance need to be improved but our depth performance seems better than this reported in (Li et al., 2013) which demonstrate the importance and the effectiveness of data representation with learning discriminant features to overcome challenging conditions. In other hand we present in table 3 our obtained results in comparison with the most performing state of the art techniques namely(Ciaccio et al., 2013; Hsu et al., 2014) and some recent works (Grati et al., 2016; Kaashki and Safabakhsh, 2018).

It is shown that our system outperforms (Ciaccio et al., 2013) (Cov+LBP) with a margin of 4 % in Yaw $\pm$30 while a gain of $\approx$ 9% in Yaw $\pm$60. The drop in the performance for the set Yaw $\pm$90 angles can be explained by the fact that CurtinFaces database contains only just two samples for left and right $\pm$90. That is, if we take one sample for testing, no corresponding pose exists anymore in the gallery. In contrary and as reported previously in the related work section, (Ciaccio et al., 2013; Hsu et al., 2014; Li et al., 2013) pre-processings allow to tackle this issue as they either augment the gallery by generating new poses, correct the pose or symmetrically filling the self occluded part in the face. Beyond these well engineered pre-processings, in this work we aim to focus on estimating optimal RGB-D data representation for face recognition applications.

In other hand, our study is compared also to (Kaashki and Safabakhsh, 2018) (labeled as HLF+SVM in the Table 3) as it uses patch representation around landmarks points. We can observe that our performance are better with a gain of 8% in Yaw $\pm$30 angle and an improvement of more than 30% in Yaw $\pm$60 angle. These results highlights the added-value of learned features to derive more discriminant representations for local features compared to the standard hand-crafted features (*i.e.,* HOG, LBP, and 3DLBP) and prove clearly the use of salient points without interpreting face structure.

On Eurecom database, the first session set is selected for training and the second one for test. The learned feature descriptor yield to 90,82% for texture images and 85,57% for depth data, and 92,70% after fusion, which is better than the recognition rate obtained in RISE (Goswami et al., 2014) (*i.e.,* 89.0% after fusion).

The second set of experiments are dedicated to compare our CNN learned features to hand-crafted features taking our system as baseline. This is to highlight the added-value of CNN learned features. Three versions of our system are tested, the only changed part is the feature extraction: HOG+SRC, LBP+SRC, and CNN+SRC. As shown in Figure 3, CNN+SRC outperforms the other systems for all the test subsets.

Based on all the obtained results and comparisons, we can conclude that CNN learned-features can achieve a competitive identification performance for person recognition from low-cost sensor and under challenging pose and expression variations and without any prior face analysis (*e.g.,* face pose estimation, facial landmarks detection, *etc.*).

5 CONCLUSION

In this paper, we proposed a data-driven representation for RGB-D face recognition. A given face is represented by a set of patches around detected salients key-points on which a CNNs transformation are applied to extract the learned local descriptor for each modality separately before performing SRC classification. The experimental results obtained on benchmark RGB-D databases highlight the added-value of deep learning local features compared to standard hand-crafted feature extractors. For future works, we plan to extend our system with learning a multimodal representation to combine texture and depth data. With an appropriate CNN, the fusion strategy of RGB-D data will take in consideration the complementarity between depth and image data to enhance the recognition performance.

REFERENCES

Abbad, A., Abbad, K., and Tairi, H. (2018). 3d face recognition: Multi-scale strategy based on geometric and

local descriptors. *Computers & Electrical Engineering*, 70:525–537.

Bay, H., Tuytelaars, T., and Van Gool, L. (2006). Surf: Speeded up robust features. In *Computer vision–ECCV 2006*, pages 404–417. Springer.

Ben-Hamadou, A., Soussen, C., Daul, C., Blondel, W., and Wolf, D. (2013). Flexible calibration of structured-light systems projecting point patterns. *Computer Vision and Image Understanding*, 117(10):1468–1481.

Boutellaa, E., Hadid, A., Bengherabi, M., and 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.

