Kinect-based 3D Face Recognition: Exploring the Potential of a CNN-CRC Hybrid Model

Ahmed Yassine Boumedine*

Laboratoire Signals et Images, D'épartement d'Electronique, Université des Sciences et de la Technologie d'Oran Mohamed Boudiaf, USTO-MB, Oran, Algérie <u>ahmedyassine.boumedine@univ-usto.dz</u>

Samia Bentaieb

Université Ain Temouchent Belhadj Bouchaib, Ain Temouchent, Algérie, Laboratoire Signals et Images, D'épartement d'Electronique, Université des Sciences et de la Technologie d'Oran Mohamed Boudiaf, USTO-MB, Oran, Algérie <u>samia.bentaieb@univ-temouchent.edu.dz</u>

Abdelaziz Ouamri

Laboratoire Signals et Images, D'epartement d'Electronique, Université des Sciences et de la Technologie d'Oran Mohamed Boudiaf, USTO-MB, Oran, Algérie abdelaziz.ouamri@univ-usto.dz

Abstract

A novel approach for enhancing face recognition is presented, utilizing a hybrid methodology tailored specifically for the affordable and lower-quality Microsoft Kinect sensor. Hybrid approaches, data augmentation and varying levels of fusions are examined in this research. The authors perform the computation of surface normals, which are then partitioned into three distinct maps, denoted as N_x , N_y , and Nz. An RGB representation is then formed by combining the three normal maps. Furthermore, from each probe face, the authors generate three different scans (frontal, 30° yaw, 60° yaw). Each of these scans is an entry to one of the three parallel hybrid systems, each system is a combination of a CNN feature extraction stage and a Collaborative Representation Classifier (CRC). For the final decision, the authors combine the scores from parallel networks. The authors achieved impressive rank one Identification Rates of 97.04% and 97.33% on the CurtinFaces and IKFDB databases, respectively.

Keywords: Classification, Collaborative Representation Classifier, Normal Maps, Point Cloud, Feature Extraction, Deep learning.

1. Introduction

Face Recognition (FR), an established biometric technology, enables the recognition and verification of individuals through their facial characteristics. The broad scope of FR applications includes access control (airports and allowed areas), human-computer interaction, law enforcement, and monitoring in public places (such as stadiums and supermarkets). Soft biometric features like gender, ethnicity, and age are critical to human recognition since they may be found on the face of a person. Changes in lighting, poses, occlusions, and even cosmetics and hairstyles can throw off 2D FR systems because they solely use intensity data. 2Dbased FR systems have made great progress over the last 50 years, but they still fall short of achieving a particular level of precision.

Although the two-dimensional image representation, the face is actually threedimensional, and its key benefits include its ability to withstand changes in lighting and pose as well as its resistance to spoofing. The development of 3D sensors in the last two decades has opened new boundaries for FR systems thanks to technical developments. Segmenting the face's background has never been easier than it is now, especially with the introduction of 3D information. In order to effectively handle the vast array of facial expressions and poses, additional representations and descriptors are employed. Due to a variety of factors including expense and accessibility, 3D face images are currently difficult to get. 3D data collecting systems are rapidly improving in quality and dropping in price, making it possible to capture 3D facial images in real time. The utilized sensor offers a commendable trade-off between capturing RGB-D data and its affordability. Leveraging the 3D facial data obtained from such sensors presents a promising opportunity to train deep learning-based methods capable of effectively handling challenges arising from variations in head orientation, illumination, and occlusions. Figure 1 provides a side-by-side comparison of facial scans taken with the Minolta VIVID 910 scanner (first row), and the Kinect (second and third).



Figure 1. First row:3D scans from FRGC dataset[1], CASIA-3D FaceV1[2] and UMB-DB[3] respectively acquired with Minolta, second row: 3D faces from CurtinFaces database [4], and third row: 3D faces from IKFDB database [5]

Despite the crucial role it plays in 3D face representation and description, the significance of surface normals has not been extensively investigated. To address this research gap, the following contributions are made:

- The exploration of how normal maps can be used as a modality for feature extraction in face recognition systems when high-quality depth sensors are not available.

- When fusing the channels of the normal maps, it is important to investigate the influence that various fusion orders have. Analyze how the accuracy of the suggested method is impacted by the various orders that are considered.

- Evaluating several combinations of signal-level fusion components to determine the optimal combination that yields the highest level of precision.

In this article, Section 2 provides a comprehensive overview of the pertinent literature. Section 3 elaborates on the proposed method. The experimental procedures and the resulting outcomes are consolidated in Section 4. Finally, a conclusion in Section 5, that offers insights and suggestions for future research directions.

