IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 12, 2021, accepted April 27, 2021, date of publication May 4, 2021, date of current version May 14, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3077479

Design of a Facial Landmark Detection System Using a Dynamic Optical Flow Approach

BING-FEI WU[1], (Fellow, IEEE), BO-RUI CHEN[1], AND CHUN-FEI HSU[2]

[1]Department of Electrical and Computer Engineering, Institute of Electrical and Control Engineering, National Yang Ming Chiao Tung University, Hsinchu 30010, Taiwan

[2]Department of Electrical Engineering, Tamkang University, New Taipei City 25137, Taiwan

Corresponding author: Chun-Fei Hsu (fei@ee.tku.edu.tw)

This work was supported by the Ministry of Science and Technology, Taiwan, under Grant MOST 108-2221-E-009-123-MY2 and Grant MOST 108-2221-E-032-039-MY2.

ABSTRACT Many facial landmark methods based on convolutional neural networks (CNN) have been proposed to achieve favorable detection results. However, the instability landmarks that occur in video frames due to CNNs are extremely sensitive to input image noise. To solve this problem of landmark shaking, this study proposes a simple and effective facial landmark detection method comprising a lightweight U-Net model and a dynamic optical flow (DOF). The DOF uses the fast optical flow to obtain the optical flow vector of the landmark and uses dynamic routing to improve landmark stabilization. A lightweight U-Net model is designed to predict facial landmarks with a smaller model size and less computational complexity. The predicted facial landmarks are further fed to the DOF approach to deal with the unstable shaking. Finally, a comparison of several common methods and the proposed detection method is made on several benchmark datasets. Experimental evaluations and analyses show that not only can the lightweight U-Net model achieve favorable landmark prediction but also the DOF stabilizing method can improve the robustness of landmark prediction in both static images and video frames. It should be emphasized that the proposed detection system exhibits better performance than others without requiring heavy computational loadings.

INDEX TERMS Facial landmark detection, lightweight U-Net, fast optical flow, dynamic routing, landmark stabilization.

I. INTRODUCTION

Over the past decade, several facial image applications have been widely developed such as facial recognition, facial enhancement, head pose estimation, facial expression recognition, face swapping, and face monitoring [1]–[3]. Facial landmark detection is a very important research topic in these applications but it poses many challenges such as blurred images, extreme lighting, artificial occlusion, extreme head posture, and data imbalance. To tackle these problems, active shape models have become one of the most popular traditional facial landmark methods [4]. Since an active shape model performs low-level matching along the edges, a high error rate of matching results gives a poor prediction performance. Furthermore, a local neural field patch expert [5], which can learn the similarity of surrounding pixels and the sparsity constraints of pixels, was proposed to solve the problem of matching failure. A convolutional expert network [6], which combined the advantages of neural architectures and mixtures of experts in an end-to-end framework, was proposed to improve the robustness of landmark prediction. However, it required time-consuming training to build a model for the appearance of each facial landmark and poor initial sample selection led to poor learning results.

Compared with traditional detection methods, facial landmark detection methods based on deep learning [7]–[13] showed better performance. There are two approaches, namely direct and indirect. Usually, convolutional neural networks (CNN) are used to detect the landmark coordinates [7]–[9] under the direct scheme, but the landmarks are obtained by post-processing the predicted heatmaps [10]–[13] under the indirect approach.

In [7], Sun *et al.* first applied CNN to facial landmark detection. A cascaded regression scheme was proposed to improve the accuracy of landmark detection based on the powerful feature extraction capabilities of CNN. In [8], a multi-task deep-learning method was proposed by

The associate editor coordinating the review of this manuscript and approving it for publication was Huanqing Wang.

optimizing the relationship between landmark coordinates and facial attributes. In [9], a fast multi-task framework was proposed with only the label of the coordinate generated. The feature map converted into coordinate values; however, the spatial information gradually disappeared or the entire landmark moved as interference from the environment occurred. On the other hand, since heatmap regression retains spatial information on the regression task, it has high spatial generalization ability and can improve the accuracy of landmark detection [10]–[13]. In [10], the landmark coordinates were obtained by using the boundary-aware face-alignment method. The boundary lines were utilized as the geometric structure of a human face to remove the ambiguity of landmark definition. In [11], a coarse-fine network coordination regression was proposed, in which the heatmap regression network branch used a spatial pyramid and attention mechanism to return better quality heatmaps. In [12] and [13], a multi-level network based on the convolutional pose architecture was proposed. Not only can the global context be merged to improve the part confidence map but also the position of each part can be executed at each stage.

Generally speaking, a large number of parameters in the network model help landmark prediction results, but the increased calculation time is not ideal for practical applications. A common approach to solve this problem is to employ parallel paths with different sizes of receptive fields [14]. Using a 1×1 convolutional layer for dimensionality reduction can greatly reduce the calculation time. A neural network compression method was proposed to increase the speed and reduce the model size [15], but it usually sacrifices some accuracy. Meanwhile, some studies focused on a design with a loss function generated by the heatmap to improve the accuracy of the model. In [16], an adaptive wing loss for the heatmap was proposed by considering the significance of every pixel on the heatmap. In [17], a fractional heatmap regression was proposed to accurately estimate the fractional part based on the 2D Gaussian function.

