information

MDPI

Article

NIRFaceNet: A Convolutional Neural Network for Near-Infrared Face Identification

Min Peng [1,2,†], Chongyang Wang [1,2,†], Tong Chen [1,2,*] and Guangyuan Liu [1,2]

1 School of Electronic and Information Engineering, Shouthwest University, Chongqing 400715, China; peng2014m@email.swu.edu.cn (M.P.); mvrjustid520@email.swu.edu.cn (C.W.); liugy@swu.edu.cn (G.L.)
2 Chongqing Key Laboratory of Nonlinear Circuit and Intelligent Information Processing, Southwest University, Chongqing 400715, China
* Correspondence: c_tong@swu.edu.cn; Tel.: +86-23-6825-4309
† These authors contribute equally to this work.

Academic Editor: Willy Susilo
Received: 16 July 2016; Accepted: 24 October 2016; Published: 27 October 2016

Abstract: Near-infrared (NIR) face recognition has attracted increasing attention because of its advantage of illumination invariance. However, traditional face recognition methods based on NIR are designed for and tested in cooperative-user applications. In this paper, we present a convolutional neural network (CNN) for NIR face recognition (specifically face identification) in non-cooperative-user applications. The proposed NIRFaceNet is modified from GoogLeNet, but has a more compact structure designed specifically for the Chinese Academy of Sciences Institute of Automation (CASIA) NIR database and can achieve higher identification rates with less training time and less processing time. The experimental results demonstrate that NIRFaceNet has an overall advantage compared to other methods in the NIR face recognition domain when image blur and noise are present. The performance suggests that the proposed NIRFaceNet method may be more suitable for non-cooperative-user applications.

Keywords: near-infrared face recognition; illumination invariance; convolutional neural network

PACS: 42.30.Sy

1. Introduction

Face recognition is one method of biometric authentication. It has attracted attention from the fields of pattern recognition and computer vision. Up to now, many methods [1–5] have been used in order to obtain higher recognition accuracy. However, most of them are concentrated on recognizing facial images in the visible spectrum, which are vulnerable to changes in environmental illumination [6–9].

Several techniques have been proposed to achieve illumination invariant face recognition, such as a 3D face scanner [10–12], hyperspectral imaging (HSI) [13–15], thermal imaging (TI) [16–18], Kinect sensors [19–22], and near-infrared (NIR) imaging techniques [23,24]. Experimental results [16–18,23,24] have shown that both TI and NIR techniques can achieve illumination invariance to some extent. TI can be used in a completely dark environment without using any active illumination. However, a TI system is more costly than a NIR imaging system, and the ambient temperature can affect recognition accuracy. For the NIR method, a NIR illumination source is needed in a dark environment. However, a NIR system costs much less.

Research on NIR face recognition has mainly focused on finding robust methods to improve recognition accuracy. Li et al. [23] established the framework of NIR face recognition and used local binary patterns (LBP) as a method. LBP can be easily calculated and is a robust method with regard to image rotation and illumination change. However, it is not robust enough with regard to sensor noise,

i.e., when there is noise in the images, the recognition rate will be low if LBP is used. Sajad et al. [25] used geometric moment (GM), Zernike moment (ZM), pseudo-Zernike moment (PZM), and wavelet moment (WM) as recognition methods, and compared the performance of the four methods on the CASIA (Chinese Academy of Sciences Institute of Automation) NIR database [23]. It was found that the best recognition performance can be achieved if ZM is employed. Using the same CASIA NIR database, Sajad et al. [26] later tested global feature extraction methods (ZM, independent component analysis, radon transform plus discrete cosine transform, radon transform plus discrete wavelet transform) and local feature extraction methods (LBP, Gabor wavelets, discrete wavelet transform, undecimated discrete wavelet transform), and found ZM and undecimated discrete wavelet transform (UDWT) can achieve the highest recognition rate among global and local feature extraction methods, respectively. To obtain better recognition performance, Sajad et al. [27,28] moved on to fuse global and local features and proposed Zernike moment undecimated discrete wavelet transform (ZMUDWT) method and the Zernike moments plus hermite kernels (ZMHK) method as the feature extraction methods for NIR face recognition.

However, the methods used in NIR face recognition so far have only been tested on the subsets of the CASIA NIR database. Moreover, all of the methods are designed for, and tested in, the cooperative-user application environment; i.e., there is no motion blur in the facial images, which is common in a non-cooperative-user application environment due to the relative motion between the object and the camera, or the focusing of the camera.

Recently, deep learning methods have been used in face recognition in the visible spectrum. The Facebook AI group presents a convolutional neural network (CNN) called DeepFace [29] for face recognition. It has eight layers and is trained on a database that contains four million facial images. In a study by Sun et al. [30], DeepID is proposed, which consists of an ensemble of small CNNs. Each small CNN has nine layers. In [31], a deep network called WebFace is proposed, which is a CNN-based network with 17 layers. All three networks have very different structures and implementation choices.

In this paper, we present a CNN called NIRFaceNet. NIRFaceNet is based on a modification of GoogLeNet [32] for NIR face recognition in non-cooperative-user applications. The experimental design focuses on one aspect of face recognition, i.e., face identification (distinguishing one face from many).

In a non-cooperative-user application, such as surveillance, the objects are in motion, and the imaging systems may be refocusing occasionally. This will lead to blur or noise in the images taken by the systems. We, therefore, added motion and Gaussian blur, salt-and-pepper, and Gaussian noise to the CASIA NIR database to simulate a non-cooperative-user application.

