Thermal Facial Landmark Detection by Deep Multi-Task Learning

Wei-Ta Chu
National Cheng Kung University
Tainan, Taiwan
wtchu@gs.ncku.edu.tw

Yu-Hui Liu
National Chung Cheng University
Chiayi, Taiwan
zoeydales@gmail.com

***Abstract*—We present a neural network to jointly consider facial landmark detection and emotion recognition for thermal face images. The first part of this network is based on the U-Net structure, targeting at extracting good features for advanced analysis. Using U-Net as the basic structure enables modeling context information based on a limited number of training data. The second part of this network contains two branches that are designed for landmark detection and emotion recognition, respectively. We propose a two-stage training mechanism to learn this network, and demonstrate the effectiveness of the proposed approach. This work is believed to be one of the few studies on thermal face image analysis.**

***Index Terms*—Thermal face images, facial landmark detection, emotion recognition, multi-task learning**

I. Introduction

Thermal image analysis attracts more and more attention in recent years because its potential in nighttime surveillance and privacy-preserving access control. In the infra-red spectrum, images are especially called thermal images when they are formed by sensing light with wavelengths ranging from 3 μm to 14 μm. Thermal face images are formed by passive thermal sensors receiving thermal signatures emitted by skin tissues. Therefore, they usually yield severe challenges because of the significant gap between the visible spectrum and the infra-red spectrum.

Currently, most thermal image studies are about thermal face recognition [1] [2]. These works originate from widely studied visible/near infra-red face recognition, and have been well recognized as an important problem. In this paper, we target at facial landmark detection and emotion recognition. Although there have been many related works for faces in the visible spectrum, to our best knowledge, very few of them were designed for thermal faces. Kopaczka et al. [3] proposed one of the earliest works on thermal facial landmark detection based on active appearance models. They later proposed a modular system for face detection, face tracking, head pose estimation, and emotion recognition for thermal faces [4]. The reason to detect landmarks on thermal faces is that it is a fundamental component of face tracking or face alignment, which are basic modules for advanced thermal image understanding. For example, Kopaczka et al. [5] studied temperature changes in specific facial regions when mental stress is induced. On the other hand, emotion recognition is an essential component for advanced behavior analysis. With the request of privacy preserving, we think analyzing emotion on thermal face images would also be a potential research topic.

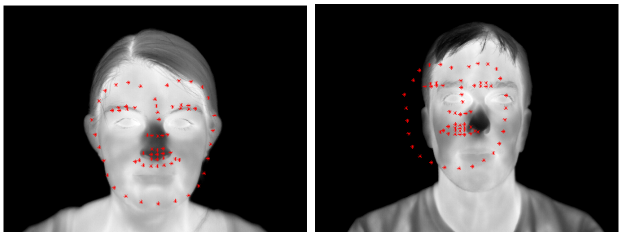

Fig. 1. Facial landmark detection results of directly employing a visible facial landmark detector. Positions of landmarks significantly deviate from the truths.

Because many facial landmark detectors have been designed for face images in the visible spectrum, ones may wonder the performance of directly applying them to thermal faces. Figure 1 shows two results of directly employing the detector proposed in [6] and implemented in the Dlib library [1]. The red dots are positions of facial landmarks. As can be seen, positions of the detected landmarks largely deviate from the truths.

As the rapid development of image transfer models, ones may also wonder the performance when we first transform thermal faces into visible faces, and then apply a visible facial landmark detector to the transformed visible faces. To check this idea, we construct a CycleGAN [7] based on the UND Collection X1 thermal face dataset[2], in order to transfer thermal faces into visible faces. After we get the transferred visible face, the detector proposed in [6] is used to detect facial landmarks. Figure 2 shows an example of this sequence of processes. Given the thermal face shown in Figure 2(a), the constructed CycleGAN transfers it into Figure 2(b). Comparing Figure 2(b) with Figure 2(d), we can see the difference between the transferred result and the real visible face. This difference remains existing even if the state-of-the-art generative model is used. Figure 2(c) shows facial landmark detection results, where red crosses indicate the ground truths, and blue crosses indicate the detected facial

