FATAUVA-Net : An Integrated Deep Learning Framework for Facial Attribute Recognition, Action Unit Detection, and Valence-Arousal Estimation

Wei-Yi Chang *[1], Shih-Huan Hsu[1], and Jen-Hsien Chien[1]

*Department of Computer Science and Information Engineering, National Cheng Kung University
[1]Emotibot Technologies Limited
{*weiyichang, cyrilhsu, kennychien*}*@emotibot.com*

Abstract

Facial expression recognition has been investigated for many years, and there are two popular models: Action Units (AUs) and the Valence-Arousal space (V-A space) that have been widely used. However, most of the databases for estimating V-A intensity are captured in laboratory settings, and the benchmarks "in-the-wild" do not exist. Thus, the First Affect-In-The-Wild Challenge released a database for V-A estimation while the videos were captured in wild condition. In this paper, we propose an integrated deep learning framework for facial attribute recognition, AU detection, and V-A estimation. The key idea is to apply AUs to estimate the V-A intensity since both AUs and V-A space could be utilized to recognize some emotion categories. Besides, the AU detector is trained based on the convolutional neural network (CNN) for facial attribute recognition. In experiments, we will show the results of the above three tasks to verify the performances of our proposed network framework.

1. Introduction

In addition to language, facial expression plays an important role for human communication in our daily life. Facial expressions could be described by the movements of numerous muscles that are located around mouth, nose, and eyes [45]. By recognizing facial expression or emotion, there are a variety of multimedia applications [8] such as human computer interaction, advertisement [25], pain recognition [35], and e-learning systems [4] that could help people to realize the response or feeling from other people. In previous studies, the facial expression typically could be divided into six popular categories (also known as universal facial expressions): anger, disgust, fear, happiness, sadness, and surprise. Moreover, Action Units (AUs) [10] were proposed to model facial behavior, and the combination of AUs also could be utilized for facial expression recognition (as shown in Table 1). In addition to AUs, the valence and arousal space (i.e., V-A space) [30] also have been widely used for facial expression recognition. The emotion categories could be recognized according to the position in V-A space (as shown in Figure 1).

Aiming at recognize facial expression, previous works have extracted different appearance features or geometry features directly from facial image [8]. Recently, deep convolutional neural networks (CNN) have shown out-

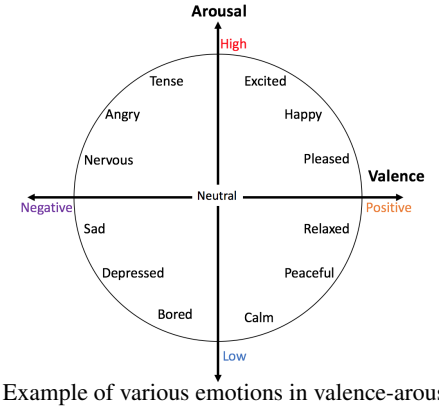


Figure 1. Example of various emotions in valence-arousal space.

Table 1. The relationship between emotion category and AUs [2].

Category	AUs
Happiness	AU12, AU25
Sadness	AU4, AU25
Anger	AU4, AU7, AU24
Disgust	AU9, AU10, AU17
Fear	AU1, AU4, AU20, AU25
Surprise	AU1, AU2, AU25, AU26

*Now is at Emotibot, and part of this work was done at National Cheng Kung University

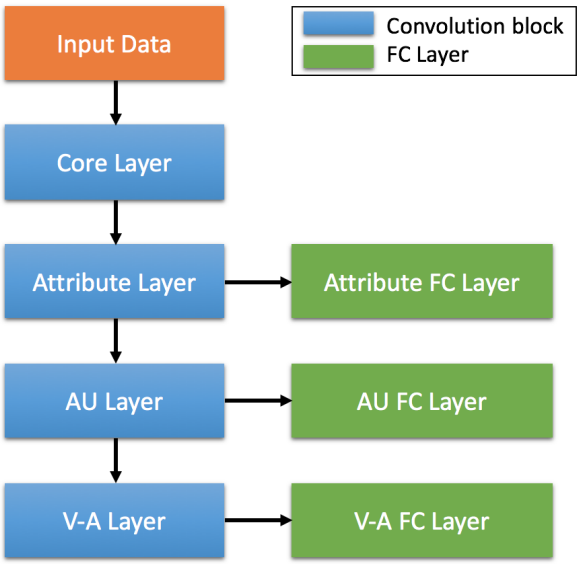


Figure 2. The basic network structure of *FATAUVA-Net*, and there are three kinds of convolutional layers (i.e., attribute layer, AU layer, and V-A layer) to solve the corresponding tasks.

standing performance in object recognition [31, 15, 42]. Moreover, CNNs also have been applied to face-related tasks [22, 29, 41] and facial expression recognition [17, 9]. For example, Jung *et al*. [17] utilized two deep models to extract temporal appearance feature and temporal geometry feature. By using a joint fine-tuning method to integrate these two networks, the improved performance could be achieved. Ding *et al*. [9] proposed FaceNet2ExpNet which pre-trained the deep network through face recognition data, and then fine-tuned the network with facial expression data.

