Region Based Parallel Hierarchy Convolutional Neural Network for Automatic Facial Nerve Paralysis Evaluation

Xin Liu, Yifan Xia, Hui Yu, *Senior Member, IEEE*, Junyu Dong, Muwei Jian and Tuan D. Pham, *Senior Member, IEEE*

Abstract— **In this paper, we propose a parallel hierarchy convolutional neural network (PHCNN) combining a Long Short-Term Memory (LSTM) network structure to quantitatively assess the grading of facial nerve paralysis (FNP) by considering the region-based asymmetric facial features and temporal variation of the image sequences. FNP, such as Bell's palsy, is the most common facial symptom of neuromotor dysfunctions. It causes the weakness of facial muscles for the normal emotional expression and movements. The subjective judgement by clinicians completely depends on individual experience, which may not lead to a uniform evaluation. Existing computer-aided methods mainly rely on some complicated imaging equipment, which is complicated and expensive for facial functional rehabilitation. Compared with the subjective judgment and complex imaging processing, the objective and intelligent measurement can potentially avoid this issue. Considering dynamic variation in both global and regional facial areas, the proposed hierarchical network with LSTM structure can effectively improve the diagnostic accuracy and extract paralysis detail from the low-level shape, contour to sematic level features. By segmenting the facial area into two palsy regions, the proposed method can discriminate FNP from normal face accurately and significantly reduce the effect caused by age wrinkles and unrepresentative organs with shape and position variations on feature learning. Experiment on the YouTube Facial Palsy Database and Extended CohnKanade Database shows that the proposed method is superior to the state of the art deep learning methods.**

Index Terms—**facial nerve paralysis, severity grade, region of interest, spatio-temporal features, LSTM**

I. Introduction

Facial nerve paralysis (FNP), such as Bell's palsy, is a neuromuscular disorder disease that causes the symptoms of facial weakness, disability of facial expressions and movements [1], [2]. The effects of FNP not only cause physical pain, but also put severe restrictions on social interaction and living quality which is often thought as eccentric disposition, depression and loneliness [3], [4]. Facial function rehabilitation is a long-term process [5]. An objective and quantitative grading assessment is required for early diagnosis and postoperative recovery. It can help patients complete self-assessment and enhance self-confidence in their rehabilitation process [6].

Over the years, various methods have been proposed for the assessment of FNP including both clinical and non-clinical ones [7]. The clinical method is a comprehensive judgement according to clinical experience, patients' performance of different expressions and various scoring standards, which is labor-intensive, subjective, biased and affected by patient's performance of the mimic facial movements. The House–Brackmann system (HBS) [8] and the Sunnybrook facial grading system [9] are the two main scales used for the evaluation of FNP. HBS is widely used for scoring facial paralysis due to its reliability, brevity, accuracy and ease of understanding [10], [11]. It grades facial palsy in six scores from normal (Grade 1) to total paralysis (Grade 6) according to functional performance of facial muscle, which is shown in Table Ⅰ. Different from clinical methods, computer-aided methods rely on specialized optical equipment and various multi-dimension imaging techniques [12], [13]. Recently, 3D reconstruction technologies and 4D imaging systems have been proposed to assess the performance of facial reanimation [14], [15]. Although, these techniques can obtain good results, they are too complicated and expensive for common patients and not convenient to use in their daily lives. Standardized 2D

This work was supported in part by Shanxi Scholarship Council of China (2020-127) and was supported in part by the EPSRC project (EP/N025849/1). (Corresponding author: Hui Yu, Email: hui.yu@port.ac.uk)

X. Liu is with the School of Electronic Information Engineering, Taiyuan University of Science and Technology, Taiyuan, 030024, China (email: liuxin@tyust.edu.cn).

H. Yu, Y. Xia and M. Jian are with the School of Creative Technologies, University of Portsmouth, Portsmouth, PO1 2DJ, UK. M. Jian is also with the School of Computer Science and Technology, Shandong University of Finance and Economics, Jinan, China. (email: hui.yu@port.ac.uk; Yifan.Xia@myport.ac.uk; jianmuweihk@163.com)

J. Dong is with the Department of Computer Science and Technology, Ocean University of China, Qingdao, China. (email: dongjunyu@ouc.edu.cn)T.D. Pham is with the Center for Artificial Intelligence, Prince Mohammad Bin Fahd University, Al Khobar, Saudi Arabia. (email: tpham@pmu.edu.sa)

photography/videography is available for objective measurement and comparison of outcomes in this special patient population. Asymmetry, shape and texture features learned from static images not only require accurate facial landmarks, but also ignore the temporal motion variation. Evaluation results also suffered from the illumination condition and patient's exhausting expression.

Since facial muscle movements are important indicators in the detection and tracking facial motor dysfunction. Dynamic facial motion features can be extracted through multiple consecutive image frames, which are beneficial to recognizing facial muscle movements. Recent development of computer vision and deep learning networks have demonstrated revolutionary capabilities in feature learning and image recognition, especially in medical imaging [16]-[19]. Some samples of different HBS grades and different facial movements are shown in Fig. 1, including closing eyes, frowning, wrinkling nose, lifting corners of mouth and pursing lips [20-22]. These movements illustrate paralysis feature in different facial areas. For example, the muscle weakness symptom around the eye region is obvious while closing eyes; eye corner droop and deformation of lip contour are revealed from wrinkling nose expression. Thus, asymmetry with FNP is a comprehensive symptom of multiple facial organs. Based on this fact, a parallel hierarchy convolutional neural network (PHCNN) with Long Short-Term Memory (LSTM) solution is proposed to objectively evaluate the severity grade of FNP by considering the correlations between facial organs and learning the spatial dependency information among key facial areas.