[1] http://dlib.net

[2] https://cvrl.nd.edu/projects/data/

978-1-7281-1817-8/19/$31.00 ©2019 IEEE

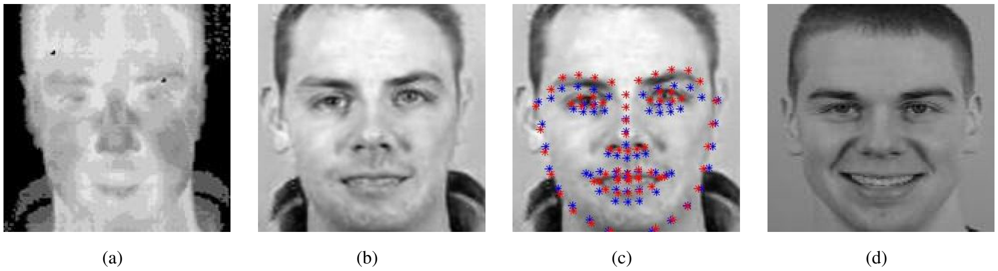

Fig. 2. An example of facial landmark detection results for a transferred visible face. (a) the given thermal face, (b) the visible face transferred from the thermal face by CycleGAN, (c) results of facial landmark detection for the transferred visible face, and (d) the real visible face corresponding to the the given thermal face.

landmarks. These results are much better than Figure 1, but the detection errors on this subject's right eye and eyebrow are still large.

Based on the discussion mentioned above, we realize that specifically designing a facial landmark detector for thermal faces is necessary. We thus focus on thermal facial landmark detection in this paper. Kopaczka's works [3] [4] were based on active appearance models. Considering the effectiveness of deep-based methods, we would like to propose a deep framework to achieve facial landmark detection and emotion recognition for thermal faces. Furthermore, motivated by the multi-task approach [8], we would jointly learn image representation and prediction models for two tasks. We will demonstrate the effectiveness of the proposed network, which is one of the earliest works on such research topics for thermal faces.

The rest of this paper is organized as follows. In Section II, we describe the framework to achieve facial landmark detection for thermal faces. This is to show the essential idea of a deep-based method. We extend this framework to multiple tasks in Section III. In Section IV, we demonstrate the overall performance and effectiveness of multi-task learning, followed by Section V concluding this work.

II. THERMAL FACIAL LANDMARK DETECTION

We first formulate the task of thermal facial landmark detection. Given an image I, we would like to find a function $\mathcal{F}$ that outputs the set of positions of K facial landmarks $\mathcal{F}(I) = L = \{\boldsymbol{\ell}_1, \boldsymbol{\ell}_2, ..., \boldsymbol{\ell}_K\}$. The position of the ith landmark $\boldsymbol{\ell}_i$ is represented as a two-dimensional vector (x_i, y_i). In this work, we will learn a deep neural network to find the function $\mathcal{F}$ in an end-to-end manner. Conceptually, this network consists of two parts: feature extraction and landmark position prediction. Particularly, given an image I with resolution $N \times N$ pixels, we flatten it as a N^2-dimensional vector $\boldsymbol{I}$ being the input of the developed neural network. The targeted output, i.e., the K facial landmarks, is represented as the concatenation of these K two-dimensional vectors, i.e., $\boldsymbol{L} = (\boldsymbol{\ell}_1, \boldsymbol{\ell}_2, ..., \boldsymbol{\ell}_K)$, which is $2K$-dimensional. The neural network acts as the function $\mathcal{F}$ to predict $\boldsymbol{L}$ based on $\boldsymbol{I}$, i.e., $\boldsymbol{L} = \mathcal{F}(\boldsymbol{I})$. In the evaluation dataset, we have 68 landmarks on each face, i.e., $K = 68$.

