IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 1, 2018, accepted November 17, 2018, date of publication December 4, 2018,
date of current version January 11, 2019.

Digital Object Identifier 10.1109/ACCESS.2018.2884969

Deep Hierarchical Network With Line Segment Learning for Quantitative Analysis of Facial Palsy

GEE-SERN JISON HSU[1], (Senior Member, IEEE), JIUNN-HORNG KANG[2,3], AND WEN-FONG HUANG[1], (Member, IEEE)

[1]Artificial Vision Laboratory, National Taiwan University of Science and Technology, Taipei 10607, Taiwan
[2]Department of Physical Medicine and Rehabilitation, School of Medicine, College of Medicine, Taipei Medical University, Taipei 11031, Taiwan
[3]Department of Physical Medicine and Rehabilitation, Taipei Medical University Hospital, Taipei 11031, Taiwan

Corresponding author: Jiunn-Horng Kang (jhk@tmu.edu.tw)

This work was supported in part by the Taiwan Ministry of Science and Technology (MOST) under Grant MOST 106-2221-E-011-144 and in part by the Joint Grant from Taipei Medical University and the National Taiwan University of Science and Technology under Grant 106-TMU-NTUST-106-05.

ABSTRACT We propose the deep hierarchical network (DHN) for the quantitative analysis of facial palsy. Facial palsy, also known as Bell's palsy, is the most common type of facial nerve palsy that results in the loss of muscle control in the affected facial regions. Typical symptoms include facial deformity and facial expression dysfunction. To the best of our best knowledge, all approaches for the automatic detection of facial palsy consider hand-crafted features. This paper reports the first deep-learning-based approach developed for the real-time quantitative analysis of facial palsy. The proposed DHN consists of three component networks: the first detects the subject's face, the second detects the facial landmarks and line segments on the detected face, and the third detects the local palsy regions. The first component network is built on the YOLO2 detector. The second component network is developed on a fused network architecture that incorporates a line segment learning network for locating the facial landmarks and line segments. The third component network is developed on an object detection network with the line-segment-embedded input that combines the landmarked region and the line segments detected by the second component network. The novelties of this research include: 1) the modification of a state-of-the-art edge detector for extracting the facial line segments; 2) the embedding of the line segment learning for the detection of facial landmarks and local palsy regions; 3) the quantitative description of the facial palsy syndrome intensity; and 4) the release of the first clinically labeled database, the YouTube Facial Palsy (YFP) database. The making of the YFP database solves the issue that previous methods were all evaluated on proprietary databases, making the comparison of different methods extremely difficult. The YFP database includes 32 videos of 21 patients collected from YouTube and labeled by clinic specialists. To enhance the robustness against facial expression variations, we include the CK+ facial expression database in the training. We show that the proposed DHN not only just detects the local palsy regions but also captures the intensity of the facial palsy syndrome over time, enabling the quantitative description of the syndrome. The experiments show that the proposed approach offers an accurate and efficient real-time solution for facial palsy analysis.

INDEX TERMS Facial palsy, bell palsy, medical image diagnosis, face alignment, face recognition.

I. INTRODUCTION

Facial palsy, also known as Bell's palsy, is a common type of facial palsy. Typical symptoms include drooping, stiffness or loss of control on the affected side of the face. Facial palsy not only significantly affects the facial appearance but also leads to impaired feeding functions and adverse psychosocial consequences [11]. The diagnosis of facial palsy usually relies on the visual inspection of facial symmetry and expression dysfunction by clinicians. Manual visual inspection has some disadvantages, for example, it is difficult to quantify the intensity and variation of the symptoms, difficult to track the symptom changes between clinic visits and difficult to compare the symptoms across patients in a quantitative way. Automatic inspection by using a camera can circumvent

these disadvantages. The approaches for automatic inspection of facial palsy have been emerging in recent years. A brief review on recent approaches is given in Sec. II. However, almost all approaches up-to-date consider handcrafted features, the deep-learning based approaches are yet to be developed. Another serious issue with the previous approaches is that their experiments were performed on proprietary databases, making benchmarking and performance comparison difficult.

Our proposed framework, called the Deep Hierarchical Network (DHN), consists of three deep convolutional neural networks (CNNs). The first CNN, called the Face Net, is for detecting the patient's face; the second CNN, called the Landmark Net, is for locating the facial landmarks on the detected face; and the third CNN, called the Region Net, is for locating the local palsy regions on the detected face. The Face Net is built on the state-of-the-art YOLO-3 network. The Landmark Net is developed with a line segment learning architecture that we explore in this study. The Region Net is developed on a relatively shallow CNN embedded with the line segments learned by the Landmark Net for fast target region detection. Given an image to the proposed DHN, the Face Net first detects the face, then the Landmark Net locates the landmarks on the detected face, and then the Region Net locates the local palsy regions with an intensity score computed based on cross-entropy loss of the network output. Our experiments were performed on the YouTube Facial Palsy (YFP) database, which is the first public database that we made for the study on the visual inspection of facial palsy symptoms. It contains 32 video clips of 22 facial palsy patients collected from YouTube. We convert the videos into sequences of images, and have the images labeled by facial palsy clinicians. As the number of patients is limited, we adopt the Leave-One-Person-Out (LOPO) protocol for performance evaluation.

This paper is an extension of our conference paper in [9], where we proposed a hybrid network with three component networks. The differences and advancements made in this extended version can be summarized as follows:

1) In the previous version [9], we used the HourGlass architecture developed by Bulat and Tzimiropoulos [2] for the Landmark Net to locate the facial landmarks. In the current version, we develop our own algorithm using a network embedded with line segment learning, which improves both the landmark localization and local palsy region detection.
2) In the previous [9], we used the Darknet framework [17] as the Region Net to locate the local palsy regions. In this extended version, we develop a different network architecture with the line segments as an important clue to locate the local palsy regions.
3) In the previous [9], we computed the frequency of the appearance of the detected palsy regions as the intensity. In this extended version, we explore the softmax probability at the output layer as the better representation of the symptom intensity.

The contributions made in this study are threefold:

1) A pioneering deep learning approach is proposed for facial palsy analysis which can accurately detect the local palsy regions and interpret the intensity of the symptom over time;
2) Incorporation of line segment learning into a deep learning framework is verified effective for the improvement of the palsy region localization and intensity estimation;
3) The first public facial palsy database, the YouTube Facial Palsy (YFP) database, is released which is composed of videos collected from YouTube and labeled by medical specialists.

