


Article

NIRExpNet: Three-Stream 3D Convolutional Neural Network for Near Infrared Facial Expression Recognition

Zhan Wu ^{1,2}, Tong Chen ^{1,2,*} , Ying Chen ^{1,2}, Zhihao Zhang ^{1,2} and Guangyuan Liu ^{1,2}

¹ Chongqing Key Laboratory of Nonlinear Circuit and Intelligent Information Processing, Southwest University, Chongqing 400715, China; zhan1994@email.swu.edu.cn (Z.W.); chenyingly@email.swu.edu.cn (Y.C.); zzh085517@email.swu.edu.cn (Z.Z); liugy@swu.edu.cn (G.L.)

² School of Electronic and Information Engineering, Southwest University, Chongqing 400715, China

* Correspondence: c_tong@swu.edu.cn; Tel.: +86-236-825-0394

Received: 28 September 2017; Accepted: 6 November 2017; Published: 17 November 2017

Abstract: Facial expression recognition (FER) under active near-infrared (NIR) illumination has the advantages of illumination invariance. In this paper, we propose a three-stream 3D convolutional neural network, named as NIRExpNet for NIR FER. The 3D structure of NIRExpNet makes it possible to extract automatically, not just spatial features, but also, temporal features. The design of multiple streams of the NIRExpNet enables it to fuse local and global facial expression features. To avoid over-fitting, the NIRExpNet has a moderate size to suit the Oulu-CASIA NIR facial expression database that is a medium-size database. Experimental results show that the proposed NIRExpNet outperforms some previous state-of-art methods, such as Histogram of Oriented Gradient to 3D (HOG 3D), Local binary patterns from three orthogonal planes (LBP-TOP), deep temporal appearance-geometry network (DTAGN), and adapt 3D Convolutional Neural Networks (3D CNN DAP).

Keywords: near-infrared facial expression recognition; 3D convolutional neural network; global and local features of facial expression; spatio-temporal features

1. Introduction

Facial expression as a carrier of emotion conveys rich behavior information [1]. Therefore, facial expression recognition (FER) has been a hot topic, and attracted attention in many fields, including human-computer interaction [2], security [3], and biometrics [4]. In early studies, FER methods focused on the still images, which did not consider the motion information of facial expression [5]. Since facial expression is a dynamic behavior, only employing still images is not sufficient for recognizing facial expressions. Presently, there are some traditional methods of extracting the facial expression dynamic features. For example, Histogram of Oriented Gradient to 3D (3D HOG) [6], as the extension of HOG, extracts the local temporal features.

Most current FER systems capture images/videos in the visible light spectrum (VIS) (380 nm–750 nm) [7]. However, different environments produce huge variation for FER [8]. For example, the varying illumination conditions, to some extent, weaken the performance of FER significantly. Though pre-processing algorithms to ease the influence of illumination have been adopted in some works recently [9], it is still difficult to be applied in different circumstances and obtain satisfying results.

An alternative method for solving the environment illumination variation problem is to use near-infrared (NIR) (780 nm–1100 nm) camera for facial expression recognition. The NIR camera with integrated NIR illumination sources were mounted in front of the users. The NIR light emitted from

the camera has a much higher intensity than the ambient NIR light (because ambient NIR is very weak). By using this kind of active light source, the ambient illumination variation problem could be solved, as long as the artificial illumination is constant, and the FER may be performed even in dark illumination conditions [10]. NIR system has applied to many research areas for solving illumination problem. For validating the idea of NIR FER, Zhao et al. [11] collected an NIR facial expression database, called Oulu-CASIA NIR facial expression database, and utilized improved LBP-TOP (Local binary patterns from three orthogonal planes) capturing the dynamic local information from the NIR video sequences. Farokhi, S. et al. [12] proposed a NIR face recognition method based on Zernike moments (ZMs) and Hermite kernels (HKs). Gejji, R.S. et al. [13] modeled the pupil's response to light by using a nonlinear delay differential equation in the NIR spectral band, which can be used to develop robust iris recognition algorithms. Son, C. et al. [14] captured NIR images and visible color images and presented a new NIR dehazing model based on color regularization, which can solve the color distortion and haze degradation problems at the same time.

Like FER in VIS, most NIR FER methods so far have focused on analyzing still images. Though there do exist some traditional methods extracting the facial expression temporal features [511] from NIR video sequences, these methods need to extract features manually. The facial muscle motion contains rich information that represents the facial expression change [15]. According to the authors of [16], video sequences are more suitable than still images for FER. Therefore, it is very necessary to make use of dynamic facial expression representations (temporal features) in video sequences for FER.

Nowadays, Convolutional Neural Networks (CNNs) with strong self-learning ability have been proved to be an effective method for action recognition [17], segmentation [18] and detection [19]. CNN can be further extended to 3D CNN [20], which can extract temporal feature in video sequences. Simonyan, K et al. [21] designed a network for extracting the spatial and temporal features separately from human action video, and achieved an outstanding performance for recognizing human action on Sports-1M dataset. Tran et al. [22] proposed a 3D CNN for extracting spatio-temporal features from video sequences. These features together with simple linear classifier can achieve best performance in different type of video analysis tasks.

To automatically extract temporal features and improve the recognition rate, we present a 3 dimensional convolutional neural network (3D CNN) structure in this research, which can extract the spatio-temporal features of facial expressions. Though 3D CNN was proposed before [20,22], to our knowledge, this study is one of the first work to explore the 3D CNN in NIR FER to achieve best recognition rate.