Experimental results show that the proposed NIRFaceNet can achieve the highest identification rate among LBP + PCA (principal component analysis), LBP Histogram, ZMUDWT, ZMHK, and GoogLeNet, and is the most robust method with regard to the added noise. NIRFaceNet is modified from GoogLeNet, but it is specifically designed for the CASIA NIR database and, thus, can achieve a 3%–5% higher identification rate with less training time (30 h < 104 h) and less processing time (0.025 s < 0.07 s). When density-0.1 salt-and-pepper noise is present, NIRFaceNet can achieve a 5.51% higher identification rate than ZMHK (96.02% > 90.51%), which has the second highest identification rate in general.

2. Convolutional Neural Networks

The structure of a CNN was first proposed by LeCun [33]. It simulates the processing system of human vision by using the local receptive field, shared weight, and subsampling. The local receptive field and shared weight can make one feature stand out in a feature map and save on the computational load. Subsampling can achieve invariance of features with regard to geometric distortion. Due to these advantages, CNN finds applications in computer vision [32–34], natural language processing [35,36], and speech recognition [37,38].

A CNN is a multi-layered non-fully-connected neural network. Figure 1 shows the general structure of a CNN. The input layer receives normalized images with identical sizes. A set of units in a small neighborhood (local receptive field) in the input layer will be processed by a convolution kernel to form a unit in a feature map (each plane in the convolutional layer in Figure 1) of the subsequent convolutional layer. One pixel in the feature map can be calculated by using:

$$C_k = f(x * W + b) \tag{1}$$

where C_k is the value of the k-th pixel in the feature map, x is the pixel-value vector of the units in the local receptive field corresponding to C_k, W and b are the coefficient vector and bias, respectively, determined by the feature map, and f is the activation function (sigmoid, tanh, ReLU, etc.). Since the results presented by Vinod et al. [39] suggest that the ReLU is superior to the sigmoid function, the ReLU function has been employed in our work. For the input t, $f(t) = \max(0, t)$ according to the definition of ReLU.

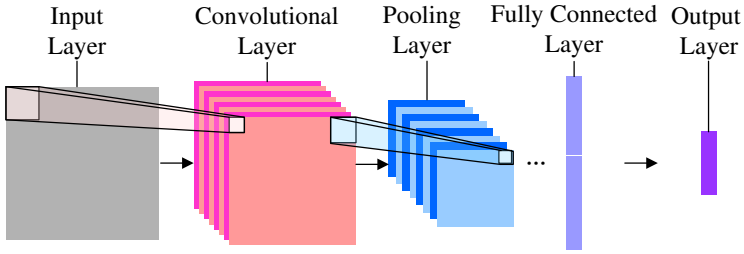

Figure 1. Structure of a convolutional neural network (CNN).

Each feature map has only one convolutional kernel, i.e., for all x in the input plane, the W and b are the same. This CNN design can largely save on calculation time and make one feature stand out in one feature map [32]. There is normally more than one feature map in a convolutional layer, so that multiple features are included in the layer.

To achieve invariance of the features with regard to geometrical shift and distortion, the convolutional layer is followed by a pooling layer to subsample the feature maps [32].

For the k-th unit in a feature map in the pooling layer, its value can be calculated by using:

$$P_k = f(\beta * down(C) + \alpha) \tag{2}$$

where P_k is the value of the k-th unit in the feature map (each plane in the pooling layer in Figure 1) in the pooling layer, C is the value vector in the feature map of the convolutional layer, β and α are the coefficient and bias, respectively, and $down(\cdot)$ is the subsampling function.

A max pooling function is used for subsampling. In that case, $down(C)$ can be written as:

$$down(C) = \max\left\{C_{s,l} \middle| |s| \leq \frac{m}{2}, |l| \leq \frac{m}{2}, s, l \in z^+\right\} \tag{3}$$

where $C_{s,l}$ is the pixel value in the unit C in the feature map, and m is the subsampling size.

The first convolutional and pooling layers extract elemental features. To obtain higher level features, more convolutional and pooling layers are often used in a CNN one after another to form a deep architecture.

Each unit in the last pooling layer will be connected as an input to a fully-connected layer that acts as a hidden layer in a normal neural network.

The fully-connected layer is followed by the output layer. The number of outputs of this layer is the number of groups to be classified. For example, if the raw data input to the CNN is expected to be divided into four groups, then there will be four outputs in this layer. The connection between the

fully-connected layer and the output layer is a softmax connection [40]. The probability of softmax regression classifying the input vector F from the previous layer into group c is:

$$p\left(y^{(F)} = c \middle| F; \theta\right) = \frac{e^{\theta_j^T F}}{\sum_{n=1}^{N} e^{\theta_n^T F}} \quad 1 \leq j \leq N \tag{4}$$

where $y^{(F)}$ is the group identity of input F, θ is the weight vector between the output layer and the previous layer, and N is the number of groups.

Finally, all coefficients, biases, and weights in the CNN are trained by Batch gradient descent [41] protocols.

3. Proposed Network Architecture

NIRFaceNet is modified from GoogLeNet [32]. GoogLeNet is a deep neural network which has 27 layers (convolution and pooling layers). It consists mostly of a CNN and won first place in the ImageNet Large Scale Visual Recognition Challenge 2014.

The success of deep neural networks, such as GoogLeNet, makes researchers believe that it is reasonable to develop deep networks trained on large datasets [42]. However, for datasets that are not large enough, a medium-sized network can achieve similar or even slightly higher recognition rates than what a large-sized network can achieve [42,43]. The CASIA NIR database used in this research contains 3940 pictures, which is much smaller than the ImageNet database. Therefore, a full-sized GoogLeNet may not perform better than a modified network with shallow structure.

We have tested the full-sized GoogLeNet on the CASIA NIR database. The results (see Section 4.3) show that the softmax0, softmax1, and softmax2 of GoogLeNet can achieve identification rates of 99.02%, 98.8%, and 98.74%, respectively, on the dataset of normal faces. Moreover, the identification rate of softmax0 is the highest, and softmax2 is the lowest, in most of the experimental conditions. Softmax0 is the classifier in the shallowest place in GoogLeNet and softmax2 is in the deepest place. This means that the deeper the GoogLeNet is, the lower the identification rate is. We will, therefore, use a shallow network by modifying GoogLeNet.

