Hierarchical Network for Facial Palsy Detection

Gee-Sern Jison Hsu ,Wen-Fong Huang
National Taiwan University of Science and Technology
Taipei, Taiwan
jison@mail.ntust.edu.tw, m10503416@mail.ntust.edu.tw

Jiunn-Horng Kang
Taipei Medical University
Taipei, Taiwan
jhk@tmu.edu.tw

Abstract

We propose the Hierarchical Detection Network (HDN) for the detection of facial palsy syndrome. This can be the first deep-learning based approach for the facial palsy detection. The proposed HDN consists of three component networks, the first detects faces, the second detects the landmarks on the detected faces, and the third detects the local palsy regions. The first and the third component networks are built on the Darknet framework, but with fewer layers of convolutions for shorter processing speed. The second component network employs the latest 3D face alignment network for locating the landmarks. The first component network employs a $N_a \times N_a$ grid over the overall input image, while the third component network employs a $N_b \times N_b$ grid over each detected face, making the HDN capable of efficiently locating the affected palsy regions. As previous approaches were evaluated on proprietary databases, we have collected 32 videos from YouTube and made the first public database for facial palsy study. To enhance the robustness against expression variations, we include the CK+ facial expression database in the training and testing phases. We show that the HDN does not just detect the local palsy regions, but also captures the frequency of the intensity, enabling the video-to-description diagnosis of the syndrome. Experiments show that the proposed approach offers an accurate and efficient solution for facial palsy detection/diagnosis.

1. Introduction

Facial palsy is a type of facial nerve paralysis that results in the loss of muscle control on the affected facial region. Typical symptoms include facial deformity and facial expression dysfunction, and the associated intensity can vary from mild to severe. Facial palsy not only significantly affects the facial appearance but also leads to impaired feeding functions and adverse psychosocial consequences [4]. The diagnosis of facial palsy is usually not difficult as it inspects the facial symmetry, and currently relies on the visual inspection by a clinician. The approaches for automatic detection/diagnosis of facial palsy have been emerging in recent years. However, most, if not all, of the approaches use handcrafted features and classifiers, the deep-learning based approaches are yet to be developed. Another issue with the previous approaches is that their experiments were performed on proprietary databases, making benchmarking and performance comparison difficult. We will give a brief review on the previous approaches in Sec. 2.

We propose the Hierarchical Detection Network (HDN), which consists of three component networks, the first is for locating the face, the second is for locating the facial landmarks on the detected faces, and the third is for locating the affected facial palsy regions. We call the first component network the *FaceNet*, the second component network the *LandmarkNet*, and the third component network the *PalsyNet*. Both FaceNet and PalsyNet are built on the Darknet framework [8], which is developed to build the YOLO detector [8]. The LandmarkNet employs the lastest 3D face alignment network that combines the state-of-the-art Hour-Glass (HG) network and residual block. Although the FaceNet and PalsyNet are built on the Darknet framework, the HDN is more advantageous than the YOLO in the customized design for the detection of a specific object, e.g., the local facial palsy area in this study. The advantages include the following: 1) The HDN is a hierarchical detector, which detects face first and then locates the palsy region on the detected face using facial landmark from the LandmarkNet as reference. This design searches for large ROIs (faces) in the first level, and then searches for relatively small ROIs (palsy regions) in the second level, making the search highly precise and efficient. 2) Both component networks, FaceNet and PalsyNet, are built with *reduced* layers in the architecture, leading to a fast processing/detection speed. Different from most of previous facial palsy diagnosis systems that identify the holistic patterns of facial asymmetry by handcrafted features, the proposed HDN converts the holistic identification problem into a component detection problem, and merge several state-of-the-art object detection networks to develop a unified solution framework.

The HDN does not just detect the affected local palsy regions, but also identifies the frequency of the syndrome intensity when the facial expression changes during talking or other facial actions.

To train and evaluate the proposed approach, we have collected 32 video clips of 22 facial palsy patients from YouTube, and labeled all the data by three specialists. When labeling the data, the affected local regions, called *palsy regions*, are annotated by bounding boxes. We call this dataset the Facial Palsy (FP) database, and will formally release the FP database after medical specialists validate the facial appearances with facial palsy related syndromes. *The alpha version of the dataset can be available upon request.*

The contributions of this work can be summarized as follows:

- We propose the first deep learning solution for the detection and diagnosis of facial palsy by using a regular camera. It can detect the affected areas, i.e., palsy regions, on a patient's face, and identify the frequency of the intensity associated with the facial palsy syndrome, and is robust to expression variation.
- We have made the first public database for facial palsy study. It is composed of videos collected from YouTube with local palsy regions labeled by specialists. Our proposed solution is validated through experiments on this database.