Though deep-learning-based facial landmark detectors can achieve favorable detection performance in static images, they usually display unnatural landmark shaking and unstable prediction results in video frames. The shaking problem is one of the most critical handicaps for applying facial landmarks to real-time applications. Some researchers proposed recurrent neural network (RNN) architectures [18]–[21] to improve the unnatural landmark shaking. Though RNNs are suitable for time-series analysis, they require a lot of resources to annotate each frame of information. Furthermore, a semi-supervised training using the correlation of optical flow as a source of supervision was proposed [22], but cannot be easily implemented. In [23], a Kalman filter was used to stabilize the facial position estimation; however, it cannot solve the landmark shaking problem when the face moves smoothly. In [24], a global and local filtering method was proposed based on the evaluation parameters of the global shaking value. It can ensure the robustness of global

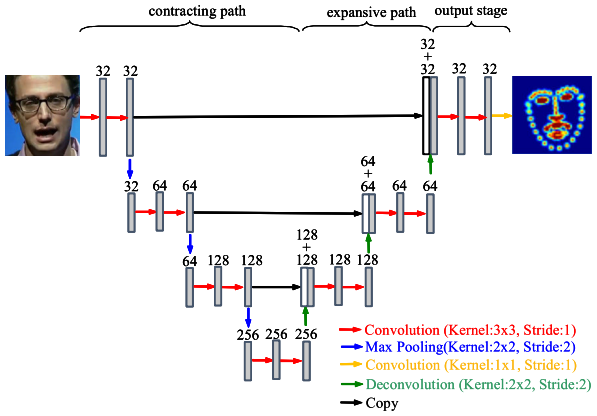

FIGURE 1. The network architecture of the proposed lightweight U-Net.

overall facial shape tracking and the adaptability of local facial part tracking.

Based on the above literature review, it can be seen that most of the existing facial landmark detection methods only consider the accuracy of face landmark location. Whether a detection system is robust against input image uncertainties is an important issue for a detection performance assessment. This study proposes a facial landmark detection method, which comprises a lightweight U-Net model and a dynamic optical flow (DOF). The DOF approach combines the advantages of fast optical flow and dynamic routing, thus it can greatly improve the accuracy and stability of landmark detection. The contributions of this study can be summarized as follows. (1) A lightweight U-Net is studied to predict facial landmark points with a simple, small, and lightweight network architecture. (2) A training trick with adaptive foreground ratio is proposed to avoid training failure. (3) A DOF approach is proposed to solve the problem of landmark shaking with high-quality detections in video frames. (4) Stability analysis is used to show the performance of the proposed DOF approach. (5) Several datasets including the 300W [25] and 300VW [26] datasets are used to show the effectiveness of the proposed facial landmark detection method.

The rest of the work is organized as follows. Section II describes the materials and Section III discusses the proposed facial landmark detection method with DOF. Some experimental tests and stability analysis are given in Section IV. The conclusion is given in Section V.

II. MATERIALS

A. LIGHTWEIGHT U-NET

It is important that facial landmark detection be simple and accurate as it is a necessity in various facial applications. Figure 1 shows the details of the lightweight U-Net architecture. It is pruned and quantized to improve the speed of the network based on the network architecture of the U-Net [27], [28]. At the last layer, a 1×1 convolution (yellow arrow) is used to map the feature vector to the number of heatmaps of the landmarks. Each heatmap can be viewed as a probability

TABLE 1. Input, output and network settings for the lightweight U-Net architecture. The kernels are described as height x width x depth.

Lightweight U-Net network setting					
Layer	Tag	Output Dim	Layer	Tag	Output Dim
0	Input	3x48x48	13	Concat1	256x12x12
1	Conv1-1	32x48x48	14	Conv5-1	128x12x12
2	Conv1-2	32x48x48	15	Conv5-2	128x12x12
3	Pool1	32x24x24	16	Dconv2	64x24x24
4	Conv2-1	64x24x24	17	Concat2	128x24x24
5	Conv2-2	64x24x24	18	Conv6-1	64x24x24
6	Pool2	64x12x12	19	Conv6-2	64x24x24
7	Conv3-1	128x12x12	20	Dconv3	32x48x48
8	Conv3-2	128x12x12	21	Concat3	64x48x48
9	Pool3	128x6x6	22	Conv7-1	32x48x48
10	Conv4-1	256x6x6	23	Conv7-2	32x48x48
11	Conv4-2	256x6x6	24	Output	68x48x48
12	Dconv1	128x12x12			

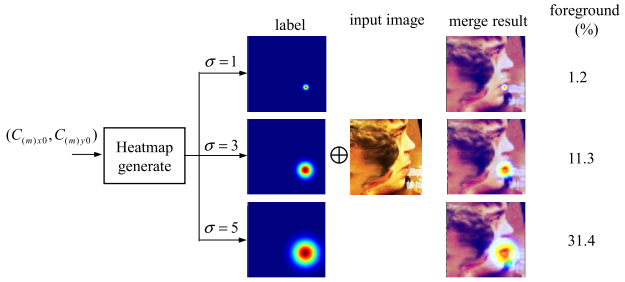

FIGURE 2. Percentage of the foreground with different values of σ.

response map. Thus, the position coordinates of the landmarks are obtained after finding the maximum value of the heatmap.

The advantage of this design is that the network can not only obtain local and global features of different proportions but also save the spatial information of each resolution. It reduces the merging steps and the number of channels in order to increase the speed of the network. Table 1 shows the detailed structure of the network and the size of the feature map. It should be emphasized that the basic topology of U-Net is not changed. The parameter amount of the original U-Net is approximately 30 million, but it is approximately 2 million in the developed lightweight U-Net. Thus, it can quickly and accurately predict facial landmarks with a simplified computation complexity.

Furthermore, Fig. 2 shows the percentages of the foreground for different values of σ. It can be seen that a large value of σ brings a large proportion of foreground into the picture. At the beginning of the training process, it often happens that the loss of certain landmarks remains fixed or the output shows all zeros. This is because the low percentage of the foreground causes the training to tend to become the background of the entire picture. Meanwhile, if the percentage of the foreground is large, it will cover too much unnecessary information and increase the uncertainty interval, which results in insufficient coordinate positioning.

