04 Mar 2019

Multi-modality Empowered Network For Facial Action Unit Detection

Peng Liu

Zheng Zhang

Huiyuan Yang
Missouri University of Science and Technology, hyang@mst.edu

Lijun Yin

Recommended Citation

P. Liu et al., "Multi-modality Empowered Network For Facial Action Unit Detection," *Proceedings - 2019 IEEE Winter Conference on Applications of Computer Vision, WACV 2019*, pp. 2175 - 2184, article no. 8659257, Institute of Electrical and Electronics Engineers, Mar 2019.
The definitive version is available at https://doi.org/10.1109/WACV.2019.00235

Multi-modality Empowered Network for Facial Action Unit Detection

Peng Liu
Aware, Inc
pliu@aware.com

Zheng Zhang Huiyuan Yang Lijun Yin
State University of New York at Binghamton
zzhang27@binghamton.edu hyang51@binghamton.edu lijun@cs.binghamton.edu

Abstract

This paper presents a new thermal empowered multi-task network (TEMT-Net) to improve facial action unit detection. Our primary goal is to leverage the situation that the training set has multi-modality data while the application scenario only has one modality. Thermal images are robust to illumination and face color. In the proposed multi-task framework, we utilize both modality data. Action unit detection and facial landmark detection are correlated tasks. To utilize the advantage and the correlation of different modalities and different tasks, we propose a novel thermal empowered multi-task deep neural network learning approach for action unit detection, facial landmark detection and thermal image reconstruction simultaneously. The thermal image generator and facial landmark detection provide regularization on the learned features with shared factors as the input color images. Extensive experiments are conducted on the BP4D and MMSE databases, with the comparison to the state of the art methods. The experiments show that the multi-modality framework improves the AU detection significantly.

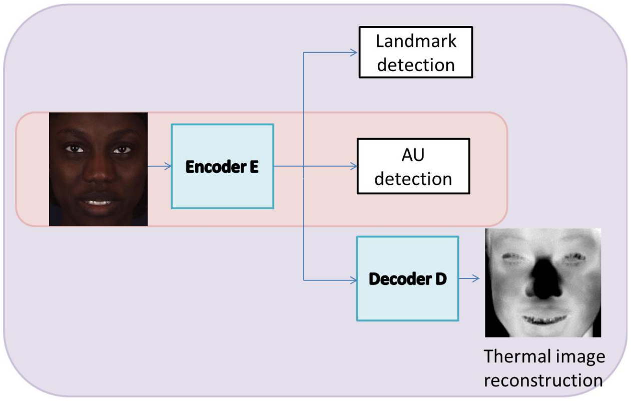

Figure 1: Overview of proposed framework with four components: an encoder, a decoder, two content-based detection (AU detection and landmark detection). Training phase: AU detection, landmark detection, and thermal image reconstruction. The AU detection task is our main target task. The landmark detection task is the task to improve the learning of AU detection. Testing phase (circled in red): AU detection only.

1. Introduction

Action units (AUs) are the local muscle movements on the face. Facial actions convey information about expression and emotions. Ekman and his collaborators designed the Facial Action Coding System (FACS) [6]. FACS relies on identifying visible local facial appearance variations called AUs. Detection and analysis of AUs is a challenging problem in computer vision. Their detection requires analyzing subtle appearance changes on the human face. Some existing work have utilized deep models to study AU detection and achieved good results [36, 19].

Many large scale datasets have multi-modality data. Most of the previous works focus on one modality. The majority of them target on 2D domain, and there are some methods targeting on other modalities such as 3D models and thermal images [9, 14]. Currently, there are few applications with multiple modalities. The multi-modality datasets are not utilized in their full capacity. One of the reason is that in most application cases, only one modality is available. Multi-modality analysis is an emerging field [4, 21, 31]. Deep learning has broken the boundaries between vision and language [7], and different modality images, such as color image and thermal image [32], color image and map image [39]. Transfer Learning aims to improve learning on a certain task with knowledge transferred from related tasks. It allows different data distributions, tasks and

978-1-7281-1975-5/19/$31.00 ©2019 IEEE
DOI 10.1109/WACV.2019.00235

representations between training and testing, and focuses on extracting knowledge that are shared between domains [3]. Transfer learning from different tasks could help reduce the number of training samples while keeping the performance nearly the same [31]. Inspired by previous work, we build a new model, which only needs multi-modality data during training to better represent the facial features. During testing, only visible images are required. Thus, our model can be easily used in real-world applications.

We propose to use a multi-task deep network to jointly learn the facial AUs based on the thermal image reconstruction and landmark detection. Multi-task learning is a machine learning approach, that learns multiple related problems simultaneously by using a common representation. In our multi-task learning deep model, the learnings of AU detection and thermal image reconstruction share a common color image input but involve different target random variables. To leverage the success of multi-task learning, we also integrate landmark detection task in our framework. The general framework of the proposed method is shown in Figure 1. The model is composed of four components: an encoder, a decoder, and two content detector. The encoder is used to extract the content feature representation of the image, which is good for face image understanding. The decoder is to reconstruct a thermal image. The AU detector is our main target task. The landmark detector is the task to improve the learning of AU. The thermal image reconstruction task is used for regularization the learning of the color image encoder parameters. We train different tasks simultaneously using multiple loss functions. This multi-task learning helps effectively exploit the complementary information from different tasks. The shared deep model features get better understanding of faces, which leads to improvements of the performances.