2. Related work

Numerous approaches have been introduced to facilitate the recognition of faces in 3D by leveraging Kinect data. The vast majority of them rely on hand-crafted and traditional feature descriptors. The authors in [4], performed their recognition task using a Sparse Representation Classifier (SRC) using 3D point cloud facial symmetry. In their study, Goswami et al.[6] proposed an automated face recognition system by incorporating the entropy of RGBD faces and a saliency feature derived from a 2D face. The researchers trained a Random Decision Forest (RDF) classifier using the extracted descriptors to facilitate the recognition process. In the work by Pamplona et al.[7], a continuous 3D face authentication system is proposed, leveraging an RGB-D camera. The system incorporates the use of the Iterative Closest Point (ICP) algorithm to detect and normalize the face images, segmenting them into distinct left and right halves, as well as the nasal area. Extracting Histogram of Oriented Gradients (HoG) features and matching them with the right ROIs is essential. While analyzing four 40-minutes videos with a wide range of facial emotions, occlusions, and positions, an EER of 0.8% is reached. To overcome the constraints caused by position and lighting changes in the rendering, color information is eliminated. Cardia Neto et al.[8] introduced an automated face recognition system using Kinect depth data. Their approach incorporates features extracted from a 3D Local Binary Pattern (LBP) and the Histogram of Averaged Oriented Gradients (HAOG), which are subsequently employed to train a Support Vector Machine (SVM) classifier. Ouloul et al.[9] developed an automated face recognition methodology that leverages the Scale Invariant Feature Transform (SIFT) descriptors, saliency map, and Local Ternary Patterns (LTP) derived from both RGB and depth maps. They utilized these extracted features to train an SVM classifier. The study conducted by Kaashki et al.[10] employed feature extraction techniques such as Local LBP, ThreeDimensional LBP (3DLBP), and HOG descriptors. These extracted features were utilized to train an SVM classifier with the objective of improving face recognition. In a previous study [11], an automated approach to facial recognition was introduced. The method involved employing the Curvelet transform on both RGB images and depth maps to extract distinctive features, followed by utilizing SVM for accurate classification. In their study, Bentaieb et al.[12] conducted an analysis using a single image from each gallery. They proposed and evaluated a 3D face recognition approach that employed the Speeded-Up Robust Feature (SURF) algorithm, applied specifically to the depth representation of the shape index map. To initiate the identification of interest points and their corresponding descriptors, the initial step involves preprocessing the 3D scans, and the second is applying SURF to the shape index map. Each 3D face scan has a unique representation based on a set of defining characteristics, and those characteristics are utilized to create a dictionary. In the recognition stage, the dictionary only contains sparse representations of the probe face scan's descriptors. The paper by Mousavi et al.[5] introduces a novel color-depth face database, specifically gathered from individuals of diverse age groups and genders in Iran. The primary aim of this work is to establish suitable database for evaluating and benchmarking existing techniques in the domains of facial identification, facial expression recognition, and age estimation. In [13], the process of identifying faces involved the preprocessing of 3D images utilizing the HOG descriptor and the Collaborative Representation Classifier (CRC). These techniques were employed to achieve successful face identification. In [14], the authors showcased an improved K-Nearest Neighbors (KNN) classifier for 3D face recognition by using the SURF features extracted from the computed shape index maps.

The significant advancements in deep neural networks have led to notable breakthroughs in face recognition. In a recent study conducted by Cui et al. [15], the authors carried out a comprehensive assessment of four fusion strategies for RGB-D face recognition. These strategies encompassed signal-level fusion, featurelevel fusion, score-level fusion, and hybrid fusion, all of which have contributed to the significant improvement made in face recognition. The first method uses a ResNet network [16] to deal with RGBD images. Moreover, the second approach entails training two separate ResNet networks, where one is trained using RGB modality and the other using depth modality. In the third approach, the RGB and depth features are merged through a fully connected layer and subjected to a unified loss function. Finally, the ultimate strategy combines fusion at the feature level and fusion at the score level. Zhang et al.[17] integrated modality activation with crossmodality matching. For each modality, Inception-v2 was utilized to train a separate feature network to guarantee complimentary feature learning. In order to effectively quantify the level of model uncertainty, Zafar et al.[18] utilized a Bayesian Deep Convolutional Neural Network (B-DCNN) in their facial recognition algorithm, as described in their work. Lin et al.[19] combined color and depth map feature extraction using two convolutional neural networks. In [20], a novel method for recognizing faces in RGB-D images is suggested; it is called Gabor-DCNN. To enhance the extraction of prominent features, the authors train a deep convolutional neural network that incorporates the Gabor transform. This transform allows them to extract significant characteristics from images, considering their diverse sizes and orientations. Subsequently, they evaluate the effectiveness of the proposed approach by testing it on the EURECOM dataset and comparing its performance to that of several state-ofthe-art techniques. In the work by Grati et al. [21], a highlevel fusion technique is proposed for face recognition, incorporating features from both image and depth modalities. The authors utilize the SURF detector to identify key points and extract CNN-based features and Binarized Statistical Image Features (BSIF) from the surrounding patches. This approach enables effective integration of information from multiple modalities in the recognition process. Uppal et al. [22] proposed a face recognition system based on a pretrained VGG. By generating an attention map from both RGB and depth images, the neural network is directed to analyze the most prominent features of the RGB image, enhancing the accuracy of the recognition process.