To solve this problem, this study proposes a novel training trick to train the network as follows. When making the output labels, generate heatmaps of different sizes of Gaussian probability distributions in different training stages. Choose a larger value of σ to begin training the networks, and then gradually reduce the value of σ according to the learning effect. From a large range to a small range of precise positioning, the foreground ratio can be reduced. This coarse-to-fine training method can effectively avoid training failure. Thus, the proposed training trick can gradually improve the learning effect and convergence speed of lightweight U-Net.

B. SOFT-ARGMAX OPERATION

Heatmap regression accurately locates key points in the image for 2D pose estimation through pixel-by-pixel prediction. It can be viewed as a probability response map, where the location coordinates of the landmark are obtained by finding the maximum value of the heatmap [10]–[13]. Because the heatmap output is a 2D image, the model can be designed as a fully convolutional network. The output feature area is large and the spatial generalization ability is strong. The correlation between the heatmaps corresponding to the input image can be used to guide network learning. The heatmap regression captures the contrast between the foreground and the background and is used to improve network learning.

For the conversion between the heatmap and the coordinates, a soft-argmax function is used to alleviate the problem of accuracy loss caused by quantification. It is used to extract the locations of image key points. The soft-argmax calculation formula of the heatmap $M \in R^{W\times H}$ is defined as follows [29]:

$$\Psi_x(M) = \sum_{i=1}^{W}\sum_{j=1}^{H} \frac{i \times M_{i,j}}{\sum_{k=1}^{W}\sum_{l=1}^{H} M_{k,l}} \tag{1}$$

$$\Psi_y(M) = \sum_{i=1}^{W}\sum_{j=1}^{H} \frac{j \times M_{i,j}}{\sum_{k=1}^{W}\sum_{l=1}^{H} M_{k,l}} \tag{2}$$

where $M_{i,j}$ is the value of the heatmap M at position (i, j), and $W \times H$ is the size of the heatmap output. Thus, the predicted regression position for the given heatmap M is:

$$P_k = (\Psi_x(M), \Psi_y(M)) \tag{3}$$

For the labelling part of the network training procedure, a heatmap with a Gaussian probability distribution can be generated in terms of the landmark coordinates as follows [29]:

$$H_{(m)}(C) = e^{-\frac{(C_x - C_{(m)x0})^2 + (C_y - C_{(m)y0})^2}{2\sigma^2}} \tag{4}$$

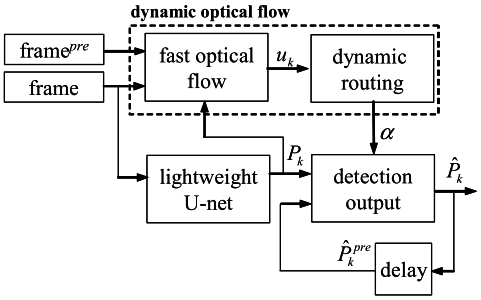

FIGURE 3. Overview of the proposed detection framework.

where $C = (C_x, C_y)$ represents the coordinates, σ represents the standard deviation, and $(C_{(m)x0}, C_{(m)y0})$ means the center coordinates of the m-th heatmap Gaussian distribution.

III. METHOD

As shown in Fig. 3, the proposed facial landmark detection system comprises a lightweight U-Net and a DOF. Since the global optical flow can determine the overall movement direction of the face and the local optical flow can avoid losing sensitivity to expressions, the DOF approach calculates the correlation information between local optical flow and global optical flow. The DOF uses the fast optical flow [30] to obtain the optical flow vector of the landmark and uses dynamic routing [31] to improve landmark stabilization. Fast optical flow can effectively reduce the coupling coefficient that is inconsistent with the overall direction. Thus, important landmarks that fit the entire vector can be selected to eliminate unstable landmarks.

According to the coarse-to-fine method, the fast optical flow establishes a scale pyramid, where the iteration starts from the first (coarsest) level θ_{ss} in a scale pyramid with a downscaling quotient of θ_{sd} to the final (finest) level θ_{sf}. A dense flow field U_s in each iteration s can be obtained as follows. First, the uniform grid patch in the image domain is initialized with the trivial zero flow. The number of patches $N_i \times N_j$ and grid density is determined through parameters $\theta_{ov} \in [0, 1)$, where $\theta_{ov} = 0$ denotes patch adjacency with no overlap and $\theta_{ov} = 1 - \varepsilon$ takes each pixel on the reference image as the center of the patch to form a dense grid. On each subsequent scale, use the flow from the previous scale $u_{ij,init} = U_{s+1}(x/\theta_{sd})\theta_{sd}$ to initialize the displacement of each patch $ij \in N_i, N_j$ at its location x and update the optical flow vectors u_{ij}. In order to obtain more robustness against outliers, reset all patches to their initial flow $u_{ij,init}$ to update the displacement $||u_{ij,init} - u_{ij}||_2$ until it exceeds the patch size. In the reference image I_t, the intensity difference between the template patch and warped image at this pixel for a given template patch T is obtained as:

$$d_{ij}(x) = I_{t+1}(x + u_{ij}) - T(x) \tag{5}$$

where u_{ij} is the estimated displacement of patch ij. Furthermore, the indicator $\lambda_{ij,x} = 1$ iff patch ij overlaps with location x in the reference image. The normalization parameter Z is given as:

$$Z = \sum_{i}^{N_i} \sum_{j}^{N_j} \frac{\lambda_{ij,x}}{\max(1, ||d_{ij}(x)||_2)} \tag{6}$$