Dong et al. [44] presented a two-stage CNN for vehicle type classification. The first stage of the network is for low-level feature extraction and the second is for high-level global feature extraction. The CNN was trained on a dataset containing 9850 vehicle images and achieved good recognition results. The size of the datasets that we used in this research is the same order of magnitude as that of the vehicle dataset. We, therefore, keep only two feature extraction modules in our NIRFaceNet.

The architecture of NIRFaceNet is shown in Figure 2. It can be seen that NIRFaceNet has only eight layers and is compact in size compared to the original GoogLeNet. NIRFaceNet has only two feature extraction modules. A common structure of the feature extraction module is shown in Figure 3.

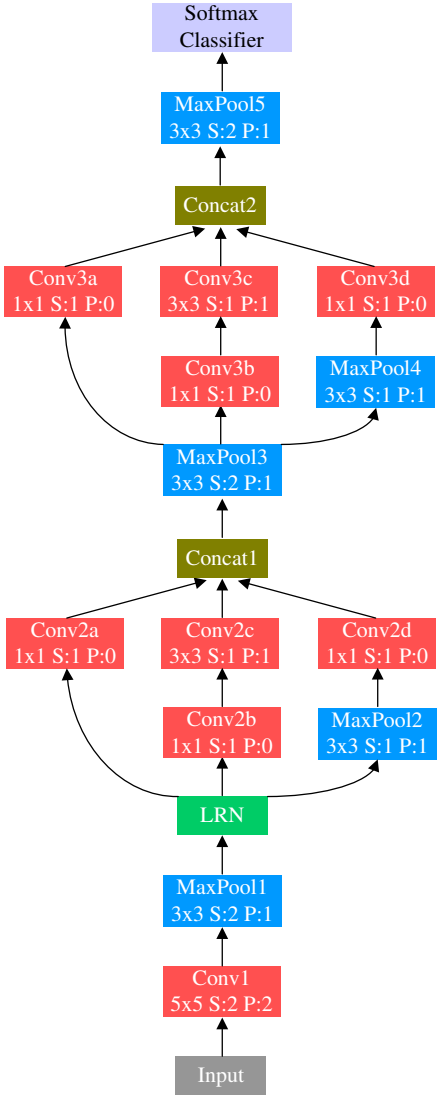

Figure 2. Proposed NIRFaceNet (LRN: local response normalization).

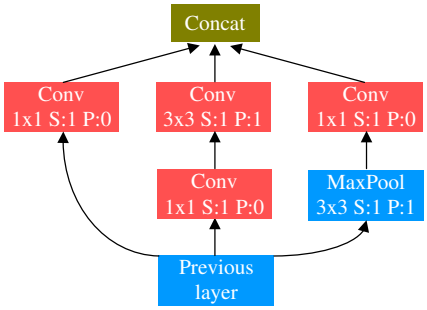

Figure 3. Common structure of a feature extraction module.

The image input to NIRFaceNet is preprocessed. To avoid diminishing small features of the image [45], the feature extraction modules leave out the 5×5 spatial filters in GoogLeNet. Since the 5×5 spatial filters also tend to consume a significant number of parameters [46]—for example, with the same number of filters, a 5×5 convolution layer needs 2.78 (25/9) times more computations than what a 3×3 convolution layer needs—the simplified modules in NIRFaceNet require less memory resources and will take less time to be trained.

In the feature extraction modules, the 1×1 convolutions play two major roles in feature extraction. Firstly, they increase the nonlinearity of the network while keeping the wealth of information from the upper layer. Secondly, the 1×1 convolutions can reduce the calculation load before we use multi-scale convolution to extract the upper features. The parallel 3×3 max pooling cell, with a one-pixel stride (S) and one-pixel padding (P), can not only maintain the resolution of the feature maps (the same resolution as that of the previous layer) but can also extract more texture details.

The output of the 3×3 convolutional filters and other related convolutional layers are stacked by the Concat [32] function to act as the input to the next layer. The local response normalization (LRN) layer [34] is inspired by a form of lateral inhibition in real neurons and can improve the generalization ability and the precision of the modules. NIRFaceNet contains no fully connected layer, which can reduce the network complexity to a great extent. The output dimensionality of each layer is shown in Table 1.

Table 1. Layers and output size.

Layers	Output Size
Input	112×112
Conv1	$64 \times 56 \times 56$
Maxpool1	$64 \times 28 \times 28$
LRN	$64 \times 28 \times 28$
Conv2a	$64 \times 28 \times 28$
Conv2b	$64 \times 28 \times 28$
Conv2c	$128 \times 28 \times 28$
Maxpool2	$64 \times 28 \times 28$
Conv2d	$64 \times 28 \times 28$
Concat1	$256 \times 28 \times 28$
Maxpool3	$256 \times 14 \times 14$
Conv3a	$128 \times 14 \times 14$
Conv3b	$128 \times 14 \times 14$
Conv3c	$192 \times 14 \times 14$
Maxpool4	$256 \times 14 \times 14$
Conv3d	$128 \times 14 \times 14$
Concat2	$448 \times 14 \times 14$
Maxpool5	$448 \times 7 \times 7$
Softmax Classifier	$197 \times 1 \times 1$

4. Experiments and Analysis

In this section, we will test NIRFaceNet, LBP + PCA [23], LBP Histogram [47], ZMUDWT [27], ZMHK [28], and GoogLeNet on the CASIA NIR database [23]. Facial expression, head pose variation, salt-and-pepper and Gaussian noise, motion and Gaussian blur are added to the dataset to compare the robustness of the algorithms. Face recognition includes face identification and verification. In this paper, we test only the algorithms in the identification case (distinguishing one face from many in the database).