Comparing to existing AU detection methods, our work has the following unique contributions:

1. We propose a novel thermal empowered multi-task Convolution Neural Network (TEMT-Net), in which task relation is captured by a shared Network, and variability across different tasks is captured by task specific networks. This way of multi-task learning helps effectively exploit the complementary information from different tasks while maximally preserving the specific information of specific tasks.

2. We present a training paradigm, which allows learning from multiple modality data, and only one modality is required during testing. Thus, our model can be used in real-world applications.

3. The experiments show that the proposed framework boosts the performances of AU detection. We demonstrate the performance improvement over the existing work. The benefits of the multi-task framework are also discussed.

2. Related Work

Facial action units are elemental components of facial expressions. Compared with categorical expressions classification, AU detection problem is more challenging which requires fine-grained features extracted. Previously, effective features on AU recognition task are most likely to be the hand-crafted ones, *i.e.*, appearance LBP features [15] and the Discrete Cosine Transform (DCT) features [8], where intuitively human design and interference are needed. With the increasing of non-linearity and introduction of parameter sharing mechanism, deep features [11, 29] gradually take the place and show superior results.

Big model capacity comes along with a large number of parameters. Deep learning approaches require a large varieties of training samples to learn features and avoid overfitting. However, AU annotation is highly labor intensive and time consuming which makes supervised methods hard to generalize. To address this challenge, there are some semi-supervised and weakly supervised learning methods, which employ partial supervision by making use of additional unannotated data [37, 34].

Fusing multiple modalities or tasks has the advantage of increasing robustness and conveying complementary information. Transfer learning [31] and joint learning [40] has shown advantages over methods that tackle individual ones. The performance of transfer learning is a useful metric as task affinity [31]. To utilize multi-modality data in single modality application or cross modality application, there are some works proposed. Wang et al. [28] proposed a multi-modality training framework which including both color image and thermal image, while only requires color image for testing and application. Liu et al. [20] proposed a multi-task deep model, that the encoder learns to extract features good for face recognition but not useful for style factor recognition. Wu et al. [30] proposed a constrained joint cascade regression framework to learn landmark detection and AU detection at the same time. Riggan et al. [24] proposed using the polarization-state information of thermal emissions to enhance the performance of cross spectrum face recognition. It has been shown that polarimetric-thermal images capture geometric and textural details of faces, which are not present in the conventional thermal facial imagery.

Several deep convolution networks have been proposed by learning spatial representation with CNNs, temporal modeling with LSTMs, and frame based spatio-temporal fusion [5, 18]. Jaiswal et al. [13] presented a novel CNN-BLSTM based approach, which learns the dynamic appearance and shape of facial regions for AU detection.

To the best of our knowledge, no prior work has attempted to estimate to use multi-modality data to do AU detection. The proposed TEMT-Net which not only learns different modalities but also learns action unit and facial

