future internet

MDPI

Article

Neurologist Standard Classification of Facial Nerve Paralysis with Deep Neural Networks

Anping Song *, Zuoyu Wu, Xuehai Ding, Qian Hu and Xinyi Di

School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China; zzzuo@i.shu.edu.cn (Z.W.); dinghai@shu.edu.cn (X.D.); ileven@shu.edu.cn (Q.H.); jsbluecat@shu.edu.cn (X.D.)
* Correspondence: apsong@shu.edu.cn; Tel.: +86-136-5167-5579; Fax: +86-6613-5550

Received: 27 October 2018; Accepted: 15 November 2018; Published: 16 November 2018

Abstract: Facial nerve paralysis (FNP) is the most common form of facial nerve damage, which leads to significant physical pain and abnormal function in patients. Traditional FNP detection methods are based on visual diagnosis, which relies solely on the physician's assessment. The use of objective measurements can reduce the frequency of errors which are caused by subjective methods. Hence, a fast, accurate, and objective computer method for FNP classification is proposed that uses a single Convolutional neural network (CNN), trained end-to-end directly from images, with only pixels and disease labels as inputs. We trained the CNN using a dataset of 1049 clinical images and divided the dataset into 7 categories based on classification standards with the help of neurologists. We tested its performance against the neurologists' ground truth, and our results matched the neurologists' level with 97.5% accuracy.

Keywords: facial image analysis; facial nerve paralysis; deep convolutional neural networks; image classification

1. Introduction

Facial nerve paralysis (FNP) is one of the most common facial neurological dysfunctions, in which the facial muscles appear to droop or weaken. Such cases are often accompanied by the patient having difficulty chewing, speaking, swallowing, and expressing emotions. Furthermore, the face is a crucial component of beauty, expression, and sexual attraction. As the treatment of FNP requires an assessment to plan for interventions aimed at the recovery of normal facial motion, the accurate assessment of the extent of FNP is a vital concern. However, existing methods for FNP diagnosis are inaccurate and nonquantitative. In this paper, we focus on computer-aided FNP grading and analysis systems to ensure the accuracy of the diagnosis.

Facial nerve paralysis grading systems have long been an important clinical assessment tool; examples include the House–Brackmann system (HB) [1], the Toronto facial grading system [2,3], the Sunnybrook grading system [4], and the Facial Nerve Grading System 2.0 (FNGS2.0) [5]. However, these methods are highly dependent on the clinician's subjective observations and judgment, which makes them problematic with regard to integration, feasibility, accuracy, reliability, and reproducibility of results.

Computer-aided analysis systems have been widely employed for FNP diagnosis. Many such systems have been created to measure facial movement dysfunction and its level of severity, and rely on the use of objective measurements to reduce errors brought about through the use of subjective methods.

Anguraj et al. [6] utilized Canny edge detection to locate a mouth edge and eyebrow, and Sobel edge detection to find the edges of the lateral canthus and the infraorbital region. Nevertheless, these edge detection techniques are very vulnerable to noise. Neely [7–9] and Mcgrenary [10] used a dynamic

video image analysis system which analyzed patients' clinical images to assess FNP. They used very simple neural networks on FNP, which validated the technology's potential. Although their results were consistent with the HB scoring system, they had a very small dataset and their system's image processing was computationally intensive. He et al. [11] used optical-flow tracking and texture analysis to solve the problem using image processing to capture the asymmetry of facial movements by analyzing the patients' video data, but this is computationally intensive. Wachtman et al. [12] measured asymmetry using static images, but their method is sensitive to extrinsic facial asymmetry caused by orientation, illumination, and shadows.

For our method, a new FNP classification standard was established based on FNGS2.0 and asymmetry. FNGS2.0 is a widely used assessment system which has been found to be highly consistent with clinical observations and judgment, achieving 84.8% agreement with neurologist assessments [13].

Using deep learning to detect facial landmarks in our previous method has shown promising results. Deep convolutional neural networks (DCNNs) [14] show potential for general and highly variable tasks on image classification [15–19]. Deep learning algorithms have recently been shown to exceed human performance in visual tasks like playing Atari games [20] and recognizing objects [16]. In this paper, we outline the development of a CNN that matches neurologist performance for human facial nerve paralysis using only image-based classification.

GoogleNet Inception v3 CNN architecture [18] was pretrained on approximately 1.28 million images (1000 object categories) for the 2014 ImageNet Large Scale Visual Recognition Challenge [16]. Sun et al. [21] proposed an effective means for learning high-level overcomplete features with deep neural networks called DeepID CNN, which classified faces according to their identities.

At the same time, DCNNs have had many outstanding achievements as diagnostic aids. Rajpurkar et al. [22] developed a 34-layer CNN which exceeds the performance of board-certified cardiologists in detecting a wide range of heart arrhythmias from electrocardiograms recorded using a single-lead wearable monitor. Hoochang et al. [23] used a CNN combined with transfer learning on computer-aided detection. They studied two specific computer-aided detection (CADe) problems, namely thoraco-abdominal lymph node (LN) detection and interstitial lung disease (ILD) classification. They achieved state-of-the-art performance on mediastinal LN detection and reported the first fivefold cross-validation classification results on predicting axial CT slices with ILD categories. Esteva et al. [15] used a pretrained GoogleNet Inception v3 CNN on skin cancer classification, which matched the performance of dermatologists in three key diagnostic tasks: melanoma classification, melanoma classification using dermoscopy, and carcinoma classification. Sajid et al. [24] used a CNN model to classify facial images affected by FNP into the five distinct degrees established by House and Brackmann. Sajid used a Generative Adversarial Network (GAN) to prevent overfitting in training. His research demonstrates the potential of deep learning on FNP classification, even though his final classification accuracy results were not very good (89.10–96.90%, depending on the class). Meanwhile, they used a traditional grading standard to directly label the data which may cause erroneous labeling. They also used four complicated image preprocessing steps, which cannot be automated and which require a lot of time and effort during the clinical diagnosis phase for data labeling.