There are numerous face recognition techniques available for processing 3D data captured by high-resolution and expensive 3D scanners such as the Minolta equipment. In a study conducted by Kim et al.[23], a Deep Convolutional Neural Network (DCNN) was constructed using transfer learning from a CNN trained on 2D face photos, while also incorporating augmentation techniques for 3D face databases. To overcome the challenge of limited training and testing data in the 3D domain, Gilani et al.[24] introduced a method to generate millions of distinct 3D facial images representing unique individuals. These synthetic images were utilized to train the proposed CNN architecture from scratch. Face recognition is a complex problem, and the authors in [25] provide a list of the most commonly used databases, as well as a discussion of face recognition issues like dealing with variations in expressions, pose and illumination. Bhople et al. [26] introduce a 3D face recognition system that leverages PointNet, a deep learningbased architecture. This system processes the acquired point cloud data and subsequently employs a siamese network [27] for the classification task. In their study, Neto et al.[28] introduced a hybrid 3D face recognition method that combines the strengths of both Streamed Attention Network (SAN) and traditional techniques. The proposed approach effectively fuses CNN features with conventional methods to enhance the accuracy and robustness of the recognition process.

The method introduced in this research paper outperforms conventional approaches and deep neural networkbased methods in 3D face identification. This superiority is evident from the results obtained on the highly demanding CurtinFaces and IKFDB RGB-D benchmark databases. It is important to note that previous studies on 3D face identification using low-quality sensors lacked the inclusion of surface normal components in their analysis.

3. Proposed approach

The subsequent description outlines the procedure employed for face recognition utilizing the Microsoft Kinect sensor, characterized by its affordability and lower quality. The authors propose to use the facial normals information rather than the original range images to highlight local variations of facial surfaces.

In the process of facial recognition, the initial preprocessing of the raw 3D scans includes the extraction of the face region and noise reduction. Following this, the facial surfaces undergo rotation around the y-axis. Subsequently, the calculation of surface normals takes place, which are then partitioned into three distinct maps denoted as N_x , N_y , and N_z . Ultimately, the three normal maps are combined to generate an RGB image.

To accomplish face recognition based on a fusion strategy, the authors construct a CNN model comprising convolutional, pooling, and fully connected layers. Through training on annotated data, this CNN model effectively extracts valuable features from the RGB normal maps of rotated faces. The extracted features are afterwards inputted into a Collaborative Representation Classifier (CRC), which utilizes a collaborative representation approach to identify the most similar match. The overall process is illustrated in Figure 2. More details about the main stages of the process are in the following subsections.



Figure 2. Diagram of the proposed approach

3.1. Yaw face rotation

Training deep learning network with a dataset enriched by rotated original probe faces is motivated by uncontrolled conditions of the image acquisition. Feeding the network with rotated samples has the primary effect of increasing classification accuracy. With this aim, two additional scans based on y-axis rotation (yaw rotation) which corresponds to a left/right movement of the head are considered. In this methodology, a 3D face is represented as a point cloud comprising a collection of 3D coordinate triplets, denoted as x, y, and z. Introducing a rotation of the face around the y-axis, with an angle α , results in updated coordinates:

$$\begin{pmatrix} x^{\alpha} \\ y^{\alpha} \\ z^{\alpha} \end{pmatrix} = \begin{pmatrix} \cos\alpha & 0 & \sin\alpha \\ 0 & 1 & 0 \\ -\sin\alpha & 0 & \cos\alpha \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix}$$
(1)

In this paper α is set to +30° and +60° as demonstrated in Figure 3.



Figure 3. Example of yaw rotation from frontal to to +30° and +60°

3.2. Normal maps estimation

Surface normals, being a fundamental measure of surface differentials, play a crucial role in defining the orientation of a surface at every point, thereby offering significant insights into the local geometry of the surface. It find their applications in several fields, such as surface defects detection [29], 3D face recognition [30], [31], 3D facial expression classification [32]. To compute normal vectors at a specific point, the authors use the method developed by Hoppe et al [33]. Given a 3D point cloud $P = \{p_1, p_2, ..., p_N\}$ and a query point $p_i \in P$, a defined neighborhood formed by the *K* nearest neighbors denoted by p_{i1} , p_{i2} , ..., p_{iK} is considered. A plan *T* is then determined by performing a least squares fit to find the bestfitting plane that minimizes the distance between the plane *T* and the neighborhood points as denoted by the equation 2.

$$min\sum_{i=1}^{K} dist(P_i, T)$$
(2)

Where *K* is set to 10. The coefficients obtained from the fitted plane serve as the representation of the normal vector associated with the point p_i . A normalization is performed to this vector to obtain the unit normal.

In this study, the authors split the normal components N_x , N_y , and N_z into three maps, as illustrated in Figure 4. To generate depth images, the authors orthogonally project these components onto a 224×224 plane. The resulting projection is then normalized within the range of 0–255 and visualized as an RGB image.



Figure 4. Examples of depth, x, y, z, xyz and yzx maps

3.3. Convolutional Neural Network

Deep convolutional neural networks (CNNs) have attained remarkable levels of performance in diverse computer vision applications [34], [35], [36], [37], establishing them as the current state-of-the-art. However, these methods exhibit a substantial number of adjustable parameters, necessitating abundant data and significant computational resources for achieving effective generalization. The initial convolution layers extract low-level characteristics from input images through the application of filters or kernels, while higher-level layers learn more intricate features.