In the following sections, we first give a review on previous work in Sec. 2. Our approach is elaborated in Sec. 3, followed by the experiments on the FP database in Sec. 4.

2. 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 three latest studies and summarize their methods and experiments in this section.

The approach based on the limited-orientation modified circular Gabor filters (LO-MCGFs) was proposed by Ngo et al. [7] for an objective and quantitative analysis of facial palsy. The LO-MCGFs use uniform passbands to remove noises and enhance the desired spatial frequencies, and use bounded filter support to specify the region of interest. These virtues make the approach effective for extracting the facial asymmetry features. The facial expression dataset composed of 75 patients and 10 participants without facial palsy, made by the Osaka Police Hospital, was used in their experiments. Each expression is composed of 60 still images per 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 expression variation, and how the intensity level is assigned.

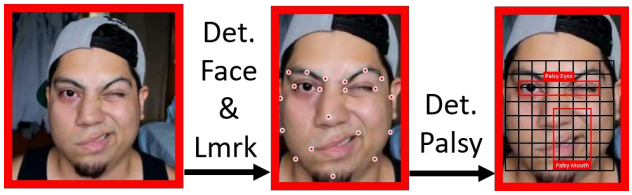

Figure 1. The proposed Hierarchical Detection Framework is composed of three component networks: Net_f for face detection, Net_m for landmark localization, and Net_p for palsy region detection. The face detection and landmark localization is shown in the middle, and the palsy region is located by Net_p at the output.

Kim et al. [3] 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 point sets that correspond to the eye-brows and mouth regions while the subjects change expression. To extract the asymmetric index, the forehead and eye regions are considered in heuristic approaches that measure the distances and ratios of different distances. The Linear Discriminant Analysis (LDA) and Support Vector Machine (SVM) are then employed for classification. The system is evaluated on a private database with 23 facial palsy patients and 13 volunteers without facial palsy.

A quantitative approach that considers both the static facial asymmetry and the speed of appearance change is proposed by Wang et al. [11]. They first trained an ASM (Active Shape Model) [11] for locating the landmarks on a 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 asymmetry 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 recent progress made on the automatic detection/diagnosis of facial palsy with the following aspects: 1) All approaches consider handcrafted features and classifiers; 2) Facial asymmetry is a core trait to identify/diagnose for facial palsy; 3) The databases used in these studies are proprietary and thus unavailable to the research community.

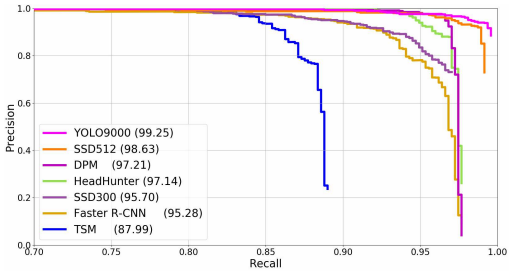

Figure 2. Comparison of approaches for face detection on the AFW benchmark. Net$_f$ is part of our framework which is made of YOLO-9000 and retrained on the Wider Face database. The numbers in the parentheses are average precision (AP).

Figure 4. The proposed Palsynet consists of 4 blocks with 11 convolution layers and 4 max-pooling layers. The last Route-Reorganization-Route and a convolution layer are for multi-block feature extraction.

3. Hierarchical Detection Framework

We formulate the facial palsy identification as an object detection problem, and consider the facial-palsy-caused deformation regions, or simply the *palsy regions*, on a patient's face as the target objects. Our proposed solution is a hierarchical network composed of three component networks. The first component network, called FaceNet and denoted as Net$_f$, is for face detection; the second component network, called LandmarkNet and denoted as Net$_m$, for facial landmark detection; and the third component network, called PalsyNet and denoted as Net$_p$, for local palsy region detection. Fig. 2 shows the outputs from these component networks. Net$_f$ is based on the YOLO-9000 and trained on the WIDER FACE dataset [12], and it attains AP 99.25% on the AFW benchmark. Net$_m$ follows the state-of-the-art Face Alignment Network (FAN) that combines the Hour-Glass (HG) network, which is one of the latest CNN architectures for human pose estimation and a cutting-edge residual block, and is trained on a large synthetically expanded 2D facial landmark dataset [2]. Given the facial landmarks detected by Net$_m$ as priors, Net$_p$ is designed with a detection window attached to the facial landmarks with anchor boxes located around the eyes and mouth regions for fast detection of local palsy regions. The proposed framework implements a top-down detection flow with face detection comes first, followed by landmark detection, and then an efficient target search for palsy regions.