In the following sections, we first present a review on the previous work in Sec. II. The development of our proposed framework is elaborated in Sec. III. The introduction to the YFP database is given in Sec. IV-B along with our experiments to study the performance of the proposed DHN approach. The conclusion of this study is given in Sec. VI.

II. RELATED WORK

Several approaches for automatic detection and diagnosis of facial palsy have been proposed in recent years. According to our survey, all of the approaches exploit handcrafted features and classifiers, with experimental results reported on proprietary databases. We select a few latest studies and summarize their methods and experiments in this section.

To conduct an objective and quantitative analysis of facial palsy, Ngo *et al.* proposed an approach based on the limited-orientation modified circular Gabor filters (LO-MCGFs) [15]. The LO-MCGFs employs uniform passbands to remove noises and enhance the desired spatial frequencies, and uses the bounded filter support to specify the region of interest. These virtues make the approach effective for extracting the facial asymmetry features. The facial dataset considered in their study is composed of image sequences of 75 facial palsy patients and 10 participants without facial palsy, made by the Osaka Police Hospital. Each image sequence is composed of 60 still images taken from the same subject and the intensity in each image is scored into 3 levels, strong, median and weak. As it is a proprietary dataset, it is not known whether the images are from continuous frames, and how the intensity level is assigned.

Kim *et al.* [10] propose a smartphone-based automatic diagnosis system that consists of three modules, namely a facial landmark detector, a feature extractor and a classifier. The incremental face alignment, proposed by Asthana *et al.* [1], is used for detecting the facial landmarks. Given the facial landmarks, they compute the asymmetric index using the displacement of shape landmark sets that correspond to the eyebrows and mouth regions while the subjects change their expressions. To extract the asymmetric index, the forehead and eye regions are considered in heuristic approaches that measure the displacements and ratios of different distances. The Linear Discriminant Analysis (LDA) and Support Vector Machine (SVM) are then employed for

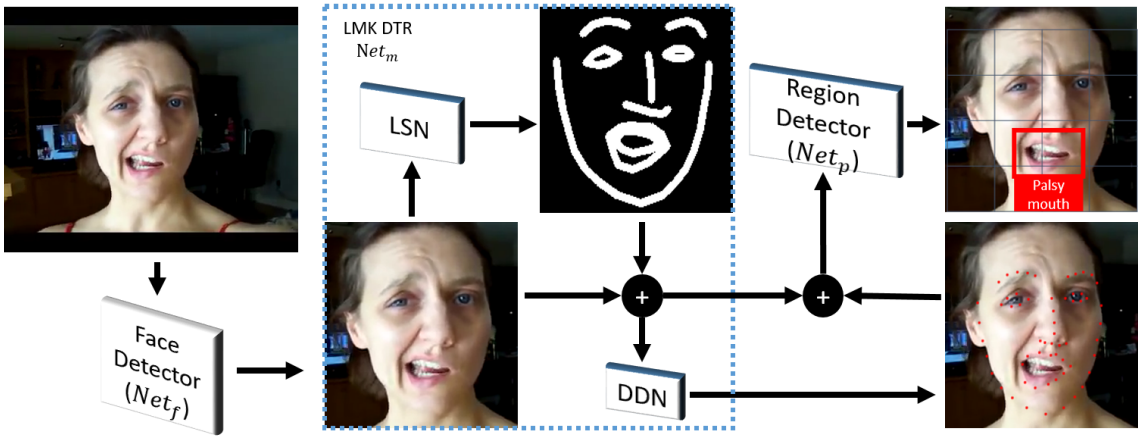

FIGURE 1. The proposed Deep Hybrid Network (DHN) is composed of a face detector Net_f, a facial landmark locator Net_m (denoted as LMK DTR) and the palsy region detector Net_p. Net_m is composed of a line segment detector LSN and a double dropout network (DDN) for landmark localization.

classification. The system is evaluated on a private database with 23 facial palsy patients and 13 volunteers without facial palsy.

A multiresolution local binary patterns (LBPs) is proposed in [5] to characterize the local and global region patterns for the analysis of facial palsy. As the facial landmark localization was not feasible then, the authors used the initial frame as the reference frame to locate four pairs of local regions in the consecutive frames and detect five apex frames by image subtraction. The asymmetry across the face is justified based on the features extracted from the temporalâĂ"spatial domain in each local region. The features are enhanced by a block processing scheme. The symmetry of facial movements is measured by the resistor-average distance (RAD) between the features extracted across the face. The SVM is used to provide quantitative evaluation of the facial palsy symptom. Their method is validated by experiments on 197 videos. No information is provided about the numbers of patients and normal subjects in the videos.

A quantitative approach that considers both the static facial asymmetry and the speed of appearance change is proposed by Wang *et al.* [23]. They first trained an ASM (Active Shape Model) [23] for locating the facial landmarks on each patient's face. The landmarks are used to segment the face into 8 regions, and the facial asymmetry is computed based on the distances between landmarks within each region and across corresponding regions. The static facial asymmetry is computed by the localization of local deformations, the extraction of asymmetric distances and the quantification of bilateral asymmetry. They use the SVM with RBF kernel to classify the degrees of facial palsy in different facial movements, and evaluate the performance on a proprietary database with 62 patients.

In summary, these methods highlight the progress made up to date on the automatic detection and analysis of facial palsy with the following aspects: 1) All previous approaches consider handcrafted features and classifiers; 2) Facial asymmetry is the core character to identify for facial palsy; 3) The databases used in the previous studies are proprietary, making performance comparison extremely difficult.