Compared with existing methods, the key contributions of this paper are:

(1) A region-based PHCNN with a LSTM structure is proposed for automatic and objective FNP evaluation. According to the extracted temporal variation of facial paralysis features from both holistic and regional areas, the proposed network can accurately distinguish the FNP from the normal face and classify the severity degree without accurate landmarks.

(2) There are two types of shallow subnets corresponding to the global face and salient paralysis region in our network. One is used to extract the general contour of facial organs and asymmetry differences between two sides of the face. The other automatically learn hierarchical nerve paralysis features through salient regions, from the low-level shape, texture to abstract sematic level information. Both holistic and regional features are then fused as input of LSTM for assessing the FNP severity degree.

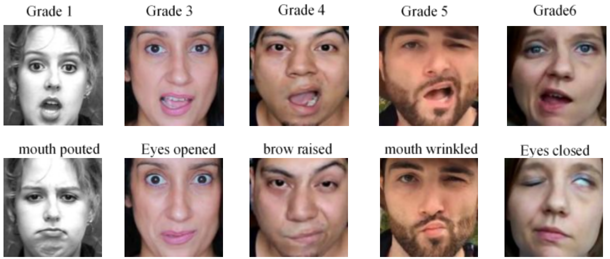

Fig. 1. Examples of different facial movements performed by facial palsy patients with different grading levels [20-22]

The remainder of this paper is structured as follows.

The related work is presented in Section 2. The proposed method is described in Section 3. The experiments and results are given in Section 4. The conclusion and related discussion are presented in Section 5.

II. Related Work

Accurate facial landmarks are usually used for computing the displacement and quantify the severity of paralysis in some researches [23-25]. Conventional facial feature point detection algorithms, such as active shape model (ASM) or active appearance model (AAM), are trained by using normal face. However, they are not effective in detecting the feature points accurately for the face of FNP. To improve the performance, Modersohn et al. [24] combined feature extraction methods based on AAM with a fast and non-linear classifier (Random Decision Forests) to predict the patients' grade of facial paralysis. Following the HBS, they obtained a prediction rate of 80%. Wang et al. [25] proposed an algorithm using ASM considering both static and dynamic facial asymmetries for recognizing facial movement patterns from facially paralyzed patients. But the intensity of FNP has not been quantitatively evaluated. Yoshihara et al. [26] proposed an automatic and accurate feature point detection method for quantitative evaluation of facial paralysis by using DCNN. Song et al. [27] proposed a neural network model called Inception-DeepID-FNP for classifying seven facial movements. Hsu et al. [21] quantitatively analyzed the intensity variation over time by using facial image sequences and the landmark outline. The accuracy of landmark directly affects the line segment map. Sajid et al. [28] proposed a CNN-based model combined with a data augmentation strategy to classify facial palsy images into five grades. It claimed to be the first study for palsy grade classification on a large dataset. In that method, since only 2000 images were available, a generative adversarial network (GAN) was applied for data augmentation. But the synthesized images might only have the palsy-specific features in a local region, which is not enough accurate compared with the real FNP images.

Generally, dynamic variations of facial muscle motion provide more information (e.g. temporal features) than static images. We incorporate LSTM in the network architecture. LSTM has been widely used in the task of facial expression and action recognition [29-32]. Zhang et al. [31] combined the time and texture information of image sequences by

TABLE I

House Brackmann system and severity level

Grade	Description	Severity Level
1	Normal	I
2	Mild dysfunction	II
3	Moderate dysfunction	II
4	Moderately severe dysfunction	II
5	Barely perceptible movement	III
6	Total paralysis	III