We adopt the U-Net structure [9] to develop the function $\mathcal{F}$. Figure 3 shows the network structure for predicting landmark positions. Taking convolutional layers as basic building components, there are a contracting path (left side) and an expansive path (right side) in the U-Net. The contracting path is a conventional convolutional neural network, where the convolutional kernel is 3×3 with stride 2, the activation function is ReLU, followed by a 2×2 max pooling for downsampling. Following the setting in [9], we double the number of feature channels at each downsampling step. On the contrary, in the expansive path we halve the number of feature channels at each upsampling step, by 2×2 deconvolution. Every step in the expansive path contains upsampling, a concatenation with the correspondingly feature map from the contracting path, and two convolutional layers with the same settings in contracting. Notice that the last convolutional layer is with 1×1 convolution kernel, and the size of feature map is $N \times N$, which is the same as the input image. The U-Net proposed in [9] was originally designed for medical image segmentation. To modify U-Net for facial landmark detection, we view the contracting and expansive paths as the components for feature extraction, and connect two fully-connected layers at the end of the expansive path to be prediction model, as shown in the top-right corner of Figure 3. These two fully-connected layers have 1024 and 136 nodes ($K \times 2$), respectively, with the sigmoid function as the activation function.

The reasons to adopt U-Net are twofold. First, U-Net was originally designed and trained based on a limited number of medical images. The basic components of U-Net are convolutional layers, and thus relatively fewer parameters need to be determined. In our case, we only have a limited number of thermal face images as well, and taking a structure similar to U-Net is advantageous to model training. Second, in U-Net the feature maps from the contracting step are concatenated with feature maps from the corresponding expansive step, and context information can be modeled. Besides, the symmetric structure enables more accurate localization [9].

To train the network shown in Figure 3, we propose a two-stage training scheme. At the first stage, we target at finding good initial parameters for the contracting path and the expansive path (excluding the last two fully-connected layers), based on the *unet loss*. At the second stage, we jointly train the U-Net structure and the last two fully-connected layers,

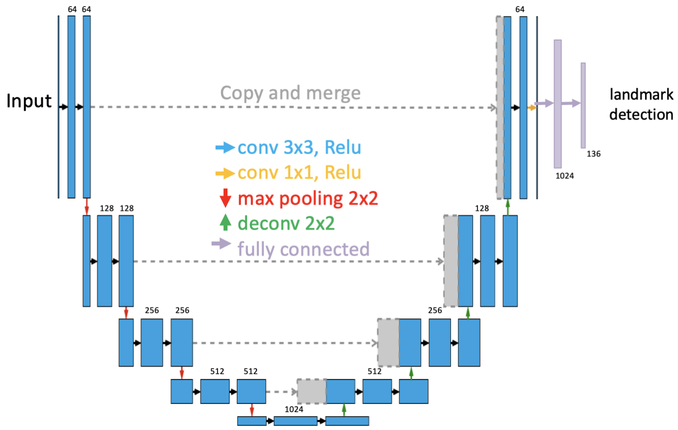

Fig. 3. The developed network structure for thermal facial landmark detection based on U-Net.

based on the *landmark loss*.

Denote the output $N \times N$ feature map of the last convolutional layer as $A = \{a_{jk}\}_{N\times N}$. Assume that the ground truth of facial landmarks is $\boldsymbol{L} = (\boldsymbol{\ell}_1, \boldsymbol{\ell}_2, ..., \boldsymbol{\ell}_K)$, we would like to transform this information into a $N \times N$ ground truth map B. The idea to train the U-Net is that we hope the output map A can be as similar to the ground truth map B as possible. We first transform the map A into $\hat{A}$. The entry $\hat{a}_{jk}$ in $\hat{A}$ is equal to $\frac{e^{a_{jk}}}{\sum e^{a_{jk}}}$. This means that $\hat{A}$ is the probability matrix indicating pixels belonging to landmarks. To construct the ground truth map, we first form a map B_i with respect to the ith facial landmark $\boldsymbol{\ell}_i = (x_i, y_i)$ by setting the (i, j)-th entry b_{jk} in B_i as