The proposed Convolutional Neural Network (CNN) architecture is composed of 4 alternating convolutional layers, alongside pooling, dropout, and fully connected layers, as illustrated in Figure 2. The convolutional layers adopt a 3×3 kernel size and stride of 1 followed by ReLU activation, while the max pooling layers utilize a 2×2 pooling window with a stride of 2. To prevent overfitting, dropout layers with varying dropout rates are incorporated. The model also includes a fully connected layer with 256 nodes. At the top layer, there is a fully connected layer comprising C nodes, representing individual subjects or classes, followed by a softmax activation layer. During the training process, batches of size 32 are used, and the model is trained using categorical cross-entropy loss and the Adam optimizer with a learning rate of 0.1 for 100 epochs. The number of parameters representing the weights of the kernels of each convolution and of the classification layer, as well as the associated biases of one branch are detailed in Table 1.

The total number of parameters is 3.78M showing its relatively small and can be trained quickly. Once the 3D faces have been described by their respective features, an appropriate classifier is needed for the identification/authentication of the 3D faces.

	Layers	Number of	Kernel Stride		Trainable
		neurons	size		parameters
Branch A	Input	224 x 224 x 3	-	-	0
	Convolution	222 x 222 x 64	3 x 3	1	1792
	Max pooling	111 x 111 x 64	2 x 2	2	0
	Convolution	109 x 109 x 64	3 x 3	1	36928
	Max pooling	54 x 54 x 64	2 x 2	2	0
	Droupout	54 x 54 x 64	-	-	
	25%				
	Convolution	52 x 52 x 32	3 x 3	1	18464
	Max pooling	26 x 26 x 32	2 x 2	2	0
	Convolution	24 x 24 x 32	3 x 3	1	9248
	Max pooling	12 x 12 x 32	2 x 2	2	0
	Droupout	12 x 12 x 32	-	-	
	25%				
	Flatten	4608	-	-	0
	Dense	256	-	-	1179904
	Droupout	256	-	-	
	50%				
	Dense	52	-	-	13364
Total	-	-	-	-	$1259700 \times 3 =$
number of					3779100
parameter					
S					

TABLE 1. A detailed structure of one branch of the proposed model

3.4. Collaborative Representation Classifier

The problem of face recognition has been effectively addressed through the successful application of the collaborative sparse classification method. It takes advantage of the fact that similar facial features belong to the same subspace, allowing for efficient reconstruction of their representations through sparse linear combination with other training data. Furthermore, CNN characteristics of input images from various image classes can serve as useful discriminators. Hence, this proposition revolves around the adoption of a cohesive framework that harnesses the CNN characteristics to learn a class dictionary. In order to enhance classification accuracy, CRC will use the data produced by the fully connected layer of each branch as input. For the training the deep CNN feature vector of size 4096 is obtained for each branch. A gallery dictionary is constructed as follows. Suppose that the gallery contains *C* classes and for the c_{th} class, *S* samples are used. Sub-dictionnaries denoted D_{c1} , D_{c2} , ... D_{c5} , are concatenated to build the sub-dictionary of the class *c*:

$$Dc = [D_{c1}, D_{c2}, ... D_{cS}] \in \mathbb{R}^{4096 \times S}$$
(3)

The sub-dictionaries for each of the C subjects are uniformly constructed and then combined to form the frontal pose gallery dictionary.

$$DF = [D_1, D_2, ... D_C] \in R^{4096 \times S \times C}$$
(4)

The gallery dictionaries of the non-frontal poses denoted $D+30^{\circ}$ and $D+60^{\circ}$ are constructed in the same way. Let Y_F be the feature vector derived from the initial CNN feature extraction stage, representing a probe face; $Y_F \in R^{4096}$. Y_F can be expresses as a linear combination of the feature vectors present in the gallery dictionary D_F , as mentioned in [38].

$$Y_F = D_F X_{0F} \tag{5}$$

To address the non-uniqueness of the solution to Equation 5, an additional L2-norm sparsity term is introduced to X_{0F} , to enhance the interpretability of the coefficient vector for identification purposes. Consequently, the modified equation becomes:

$$X_{0F} = \arg\min_{X} \{ \|Y_F - D_F X\|_2^2 + \lambda \|X\|_2^2 \}$$
(6)

The equation above (Eq. 6) can be solved using a closedform solution:

$$X_{0F} = (D_F^T D_F + \lambda I)^{-1} D_F^T Y_F$$
(7)

 A_F can be computed as $A_F = (D_F^T D_F + \lambda I)^{-1} D_F^T$, D_F is denoted as the dictionary for frontal gallery images, λ as a scalar weight, and I as the identity matrix. Importantly, it should be emphasized that A_F is unrelated to Y_F and can be precomputed solely based on the frontal gallery dictionary. After obtaining A_F , the residuals can be calculated as follows:

$$Res_{cF} = \|Y_F - D_F \delta_{cF}(X_{0F})\|_2$$
(8)