3.1. Face and Facial Landmark Detection

The FaceNet, Net$_f$, is built on the YOLO-9000 and retrained on the Wider Face database [12]. The YOLO-9000, proposed by Redmon and Farhadi, is one of the few state-of-the-art real-time object detectors [8]. 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 methods, including the Faster RCNN with ResNet [9] and the SSD [5]. For face detection, we train the YOLO-9000 using the WIDER FACE database [12], 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 [12], 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 some settings of the YOLO-9000, including the partition of the 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, the results are shown in Fig.3. FaceNet (or YOLO-9000 face detector) achieves AP (Average Precision) 99.25% on AFW benchmark, better than the DPM (97.2%) [10], the HeadHunter (97.1%) [6], SSD-512 (98.6%) [5] and the Faster RCNN (95.3%) [9]. 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 the FaceNet.

Given a detected face, we exploit the FAN (Face Alignment Network) proposed by Bulat and Tzimiropoulos [2] for locating facial landmarks on a detected face. The FAN is built on the state-of-the-art Hour-Glass (HG) network [2] with the bottleneck block replaced by a residual block. The architecture of the HG network is illustrated in Fig. 3.1. It consists of multiple stacked hourglass modules which allow for repeated bottom-up, top-down inference. A single hourglass module contains boxes of different scales and specific connections between the blocks. The hourglass module before stacking is related to fully convolutional networks and other designs that process spatial information at multiple scales, and also related to conv-deconv and encoder-decoder architectures. The symmetric topology of the hourglass module and the conv-deconv and encoder-decoder networks are similar, but the operations are different in that the unpooling or deconv layers are removed from the hourglass module. The hourglass module relies on simple nearest neighbor upsampling and skips the connections for top-down processing. Another major difference of the HG network is that it performs repeated bottom-up, top-down inference by stacking multiple hourglass modules.

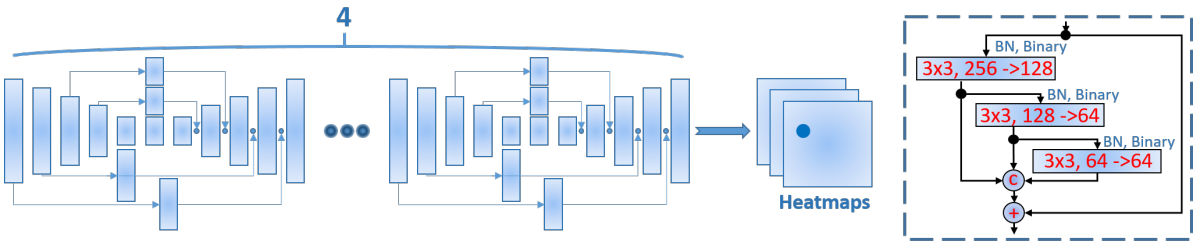

Figure 3. The LandmarkNet, Net$_m$, exploits the latest FAN (Face Alignment Network) proposed by Bulat and Tzimiropoulos [2].

3.2. Detection of Facial Palsy Regions

The reason that we train the YOLO-9000 for face detection is that face is a specific object which can appear anywhere in an image. This is different from the detection of palsy regions, which only appear on a face, and particularly on the eyes and mouth regions. We therefore consider the grid associated with the landmarks for the palsy region detection. Figures 3 shows the grid with 8×8 cells that covers the facial area for palsy region detection. The proposed approach consists of the following steps.

1. The facial landmarks are used as the references to implement the grid of 8×8 cells, as shown in Fig. 3. The 8×8 cells are designed to be able to capture the smallest palsy region.

2. 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 eyes region and the latter for the mouth region.

3. The core network, called Palsynet, 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). 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.

4. We train and evaluate the network by the alpha version of the FP (Facial Palsy) database. Experimental details are reported in Sec. 4.