$$b_{jk} = 0.5^{\max(|x_i - j|, |y_i - k|)}. \qquad (1)$$

The overall ground truth map B is then formed by adding all maps with respect to all facial landmarks: $B = \sum_{i=1}^{K} B_i$. Figure 4 shows two sample truth maps. Basically, values of the points for the index (x_i, y_i) are the largest, and the points farther away from (x_i, y_i) have smaller values. With the truth map B and the feature map A of the last convolutional layer, the *unet loss* is defined as

$$\mathcal{L}_u = -\sum B \log(\hat{A}). \qquad (2)$$

At the first-stage training, the adopted optimization algorithm is Adam, and the learning rate is set as 0.0001. Network parameters are initialized by a truncated normal distribution. We find the best parameters of the U-Net by minimizing the *unet loss* for 40 epochs. According to our preliminary experiments, we found the first-stage training is important to make performance reliable.

With the parameters determined at the first stage, in the second-stage training we include last two fully-connected layers. Output of the last fully-connected layer is a 136-dimensional vector $\hat{\boldsymbol{\ell}} = (\hat{x}_1, \hat{y}_1, \hat{x}_2, \hat{y}_2, ..., \hat{x}_{68}, \hat{y}_{68})$ indicating the predicted positions of the 68 facial landmarks. The ground truth positions are $\boldsymbol{\ell} = (\boldsymbol{\ell}_1 = (x_1, y_1), \boldsymbol{\ell}_2 = (x_2, y_2), ..., \boldsymbol{\ell}_{68} = (x_{68}, y_{68}))$. The *landmark loss* is defined as the mean square error between $\hat{\boldsymbol{\ell}}$ and $\boldsymbol{\ell}$. That is

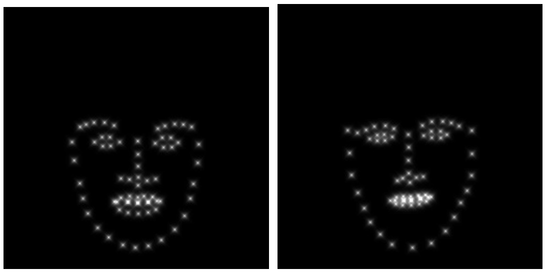

Fig. 4. Two sample ground truth maps representing facial landmark information.

$$\mathcal{L}_\ell = \frac{1}{K}\sum_{i=1}^{K}\Big((x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2\Big). \qquad (3)$$

With the second-stage training, we fine-tune the U-Net and find the best parameters of the last two fully-connected layers. In fact, the concepts of *unet loss* and *landmark loss* are similar, both based on the difference between the predicted ones and the ground truth but from different representations.

III. Multi-Task Learning

The network mentioned in Sec. II can be modified to do other works like emotion recognition, by changing the last two fully-connected layers. However, both facial landmark detection and emotion recognition are based on features extracted from face images. Intuitively, knowing a person being smiling, for example, may be helpful to more accurately detect facial landmarks. Therefore, we would jointly consider these two tasks and attempt to learn better features for obtaining performance better than single tasks [8].

To build the multi-task network, we change the last fully-connected layer of Figure 3 into two branches. The first branch is a fully-connected layer containing 136 nodes, with sigmoid as the activation function. This branch is for predicting landmark positions, and is basically the same as mentioned in Sec. II. The second branch is a fully-connected layer containing 8 nodes, with softmax as the activation function. This branch is for estimating the probability of the considered eight types of emotion.