Moreover, in order to describe the pattern of facial muscle movements, Action Units (AUs) were defined by Facial Action Coding System [10] (e.g., AU12 represents *lip corner puller*), and AUs have been utilized for facial expression recognition [2, 44, 21] and pain intensity estimation [24, 35]. By recognizing the activation of these AUs, basic and compound emotion categories could also be recognized according to the combination of AU [2]. For example, happy can be recognized if both AU12 and AU25 are activated. On the other hand, in order to annotate a large number of image for facial expression, Benitez-Quiroz *et al*. [2] proposed a real-time algorithm for action unit detection. Moreover, they also release a database with a million face images in the wild with the annotations of AU occurrences and emotion categories.

In addition to apply AUs for facial expression recognition, recently, psychologists and researchers in computer vision often focus on the analysis of valence and arousal space (i.e., V-A space) [30] for emotion recognition [28, 12, 19]. Valence represents the positive or negative experience, while arousal represents exciting or calm experience. By estimating these two values, the emotion could be recognized according to the position in V-A space (as shown in Figure 1). Although AUs and V-A space are two popular models in facial expression recognition, there is an interesting point that has less discussion: can we utilize AUs for V-A estimation? Since both AUs and V-A space could be employed to recognize the basic six emotions, this implies that the combination of AUs could be mapped to a position in V-A space (through the same emotion category). Thus, by extending this concept, we would like to know: is it possible to estimate the intensity of valence and arousal through AUs?

In this paper, we propose an integrated deep learning framework named *FATAUVA-Net* to achieve the goal of *F*acial *AT*tribute recognition, *A*ction *U*nit detection, and *V*alence-*A*rousal estimation. The basic network structure is illustrated in Figure 2. The main idea in this paper is utilizing AUs to estimate the intensity of valance and arousal while the relationship between AUs and V-A space has been investigated by [27]. Thus, in our framework, the V-A layers are placed on the top of AU layers so that the properties for AUs could be directly applied to V-A estimation. On the other hand, in order to train AU layers, we adopt the concept of enhance layer in [21] to focus on the regions that are corresponding to AU (e.g., lip corner for AU12). Here, since the response maps in face attribute recognition CNN can reveal the activated region of facial parts [37], we train a part-based CNN (similar to [37]) for facial attribute recognition to learn the core layer and attribute layer in *FATAUVA-Net*. After training the core layer, AU layer, and V-A layer sequentially, fully-connected (FC) layers are added to estimate the value in V-A space. Moreover, we also examine the performance of two kinds of loss layers in training-phase: class-based and regression-based. The contributions of this paper are summarized as:

- We present a deep learning framework for V-A estimation by employing AUs as mid-level representation. In experiments, we will show the advantage of using AUs for V-A estimation, and compare the method without using AUs.
- We propose a novel AU detection method that learns the response maps based on the CNN for facial attribute recognition. Thus, the activated facial parts would be emphasized for AU detection.
- We train a multi-task CNN with respect to facial part for facial attribute recognition.
- In experiment, we will show the performance of our single network for facial attribute recognition, action unit detection, and valence-arousal estimation.

2. Related Work

2.1. Multi-task CNN

With the concept of sharing information, multi-task learning could solve many related problems at the same time. For example, Chen *et al*. [7] integrated face detection and alignment into the same model by learning random forest, and Zhang *et al*. [41] proposed a deep cascaded multi-task framework to achieve the same goal. As to face analysis with CNN models, since the lower layers of CNN would learn common features [40], Ranjan *et al*. [29] proposed an all-in-one learning framework which shared the parameters in lower layers and learned task-specific layers to make predictions. Thus, this CNN architecture could simultaneously perform face detection, landmark localization, pose estimation, attribute recognition, and face identification.

2.2. Facial attribute recognition

Facial attributes are mid-level representations, and could be utilized to give more descriptions about face. With these semantic features, facial attribute could be applied to face recognition and verification [20, 3]. In order to extract these features, there have been some works proposed to recognize facial attribute (e.g., age, gender, and race). For example, Liu *et al*. [22] proposed a cascaded deep learning framework for attribute prediction. Torfason *et al*. [32] investigated classic hand-crafted feature and deep feature for attribute recognition, and showed that using combination of these features could achieve better result. Chen *et al*. [6] proposed an approach for cross-age face recognition, and also released a large-scale dataset called cross-age celebrity dataset (CACD).

By incorporating the property of shared feature into CNN, the multi-task CNN has been applied to face attribute recognition. Hand and Chellappa [14] proposed a multi-task deep CNN (MCNN) for face attribute recognition. The first two layers (Conv1 and Conv2) in their network were shared for all attribute. After Conv2, they divided 40 facial attributes into several groups according to the location on face, and generated Conv3s based on these groups. Then, the Conv3s were followed by fully connected layers. Different to MCNN [14], Lu *et al*. [23] proposed a dynamic branching procedure (to make task grouping decisions) in multi-task deep CNN by taking into account both task relatedness and complexity of the model.