In the process of realizing a reliable computer-aided analysis system, we also proposed a method for FNP quantitative assessment [25]. We used a DCNN to obtain facial features, then we used asymmetry algorithm to calculate FNP degree. In this work, we validated the effectiveness of DCNN. However, there is currently no work related to the hierarchical classification of FNP using DCNN.

The difficulty of FNP classification lies first and foremost in image classification, followed by face recognition. To design a responsive and accurate CNN for FNP classification, we combined a GoogleNet Inception v3 CNN and a DeepID CNN to design a new CNN called Inception-DeepID-FNP (IDFNP) CNN. As it is difficult to obtain a large enough training dataset, direct training of our model would cause overfitting results, so we need to use transfer learning methods [26] to eliminate overfitting, as given the amount of expected data available, this was considered to be the optimal

choice. We trained the IDFNP CNN by training on ImageNet with no final classification layer and then retrained it using our dataset. This method is optimal given the amount of data available.

Compared with other classification methods, we set up our own dataset classification standards. We used deep learning to directly classify FNP, which allows each FNP image to be processed more quickly, has more accurate classification, and has lower image quality requirements. In order to improve the liability and accuracy of our labeling results, we used a triple-check method to complete the labeling of the image dataset. At the same time, we combined image classification with face recognition.

Using the proposed system, clinicians can quickly obtain the degree of facial paralysis under different movements and make a prediagnosis of facial nerve condition, which can then be used as a reference for final diagnosis. At the same time, we also developed a mobile phone application that enables patients to perform self-evaluations, which can help them avoid unnecessary visits to hospitals.

The remainder of this paper is structured as follows.

The proposed methodology is presented in Section 2. The experiments and results are given in Section 3. The results and related discussion are presented in Section 4. The conclusions about this study are given in Section 5.

2. Materials and Methods

2.1. Data Sources

We used two types of data sources, a fixed camera in a hospital and a mobile application.

2.1.1. Hospital Camera

In order to establish a novel method for quantitative FNP assessment, we prepared a fixed scene in the Department of Rehabilitation at the Shanghai Tenth People's Hospital in order to obtain FNP images with the neurologists' help. We captured front-view facial images of the patients using reasonable illumination to reduce any adverse illumination effects. The procedure for obtaining the images was standardized; photography was executed while the participant was seated in a chair, and a reference background was placed behind. The camera was mounted on a sturdy tripod at a distance of 1.5 m from the participant, and the latter was instructed to look directly at the camera with their chin raised. Then, digital images were acquired as each participant performed each of the different movements.

2.1.2. Mobile Application

For the purposes of the present study, we developed a mobile application for both iPhone and Android devices, with the end-goal being that patients would be able to obtain an automated preassessment of the extent of their FNP using their mobile phone camera. Participants were asked to download the application, which used the phone's camera and suitable prompts to obtain the relevant images of the participant.

2.2. Dataset

Our dataset came from a combination of an FNP dataset and a normal dataset. The FNP dataset came from clinical images from the Department of Rehabilitation at the Shanghai Tenth People's Hospital. The FNP dataset was composed of 377 male images and 483 female images, of which 136 were of patients less than 40 years old, 302 were middle-aged (between 40 and 65 years old), and 422 were elderly (greater than 65 years old). The normal dataset was composed of recovered patients, volunteers to our research group, and healthy neurologists from the hospital's Department of Rehabilitation. The normal dataset was composed of 86 male normal images and 103 female images, of which 38 were less than 40 years old, 82 were between 40 and 65 years old, and 69 were elderly (Table 1). Our dataset covers patients of all ages and genders, while patient data are relatively evenly distributed.

Table 1. Dataset Distribution.

Dataset	Young	Middle-Aged	Elderly	Male	Female	Total
FNP images	136	302	422	377	483	860
Normal images	38	82	69	86	103	189
Total	174	384	491	463	586	1049

Figures 1 and 2, respectively, show example facial images of the control and the patient groups taken as each group was performing seven facial movement types: at rest, eyes closed, eyebrows raised, cheeks puffed, grinning, nose wrinkled, and whistling. Table 2 contains a description of each movement. These images were used for our model's training.

Table 2. Taxonomy movements table.

Notation	Movement	Affected Facial Muscle
MV0	At rest	All facial muscles
MV1	Eyes closed	Orbicularis oculi muscle
MV2	Eyebrows raised	Orbicularis oculi muscle, frontalis muscle
MV3	Cheeks puffed	Orbicularis oculi muscle, buccinator muscle, zygomatic muscle
MV4	Grinning	Orbicularis oris muscle
MV5	Nose wrinkled	Nasalis muscle
MV6	Whistling	Orbicularis oris muscle

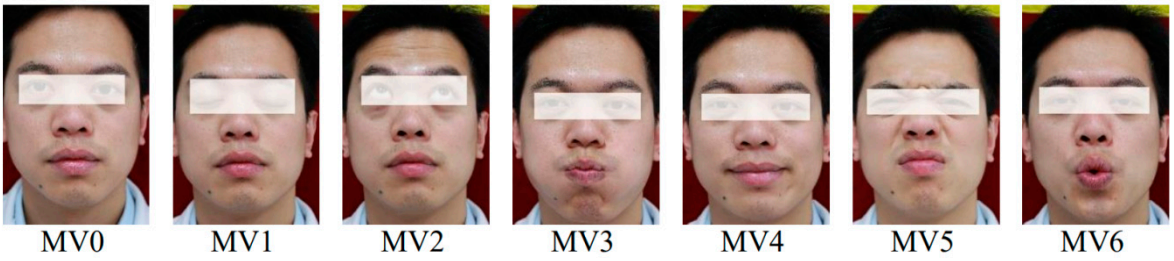

Figure 1. Facial images of the control subjects during seven movement types.