Applying the weighted average method, the estimated displacements of all patches overlapping at each pixel x in the reference image can be used to obtain the dense flow field U_s as follows:

$$U_s(x) = \frac{1}{Z} \sum_{i}^{N_i} \sum_{j}^{N_j} \frac{\lambda_{ij,x}}{\max(1, ||d_{ij}(x)||_2)} u_{ij} \tag{7}$$

Furthermore, the dynamic routing can effectively distribute the output vector of each node to the next node through the coupling coefficient [31]. First, initialize dynamic routing as follows:

$$b_k \leftarrow 0, c_k = \frac{1}{N} \tag{8}$$

where b_k is the initial base, c_k is the coupling coefficient of each vector and N is the number of landmarks. Obtain each optical flow vector u_k from the dense flow field U_s and then use initialization to start the iteration of dynamic routing, as follows:

$$s = \sum_{k=1}^{N} c_k u_k \tag{9}$$

where s is the sum of the inputs of all input nodes. The coupling coefficient is used to update the coefficient through the inner product of the prediction vector and the output vector of each node. The output vector of the node v_g can be obtained as follows:

$$v_g = \frac{s}{\|s\|} \tag{10}$$

The k-th base vector can be further updated as:

$$b_k \leftarrow b_k + \frac{u_k^T v_g}{\|u_k\|} \tag{11}$$

and the k-th coupling coefficient c_k is given as:

$$c_k = \frac{e^{b_k}}{\sum_{l=1}^{N} e^{b_l}} \tag{12}$$

It can be found that the coupling coefficient of this node will be increased through feedback from top to bottom if the inner product is large, yet the coupling coefficients of other nodes will be reduced. After continuous iterative approximation, the best coupling coefficient c_k can be obtained with dynamic selection. The weight α can be obtained using the best coupling coefficient between the front and rear frames as follows:

$$\alpha = \begin{cases} 1 - \sum_{k=1}^{N} c_k \|u_k\| & \text{if } \sum_{k=1}^{N} c_k \|u_k\| > 0 \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

TABLE 2. Details of the used benchmark databases.

Name	Year	Images	Subjects	Landmarks	Description
BP4D	2014	-	41	49 pts	
300VW	2015	218,595	-	68 pts	Video base
300W	2016	689	-	68 pts	AU code, 2D/3D Landmark, head pose
300W-LP	2016	61225	-	68 pts	Larger pose
BP4D+	2016	-	140	49 pts	AU code, 2D/3D/IR Landmark, head pose, physiological signal

The k-th estimated landmark coordinate $\hat{P}_k$ can be obtained as follows:

$$\hat{P}_k = (1-\alpha)P_k + \alpha\hat{P}_k^{pre} \tag{14}$$

where $w_i^k P_k$ is the k-th predicted landmark coordinate and $\hat{P}_k^{pre}$ is the k-th estimated landmark coordinate of the previous frame image. It should be emphasized that the DOF method can be used for other facial landmark detections.

In summary, the steps of the proposed facial landmark detection method are described below:

Step 1: Load an image into the facial landmark detection system.

Step 2: Use a learned lightweight U-Net to find the predicted facial landmark coordinate P_k.

Step 3: Obtain the optical flow vector u_k of each landmark between frames by using fast optical flow.

Step 4: Use dynamic routing to obtain the global optical flow vector v_g for iteration of each vector.

Step 5: Calculate the weight α between the front and rear frames in (13).

Step 6: Obtain the estimated facial landmark coordinate $\hat{P}_k$ in (14).

IV. EXPERIMENTAL RESULTS

A. DATASETS AND EXPERIMENTAL SETUP

Five benchmark datasets for training and testing, 300W [25], 300VW [26], BP4D [32], BP4D+ [33], and 300W-LP [34], are utilized as summarized in Table 2. On the 300W dataset, each face has 68 landmarks and the dataset is divided into four groups: training set, common set, challenging set, and full set. The 300W-LP dataset is an extended version of 300W to provide more images with larger poses. The 300-VW dataset includes 114 lengthy videos (approx. 1 min each) with 68 landmarks annotated densely. In the BP4D dataset, eight tasks in the form of interviews are designed to generate spontaneous emotions and the BP4D+ dataset increases the number of tasks to ten. Furthermore, this study utilizes the Caffe framework to train the lightweight U-Net model with the optimal learning rate between 10^{-4} and $10^{-6}10^{-6}$ and a batch size setting of 256. The input image is 48 $\times$ 48 pixels with several data augmentation methods that include blur, zoom, translation, rotation, light and dark, and flip, crop and zoom. When the loss function uses adaptive wing loss, the penalty loss for foreground pixels is larger, but the loss for background pixels is smaller, so the training effect is better.

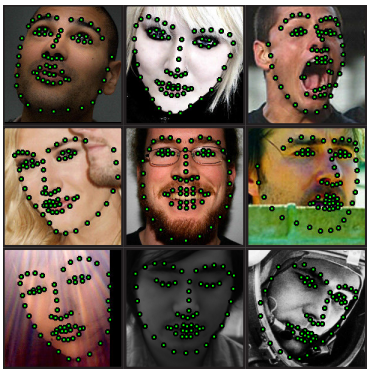

FIGURE 4. Results of predicted landmarks on 300W dataset.

To evaluate the proposed method, a normalized mean error (NME%) is defined as follows:

$$NME\% = \frac{\sum_{k=1}^{N} \|x_k - x_k^*\|_2}{Nd} \tag{15}$$

where x_k is the predicted coordinates of the k-th landmark, x_k^* is the ground truth of the k-th landmark, N is the total number of landmarks, and d is defined as the distance between the left and right corners of the eye and the average distance between twelve landmarks in the binocular region for ION and IPN, respectively.