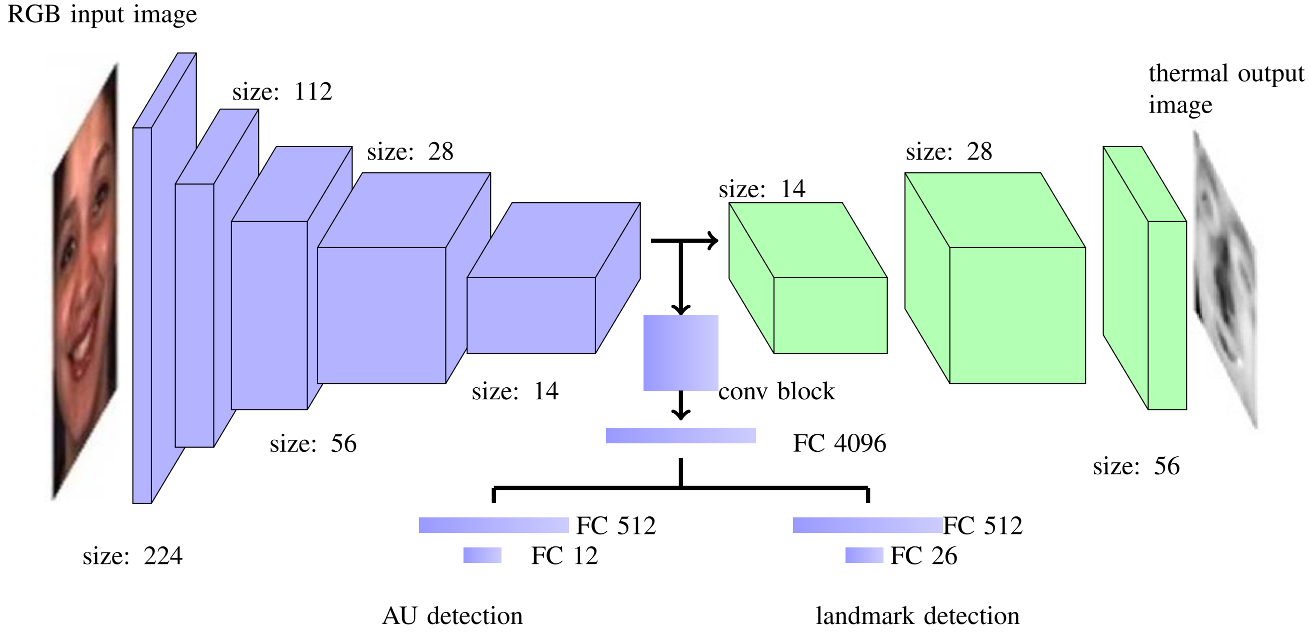

Figure 2: Framework for simultaneous training facial action unit detection, thermal image reconstruction, and landmark detection

landmark at the same time.

3. TEMT Network

Due to the availability of training data, most deep learning face models are designed and trained on color images. As the cost of 3D and thermal sensor decreases, there are more and more multi-modality databases such as BP4D database [33], MAHNOB laughter database [22], and USTC-NVIE database [27]. Those databases have not achieved their full potential. To better utilize the multi-modality databases, We propose a multi-task framework to detect AUs. The network architecture is shown in Figure 2. The model is composed of four modules: an encoder, a decoder, two content-based detector (AU detector and landmark detector). The first module is the encode. The second module is a decoder network to reconstruct corresponding thermal image to improve the shared 2D facial feature learning. We apply a deconvolution network as the decoder. The third module is the AU detector. The forth module is the network that additionally provides facial landmarks locations to improve the AU detection. Note that the target here is not to regenerate a thermal image or get better landmark detection, but to improve the learning of the shared model.

3.1. Architecture

Learning multiple correlated tasks simultaneously can improve the performance of individual tasks [23, 2]. We applied the multi-task framework to handle color image and thermal image in a unified way. The TEMT-Net is trained on pairs of corresponding color and thermal images that are from a publicly available MMSE database [35].

We have a color image **I** and its corresponding thermal image **T**, AU labels **Y** and landmark coordinates **P**. The regression function f is given by the following equation:

$$(\mathbf{Y}, \mathbf{T}, \mathbf{P}) = f(\mathbf{I}, \Theta) \tag{1}$$

where Θ is a set of trainable parameters of function f. The objective is to optimize the parameters in Θ to minimize: the error between the estimated AU $\hat{\mathbf{Y}}$ and the ground truth AU $\mathbf{Y}$; the error between the estimated corresponding thermal images $\hat{\mathbf{T}}$ and the ground truth thermal images $\mathbf{T}$; and the error between the estimated facial landmarks $\hat{\mathbf{P}}$ and the ground truth facial landmarks $\mathbf{P}$.

The size and the number of the feature maps are given in the figure. The input is an aligned RGB face image. The convolution encoder are generated by convolution blocks. The blocks can be any convolution blocks, eg, VGG 3 layers 3×3 convolution block [25], Inception block [26], and Resnet block [12]. Pooling operations are performed between convolution blocks. Batch normalization is applied after each convolution layer. ReLU is used as the activation function after batch normalization. The output of encoder is 14×14. There are two branches of output. One branch for thermal image reconstruction. One branch for AU and landmark detection.

For the AU and landmark detection branch, there is another 7×7 convolution block. At the end, one fully connected layer sized of 4096 is augmented. We add extra two fully connected layers sized of 512 for AU detector and landmark detector. Another two fully connected layers f_{au} and $f_{landmarks}$, for AU and landmarks detection, respectively.

The decoder of the thermal image is a deconvolution model. The input of decoder is 14×14 The decoder network are generated by 3 deconvolution blocks. The 3 deconvolution blocks are mirrored version of the encoder convolution network, and has multiple series of unpooling, deconvolution, and rectification layers. To reduce the size of the deconvolution model, we reduce the thermal target image to 56×56. The deconvolution network enlarges the activations through the combination of unpooling and deconvolution operations. The target is an aligned thermal face image.

3.2. Training

We use MMSE [35] dataset for training our network. It contains face images with full pose, expression, ethnicity, age, and gender variations. It provides annotations for 49 landmarks per RGB face, and 28 landmarks per thermal face. Different loss functions are used for training the tasks of AU detection, thermal image reconstruction, and landmark localization.

3.2.1 AU detection

AU detection can be treated as a multi-label classification problem. The f_{au} layer is extracted as a feature vector for each image. Let the number of AUs be C. The output layer was designed as a multi-label sigmoid cross-entropy loss:

$$\begin{aligned}\ell_{au}(y, \hat{y}) = -\sum_{n=1}^{C}(y_n \times \log \hat{y}_n \\ + (1 - y_n) \times \log(1 - \hat{y}_n)),\end{aligned} \tag{2}$$

where $\hat{y}_n$ is the estimated probability of the n_{th} AU. If the n_{th} AU is labeled, $y_n = 1$. Otherwise, $y_n = 0$. f_{au} is the two dimensional probability vector computed from the network. If $\hat{y}_n \geq 0.5$, the corresponding AU is estimated as detected. Otherwise, that AU is estimated as undetected.