2.3. Facial action unit detection

With the development of deep learning in computer vision, there have been some research works that applied deep learning to AU detection. In [18], Khorrami *et al*. showed that the features learning by CNNs for the emotion recognition task were strongly related to the AU. Based on this observation, Zhao *et al*. [44] proposed DRML which integrated region learning and multi-label learning into the same CNN to detect AU, while the proposed *region layer* was enable to capture the appearance changes in different subregions. Han *et al*. [13] utilized incremental boosting layer to integrate boosting into CNN for AU detection. Li *et al*. [21] included the enhancing layer and cropping layer into CNN structure so that the region of interest (e.g., eyes, eyebrows, and mouth) could be extracted for AU detection. However, they need extra information about facial landmarks to generate enhance layer. In this paper, we employ the CNN learned from the task of facial attribute recognition to guide the learning of AU detection, and it does not require the positions of facial landmarks.

2.4. Valence-arousal estimation

Within the past decades, researchers have investigated and developed various methods for facial affect analysis. In addition to AU detection, there have been some approaches focusing on estimating valence-arousal values from images or videos [33]. Baveye *et al*. [1] found that the fine-tuned CNN framework was a promising solution for predicting dimensional affective scores. Brady *et al*. [5] proposed a multi-modal solution by fusing the estimation from audio, video and physiological sensor. However, the data used in most prior work on valence-arousal estimation was in laboratory settings. Recently, in oder to show the generality of methods that were trained in controlled environments, some in-the-wild databases [39, 19] have been established. In [19], they have compared the performances of various features and different machine learning approaches on this new database.

Although V-A space has been widely adopted for emotion recognition, to the best of our knowledge, there are only few researches that applied AU for V-A estimation, for example, *Affectiva* [1] calculates valence likelihood based on AUs. Moreover, in [27], they have discovered that some AUs were discriminative to V-A values, and it would be more reasonable for V-A estimation since the V-A value could be directly interpreted by AU. For example, AU12 (*lip corner puller*) is related to positive valence. Thus, in this paper, we try to employ AUs for V-A estimation based on the investigation in [27].

3. Our Proposed Method

As illustrated in Figure 2, the proposed *FATAUVA-Net* is composed of convolutional layers (or blocks) from three tasks: attribute layer, AU layer and V-A layer, and the core layer is the foundation for the above three tasks. Since these tasks are related to face, the lower layers of CNN would learn a general set of feature for face analysis [29]. Based on the core layer, we add layers on the top of it to solve

[1] http://developer.affectiva.com/emotion_mapping/

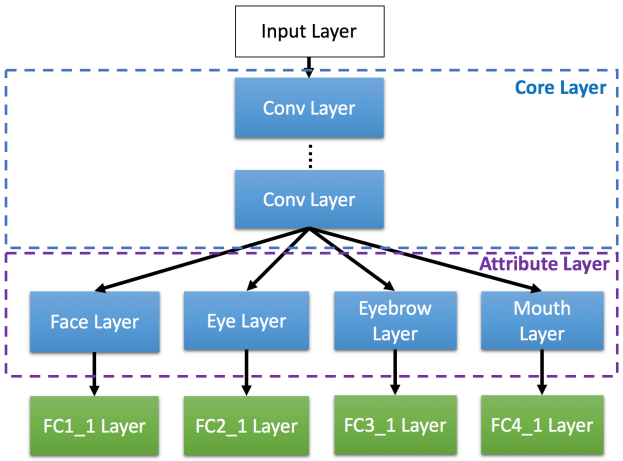


Figure 3. The network structure for facial attribute recognition.

Table 2. The relationship between attribute layer and attribute FC layer, while each attribute in FC layers is attached to the corresponding attribute layer.

Attribute Layer	Attribute FC Layer
Face	Attractive, Male, No Beard, and Young
Eyebrows	Arched Eyebrows, and Bushy Eyebrows
Eyes	Eyeglasses, Narrow Eyes
Mouth	Mouth Slightly Open, and Smile

different tasks, and the training step is in the order of: attribute recognition (learning core layer and attribute layer), AU detection (learning AU layer) and finally, V-A estimation (leaning V-A layer). In the following subsections, we will describe the detailed training procedure for each task.

3.1. Facial attribute recognition

Given the face region by face detection method (e.g., MTCNN [41]), we first train a CNN for attribute recognition to localize the region of facial parts (e.g., eyes, eyebrows, and mouth). Here, we select 10 attributes from CelebA dataset [22], and train our core layer and attribute layer with CelebA dataset while most of these attributes are related to eyes or eyebrows or mouth (these parts are also related to AU). Figure 3 shows the network structure for facial attribute recognition, and the core layer and the attribute layer are composed of several convolutional blocks. Then, we add FC layers on the top of attribute layers to recognize the selected 10 facial attributes, and the relationships between attribute layers and FC layers are listed in Table 2. In attribute layer, we utilize four different layers to capture the properties of each facial part, and the learned attribute layers further could be employed to lead the training process of AU detection.