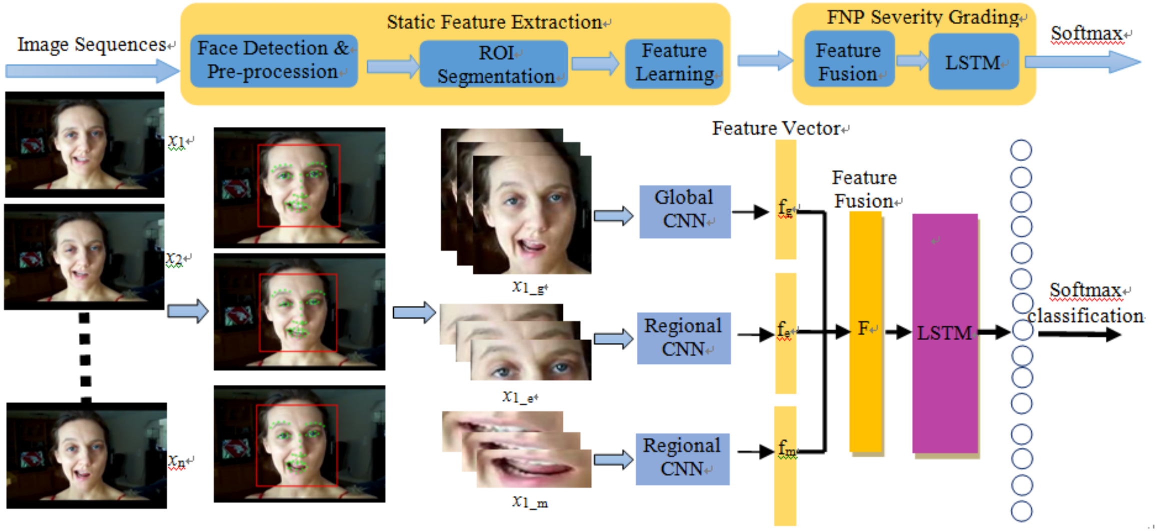

Fig. 2. Overview of the framework for facial palsy severity grading

double-channel WMCNN-LSTM for improving recognition rate. At present, there are few studies using image sequences for facial paralysis grading. Storey et al. [32] proposed a 3D CNN architecture with a ResNet framework to recognize mouth motion and assess facial palsy.

Different with facial expression recognition, FNP assessment is to evaluate the intensity of facial deformation through the landmark distance features and spatio-temporal variation information involved in patient's facial salient areas. Some researchers also focused on the local symmetric features of facial palsy patients [33-35]. Liu et al. [33] compared two sides of the face and represented the severity of the paralysis by calculating four ratios. The difference between the affected paralysis side and the unaffected side with various expressions was calculated in [34]. In [35], iris segmentation and LAC-based key point detection were employed to extract the symmetry features in three regions of the face image to quantitatively classify and assess facial paralysis. However, that method was sensitive to the quality of the facial image in the iris extraction. Compared to existing works, our method uses a parallel CNN network structure to extract different level of paralysis features in static images and then use LSTM to extract the sequential motion variation features to evaluate the FNP intensity. The parallel inputs to the CNN are multiple facial areas, which don't require accurate facial landmark detection and reduce the sensitivity of the facial tissue on asymmetry feature learning (e g. facial skin texture, age wrinkles, shapes and position variations of facial organs).

III. METHOD

All training FNP image sequences and normal face frames are from the public available YFP Database [20], [21], and Extended CohnKanade (CK+) Database [22]. The quantitative outcome in [21] is just a regional characteristic, not a comprehensive diagnosis. The detection accuracy of landmark location was directly related to the recognition accuracy of intensity outcomes.

Considering CNN can subsample a high-dimensional image without losing important information, we propose the region-based PHCNN with LSTM (PHCNN-LSTM) for facial palsy severity grading evaluation. Fig. 2 shows a diagram of the proposed method, which includes two distinct stages: the first stage, named static feature extraction, relates to pre-process employing face detection and crucial region segmentation from each frame of 2D image sequences. A region-based PHCNN structure is used in this stage for each image frame to learn multi-level asymmetry feature. The goal of first stage is to obtain a feature vector representation for each frame of a sequence. The second stage, refer to FNP severity grading, consists of feature fusion and a LSTM structure to capture discriminative features of spatio-temporal variation. The output features of each frame are connected to form a feature sequence, and then the time dynamic variations are extracted by LSTM. Based on the appearance and temporal features, the network can provide a given facial image sequence an objective FNP severity level. Images displayed in this paper have been permitted by the owner of the YFP Database.

A. *Face Detection and Image Frame Pre-processing*

Accurate face detection is essential to ensure the face extraction from the database. AdaBoost learning algorithm was used to select critical visual features for the detector [36]. Because only the interested face region is as the input of the CNN, other unnecessary parts of the original images should be removed. Formatting facial images prior to objective FNP assessment may significantly reduce influence of environmental factors to some extent. The input image sequence x_i ($i=1,2,\ \ldots\ldots n$) is pre-processed to obtain standard format with an image size of 128×128×3 pixels. For patients with fewer sequences, the image translating and rotating are used in a certain scale for data augmentation.

B. Landmark Location and ROI Detection

Facial landmarks play a prominent role in face recognition, facial expression analysis and feature-based face registration [37]. A person who has a symptom of Bell's FNP is likely to have deformity symptoms on crucial regions, such as droopy brow, inability to close the eye fully or blink, drooping of the lower eyelid, the corner of the mouth pulls down, inability to pout [38], [39]. Because we don't need precise notation for facial salient points, IntraFace [40] trained on normal faces is used to detect 49 facial landmarks.

Facial deformation can be recognized by extracting the geometric features of facial organs, including shape, profile, position and distance between salient points. Guo, et al. [12] proved that features from mouth region had obvious correlates with HBS score. Barbosa et al. [13] detected iris boundaries as the region-based feature for facial paralysis classification. Plastic surgery can improve the facial symmetry by brow lifting, interventions treatment of eyes and lips regions [39]. In order to reduce the interference of unrepresentative regions, especially the nose region, two ROIs with 40×120×3 pixels, including the eye-brow region and mouth region, are cropped according to the edge landmarks of facial component for ensuring enough facial nerve disfunction information.

C. ROI-based Feature Learning

For a given subject, different facial expression images correspond to a uniform FNP label. The proposed parallel hierarchy structure can learn the spatial feature of regional nerve paralysis, general contour and asymmetry from discontinuous multi-frame images with different facial expressions. A unilateral facial droop symptom not only manifests as surface profile, shape and skewing, but also contains some deeper features in crucial regions. Inspired by VGG-Nets [41], we introduce two types of subnet branch for global and regional areas, named N_1 and N_2. Branch N_1 is for face images and two N_2 branches are for mouth and eye-brow regions. Fig. 3 illustrates the framework of the proposed PHCNN.