Besides spatial and temporal features, facial expression representation can also be divided into global and local features. Researchers generally focused on the separated global or local feature extraction methods of facial expressions alone [11,23]. Few studies tried global and local feature fusion methods in NIR FER. To make full use of the facial expression characteristics, we extract global features from the whole face and local features from the partial faces (upper and lower regions) in temporal and spatial dimensions and then fuse the features for FER by using three-stream 3D CNN.

One of the problems of CNN is that over-fitting would happen if the size of the network and the size of the database were not well matched. Min et al. [24] found that a medium-sized network could achieve higher accuracy than what a large-sized network can achieve under the circumstance of middle size dataset. In NIR FER, the database (Oulu-CASIA NIR facial expression database) is not too large, too many layers of network are not suitable for the recognition neither. We therefore design a medium-sized network for NIR FER in this research in order to avoid over-fitting problems.

As described above, the main contributions of this paper are three-fold. First, we present a 3D CNN that can automatically extract spatio-temporal features from NIR video for FER. Secondly, we design a three-stream 3D CNN for extracting global features and local features for further improving the recognition rate. Thirdly, we design the network with a medium-size specifically for Oulu-CASIA NIR facial expression database, which may prevent over-fitting to some extent. Experiment results show that our proposed methods for FER can achieve 78.42% recognition accuracy, which is higher

than other recognition methods, such as Histogram of Oriented Gradient to 3D (HOG 3D) (60%), Local binary patterns from three orthogonal planes (LBP-TOP) (72.33%), deep temporal appearance-geometry network (DTAGN) (66.67%), and adapt 3D Convolutional Neural Networks (3D CNN DAP) (72.12%).

2. Materials and Methods

2.1. 3D CNN

Deep learning is becoming popular, which has outperformed traditional methods in many fields, such as speech recognition [25] and face recognition [26]. In deep learning, CNN capable of strong self-learning is one of the most successful methods for image classification [27]. Different from the traditional neural network, the 2D CNN model optimizes the neural network structure through the local receptive field, shared weights, and sampling. The 2D CNN is suitable for various analysis of still pictures [28,29], whereas it will lose much dynamic information if it is used for analysis of video sequences.

Later on, 2D CNN is extended to 3D CNN [20] in order to solve action recognition problems based on video sequences. The 3D CNN can extract spatio-temporal features from video sequences [22]. The dynamic change of facial expression consists of facial muscle motion. Therefore, 3D CNN can be employed to extract the spatio-temporal features of facial expression.

To extract temporal features, 3D CNN convolves a 3D kernel to the image cube that is generated by stacking several contiguous frames. By using this construction, the feature maps obtain the information of the contiguous frames of previous layers and the thus capture the temporal information. The basic structure of 3D CNN consists of input layer, 3D convolution layer, 3D pooling layer, and fully connection layer.

Input Layer: this layer is composed of normalized video sequences in spatial and temporal dimensions. The dimension of the video sequences is represented as $c \times f \times h \times w$, where c is the number of channels of the video, f is the number of frames of the video, d and k are the height and width of each frame image.

Convolutional Layers $C(d, k, k)$: these layers extract features of the upper layer by several 3D convolution kernels with d and k as temporal and spatial dimension, respectively. A convolutional value is computed by convolving local receptive field $k \times k$ of continuous frames with input feature maps. The output of these layers is passed through a leaky rectified nonlinearity unit (ReLU).

Pooling Layers $P(m, n)$: these layers reduce the computational complexity and avoid possibility of over-fitting. A pooling value is computed by substituting $m \times m \times n$ kernel for maximum or average. In the conventional CNN model, in order to learn more abstract temporal and spatial features, convolutional layers and pooling layers appear alternately, which constitutes the deep CNN model.

Fully Connected Layer $FC(c)$: each unit of feature maps in upper layer will be connected with c units of fully connected layer. The fully connected layer is followed by output layer. The number of outputs corresponds to the number of class labels and a softmax nonlinearity is used to provide a probabilistic output.

2.2. The Proposed System

Facial expression information can be decomposed into global and local information. Global information includes the integral geometry and appearance property of facial expression. Local information focuses on the detail and texture of facial expression. For CNN, the neurons in higher layers capture global abstract features and neurons in lower layers capture local detailed information [26]. We, accordingly, divide the whole structure into two networks: global network and local network. In this section, we first describe the proposed global and local network, and then fuse the two networks into our proposed NIRExpNet for NIR FER. The whole NIRExpNet structure is given in Figure 1.