Again we adopt the two-stage scheme to train the multi-task network. At the first-stage training, the best parameters of the U-Net are found by minimizing the *unet loss*. At the second-stage training, in addition to the *landmark loss* with respect to facial landmark detection, we further integrate *emotion loss* with respect to emotion recognition. For an input image I, the second branch outputs a 8-dimensional probability vector $\hat{\boldsymbol{e}} = (\hat{e}_1, ..., \hat{e}_8)$. The ground truth of I is encoded as a one-hot vector $\boldsymbol{e} = (e_1, ..., e_8)$. The *emotion loss* is defined as the mean square error $\mathcal{L}_e$ between $\boldsymbol{e}$ and $\hat{\boldsymbol{e}}$. Overall, the loss for the second-stage training is the combination of *landmark loss* and *emotion loss*: $\mathcal{L} = \mathcal{L}_\ell + \mathcal{L}_e$. By minimizing the integrated loss, we fine-tune the parameters of the U-Net and find the best parameters of the two connected branches.

The idea of multi-task learning mentioned above is to simultaneously optimize two tasks based on the same features and network structure. However, this joint optimization is not a trivial because different tasks have different learning difficulties. As the training proceeds, the task A may converge before the task B does. In this case, if we keep updating network parameters by jointly considering two tasks, the task A may eventually overfit. To solve this problem, we modify the early stopping approach proposed in [8] to "early stop" the task that has been converged.

In our work, facial landmark detection is the main task, and emotion recognition is the auxiliary task. We thus focus on early stopping the emotion recognition task. Let $E(i)$ denote the loss of the emotion recognition task at the ith iteration. We stop the emotion recognition task if the following criterion meets:

$$\frac{k \cdot \mathrm{med}_{i=t-k}^{t} E(i)}{\left(\sum_{i=t-k}^{t} E(i)\right) - k \cdot \mathrm{med}_{i=t-k}^{t} E(i)} > \epsilon, \tag{4}$$

where t denotes the current iteration, and k controls the number of iterations in the past to be considered. The "med" denotes the function calculating the median of loss values in the considered interval. If the loss drops rapidly within a period of length k, the value in Eqn. (4) would be small. Otherwise, if the training process tends to be converged, the value in Eqn. (4) would be large. The threshold ϵ is empirically set to appropriately stop the auxiliary task.

IV. EVALUATION

A. Evaluation Dataset

We evaluate the proposed approach based on the dataset provided in [10]. For each of the 2,935 thermal face images, positions of 68 landmarks are manually annotated. The resolution is 1024×768 pixels. Parts of images were captured in nine different head poses, and eight different emotions are included: *angry*, *contempt*, *disgust*, *fear*, *happy*, *sad*, *surprise*, and *neural*.

To focus on facial landmark detection and emotion recognition, we intentionally select frontal faces only from the dataset. Table I shows the numbers of faces with different emotions. Totally we pick 2,190 images of 64 individuals for evaluation. Because the data are quite limited, from each emotion category, we randomly select seven eighths of the images as the training data, and the remaining one eighth of images are used for testing. This training/testing scheme is applied five times, and average performance is calculated.

B. Evaluation Metric

Performance of facial landmark detection is measured by normalized mean error (NME) mentioned in [11]. NME is calculated by the distances between the predicted landmarks and the ground truths, normalized by the interpupil distance, i.e.,

$$NME = \frac{1}{N} \sum_{i=1}^{N} \frac{\|\boldsymbol{\ell} - \hat{\boldsymbol{\ell}}\|}{K \times D_i} \times 100\%, \tag{5}$$

where $\boldsymbol{\ell}$ and $\hat{\boldsymbol{\ell}}$ are ground truths and predicted coordinates, respectively. The term $\|\boldsymbol{\ell} - \hat{\boldsymbol{\ell}}\|$ is the L2 norm between $\boldsymbol{\ell}$ and $\hat{\boldsymbol{\ell}}$. The value D_i is the distance between two eyes, the value K is the number of facial landmarks, and N is the number of test images. Conceptually, the NME value denotes the prediction errors adaptively normalized by the interpupil distance of each individual.