4.1. CASIA NIR Database

The CASIA NIR database was established by Li et al. [23]. It contains 3940 pictures (resolution 640×480) of 197 persons with different expressions, different head poses, and with or without glasses. In this study, we tested the algorithms using 3330 pictures, including all pictures with normal faces, different expressions, and different head poses. The other 610 pictures with glasses were not considered in this research. Figure 4 shows the pictures of one person in the database in normal, expression variation, and head pose variation conditions.

(**a**) (**b**) (**c**)

Figure 4. NIR pictures of one person under normal (**a**), expression variation (**b**) and head pose variation (**c**) conditions.

4.2. Data Analysis

Before identification, we used the Viola–Jones [48] function in MATLAB 2015a to detect the face, and then normalized the facial images into 112 × 112 pixels in size and 0–255 in terms of pixel dynamic range.

We tested the algorithms on nine test sets. The training sets of all nine test sets were the same. Three pictures of the normal faces of each person were selected to form the training set. Therefore, there were 591 (197 × 3) pictures in the set. There were no overlapping pictures between the training set and the nine-test sets.

The methods to generate the nine test sets are described in Table 2. Test Set 1 is made up of pictures of normal faces (norm face) other than the ones in the training set. Test Set 2 is made up of pictures of normal faces, faces with different expressions, and faces with different head poses.

Table 2. Methods to generate testing datasets.

Test Set ID	Method to Generate
1	Exclude the training set For each person, select three pictures of normal face Exclude the person if there is less than three pictures left 459 pictures from 153 persons are selected to form the test set
2	Exclude the training set Select all the other pictures, except the pictures of persons with glasses 2739 pictures are selected to form the test set
3	Add motion blur to Test Set 2, with a length of nine pixels and an angle randomly sampled in the range of 0–360°
4	Add Gaussian blur to Test Set 2, with standard deviation of 0.5
5	Add Gaussian blur to Test Set 2, with standard deviation of 2
6	Add salt-pepper noise to Test Set 2, with density of 0.01
7	Add salt-pepper noise to Test Set 2, with density of 0.1
8	Add Gaussian noise to Test Set 2, with mean of 0 and variance of 0.001
9	Add Gaussian noise to Test Set 2, with mean of 0 and variance of 0.01

In non-cooperative-user applications, there may be blur and noise in the images taken by NIR cameras. The blur comes from the relative motion between the object and camera or the refocusing of the camera, which are common in a surveillance application. The noise is mainly salt-and-pepper noise and Gaussian noise. Therefore, Test Sets 3–9 were generated from Test Set 2 by adding different levels of noise and blur to simulate a non-cooperative-user application environment. Figure 5 shows images of one participant in Test Sets 2–9.

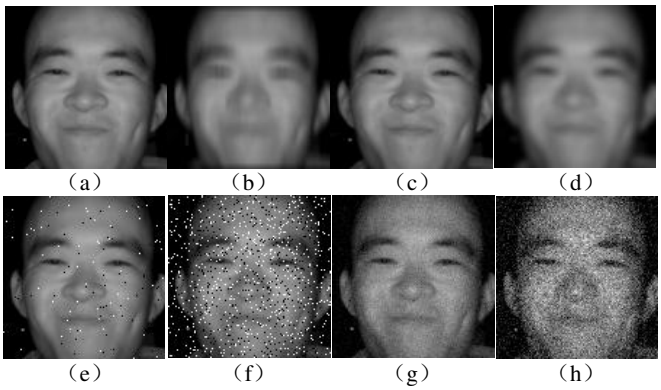

Figure 5. Images of one participant in Test Sets 2–9: (**a**)–(**h**) are the images in the datasets 2–9, respectively.

This approach to selecting the training set and developing the test sets is designed to simulate a more realistic surveillance application (in which the face is expected to be recognised when an object may be in motion or be obscured by image noise) by using a limited training set (three pictures of the normal face for each person in this research).

The NIRFaceNet model and five algorithms were tested on every test set. The parameters used for the model and algorithms are described below.

The mini-batch sizes and dropout ratios of the NIRFaceNet model were set to 35 and 0.5, respectively. The training process and the testing process were implemented by using Caffe [49].

For the LBP + PCA, the raw image was divided into 4 × 4 blocks. The LBP feature vector was extracted by using the 3 × 3 neighbourhood uniform LBP ($LBP_{8,1}^{U_2}$). PCA (principal component analysis) was used to extract the most important components, up to 100, from the feature vector. The 100 components were input into linear discriminant analysis (LDA) for the identification.

For the LBP histogram, the raw image was divided into 4 × 4 blocks. The LBP feature vector was extracted by using the 3 × 3 neighbourhood uniform LBP. The classifier was support vector machine (SVM) (using the "svmclassify" function in MATLAB 2015a with default settings).

For the ZMUDWT, n = 10 in the ZM. There were 66 moment features, each of which included imaginary and real parts, and modulus values. The raw image was divided into 12 blocks according to [27]. The DB3 wavelet was then used to perform a three-layer non-sampling discrete wavelet transform. The wavelet coefficients of low and high frequency in the third layer were used to form the feature vector. The feature fusion and classification methods in [27] were used.

For the ZMHK, the parameter settings in the ZM were the same as those in the ZMUDWT. The γ and σ in the HK were set to 13 and two, respectively. The image was divided into eight blocks, according to [28], to extract features. The feature fusion and classification methods in [28] were used.