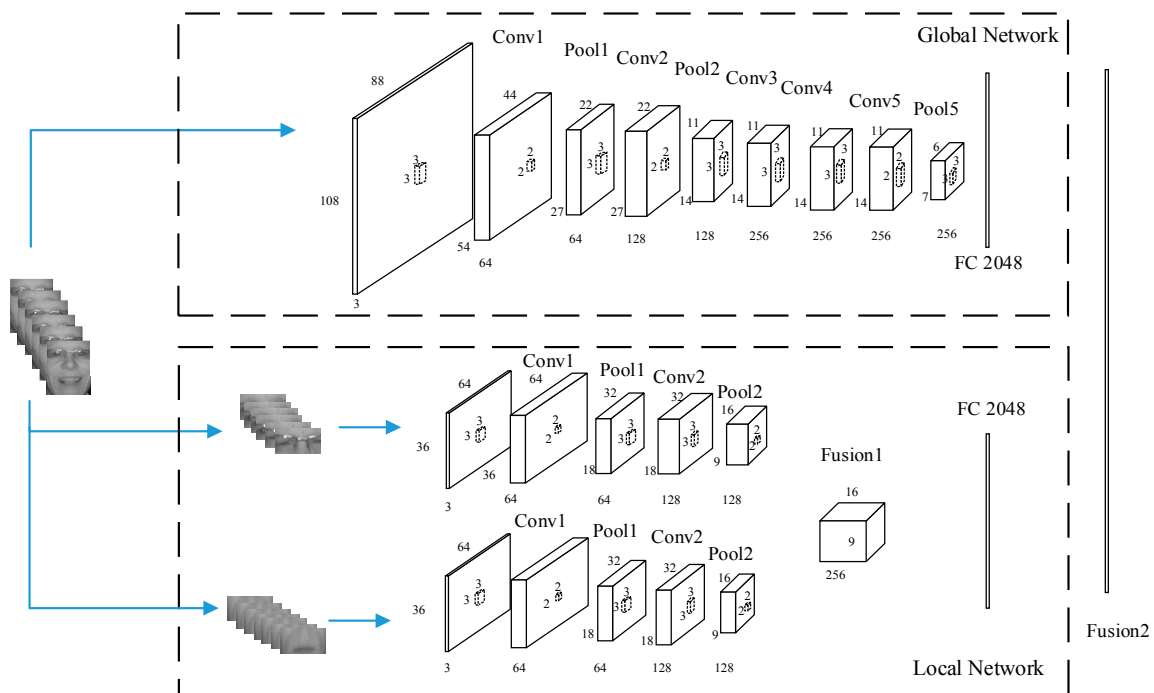


Figure 1. The proposed NIRExpNet for near-infrared facial expression recognition (NIR FER). Global network: the preprocessed whole face acts as input flow of network. The network employed medium-size VGG-M-2048 models. Local network: the two streams using same network that has two convolutional layers and two pool layers and are fused by Fusion 1. The global and local network are fused in Fusion 2.

2.2.1. Global Network

Global network focuses on global information in both temporal and spatial dimension. The network directly takes preprocessed facial expression video frames as network inputs.

The network extracts temporal and spatial features by layers. Min et al. [24] proposed a medium-sized network could achieve higher accuracy than what a large size network can achieve under the circumstances of a medium-sized dataset. The Oulu-CASIA NIR facial expression database used in this research contains 480 videos of 80 persons, which is similar in size to the database used in Min's work [24]. Therefore, we also used a well-designed medium-size network that is VGG-M-2048 [30] with five convolutional layers, three pooling layers, and one fully connected layer that has 2048 neurons. The detailed configurations of global network are given in Table 1.

Table 1. Configurations of global network.

Layers	Patch Size	Output Size	Output
Data	-	-	$108 \times 88 \times 32$
Conv1	$5 \times 5 \times 3$	64	$54 \times 44 \times 32$
Pool1	$2 \times 2 \times 2$	64	$27 \times 22 \times 16$
Conv2	$3 \times 3 \times 3$	128	$27 \times 22 \times 16$
Pool2	$2 \times 2 \times 2$	128	$14 \times 11 \times 8$
Conv3	$3 \times 3 \times 3$	256	$14 \times 11 \times 8$
Conv4	$3 \times 3 \times 3$	256	$14 \times 11 \times 8$
Conv5	$3 \times 3 \times 3$	256	$14 \times 11 \times 8$
Pool5	$2 \times 2 \times 2$	256	$7 \times 6 \times 1$
FC		2048 neurons	

2.2.2. Local Network

Eyes, eyebrows, and mouth are mainly involved in facial expression displays [15]. Therefore, we divide the face into upper region and lower region. The upper region includes the eyes and eyebrows, and lower region includes the mouth. Cropped upper and lower region can reduce intra-class differences [31] and strengthen dynamic and spatial features of eyes, eyebrows, and mouth.

The input data flows are the cropped upper and lower regions of facial expression video sequences. The two input flows in the local network correspond to two streams. The input to the local network has lower spatial resolution than that of the global network, because the shallower network can adapt to the input data flow of relatively lower resolution and extract local features of a partial face. In the local network, it may be better to use a shallower structure than that of the global network. In CNN design, a convolutional layer is normally followed by a pooling layer, and the fully connected layer is at the end of the convolutional and pooling layer pairs. In design of the local network, we adopted this widely used structure. To determine the number of convolutional layers (convolutional-pooling-layer pair), we tried different numbers of convolutional layers (experiment results are given in Section 3.1), and found two convolutional layers gave best result. Thus each stream of local network includes two convolutional layers, two pooling layers and a fully connected layer. Meanwhile, in order to achieve the effect of network symmetry [32], the two streams employ the same network parameters. Finally, the two streams are fused and followed by fully connected layer with the same number neurons as that of global network in order to keep the balance between global and local information. Different fusion strategies can have an impact on the final performance of the network. We compared five different fusion methods and found that concatenation had the best results. The detailed configurations of local network are given in Table 2.

Table 2. Configurations of local network.