C. Performance of Landmark Detection

We compare the proposed approach with several baselines:

- Detection on transformed faces: As mentioned in Sec. I, we construct a CycleGAN to transform thermal faces into visible faces. We then employ the Dlib library to detect facial landmarks on transformed visible faces.
- Basic CNN approach: We build a basic convolutional neural network consisting of six convolutional layers. The convolution kernel is 3×3 and the ReLU activation function is used for all layers. After the first, the third, the fourth, and the sixth convolutional layers, a 2×2 max pooling is applied. After the sixth convolutional layers, two fully-connected layers respectively consisting of 1024 nodes and 136 nodes are connected. To train this network, the mean square error between predicted coordinates and ground truth coordinates is calculated as the loss function. Other training settings are the same as the first-stage training mentioned in Sec. II.
- Active appearance model [3]: The work in [3] proposed a series of pre-processes and post-processes on thermal face images. The main idea is training an active appearance model based on visual features like SIFT and HOG. For a given face image, the model adaptively fits this face and estimates positions of facial landmarks.
- Our U-Net-based approach (single task): The proposed U-Net-based approach that considers facial landmark detection only.
- Our U-Net-based approach (multi task, with or without early stopping): The proposed U-Net-based approach that jointly considers facial landmark detection and emotion recognition.

Table II shows performance comparison of facial landmark detection. The first row shows the performance of the Dlib facial landmark detector on transformed visible faces. The performance is not bad, but a method dedicated for thermal faces like [3] can work better. The basic CNN approach works much better than the Dlib on transformed visible faces, but is still not satisfactory. Performance of [3] is promising (5.20), while our proposed U-Net-based approach (single task) works even better (4.57). If we jointly consider two tasks, more performance gain can be obtained (4.31). This shows the value of multi-task learning. If the early stopping approach is further applied, the best performance can be obtained (4.03).

TABLE I
THE NUMBERS OF FACE IMAGES WITH DIFFERENT EMOTIONS.

angry	contempt	disgust	fear	happy	sad	surprise	neural	Total
218	179	211	238	301	224	323	496	2190

TABLE II
PERFORMANCE OF LANDMARK DETECTION, IN TERMS OF NME.

	single task	multiple tasks (w/o early stopping)	multiple tasks (w. early stopping)
Dlib on transformed faces	9.54	–	–
CNN	5.37	–	–
AAM [3]	5.20	–	–
Ours	4.57	4.31	**4.03**

TABLE III
PERFORMANCE COMPARISON OF LANDMARK DETECTION, BASED ON THE SETTINGS WITH/WITHOUT TWO-STAGE TRAINING.

	single task (without two-stage training)	single task (with two-stage training)
NME	6.84	4.57
	multiple tasks (without two-stage training)	multiple tasks (with two-stage training)
NME	5.25	4.31

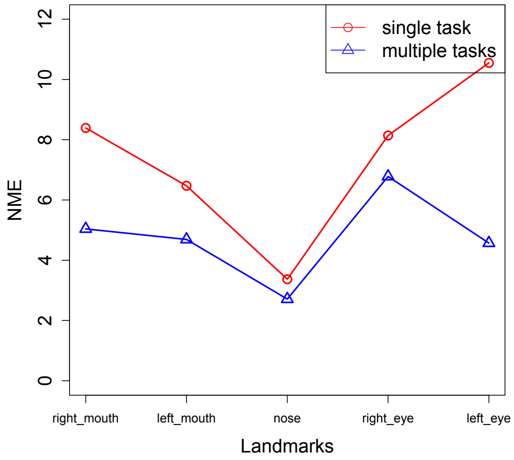

Fig. 5. Values of NME for five selected facial landmarks.