As shown in Fig. 3, there are three independent network branches for each image region. Branch N_1 consists of six convolutional layers (double layer with 3×3-16, double layer with 3×3-32 and double layer with 3×3-64), three max pooling layers with 2×2 and one fully connected (FC) layer. Branch N_2 and N_3 both consist of six convolutional layers (double layer with 3×3-8, double layer with 3×3-32 and double layer with 3×3-128), three max pooling layers with 2×2 and three fully connected (FC) layers. The convolutional layers alternate sub-sampling layers for feature extraction from the preprocessed images. These stacks of convolution layers are used to learn multi-level asymmetry from low level features to semantic level features from different images for each subject. Extracted features for each level are flattened to 1024 feature vectors in N_2 and N_3 networks. ReLU is used as an activation function for each subnet. In order to avoid over-fitting problem, a batch normalization layer is added before each hidden layer. Dropout is also applied on the FC layer, which is set as 0.5.

D. Spatio-temporal Feature Fusion

The output feature vectors of each PHCNN are fused in this stage. Spatial features of consecutive frames are connected to a feature sequence, which is fed into LSTM. LSTM is specialized for processing sequential inputs, mapping an input sequence to complex temporal dynamics. LSTM has a three-gate controlled cell state. They are input gates, output gates and forgetting gates, that determines how much information should be passed [42]. The feature vectors obtained by PHCNN-LSTM are used for evaluating the severity of facial paralysis.

Removing the subjects with incomplete face, there are 26 patients in the YFP Database can be used. Because the FNP grading labels of YFP Database are not offered, the clinician should provide the grade score for each subject according to the image sequence.

The grading score provided by a clinician according to HBS is used as the ground truth for training. The results show that distribution of patients with different scores is quite uneven. In this paper, we divided the grade of FNP into three levels marked Ⅰ, Ⅱ, Ⅲ shown as Table I. Level I expressed normal face responds to the grade score 1, Level Ⅱ expressed mild-to-moderate dysfunction responds to the grade scores 2 to 4. Level Ⅲ expressed severe paralysis responds to the grade scores 5 and 6. The concentrated feature sequences from the whole face, eye-brow and mouth regions are fed into LSTM for analyzing the temporal feature and spatial position changes among different facial movements. Due to the imbalanced distribution for each class, we introduce a class-weight coefficient K_j for the loss function in the training process. We calculate the weights for each class. The value of K_j is inversely proportional to the number of class samples, where j is the class index.

As shown in Fig. 3, the static feature vector of the i^{th} frame f_i consists of global and regional features, which is shown as

$$f_i = f_{i\_g} + f_{i\_e} + f_{i\_m} \qquad (1)$$

The feature sequence $F = \{f_1, \cdots, f_N\}$ would be input into LSTM. The memory cells in the LSTM layer will produce a representation sequence for the final FNP severity classification.

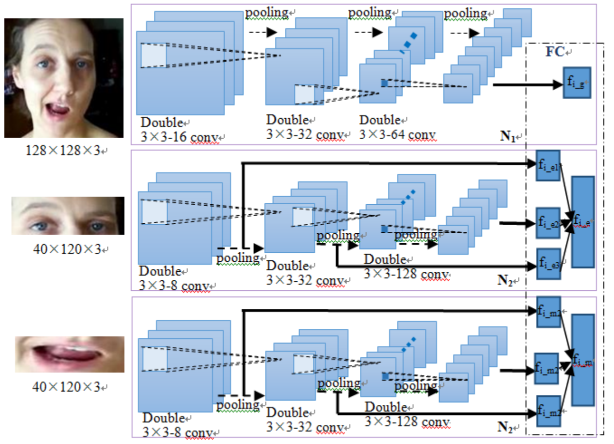

Fig. 3. Framework of the proposed PHCNN

TABLE II
DISTRIBUTIONS OF THE SUBJECTS IN TRAINING

HBS grade score	1	2	3	4	5	6
Patient number	9	0	2	7	15	2
Sequence number	332	0	180	1204	3532	172

E. FNP Severity Grading

A classifier *f* is a function that maps input feature space to the class space.

$$f: X \to \mathbb{R}^c \qquad (2)$$

where, X is the input feature vector for the top FC layer. $\mathbb{R}^c$ is the class space. In this step, the blended feature vector learned from the LSTM with PHCNN is taken by the FC layer for the FNP grade classification for a given image sequence. In this final layer, soft-max and weighted cross entropy (WCE) are the activation function and loss, respectively. A conditional probability is output by each neuron which corresponds to the specific class for each input feature sequence. By using soft-max function, the activation y_i of the i^{th} output neuron is

$$y_i = f(c)_i = \frac{e^{c_i}}{\sum_j^M e^{c_j}} \qquad (3)$$

where, M is the number of output neuron. c_i is the output score inferred by the linear prediction net for the i^{th} class. Each output component is normalized in the interval (0,1). So, each activation output can be considered as a probability value for each class. The general form of the WCE loss is as follows

$$\mathcal{L} = -\sum_i^M K_i \cdot y_i^* \cdot \log{(y_i)} \qquad (4)$$

where, y_i^* is the ground truth which can only be 1 or 0. And y_i is the CNN predicted score for each class in M. To get more precise classification, models are trained with the stochastic gradient descent (SGD) by minimizing the WCE loss.