III. DEEP HIERARCHICAL NETWORK

We formulate the facial palsy identification as a region detection problem, and consider the facial-palsy-caused deformation regions, or simply the *palsy regions*, on a patient's face as the target regions to locate. Our proposed solution is the Deep Hierarchical Network (DHN), which is composed of three component networks. The first component network is a face detector, denoted as Net_f; the second component network is a facial landmark locator, denoted as Net_m; and the third component network is a facial palsy region detector, denoted as Net_p. Figure 1 shows the overall configuration of the proposed DHN with the component networks and the outputs from each component network. Given an image, Net_f first detects the face, then Net_m locates the landmarks on the face, and then Net_p locates the facial palsy regions. The landmark detector Net_m has two sub-networks, the Line Segment Network (LSN) and the Double Dropout Network (DDN). The former detects the facial line segments, and the latter locates the facial landmarks using the detected face image and the associated facial line segments as the fused input. The fused input is also used by the region detector Net_p for locating the local palsy regions. The three component networks, Net_f, Net_m and Net_p, are elaborated in the following sections.

A. FACE DETECTION

The Face Detector Net_f is built on the pretrained YOLO2 and retrained on the Wider Face database [25]. The YOLO2, also known as YOLO-9000, proposed by Redmon and Farhadi, is a state-of-the-art real-time object detector [17]. It reports 76.8 mAP (mean Average Precision) on the benchmark VOC 2007 (the Pascal Visual Object Classes Challenge) at processing speed 67 FPS, and 78.6 mAP at 40 FPS, outperforming many state-of-the-art approaches, including the Faster RCNN with ResNet [19] and the SSD [12]. For face detection, we train the YOLO2 using the WIDER FACE database [25],

which offers 393,703 labeled faces in 32,203 images with a large variation in pose, illumination, expression, scale and occlusion. Following the data partition specified in [25], the WIDER FACE is split into a training and validation set with 199k faces in 16,106 images and a test set with 194k faces in 16,097 images. We change default settings of YOLO2, including the partition of input image into a grid of 11×11 cells, each cell associated with 2 bounding boxes for prediction, and only one class (face) is considered. Compared with other contemporary approaches on the benchmark AFW database, Net_f achieves AP (Average Precision) 99.25% on AFW benchmark, better than the DPM (97.2%) [20], the HeadHunter (97.1%) [14], SSD-512 (98.6%) [12] and the Faster RCNN (95.3%) [19]. Note that the Faster RCNN and SSD are proposed for object detection, we tailored them for face detection the same way as we did for Net_f.

B. FACIAL LANDMARK LOCALIZATION

The Landmark Locator Net_m is composed of 1) the Line Segment Network (LSN) for estimating the line segments that connect the facial landmarks and 2) the Double-Dropout Network (DDN) for locating the facial landmarks.

1) LINE-SEGMENT NETWORK

Many facial landmarks are located on the edges of the whole face and of the facial components, e.g., eyes, nose and mouth. The Line Segment Network (LSN) is proposed to take the advantage of this observation by detecting the line segments that connect all of the landmarks. The line segments are obtained by connecting the neighboring landmarks in each training image, generating a target line segment map for learning. Given a training face sample U_k with landmarks labeled, we connect the neighboring landmarks that follow the shapes of the facial components and the whole face, and generate V_i, the binary image of the facial line segments. This can be readily done as most databases offer landmarks numbered in a predefined order. This way of forming a landmarked line segment map can be applied to the whole training set and end up with the image pair set $[U_k, V_k]_{k=1,\cdots,K_u}$, i.e., each training sample is composed of an image and a landmarked facial line segment image. Given $[U_k, V_k]_k$, the LSN is designed to take U_k as input and generate $\hat{V}_k$ as output, so that the error between $\hat{V}_k$ and V_k is minimized by the end-to-end training.

The design of the LSN network considers the state-of-the-art RCF (Richer Convolutional Feature) network [13] as the base net, and improves it with shallower convolution layers for better scaled and leveled features. The shallower convolution layers account for the fact that our targets, including eyes, mouth and the whole face, are in the same scale as the input is a face cropped by the face detector, instead of the multi-scaled objects considered in the general edge detection.

The proposed LSN is structured as that shown in Figure 2. As the RCF network is the base of the LSN, and the VGG-16 network [16] is the base of the RCF, the LSN can be well

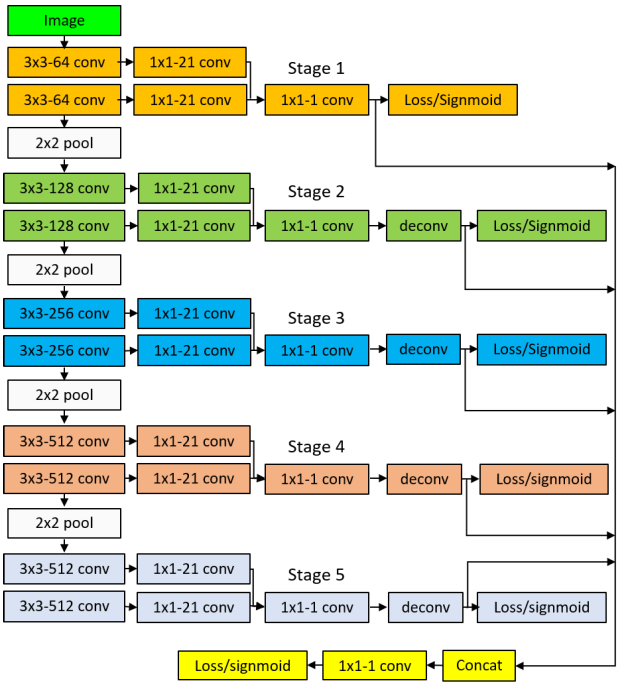

FIGURE 2. Network structure and parameter settings of the proposed Line Segment Network (LSN), modified from VGG-16 with all fully-connected layers removed and all convolution blocks made of two convolution layers.

explained by looking into the architecture of the VGG-16. The VGG-16 consists of two double-convolution blocks, three triple-convolution blocks and three fully-connected layers. The fully-connected layers are all removed in the RCF, which keeps the five convolution blocks with 13 convolution layers. The convolution layers are commonly denoted as conv-1-1, conv-1-2, conv-2-1, ... conv-5-2 and conv-5-3, where conv-*i*-*k* denotes the *k*-th convolution layer at Block-*i*. A pooling layer with 2 × 2 window is implemented between the convolution blocks.