Layers	Upper Stream			Lower Stream		
	Patch Size	Output Size	Output	Patch Size	Output Size	Output
Data	-	-	$36 \times 64 \times 32$	-	-	$36 \times 64 \times 32$
Conv1	$3 \times 3 \times 3$	64	$36 \times 64 \times 32$	$3 \times 3 \times 3$	64	$36 \times 64 \times 32$
Pool1	$2 \times 2 \times 2$	64	$18 \times 32 \times 16$	$2 \times 2 \times 2$	64	$18 \times 32 \times 16$
Conv2	$3 \times 3 \times 3$	128	$18 \times 32 \times 16$	$3 \times 3 \times 3$	128	$18 \times 32 \times 16$
Pool2	$2 \times 2 \times 2$	128	$9 \times 16 \times 8$	$2 \times 2 \times 2$	128	$9 \times 16 \times 8$
Fusion 1						
FC	2048 neurons					

2.2.3. Fusion of the Global and Local Networks

Different fusion methods may produce different recognition rates. In this paper, we compared five fusion methods, i.e., concatenation, product, sum, subtract and max fusion method. For clarity, we define a fusion function f , two feature maps x_t^a and x_t^b , and a fused feature maps y , where $x^a \in \mathbb{R}^{H \times W \times D}$, $x^b \in \mathbb{R}^{H \times W \times D}$, $y \in \mathbb{R}^{H' \times W' \times D'}$, where W , H , and D are the width, height, and number of channels of feature maps. The five fusion methods are described as follows:

Concatenation fusion. $y = f_{\text{cat}}(x^a, x^b)$ stacks the two features at the same location i, j across the feature channels d :

$$y_{i,j,d} = x_{i,j,d}^a, y_{i,j,D+d} = x_{i,j,d}^b \quad (1)$$

where $y \in \mathbb{R}^{H \times W \times 2D}$.

Product fusion. $y = f_{\text{product}}(x^a, x^b)$ computes the product of the two feature maps at the same location i, j and feature channels d :

$$y_{i,j,d} = x_{i,j,d}^a \times x_{i,j,d}^b \quad (2)$$

where $1 < i < H, 1 < j < W, 1 < d < D, y \in \mathbb{R}^{H \times W \times D}$.

Sum fusion. $y = f_{\text{sum}}(x^a, x^b)$ computes the sum of the two feature maps at the same location i, j and feature channels d :

$$y_{i,j,d} = x_{i,j,d}^a + x_{i,j,d}^b \quad (3)$$

where $1 < i < H, 1 < j < W, 1 < d < D, y \in \mathbb{R}^{H \times W \times D}$.

Subtract fusion. $y = f_{\text{subtract}}(x^a, x^b)$ computes the subtract of the two feature maps at the same location i, j and feature channels d :

$$y_{i,j,d} = x_{i,j,d}^a - x_{i,j,d}^b \quad (4)$$

where $1 < i < H, 1 < j < W, 1 < d < D, y \in \mathbb{R}^{H \times W \times D}$.

Max fusion. $y = f_{\text{max}}(x^a, x^b)$ computes the max of the two feature maps at the same location i, j and feature channels d :

$$y_{i,j,d} = \max\{x_{i,j,d}^a, x_{i,j,d}^b\} \quad (5)$$

where $1 < i < H, 1 < j < W, 1 < d < D, y \in \mathbb{R}^{H \times W \times D}$.

Through experiments, we found that concatenation fusion method can achieve better recognition rate than other methods. The concatenation fusion method is used in this research. In this section, we adopt the same fusion method with Fusion 1 to fuse global network and local network as Fusion 2 followed by fully connected layer with 2048 neurons. We adopt a softmax layer with dimension of six to recognize facial expression.

For the whole network, global and local networks are complementary. Either global network or local network cannot represent facial expression information solely, as their fusion significantly improves on both (28.29% over local and 6.37% over global network).

2.2.4. Learning Parameters

In order to achieve promising results, it is key to train a good model with many important parameters, which can influence performance of entire structure. During the process of training, the network weight parameters are learnt using mini-batch stochastic gradient descent with momentum (set to 0.9). Each 10 video batch is sent to the network with weight decay of 0.0005. The base learning rate is 10^{-3} and the value is further dropped when the loss stops changing.

2.3. Experiments

The proposed network is evaluated on Oulu-CASIA NIR facial expression database. The NIRExpNet was implemented in Caffe deep learning framework, which runs on a PC with NVIDIA GeForce GTX 1080 GPU (8 G) (NVIDIA CUDA framework 8.0, and cuDNN library).

2.3.1. Oulu-CASIA NIR Facial Expression Database

Oulu-CASIA NIR facial expression database [11] includes facial expression videos captured in three different illumination conditions: normal, weak, and dark. The database was collected in an experiment room. The subjects were asked to sit on a chair in the observation room that he/she was in frontal direction to the camera. The images/videos of the database were captured by using a USB 2.0 PC Camera (SN9C201&202) that includes a VIS camera and a NIR camera. Eighty NIR light-emitting diodes with wavelength of 850 nm were used as active illumination sources. The NIR camera thus mainly receives NIR light at around 850 nm. The camera-face distance is about 60 cm. The subjects were asked to make a facial expression according to an expression example shown in picture sequences. The imager worked at a rate of 25 frames per second and the image resolution was 320×240 . The usability of Oulu-CASIA NIR facial expression database has been confirmed by many researchers and widely used in FER research [33,34]. Zhao et al. [11] demonstrates that NIR FER under dark illumination is most challenge in all three illumination conditions since the facial images captured in the dark illumination often lose much useful texture information. Therefore, in this paper, we tested our methods on the subset of the database (dark illumination condition). This subset consists of 6

expressions (anger, disgust, fear, sadness, surprise, happiness) of 80 persons between 23 and 58 years old. The subjects differ in sex, age, background, and region with glasses or without glasses. There is a collection of 480 (80×6) video sequences. Each video sequence starts at neutral state and ends at the expression's apex. Figure 2 shows six facial expression images of one person in the database in apex condition.