IV. EXPERIMENTAL EVALUATION

This section presents a thorough experimental procedure of the proposed PHCNN-LSTM framework for FNP grading. All experiments are conducted using Python 3.5 on a computer with system with Windows 10 with a Nvidia GTX 1080 GPU.

A. Facial Palsy Images Collection

The images of YFP Database were captured from Youtube videos which recorded the facial expression variations in conversations. Facial movements of talking include “at rest”, “open mouth”, “closure the eyes lightly”, “Elevation of eyebrows”, “pursing lips”, etc. A professional clinician acquired the front view of the participants’ image sequences to assess the FNP grading with no constraints. The experiment was under normal office fluorescent lighting conditions. All FNP patients were manually labeled according to the HBS. As the video sampling rate of the YFP Database is 6FPS, each facial image is equivalent to the duration of 1/6 second leading to a big gap between consecutive images in the sequence. We removed the incoherent and blurred images in the pre-processing stage. The result shows that both the grade score distribution and frame number for each subject are imbalanced enormously. Patients with grade 5 are much more than patients with grade 3 and 6. And the image number for each subject in our experiment various enormously. For example, there are only 55 images corresponding to patient 8 and 2664 images for patient 21. Due to the limitations mentioned above, we segment the long sequence into short one with 4 consecutive frames. For the patient whose sequence number is less than 16, we apply the translation and rotation operations for data augmentation. The label score distribution and sequence number for each grade are shown in Table Ⅱ.

TABLE III
DATA SUBSETS USED IN THE CLASSIFICATION FOR 5-FOLD CROSS-VALIDATION EXPERIMENTS

Subset	Training			Validation		
No.	level- Ⅰ	level- Ⅱ	level-Ⅲ	level- Ⅰ	level- Ⅱ	level-Ⅲ
1	304	1256	2880	28	128	824
2	298	490	3550	34	894	154
3	266	1316	2934	66	68	770
4	180	1022	3596	152	362	108
5	280	1384	1856	52	0	1848

Apart from 26 patients (5088 sequences in total) from YFP Database, 9 normal subjects (332 images sequences in total) from S022, S037, S045, S050, S077, S102, S105, S124 and S130 folders of CK+ Database are considered as normal face in our experiments for increasing the diversity of normal faces. The study protocol was reviewed and approved by the ethics committee at our institution (ER Number: ETHIC-2019-799).

B. Experiment Set-up

To test the model accuracy for the multi-classification tasks, K-fold cross-validation protocol is adopted in our experiment, where the k value is chosen as 5. This is to allow for the testing on unseen FNP images thus reducing the possibility of overfitting to previously seen images. The 5-fold cross-validation method divides the dataset into 5 subsets. Each single subset is retained as validation data, and the other 4 subsets are used as training data. It can ensure the test data is untouched in each session of experiment. The experiment is repeated 5 times, and each subject has the same probability for validation. In order to reduce the processing time, all the original facial images are cropped to include only the face region as shown in Fig. 2. Table Ⅲ displays the number of facial image sequences for each grade level for 5-fold cross-validation experiment.

We adopt a batch-based SGD method to optimize the model. SGD is an iterative optimizer, which has a parameter called learning rate (LR) to achieve the loss function minimum. This parameter determines the current value of the weights in each updating process. The base LR is set as 0.0001 and is reduced by polynomial policy with gramma of 0.1. The momentum is equal to 0.9 and the weight decay is set as 0.00002. During the training process, we set the actual batch size as 64 for each session. Finally, the best model is chosen after 50 epochs to evaluate the network performance in classifying the severity grade of FNP frame sequences.

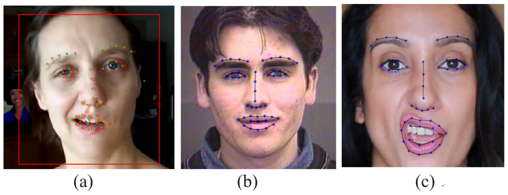

Fig. 4 Facial recognition and 49 facial landmarks detection [20-22]. (a) Face detection and landmarks coordinates of FNP from YFP database; (b) Landmarks of normal face from CK+ database; (c) Landmarks of patient with FNP from YFP database

C. Result and discussion

The proposed method can solve two problems: (1) discrimination of FNP patients from normal faces; (2) severity grade classification of FNP. The results of facial recognition and 49 landmarks detection are shown in Fig. 4. We separate out the eye-brow region and mouth region following this detection results. These images of sequences are then fed into the PHCNN for asymmetry static features.

These 49 points landmarks shown as Fig. 4 (a) include 10 marks for eyebrow, twelve marks for two eyes, 9 marks for nose and 18 marks for mouth lip. The landmarks of left and right inner corners for mouth are not available. Fig. 4 (b) and (c) show the facial landmarks for normal face and FNP respectively. We can see that the landmarks on the mouth lip are not very accurate for patient with FNP. If a given facial image has a low quality caused by illumination or the facial feature outline is not obvious, the detection results may be worse. Fig. 5 displays the two ROIs cropped from the facial palsy image, which are the eye-brow region and mouth region. In the model training procedure, face image, eye-brow image and mouth image are as the inputs of the proposed parallel subnets.