3.2. Facial action unit detection

According to the investigation in [27], the relationship between AUs and V-A intensity can be modeled. Thus, in this paper, we propose an approach that utilize AUs as mid-level representation for V-A estimation, and the intensity of V-A would be estimated more reasonable by the interpretation from AUs. In order to detect AUs, we add AU layer on the top of attribute layer as illustrated in Figure 4, while the connection between attribute layer and AU layer is according to the corresponding facial part. For example, AU1 (*inner brow raiser*), AU2 (*outer brow raiser*), and AU4 (*brow lowerer*) are attached to eyebrow layer. On the other hand, since the corresponding areas among AU1, AU2, and AU4 are slightly different, we adopt three branches to model this difference. Similarly, mouth layer is divided into three more precise branches, and eye layer is divided into two branches. Finally, we attach FC layers to each AU layer for AU detection, while Table 3 shows the relationship between AU layer and FC layer. For example, we add FC_AU6 and FC_AU7 on the top of AU6_7. In training phase, in order to transfer the features from attribute layer, we freeze core layer and attribute layer in Figure 2, and only learn AU layers and the corresponding FC layers for AU detection.

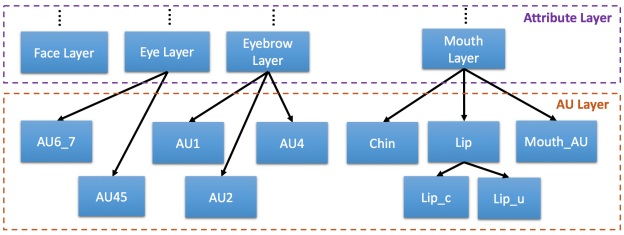


Figure 4. The network structure for AU detection.

Table 3. The relationship between AU layer and AU FC layer, while each AU in FC layers is attached to the corresponding AU layer. Here, the total number of selected AU is 14.

AU Layer	AU FC Layer
AU6_7	AU6, and AU7
AU45	AU45
AU1	AU1
AU2	AU2
AU4	AU4
Chin	AU17, and AU26
Lip_c	AU12, and AU15
Lip_u	AU10
Mouth_au	AU23, AU24, and AU25

3.3. Valence-arousal estimation

After learning features from AU layers, the next step is to train the network for V-A estimation. Based on the analysis in [27], we select different discriminative AUs to estimate the intensity of valence and arousal. Here, we concatenate the layer of AU6_7, AU45, AU4, Lip_c (focus on AU12), and Mouth_AU (focus on AU25) to generate Con_Val layer which further would be utilized to estimate the intensity of valance, while the conv layer between Con_Val layer and

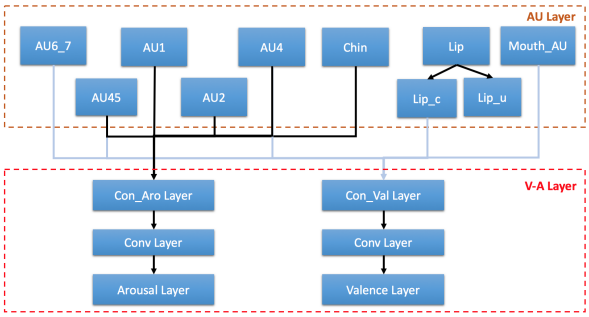


Figure 5. The network structure for V-A estimation.

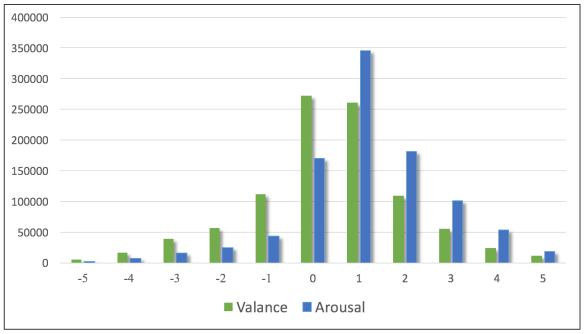


Figure 6. The histogram of V-A dataset with quantized values.

Valence layer is for the purpose of dimension reduction. Similarly, we concatenate the layer of AU45, AU1 AU2, AU4, and Chin (focus on AU26) to generate Con_Aro layer which further would be utilized to estimate the intensity of arousal, and a conv layer is inserted between Con_Aro and arousal layer for dimension reduction. Finally, we append FC layers on the top of valence layer and arousal layer, respectively. Figure 5 shows the detailed network structure of V-A layer. In training phase, we adopt the similar procedure used in training AU layer. That is, we freeze core layer, attribute layer, and AU layer in Figure 2, and only learn V-A layers and the corresponding FC layers.

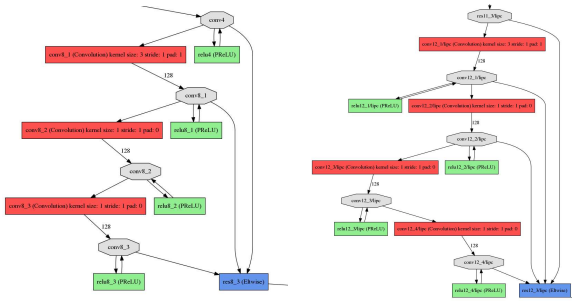
4. Experimental Results

In this section, we will show the performances of the proposed framework in the tasks of facial attribute recognition, AU detection, and V-A estimation. Although the main goal of this paper is to estimate the intensity of V-A, the main idea comes from AU detection. Thus, we also show the results on both AU detection and attribute recognition, and compare our proposed method with some recent works to demonstrate the effectiveness of our method.