Figure 2. Anger, disgust, fear, happiness, sadness, surprise images of one person in apex condition of Oulu-CASIA NIR facial expression database.

2.3.2. Data Preprocessing

Face detection and segmentation of upper face and lower face needs to be performed before FER. We use Conditional with Local Neural Fields (CLNF) with OpenFace toolkit [35] to detect the face locate the key points on the face (show in Figure 3a). Every frame is aligned and cropped according to the key points. The upper face region (R1 in Figure 3b) and lower face region (R2 in Figure 3b) are rectangular regions, and are segmented by using location of eyes points, nose point, mouth points (shown as the specific red point position in Figure 3a). Assume that the two eye points are $E_1(X_1, Y_1)$ and $E_2(X_2, Y_2)$, nose point are $N(X_3, Y_3)$, central point of mouth is $M(X_4, Y_4)$, left point of mouth is $ML(X_6, Y_6)$, right point mouth is $MR(X_7, Y_7)$. Then the centroid of R1 is $R1c(X_5, Y_5) = R1c((X_1 + X_2)/2, (Y_1 + Y_2)/2)$, and the centroid of R2 is $R2c = M(X_4, Y_4)$. The height and width of R1 are $|Y_5 - Y_3|$ and $\frac{5}{3}|X_2 - X_1|$, respectively. The height and width of R2 is $\frac{5}{3}|Y_4 - Y_3|$ and $|Y_7 - Y_6|$, respectively.

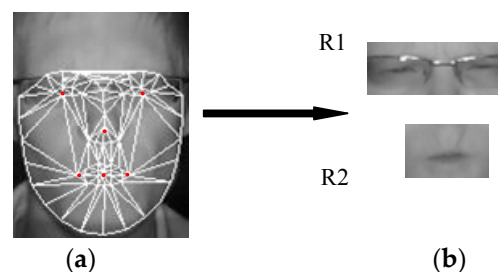


Figure 3. (a) The 68 facial points tracked by Conditional with Local Neural Fields (CLNF) (b) cropped upper and lower regions.

For local networks, we divide the face into two non-overlapping regions including upper and lower regions. The upper region contains eyebrows and eyes and the lower region contains the mouth. We crop the upper region and lower region according to eyes and mouth key points, as shown by the specific red point position in Figure 3a, the cropped upper and lower regions is shown in Figure 3b.

Each video sequence was normalized for 32 frames by using the linear interpolation method [36]. For global input data flow, each image was resized to 108×88 by using bilinear interpolation method [37]. For local input data flows, each image was resized to 36×64 . Then all the preprocessed images were converted from RGB into 8-bit gray scale images to reduce computation complexity.

Data augmentation has always been used as an effective way of reducing over-fitting and improving recognition rates in the case of a small dataset [32]. In the training dataset, we cropped each image along its corner (random location) and changed it from 108×88 into 98×78 . Then, the cropped image is recovered into original size 108×88 by using the bilinear interpolation method. After augmentation, we obtain a training set that is 50 times larger than original dataset.

The 10-fold cross-validation is used. Therefore, there are 72 people as the training set and 8 people as the test set. In all experiments, there are no overlapping images between the training sets and test sets.

3. Results and Discussion

3.1. Comparisons of Different Fusion Strategies and Structures

We used VGG-M-2048 directly as the global network. Our work focused on designing the local network. Because fusion methods and numbers of convolutional layer can both influence the recognition rate. We compared different numbers of the convolutional layer and fusion methods to determine the network structure.

Figure 4 gives the comparison results. The recognition rate given in Figure 4 is the final recognition rate of the whole network. It is seen that in all cases of varying number of layers, the concatenation method can give the best final recognition rate. This may indicate that the concatenation method can better fuse the global and local features and is more suitable in this research. It is also observed that when the number of convolutional layers reaches two, the recognition rate has the maximum value. From the point of two layers, the recognition rate will decrease slightly if the number of layers increases. This may be because a two layer convolutional network is enough for the medium-sized database in this research, and therefore, more layers become redundant. Thus, we chose two convolutional layers structure in the local network and concatenation fusion methods in this research. This design is based on the experiment results and also accords with the inference (in Section 2.2.2) that the local network have shallower structure than that of the global network.

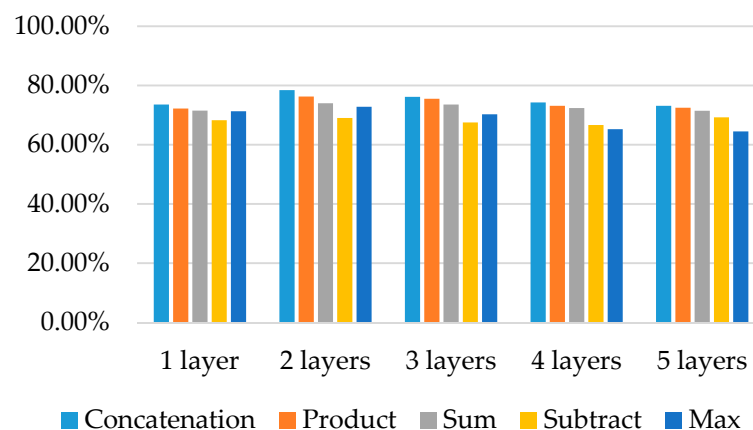


Figure 4. Performance comparison of five kinds of fusion strategies and five kinds of depth of local network.