To illustrate the prediction ability of the proposed method with parallel input, we have compared with several combination of networks with LSTM. The first is N₁ subnet structure (FACE-N₁) combined with LSTM, which only focuses on the whole face. The second is LSTM combined Visual Geometry Group Network with 16 layers (LSTM-VGG16). The third is Squeeze-and-Excitation Networks (SENet) and the fourth is the PHCNN. The final one is the LSTM itself only. We use pre-processed facial images as the inputs of SENet and PHCNN. The other methods have the same image sequence inputs.

The performance metrics like accuracy, precision, recall and F1 score are utilized for assessing the performance of the proposed methodology, which are defined as follows:

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad (5)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (6)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (7)$$

$$\text{F1 Score} = \frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision+Recall}} \quad (8)$$

where, TP stands for the positive sequence samples that the model predicted as positive. FN stands for positive sequence samples that the model predicted negative wrongly. FP stands for the negative sample that the model predicted positive. TN stands for the negative sample that the model predicted negative.

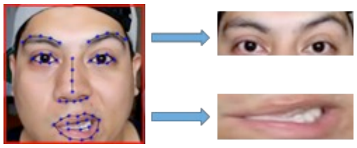

Fig. 5 Image segmentation of facial ROIs from YFP database [20], [21]

The classification accuracy compared with different network structures in 5-fold cross-validation experiment is shown in Fig. 6. The accuracy of a model shows the percentage of correctly predicted samples. It is found that the proposed region-based network model has achieved excellent performance in FNP recognition and classification. The recognition rates on 5-fold cross-validation experiment with PHCNN-LSTM are 91.02%, 93.98%, 95.13%, 96.14%, 97.79%.

It is shown that spatio-temporal features extracted by a recurrent neural network combined with a CNN architecture could provide dynamic detailed variation of facial critical areas. Th proposed PHCNN-LSTM has the highest classification accuracy in 5-fold cross-validation experiment, that is 94.81% on average. The accuracy of LSTM-FACE-N₁ is also over 90%, but lower than PHCNN-LSTM. That is, region-based input uses hierarchical subnets to learn asymmetry features from the low level to a semantic level, showing superior classification accuracy over simple face input. Then the LSTM can capture the dynamic changes contained in the feature sequences composed of these asymmetric features. SENet and PHCNN which only focus the facial spatial features and ignore the movement variation have moderate classification accuracies. The results show that it is difficult to learn the facial asymmetry features only from the still image. Besides, PHCNN architecture has the advantage of removing unrepresentative regions with redundant information and effectively reducing interference caused by age wrinkles and the variations of facial organ's position. Due to the ability to explore the co-occurrence relationship between spatial and temporal domains, our proposed method outperforms by 6.9% than PHCNN. The LSTM method just uses pixel vectors of the image sequence, which ignores the appearance feature of FNP, such as texture, contour and shape. It has the lowest accuracy.

Table Ⅳ presents the average performance of these methods. The precision denotes the fraction of correctly predicted positive samples from the positive predicted samples, and the recall means the ratio of real positive samples that are truly discovered by the model. F1 is the harmonic mean of precision and recall, which is a meaningful criterion. All these criteria are over 90% with the proposed method. From the assessment, it indicates that the proposed network structure achieves the best recognition rate when measured with the existing methods.

Each parallel subnet in PHCNN-LSTM is a shallow network structure with a low number of parameters and hidden layers, which can speed up the model training procedure

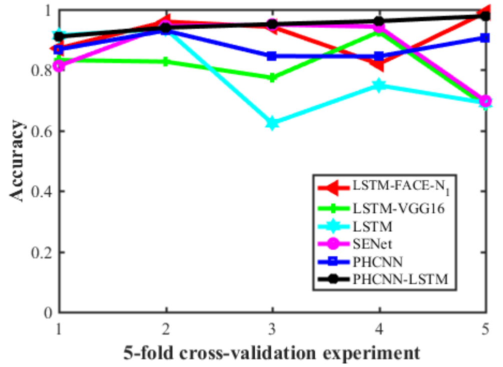

Fig. 6 Accuracy comparison of different networks in 5-fold cross-validation experiment

compared with other existing networks. We compared the training time among these methods, which are shown in Table Ⅳ. LSTM has the shortest training time due to lack of feature extraction process. LSTM-FACE-N_1 also takes less training time with lower robustness due to the only one stage network. SENet and LSTM-VGG16 have more training time because of the complex network architecture. PHCNN-LSTM can achieve higher recognition rate with a moderate amount of time. Compared with current deep learning network structures, the proposed hierarchical model in this paper is more effective and robust.

V. Conclusion

In this paper a ROI-based hierarchy network combining LSTM model is presented to classify FNP image sequences into 3 severity levels based on spatio-temporal variation of salient facial regions. Based on the hierarchy CNN network branches, this method learns automatically ROI features, including the low-level contour and shape characteristics and higher semantic features, reducing the interference of age wrinkles and unrepresentative organs with shape and position variations, such as the nose. The proposed method can also distinguish the difference between normal faces and faces carrying FNP, showing superior consistency and robustness in FNP classification on a large database.

The current research still has some limitations. Due to ethical and personal privacy reasons, there are few public FNP databases available for research. The uneven distribution of the patient's sequence length with different grading benchmarked by HBS and lack of various facial expressions limit the dynamic features learning of muscle movement variation and the optimization of the network. Additionally, the results are promising as there are many areas in which this research can be taken forward. Firstly, if we have enough FNP data, there is a potential to investigate the effect of different sequence length on model accuracy. Secondly, the asymmetry quantization of the facial paralysis area and the FNP image synthesis method based on the Facial Action Unit for data augmentation can be considered as future research problems.