 $\delta_{cF}(X_{0F})$ represents entries that are almost zero except for the ones associated with person *c*. The residuals of the other parallel networks denoted Res_{c+30} ° and Res_{c+60} ° are computed in the same way. The authors determine the identity of the person depicted in the probe face by adding up the residuals generated by the three parallel networks:

$$Res_{cT}(Y_F, Y_{+30^{\circ}}, Y_{+60^{\circ}}) = Res_{cF} + Res_{c+30^{\circ}} + Res_{c+60^{\circ}}$$
(9)

Then selecting the person with the lowest residual value as the final decision:

identity(Y) = $argmin_{c} \{Res_{cT}(Y_{F}, Y_{+30^{\circ}}, Y_{+60^{\circ}})\}$ (10)

4. Experimental results and analysis

All this study employs the CurtinFaces [4] and Iranian Kinect Face Database (IKFDB) [5] public databases to evaluate the efficacy of the proposed method across various modes. Two primary indicators are utilized to assess the effectiveness of the method in different scenarios. In the identification mode, Identification Rate (IR) is measured using the ratio of tested faces that were correctly identified to the total number of tested faces. To demonstrate the measured accuracy performance of a proposed system, the authors performed identification task comparison of the rank 1, 5, 10, etc. against the Identification Rate and plotted a Cumulative Match Characteristic (CMC) curve. Additionally, for the verification mode, the authors used Verification Rate (VR) at 0.1% False Acceptance Rate (VR@FAR=0.1%) associated with the Receiver Operating Characteristic (ROC) curve. In the subsequent subsections, the authors thoroughly examine the performance of various fusion levels, data augmentation techniques, as well as modern and traditional methods on a database obtained from Kinect captures. Subsequently, the authors propose an approach that is built upon the most effective systems identified through this investigation.

4.1. CurtinFaces database

This database comprises a total of 5044 scans collected from 52 unique subjects. These scans were acquired using a Kinect V2 sensor. In this paper S = 18 scans per person are selected for training, 12 of them are frontal head orientation scans with various facial expressions, while the rest 6 scans have a neutral facial expression with various head orientation (non frontal).

For testing two subsets are used, the first one with various head orientation and facial expressions, while the second has a variation of illumination and facial expressions. Figure 5 visually demonstrates a representative instance from the CurtinFaces database, showcasing various facial poses and expressions.



Figure 5. An example of a face from the CurtinFaces database with different poses and expressions

4.1.1. Exploring level fusion techniques for normal maps:

After obtaining the estimated normal map, various fusion methods can be explored. These approaches encompass signal level fusion, feature level fusion, and score level fusion. Signal level fusion involves combining the three components of the normal maps to produce an RGB image. On the other hand, in feature level fusion, each normal map is used as an input for one of the three parallel CNN feature extraction stages, then the three stages are fused and flowed by a CNN classification layers. In the score level fusion, each of the normal maps is an entry to one of a three parallel CNN branches, then the resulted scores of the networks are added for the final decision.

To determine the best signal-level fusion components that can be combined, the authors perform extensive experiments that involve fusing the channels of normal maps in a variety of orders such as XYZ, XZY, etc. Figure 6 reports the Identification Rates versus using different normal maps order to form the RGB image. The highest rate 84.29% is achieved using YZX order, it should be noted that using the depth map results in a low Identification Rate of 54.48%. Figure 7 reports the CMC curves of the different fusion levels of normal maps. clearly the signal level fusion has the highest performance, with a Identification Rate of 84.29%.



Figure 6. Identification Rates using different maps



Figure 7. Signal fusion vs feature fusion vs score fusion

4.1.2. Investigating different scenarios of data augmentation:

The authors construct and evaluate a variety of scenarios within the face recognition system in order to conduct an analysis of the effect that data augmentation has on a database obtained through the use of Kinect. In the first scenario, the head orientation is corrected to frontal in all 3D scans of both training and testing sets. To augment data in training set, new scans are generated (Figure 8,e) by cropping a region from the originally frontal face (Figure 8,a) in order to simulate a non frontal scan corrected to frontal (Figure 8,d). In the second scenario, the original head orientation is preserved without any correction in all 3D scans present in both the training and testing sets. To enrich the training data, additional scans are generated by rotating the head of a frontal scan (Figure 8,a) to non frontal direction (Figure 8,c) in order to simulate an original non frontal scan (Figure 8,b). The third scenario is a hybrid of both first and second, three parallel CNN's with a score level fusion are used, the first CNN is trained on only frontal scans, while the other two are trained on yaw + 30° and yaw + 60°.



Figure 8. Examples of generated scans for data augmentation

Table 2 compares the rank one Identification Rates achieved by this investigation on on both subsets on CurtinFaces database. As can be seen the results obtained using data augmentation consistently outperform those without employing this strategy highlighting its effectiveness.