3.2.2 Thermal image reconstruction

The thermal image reconstruction network is a deconvolution network. The model records the locations of maximum activations selected during pooling operation in switch variables, which are employed to place each activation back to its original pooled location. Similar to the convolution network, a hierarchical structure of deconvolution layers is used to capture different levels of shape details. The filters in lower layers tend to capture overall shape of an object while the task-specific details are encoded in the filters in higher layers. We construct the normalized pixel-wise grey thermal image by pixel-wise color image using equation (3), which is given by

$$\ell_{tr} = \frac{1}{W \times H} \sum_{w=1}^{W} \sum_{h=1}^{H} \| D(E(I_{wh})) - T_{wh} \|_2, \tag{3}$$

where I, T are the normalized input color image and corresponding thermal image, $W \times H$ is the dimension of the input image, D, E stand for the decoder and encoder network in Figure 1. We omit their parameters here for simplicity.

3.2.3 Landmarks detection

The face image is labeled by landmarks $\{c_x, c_y, w, h\}$, where c_x and c_y are the coordinates of the center of the region, and w and h are the width and height of the face region, respectively. Each landmark is shifted with respect to the region center (c_x, c_y), and normalized by w and h as given in (4).

$$(x_i^{norm}, y_i^{norm}) = \left(\frac{x_i - c_x}{w}, \frac{y_i - c_y}{h} \right), \tag{4}$$

(x_i, y_i) and (x_i^{norm}, y_i^{norm}) are the original and normalized coordinates of landmark i. Assuming that we have N landmarks in consideration, the loss is therefore computed by

$$\ell_{landmark} = \frac{1}{N} \sum_{i=1}^{N} \| \psi(E(I))_i - P_i \|_2, \tag{5}$$

where P are the ground truth coordinates after normalization. $\psi(\cdot)$ is the prediction network which maps encoded image to landmark coordinates.

3.2.4 Total loss function

As shown in Figure 2, the total loss is a combination of AU detection loss, facial landmark detection loss, and thermal image reconstruction loss. We further introduce an regularization term to penalize the complexity of weights.

$$\begin{aligned}\ell_{total} =& \lambda_{au}\ell_{au} + \lambda_{tr}\ell_{tr} \\ &+ \lambda_{landmark}\ell_{landmark} \\ &+ \lambda \|\Theta\|_2^2.\end{aligned} \tag{6}$$

where $\lambda_{au}, \lambda_{tr}, \lambda_{landmark}, \lambda$, are the factors of each loss components and L2 regularizer.

4. Experiment results

4.1. Database

We evaluate our proposed method on two public databases, BP4D database [33] and MMSE (a.k.a. BP4D+) database [35]. Our target task is to detect if the AUs are active or not, which is a multi-label binary classification problem. F1-frame is the harmonic mean of precision and recall, and widely used in AU detection [18, 36]. In our evaluation, we compute F1 scores for 12 AUs. They are AU 1: Inner Brow Raiser; AU 2: Outer Brow Raiser; AU 4: Brow Lowerer; AU 6: Cheek Raiser; AU 7: Lid Tightener; AU 10: Upper Lip Raiser; AU 12: Lip Corner Puller; AU 14: Dimpler Buccinator; AU 15: Lip Corner Depressor; AU 17: Chin Raiser; AU 23: Lip Tightener; and AU 24: Lip Pressor.

(1) **BP4D database**: This is a spontaneous 3D dynamic model database. The database contains 3D video sequences of 41 subjects with spontaneous head movement and spontaneous facial expressions, including happiness, disgust, pain, surprise, etc. For each subject, there are 8 tasks corresponding to 8 authentic emotions. Each task lasts around 1-2 minutes. There are around 140,000 images with AU labels.

(2) **MMSE (a.k.a., BP4D+) database**: This is a multimodal spontaneous emotion corpus (MMSE), including synchronized 3D, 2D, thermal, and physiological data sequences (e.g., heart rate, blood pressure, skin conductance (EDA), and respiration rate) from 140 subjects (58 males and 82 females) with ages ranging from 18 to 66 years old. Facial expression was annotated for both the occurrence and intensity of facial action units from 2D video by experts in the Facial Action Coding System (FACS).

The resolution of texture color image of the BP4D and MMSE is 1040×1392. The color image face region is resized to 224×224 for training and testing. The resolution of thermal image of MMSE is 726×480. The face region size is around 100×100 to 150×150. The thermal face region are resized to 56×56 in this work. The reasons we do not regenerate larger thermal image are: First, our target is AU detection. Thermal image regeneration is only to help the model learn the 2D features. Second, we want to control the size of the model and the time for training. Increasing the size of the regeneration thermal image would increase the model size, decrease the input batch size and extend the convergence time. To control the size of the model, here we apply the 56×56 gray level thermal image as our regeneration target.