4.1. Implement details

4.1.1 Data preprocessing

For attribute and AU dataset, we perform MTCNN [41] face detector to crop the face area, and we use the given bounding box for V-A dataset. In AU dataset, we do binary classification for each AU, and the number of positive and negative sample are very imbalanced (i.e., negative samples are too much with respect to positive samples). Thus, we do over-sampling for rare AUs, and down-sampling for some negative samples. In V-A dataset [38], we first quantize the V-A value into the range of -5 and 5, and then we analysis the histogram of V-A values as shown in Figure 6. Here, we also discover the imbalanced distribution that would produce bias for V-A estimation. Thus, we also do over-sampling for some samples with rare V-A value, and down-sampling for some samples with frequent V-A values (e.g., valence with value of 0 and 1, arousal with value of 1).



(a) (b)

Figure 7. The block structure of (a) *rPoly-2* and (b) *rPoly-3*.

4.1.2 Convolutional blocks

In previous sections, we have described the network structure of the proposed *FATAUVA-Net*. As to the convolutional blocks in core layer, attribute layer, AU layer, and V-A layer, they could be replaced with only one convolutional layer or some popular convolutional blocks like Inception block [31] or Residual block [15]. Here, we design two blocks named *rPoly-2* and *rPoly-3*, which could be viewed as reduced blocks of *Poly-2* and *Poly-3* in PolyNet [42], and the block structures are shown in Figure 7(a) and Figure 7(b), respectively. The proposed mixed *rPoly-2* and *rPoly-3* network (*mrPoly-Net*) is defined as: we employ 8 *rPoly-2* blocks for core layer, 2 *rPoly-2* blocks for attribute layer, 2 *rPoly-3* blocks for AU layer, and 2 *rPoly-3* blocks V-A layer. Note that in this paper, the main contribution is the proposed framework of V-A estimation based on AU features. Thus, we only perform this framework with *mrPoly-Net* and all implementations were based on Caffe [16]. As to the performances of Inception block [31] or Residual block [15] for the above tasks, they are out of the scope of this paper.

4.1.3 Loss layer in training phase

In the training phase of facial attribute recognition and AU detection, we append three FC layers to attribute layer and

Table 4. The performance of different approaches for facial attribute recognition in terms of accuracy.

Attribute	[32]	[14]	Ours
arched eyebrows	0.8408	0.8342	0.8410
attractive	0.8262	0.8306	0.8143
bushy eyebrows	0.9075	0.9284	0.9228
eyeglasses	0.9889	0.9963	0.9938
male	0.9774	0.9817	0.9757
mouth slightly open	0.8927	0.9374	0.9339
narrow eyes	0.8599	0.8723	0.9072
no beard	0.9638	0.9605	0.9503
smile	0.9265	0.9273	0.9243
young	0.8892	0.8848	0.8703
AVG	0.9073	0.9154	0.9134

AU layer (as described in Section 3.1 and Section 3.2), and then we append the loss layer (softmax) to the last FC layer. Here, we apply binary classification for each attribute and AU, so the output dimension of the last FC layer is 2. In the training phase of V-A estimation. We append three FC layers to V-A layer, and employ two types of loss layer : class-based and regression-based. The loss layer for class-based is softmax, while the classes are the discrete number in the range -5 to 5 (as shown in Figure 6). As to the prediction function for class-based loss, we first select the top-3 classes according to the predicted probabilities. If the three classes are sequential (e.g., 1,2,3 or 1,3,2), we apply weighted sum (probability as weight) to estimate the V-A intensity. If the top-3 classes are not sequential, we only predict the top-1 class as result. For regression-based loss, similar to [35], we utilize the combination of center loss [36] and smooth $\ell 1$ loss [11]. The center loss is defined as:

$$loss_c = -\|x - c_{\tilde{y}}\|_p^p \,, \tag{1}$$

where x is the deep feature of the second last FC layer, $c_{\tilde{y}}$ is the deep features of $\tilde{y}$ class center. The smooth $\ell 1$ loss is utilized to measure the regression loss which is less sensitive to outliers than the $\ell 2$ loss:

$$loss_r = \begin{cases} 0.5|y - \tilde{y}|^2 & \text{if } |y - \tilde{y}| < t \\ |y - \tilde{y}| - t + 0.5t^2, & \text{otherwise} \end{cases} \tag{2}$$

Here, y and $\tilde{y}$ are the prediction and the ground truth value, respectively, t is the turning point between $\ell 2$ distance and $\ell 1$ distance. In following subsection, we will show the performance of class-based loss and regression-based loss on V-A estimation.