4.3. Experimental Results Using Normal Faces

The identification rates of every method tested on Test Set 1 are shown in Table 3. It can be seen that NIRFaceNet achieves 100% accuracy when used to recognize normal faces (without expression and posture changes). The identification rate of GoogLeNet is lower than that of NIRFaceNet, and the deeper the GoogLeNet is, the lower the identification rate is. This confirms that a shallow network is better for a small-sized dataset. With respect to the identification performance of traditional algorithms, the methods fusing global and local features (95.64% for the ZMUDWT and 100% for the ZMHK) outperform LBP (89.76% and 87.34%), and the ZMHK outperforms the ZMUDWT. This result is in accordance with that in [28].

Table 3. Identification rates of various methods tested on Test Set 1.

	LBP + PCA	LBP Histogram	ZMUDWT	ZMHK	GoogLeNet softmax0	GoogLeNet softmax1	GoogLeNet softmax2	NIRFaceNet
Identification Rate (%)	89.76	87.34	95.64	100	99.02	98.8	98.74	100

4.4. Experimental Results Using Images with Facial Expressions and Head Rotations

The identification rates of every method tested on Test Set 2 are shown in Table 4. Due to the large variations in expression and posture in Test Set 2, the identification rates for this set are lower than those for Test Set 1. Nevertheless, NIRFaceNet still outperforms the other algorithms and achieves an identification rate of 98.28%. This result shows that NIRFaceNet is a robust identification method with regard to variations in expression and posture. Again, the identification rate of GoogLeNet shows that the shallow net is more suitable for the CASIA NIR database (softmax0's 95.64% > softmax1's 95.15% > softmax2's 94.73%). The LBP histogram (87.34%) outperforms LBP + PCA (80.94%) under this experimental condition.

Table 4. Identification rates of various methods tested on Test Set 2.

	LBP + PCA	LBP Histogram	ZMUDWT	ZMHK	GoogLeNet softmax0	GoogLeNet softmax1	GoogLeNet softmax2	NIRFaceNet
Identification Rate (%)	80.94	87.34	90.18	96.5	95.64	95.15	94.73	98.28

4.5. Experimental Results Using Images with Blur and Noise

The identification rates of every method tested on Test Sets 3–9 are shown in Table 5. Compared to the identification rates achieved on Test Set 2, which includes no noise or blur, the identification rates achieved on Test Sets 3–9 are generally lower due to the addition of noise and blur. LBP + PCA can only achieve rates of 30.92%, 30.27%, and 20.45% when motion blur, density-2 Gaussian blur, and density-0.1 salt-and-pepper noise were present, respectively; the identification rate drops to 0.99% and 0.66% when density-0.001 and density-0.01 Gaussian noise were present, respectively. The LBP histogram is more robust than LBP + PCA with regard to blur and noise. Except under the density-0.1 salt-pepper noise condition, the LBP histogram has 1%–20% higher identification rates than those of LBP + PCA. The performance of LBP + PCA observed in this experiment are in accordance with that in [50].

Table 5. Identification rate of algorithms tested on Test Sets 3–9 with different levels of blur and noise.

Identification Rate (%)	**Motion Blur**	**Gaussian Blur**		**Salt-Pepper Noise**		**Gaussian Noise**	
	Test Set 3 (density 9)	***Test Set 4 (density 0.5)***	***Test Set 5 (density 2)***	***Test Set 6 (density 0.01)***	***Test Set 7 (density 0.1)***	***Test Set 8 (density 0.001)***	***Test Set 9 (density 0.01)***
LBP + PCA	30.92	76.85	30.27	78.02	20.45	0.99	0.66
LBP Histogram	54.14	82.99	46.4	81.38	12.6	1.61	1.35
ZMUDWT	88.35	90.03	88.97	89.89	82.48	89.63	87.77
ZMHK	95.14	96.50	95.25	95.87	90.51	96.20	94.56
GoogLeNet softmax0	94.33	94.98	93.25	95.2	94.79	95.79	92.41
GoogLeNet softmax1	93.79	95.15	93.03	95.05	94.04	96.03	92.20
GoogLeNet softmax2	93.04	94.20	92.04	94.88	93.73	95.25	91.73
NIRFaceNet	98.12	98.48	98.24	98.32	96.02	98.36	97.48

ZMUDWT and ZMHK are more robust than LBP with regard to noise and blur. They can still achieve identification rates greater than 80%. The lowest identification rates for ZMUDWT and ZMHK were 82.48% and 90.51%, respectively, when density-0.1 salt-and-pepper noise was present. GoogLeNet has lower identification rates than those of ZMHK in most cases, the one exception being the density-0.1 salt-and-pepper noise condition.

NIRFaceNet is the most robust method. It achieves the highest identification rate on every test set, which is at least 2% more than the second highest rate. When density-0.1 salt-and-pepper noise was present, its identification rate of 96.02% was 5.51% higher than that of the best traditional method, ZMHK (90.51%).

The results in Table 5 are graphically illustrated by using the line chart shown in Figure 6.

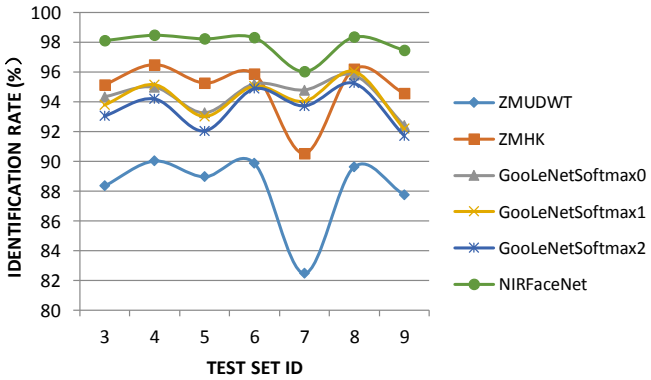

Figure 6. The identification rates of algorithms tested on Test Sets 3–9.