The modifications made on the VGG-16 include the following:

- Each conv layer is connected to a conv layer with kernel size 1 × 1 and channel depth 21 (denoted as 1 × 1-21). The resulting layers in each block are accumulated using an *eltwise* layer to form hybrid features.
- Each eltwise layer is connected to a 1 × 1-1 conv layer, followed by a deconv layer for feature map upsampling. The deconv layer is connected to a sigmoid layer for minimizing the cross-entropy loss from the target.
- All upsampling layers are concatenated and followed by a 1 × 1-1 conv layer for fusing the feature maps from each block. The fused feature is connected to a sigmoid layer for minimizing the cross-entropy loss.
- The above three follow the RCF setups [13]. In addition, we further modify the network to keep two convolution layers in each block as the line segment features are of high spatial frequencies, which will be weakened by deeper convolution.

Given the LSN configuration shown in Figure 2, the following loss is computed at each pixel with respect to the pixel label as

$$L(X_i; W) = \begin{cases} \alpha \cdot log(1 - P(X_i); W), & \text{if } y_i = 0 \\ 0 & \text{if } 0 < y_i \leq \eta \\ \beta \cdot \log P(X_i; W), & otherwise \end{cases}$$

$$\text{where} \quad \alpha = \gamma \cdot \frac{|Y^+|}{|Y^+| + |Y^-|}, \quad \beta = \frac{|Y^-|}{|Y^+| + |Y^-|} \quad (1)$$

Y^+ and Y^- denote the set of line segment pixels (or positive pixels) and the set of background pixels (or negative pixels), respectively. The hyper-parameter γ is a weight coefficient chosen to balance the positive and negative sets. The activation value fired by the network and the ground-truth line segment probability at pixel i are presented by X_i and y_i, respectively. As the VGG-16 is employed as the backbone network, $P(\cdot)$ is the sigmoid function, and W denotes the network parameters to be learned from training.

Summing up the above loss from each convolution block and the fused loss contributed by all convolution blocks, the total loss considered in the proposed framework can be written as follows

$$L_T = \sum_{i=1}^{|I|} \left(\sum_{k=1}^{K} L(X_i^k; W) + L(X_i^{fuse}; W) \right) \quad (2)$$

where $|I|$ is the number of pixels in image I, K is the number of convolution blocks (5, in this case), X_i^k is the activation value from block k, and X_i^{fuse} is the fused activation output.

2) DOUBLE DROPOUT NETWORK

The Double Dropout Network (CDN) is derived from the Multiple Dropout Network (MDN) that we proposed for facial landmark localization [7]. The following are experimentally verified in [7]:

1) Dropout added to the convolutional layers can better prevent the *regression* network training from overfitting than the general practice with dropout added to the fully-connected layers;
2) Two dropouts can better balance the training time and stability than other setups with fewer or more dropouts;
3) Shallow network can better balance the runtime speed and accuracy than deeper network.

Due to these advantages, the DDN is designed as a VGG-10 with two double-convolution blocks, one triple-convolution block and three fully-connected layers, with two dropout layers implemented next to the second and third convolution blocks. We keep the same network settings as of the MDN landmark detector in [7]. Note the following big differences between our DDN and the MDN in [7]:

1) As we do not have to deal with non-frontal faces for facial palsy analysis, at least for the time being, we do not have the pose classifier as in the framework of the MDN;
2) The input to the MDN is a facial image only, but we consider the facial line segment image given by the LSN as an addition input to the facial image, i.e., the DDN considers both the image and associated line segments as input.

C. DETECTION OF FACIAL PALSY REGION

The Facial Palsy Region Detector Net_p takes the facial image combined with the associated facial line segment image as input and delivers the local palsy regions in the output. The network structure is similar to that of the Darknet [17] but with landmark-oriented 8×8 cells to cover the facial area where the local palsy regions can appear. The approach consists of the following steps.

1) The input combines the face detected by Net_f and the facial line segment image rendered by LNS, i.e., the input includes the three RGB channels and the additional fourth channel for the line segment image.
2) The facial landmarks given by Net_m are used as the references to implement the grid of 8×8 cells, as shown in Figure 3. The 8×8 cells are designed for detecting all possible sizes of palsy regions.
3) Each cell is associated with 2 bounding boxes for predicting the palsy regions of two classes, which are classes *Eyes* and *Mouth*. The former captures the palsy regions at the eyes region and the latter for the mouth region.
4) The core part of the network is modified from the Darknet-19, and it consists of 4 blocks with 11 convolution layers and 4 max-pooling layers (v.s. 7 blocks, 19 convolution layers and 5 max-pooling in Darknet-19). As the configuration shown in Figure 3, it operates on the input firstly by 2 single-convolution blocks, then 2 double-convolution blocks, then 1 triple-convolution blocks, then 1 convolution layer followed by a Route-Reorganization-Route and another convolution layer for multi-block feature extraction. A 2×2 max-pooling is implemented at each of the first 4 blocks.
5) We train and evaluate the network on the YFP (YouTube Facial Palsy) database. The details are reported in Sec. V.

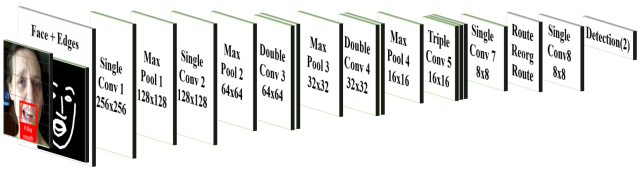

FIGURE 3. The Facial Palsy Region Detector Net_p consists of 4 blocks with 11 convolution layers and 4 max-pooling layers. The last Route-Reorganization-Route and the convolution layer Conv8 are for multi-block feature extraction. The input is the facial image coupled with the facial line segment image.