B. EVALUATION ON DATASETS

The 300W and 300VW datasets are employed to evaluate the effectiveness of the proposed detection method. First, the proposed facial landmark detection is applied to the 300W dataset and the experimental results are given in Fig. 4. It can be seen that the proposed method is robust to partial occlusion, head pose, and extreme expressions. Furthermore, a comparison of cascade regression [35]–[37], the coordinate or heatmap regression [38], [39], the 3D shape model [38], and the proposed method is made in Table 3. It shows that the NME% of the proposed method is better than that of most methods, reaching a similar level as [37], but worse than that of [38] and [39]. It should be emphasized that the used network model only requires 7 MB and does not occupy too much space on the storage device, while the network size in [39] is 798.5 MB. This simple design significantly helps to reduce the computational burden.

Next, the proposed facial landmark detection is applied to the 300VW dataset and the experimental results are shown in Fig. 5. It can be seen that the proposed method achieves satisfactory prediction results for each frame even if the head posture and light and shadow change between video frames. Table 4 summarizes the experimental results of SDM [41], CFSS [37], TCDCN [8], DSRN [42], and the proposed method. Compared with other methods, the proposed method can achieve better measurement values,

TABLE 3. The NME(%) of facial landmark detection results on 300W. The lower value is better.

Method	Common	Challenging	Fullset
Inter-pupil Normalization(IPN)			
RCPR [35]	6.18	17.26	8.35
TCDCN [8]	4.80	8.60	5.54
CFAN [36]	5.50	16.78	7.69
CFSS [37]	4.73	9.98	5.76
3DDFA [34]	6.15	10.59	7.01
Openpose[12]	8.4	12.67	9.23
DeFA [40]	5.37	6.38	6.10
LAB(8-stack) [10]	3.42	6.98	4.12
DAN [38]	3.19	5.24	3.59
SAN [39]	3.34	6.60	3.98
Ours	**4.68**	**9.13**	**5.55**

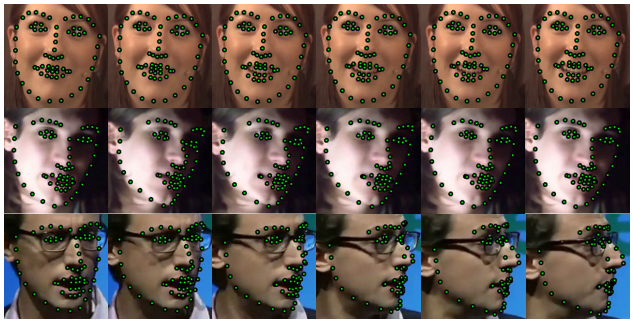

FIGURE 5. Results of predicted landmarks on 300VW dataset.

TABLE 4. The NME(%) of facial landmark detection results on 300VW. The lower value is better.

Method	Category1	Category2	Category3
Inter-ocular Normalization(ION)			
SDM [41]	7.41	6.18	13.04
Openpose [12]	8.5	7.42	21.69
CFSS [37]	7.68	6.42	13.67
TCDCN [8]	7.66	6.77	14.98
AAN [43]	5.03	4.82	7.98
DSRN [42]	5.33	4.92	8.85
SA [44]	3.85	3.46	7.51
DeCaFA [45]	3.82	3.63	6.67
AND [46]	4.69	4.34	6.72
Ours	**3.64**	**3.8**	**5.03**

and the most difficult Category 3 can also be increased by approximately 30%. According to these results, the proposed detection method can efficiently improve inference speed and maintain performance.

C. STABILITY ANALYSIS

Two testing scenarios are considered to analyze the stability performance of the DOF approach. Scenario 1 is a static testing scenario used to observe the effects of processing by the DOF method in static images, and Scenario 2 is a dynamic testing scenario used to observe the effect of the DOF method in video frames. In both testing scenarios, comparisons with the Dlib [47], mean filter (MF), first-order filter (FF), second-order filter (SF), and DOF are made. For the static testing scenario, a stationary image is obtained from the photo recording experiment. The experimental results of the static testing scenario are shown in Figs. 6 and 7, where the landmark average of the entire video is used as

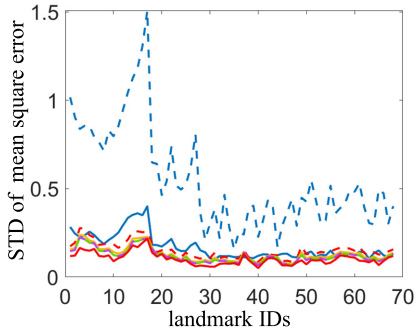

FIGURE 6. Comparison variation of each landmark mean square error in static experiment (▬ ▬: Dlib, ▬: Dlib with DOF, ▬ ▬: lightweight UNet, ▬: lightweight U-Net with MF, ▬ ▬: lightweight U-Net with FF, ▬ ▬: lightweight U-Net with SF, ▬: lightweight U-Net with DOF).