4.2. Result of facial attribute recognition

As to facial attribute recognition, we select 10 attributes from CelebA [22] dataset to train the core layer and attribute layer in *FATAUVA-Net*. The number of image in training phase is about 160,000, and the remains (about 40,000) are used for evaluation (the partitions are given by the dataset). Here, we compare the performance of our proposed net with two state-of-the-art approaches [32, 14]. As shown in Table 4, the proposed approach can achieve promising results in terms of accuracy, and we can observe that the facial part-based approach is suitable for attribute recognition. Moreover, the trained core layer and attribute layer further could be utilized to guide the training process of AU detection.

Table 5. The performance of different approaches for AU detection in terms of AUC.

AU	DRML [44]	EAC-Net [21]	Ours
1	0.557	0.689	0.548
2	0.545	0.739	0.552
4	0.588	0.781	0.721
6	0.566	0.785	0.780
7	0.610	0.690	0.621
10	0.536	0.776	0.787
12	0.608	0.846	0.861
15	0.562	0.781	0.678
17	0.500	0.706	0.688
23	0.539	0.810	0.636
24	0.539	0.824	0.803
AVG	0.562	0.774	0.698

Table 6. The performance of different approaches for AU detection in terms of F1-score.

AU	DRML [44]	EAC-Net [21]	Ours
1	0.364	0.390	0.318
2	0.418	0.352	0.248
4	0.430	0.486	0.455
6	0.550	0.761	0.729
7	0.670	0.729	0.737
10	0.663	0.819	0.802
12	0.658	0.862	0.815
15	0.332	0.375	0.384
17	0.480	0.591	0.556
23	0.317	0.359	0.310
24	0.300	0.358	0.407
AVG	0.462	0.541	0.516

4.3. Result of AU detection

As to the training phase for AU detection, we learn AU layers for AU detection from the dataset in FERA2015 [34], and we select 14 AUs (as shown in Table 3) from both BP4D database [43] and SEMAINE database [26]. Here, we only include the training partition to learn the AU layer. In testing phase, we evaluate the performance only on 11 AUs in BP4D database (there are no results for the comparisons on SEMAINE database). In BP4D database, 41 participants

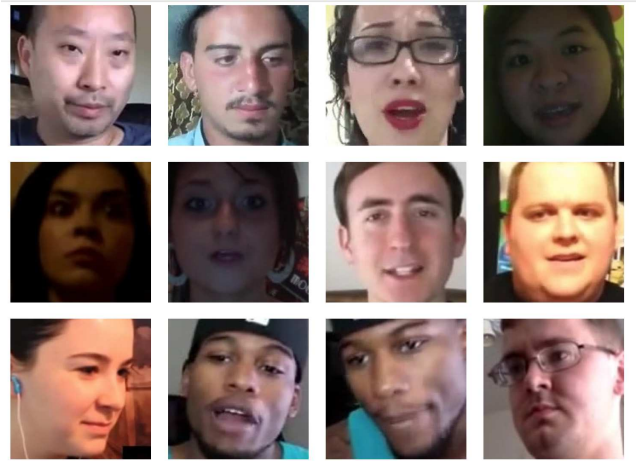


Figure 8. The example of face images in AFF-Wild [38] dataset, and each row shows the variances in gender, light condition, and head pose, respectively.

were recorded with different expressions in the format of videos, and each subject participated 8 sessions of experiments. Here, we follow the partition given by this dataset. It means that we use 21 participants' videos for training, and 20 participants' videos for evaluation. In SEMAINE database, the training partition contains 16 sessions, and there are 15 sessions in development partition. The number of image in training partition is about 48,000, and there are about 45,000 images in development partition. Moreover, both of these databases are in laboratory setting. Table 5 and Table 6 show the results of AU detection in terms of AUC and F1-score, respectively, and we compare our approach with two CNN-based approaches: DRML [44] and EAC-Net [21]. Although the training-testing partitions in DRML and EAC-Net are slightly different to ours, our approach can achieve competitive results. Furthermore, the learned AU layer could be viewed as mid-level representation to connect face attribute and V-A intensity, and AU layer further could be utilized to help V-A estimation.

4.4. Result of V-A estimation

After learning AU layer, we adopt the data from AFF-Wild Challenge [38] to train V-A layer for V-A estimation. In this dataset, there are 253 videos with frame-by-frame annotations for training and 47 videos for testing, and this is the first benchmark for estimating valence and arousal "in-the-wild". As shown in Figure 8, there are various conditions on faces in this dataset, for example, gender, light condition, and head pose.

Here, we examine the performance of two network structures (i.e., with or without AU layer) and two kinds of loss layers (i.e., class-based and regression-based), while the evaluation metrics are the Mean Squared Error (MSE) and Concordance Correlation Coefficient (CCC). Here, we

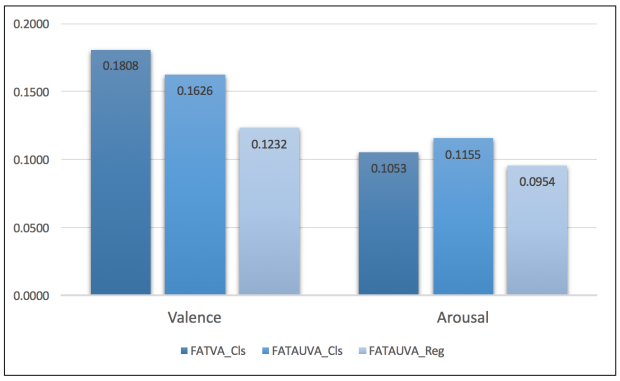


Figure 9. The performances of three approaches for V-A estimation in terms of MSE.