The palsy region detector Net_p aims at the minimization of the prediction loss, L_p, which is the sum of the following

three losses, the location loss L_n, the region confidence loss L_o and the class probability loss L_c.

$$\begin{aligned} L_n = {} & \lambda_{ob}^{loc} \sum_{i}^{S^2} \sum_{j}^{N_b} \mathbb{I}_{ij}^{ob} [(x_{ij}^{pr} - x_{ij}^{ob})^2 + (y_{ij}^{pr} - y_{ij}^{ob})^2 \\ & + (w_{ij}^{pr} - w_{ij}^{ob})^2 + (h_{ij}^{pr} - h_{ij}^{ob})^2] \\ & + \lambda_{no\_ob}^{loc} \sum_{i}^{S^2} \sum_{j}^{N_b} \mathbb{I}_{ij}^{no\_ob} [(x_{ij}^{pr} - x_{ij}^{c})^2 + (y_{ij}^{pr} - y_{ij}^{c})^2 \\ & + (w_{ij}^{pr} - w_{ij}^{c})^2 + (h_{ij}^{pr} - h_{ij}^{c})^2] \end{aligned} \tag{3}$$

where x_{ij}^{ob}, y_{ij}^{ob}, w_{ij}^{ob}, h_{ij}^{ob} are respectively the center coordinates and the width and height of the target, i.e., the palsy region, associated with Cell-(i, j). x_{ij}^{pr}, y_{ij}^{pr}, w_{ij}^{pr}, h_{ij}^{pr} are respectively the coordinates and the width and height of the anchor-based predicted box. x_{ij}^{c}, y_{ij}^{c}, w_{ij}^{c}, h_{ij}^{c} are respectively the center coordinates and the width and height of the cell without overlap with any palsy region. λ_{ob}^{loc} and $\lambda_{no\_ob}^{loc}$ are the weights imposed on the palsy region (target) and non-palsy region (background).

$$\begin{aligned} L_o = {} & \lambda_{ob}^{conf} \sum_{i}^{S^2} \sum_{j}^{N_b} \mathbb{I}_{ij}^{ob} \left[Conf_{ij}^{pr} - IOU(B_{ij}^{pr}, B_{ij}^{tr}) \right]^2 \\ & + \lambda_{no\_ob}^{conf} \sum_{i}^{S^2} \sum_{j}^{N_b} \mathbb{I}_{ij}^{no\_ob} (Conf_{ij}^{pr})^2 \end{aligned} \tag{4}$$

where $Conf_{ij}^{pr}$ is the confidence of the predicted box based on Cell-(i, j), $IOU(B_{ij}^{pr}, B_{ij}^{tr})$ is the Intersection-over-Union of the predicted bounding box B_{ij}^{pr} and the ground-truth bounding box B_{ij}^{tr} of Cell-(i, j). λ_{ob}^{conf} and $\lambda_{no\_ob}^{conf}$ are the weights to compromise the cells overlapped with targets and those without.

$$L_c = \sum_{i}^{S^2} \sum_{j}^{B} \mathbb{I}_{ij}^{ob} \{p_{ij}^{pr}(C_k) - p_{ij}^{tr}(C_k)\}^2 \tag{5}$$

where $p_{ij}^{pr}(C_k)$ and $p_{ij}^{tr}(C_k)$ are respectively the probabilities of the predicted box and of the ground-truth box being with the region class C_k.

As we implement an 8×8 grid, the output of Net_p is an 8×8 tensor, due to the design with 2 bounding boxes for each cell, 4 numbers for the coordinates of each bounding box, the probability that each bounding box confines or overlaps a local palsy region, and the probabilities that each bounding box classified to Class-Eyes and Class-Mouth.

IV. DATA PREPARATION

Among the three component networks, the training and testing of the face detector Net_f is addressed in Sec. III-A, this section presents the data used for training and testing the landmark detector Net_m and the palsy region detector Net_p.

A. FACIAL LANDMARK DATABASES

We consider the 300W and Menpo databases for training and evaluating the landmark detector Net_m [22], [26]. Both databases offer specific training and testing sets. The 300W consists of nearly frontal images with 68 annotated landmarks. The 300W training set is composed of several popular datasets, for example, AFW and HELEN [21]; the testing set provides 300 indoor and 300 outdoor face images. The Menpo database has 8979 training images (6679 nearly-frontal and 2300 nearly-profile faces) and 7281 test images (5335 nearly-frontal and 1946 nearly-profile faces). 68 landmarks are annotated on each nearly-frontal face, and 39 landmarks on each nearly-profile face. Because our facial palsy analysis focuses on nearly-frontal faces, the Menpo profile faces are not considered in our experiments. All Menpo nearly-frontal faces are put together with the 300W training set for training the Net_m, and the performance evaluated on the 300W test set. The training phase involves the following steps:

1) Connect the neighboring landmarks on each training face image to make the associated line segment image;
2) Train the LSN by using the training set as input and the associated line segment images as output;
3) Use the above trained LSN as the pretrained component in Net_m, composed of LSN and DDN, and train Net_m by using the training set as input and the associated landmark locations as output.

In addition to the 300W test set, we also select 500 faces randomly from the YFP database as another test set. The performance of Net_m is reported in Sec. V with a comparison to other contemporary approaches.

B. YOUTUBE FACIAL PALSY (YFP) DATABASE

We have collected 32 videos of 21 facial palsy patients from YouTube, and a few patients have multiple videos. The patient in each video speaks to the camera and the camera records the facial expression variation across time. Depending on different patients at different time of recording, some images show the syndrome of the palsy-caused deformation with high intensity and some with median or low intensity, justified by the severity revealed by the deformation pattern. The images with very low intensity may appear similar to a normal face, and in some cases, even the clinician can hardly tell whether the face is with the palsy syndrome if only looking at one single image without referencing other frames. For some patients, the palsy-induced facial asymmetry is easy to observe even when the patient stops talking and keeps neutral in the expression.

As the duration of the shortest facial palsy syndrome usually lasts for a second or so, we converted each video into an image sequence with sampling rate 6FPS. For each image, we manually cropped the local palsy regions when the facial palsy intensity was considered sufficiently high by a clinician. The palsy regions were labeled by three independent clinicians, and we used the intersection of the independently

cropped regions as the ground truth. When cropping on each image, we labeled the intensity observed in each palsy region as 0.5 for *low* or 1.0 for *high*, and the ground truth was defined by averaging. In addition to the syndrome intensity, we also labeled the palsy regions into Classes Eyes or Mouth, depending on whether the palsy region appeared at the eyes or mouth area. This part of labeling was performed semi-automatically by using the facial landmarks. Since the facial landmarks are numbered in a specific order, the class labels *Eyes* or *Mouth* for the palsy regions were given directly from the numbered landmarks that are in these regions. The YFP database can be available to the research community by request.