4.6. Traning Time and Processing Time

CNN-based methods have to be trained before they are used for identification. The training times for GoogLeNet and NIRFaceNet are listed in Table 6. Caffe [49] was used for training the networks. All of the settings for the training sessions were the same (e.g., 320,000 iteration steps). All training sessions were run on a DELL PRECISION T3600 (CPU: Xeon E5-1620 3.6 GHz, Memory: 64 GB, Graphic Card: nVIDIA Quadro 600) (Dell, Chongqing, China).

Table 6. Training times of CNN-based methods.

Method	Time (h)
GoogLeNet	104
NIRFaceNet	30

The processing times of all the methods are listed in Table 7. The processing time of each method is the average time used by the method to process one face image (i.e., to identify each face image). Since Caffe was used for training the CNN-based methods, it was also used for implementing the CNN-based methods. MATLAB 2015a was used for implementing the other methods. All of the codes were run on the DELL PRECISION T3600.

Table 7. Processing times of all the methods.

Methods	Processing Time(s)
LBP + PCA	0.078
LBP histogram	0.069
ZMUDWT	0.315
ZMHK	0.214
GoogLeNet	0.07
NIRFaceNet	0.025

It can be seen from Table 7 that the LBP-based methods need much less processing time than ZMUDWT and ZMHK, and that NIRFaceNet needs less processing time than GoogLeNet. Since the traditional methods and CNN-based methods were implemented in different languages (MATLAB and Caffe), the processing times of these two method types cannot be compared directly. However, Table 7 shows that NIRFaceNet could process an input image in real-time (0.025 s per image) if the appropriate implementation method was chosen.

5. Discussion and Conclusion

In this paper, we proposed a CNN-based method called NIRFaceNet to recognize NIR faces. The strong self-learning ability of a CNN was used to achieve robust NIR face identification in this research. We tested NIRFaceNet on the CASIA NIR database. In contrast with previous work, we included not only faces with expression and posture variations but also faces with different types of blur and noise and different intensities of blur and noise for the testing. Experimental results demonstrated that NIRFaceNet can achieve the highest identification rate among the LBP, ZMUDWT, and ZMHK methods, and is the most robust method with regard to expression and posture variation and with regard to noise and blur.

NIRFaceNet is modified from GoogLeNet. However, it is much more compact in size than GoogLeNet. Compared to the 27 layers in GoogLeNet, NIRFaceNet has only eight layers. This reduction in complexity of structure enables NIRFaceNet to be trained in much less time and to process an input image in less time. For instance, it takes 30 h to train NIRFaceNet, whereas it takes 104 h to train GoogLeNet. It takes 0.025 s for NIRFaceNet to process one image, compared with 0.07 s for GoogLeNet. Since NIRFaceNet is designed specifically for the CASIA NIR dataset, it can achieve a 3%–5% higher identification rate than GoogLeNet.

With respect to the traditional methods of NIR face identification, ZMHK can achieve the highest identification rate. Its performance is even better than GoogLeNet in most cases. However, in the case of density-0.1 salt-pepper noise, the performance of ZMHK decreases sharply. Its identification rate drops from 96.50% under the non-noise condition (Test Set 2) to 90.51% under the noise condition (Test Set 7). On the other hand, NIRFaceNet is much more robust than ZMHK in this case: the identification rate of NIRFaceNet drops from 98.28% (Test Set 2) to 96.02% (Test Set 7). The drop in identification rates of ZMHK and NIRFaceNet are 6.21% and 2.30%, respectively. This suggests that NIRFaceNet may be more suitable for recognizing faces under very noisy conditions, such as in real non-cooperative NIR face identification applications.

It can be seen from Tables 4 and 5 that the adding of density-0.5 Gaussian blur (Test Set 4), density-0.01 salt-and-pepper noise (Test Set 6), and density-0.001 Gaussian noise (Test Set 8) does not decrease the identification rates (98.48%, 98.32%, and 98.36%, respectively) of NIRFaceNet, but increases them compared to the identification rate (98.28%) under the non-noise condition (Test Set 2). In the case of GoogLeNet, the adding of density-0.001 Gaussian noise (Test Set 8) increases the identification rates of softmax0, softmax1, and softmax2, whilst density-0.5 Gaussian blur (Test Set 4) increases the identification rate of softmax1, and density-0.01 salt-and-pepper noise (Test Set 6) increases the identification rate of softmax2. These small increases in identification rates can only be observed in low-density settings of all types of noise. This may be due to the robustness of the CNN; i.e., the adding of low density noise may not affect the overall performance of the CNN, but causes identification rates to vary randomly to a small extent. In the case of NIRFaceNet, the identification rates happen to vary towards larger values. However, in the case of GoogLeNet, the identification rates vary to larger or smaller values.

As the CASIA NIR database is built incrementally, the structure of NIRFaceNet may need to be redesigned and retrained again (by updating parameters). Since NIRFaceNet was designed specifically for the CASIA NIR database, the enlargement of the database may require changes to the network; i.e., more feature extraction modules may be required. This redesign of the network will not stop until the training dataset reaches a large enough size, such as a planet-scale size as an extreme example.

Of course, building a dataset containing seven billion identities is inconceivable. Additionally, it is hard to tell where the boundary of the size lies, beyond which the network's structure can be constant.

However, the MegaFace Challenge [51] does start to investigate what happens to the performance of face recognition algorithms when the person to be recognized is mixed with up to a million distractors that were not in the training set. It was found that all algorithms had lower recognition accuracy when they were tested on the MegaFace dataset. However, the algorithms that were trained on larger sets had a higher accuracy, and FaceN, trained on 18 million images, performed similarly to FaceNet, trained on 500 million images.

The building of large datasets for training is as equally important as algorithm development. In terms of CNN-based NIR face identification, the design of NIRFaceNet may be just a starting point. Building a database including more identities could be a project for the future. Age variation is also a factor that affects identification accuracy. NIR images of each person at different ages could be included in the database.