BP4D and MMSE databases both have 49 2D landmarks on the face. Landmark detection is a very challenge problem. Our model is not design to detect landmark as main target. If we detect all of landmarks on the face, the $\ell_{landmark}$ will occupy the ℓ_{total} and make the loss hard to converge. In order to balance the benefit of landmark detection on AU detection and the loss of landmark detection on the multitask learning, we only choose some key landmarks in the training set. The landmarks can be seen on figure 3.

4.2. Implementation

The entire network is trained on a Nvidia GTX 1080 GPU using the Tensorflow framework [1]. We choose $\lambda_{au} = 1$ for the AU detection loss, $\lambda_{landmark} = 0.001$ for the perceptual loss, $\lambda = 0.001$ for L2 regularizer, and $\lambda_{tr} = 0.0001$ for the thermal image reconstruction loss. During training, we use ADAM [16] as the optimization algorithm with learning rate of 1×10^{-3} and batch size of 16 images. We randomly initialize our network by using Xavier initializer [10].

4.3. Ablation Study

In order to better demonstrate the improvements obtained by different modules in the proposed network, we perform an ablation study on MMSE database. We divide the MMSE database for five folder. Three folder are used for training, one folder is used for validation, and another one is used for testing. We compare six architectures to demonstrate the advantages of the proposed network. The AU detection ablation study involves the following experiments:

(1) VGG-RGB: VGG-16 model trained on RGB color images. The input color image dimension is $224 \times 224 \times 3$.

(2) VGG-thermal: VGG-16 model trained on gray color thermal images. The input thermal image dimension is 224×224.

(3) VGG-RGB-landmarks: VGG-16 model trained on AU detection and facial landmark detection at the same time on RGB color images. The input color image dimension is $224 \times 224 \times 3$.

(4) VGG and deconvolution network: VGG-16 convolution model on RGB color image and deconvolution network reconstruct the gray color thermal images. The input color image is $224 \times 224 \times 3$. The target thermal reconstruction image is 56×56.

(5) TEMT-Net with VGG block: All the hyperparameters of the training are the same as section 4.2. The input color image is $224 \times 224 \times 3$. The target thermal reconstruction image is 56×56.

Table 1 shows the results of different architectures. We can see that each task loss function has a performance improvement compared to the baseline VGG-RGB network, and VGG-thermal network in terms of average F1 score. We also compare VGG-thermal with VGG-RGB. The result shows that thermal image input performs worse than color image on average, but it improve the result of action unit detection when it works together with color image on average. It is also shown that training AU detection and

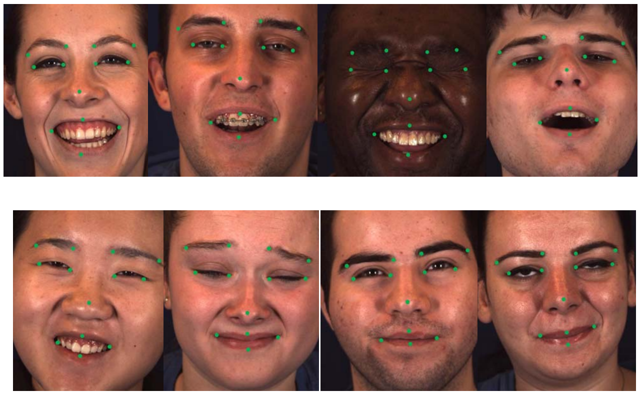

Figure 3: Examples of image frames with detected facial landmarks.

facial landmark detection tasks together can improve the result of action unit detection on average. In sum, the proposed multi-task framework can improve the AU detection.

4.3.1 Evaluation on landmark detection

Facial landmark detection is a challenge task. There are many models specifically targeting on landmark detection. Here we only evaluate the benefit of AU detection on the landmark detection. Because our model is not targeting on the landmark detection, but using landmarks to enhance the performance of AU detection, we only use 13 key landmarks in our model. Figure 3 shows those 13 landmarks. Those 13 landmarks are not very hard to allocated and can help the training of AU detection. We don't want to involve those landmarks are hard to detect and bring in a lot of errors for the AU detection. In the total loss function, the weight of the landmark detection task is relatively small. Figure 3 shows some sample images of our results. Table 2 shows the comparison between the model having only the landmark detection task and the model having landmark and AU detection tasks on MMSE database. The loss is calculated as the total root mean squared error (RMSE) in terms of pixel. Table 2 shows that AU detection task also has a positive effect on the performance of landmark detection.

4.3.2 Evaluation on thermal image reconstruction

Figure 4 illustrates several samples of the generated neutral face image on MMSE database. The left column is the input color image, the right column is the regenerated thermal image by the deconvolution network, and the middle is the corresponding ground-truth thermal face image. Comparing to color image thermal image are less sensitive to illumination and irrelevant to the face color. As shown in Figure. 4, a small weight in the total loss of the deconvolution network could generate good quality thermal images. The average pixel value difference in the test set is 8.2. By decoding the corresponding thermal image, the encoder network could learn the features irrelevant to face color and illuminations.