Bowyer, K. W., Chang, K., and Flynn, P. (2006). A survey of approaches and challenges in 3d and multi-modal 3d+ 2d face recognition. *Computer vision and image understanding*, 101(1):1–15.

Ciaccio, C., Wen, L., and Guo, G. (2013). Face recognition robust to head pose changes based on the rgb-d sensor. In *Biometrics: Theory, Applications and Systems (BTAS), 2013 IEEE Sixth International Conference on*, pages 1–6. IEEE.

Dai, X., Yin, S., Ouyang, P., Liu, L., and Wei, S. (2015). A multi-modal 2d+ 3d face recognition method with a novel local feature descriptor. In *Applications of Computer Vision (WACV), 2015 IEEE Winter Conference on*, pages 657–662. IEEE.

Fanelli, G., Gall, J., and Van Gool, L. (2011). Real time head pose estimation with random regression forests. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 617–624. IEEE.

Goswami, G., Vatsa, M., and Singh, R. (2014). Rgb-d face recognition with texture and attribute features. *Information Forensics and Security, IEEE Transactions on*, 9(10):1629–1640.

Grati, N., Ben-Hamadou, A., and Hammami, M. (2016). A scalable patch-based approach for rgb-d face recognition. In *International Conference on Neural Information Processing*, pages 286–293. Springer.

Hayat, M., Bennamoun, M., and El-Sallam, A. A. (2016). An rgb–d based image set classification for robust face recognition from kinect data. *Neurocomputing*, 171:889–900.

Hsu, G.-S. J., Liu, Y.-L., Peng, H.-C., and Wu, P.-X. (2014). Rgb-d-based face reconstruction and recognition. *Information Forensics and Security, IEEE Transactions on*, 9(12):2110–2118.

Huynh, T., Min, R., and Dugelay, J.-L. (2012). An efficient lbp-based descriptor for facial depth images applied to gender recognition using rgb-d face data. In *Computer vision-ACCV 2012 workshops*, pages 133–145. Springer.

Kaashki, N. N. and 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.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Li, B. Y., Mian, A., Liu, W., and Krishna, A. (2013). Using kinect for face recognition under varying poses, expressions, illumination and disguise. In *Applications of Computer Vision (WACV), 2013 IEEE Workshop on*, pages 186–192. IEEE.

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

Rekik, A., Ben-Hamadou, A., and Mahdi, W. (2015a). Human Machine Interaction via Visual Speech Spotting. In *Advanced Concepts for Intelligent Vision Systems*, number 9386 in Lecture Notes in Computer Science, pages 566–574. Springer International Publishing.

Rekik, A., Ben-Hamadou, A., and Mahdi, W. (2015b). Unified System for Visual Speech Recognition and Speaker Identification. In *Advanced Concepts for Intelligent Vision Systems*, number 9386 in Lecture Notes in Computer Science, pages 381–390. Springer International Publishing.

Rekik, A., Ben-Hamadou, A., and Mahdi, W. (2016). An adaptive approach for lip-reading using image and depth data. *Multimedia Tools and Applications*, 75(14):8609–8636.

Sang, G., Li, J., and Zhao, Q. (2016). Pose-invariant face recognition via rgb-d images. *Computational intelligence and neuroscience*, 2016:13.

Schroff, F., Kalenichenko, D., and Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823.

Szegedy, C., Toshev, A., and Erhan, D. (2013). Deep neural networks for object detection. In *Advances in neural information processing systems*, pages 2553–2561.

Wright, J., Yang, A. Y., Ganesh, A., Sastry, S. S., and Ma, Y. (2009). Robust face recognition via sparse representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(2):210–227.

Wu, D., Zhu, F., and Shao, L. (2012). One shot learning gesture recognition from rgbd images. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, pages 7–12. IEEE.

Zhu, X. and Ramanan, D. (2012). Face detection, pose estimation, and landmark localization in the wild. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2879–2886. IEEE.