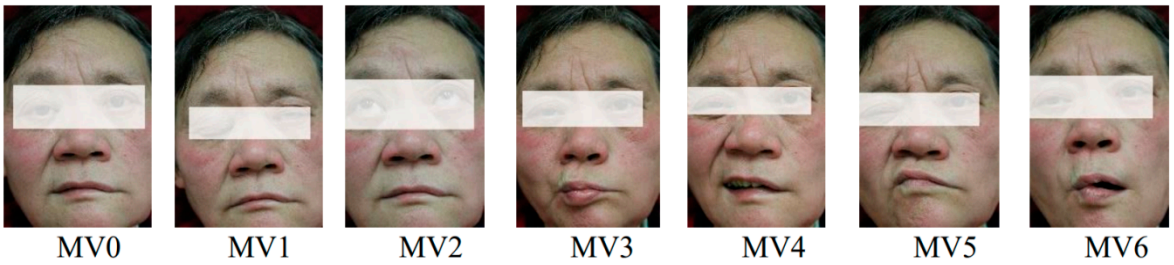

Figure 2. Facial images of patients with paralysis during the seven movement types.

2.3. Taxonomy

2.3.1. Classification Standard

Since FNP causes barriers to the movement of facial muscles, we can evaluate the degree of FNP by calculating the asymmetry of facial features for different facial movements. This method was chosen because simultaneous bilateral FNP is highly improbable. Our method is based on facial image analysis. Considering our dataset consists of FNP images and not video, in order to reduce subjective factors and the difficulty of diagnosis, the new classification standard divides the dataset into seven categories. These are: normal, left mild dysfunction, left moderate dysfunction, left severe dysfunction, right mild dysfunction, right moderate dysfunction, and right severe dysfunction (Table 3).

Table 3. Taxonomy characteristics.

Taxonomy	Symbol	Characteristics
normal	N	Normal function in all facial nerve areas.
left mild dysfunction	L1	Slight muscular weakness observed on examination of the side of the face. Facial image appears symmetrical and with tones in MV0, eye area exhibits mild asymmetry in MV1, eyebrow area exhibits mild asymmetry in MV2, cheek area exhibits mild asymmetry in MV3, mouth area exhibits mild asymmetry in MV4 and MV6, nose and cheek areas exhibit asymmetry in MV5.
right mild dysfunction	R1	
left moderate dysfunction	L2	In the side of the face, there is a clear difference between the two hemifaces but this is not total asymmetry. Nonserious disordered movement may be observed. Facial image exhibits mild asymmetry in MV0, eye area exhibits moderate asymmetry in MV1, eyebrow area exhibits moderate asymmetry in MV2, cheek area exhibits moderate asymmetry in MV3, mouth area exhibits moderate asymmetry in MV4 and MV6, nose and cheek areas exhibit moderate asymmetry in MV5.
right moderate dysfunction	R2	
left severe dysfunction	L3	In the side of the face, there is clear weakness and total asymmetry, with hardly any observable mobility. Facial image exhibits severe asymmetry in MV0, eye area exhibits severe asymmetry in MV1, eyebrow area exhibits severe asymmetry in MV2, cheek area exhibits severe asymmetry in MV3, mouth area exhibits severe asymmetry in MV4 and MV6, nose and cheek areas exhibit severe asymmetry in MV5.
right severe dysfunction	R3	

2.3.2. Frequencies in Dataset Taxonomy

Our taxonomy represents seven different classes of FNP and their frequency for the study sample is given in Table 4. This aspect of the taxonomy is useful for generating training classes that are well suited for machine learning classifiers. We obtained 664 images from the hospital camera and 385 images from the application.

Table 4. Frequencies in dataset taxonomy.

Taxonomy	**N**	**L1**	**L2**	**L3**
Frequency	189	133	161	146
Taxonomy	**R1**	**R2**	**R3**	
Frequency	129	151	140	

2.3.3. Labeling

In order to objectively divide image database into those seven categories, we used a triple-check method to complete the labeling of the image dataset.

To start with, neurologists labeled images into seven different categories twice, and only coinciding labels were retained for subsequent steps. This was the first check in the process.

Then, we measured the degree of bilateral face FNP difference using asymmetry [25]. In order to measure the asymmetry of patients during different facial movements, we assessed eye asymmetry (EAs), eyebrow asymmetry (EBAs), nose asymmetry (NAs), mouth asymmetry (MAs), mouth angle (MAn), nose angle (NAn), and eyebrow angle (EbAn). We quantified this assessment using two variables, regional asymmetry (RgAs) and angular asymmetry (AnAs), which were calculated using the following equation:

$$RgAs = EAs + EBAs + NAs + MAs \quad (1)$$

$$AnAs = MAn + NAn + EbAn \quad (2)$$

Based on the results of the first check, we obtained the range of $RgAs$ and $AnAs$ for every movement type in the same manner for the seven categories.