Data augmentation scenarios	Head rotation	Expression	All		
		variation			
Scenario 1:					
without data augmentation	82.95%	100%	90.58%		
with data augmentation	88.30%	99.55%	93.34%		
Scenario 2:					
without data augmentation	44.80%	94.23%	66.93%		
with data augmentation	52.44%	93.33%	70.75%		
Scenario 3:					
	90.12%	100 %	94.54 %		

Figure 9 displays the corresponding Identification Rates across different ranks. Its obvious with data augmentation there is an improvement of the performance, the first scenario performs better than the second, which means correcting the head orientation to frontal yields better performance. The highest performance is achieved by third scenario which merge both the two scenarios.



Figure 9. CMC curves of different data augmentation scenarios

4.1.3. Comparing between traditional and deep Learningbased methods:

Different methods are evaluated on the Kinect captured CurtinFaces database in order to investigate their performance. The traditional ones use hand-crafted feature descriptors (HOG, SURF) and classifiers (KNN, CRC), while the modern ones use deep learning methods like CNNs and transfer learning. There is also the hybrid ones that use both traditional and modern methods. Each of these methods performance is reported in Table 3 and Figure 10. The highest performance is achieved by the hybrid method that merges both CNN features and Collaborative Representation Classifier. On the contrary, employing the VGG16 network [39] yielded the least performance.

Methods	Head rotation	Expression	All		
		variation			
Traditional <i>methods:</i>					
$HOG \rightarrow Colaborative$	79.62%	99.93%	88.71%		
$SURF \rightarrow KNN$	78.79	% 99.35%	88.14%		
Modern methods:					
VGG16 Network	31.54%	79.42%	52.98%		
CNN	82.95%	100%	90.58%		
Hybrid methods:					
CNN features \rightarrow Colaborative	90.48%	100%	94.74%		

TABLE 3. Rank one Identification Rates of some of the implemented modern and traditional methods on Curtinfaces database



Figure 10. CMC curves of traditional and modern methods

4.1.4. Detailed results of the proposed approach:

To assess the efficacy of the proposed system, the authors employ two modes of evaluation: identification mode, which involves one-to-many identification, and verification mode, which focuses on one-to-one comparison. This comprehensive approach allows us to thoroughly evaluate the effectiveness of the system in different scenarios. In Figure 11 a, the CMC curves clearly show that the system is robust on both testing sets.

In Table 4, more details about the process performance is presented, the highest results are achieved when fusing the scores from the three parallel systems, with a rank one Identification Rate of 97.04% on all the testing set. Figure 11.b presents the Verification Rate versus to the False Acceptance Rate. The system achieves a Verification Rate of approximately 70% at a False Acceptance Rate of 1%. Table 5 presents a comparison of Identification Rates achieved by this proposed method and other existing methods on the CurtinFaces dataset.

Subsets	trained on	trained on	trained on	Score level	
	generated	generated	generated	fusion	
	frontal scans	30°yaw scans	60°yaw scans		
		Subset 1:			
Frontal	100%	100%	100%	100%	
30°yaw	95.67%	93.26%	95.67%	98.23%	
60° yaw	85.73%	86.53%	84.77%	93.58%	
60° pitch	89.26%	84.45%	83.17%	91.66%	
All	90.48%	88.40%	88.20%	94.64%	
Subset 2:					
Low	100%	99.67%	99.67%	100%	
Front	100%	99.03%	99.51%	100%	
Back	100%	98.23%	99.51%	100%	
All	100%	98.84%	99.55%	100%	
Both subsets:					
	94.74%	93.08%	93.28%	97.04%	

TABLE 4. Detailed results of the proposed approach on curtinfaces database



Figure 11. CMC curves (a) and ROC curves (b) of the proposed approache on Curtinfaces database

This proposed methodology demonstrates superior performance compared to the majority of existing methods, especially when confronted with variations in facial expressions (100%). It is important to note that all the techniques listed in Table 5 employ the evaluation method initially introduced by Li et al. [4].

Approach	Head rotation	Expression
		variation
Li et al.[4]	86.02%	92.8%
Kaashki et al.[9]	69.7%	96.1%
Boumedine et al.[13]	83.00%	99.87%
Grati et al.[20]	97.47%	98.1%
Proposed approach	94.64%	100%

TABLE 5. Rank one Identification Rates of some of the implemented modern and traditional methods on Curtinfaces database

4.2. IKFDB database

To further validate the results obtained by the proposed approach, an additional database is used. The Iranian Kinect face database (IKFDB) [5] encompasses a population of 40 individuals, featuring an extensive assortment of over 100,000 recorded color and depth frames. Specifically, within frames 150-250, seven primary facial expressions have been meticulously captured. Furthermore, to address the challenge of recognition from various angles, the database incorporates variations in pitch and yaw movements. In this paper 18 scans per person are selected for training (S=18), 6 scans with frontal head orientation, 6 right 45°yaw scans, and 6 left 45°yaw scans. For testing, 50 scans per person with various pitch and yaw head orientation are selected.

In Figure 12, both the CMC and ROC curves clearly show that the system is robust with a Verification Rate of 73.84% for a 1% False Acceptance Rate. In Table 6, more details about the process performance is presented, the highest results are achieved when fusing the scores.