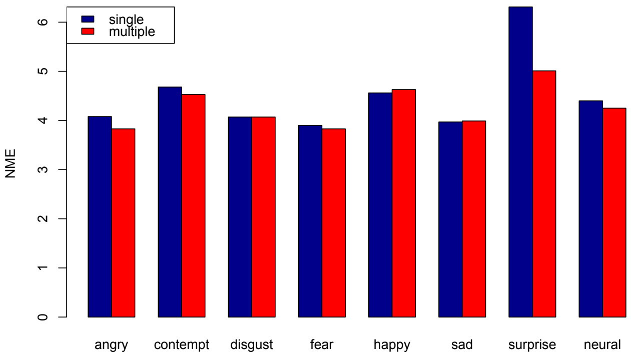

Fig. 6. NMEs of landmark detection for faces in eight different emotions.

We proposed the two-stage training scheme in Sec. II. Recall that the setting of "without two-stage training" means that the framework depicted in Figure 3 is trained in an end-to-end manner. The entire network is trained from the beginning to the end. The setting of "with two-stage training" means that the Unet part is first trained based on the *unet loss* for 40 epochs. With the obtained initial parameters, the entire network is then trained from the 41th epochs. Table III shows performance comparison between two schemes. As can be seen, the network trained by the two-stage training scheme works significantly better than that without it.

We intentionally select five facial landmarks and compare prediction errors obtained by the single-task approach and the multi-task approach. Figure 5 shows prediction errors for the five landmarks, which are at right mouth corner, left mouth corner, the center of nose, the center of right eye, and the center of left eye. This figure clearly shows that the multi-task approach works better than the single-task approach.

As we think facial landmark detection is related to emotion, we check if performance of landmark detection varies for faces with different emotions. Figure 6 shows variations of NMEs for faces in eight different emotions. Generally, performance of faces with different emotions is similar, which shows that the proposed method is robust. The single-task model relatively yields a larger NME for faces with surprise. This may be because faces with surprise deviate more from the neural faces. On the other hand, the multi-task model largely improves performance of faces with surprise.

Figure 7 shows sample thermal faces of three individuals. From left to right, they are actually with emotions angry, happy, and sad, respectively. As can be seen, the visual appearance of thermal faces is significantly different from visible faces, and thus dedicated techniques should be designed to conduct thermal face analysis.

Finally, we show performance of the auxiliary task, i.e., emotion recognition, in Table IV. When we solely work on the emotion recognition task, 73.80% recognition accuracy is obtained. Few previous works were proposed on thermal

TABLE IV
RECOGNITION ACCURACY OF EMOTION RECOGNITION BASED ON TWO SETTINGS.

	single task	multiple tasks (w/o early stopping)	multiple tasks (w. early stopping)
Accuracy	73.80	69.37	65.68

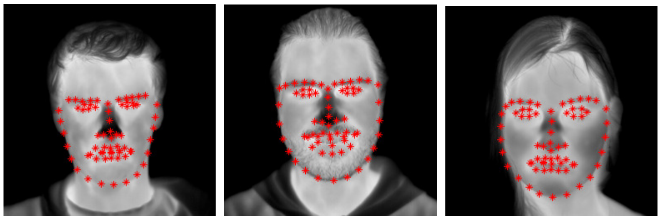

Fig. 7. Sample thermal faces of three individuals.

face emotion recognition. To our best knowledge, Kopaczka et al. [12] are still the only team working on this topic. In [12], they extracted dense SIFT features and constructed an SVM classifier to achieve emotion recognition. Also based on the dataset in [10], the average recognition accuracy for eight emotions mentioned above is 45.83%. This shows the effectiveness of the deep-based method.

Table IV also shows that, in the multi-task learning, either with or without early stopping, performance of emotion recognition degrades. In Sec. III, we view facial landmark detection as the main task and emotion recognition as the auxiliary task. All designs and learning processes are optimized for landmark detection. On the other hand, how to systematically analyze correlations between tasks and make multitask learning effective is still an ongoing research. The works in [13] and [14] pointed out that multitask learning is not guaranteed to always perform better than the single-task counterpart on each task. The performance trend shows this case and motivates us to investigate the related issues in the future.