Since the results of this work are not accurate enough, the work on the classification of the face can only be used as a reference, so we still need to optimize the results to ensure the accuracy of the labeling. We compared the results of the asymmetrical algorithm with the first-check results as reference and kept the coinciding results to obtain the second-check result. Neurologists will take the results of the asymmetrical algorithm as reference to analyze the different part above. Finally, neurologists will obtain the final classification results for the third check.

Using this approach, the results of the first check reached 97% agreement, and for the second check, we achieved 93% agreement.

2.3.4. Data Preparation

Since our data came from two different sources, data transformation was the first step of our method. The biggest difference between the two data sources were the environmental factors. The FNP images taken on the mobile phone application suffered from problems with face angle and image size. We therefore preprocessed the images to obtain a standardized format of the face image. In order to eliminate the influence of environmental factors, we cropped every image. To make them compatible with the IDFNP CNN architecture, we resized each image to 299 × 299 × 3 pixels, which were used as the input to IDFNP. However, because the image size was fixed at 299 × 299, and image cropping may have resulted in loss of facial nerve information, cropping was adjusted according to the specific facial movement being captured. In order to retain as much facial nerve muscle information as possible, cropping retained all parts of the muscle for a specific movement. Pictures were cropped automatically and the results were visually inspected and, if necessary, corrected manually to ensure that no useful information was discarded.

Blurry images and distant images were removed from the test and validation sets, but were still used for training. While this is useful training data, extensive care was taken to ensure that these sets were not split between the training and validation sets. No overlap (that is, same lesion, multiple viewpoints) existed between the test sets and the training/validation data.

Based on the above principles, the 1049 images selected after filtering were randomly and evenly divided using a 7:2:1 ratio for the training, verification, and test sets, respectively. The training set batch size was 60, the cross-validated batch size was 100, and for k-fold cross-validation we used $k = 10$.

2.4. Model Architecture

The difficulty of FNP classification lies first and foremost in image classification, followed by face recognition. Inception v3 CNN [18] shows great performance on image classification and won first prize during the 2015 ImageNet Large Scale Visual Recognition Challenge [16]. At the same time, DeepID CNN [21] is the top model in the field of face recognition. In order to design a model for FNP classification, we combined the best image classification CNN model and the best face recognition CNN model for the learning task. In order to combine GoogleNet Inception v3 CNN and DeepID CNN, and thereby create IDFNP CNN, we must identify their essential components and utilize them.

The complete model is based on the Inception-v3 architecture. Apart from the essential components of Inception-v3 and DeepID, IDFNP used a concat layer to concatenate the parameters of the two parts. After the above, the FNP grade classification task is performed by the softmax layer.

The network's high-level architecture is shown in Figure 3.

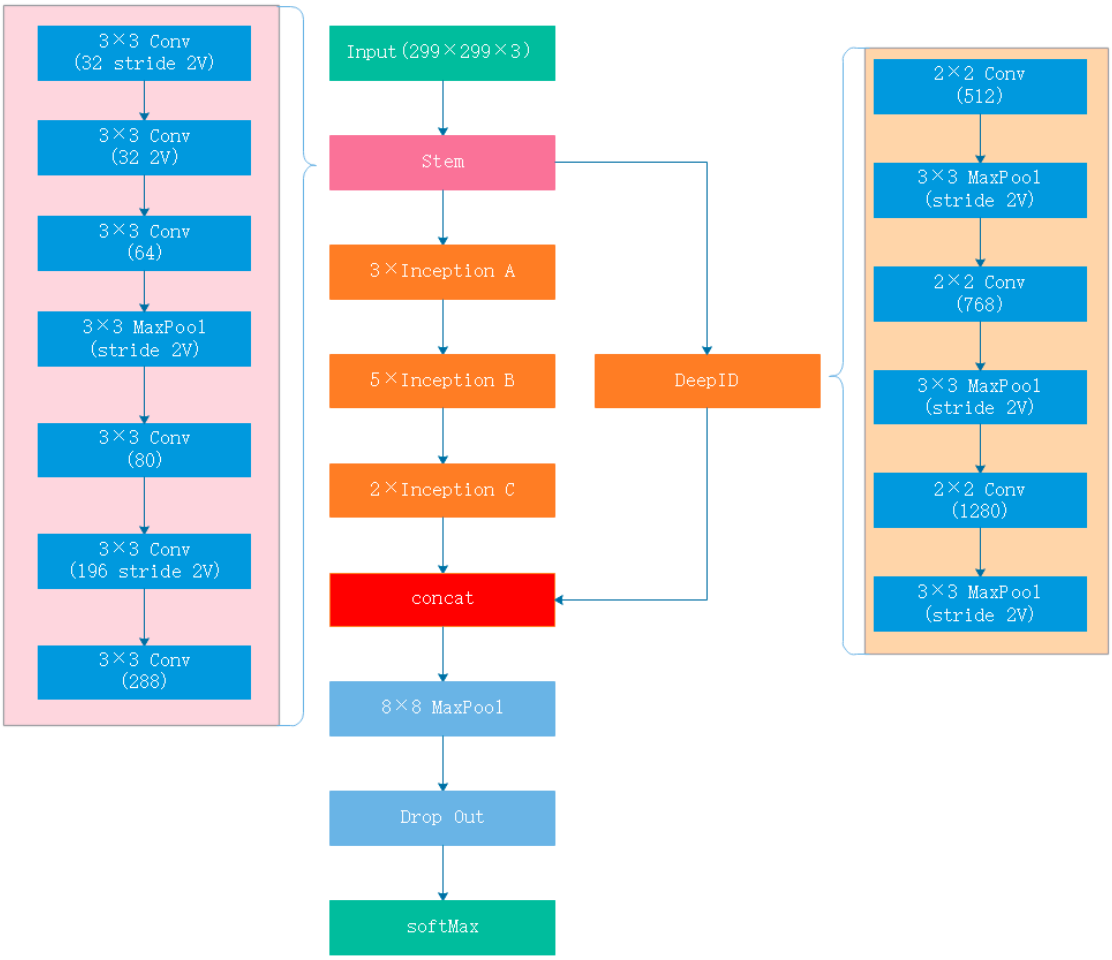

Figure 3. Inception-DeepID-FNP CNN.