The proposed Palsynet 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.

$$\begin{aligned} L_n &= \lambda_{ob}^{loc} \sum_{i}^{S^2} \sum_{j}^{N_b} \mathbb{I}_{ij}^{ob} [(x_{ij}^{pr} - x_{ij}^{ob})^2 + (y_{ij}^{pr} - y_{ij}^{ob})^2 \\ &+ (w_{ij}^{pr} - w_{ij}^{ob})^2 + (h_{ij}^{pr} - h_{ij}^{ob})^2] \\ &+ \lambda_{no\_ob}^{loc} \sum_{i}^{S^2} \sum_{j}^{N_b} \mathbb{I}_{ij}^{no\_ob} [(x_{ij}^{pr} - x_{ij}^{c})^2 + (y_{ij}^{pr} - y_{ij}^{c})^2 \\ &+ (w_{ij}^{pr} - w_{ij}^{c})^2 + (h_{ij}^{pr} - h_{ij}^{c})^2] \end{aligned} \quad (1)$$

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 object, 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 (object) and non-palsy region (no object).

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

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.

$$L_c = \sum_{i}^{S^2} \sum_{j}^{B} \mathbb{I}_{ij}^{ob} \{(p_{ij}^{pr}(C_k) - p_{ij}^{tr}(C_k)\}^2 \quad (3)$$

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 object class C_k.

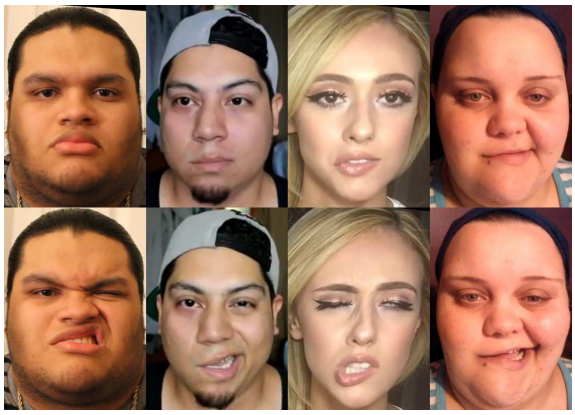

Figure 5. The top row shows the patients when pausing on talking, the bottom row shows the syndromes at talking.

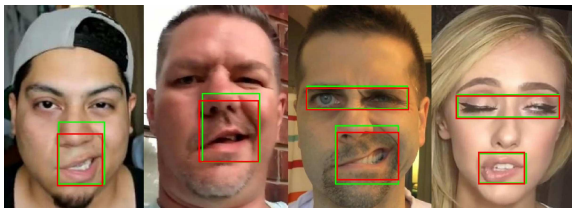

Figure 6. Samples with ground-truth bbox in red and detected bbox in green.

As we implement an 8×8 grid, the output of the Palsynet is an 8×8 tensor, due to the design with 2 bounding boxes (bboxs) for each cell, and for each cell, there are 4 numbers for the coordinates and size of a bbox, the probability that the bbox confines or overlaps a palsy region and the probabilities that the bbox being in Class Eyes or Class Mouth.

4. Experimental Evaluation

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 previous 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. Figure 5 shows a few cases during talking and taking a pause.

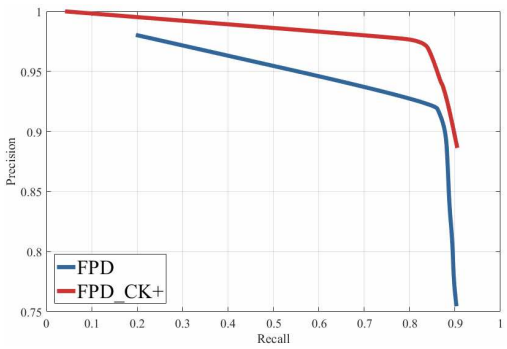

Figure 7. Performance comparison between training/testing sets with and without CK+ expression dataset. When CK+ excluded, the overall performance degrades, with precision 89% at recall 87%. When CK+ included, the accuracy is improved, with precision 93% at recall 88%

As the duration of the shortest facial palsy syndrome usually lasts for a second or so, we convert each video into an image sequence with 6FPS. For each image, we manually cropped the local palsy regions when the facial palsy intensity was considered sufficiently high by a specialist. The palsy regions were labeled by three independent specialists, and we use 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 *low* or *high*, and the ground truth was determined by majority voting. In addition to the syndrome intensity, we also labeled the palsy regions into Classes Eyes or Mouth, depending on whether the palsy region is close to 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 and close to these regions, and then confirmed by us. Figure 6 shows samples with ground truth and the detection outcomes obtained by the proposed HDN. This dataset is currently under examination by clinicians for labeling other facial syndromes, and will be released to the research community thereafter. However, the alpha version of the dataset can be available upon request.