V. EXPERIMENTAL EVALUATION

A. PERFORMANCE OF FACIAL LANDMARK DETECTOR

The experimental results for the data and setups reported in the previous sections are presented in this section. All experiments were run on a Ubuntu 14.04 with Titan X GPU, and CUDA 7.5 with cuDNN 4.0 on Caffe. The accuracy of the facial landmarks is measured by the common Normalized Mean Error (NME), which is the average point-to-point Euclidean distance normalized by the interocular distance (the distance between the outer corners of the eyes) [21]. The NME can be written as follows.

$$NME = \frac{1}{N}\sum_{i=1}^{N}\frac{||z_i - s_i||_2}{d_i} \tag{6}$$

where z_i denotes the ground-truth coordinates of the landmarks of the face$-i$, s_i is the estimated coordinates and d_i is the interocular distance.

The performance of the landmark detector Net_m on the 300W is shown in Tabel 1, together with the performance of other contemporary approaches. The proposed Net_m outperforms others for facial landmark localization. As the configuration elaborated in Sec. III-B, Net_m is actually the DDN (Double Dropout Network) with line segment learning added in. Without the LSN for line segment learning, Net_m is simply the DDN, whose performance is the same as of the MDN [7], as shown in the table. Therefore, the performance difference between Net_m and MDN in the table reveals the contribution made by the inclusion of the line segment learning.

TABLE 1. Landmark accuracy on 300W in the normalized mean error. The best three in each category column are in boldface.

Method	Common	Challenge	Full Set
SDM [24]	5.57	15.40	7.50
RLBF [18]	4.95	11.98	6.32
CMC-CNN [6]	**4.91**	12.03	**6.30**
RTSM [8]	6.02	16.52	8.06
RCPR [3]	6.18	17.26	8.35
ESR [4]	5.28	17.00	7.58
3DDFA [27]	6.15	10.59	7.01
3DDFA+SDM [27]	5.53	**9.56**	6.31
MDN [7]	**4.95**	**10.52**	**6.01**
Net_m	**4.85**	**9.81**	**5.83**

Fig. 4 shows several samples of the landmarks located on the testing set of 500 palsy faces. The NME on this set is 9.22, close to the performance obtained on the challenge subset of 300W. Because it is not a central concern to locate the landmarks on palsy faces in this study, we use the same Net_m which was tested on the above 300W and does not consider any facial palsy samples in the training. Although the inclusion of palsy samples in training may improve the landmark accuracy on palsy faces, the selection of training samples, e.g., with different portions of strong and weak palsy patterns, can result in different capacities of handling different palsy patterns. This task is beyond the scope of this paper, and can be considered in the continuing work.

B. PERFORMANCE OF LOCAL PALSY REGION DETECTOR

As only 21 patients are available in the YFP dataset, we adopt the leave-one-person-out (LOPO) protocol that

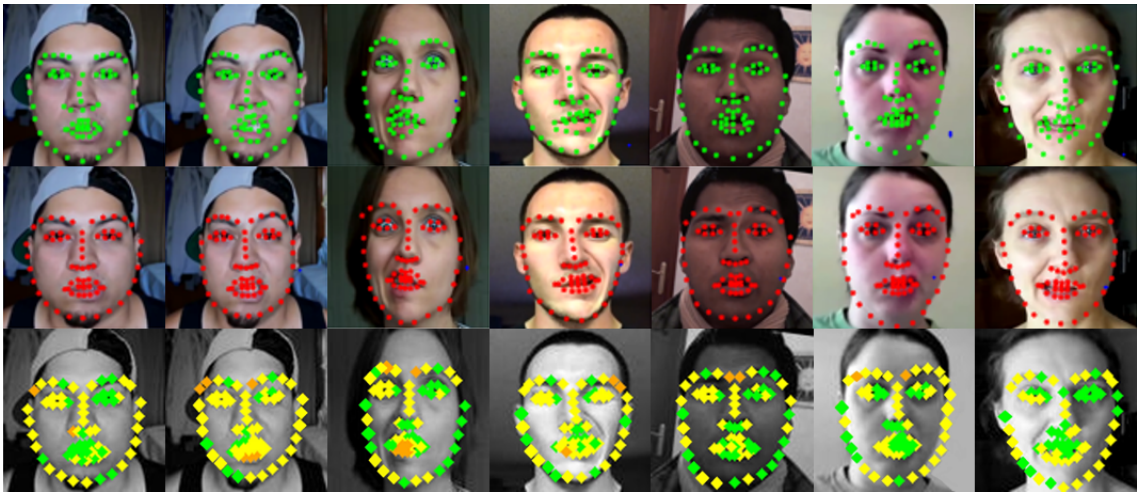

FIGURE 4. Samples of landmark detection on palse faces. The top row are ground truth landmarks, the middle row are the landmarks detected by the proposed Net_m, and the bottom row are the errors between the detected and ground-truth. Errors $\leq$ 2 pixels are in green, $\leq$ 4 pixels in yellow, $>$ 4 pixels in orange.

takes 20 patients for training and the remaining one for testing in one session, and in the next session the one in testing is replaced by one who was in the previous training. This process is repeated for all 21 patients, and the performance is measured by the average. To make our solution robust against expression variation, we have included the CK+ database in our training/testing sets, and compared with setup that excludes the CK+. In the experiment with CK+ included in the training, we randomly split the CK+ into five subject-independent subsets, and run 5-fold cross validation together with the 21 LOPO tests on the YFP dataset. In the experiment without CK+ in the training, we run the same tests on the same testing subsets.

To better understand the contribution of the line segment learning, we built another network without the line segment network included in the pipeline, i.e., the configuration in Figure 1 with the LSN removed. The palsy region detector Net_p in Figure 3 was also modified with the facial line segment image removed and the facial image was used as the only input. Figure 5 shows the performances with and without CK+ considered in the training. When the CK+ is not included and the framework does not consider facial line segments, the palsy region detector Net_p detects many false positives on the CK+ test set, and the overall performance gives EER (Equal Error Rate) 77.7%. When the CK+ is included, the accuracy is improved to EER 87.5%. When the facial line segment is included in the input, the performance reports EER 82.3% for the case without CK+ included in the training. The best performance, EER 91.2%, is given by the setup with the facial line segment included in the input and the CK+ is included in the training. The experiments reveal the following observations:

- The inclusion of the CK+ substantially improves the robustness against expression variations, which can be a major cause for false positives.