TABLE Ⅳ
Comparison of Proposed Method with State of the Art on the Criteria of Accuracy, Precision, Recall and F1 Score in 5-fold Cross-validation Experiment

Method	Accuracy	Precision	Recall	F1	Training time
LSTM-FACE-N_1	0.9167	0.896	0.916	0.902	1.94h
LSTM-VGG16	0.8086	0.892	0.808	0.838	7.86h
LSTM	0.7829	0.92	0.782	0.83	1.24h
SENet	0.8698	0.94	0.87	0.894	16.68h
PHCNN	0.8791	0.896	0.88	0.88	2.96h
PHCNN-LSTM	**0.9481**	**0.956**	**0.948**	**0.942**	**3.51h**

References

[1] A. Y. Fattah et al., "Facial nerve grading instruments: systematic review of the literature and suggestion for uniformity," *Plastic and Reconstructive Surgery*, vol. 135, no. 2, pp. 569-579, Feb. 2015.

[2] D. Jayatilake, T. Isezaki, Y. Teramoto, K. Eguchi and K. Suzuki, "Robot Assisted Physiotherapy to Support Rehabilitation of Facial Paralysis," *IEEE Transactions on Neural Systems and Rehabilitation Engineering,"* vol. 22, no. 3, pp. 644-653, May 2014.

[3] J. Lou, H. Yu and F. Wang, "A Review on Automated Facial Nerve Function Assessment from Visual Face Capture," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 28, no. 2, pp. 488-497, Feb. 2020.

[4] J. Tavares-Brito, M. van Veen, J. Dusseldorp, F. Fayez Bahmad Jr and T. Hadlock, "Facial Palsy-Specific Quality of Life in 920 Patients: Correlation with Clinician-Graded Severity and Predicting Factors," *Laryngoscope*, vol.129, pp. 100-104, Jan. 2019.

[5] A. Joseph, J. Kim. "Management of Flaccid Facial Paralysis of Less Than Two Years' Duration," *Otolaryngologic Clinics of North America*, vol. 51, no. 6, pp.1093-1105, 2018.

[6] W. Samsudin, K. Sundaraj, "Review: evaluation and grading systems of facial paralysis for facial rehabilitation," *Journal of Physical Therapy Science*, vol. 25, no. 4, pp. 515-519, 2013.

[7] W. Samsudin, K. Sundaraj, "Clinical and non-clinical initial assessment of facial nerve paralysis: A qualitative review. Biocybernetics and Biomedical Engineering," vol. 34, no. 2, pp. 71-78, 2014.

[8] J. W. House and D. E. Brackmann. "Facial nerve grading system," *Otolaryngol Head Neck Surgery*, vol. 93, no. 2, pp. 146-147,1985.

[9] B. G. Ross, G. Fradet and J.M. Nedzelski. "Development of a sensitive clinical facial grading system". *Otolaryngol Head Neck Surgery,* vol. 114, no. 3, pp. 380-386, March 1996.

[10] R. Niziol, F. Henry, J. Leckenby and A. Grobbelaar. "Is there an ideal outcome scoring system for facial reanimation surgery? A review of current methods and suggestions for future publications," *Journal of Plastic, Reconstructive & Aesthetic Surgery*, vol. 68, pp. 447-456, 2015.

[11] I. Song, J. Vong, N. Yen J. Diederich and P. Yellowlees. "Profiling Bell's Palsy Based on House-Brackmann Score," *JAISCR*, vol. 3, no. 1, pp.41-50, 2013.

[12] Z. Guo et al. "An Unobtrusive Computerized Assessment Framework for Unilateral Peripheral Facial Paralysis," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 3, pp.835-841, 2018.

[13] J. Barbosa, W. K. Seo and J. Kang. "paraFaceTest: an ensemble of regression tree-based facial features extraction for efficient facial paralysis classification," *BMC Medical Imaging*, 19, 30, 2019.

[14] D. Gibelli, F. Tarabbia, S. Restelli et al. "Three-dimensional superimposition for patients with facial palsy: an innovative method for assessing the success of facial reanimation procedures," *British Journal of Oral and Maxillofacial Surgery*. vol. 56, no. 1, pp. 3-7, January 2018.

[15] M. Alagha, X. Ju, S. Morley, A. Ayoub, "Reproducibility of the dynamics of facial expressions in unilateral facial palsy," *International Journal of Oral and Maxillofacial Surgery*, vol. 47, no. 2, pp.268-275, February 2018.

[16] T. Wang, S. Zhang, L. Liu, G. Wu and J. Dong. "Automatic Facial Paralysis Evaluation Augmented by a Cascaded Encoder Network Structure," *IEEE Access.*, vol. 7, pp. 135621-135631, Oct., 2019.

[17] A. Zhao, L. Qi, J. Li, J. Dong and H. Yu. "A Hybrid Spatio-temporal Model for Detection and Severity Rating of Parkinson's Disease from Gait Data," *Neurocomputing*, vol. 315, pp. 1-8, Nov. 2018.

[18] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger and H.

Greenspan. “GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification,” *Neurocomputing*, vol. 321, no.10, pp. 321-331, Dec. 2018.

[19] U. Côté-Allard et al. “Deep Learning for Electromyographic Hand Gesture Signal Classification Using Transfer Learning,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 4, pp.760-771, April 2019.