Because FNP classification counts as image classification, putting the DeepID CNN part into GoogleNet Inception v3 CNN was our strategy of choice. Since the DeepID CNN has much fewer characteristics than GoogleNet Inception v3 CNN, we fine-tuned the parameters across multiple layers in order to enhance the human face component.

2.5. Training Algorithm

As it is difficult to obtain a large enough training dataset, direct training of our model would cause overfitting results, so we needed to use migration study methods to eliminate overfitting. Given the amount of expected data available, transfer learning was considered to be the optimal choice.

The ImageNet Challenge Database is a 1000 object class (1.28 million images) image database. Pretraining the model on ImageNet Challenge Database will increase the model's sensitivity to image classification. FNP image classification is based on the details and characteristics of facial muscles, while ImageNet classification is based on the details and characteristics of the classification for which it is trained. The data distribution of the FNP database and ImageNet Challenge Database are similar and, in this case, we transferred the model from a source domain (pretrained model) to a target domain (final model).

The IDFNP CNN is based on Inception-v3 CNN, which has very good performance in the ImageNet Challenge Database. Therefore, we pretrained the IDFNP CNN on the ImageNet Challenge Database and achieved a 93.33% classification accuracy, ranking top-five compared with other CNNs. We then removed the final classification layer from the network, retrained it with our own dataset, and leveraged the natural-image features already learned by the ImageNet pretrained network.

The classification task is performed by the softmax layer, and we used back propagation to update the network weights for training. All layers of the network were fine-tuned using the same global learning rate of 0.001 and a decay factor of 16 every 30 epochs. We used RMSProp [27], which can speed up first-order gradient descent methods, with a decay of 0.9, momentum of 0.9, and epsilon of 0.1. We used Google's TensorFlow deep learning framework to train, validate, and test our network.

3. Results

3.1. Confusion Matrix

Precision: The precision metric represents the correctly predicted labels out of the total true predictions. The precision achieved for every label is shown in Table 5.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3}$$

where TP and FP represent true positive and false positive.

Table 5. Precision of IDFNP for Every Taxonomy.

Taxonomy	N	L1	L2	L3	R1	R2	R3
Precision	0.974	0.956	0.980	0.960	0.969	0.959	0.951

Sensitivity: The sensitivity metric is used to quantify the cases that are predicted correctly (i.e., the number of predicted labels over all positive observations). IDFNP's sensitivity of every label is shown in Table 6.

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{4}$$

where TP and FN represent true positive and false negatives, respectively.

Table 6. Sensitivity for Every Taxonomy.

Taxonomy	N	L1	L2	L3	R1	R2	R3
Sensitivity	0.974	0.977	0.937	0.993	0.961	0.933	0.978

Accuracy: This metric which represents correct predictions out of total predictions:

$$\text{Accuracy} = \frac{TP + FN}{TP + FN + FP + TN} = 97.5\% \tag{5}$$

where TP, TN, FP, and FN represent true positive, true negative, false positive and false negatives, respectively.

Figure 4 shows the confusion matrix of our method over the seven classes of predicted labels. Element of each confusion matrix represents the empirical probability of predicting class given that the ground truth

By analyzing the confusion matrix, one can observe that the proposed method can predict the FNP types well. The highest classification accuracy was 0.993, achieved for L3, while the lowest classification accuracy was 0.933 for R2. It can be seen that the accuracy is very high for the most serious disease conditions (R3 and L3), but the accuracy is not very high for intermediate disease conditions (R2 and L2). The overall accuracy was 97.5%.was class.

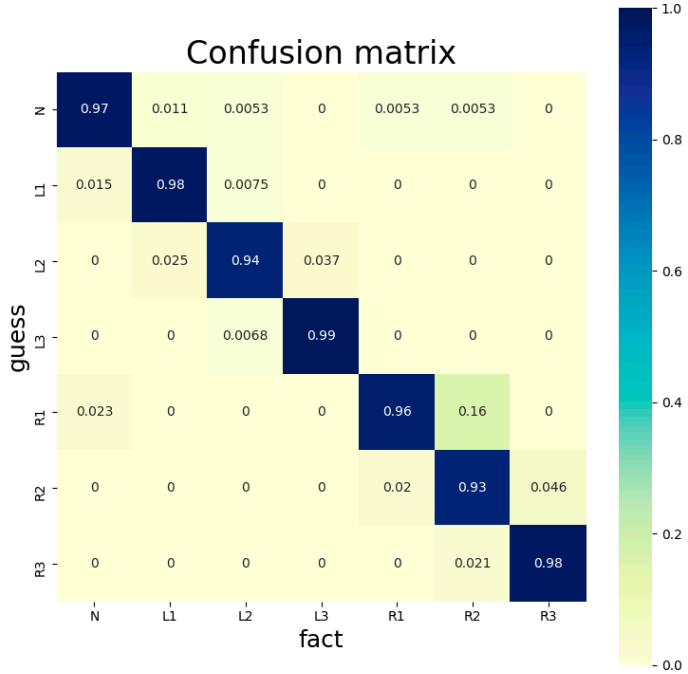

Figure 4. Confusion Matrix.