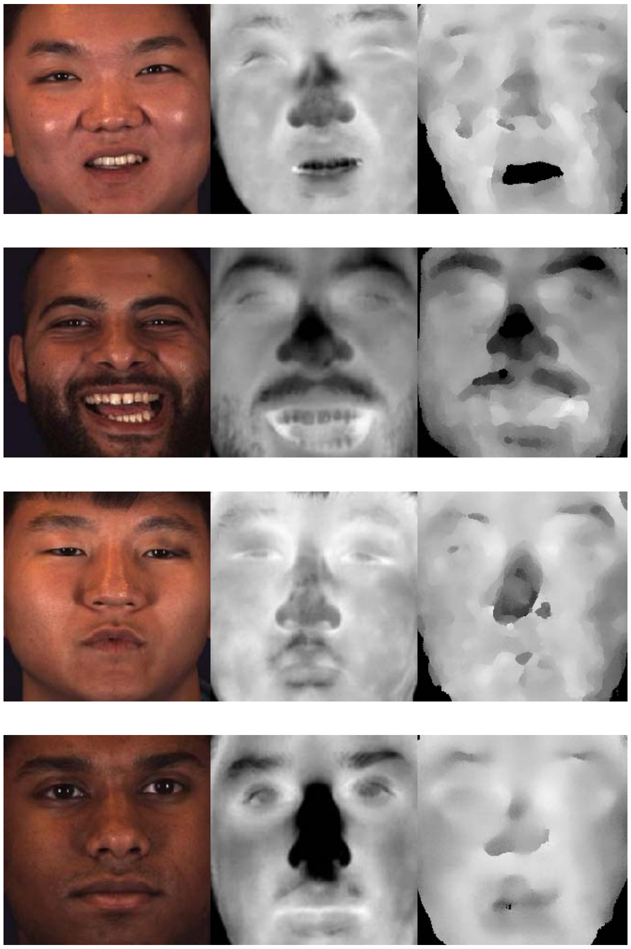

Figure 4: Examples of thermal image reconstruction. The left column is the input color image, the right column is the regenerated thermal image by the deconvolution network, and the middle is the corresponding ground-truth thermal face image.

4.4. Experimental comparison

In this section, we compare our model with two baseline methods and three state of the art methods. Baseline methods include AlexNet [17] and RGB and thermal AlexNet late fusion net. Color and gray level thermal AlexNet late fusion network is shown in figure 5. The color and thermal AlexNet outputs are fused in the fully connected layer. Two FC layers are size of 4096. The last FC layer is size of 12. We also compare two state of the art methods: DRML-Net [38], ResNet-34 [12] and EAC-Net [19] AU detection deep networks. DRML is an end to end unified architec-

Table 1: Ablation Study on MMSE considering different tasks.

Method	VGG-RGB	VGG-thermal	VGG-landmarks	VGG and deconv	TEMT-Net (VGG-block)
AU 1	21.6	23.2	30.6	20.5	41.9
AU 2	13.6	10.2	27.0	15.0	30.9
AU 4	15.7	16.7	10.6	9.2	15.4
AU 6	87.0	72.2	81.7	88.3	82.8
AU 7	87.4	85.7	90.9	88.6	91.0
AU 10	94.1	82.5	93.4	91.2	93.5
AU 12	80.9	76.3	85.9	84.9	86.1
AU 14	74.4	77.1	85.3	81.5	85.1
AU 15	14.0	32.1	32.7	28.1	29.2
AU 17	30.1	22.6	27.3	32.2	28.1
AU 23	45.6	33.5	49.8	49.4	47.0
AU 24	18.5	10.7	30.2	16.5	29.3
Avg.	48.6	45.3	53.8	50.1	55.0

Table 2: root mean squared error comparison of landmark detection

Model	RMSE
VGG landmark detection	2.462
VGG multi-task	2.163

ture for AU detection. In DRML, region learning (RL) and multi-label learning (ML) interact directly. EAC-Net is a combination of two nets: an enhancing net (E-Net) to force the neural network to pay more attention to AU interest regions on face images, and a cropping net (C-Net) to ensure that the network learns features in "aligned" facial areas.

As shown in Table 3, the second column is the RGB and thermal AlexNet fusion network. The thermal and color image data AlexNet fusion model gets better results than just using RGB image on AlexNet on average. Comparing to the state of the art methods, we get better results than DRML and Resnet-34 network on average, and we get comparable results with the EAC-Net. However, EAC-Net has a shortage that, during testing, EAC-Net needs landmark detection and must crop the AU related regions on the face. Many of those landmarks are not general facial landmarks, such as the landmarks on the forehead, cheek and chin region. Those landmarks are hard to localize. While our model does not require these operations.