Subsets	trained on generated frontal scans	trained on generated 30°yaw scans	trained on generated 60°yaw scans	Score level fusion
Testing set	82.51%	90.15%	85.94%	97.33%

TABLE 6. Detailed results of the proposed approach on IKFDB database



Figure 12. CMC curves (a) and ROC curves (b) of the proposed approach on IKFDB database

5. Conclusion

In this research, the authors have devised a 3D face identification system that harnesses the potential of low-quality 3D data captured by the Kinect sensor. This study addresses various aspects overlooked by previous researchers, including the evaluation of normal components' efficacy and the assessment of verification mode using depth sensor data of inferior quality. To explore the potential of data augmentation and signallevel fusion derived from normal maps, the authors have undertaken a thorough investigation. This proposed methodology involves the computation of three distinct normal maps, denoted as N_x , N_y , and N_z , from meticulously preprocessed 3D data, which are subsequently fused to generate RGB images. the authors have investigated the sequence of components that results in the highest accuracy, and it was found to be YZX. The hybrid approach the authors developed combines a modern Convolutional Neural Network (CNN) for feature extraction with the traditional Collaborative Representation Classifier (CRC). This implemented method exhibits exceptional performance, achieving the highest accuracy among the evaluated techniques. By exclusively utilizing depth information for both training and testing, this proposed method achieves an impressive accuracy of 97.04% on the CurtinFaces database and 97.33% on the IKFDB database. In the future, the authors plan to integrate RGB data from the Kinect sensor and examine the potential of deep learning architectures created explicitly for point cloud 3D data. Furthermore, the authors will assess the method's effectiveness using additional publicly accessible databases.

References

[1] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek, "Overview of the face recognition grand challenge," in 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05), vol. 1. IEEE, 2005, pp. 947–954.

[2] "CASIA-3D FaceV1." database collected by the Chinese Academy of Sciences' Institute of Automation (CASIA). <u>http://biometrics.idealtest.org</u>.

[3] A. Colombo, C. Cusano, and R. Schettini, "UMB-DB: A database of partially occluded 3D faces," in 2011 IEEE international conference on computer vision workshops (ICCV workshops). IEEE, 2011, pp. 2113–2119.

[4] B. Y. Li, A. S. Mian, W. Liu, and A. Krishna, "Using kinect for face recognition under varying poses, expressions, illumination and disguise," in 2013 IEEE workshop on applications of computer vision (WACV). IEEE, 2013, pp. 186–192.

[5] S. M. H. Mousavi and S. Y. Mirinezhad, "Iranian kinect face database (IKFDB): a color-depth based face database collected by kinect v. 2 sensor," SN Applied Sciences, vol. 3, no. 1, p. 19, 2021.

[6] G. Goswami, S. Bharadwaj, M. Vatsa, and R. Singh, "On RGBD face recognition using Kinect," in 2013 IEEE sixth international conference on biometrics: Theory, applications and systems (BTAS). IEEE, 2013, pp. 1–6.

[7] M. Pamplona Segundo, S. Sarkar, D. Goldgof, L. Silva, and O. Bellon, "Continuous 3D face authentication using RGB-D cameras," in Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2013, pp. 64–69.

[8] J. B. Cardia Neto and A. N. Marana, "3DLBP and HAOG fusion for face recognition utilizing Kinect as a 3D scanner," in Proceedings of the 30th annual ACM symposium on applied computing, 2015, pp. 66–73.

[9] M. Ouloul, Z. Moutakki, K. Afdel, and A. Amghar, "An efficient face recognition using SIFT descriptor in RGB-D images," International Journal of Electrical and Computer Engineering, vol. 5, no. 6, 2015.

[10] N. N. Kaashki and R. Safabakhsh, "RGB-D face recognition under various conditions via 3D constrained local model," Journal of Visual Communication and Image Representation, vol. 52, pp. 66– 85, 2018.

[11] S. Mohammadi and O. Gervei, "Using nonlocal filtering and feature extraction approaches in three-dimensional face recognition by Kinect," International Journal of Advanced Robotic Systems, vol. 15, no. 4, p. 1729881418787743, 2018.

[12] S. Bentaieb, A. Ouamri, A. Nait-Ali, and M. Keche, "Face recognition from unconstrained three-dimensional face images using multitask sparse representation," Journal of Electronic Imaging, vol. 27, no. 1, pp. 013 008–013 008, 2018.

[13] A. Y. Boumedine, S. Bentaieb, A. Ouamri, and A. Mallek, "3D Face Identification Using HOG Features and Collaborative Representation," in Advances in Computing Systems and Applications: Proceedings of the 4th Conference on Computing Systems and Applications. Springer, 2021, pp. 3–13.

[14] A. Y. Boumedine, S. Bentaieb, and A. Ouamri, "An Improved KNN Classifier for 3D Face Recognition Based on SURF Descriptors," Journal of Applied Security Research, pp. 1–19, 2022.

[15] J. Cui, H. Han, S. Shan, and X. Chen, "RGB-D face recognition: A comparative study of representative fusion schemes," in Biometric Recognition: 13th Chinese Conference, CCBR 2018, Urumqi, China, August 11-12, 2018, Proceedings 13. Springer, 2018, pp. 358–366.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[17] H. Zhang, H. Han, J. Cui, S. Shan, and X. Chen, "RGB-D face recognition via deep complementary and common feature learning," in 2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018). IEEE, 2018, pp. 8–15.