V. CONCLUSION

We have presented a thermal facial landmark detection based on deep multi-task learning. In the proposed network, a U-Net structure is constructed to extract features from thermal faces. Two tasks, i.e., facial landmark detection and emotion recognition, are then jointly considered by two network branches. The entire network is trained in an end-to-end manner with the idea of multi-task learning. Experimental results show that the multi-task approach works better than the single-task approach and provides robust performance for faces with different emotions. Furthermore, we verify that the early stop mechanism brings performance gain. In the future, more tasks can be jointly considered together, and more evaluations can be conducted, such as performance variations on thermal faces of different spatial resolutions or different temperature resolutions.

ACKNOWLEDGMENT

This work was partially supported by the Ministry of Science and Technology under the grant 108-2221-E-006-227-MY3, 107-2221-E-006-239-MY2, 107-2923-E-194-003-MY3, 107-2627-H-155-001, and 107-2218-E-002-055.

REFERENCES

[1] M. Saquib Sarfraz and Rainer Stiefelhagen, “Deep perceptual mapping for thermal to visible face recognition,” in *Proceedings of British Machine Vision Conference*, 2015.

[2] Wei-Ta Chu and Jo-Ning Wu, “A parametric study of deep perceptual model on visible to thermal face recognition,” in *Proceedings of IEEE International Conference on Visual Communications and Image Processing*, 2018.

[3] Marcin Kopaczka, Kemal Acar, and Dorit Merhof, “Robust facial landmark detection and face tracking in thermal infrared images using active appearance models,” in *Proceedings of Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 2016, pp. 150–158.

[4] Marcin Kopaczka, Justus Schock, Jan Nestler, Kevin Kielholz, and Dorit Merhof, “A combined modular system for face detection, head pose estimation, face tracking and emotion recognition in thermal infrared images,” in *Proceedings of IEEE International Conference on Imaging Systems and Techniques*, 2018.

[5] Marcin Kopaczka, Thomas Jantos, and Dorit Merhof, “Towards analysis of mental stress using thermal infrared tomography,” in *Proceedings of Bildverarbeitung fur die Medizin*, 2018.

[6] Vahid Kazemi and Josephine Sullivan, “One millisecond face alignment with an ensemble of regression trees,” in *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2014.

[7] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in *Proceedings of IEEE International Conference on Computer Vision*, 2017.

[8] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang, “Facial landmark detection by deep multi-task learning,” in *Proceedings of European Conference on Computer Vision*, 2014.

[9] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015.

[10] Marcin Kopaczka, Raphael Kolk, and Dorit Merhof, “A fully annotated thermal face database and its application for thermal facial expression recognition,” in *Proceedings of IEEE International Instrumentation and Measurement Technology Conference*, 2018.

[11] Hanjiang Lai, Shengtao Xiao, Yan Pan, Zhen Cui, Jiashi Feng, Chunyan Xu, Jian Yin, and Shuicheng Yan, “Deep recurrent regression for facial landmark detection,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 5, pp. 1144–1157, 2016.

[12] Marcin Kopaczka, Raphael Kolk, Justus Schock, Felix Burkhard, and Dorit Merhof, “A thermal infrared face database with facial landmarks and emotion labels,” *IEEE Transactions on Instrumentation and Measurement*, 2018.

[13] Hector Martinez Alonso and Barbara Plank, “When is multitask learning effective? semantic sequence prediction under varying data conditions,” in *Proceedings of Conference of the European Chapter of the Association for Computational Linguistics*, 2017, pp. 44–53.

[14] Yu Zhang and Qiang Yang, “A survey on multi-task learning,” *CoRR*, 2017.