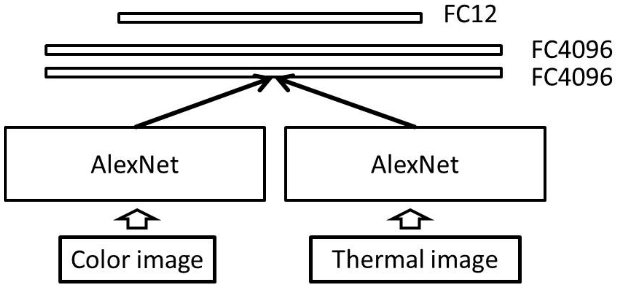

Figure 5: Framework of thermal and color image fusion network

We also conduct the same testing on the BP4D database. Since BP4D database only has 2D and 3D data. We use the MMSE color and thermal image training set and test the model in BP4D dataset color image. We randomly select half subjects of BP4D database for testing. The results reported can approximate the performance on the entire dataset. The other methods to be compared use the 2D color image in MMSE database for training and the models are tested on the same BP4D database test set. Table 4 shows the comparison results in BP4D dataset. We obtain experimental results similar to MMSE dataset on BP4D dataset. Table 4 shows that We get better results than DRML on average, and comparable results with the EAC-Net on average.

The effectiveness of our method relies on two main characteristics: First, we trained facial landmark detection and facial action unit detection together. The learning correlated tasks improved the each others' learning. Second, we trained color and thermal image together. The learning correlated modality improved the training of color features in the deep model.

5. Conclusion

This paper targets on the situation that the training set has multi-modality data but the application situation only has one modality data available. We have proposed a multi-

Table 3: Comparison experiment on MMSE dataset.

Method	AlexNet [17]	RGB thermal Fusion	DRML [38]	Resnet-34 [12]	EAC-Net [19]	**TEMT-Net (VGG-block)**	**TEMT-Net (resnet-block)**
AU 1	13.8	22.0	22.6	16.8	35.9	32.4	36.4
AU 2	13.8	19.5	18.9	11.9	30.9	29.4	30.1
AU 4	10.5	12.6	9.3	10.9	32.4	11.9	26.6
AU 6	85.4	79.7	87.5	83.3	82.8	82.9	81.8
AU 7	92.6	89.0	88.5	90.1	91.0	89.4	89.7
AU 10	94.4	92.7	92.5	94.0	93.5	93.5	91.0
AU 12	90.6	86.3	81.8	83.7	86.1	85.9	82.0
AU 14	86.1	83.1	80.2	84.4	85.1	86.3	83.8
AU 15	14.9	29.3	17.8	21.0	35.2	37.8	29.1
AU 17	31.4	26.5	36.0	30.4	34.1	35.0	37.6
AU 23	31.4	42.8	41.6	50.1	47.0	50.6	57.9
AU 24	19.3	23.5	17.7	27.2	35.3	33.2	27.3
Avg.	48.2	50.5	49.9	50.3	58.0	55.0	56.1

Table 4: Comparison experiment on BP4D dataset.

Method	AlexNet [17]	RGB thermal Fusion	DRML [38]	EAC-Net [19]	**TEMT-Net (VGG block)**
AU 1	34.2	28.1	35.4	39.0	41.8
AU 2	22.4	21.9	26.7	35.2	30.4
AU 4	27.5	30.6	26.9	43.6	32.4
AU 6	59.6	61.1	64.9	76.1	71.8
AU 7	60.7	61.5	68.1	72.9	71.6
AU 10	64.1	71.7	75.3	81.9	80.7
AU 12	70.9	70.0	76.9	86.2	83.3
AU 14	57.2	61.1	57.2	58.8	61.6
AU 15	21.9	25.0	30.1	37.5	32.9
AU 17	33.8	38.4	38.3	47.1	41.6
AU 23	33.3	37.8	36.3	35.9	37.9
AU 24	9.7	11.7	11.8	14.8	12.2
Avg.	40.1	43.2	45.7	52.4	49.8

task framework for AU detection, which jointly learns facial landmark detection and thermal image reconstruction to enhance the performance of AU detection. This multimodality training paradigm can be used in other real world applications. The proposed approach is evaluated on two datasets: BP4D and MMSE. The experiments show that the multi-modality framework could improve the AU detection significantly. Our future work will be focused on facial analysis by adding extra modality (e.g., 3D geometric data, depth maps, and physiological data).

6. Acknowledgements

The material is based upon the work supported by the National Science Foundation under grants CNS-1629898 and CNS-1205664.

References

[1] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al. Tensorflow: A system for large-scale machine learning. In *OSDI*, volume 16, pages 265–283, 2016.

[2] J. Cao, Y. Li, and Z. Zhang. Partially shared multi-task convolutional neural network with local constraint for face attribute learning. June 2018.

[3] H. Chen, S. Cui, and S. Li. Application of transfer learning approaches in multimodal wearable human activity recognition. *arXiv preprint arXiv:1707.02412*, 2017.

[4] M. Chen, S. Wang, P. P. Liang, T. Baltrušaitis, A. Zadeh, and L.-P. Morency. Multimodal sentiment analysis with word-level fusion and reinforcement learning. In *Proceedings of*

the 19th ACM International Conference on Multimodal Interaction, pages 163–171. ACM, 2017.

[5] W.-S. Chu, F. De la Torre, and J. F. Cohn. Learning spatial and temporal cues for multi-label facial action unit detection. In *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*, pages 25–32. IEEE, 2017.