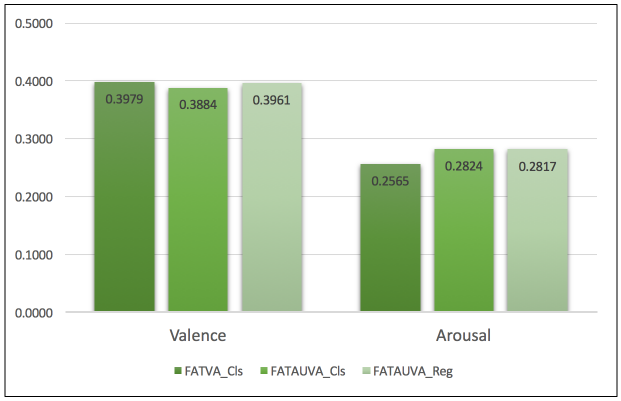


Figure 10. The performances of of three approaches for V-A estimation in terms of CCC.

naively map the results of our method (in range of -5 to 5) to the range of -1 to 1 for evaluation. As shown in Figure 9 and Figure 10, we can observe that the performance of AU layer method (FATAUVA_Cls) could achieve approximate or better result than the method without AU layer (FATVA_Cls). It means that AUs are useful for V-A estimation. Moreover, we also notice that the appoach with regression-based loss (FATAUVA_Reg) could achieve better result than class-based loss (FATAUVA_Cls).

5. Conclusion and Future Work

In this paper, we propose an integrated deep learning framework for V-A estimation. In our framework, we train core layer, attribute layer, AU layer, and V-A layer sequentially, and these layers further could be utilized to solve the task of facial attribute recognition, AU detection, and V-A estimation. We first learn the facial part-based response through attribute recognition CNNs, and then apply these layers to supervise the learning of AU detection. Finally, we employ AUs as mid-level representation to estimate the intensity of valence and arousal. In experiments, we have shown the promising performances of our proposed frame-

work in the above three tasks, and we also have shown that AUs are useful for V-A estimation. Moreover, with suitable loss layer, our proposed method for V-A estimation could achieve competitive performance.

For future work, since the core layer and attribute layer are only trained by CelebA dataset and these two layers are important for AU detection and V-A estimation, we would like to include more images to fine-tune these two layers. As to AU layer, the images for training AU detector were captured in constrained condition and the number of image was not enough. Thus, we would like to train new model with a large number of in-the-wild images with AU annotations. For V-A estimation, there should be more investigations to examine the relationship between AUs and V-A intensity, and the experiments of using various combinations of AUs for constructing V-A layer would be performed.

References

[1] Y. Baveye, E. Dellandreay, C. Chamaret, and L. Chen. Deep learning vs. kernel methods: Performance for emotion prediction in videos. In *ACII*, 2015.

[2] C. F. Benitez-Quiroz, R. Srinivasan, and A. M. Martinez. EmotioNet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *IEEE CVPR*, 2016.

[3] T. Berg and P. N. Belhumeur. POOF: Part-based one-vs-one features for fine-grained categorization, face verification, and attribute estimation. In *IEEE CVPR*, 2013.

[4] N. Bosch, S. K. D'mello, J. Ocumpaugh, R. S. Baker, and V. Shute. Using video to automatically detect learner affect in computer-enabled classrooms. *ACM Transactions on Interactive Intelligent Systems*, 6(2):17:1–17:26, 2016.

[5] K. Brady, Y. Gwon, P. Khorrami, E. G. adn W. Campbell, C. Dagli, and T. S. Huang. Multi-modal audio, video and physiological sensor learning for continuous emotion prediction. In *ACM AVEC*, 2016.

[6] B.-C. Chen, C.-S. Chen, and W. H. Hsu. Face recognition using cross-age reference coding with cross-age celebrity dataset. *IEEE Transactions on Multimedia*, 17(6):804 – 815, 2015.

[7] D. Chen, S. Ren, Y. Wei, X. Cao, and J. Sun. Joint cascade face detection and alignment. In *ECCV*, 2014.

[8] C. A. Corneanu, M. Oliu, J. F. Cohn, and S. Escalera. Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8):1548 – 1568, 2016.

[9] H. Ding, S. K. Zhou, and R. Chellappa. FaceNet2ExpNet: Regularizing a deep face recognition net for expression recognition. *arXiv:1609.06591*, 2016.

[10] P. Ekman and E. L. Rosenberg. What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS), 2nd Edition. *Oxford University Press*, 2015.

[11] R. Girshick. Fast r-cnn. In *IEEE ICCV*, 2015.