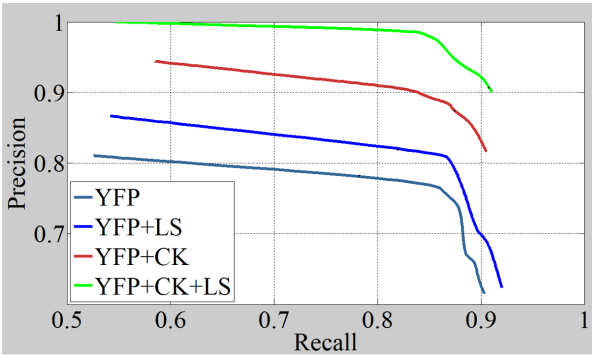

FIGURE 5. Performance comparison between different setups: the model without line segments and trained on YFP gives 77.7% EER (Equal Error Rate), with line segments gives 82.3% (YFP+LS), without line segments and with CK+ included in training gives 87.5% (YFP+CK), with line segments and with CK+ included gives 91.2% (YFP + CK + LS).

- The incorporation of facial line segment learning can enhance the network's capability of detecting the target local regions.

To identify the intensity variation of the facial palsy syndrome over time, we extract the softmax probability from the output of the DHN when processing each frame as the intensity indicator. As the data in our training set are labeled score 1 for *high* intensity, 0.5 for *low* intensity and 0 for no intensity or the normal face. All scores are labeled on the training data and considered in the DHN training phase. Figures 6 and 7 show the intensity variation of the facial palsy syndrome over time for Subject-1 and Subject-6 in our database. The intensity is expressed in % with the frame number in parenthesis (·). It can be seen that the intensity is *quantitatively* captured by the softmax probability output of the DHN for each frame. When the intensity is low, the mouth shape appears close to normal; when it is high, the mouth deformation shows a strong asymmetric pattern.

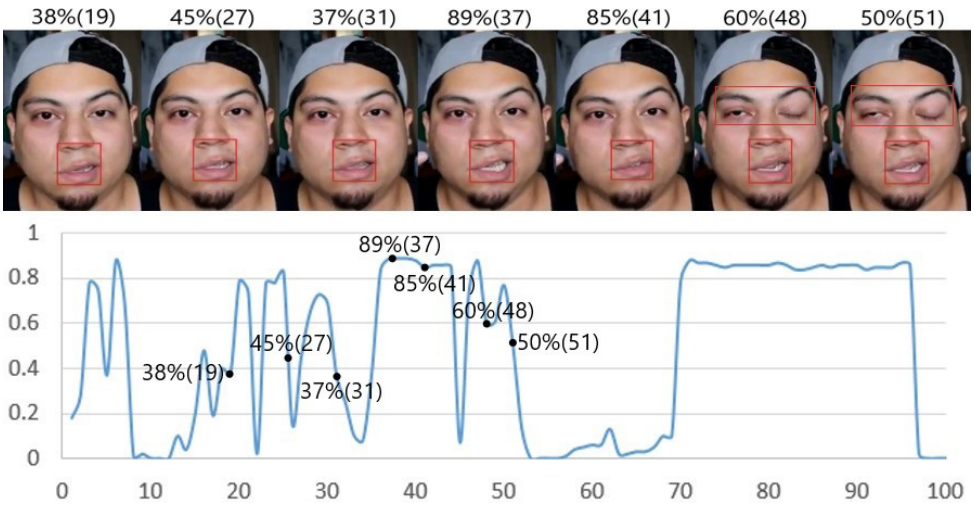

FIGURE 6. Intensity variation over time for Subject-1, intensity expressed in % with the frame number in parenthesis (·). Top row shows the faces captured at the time specified on the intensity variation at the bottom row.

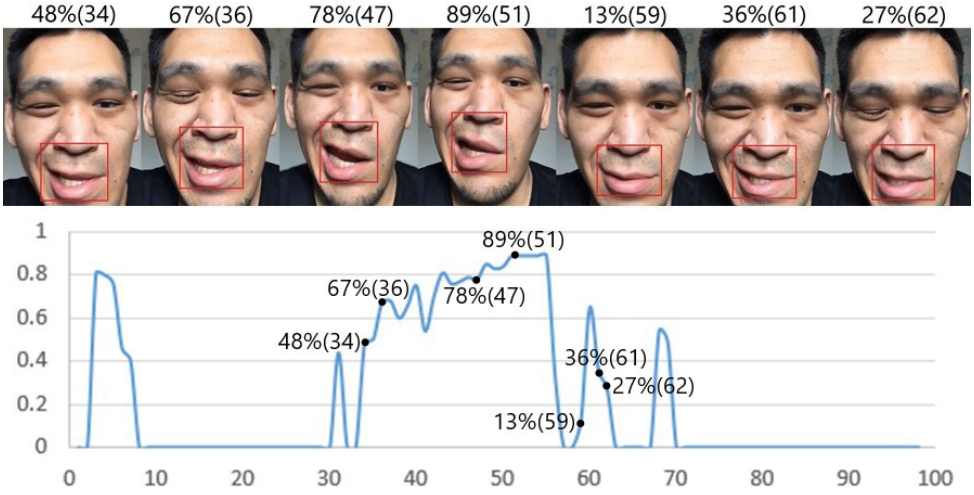

FIGURE 7. Intensity variation over time for Subject-6, intensity expressed in % with the frame number in parenthesis (·). Top row shows the faces captured at the time specified on the intensity variation at the bottom row.

VI. CONCLUSION

We present the development of a pioneering deep learning framework, the Deep Hierarchical Network (DHN), for quantitative analysis of facial palsy. The proposed hierarchical framework is composed of a face detector, a facial landmark detector and a local palsy region detector. We have experimentally verified that the line segment learning in the framework leads to an important part of deep features able to improve the accuracy of facial landmark and palsy region detection. To enhance the robustness against facial expression variations, we include the CK+ expression database in the learning phase so that the framework is trained to distinguish common facial expressions from facial palsy patterns. The novelties of this study include the modification of a state-of-the-art edge detector for extracting the facial line segments, the embedding of the line segment learning for the detection of facial landmarks and local palsy regions, the quantitative description of the syndrome intensity, and the release of the first clinically labeled YFP (YouTube Facial Palsy) database. Experiments show that the proposed framework can be a highly effective solution for the automatic quantitative analysis of facial palsy.