[6] P. Ekman and E. L. Rosenberg. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA, 1997.

[7] H. Fan and J. Zhou. Stacked latent attention for multimodal reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1072–1080, 2018.

[8] T. Gehrig and H. K. Ekenel. A common framework for real-time emotion recognition and facial action unit detection. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2011 IEEE Computer Society Conference on*, pages 1–6. IEEE, 2011.

[9] J. M. Girard, J. F. Cohn, L. A. Jeni, S. Lucey, and F. De la Torre. How much training data for facial action unit detection? In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, volume 1, pages 1–8. IEEE, 2015.

[10] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256, 2010.

[11] S. Han, Z. Meng, A.-S. Khan, and Y. Tong. Incremental boosting convolutional neural network for facial action unit recognition. In *Advances in Neural Information Processing Systems*, pages 109–117, 2016.

[12] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[13] S. Jaiswal and M. Valstar. Deep learning the dynamic appearance and shape of facial action units. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, pages 1–8. IEEE, 2016.

[14] S. Jarlier, D. Grandjean, S. Delplanque, K. N'diaye, I. Cayeux, M. I. Velazco, D. Sander, P. Vuilleumier, and K. R. Scherer. Thermal analysis of facial muscles contractions. *IEEE transactions on affective computing*, 2(1):2–9, 2011.

[15] B. Jiang, M. F. Valstar, and M. Pantic. Action unit detection using sparse appearance descriptors in space-time video volumes. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 314–321. IEEE, 2011.

[16] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[18] W. Li, F. Abtahi, and Z. Zhu. Action unit detection with region adaptation, multi-labeling learning and optimal temporal fusing. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6766–6775. IEEE, 2017.

[19] W. Li, F. Abtahi, Z. Zhu, and L. Yin. Eac-net: Deep nets with enhancing and cropping for facial action unit detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[20] Y. Liu, Z. Wang, H. Jin, and I. Wassell. Multi-task adversarial network for disentangled feature learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3743–3751, 2018.

[21] K. Nakamura, S. Yeung, A. Alahi, and L. Fei-Fei. Jointly learning energy expenditures and activities using egocentric multimodal signals. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, number EPFL-CONF-230255, 2017.

[22] S. Petridis, B. Martinez, and M. Pantic. The mahnob laughter database. *Image and Vision Computing*, 31(2):186–202, 2013.

[23] R. Ranjan, V. M. Patel, and R. Chellappa. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[24] N. Short, S. Hu, P. Gurram, K. Gurton, and A. Chan. Improving cross-modal face recognition using polarimetric imaging. *Optics letters*, 40(6):882–885, 2015.

[25] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[26] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[27] S. Wang, Z. Liu, S. Lv, Y. Lv, G. Wu, P. Peng, F. Chen, and X. Wang. A natural visible and infrared facial expression database for expression recognition and emotion inference. *IEEE Transactions on Multimedia*, 12(7):682–691, 2010.

[28] S. Wang, B. Pan, H. Chen, and Q. Ji. Thermal augmented expression recognition. *IEEE Transactions on Cybernetics*, 2018.

[29] S. Wu, S. Wang, B. Pan, and Q. Ji. Deep facial action unit recognition from partially labeled data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3951–3959, 2017.

[30] Y. Wu and Q. Ji. Constrained joint cascade regression framework for simultaneous facial action unit recognition and facial landmark detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3400–3408, 2016.

[31] A. R. Zamir, A. Sax, W. Shen, L. Guibas, J. Malik, and S. Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3712–3722, 2018.

[32] H. Zhang, V. M. Patel, B. S. Riggan, and S. Hu. Generative adversarial network-based synthesis of visible faces from polarimetric thermal faces. In *IJCB*, 2017.

[33] 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 and Vision Computing*, 32(10):692–706, 2014.

[34] Y. Zhang, W. Dong, B.-G. Hu, and Q. Ji. Weakly-supervised deep convolutional neural network learning for facial action unit intensity estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[35] Z. Zhang, J. Girard, Y. Wu, X. Zhang, P. Liu, U. Ciftci, S. Canavan, M. Reale, A. Horowitz, H. Yang, Q. Ji, J. Cohn, and L. Yin. Multimodal spontaneous emotion corpus for human behavior analysis. In *CVPR*, 2016.

[36] K. Zhao, W.-S. Chu, F. De la Torre, J. F. Cohn, and H. Zhang. Joint patch and multi-label learning for facial action unit detection. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 2207–2216. IEEE, 2015.

[37] K. Zhao, W.-S. Chu, and A. M. Martinez. Learning facial action units from web images with scalable weakly supervised clustering. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[38] K. Zhao, W.-S. Chu, and H. Zhang. Deep region and multi-label learning for facial action unit detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3391–3399, 2016.

[39] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networkss. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.

[40] Y. Zou, Z. Luo, and J.-B. Huang. Df-net: Unsupervised joint learning of depth and flow using cross-task consistency. In *European Conference on Computer Vision*, pages 38–55. Springer, 2018.