3.2. Comparison with Previous Methods and Neurologist Classification

In this study, we divided all the movements (MV0, MV1, etc.) into different levels (N, L1, L2, et al.). In the process of specific training, we did not separate the different movements and did not test them accordingly. We believe that the FNP grading should not be performed by the movements. When FNP images are input into our system, the movement type does not need to be identified, as this is another deep learning topic; the output of our system is the FNP grading of the image. In our case, the accuracy for all movements was 97.5%.

To conclusively validate the algorithm, we used our previous method [25] for FNP quantitative assessment to compare validity with IDFNP. Meanwhile, neurologists classified the unlabeled FNP images. In this task, the IDFNP achieved 97.5% classification accuracy based on all movement, while our previous method for FNP quantitative assessment achieved 79.2–98.7% accuracy. Apart from MV0 RgAs, this method achieves a maximum of 94.4% in the other 13 ways of measuring FNP (Table 7).

Table 7. Comparison with previous method and neurological agreement.

Movements	Previous Method Accuracy on RgAs	Previous Method Accuracy on AnAs	IDFNP CNN Accuracy	Neurological Agreement
MV0	98.7%	80.6%		98.0%
MV1	94.4%	81.9%		97.3%
MV2	93.1%	80.6%		97.5%
MV3	94.4%	79.2%	97.5%	97.4%
MV4	93.1%	81.9%		97.1%
MV5	94.4%	84.7%		97.7%
MV6	94.4%	96.1%		98.0%

We asked neurologists to diagnose each FNP image again when we went through the whole set; the double diagnosis agreement for the side affected by FNP reached 100%, while the double diagnosis agreement for the FNP degree ranged between 97.1% and 98.0%. Neurological agreement represents consistent neurological classification for FNP. As the images in the validation set were labeled by neurologists, but not necessarily confirmed by them, this metric is inconclusive, and instead actually shows that the CNN is learning relevant information.

3.3. Comparison with Other Computer-Aided Analysis Systems

Sajid et al. [24] used a CNN model to classify face images with FNP into the five distinct degrees established by House and Brackmann. Sajid used GAN to prevent overfitting in training (Column 3, VGG-16 Net with GAN). Neely [28] used a computerized objective measurement of facial motion to obtain diagnosis of facial paralysis; using a standardized classification method, he achieved an accuracy of 95% (Columns 4). HC et al. [23] used optical-flow tracking and texture analysis methods to solve the problem. They used advanced image processing technology to capture the asymmetry of facial movements by analyzing the patients' video data and then used several different classification methods to diagnose FNP. The result is shown in Table 2 (Columns 5–6, RBF with 0/1 disagreement). Wang et al. [29,30] presented a novel method for grading facial paralysis integrating both static facial asymmetry and dynamic transformation factors. Wang used an SVM with the RBF kernel function to quantify the static facial asymmetry on images using five of the six facial movements (MV1–6), but they did not measure accuracy of MV0. The results are shown in column 7 of Table 8.

Table 8. Comparison with the other computer-aided analysis systems.

Movements	IDFNP CNN Accuracy	VGG-16 Net with GAN	Neely's Method	RBF with 0 Disagreement	RBF with 1 Disagreement	SVM with the RBF
MV0				44.36%	92.26%	
MV1				41.27%	93.11%	97.56%
MV2				68.71%	93.23%	92.55%
MV3	97.5%	92.60%	95%	49.80%	94.18%	91.36%
MV4				49.80%	94.18%	95.40%
MV5				61.78%	86.31%	93.36%
MV6				49.80%	94.18%	95.40%

3.4. Comparison with Other Deep Convolution Neural Networks Models

Because our dataset's scale is not large enough to train models directly, for every model compared, we removed the final classification layer from the network, retrained it with our dataset, and leveraged the natural image features learned by the ImageNet pretrained network, a technique known as transfer learning. We chose Inception-v3, Inception-v4, Inception-ResNet-v1, Inception-ResNet-v2, DeepID, and ResNet, which in recent years have shown the best results in image classification. The results are shown in Table 9.

Table 9. Comparison of IDFNP with other classification CNNs.

Network	Accuracy for FNP Dataset	Network	Accuracy for FNP Dataset
Inception-ResNet-v1	95.2%	Inception-ResNet-v2	95.7%
Inception-v3	93.3%	DeepID	92.5%
Inception-v4	95.3%	ResNet	95.0%
IDFNP		**97.5%**	

For accuracy, IDFNP CNN outperforms all the other CNNs for the FNP dataset. All the other CNNs were designed for the ImageNet Challenge Database, which has 1000 object classes and are optimized for image classification, which is quite relevant for the present application. Our original plan for diagnosing FNP was to use transfer learning with Inception-ResNet-v2 directly. However, the result did not match the accuracy of neurologists. Considering that FNP classification is a face classification, combining DeepID CNN with Inception-v3 CNN improves accuracy.

4. Discussion

As we see from Table 7, neurological agreement exceeds our method in MV2, MV5, and MV6. However, neurologists take too long examining FNP images, as each such examination takes at least

10 s. Our method takes a few milliseconds per FNP image and is thus more efficient, while its accuracy is comparable to that of neurologists. Our previous method takes much longer per FNP image by calculating facial asymmetry with traditional computational methods, while only its accuracy in MV0 on RgAs is higher. Furthermore, our previous method requires more standard images like face angle, image clarity, and lighting conditions.