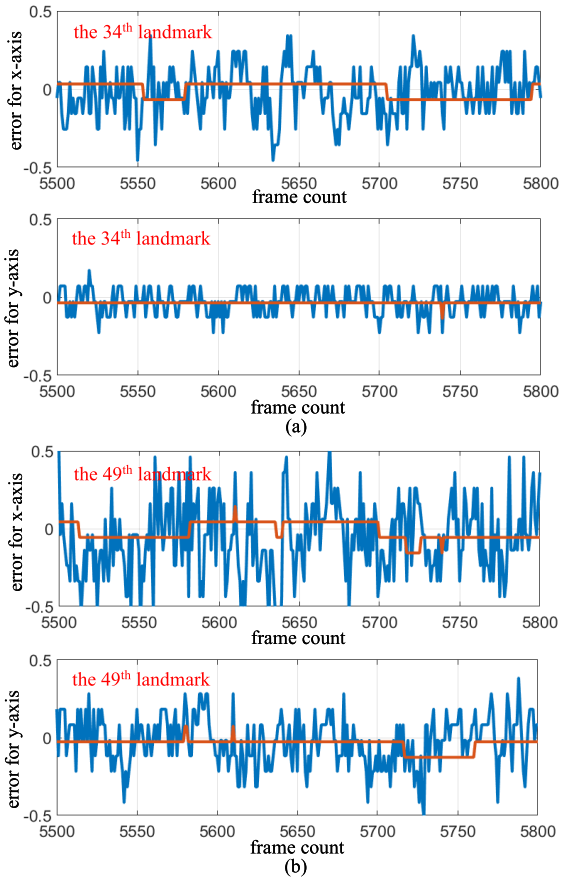

FIGURE 7. Landmark error for static testing scenario (▬: lightweight U-Net, ▬: lightweight U-Net with DOF).

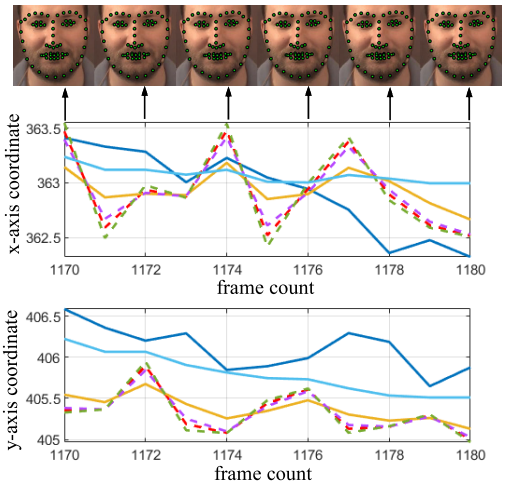

FIGURE 8. Stability analysis for dynamic testing scenario (▬: ground truth, ▬ ▬: lightweight U-Net, ▬: lightweight U-Net with MF, ▬ ▬: lightweight U-Net with FF, ▬ ▬: lightweight U-Net with SF, ▬: lightweight U-Net with DOF).

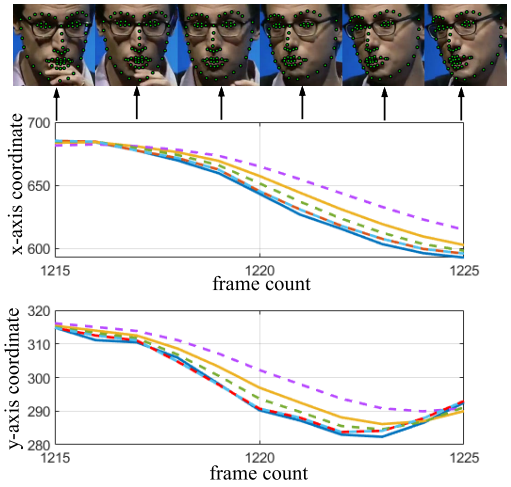

FIGURE 9. Time-delay analysis for dynamic testing scenario (▬: ground truth, ▬ ▬: lightweight U-Net, ▬: lightweight U-Net with MF, ▬ ▬: lightweight U-Net with FF, ▬ ▬: lightweight U-Net with SF, ▬: lightweight U-Net with DOF).

the ground truth. Figure 6 shows that both the Dlib and lightweight U-Net cause unstable shaking for each point but the shaking amplitude of the lightweight U-Net is smaller than that of Dlib. It can be seen that the DOF approach achieves better performances than other approaches such as MF, FF, and SF. Meanwhile, the time responses of the 34th and 49th landmark points for the x-axis and y-axis are shown in Figs. 7(a) and (b), respectively. It can be seen that the DOF approach can solve the problem of facial landmark shaking. It should be emphasized that the DOF method can be used for other facial landmark detections. The detection performance of the Dlib also can be improved by using the DOF approach.

For the dynamic testing scenario, the 300VW dataset is used to observe landmark shaking. The experimental results for the dynamic testing scenario are shown in Figs. 8 and 9 for stability analysis and time-delay analysis, respectively. Figure 8 shows the results of stability analysis of No. 114 in Category 1. The original prediction has a serious shaking problem, but this can be improved after considering with DOF. Compared with the ground truth and the lightweight U-Net without DOF, the curve between frames becomes more stable by using the lightweight U-Net with DOF. Meanwhile, Figure 9 shows the time-delay analysis of No. 411 in Category 3. The original prediction has a serious time-delay problem, and the lightweight U-Net with SF cannot handle instantaneous changes. Since the DOF can dynamically adjust the weight according to the degree of movement, the lightweight U-Net with DOF can reduce time delay and can get closer to the ground truth data.

V. CONCLUSION

Accurate and stable facial landmark detection presents a considerable challenge for researchers in the field of computer vision. This study proposed a simple yet effective solution for facial landmark detection. The proposed facial landmark detection system comprised a lightweight U-Net and a dynamic optical flow (DOF) to produce high-quality detections. The lightweight U-Net was designed to predict facial landmarks and reduce model size and computational complexity without sacrificing model accuracy. Furthermore, the proposed DOF integrated the global and local structural constraints of facial landmarks to calculate the motion of objects between frames. Although the lightweight U-Net model achieved good prediction performance in static images, unnatural landmark shaking occurred in video frames. To deal with the unstable shaking, the predicted facial landmarks using lightweight U-Net were further inputted to the DOF approach. Finally, two benchmark datasets (300W and 300VW) were applied to verify the effectiveness of the proposed facial landmark detection system. Experimental evaluations and analyses showed that the proposed detection method not only needed less memory space but also effectively suppressed landmark shaking.