Acknowledgments: We would like to thank for their support the National Natural Science Foundation of China (Grant No. 61301297 and No. 61472330), the Fundamental Research Funds for the Central Universities (No. XDJK2013C124), and the Southwest University Doctoral Foundation (No. SWU115093).

Author Contributions: Min Peng, Tong Chen, and Guangyuan Liu conceived and designed the experiments; Min Peng and Chongyang Wang performed the experiments; Min Peng and Chongyang Wang analyzed the data; all authors wrote the paper. All authors have read and approved the final manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Wolf, L.; Hassner, T.; Maoz, I. Face recognition in unconstrained videos with matched background similarity. In Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition, Colorado Springs, CO, USA, 20–25 June 2011.
2. Arandjelović, O.; Hammoud, R.; Cipolla, R. Thermal and reflectance based personal identification methodology under variable illumination. *Pattern Recognit.* **2010**, *43*, 1801–1813. [CrossRef]
3. Wright, J.; Yang, A.Y.; Ganesh, A.; Sastry, S.S.; Ma, Y. Robust face recognition via sparse representation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2009**, *31*, 210–227. [CrossRef] [PubMed]
4. Arandjelović, O.; Cipolla, R. Face set classification using maximally probable mutual modes. In Proceedings of the 18 International Conference on Pattern Recognition, Hong Kong, China, 20–24 August 2006.
5. Yin, Q.; Tang, X.; Sun, J. An associate-predict model for face recognition. In Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition, Colorado Springs, CO, USA, 20–25 June 2011.
6. Adini, Y.; Moses, Y.; Ullman, S. Face recognition: the problem of compensating for changes in illumination direction. *IEEE Trans. Pattern Anal. Mach. Intell.* **1997**, *19*, 721–732. [CrossRef]
7. Braje, W.L.; Kersten, D.; Tarr, M.J.; Troje, N.F. Illumination effects in face recognition. *Psychobiology* **1998**, *26*, 371–380.
8. Phillips, P.J.; Flynn, P.J.; Scruggs, T.; Bowyer, K.W.; Chang, J.; Hoffman, K.; Marques, J.; Min, J.; Worek, W. Overview of the face recognition grand challenge. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA, 20–25 June 2005.
9. Ochoa-Villegas, M.A.; Nolazco-Flores, J.A.; Barron-Cano, O.; Kakadiaris, I.A. Addressing the illumination challenge in two-dimensional face recognition: A survey. *IET Comput. Vis.* **2015**, *9*, 978–992. [CrossRef]
10. Gökberk, B.; Salah, A.A.; Akarun, L.; Etheve, R.; Riccio, D.; Dugelay, J.-L. 3D face recognition. In *Guide to Biometric Reference Systems and Performance Evaluation*; Petrovska-Delacrétaz, D., Dorizzi, B., Chollet, G., Eds.; Springer: London, UK, 2009; pp. 263–295.
11. Drira, H.; Amor, B.B.; Srivastava, A.; Daoudi, M.; Slama, R. 3D face recognition under expressions, occlusions, and pose variations. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 2270–2283. [CrossRef] [PubMed]
12. Liang, R.; Shen, W.; Li, X.-X.; Wang, H. Bayesian multi-distribution-based discriminative feature extraction for 3D face recognition. *Inf. Sci.* **2015**, *320*, 406–417. [CrossRef]
13. Shen, L.; Zheng, S. Hyperspectral face recognition using 3d Gabor wavelets. In Proceedings of the 21st International Conference on Pattern Recognition, Tsukuba, Japan, 11–15 November 2012.

14. Uzair, M.; Mahmood, A.; Mian, A. Hyperspectral face recognition with spatiospectral information fusion and PLS regression. *IEEE Trans. Image Process.* **2015**, *24*, 1127–1137. [CrossRef] [PubMed]
15. Cho, W.; Koschan, A.; Abidi, M.A. Hyperspectral face databases for facial recognition research. In *Face Recognition across the Imaging Spectrum*; Bourlai, T., Ed.; Springer: Cham, Switzerland, 2016; pp. 47–68.
16. Hermosilla, G.; Ruiz-del-Solar, J.; Verschae, R.; Correa, M. A comparative study of thermal face recognition methods in unconstrained environments. *Pattern Recognit.* **2012**, *45*, 2445–2459. [CrossRef]
17. Ghiass, R.S.; Arandjelović, O.; Bendada, A.; Maldague, X. Infrared face recognition: A comprehensive review of methodologies and databases. *Pattern Recognit.* **2014**, *47*, 2807–2824. [CrossRef]
18. Choraś, R.S. Thermal face recognition. In *Image Processing and Communications Challenges 7*; Choraś, R.S., Ed.; Springer: Cham, Switzerland, 2016; pp. 37–46.
19. Li, B.Y.L.; Mian, A.S.; Liu, W.; Krishna, A. Using kinect for face recognition under varying poses, expressions, illumination and disguise. In Proceedings of 2013 IEEE Workshop on Applications of Computer Vision, Clearwater Beach, FL, USA, 15–17 January 2013; pp. 186–192.
20. Goswami, G.; Bharadwaj, S.; Vatsa, M.; Singh, R. On RGB-D face recognition using Kinect. In Proceedings of IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems, Arlington, VA, USA, 29 September–2 October 2013.
21. Li, B.Y.L.; Mian, A.S.; Liu, W.; Krishna, A. Face recognition based on Kinect. *Pattern Anal. Appl.* **2016**, *19*, 977–987. [CrossRef]
22. Goswami, G.; Vatsa, M.; Singh, R. Face recognition with RGB-D images using Kinect. In *Face Recognition across the Imaging Spectrum*; Bourlai, T., Ed.; Springer: Cham, Switzerland, 2016; pp. 281–303.
23. Li, S.Z.; Chu, R.; Liao, S.; Zhang, L. Illumination invariant face recognition using near-infrared images. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 627–639. [CrossRef] [PubMed]
24. Li, S.Z.; Yi, D. Face recognition, near-infrared. In *Encyclopedia of Biometrics*; Li, S.Z., Jain, A., Eds.; Springer: berlin/Heidelberg, Germany, 2015.
25. Farokhi, S.; Shamsuddin, S.M.; Sheikh, U.U.; Flusser, J. Near infrared face recognition: A comparison of moment-based approaches. In *Innovation Excellence towards Humanistic Technology*; Springer: Singapore, 2014; pp. 129–135.
26. Farokhi, S.; Sheikh, U.U.; Flusser, J.; Shamsuddin, S.M.; Hashemi, H. Evaluating feature extractors and dimension reduction methods for near infrared face recognition systems. *Jurnal Teknologi* **2014**, *70*, 23–33. [CrossRef]
27. Farokhi, S.; Shamsuddin, S.M.; Sheikh, U.U.; Flusser, J.; Khansari, M.; Jafari-Khouzani, K. Near infrared face recognition by combining Zernike moments and undecimated discrete wavelet transform. *Digit. Signal Process.* **2014**, *31*, 13–27. [CrossRef]
28. Farokhi, S.; Sheikh, U.U.; Flusser, J.; Yang, B. Near infrared face recognition using Zernike moments and Hermite kernels. *Inf. Sci.* **2015**, *316*, 234–245. [CrossRef]
29. Taigman, Y.; Yang, M.; Ranzato, M.; Wolf, L. DeepFace: Closing the gap to human-level performance in face verification. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1701–1708.
30. Sun, Y.; Wang, X.; Tang, X. Deep learning face representation from predicting 10,000 Classes. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1891–1898.
31. Yi, D.; Lei, Z.; Liao, S.; Li, S.Z. Learning face representation from scratch. 2014, arXiv:1411.7923.
32. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
33. Lécun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
34. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. In Proceedings of the Twenty-sixth Annual Conference on Neural Information Processing Systems (NIPS), Stateline, NV, USA, 3–8 December 2012.
35. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.