3.2. Comparisons of Different Streams and Their Combinations

Table 3 shows 10-fold cross-validation results using separate networks and fused networks respectively, where ‘U’ represents upper face network, ‘L’ represents lower face network, ‘U+L’ represents fused local network.

Table 3. Comparisons of different streams combinations.

Architecture	Accuracy
U	31.27%
L	43.87%
U+L (Local)	50.13%
Global	72.05%
NIRExpNet (Global + Local)	78.42%

Upper face network extracting features of eyes and eyebrows region achieves an accuracy of 31.27% and lower face network extracting features of mouth region achieves an accuracy of 43.87%. It can be seen that the performance of the upper face network is lower than that of the lower face network. This could be due to the effect of glasses, i.e., if the subjects have glasses, the features of eyes could be lost to some extent.

Meanwhile, we can see that the performance of the global network (72.05%) outperforms the local network (50.13%). This may be because the local features extracted in this research do not cover all features of facial expressions. However, the local network fused with the global network can improve the recognition accuracy. It is seen that NIRExpNet (global network + local network) can achieve a recognition accuracy of 78.42%, which is the highest in all kinds of networks. This may indicate that either global or local networks cannot represent facial expression information completely, that the global features and local features are complementary to each other, and that the combination of them can improve the recognition rate.

3.3. Confusion Matrixes of NIRExpNet and Global Network

In order to compare NIRExpNet and global network further, we list the confusion matrix of global network and NIRExpNet in Figure 5. It is seen that the NIRExpNet can achieve a higher recognition rate on every type of expression. The increases of the recognition rates are from 5% to 9%, except in the case of the recognition rate on happiness, which only increases 0.01%. This may be due to the fact that happiness already has a very high recognition rate (96.00%).

It is observed from Figure 5 that anger is confused with disgust and sadness, disgust is confused with anger and fear, fear is confused with disgust and surprise, happiness is confused with surprise, sadness is confused with anger and fear, and surprise is confused with fear. This may indicate the movement pattern of one type of expression being somewhat similar to that of the other types of expression. However, the NIRExpNet can decrease the confusion on all types of expression, especially in the case of recognizing disgust. The global network confused disgust with anger and fear. However, the NIRExpNet only confused disgust with anger.

One type of expression can be confused with two other types of expression. In some cases NIRExpNet selectively decreased one type of confusion more. In the case of recognizing fear, NIRExpNet mainly decreased the situation of confusing fear with disgust (from 17.50% to 8.00%). In the case of recognizing surprise, NIRExpNet mainly decreased the situation of confusing surprise with fear (from 17.56% to 9.41%). This may be because the movement pattern of fear is less similar to that of disgust, and the pattern of surprise is less similar to that of fear.

Anger	71.01% (64.78%)	14.43% (17.66%)	0 (0)	0 (0)	14.56% (17.56%)	0 (0)
Disgust	20.56% (14.50%)	79.44% (70.94%)	0 (14.56%)	0 (0)	0 (0)	0 (0)
Fear	0 (0)	8.00% (17.50)	62.44% (53.17%)	0 (0)	0 (0)	29.56% (29.33%)
Happiness	0 (0)	0 (0)	0 (0)	96.01% (96.00%)	0 (0)	3.99% (4.00%)
Sadness	10.44% (14.43%)	0 (0)	14.44% (15.46%)	0 (0)	75.12% (70.11%)	0 (0)
Surprise	0 (0)	0 (0)	9.41% (17.56%)	4.04% (5.11%)	0 (0)	86.55% (77.33%)
	Anger	Disgust	Fear	Happiness	Sadness	Surprise

Figure 5. Confusion matrix of NIRExpNet and global network for NIR FER (Numbers inside brackets represent the results of global network).

3.4. Comparisons between NIRExpNet and State-of-the-Art Methods

The NIRExpNet and several state-of-the-art methods were compared in the Oulu-CASIA NIR facial expression database under dark illumination. All the methods employ the 10-fold cross-validation, and the parameters of the methods were set according to the original works.

It is observed that LBP-TOP outperforms 3D HOG, which are both hand-crafted features, in terms of achieving higher recognition rates. This may suggest LBP-TOP is more suitable for NIR FER.

NIRExpNet (78.42%) outperforms the LBP-TOP (72.33%) and 3D HOG (60.00%) that use hand-crafted features. This result verifies that NIRExpNet can automatically extract features which help to improve the recognition rate.

It is also seen that NIRExpNet (78.42%) outperforms other CNN-based methods, such as DTAGN (66.67%) [38] and 3D CNN DAP (72.12%) [31]. The DTAGN and 3D CNN DAP both use temporal features. The higher recognition rate achieved by NIRExpNet could be because global and local features were fused in our design, which provide more information of facial expression. The results may also imply that the structure of NIRExpNet is more suitable for NIR FER.

LBP-TOP can achieve the second highest recognition rate according to Table 4. In Table 5, we list the recognition rates of NIRExpNet and LBP-TOP on every type of expression for comparison. It is seen that NIRExpNet can achieve higher recognition rate nearly on every expression. One exception is the fear expression: LBP-TOP has higher recognition rate (68.18%) than that of NIRExpNet (62.44%). This may be due to the limited size of training data. If large database is available for training, a deeper network can be designed, which may improve the recognition rate of recognizing fear expression.