REFERENCES

[1] S. Gu, J. Bao, H. Yang, D. Chen, F. Wen, and L. Yuan, "Mask-guided portrait editing with conditional GANs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3431–3440.

[2] Y.-C. Tsai, P.-W. Lai, P.-W. Huang, T.-M. Lin, and B.-F. Wu, "Vision-based instant measurement system for driver fatigue monitoring," *IEEE Access*, vol. 8, pp. 67342–67353, 2020.

[3] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE Trans. Affect. Comput.*, early access, Mar. 17, 2020, doi: 10.1109/TAFFC.2020.2981446.

[4] J. M. Saragih, S. Lucey, and J. F. Cohn, "Deformable model fitting by regularized landmark mean-shift," *Int. J. Comput. Vis.*, vol. 91, no. 2, pp. 200–215, Jan. 2011.

[5] T. Baltrusaitis, P. Robinson, and L.-P. Morency, "Constrained local neural fields for robust facial landmark detection in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Dec. 2013, pp. 354–361.

[6] A. Zadeh, T. Baltrusaitis, and L.-P. Morency, "Convolutional experts constrained local model for facial landmark detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 2051–2059.

[7] Y. Sun, X. Wang, and X. Tang, "Deep convolutional network cascade for facial point detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3476–3483.
[8] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Facial landmark detection by deep multi-task learning," in *Proc. Eur. Conf. Comput. Vis., Sept.*, vol. 2014, pp. 94–108.
[9] Z. Cai, Q. Liu, S. Wang, and B. Yang, "Joint head pose estimation with multi-task cascaded convolutional networks for face alignment," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2018, pp. 495–500.
[10] W. Wu, C. Qian, S. Yang, Q. Wang, Y. Cai, and Q. Zhou, "Look at boundary: A boundary-aware face alignment algorithm," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2129–2138.
[11] S. Zheng, X. Bai, L. Ye, and Z. Fang, "HafaNet: An efficient Coarse-to-Fine facial landmark detection network," *IEEE Access*, vol. 8, pp. 123037–123043, 2020.
[12] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 172–186, Jan. 2021, doi: 10.1109/TPAMI.2019.2929257.
[13] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4724–4732.
[14] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.
[15] M. Rastegari, V. Ordonez, J. Redmon, and A. Farhadi, "XNOR-Net: Imagenet classification using binary convolutional neural networks," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 525–542.
[16] X. Wang, L. Bo, and L. Fuxin, "Adaptive wing loss for robust face alignment via heatmap regression," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6971–6981.
[17] Y. Tai, Y. Liang, X. Liu, L. Duan, J. Li, C. Wang, F. Huang, and Y. Chen, "Towards highly accurate and stable face alignment for high-resolution videos," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 8893–8900.
[18] H. Liu, J. Lu, J. Feng, and J. Zhou, "Two-stream transformer networks for video-based face alignment," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 11, pp. 2546–2554, Nov. 2018.
[19] R. Belmonte, N. Ihaddadene, P. Tirilly, I. M. Bilasco, and C. Djeraba, "Video-based face alignment with local motion modeling," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2019, pp. 2106–2115.
[20] Y. Chen, J. Qian, J. Yang, and Z. Jin, "Face alignment with cascaded bidirectional LSTM neural networks," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2016, pp. 313–318.
[21] T. Yang, S. Qin, J. Yan, and W. Zhang, "Multi-label dilated recurrent network for sequential face alignment," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2018, pp. 1–6.
[22] X. Dong, S.-I. Yu, X. Weng, S.-E. Wei, Y. Yang, and Y. Sheikh, "Supervision-by-registration: An unsupervised approach to improve the precision of facial landmark detectors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 360–368.
[23] M. Uricar, V. Franc, and V. Hlavac, "Facial landmark tracking by tree-based deformable part model based detector," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW)*, Dec. 2015, pp. 10–17.
[24] X. Guo, Y. Jin, Y. Li, J. Xing, and C. Lang, "Stabilizing video facial landmark detection and tracking via global and local filtering," in *Proc. 10th Int. Conf. Internet Multimedia Comput. Service (ICIMCS)*, 2018, pp. 1–7.
[25] 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.
[26] J. Shen, S. Zafeiriou, G. G. Chrysos, J. Kossaifi, G. Tzimiropoulos, and M. Pantic, "The first facial landmark tracking in-the-wild challenge: Benchmark and results," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW)*, Dec. 2015, pp. 1003–1011.
[27] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2015, pp. 234–241.
[28] A. Kumar and R. Chellappa, "S2LD: Semi-supervised landmark detection in low resolution images and impact on face verification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 3275–3283.
[29] D. C. Luvizon, H. Tabia, and D. Picard, "Human pose regression by combining indirect part detection and contextual information," Oct. 2017, *arXiv:1710.02322*. [Online]. Available: http://arxiv.org/abs/1710.02322
[30] T. Kroeger, R. Timofte, D. Dai, and L. V. Gool, "Fast optical flow using dense inverse search," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 471–488.
[31] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 3859–3869.
[32] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, and J. M. Girard, "BP4D-spontaneous: A high-resolution spontaneous 3D dynamic facial expression database," *Image Vis. Comput.*, vol. 32, no. 10, pp. 692–706, Oct. 2014.
[33] Z. Zhang, J. M. Girard, Y. Wu, X. Zhang, P. Liu, U. Ciftci, S. Canavan, M. Reale, A. Horowitz, H. Yang, J. F. Cohn, Q. Ji, and L. Yin, "Multimodal spontaneous emotion corpus for human behavior analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3438–3446.
[34] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li, "Face alignment across large poses: A 3D solution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 146–155.
[35] X. P. Burgos-Artizzu, P. Perona, and P. Dollár, "Robust face landmark estimation under occlusion," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1513–1520.
[36] J. Zhang, S. Shan, M. Kan, and X. Chen, "Coarse-to-fine auto-encoder networks (CFAN) for real-time face alignment," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 1–16.
[37] S. Zhu, C. Li, C. C. Loy, and X. Tang, "Face alignment by coarse-to-fine shape searching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4998–5006.
[38] M. Kowalski, J. Naruniec, and T. Trzcinski, "Deep alignment network: A convolutional neural network for robust face alignment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 2034–2043.
[39] X. Dong, Y. Yan, W. Ouyang, and Y. Yang, "Style aggregated network for facial landmark detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 379–388.
[40] Y. Liu, A. Jourabloo, W. Ren, and X. Liu, "Dense face alignment," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 1619–1628.
[41] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 532–539.
[42] X. Miao, X. Zhen, X. Liu, C. Deng, V. Athitsos, and H. Huang, "Direct shape regression networks for end-to-end face alignment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5040–5049.
[43] L. Yue, X. Miao, P. Wang, B. Zhang, X. Zhen, and X. Cao, "Attentional alignment networks," in *Proc. BMVC*, 2018, pp. 1–14.
[44] Z. Liu, X. Zhu, G. Hu, H. Guo, M. Tang, Z. Lei, N. M. Robertson, and J. Wang, "Semantic alignment: Finding semantically consistent ground-truth for facial landmark detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3467–3476.
[45] A. Dapogny, M. Cord, and K. Bailly, "DeCaFA: Deep convolutional cascade for face alignment in the wild," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6893–6901.
[46] M. Sadiq, D. Shi, M. Guo, and X. Cheng, "Facial landmark detection via attention-adaptive deep network," *IEEE Access*, vol. 7, pp. 181041–181050, 2019.
[47] D. E. King, "Dlib-ml: A machine learning toolkit," *J. Mach. Learn. Res.*, vol. 10, pp. 1755–1758, Jan. 2009.