As only 21 patients are available, we adopt the leave-one-person-out (LOPO) protocol that 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 cases excluding the CK+. In the experiment with CK+ included in the training,

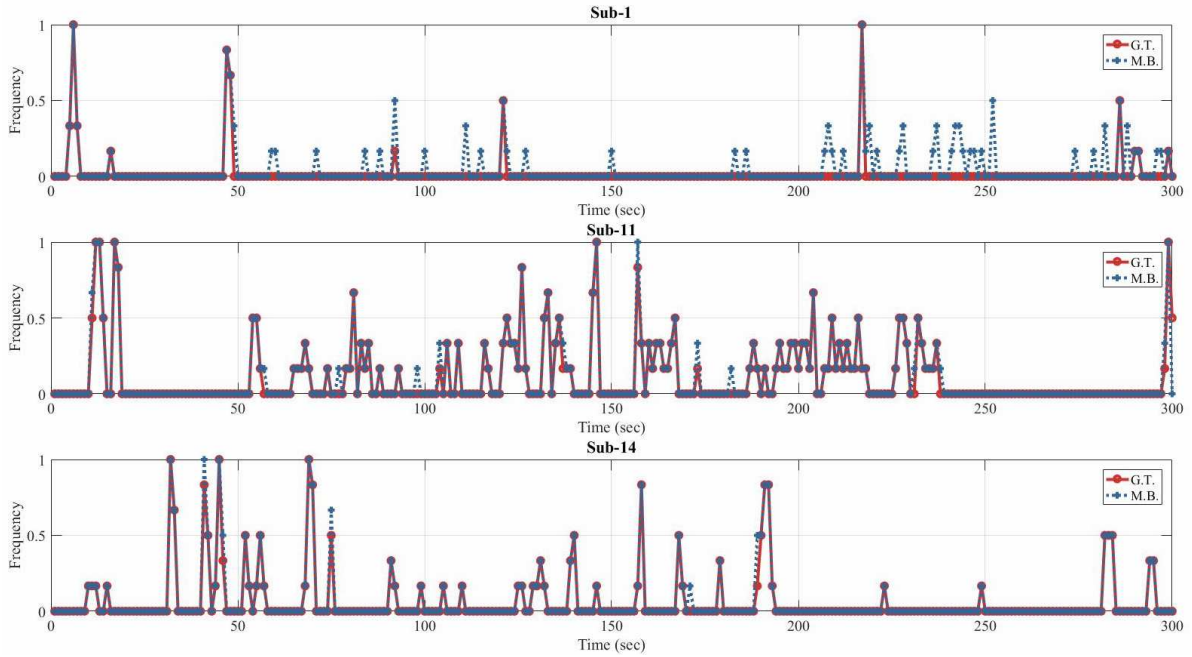

Figure 8. The normalized intensity frequency plot of three subjects selected from the test set of our database.

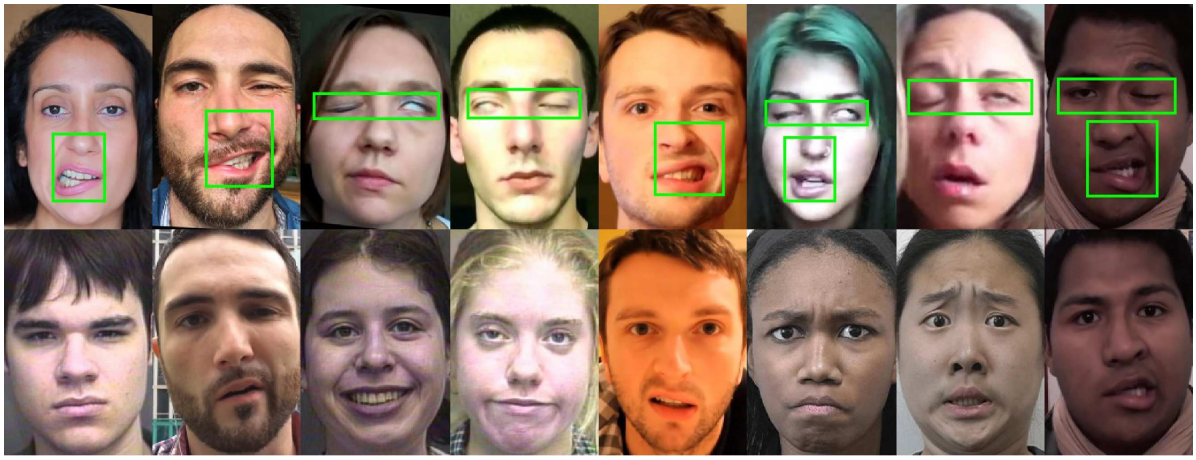

Figure 9. The detection results of applying the proposed solution on the patients in the test set of our database (the top row), and on those randomly selected from the CK+ test set (the bottom row). The bottom row contains a few examples of the patients when pausing on talking and no significant syndrome visible, in line with the output of our system.