[18] U. Zafar, M. Ghafoor, T. Zia, G. Ahmed, A. Latif, K. R. Malik, and A. M. Sharif, "Face recognition with Bayesian convolutional networks for robust surveillance systems," EURASIP Journal on Image and Video Processing, vol. 2019, pp. 1–10, 2019.

[19] T.-Y. Lin, C.-T. Chiu, and C.-T. Tang, "RGB-D based multimodal deep learning for face identification," in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020, pp. 1668–1672.

[20] Y. Xiao and X. Xie, "Application of Novel Gabor-DCNN into RGBD Face Recognition." Int. J. Netw. Secur., vol. 22, no. 3, pp. 532–539, 2020.

[21] N. Grati, A. Ben-Hamadou, and M. Hammami, "Learning local representations for scalable RGB-D face recognition," Expert Systems with Applications, vol. 150, p. 113319, 2020.

[22] H. Uppal, A. Sepas-Moghaddam, M. Greenspan, and A. Etemad, "Depth as attention for face representation learning," IEEE Transactions on Information Forensics and Security, vol. 16, pp. 2461–2476, 2021.

[23] D. Kim, M. Hernandez, J. Choi, and G. Medioni, "Deep 3D face identification," in 2017 IEEE international joint conference on biometrics (IJCB). IEEE, 2017, pp. 133–142.

[24] S. Z. Gilani and A. Mian, "Learning from millions of 3D scans for large-scale 3D face recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1896–1905.

[25] G. Guo and N. Zhang, "A survey on deep learning based face recognition," Computer vision and image understanding, vol. 189, p. 102805, 2019.

[26] A. R. Bhople, A. M. Shrivastava, and S. Prakash, "Point cloud based deep convolutional neural network for 3D face recognition," Multimedia Tools and Applications, vol. 80, pp. 30 237–30 259, 2021.

[27] G. Koch, R. Zemel, R. Salakhutdinov et al., "Siamese neural networks for one-shot image recognition," in ICML deep learning workshop, vol. 2, no. 1. Lille, 2015.

[28] J. B. C. Neto, C. Ferrari, A. N. Marana, S. Berretti, and A. Del Bimbo, "Learning Streamed Attention Network from Descriptor Images for Cross-Resolution 3D Face Recognition," ACM Transactions on Multimedia Computing, Communications and Applications, vol. 19, no. 1s, pp. 1–20, 2023.

[29] E. T. Lee, Z. Fan, and B. Sencer, "A new approach to detect surface defects from 3D point cloud data with surface normal Gabor filter (SNGF)," Journal of Manufacturing Processes, vol. 92, pp. 196–205, 2023.

[30] V. Vijayan, K. W. Bowyer, P. J. Flynn, D. Huang, L. Chen, M. Hansen, O. Ocegueda, S. K. Shah, and I. A. Kakadiaris, "Twins 3D face recognition challenge," in 2011 international joint conference on biometrics (IJCB). IEEE, 2011, pp. 1–7.

[31] H. Li, D. Huang, J.-M. Morvan, L. Chen, and Y. Wang, "Expressionrobust 3D face recognition via weighted sparse representation of multi-scale and multi-component local normal patterns," Neurocomputing, vol. 133, pp. 179–193, 2014.

[32] H. Ujir, M. Spann, and I. H. M. Hipiny, "3D facial expression classification using 3D facial surface normals," in The 8th International Conference on Robotic, Vision, Signal Processing & Power Applications: Innovation Excellence Towards Humanistic Technology. Springer, 2014, pp. 245–253.

[33] H. Hoppe, T. DeRose, T. Duchamp, J. McDonald, and W. Stuetzle, "Surface reconstruction from unorganized points," in Proceedings of the 19th annual conference on computer graphics and interactive techniques, 1992, pp. 71–78.

[34] W. Zhou, H. Wang, and Z. Wan, "Ore image classification based on improved CNN," Computers and Electrical Engineering, vol. 99, p. 107819, 2022.

[35] I. Nouicer and H. Fizazi, "Multiple Convolution Neural Network for Supervised Image Classification," YMER International Open Access Journal vol. 22, no. 2, pp. 96-104, 2023.

[36] Y. Dong, Q. Liu, B. Du, and L. Zhang, "Weighted feature fusion of convolutional neural network and graph attention network for hyperspectral image classification," IEEE Transactions on Image Processing, vol. 31, pp. 1559–1572, 2022.
[37] S. Telsang, R. Sasne, R. Loya, R. Mandake, T. Rokade, and R. Lohe, "Unique Dog Identification Using Convolutional Neural Network," YMER International Open Access Journal vol. 21, no. 11, pp. 1870-1882, 2022.

[38] L. Zhang, Y. Shen, H. Li, and J. Lu, "3D palmprint identification using block-wise features and collaborative representation," IEEE transactions on pattern analysis and machine intelligence, vol. 37, no. 8, pp. 1730–1736, 2014.

[39] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.