[20] https://sites.google.com/view/yfp-database

[21] G. Hsu, J. Kang and W. Huang. “Deep Hierarchical Network with Line Segment Learning for Quantitative Analysis of Facial Palsy,” *IEEE Access*, vol.7, pp. 4833-4842, Dec. 2018.

[22] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. “The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression,” in Proceedings *IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, pp. 94-101, 2016.

[23] H. Kim, S. Kim, Y. Kim and K. Park. “A Smartphone-Based Automatic Diagnosis System for Facial Nerve Palsy,” *Sensors*, vol. 15, no, 10, pp. 26756-26768, Oct. 2015.

[24] L. Modersohn and J. Denzler. “Facial Paresis Index Prediction by Exploiting Active Appearance Models for Compact Discriminative Features,” in Proceedings *the 11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, vol. 4, pp. 271-278, 2016.

[25] T. Wang, S. Zhang, J. Dong, L. Liu and H. Yu. “Automatic evaluation of the degree of facial nerve paralysis”, *Multimedia Tools and Applications*, vol.7, no. 19, pp. 11893-11908, Oct. 2016.

[26] H. Yoshihara, M. Seo, T. Ngo, N. Matsushiro and Y. Chen. “Automatic Feature Point Detection Using Deep Convolutional Networks for Quantitative Evaluation of Facial Paralysis,” in Proceedings *9th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics*, pp. 811-814, 2016.

[27] A. Song, Z. Wu, X. Ding,Q. Hu and X. Di. “Neurologist Standard Classification of Facial Nerve Paralysis with Deep Neural Networks,” *Future Internet*, vol. 10, no. 11:111, Nov. 2018. doi:10.3390/fi10110111.

[28] M. Sajid, T. Shafique, M. Baig, I. Riaz, S. Amin, and S. Manzoor. “Automatic Grading of Palsy Using Asymmetrical Facial Features: A Study Complemented by New Solutions,” *Symmetry*, vol. 10, no. 7, 242, 2018. https://doi.org/10.3390/sym10070242

[29] Z. Yu, G. liu, Q. Liu and J. Deng. “Spatio-temporal convolutional features with nested LSTM for facial expression recognition,” *Neurocomputing*, vol. 317, pp. 50-57, 2018.

[30] X. Ouyang, S. Xu, C. Zhang. et al., “A 3D-CNN and LSTM based Multi-task Learning Architecture for Action Recognition,” *IEEE Access*, vol. 7, pp. 40757-40770, March, 2019.

[31] H. Zhang, B. Huang, G. Tian, “Facial expression recognition based on deep convolution long short-term memory networks of double-channel weighted mixture,” *Pattern Recognition Letters*, vol. 131, pp. 128-134, 2020.

[32] G. Storey, R. Jiang, S. Keogh et al., “3DPalsyNet: A Facial Palsy Grading and Motion Recognition Framework using Fully 3D Convolutional Neural Networks,” *IEEE Access*, 2019, arXiv: 1905.13607.

[33] L. Liu, G. Cheng, J. Dong et al. “Evaluation of facial paralysis degree based on regions,” in Proceedings *the Third International Conference on Knowledge Discovery and Data Mining*, pp. 514-517, 2010.

[34] K. Anguraj and S.Padma. “Evaluation and Severity Classification of Facial Paralysis using Salient Point Selection Algorithm,” *International Journal of Computer Applications*, vol. 123, no.7, pp. 23-29, Aug. 2015.

[35] J. Barbosa, K. Lee, S. Lee, B. Lodhi, J. Cho, W. Seo and J. Kang. “Efficient quantitative assessment of facial paralysis using iris segmentation and active contour-based key points detection with hybrid classifier,” *BMC Medical Imaging*, vol.16, 23, March, 2016.

[36] P. Viola and M. Jones. “Robust real-time face detection” *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137-154, May 2004.

[37] J. Lou, X. Cai, Y. Wang, H. Yu and S. Canavan, “Multi-subspace supervised descent method for robust face alignment,” *Multimedia Tools and Applications*. vol. 78, 35455, 2019. [Online] Available: https://doi.org/10.1007/s11042-019-08129-4.

[38] W. Samsudin, R. Samad, M. Ahmad and K. Sundaraj. “Forehead Lesion Score for Facial Nerve Paralysis Evaluation,” in Proceedings *IEEE International Conference on Automatic Control and Intelligent Systems*, pp. 102-107, June, 2019.

[39] M. Lafer, M. Teresa, “Management of Long-Standing Flaccid Facial Palsy Static Approaches to the Brow, Midface, and Lower Lip,” *Otolaryngologic Clinics of North America*, vol. 51, no. 6, pp. 1141-1150, 2018.

[40] F.Torre, W. Chu, X. Xiong, F. Vicente, X. Ding and J. Cohn. “IntraFace,” in Proceedings *11th IEEE International Conference and workshops on Automatic Face and Gesture Recognition*, pp. 1-30, June, 2016.

[41] K. Simonyan, A. Zisserman, “Very deep convolutional networks for large-scale image recognition”, arXiv: 1409.1556 [cs.CV]

[42] C. Si, W. Chen, W. Wang, et al., “An Attention Enhanced Graph Convolutional LSTM Network for Skeleton-Based Action Recognition,” in Proceedings *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. DOI: 10.1109/CVPR.2019.00132.