Table 4. Comparisons between our approach and state-of-the-art methods.

Method	Accuracy
LBP-TOP [11]	72.33%
3D HOG [6]	60.00%
DTAGN [38]	66.67%
3D CNN DAP [31]	72.12%
NIRExpNet	78.42%

Table 5. Comparisons of different methods on the Oulu-CASIA NIR database.

Method	Expression						
	Anger	Disgust	Fear	Happiness	Sadness	Surprise	Average
NIRExpNet	71.01%	79.44%	62.44%	96.01%	75.12%	86.55%	78.42%
LBP-TOP [11]	65.67%	56.60%	68.18%	78.75%	71.21%	86.25%	72.33%
DTAGN [38]	69.25%	70.32%	59.32%	71.13%	60.21%	71.12%	66.67%
3D CNN DAP [31]	69.82%	73.41%	60.21%	83.23%	64.30%	81.55%	72.12%

We also compare the performance of other CNN-based methods in Table 5. It is seen that 3D CNN DAP can achieve higher recognition rate on every type of expression than DTAGN. Furthermore, NIRExpNet can achieve the highest recognition rate on every type of expression in all three CNN-based methods. This may be because NIRExpNet is well designed and suitable for solving the NIR FER problems. The 3D CNN DAP has very similar recognition performance with LBP-TOP. However, it is observed that 3D CNN DAP has much higher recognition rate on the disgust expression, and also much lower recognition rate on the fear expression. This may again indicate that the size of training data is not sufficient, in that for some expressions, the feature extracted by 3D CNN DAP is not general enough. A larger database for NIR FER research is needed.

4. Conclusions

In this paper, we proposed a three-stream 3D convolutional network called NIRExpNet to recognize NIR facial expression from video sequences. The NIRExpNet considers the motion information of facial expression and extracts spatio-temporal features from the video. We divided the whole network into two components: global network and local network to consider both global and local information of facial expressions. For the global network, we input the complete facial video sequences and employed VGG-M-2048 model to extract global dynamic information. For the local network, the upper and lower regions of facial video sequences were segmented as two data flows input into the local network. We fused two streams by using the concatenation fusion followed by a fully connected layer with 2048 neurons. We further fused the global and local networks to extract facial expression features in motion. Experimental results on the Oulu-CASIA NIR facial expression database showed that the NIRExpNet can achieve an average recognition rate of 78.42%, and therefore, outperform some state-of-the-art methods, such as LBP-TOP [11], 3D HOG [6], DTAGN [38], and 3D CNN DAP [31].

Acknowledgments: This work was partially funded by the National Natural Science Foundation of China (Grant No. 61301297), and the Southwest University Doctoral Foundation (No. SWU115093).

Author Contributions: T.C. conceived the research, Z.W. performed the data analysis, and T.C., Z.W., Y.C., Z.Z., and G.L. wrote the paper.

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

References

1. Knutson, B. Facial expressions of emotion influence interpersonal trait inferences. *J. Nonverbal Behav.* **1996**, *20*, 165–182. [[CrossRef](#)]
2. Vinciarelli, A.; Pantic, M.; Bourlard, H. Social signal processing: Survey of an emerging domain. *Image Vis. Comput.* **2009**, *27*, 1743–1759. [[CrossRef](#)]
3. Pantic, M.; Patras, I. Dynamics of facial expression: Recognition of facial actions and their temporal segments from face profile image sequences. *IEEE Trans. Syst. Man Cybern. Syst.* **2006**, *36*, 433–449. [[CrossRef](#)]
4. Tulyakov, S.; Slowe, T.; Zhang, Z.; Govindaraju, V. Facial expression biometrics using tracker displacement features. In Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, USA, 18–23 June 2007; pp. 1–5.
5. Corneanu, C.A.; Simon, M.O.; Cohn, J.F.; Guerrero, S.E. Survey on RGB, 3D, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 1548–1568. [[CrossRef](#)] [[PubMed](#)]
6. Klaser, A.; Marszałek, M.; Schmid, C. A Spatio-Temporal Descriptor Based on 3D-Gradients. In Proceedings of the BMVC 2008—19th British Machine Vision Conference, Leeds, UK, 1–4 September 2008; pp. 1–10.
7. Zhang, Z.; Wang, Y.; Zhang, Z. Face synthesis from low-resolution near-infrared to high-resolution visual light spectrum based on tensor analysis. *Neurocomputing*. **2014**, *140*, 146–154. [[CrossRef](#)]
8. Wang, S.; Liu, Z.; Lv, S.; Lv, Y.; Wu, G.; Peng, P.; Wang, X. A natural visible and infrared facial expression database for expression recognition and emotion inference. *IEEE Trans. Multimed.* **2010**, *12*, 682–691. [[CrossRef](#)]
9. Tan, X.; Triggs, B. Enhanced local texture feature sets for face recognition under difficult lighting conditions. *IEEE Trans. Image Process.* **2010**, *19*, 1635–1650. [[PubMed](#)]
10. Qiao, Y.; Lu, Y.; Feng, Y.S.; Li, F.; Ling, Y. A new method of NIR face recognition using kernel projection DCV and neural networks. In Proceedings of the 2013 Fifth International Symposium on Photoelectronic Detection and Imaging, Beijing, China, 25 June 2013; pp. 89071M1–6.
11. Zhao, G.; Huang, X.; Taini, M.; Li, S.Z.; Pietikäinen, M. Facial expression recognition from near-infrared videos. *Image Vis. Comput.* **2011**, *29*, 607–619. [[CrossRef](#)]
12. 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](#)]
13. Gejji, R.S.; Clark, A.D.; Crihalmeanu, S.; Rossy, A.A. Understanding the subject-specific effects of pupil dilation on iris recognition in the NIR spectrum. In Proceedings of the 2015 IEEE International Symposium on Technologies for Homeland Security (HST), Waltham, MA, USA, 14–16 April 2015; pp. 1–6.
14. Son, C.; Zhang, X. Near-Infrared Image Dehazing Via Color Regularization. In Proceedings of the 2016 IEEE Computer Vision and Pattern Recognition, Seattle, WA, USA, 27–30 June 2016.
15. Fasel, B.; Luetttin, J. Automatic facial expression analysis: A survey. *Pattern Recognit.* **2003**, *36*, 259–275. [[CrossRef](#)]
16. Bassili, J. Emotion recognition: The role of facial movement and the relative importance of upper and lower areas of the face. *J. Personal. Soc. Psychol.* **1979**, *37*, 2049–2059. [[CrossRef](#)]
17. Chéron, G.; Laptev, I.; Schmid, C. P-CNN: Pose-based CNN features for action recognition. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, December 2015; pp. 3218–3226.
18. Liu, F.; Lin, G.; Shen, C. CRF learning with CNN features for image segmentation. *Pattern Recognit.* **2015**, *48*, 2983–2992. [[CrossRef](#)]
19. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. In Proceedings of the NIPS 2015 Advances in Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; pp. 91–99.
20. Ji, S.; Xu, W.; Yang, M.; Yu, K. 3D convolutional neural networks for human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 221–231. [[CrossRef](#)] [[PubMed](#)]
21. Simonyan, K.; Zisserman, A. Two-stream convolutional networks for action recognition in videos. In Proceedings of the NIPS 2014 Advances in Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; pp. 568–576.

22. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning spatiotemporal features with 3D convolutional networks. In Proceedings of the 2015 IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 4489–4497.
23. Gu, W.; Xiang, C.; Venkatesh, Y.V.; Huang, D.; Lin, H. Facial expression recognition using radial encoding of local Gabor features and classifier synthesis. *Pattern Recognit.* **2012**, *45*, 80–91. [[CrossRef](#)]
24. Peng, M.; Wang, C.; Chen, T. NIRFaceNet: A Convolutional Neural Network for Near-Infrared Face Identification. *Information* **2016**, *7*, 61. [[CrossRef](#)]
25. Zhang, Z.; Geiger, J.; Pohjalainen, J.; Mousa, A.E.D.; Schuller, B. Deep Learning for Environmentally Robust Speech Recognition: An Overview of Recent Developments. *arXiv*, 2017.
26. Sun, Y.; Chen, Y.; Wang, X.; Tang, X. Deep learning face representation by joint identification-verification. In Proceedings of the NIPS 2014 Advances in Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; pp. 1988–1996.
27. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In Proceedings of the NIPS 2012 Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–6 December 2012; pp. 1097–1105.
28. Chatfield, K.; Simonyan, K.; Vedaldi, A.; Zisserman, A. Return of the devil in the details: Delving deep into convolutional nets. *arXiv*, 2014.
29. Sun, Y.; Liang, D.; Wang, X.; Tang, X. Deepid3: Face recognition with very deep neural networks. *arXiv* **2015**, arXiv:1502.00873.
30. Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y. Towards good practices for very deep two-stream convnets. *arXiv* **2015**, arXiv:1507.02159.
31. Liu, M.; Li, S.; Shan, S.; Wang, R.; Chen, X. Deeply learning deformable facial action parts model for dynamic expression analysis. In Proceedings of the 12th Asian Conference on Computer Vision (ACCV), Singapore, 1–5 November 2014; pp. 143–157.
32. Gens, R.; Domingos, P.M. Deep symmetry networks. In Proceedings of the NIPS 2014 Advances in Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; pp. 2537–2545.
33. Jung, H.; Lee, S.; Park, S.; Lee, I.; Ahn, C.; Kim, J. Deep temporal appearance-geometry network for facial expression recognition. *arXiv* **2015**, arXiv:1503.01532.
34. Rivera, A.R.; Chae, O. Spatiotemporal directional number transitional graph for dynamic texture recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 2146–2152. [[CrossRef](#)] [[PubMed](#)]
35. Baltrušaitis, T.; Robinson, P.; Morency, L.P. Openface: An open source facial behavior analysis toolkit. In Proceedings of the 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Placid, NY, USA, 7–9 March 2016; pp. 1–10.
36. Smolic, A.; Muller, K.; Dix, K.; Merkle, P.; Kauff, P.; Wiegand, T. Intermediate view interpolation based on multiview video plus depth for advanced 3D video systems Image Processing. In Proceedings of the 15th IEEE International Conference on Image Processing, San Diego, CA, USA, October 2008; pp. 2448–2451.
37. Prashanth, H.S.; Shashidhara, H.L.; KN, B.M. Image scaling comparison using universal image quality index. In Proceedings of the IEEE International Conference on Advances in Computing, Control & Telecommunication Technologies, Kyoto, Japan, 27 September 2009; pp. 859–863.
38. Jung, H.; Lee, S.; Yim, J.; Park, S.; Kim, J. Joint fine-tuning in deep neural networks for facial expression recognition. In Proceedings of the 2015 IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 2983–2991.