we randomly split the CK+ into five subject-independent subsets, and run 5-fold cross validation together with the 21 LOPO tests on the palsy dataset. In the experiment without CK+ in the training, we run the same tests on the same testing subsets. Figure 7 shows the performance with and without CK+ in the training sets. When the CK+ is not included, our Palsynet detects quite a few false positives on the CK+ test set, and the overall performance degrades substantially with precision 89% at recall 87%. When the CK+ is included, the accuracy is significantly improved, with precision 93% at recall 88%.

To identify the frequency of the intensity of the facial palsy syndrome, we consider a moving window that counts the frames with local palsy regions detected over a short time period and up to the sampling time. It is difficult (and probably not realistic) to define one cycle of the palsy-caused deformation as the associated intensity changing from the lowest to the next lowest, especially when the patient keeps talking and the lowest intensity changes in one cycle to another. As mentioned above, some patients show

normal faces without any observable palsy syndrome when talking paused and showing a neutral expression; however, some patients show clear palsy syndrome with their neutral expressions. When the latter are asked to keep neutral and normal expression, their faces show a clear sign of palsy, due to the lack of facial muscle control. Considering the frequency meant to measure the severity of the facial palsy syndrome over time, we define it as the number of frames with a sufficient intensity of the palsy-caused deformation detected over a short interval T_s. We normalize the intensity so that the maximum intensity over the sampling interval is made to unity. Figure 8 shows three cases of using our solution for identifying the intensity frequencies. To demonstrate the robustness against expression variations, Figure 9 shows the outcomes of applying the proposed solution on the patients in the test set of our database and on those randomly selected from the CK+ test set. A video clip that was recorded from our test is available on YouTube, https://www.youtube.com/watch?v=1wXeCfFUxd8.

5. Conclusion

We formulate the problem of identifying facial palsy as the detection of local palsy regions, and propose a top-down hierarchical framework composed of three component detection networks. Net_f detects the face, then Net_m detects facial landmarks, and then Net_p detects local palsy regions. To train and evaluate our solution, we collect 32 video clips of 22 facial palsy patients from YouTube, labeled all the data by three clinicians, and will make this database available to the research community. Different from the YOLO-9000 that aims at the detection of generic objects, our solution is tailored made for medical diagnosis and can be extended to the detection of other specific objects. This study can serve as a valuable sample study for the application/modification of deep learning framework for automatic detection/diagnosis of medical disorders.

References

[1] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic. Incremental face alignment in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1859–1866, 2014.

[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 *International Conference on Computer Vision*, volume 1, page 8, 2017.

[3] H. S. Kim, S. Y. Kim, Y. H. Kim, and K. S. Park. A smartphone-based automatic diagnosis system for facial nerve palsy. *Sensors*, 15(10):26756–26768, 2015.

[4] A. M. Kosins, K. A. Hurvitz, G. R. Evans, and G. A. Wirth. Facial paralysis for the plastic surgeon. *Canadian Journal of Plastic Surgery*, 15(2):77–82, 2007.

[5] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

[6] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool. Face detection without bells and whistles. In *European Conference on Computer Vision*, pages 720–735. Springer, 2014.

[7] 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 *Pattern Recognition (ICPR), 2016 23rd International Conference on*, pages 349–354. IEEE, 2016.

[8] J. Redmon and A. Farhadi. Yolo9000: better, faster, stronger. *arXiv preprint arXiv:1612.08242*, 2016.

[9] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[10] M. A. Sadeghi and D. Forsyth. 30hz object detection with dpm v5. In *European Conference on Computer Vision*, pages 65–79. Springer, 2014.

[11] T. Wang, J. Dong, X. Sun, S. Zhang, and S. Wang. Automatic recognition of facial movement for paralyzed face. *Bio-medical materials and engineering*, 24(6):2751–2760, 2014.

[12] S. Yang, P. Luo, C.-C. Loy, and X. Tang. Wider face: A face detection benchmark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5525–5533, 2016.