As we see from Table 8, the accuracy of FNP classification when using Sajid's method was 92.6%. The accuracy of FNP in Neely's method [28] is 95%, which is lower than our method. HC [28] used RBF with 0/1 disagreement to measure accuracy of FNP movements. Even with 1 disagreement, which allows for more experimental errors, the result is significantly worse than ours. Wang [29,30] used SVM with RBF to measure accuracy. The result showed our method is better than their method in MV2–6. In MV1, their accuracy is not much higher than ours. Although they didn't calculate the accuracy of MV0, we can still see from the rest of the results that our method yields superior results.

As we see from Table 9, these models have strong generalization ability for different datasets, but because their design was optimized for their main, that is, image classification, the final training results of these models are not as good as our model. We also see that Inception-v3, upon which our own design was based, achieved only 93.3% accuracy. Therefore, there is still considerable potential for the optimization of this excellent image classification model for specific applications, especially with residual network derivatives like Inception-ResNet-v2.

Meanwhile, on the basis of our findings, clinicians can quickly obtain the degree of facial paralysis according to different facial movements. Clinicians can make a prediagnosis of facial nerve paralysis based on patients' facial movements, which will be used as a reference for their final diagnosis. For example, the result of one patient in MV1 (Eye closed), MV2 (Eyebrows raised), and MV4 (Grinning) was L3, and the result of the patient in other movements was N or L1, which corresponds to a prediagnosis that severe paralysis is present in the in left orbicularis oculi muscle.

5. Conclusions

In this paper, we presented a neural network model called IDFNP for FNP image classification, which uses a deep neural network and can achieve accuracy which is comparable to that of neurologists. Key to the performance of the model is an FNP annotated dataset and a deep convolutional network which can classify facial nerve paralysis and facial nerve paresis effectively and accurately. IDFNP combines Inception-v3, which achieves a great result in image classification, and DeepID, which is highly efficient in facial recognition.

The contributions of our method can be summarized as follows: Firstly, a symmetry-based annotation scheme for FNP images with seven different classes is presented. Secondly, using deep neural network on FNP images and cropping the face from the FNP images can eliminate facial deformation for FNP patients and minimize the influence of environmental factors. Thirdly, transfer learning avoids overfitting effectively for a limited range of FNP images. Combining an image classification CNN, such as Inception-v3, and a face recognition CNN like DeepID improves accuracy for the FNP dataset and achieves the same diagnostic accuracy as a neurologist. Fourthly, our method is validated against the performance of other well-known methods, which serves as proof that IDFNP is suitable for FNP classification and can effectively assist neurologists in clinical diagnosis.

In terms of clinical diagnosis, future work will be needed to apply IDFNP performance to other facial diseases or diseases which can be identified visually. On the one hand, more detailed diagnosis of facial paralysis would further aid neurologists in their work. In the future, we plan to undertake a more in-depth study of the position and the degree of disease. On the other hand, we can extend our findings to other conditions. For example, one of the symptoms of a stroke is facial asymmetry, which is very similar to the symptoms of FNP. If IDFNP can diagnose strokes and distinguish various degrees of facial stroke images and facial nerve paralysis images, then preventive treatment for strokes based on facial images can be realized. Given that modern smartphones and PCs are power tools of

deep learning, with the help of the IDFNP results, citizens will have an enhanced ability to obtain an automated assessment for these diseases that may prompt them to visit a specialized physician.

The evaluation results produced by our methods are mostly consistent with the subjective assessment of doctors. Our methods can help clinicians to decide on a specific therapy for each patient, and for the most affected region of the face as reference.

Given that more and more FNP patients are being treated, high-accuracy diagnosis from FNP images can save expert clinicians and neurologists considerable time and decrease the frequency of misdiagnosis. Furthermore, we hope that this technology will enable greater widespread use of FNP images through photography as a diagnostic tool in places where access to a neurologist is limited.

Author Contributions: Conceptualization, A.S. and Z.W.; Data curation, A.S., Z.W., and X.D. (Xuehai Ding); Formal analysis, X.D. (Xuehai Ding); Funding acquisition, A.S.; Investigation, X.D. (Xuehai Ding); Methodology, A.S.; Project administration, A.S.; Resources, A.S. and X.D.; Software, Z.W. and Q.H.; Supervision, X.D. (Xuehai Ding); Validation, Z.W., Q.H., and X.D. (Xinyi Di); Visualization, Q.H.; Writing—original draft, Z.W.; Writing—review & editing, Z.W.

Funding: This research was funded by the National Natural Science Foundation of China (Grant No. 91630206).

Acknowledgments: This study is financially supported by the National Natural Science Foundation of China (Grant No. 91630206). Their support is greatly appreciated.

Conflicts of Interest: All authors declared that they have no conflict of interest.