[12] H. Gunes and B. Schuller. Categorical and dimensional affect analysis in continuous input: Current trends and future direction. *Image and Vision Computing*, 31(2):120–136, 2013.

[13] S. Han, Z. Meng, A. S. Khan, and Y. Tong. Incremental boosting convolutional neural network for facial action unit recognition. In *NIPS*, 2016.

[14] E. M. Hand and R. Chellappa. Attributes for improved attributes: A multi-task network utilizing implicit and explicit relationships for facial attribute classification. In *AAAI*, 2017.

[15] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE CVPR*, 2016.

[16] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv:1408.5093*, 2014.

[17] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim. Joint fine-tuning in deep neural networks for facial expression recognition. In *IEEE ICCV*, 2015.

[18] P. Khorrami, T. L. Paine, and T. S. Huang. Do deep neural networks learn facial action units when doing expression recognition? In *IEEE ICCV Workshop*, 2015.

[19] J. Kossaifi, G. Tzimiropoulos, S. Todorovic, and M. Pantic. AFEW for valence and arousal estimation in-the-wild. *Image and Vision Computing*, 2017.

[20] N. Kumar, A. Berg, P. N. Belhumeur, and S. Nayar. Describable visual attributes for face verification and image search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(10):1962–1977, 2011.

[21] W. Li, F. Abtahi, Z. Zhu, and L. Yin. EAC-Net: A region-based deep enhancing and cropping approach for facial action unit detection. *arXiv: 1702.02925*, 2017.

[22] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *IEEE ICCV*, 2015.

[23] Y. Lu, A. Kumar, S. Zhai, Y. Cheng, T. Javidi, and R. Feris. Fully-adaptive feature sharing in multi-task networks with applications in person attribute classification. *arXiv:1611.05377*, 2016.

[24] P. Lucey, J. F. Cohn, K. M. Prkachin, P. E. Solomon, and I. Matthews. Painful data: The UNBC-McMaster shoulder pain expression archive database. In *IEEE FG*, 2011.

[25] D. McDuff, R. E. Kaliouby, J. F. Cohn, and R. Picard. Predicting ad liking and purchase intent: Large-scale analysis of facial responses to ads. *IEEE Transactions on Affective Computing*, 6(3):223 – 235, 2015.

[26] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder. The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing*, 3(1):5–17, 2012.

[27] M. Mehu and K. R. Scherer. Emotion categories and dimensions in the facial communication of affect: An integrated approach. *Emotion*, 15(6):798–811, 2015.

[28] M. A. Nicolaou, H. Gunes, and M. Pantic. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE Transactions on Affective Computing*, 2(2):92–105, 2011.

[29] R. Ranjan, S. Sankaranarayanan, C. D. Castillo, and R. Chellappa. An all-in-one convolutional neural network for face analysis. *arXiv:1611.0085*, 2016.

[30] J. A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 1980.

[31] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *IEEE CVPR*, 2015.

[32] R. Torfason, E. Agustsson, R. Rothe, and R. Timofte. From face images and attributes to attributes. In *ACCV*, 2016.

[33] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. T. Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic. AVEC 2016: Depression, mood, and emotion recognition workshop and challenge. In *ACM AVEC*, 2016.

[34] M. F. Valstar, T. Almaev, J. M. Girard, G. McKeown, M. Mehu, L. Yin, M. Pantic, and J. F. Cohn. FERA 2015 - Second facial expression recognition and analysis challenge. In *IEEE FG Workshop*, 2015.

[35] F. Wang, X. Xiang, C. Liu, T. D. Tran, A. Reiter, G. D. Hager, H. Quon, J. Cheng, and A. L. Yuille. Transferring face verification nets to pain and expression regression. *arXiv:1702.06925*, 2017.

[36] Y. Wen, K. Zhang, Z. Li, and Y. Qiao. A discriminative feature learning approach for deep face recognition. In *ECCV*, 2016.

[37] S. Yang, P. Luo, C. C. Loy, and X. Tang. From Facelet-Net: Face detection through deep facial part responses. *arXiv:1701.08393*, 2017.

[38] S. Zafeiriou, M. Nicolao, I. Kotsia, F. Benitez-Quiroz, and G. Zhao. Aff-wild: Valence and arousal in-the-wild challenge. In *IEEE CVPR Workshop*, 2017.

[39] S. Zafeiriou, A. Papaioannou, I. Kotsia, M. Nicolaou, and G. Zhao. Facial affect in-the-wild: A survey and a new database. In *IEEE CVPR Affect in-the-wild Workshop*, 2016.

[40] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *ECCV*, 2014.

[41] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multi-task cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.

[42] X. Zhang, Z. Li, C. C. Loy, and D. Lin. PolyNet: A pursuit of structural diversity in very deep networks. *arXiv:1611.05725*, 2016.

[43] 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.

[44] K. Zhao, W.-S. Chu, and H. Zhang. Deep region and multi-label learning for facial action unit detection. In *IEEE CVPR*, 2016.

[45] L. Zhong, Q. Liu, P. Yang, B. Liu, J. Huang, and D. N. Metaxas. Learning active facial patches for expression analysis. In *IEEE CVPR*, 2012.