REFERENCES

[1] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, "Incremental face alignment in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1859–1866.

[2] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks)," in *Proc. Int. Conf. Comput. Vis.*, 2017, pp. 1021–1030.

[3] X. P. Burgos-Artizzu, P. Perona, and P. Dollár, "Robust face landmark estimation under occlusion," in *Proc. ICCV*, Dec. 2013, pp. 1513–1520.

[4] X. Cao, Y. Wei, F. Wen, and J. Sun, "Face alignment by explicit shape regression," *Int. J. Comput. Vis.*, vol. 107, no. 2, pp. 177–190, 2014.

[5] S. He, J. J. Soraghan, B. F. O'Reilly, and D. Xing, "Quantitative analysis of facial paralysis using local binary patterns in biomedical videos," *IEEE Trans. Biomed. Eng.*, vol. 56, no. 7, pp. 1864–1870, Jul. 2009.

[6] Q. Hou, J. Wang, L. Cheng, and Y. Gong, "Facial landmark detection via cascade multi-channel convolutional neural network," in *Proc. ICIP*, Sep. 2015, pp. 1800–1804.

[7] G.-S. Hsu and C.-H. Hsieh, "Cross-pose landmark localization using multi-dropout framework," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Oct. 2017, pp. 390–396.

[8] G.-S. J. Hsu, K.-H. Chang, and S.-C. Huang, "Regressive tree structured model for facial landmark localization," in *Proc. ICCV*, 2015, pp. 3855–3861.

[9] G.-S. J. Hsu, W.-F. Huang, and J.-H. Kang, "Hierarchical network for facial palsy detection," in *Proc. CVPRW*, 2018, pp. 1–7.

[10] H. S. Kim, S. Y. Kim, Y. H. Kim, and K. S. Park, "A smartphone-based automatic diagnosis system for facial nerve palsy," *Sensors*, vol. 15, no. 10, pp. 26756–26768, 2015.

[11] A. M. Kosins, K. A. Hurvitz, G. R. Evans, and G. A. Wirth, "Facial paralysis for the plastic surgeon," *Can. J. Plastic Surgery*, vol. 15, no. 2, pp. 77–82, 2007.

[12] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 21–37.

[13] Y. Liu, M.-M. Cheng, X. Hu, K. Wang, and X. Bai, "Richer convolutional features for edge detection," in *Proc. CVPR*, 2017, pp. 5872–5881.

[14] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool, "Face detection without bells and whistles," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2014, pp. 720–735.

[15] T. H. Ngo, M. Seo, N. Matsushiro, W. Xiong, and Y.-W. Chen, "Quantitative analysis of facial paralysis based on limited-orientation modified circular Gabor filters," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2016, pp. 349–354.

[16] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. BMVC*, 2015, pp. 1–6.

[17] J. Redmon and A. Farhadi. (2016). "YOLO9000: Better, faster, stronger." [Online]. Available: https://arxiv.org/abs/1612.08242

[18] S. Ren, X. Cao, Y. Wei, and J. Sun, "Face alignment at 3000 FPS via regressing local binary features," in *Proc. CVPR*, Jun. 2014, pp. 1685–1692.

[19] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. ANIPS*, 2015, pp. 91–99.

[20] M. A. Sadeghi and D. Forsyth, "30 Hz object detection with DPM V5," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2014, pp. 65–79.

[21] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: Database and results," *Image Vis. Comput.*, vol. 47, pp. 3–18, Mar. 2016.

[22] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: The first facial landmark localization challenge," in *Proc. CVPRW*, 2013, pp. 397–403.

[23] T. Wang, J. Dong, X. Sun, S. Zhang, and S. Wang, "Automatic recognition of facial movement for paralyzed face," *Bio-Med. Mater. Eng.*, vol. 24, no. 6, pp. 2751–2760, 2014.

[24] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *Proc. CVPR*, 2013, pp. 532–539.

[25] S. Yang, P. Luo, C.-C. Loy, and X. Tang, "Wider face: A face detection benchmark," in *Proc. CVPR*, 2016, pp. 5525–5533.
[26] S. Zafeiriou, G. Trigeorgis, G. Chrysos, J. Deng, and J. Shen, "The Menpo facial landmark localisation challenge: A step towards the solution," in *Proc. CVPRW*, Jul. 2017, pp. 2116–2125.
[27] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li, "Face alignment across large poses: A 3D solution," in *Proc. CVPR*, 2016, pp. 146–155.

GEE-SERN JISON HSU (M'07–SM'14) received the dual M.S. degree in electrical and mechanical engineering and the Ph.D. degree in mechanical engineering from the University of Michigan, Ann Arbor, in 1993 and 1995, respectively.

From 1995 to 1996, he was a Post-Doctoral Fellow with the University of Michigan. From 1997 to 2000, he was a Senior Research Staff with the National University of Singapore. In 2001, he joined Penpower Technology, where he led research on face recognition and intelligent video surveillance. In 2007, he joined the Department of Mechanical Engineering, National Taiwan University of Science and Technology (NTUST), where he is currently an Associate Professor. His research interests include computer vision and pattern recognition. He is a Senior Member of the IEEE and IAPR. His team at Penpower Technology was a recipient of the Best Innovation and Best Product Award at the SecuTech Expo for three consecutive years (2005–2007). He received several best papers awards after joining NTUST, including ICMT 2011, CVGIP 2013, CVPRW 2014, ARIS 2017, and CVGIP 2018.

JIUNN-HORNG KANG received the M.D. degree from the School of Medicine, in 1998, the M.M.S. degree from the Medicine Graduate Institute of Clinical Medicine, in 2008, and the Ph.D. degree from the Institute of Biomedical Engineering, National Taiwan University, Taipei, Taiwan, in 2011. He is currently an Associate Professor with the School of Medicine, Taipei Medical University, Taipei. His main interests include biosignal procession and non-linear system analysis.

WEN-FONG HUANG received the B.S. degree in mechanical engineering from the National Kaohsiung University of Science and Technology, Kaohsiung, Taiwan, in 2016, and the M.S. degree in mechanical engineering from the National Taiwan University of Science and Technology, Taipei, Taiwan, in 2018. His research interests include deep learning and facial landmark.