References

1. House, J.W.; Brackmann, D.E. Facial nerve grading system. *Laryngoscope* **2010**, *93*, 1056–1069.
2. Fields, M.J.; Peckitt, N.S. Facial nerve function index: A clinical measurement of facial nerve activity in patients with facial nerve palsies. *Oral Surg. Oral Med. Oral Pathol.* **1990**, *69*, 681–682. [CrossRef]
3. Ross, B.R.; Fradet, G.; Nedzelski, J.M. *Development of a Sensitive Clinical Facial Grading System*; Springer: Berlin/Heidelberg, Germany, 1994; pp. 380–386.
4. Facs, J.G.N.M.; Cherian, N.G.; Dickerson, C.B.; Nedzelski, J.M. Sunnybrook facial grading system: Reliability and criteria for grading. *Laryngoscope* **2010**, *120*, 1038.
5. Vrabec, J.T.; Backous, D.D.; Djalilian, H.R.; Gidley, P.W.; Leonetti, J.P.; Marzo, S.J.; Morrison, D.; Ng, M.; Ramsey, M.J.; Schaitkin, B.M. Facial Nerve Grading System 2.0. *Otolaryngol. Head Neck Surg.* **2009**, *140*, 445–450.
6. Anguraj, K.; Padma, S. Analysis of Facial Paralysis Disease using Image Processing Technique. *Int. J. Comput. Appl.* **2013**, *54*, 1–4. [CrossRef]
7. Neely, J.G.; Joaquin, A.H.; Kohn, L.A.; Cheung, J.Y. Quantitative assessment of the variation within grades of facial paralysis. *Laryngoscope* **1996**, *106*, 438–442. [CrossRef]
8. Helling, T.D.; Neely, J.G. Validation of objective measures for facial paralysis. *Laryngoscope* **1997**, *107*, 1345–1349. [CrossRef]
9. Neely, J.G. Advancement in the evaluation of facial function. *Adv. Otolaryngol.* **2002**, *15*, 109–134.
10. Mcgrenary, S.; O'Reilly, B.F.; Soraghan, J.J. Objective grading of facial paralysis using artificial intelligence analysis of video data. In Proceedings of the IEEE Symposium on Computer-Based Medical Systems, Dublin, Ireland, 23–24 June 2005; pp. 587–592.
11. He, S.; Soraghan, J.; O'Reilly, B. Biomedical Image Sequence Analysis with Application to Automatic Quantitative Assessment of Facial Paralysis. *Eurasip J. Image Video Process.* **2007**, *2007*, 081282. [CrossRef]
12. Wachtman, G.S.; Liu, Y.; Zhao, T.; Cohn, J.; Schmidt, K. Measurement of Asymmetry in Persons with Facial Paralysis. Available online: https://www.ri.cmu.edu/publications/measurement-of-asymmetry-in-persons-with-facial-paralysis/ (accessed on 15 November 2018).
13. Lee, H.Y.; Park, M.S.; Byun, J.Y.; Ji, H.C.; Na, S.Y.; Yeo, S.G. Agreement between the Facial Nerve Grading System 2.0 and the House-Brackmann Grading System in Patients with Bell Palsy. *Clin. Exp. Otorhinolaryngol.* **2013**, *6*, 135–139. [CrossRef]
14. Yann, L.C.; Yoshua, B.; Geoffrey, H. Deep learning. *Nature* **2015**, *521*, 436–444.
15. Esteva, A.; Kuprel, B.; Novoa, R.A.; Ko, J.; Swetter, S.M.; Blau, H.M.; Thrun, S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **2017**, *542*, 115–118. [CrossRef]

16. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
17. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. In Proceedings of the International Conference on Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–6 December 2012; pp. 1097–1105.
18. 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 Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
19. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
20. Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A.A.; Veness, J.; Bellemare, M.G.; Graves, A.; Riedmiller, M.; Fidjeland, A.K.; Ostrovski, G. Human-level control through deep reinforcement learning. *Nature* **2015**, *518*, 529. [CrossRef]
21. Sun, Y.; Wang, X.; Tang, X. Deep Learning Face Representation from Predicting 10,000 Classes. In Proceedings of the Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1891–1898.
22. Rajpurkar, P.; Hannun, A.Y.; Haghpanahi, M.; Bourn, C.; Ng, A.Y. Cardiologist-Level Arrhythmia Detection with Convolutional Neural Networks. *arXiv*, 2017; arXiv:1707.01836.
23. Hoochang, S.; Roth, H.R.; Gao, M.; Lu, L.; Xu, Z.; Nogues, I.; Yao, J.; Mollura, D.; Summers, R.M. Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. *IEEE Trans. Med Imaging* **2016**, *35*, 1285.
24. Sajid, M.S.T.; Baig, M.; Riaz, I.; Amin, S.; Manzoor, S. Automatic Grading of Palsy Using Asymmetrical Facial Features: A Study Complemented by New Solutions. *Symmetry* **2018**, *10*, 242. [CrossRef]
25. Song, A.; Xu, G.; Ding, X.; Song, J.; Xu, G.; Zhang, W. Assessment for facial nerve paralysis based on facial asymmetry. *Australas. Phys. Eng. Sci. Med.* **2017**, *40*, 1–10.
26. Pan, S.J.; Yang, Q. A Survey on Transfer Learning. *IEEE Trans. Knowl. Data Eng.* **2010**, *22*, 1345–1359. [CrossRef]
27. Tieleman, T.; Hinton, G. *Lecture 6.5-RMSProp, COURSERA: Neural Networks for Machine Learning*; University of Toronto: Toronto, ON, USA, 2012.
28. Neely, J.G.; Wang, K.X.; Shapland, C.A.; Sehizadeh, A.; Wang, A. Computerized Objective Measurement of Facial Motion: Normal Variation and Test-Retest Reliability. *Otol. Neurotol.* **2010**, *31*, 1488–1492. [CrossRef]
29. Wang, T.; Zhang, S.; Dong, J.; Liu, L.A.; Yu, H. Automatic evaluation of the degree of facial nerve paralysis. *Multimed. Tools Appl.* **2016**, *75*, 11893–11908. [CrossRef]
30. Wang, T.; Dong, J.; Sun, X.; Zhang, S.; Wang, S. Automatic recognition of facial movement for paralyzed face. *Biomed. Mater. Eng.* **2014**, *24*, 2751–2760.