BING-FEI WU (Fellow, IEEE) received the B.S. and M.S. degrees in control engineering from National Chiao Tung University (NCTU), Hsinchu, Taiwan, in 1981 and 1983, respectively, and the Ph.D. degree in electrical engineering from the University of Southern California, Los Angeles, in 1992.

Since 1992, he has been with the Department of Electrical and Computer Engineering, where he was promoted to be a Professor, in 1998, and the Chair Professor, in 2020. He is also the President of the Taiwan Association of System Science and Engineering, in 2019, and the Director of the Control Engineering Program, Ministry of Science and Technology, Taiwan, in 2019. He has served as the Director of the Institute of Electrical and Control Engineering, NCTU, in 2011. His research interests include image recognition, physiological informatics, vehicle driving safety and control, intelligent robotic systems, and intelligent transportation systems. He received many research honors and awards, including the FutureTech Breakthrough Award, from 2017 to 2019, and the Outstanding Research Award, in 2015 and 2019, all from the Ministry of Science and Technology, Taiwan; the Technology Invention Award of Y. Z. Hsu Scientific Award from Y. Z. Hsu Foundation, in 2014; the National Invention and Creation Award of Ministry of Economic Affairs, Taiwan, in 2012 and 2013; the Outstanding Research Award of Pan Wen Yuan Foundation, in 2012; the Best Paper Award in the 12th International Conference on ITS Telecommunications, in 2012; the Best Technology Transfer Contribution Award from National Science Council, Taiwan, in 2012 and 2020; and the Outstanding Automatic Control Engineering Award from the Chinese Automatic Control Society, in 2007. He founded the Taipei Chapter of IEEE Systems, Man and Cybernetics Society (SMCS), in 2003, and was also the Chair of the Technical Committee on Intelligent Transportation Systems of IEEE SMCS, in 2011. He has been the Director of Board of Governors of IEEE Taipei Section, since 2012. He is also an Associate Editor of the IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS: SYSTEMS.

BO-RUI CHEN received the B.S. and M.S. degrees in electrical engineering from Tamkang University, New Taipei City, Taiwan, in 2016 and 2018, respectively. He is currently pursuing the Ph.D. degree in electrical and control engineering with National Yang Ming Chiao Tung University. His research interests include computer vision, machine learning, and intelligent control using fuzzy systems.

CHUN-FEI HSU received the B.S., M.S., and Ph.D. degrees in electrical engineering from Yuan-Ze University, Taiwan, in 1997, 1999, and 2002, respectively. From 2007 to 2011, he joined the faculty of the Department of Electrical Engineering, Chung Hua University, Taiwan, where he was appointed as an Associate Professor. In 2012, he joined the faculty of the Department of Electrical Engineering, Tamkang University, Taiwan, where he is currently a Professor of electrical engineering. His research interests include intelligent robotic systems and intelligent control systems using fuzzy systems and neural network technologies. The outstanding achievement of his research is for contributions to real-time intelligent control in practical applications. He was a recipient of the Young Automatic Control Engineering Award from the Chinese Automatic Control Society, in 2007, and the Outstanding Young Award from the Taiwan Association of Systems Science and Engineering, in 2011.