36. Collobert, R.; Weston, J. A unified architecture for natural language processing: Deep neural networks with multitask learning. In Proceedings of the 25th international conference on Machine learning, Helsinki, Finland, 5–9 July 2008.
37. Abdel-Hamid, O.; Mohamed, A.-R.; Jiang, H.; Penn, G. Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition. In Proceedings of 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, Japan, 25–30 March 2012.
38. Graves, A.; Mohamed, A.R.; Hinton, G. Speech recognition with deep recurrent neural networks. In Proceeding of 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vancouver, BC, Canada, 26–31 May 2013.
39. Nair, V.; Hinton, G.E. Rectified linear units improve restricted Boltzmann machines. In Proceedings of the 27th International Conference on Machine Learning, Haifa, Israel, 21–24 June 2010.
40. Bouchard, G. Efficient bounds for the softmax function and applications to approximate inference in hybrid models. In Proceedings of NIPS 2007 workshop for approximate Bayesian inference in continuous/hybrid systems, Whistler, BC, Canada, 7–8 December 2007.
41. Wilson, D.R.; Martinez, T.R. The general inefficiency of batch training for gradient descent learning. *Neural Netw.* **2003**, *16*, 1429–1451. [CrossRef]
42. McDonnell, M.D.; Tissera, M.D.; Vladusich, T.; van Schaik, A.; Tapson, J. Fast, simple and accurate handwritten digit classification by training shallow neural network classifiers with the "Extreme Learning Machine" algorithm. *PLoS ONE* **2015**, *10*, e0134254. [CrossRef] [PubMed]
43. Hu, G.; Yang, Y.; Yi, D.; Kittler, J.; Christmas, W.; Li, S.Z.; Hospedales, T. When face recognition meets with deep learning: An evaluation of convolutional neural networks for face recognition. In Proceedings of the 2015 IEEE International Conference on Computer Vision Workshops, Santiago, Chile, 13–16 December 2015; pp. 142–150.
44. Dong, Z.; Wu, Y.; Pei, M.; Jia, Y. Vehicle type classification using a semi supervised convolutional neural network. *IEEE Trans. Intell. Transp. Syst.* **2015**, *16*, 2247–2256. [CrossRef]
45. Hayder, M.; Haider, A.; Naz, E. Robust Convolutional Neural Networks for Image Recognition. *Int. J. Adv. Comput. Sci. Appl.* **2015**, *6*, 105–111.
46. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. 2015, arXiv:1512.00567.
47. Ahonen, T.; Hadid, A.; Pietikinen, M. Face description with local binary patterns: Application to face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2006**, *28*, 2037–2041. [CrossRef] [PubMed]
48. Castrillón, M.; Déniz, O.; Guerra, C.; Hernández, M. ENCARA2: Real-time detection of multiple faces at different resolutions in video streams. *J. Vis. Commun. Image Represent.* **2007**, *18*, 130–140.
49. Jia, Y.; Shelhamer, E.; Donahue, J.; Karayev, S.; Long, J.; Girshick, R.; Guadarrama, S.; Darrell, T. Caffe: Convolutional architecture for fast feature embedding. In Proceedings of the 22nd ACM International Conference on Multimedia, Orlando, FL, USA, 3–7 November 2014; pp. 675–678.
50. Ren, J.; Jiang, X.; Yuan, J. Noise-resistant local binary pattern with an embedded error-correction mechanism. *IEEE Trans. Image Process.* **2013**, *22*, 4049–4060. [CrossRef] [PubMed]
51. Kemelmacher-Shlizerman, I.; Seitz, S.; Miller, D.; Brossard, E. The MegaFace benchmark: 1 million faces for recognition at scale. 2016, arXiv:1512.00596.