Estimating Multiple Emotion Descriptors by Separating Description and Inference

Didan Deng[1], Bertram E. Shi[2]
Department of Electronic and Computer Engineering,
Hong Kong University of Science and Technology, Kowloon, Hong Kong
ddeng@connect.ust.hk[1] eebert@ust.hk[2]

Abstract

To describe complex emotional states, psychologists have proposed multiple emotion descriptors: sparse descriptors like facial action units, continuous descriptors like valence and arousal, and discrete class descriptors like the expressions of happiness and anger. According to Cohn et al. [1], facial action units are sign vehicles that convey the emotion message, while discrete or continuous emotion descriptors are the messages perceived by observers. They differ in their focuses. Sign vehicles focus on describing facial behavior. Emotion messages focus on an observer's inference about the underlying state of the subject from facial behavior.

We describe a novel architecture for multiple emotion descriptor estimation that incorporates this prior knowledge about the differences between descriptive labels (sign vehicles, like facial action units) and inferential labels (emotion messages like discrete emotion expressions, valence, and arousal). In our multi-level architecture, a common set of low-level features of facial regions are fed into two separate branches: one for descriptive labels and the other for inferential labels. The differences between these two branches reflects the differences between the two types of labels. Sign vehicles are typically more specific and spatially localized. Emotion messages are reflected across the entire face. Our experiments on the ABAW3 challenge [9] dataset demonstrate this approach outperforms all other submitted approaches to multi-task learning. Code is available at https://github.com/HKUST-NISL/ABAW3_MultiEmotionNet.

1. Introduction

To study facial behavior, one can use two approaches: describing the surface manifestations, *e.g.*, how the facial muscles move, or making inferences about the underlying causes, *e.g.*, emotions [1]. The first approach corresponds to measuring sign vehicles, *e.g.*, facial action units. Observers using a sign-based approach are sometimes referred to as "coders" because they ideally describe changes in facial appearance objectively. Although individual differences among coders, *e.g.*, annotation skills, do affect labels, the sign-based approach focuses on the facial behavior itself, rather than the underlying causes. The second approach corresponds to inferring messages (judgments). These messages might be discretized to different categories (*e.g.*, expressions) or measured along different continuous dimensions (*e.g.*, valence and arousal). Observers following a message-based approach are often referred to as "judges" or "raters" because they need to infer emotional messages from facial appearance. The measurement of messages is much more affected by individual differences among raters, such as gender, personality, and culture, than the measurement of sign vehicles, because inferring messages relies upon interpretation of the facial appearance. Understanding the differences between these two approaches is the key to designing an effective architecture for estimating them simultaneously.

Sign vehicles like facial action units are external manifestations of an underlying emotional state. However, inferring emotional states purely from facial action units is usually not accurate. Facial action units are an incomplete and often heavily quantized (*e.g.*, on a binary or 0 to 5 integer scale) descriptions of the facial muscles' activation. They focus on specific facial regions while ignoring other parts. Thus, they cannot capture all of the nuances about the facial expression. This suggests that it is essential to extract more low-level features than the number of AUs described by the facial action coding system to make accurate inferences about emotion messages.

Measuring emotional messages relies less on a detailed description of certain facial regions, and more on integration of information from across multiple facial regions. While an observer can infer the mental state of a person from a small part of face, such as eyes, the inference can be more accurate if the observer gathers the evidence from the entire face. Therefore, our architecture projects visual

features from different facial regions into a shared feature space. The feature learning of these facial regions is partially guided by AU labels, so that we can better exploit the richer information available in a multi-task framework, but also includes additional information. After projection, the features are integrated by averaging to obtain a consensus feature vector. This consensus feature vector is used as the basis of inference about the emotion messages (*i.e*., facial expression, valence, and arousal).

Our main contributions are as follows:

- We propose a novel model architecture, the Sign-and-Message Multi-Emotion Net (SMM-EmotionNet) that simultaneously estimates both descriptive labels (*i.e*., the facial action units) and inferential labels (*i.e*., the facial expressions, valence, and arousal).

- We employ the psychological prior knowledge in our architecture design to regularize the multiple emotion descriptor estimation. This is particularly helpful in the absence of complete emotion annotations, and enables us to integrate information across more datasets to improve performance.

2. Related Work

Multitask emotion models are scarcer than uni-task emotion models, mainly because of the lack of datasets with complete annotations of multiple emotion descriptors. Recently, since the release of the Aff-wild and the Aff-wild2 dataset [13, 14, 24], there has been an increasing focus on the simultaneous prediction of three emotion descriptors: facial action units (AU), facial expressions (EXPR) and valance/arousal (VA) [15].

In the ABAW2 Challenge [10], Deng *et al*. [3] proposed a light-weight CNN-RNN model to predict three emotion descriptors in video sequences. In their uni-modal (visual modality) approach, they used a common architecture: a feature extractor shared by multiple tasks, followed by several branches. Each branch corresponds to one task. Since they did not recognize the different properties of each task, they used similar branches for all of them. A similar architecture was used in their approach [2] submitted to the first ABAW Challenge.

Zhang *et al*. [25] designed a dedicated architecture for three emotion descriptors prediction in the second ABAW Challenge. Making and assumption about the relationship between them, they designed a serial recognizer: AU (action unit) $\rightarrow$ EXPR (facial expressions) $\rightarrow$ VA (valence and arousal), which proceeded from local action units to global emotion states. Their work and our work both aim to employ prior knowledge about the relationship of multiple emotion descriptors in architecture design. However, since we utilize the theory proposed by [1], the data flows in our model along two parallel branches: one from the facial regions to the sign vehicles space (AU) and another from the facial regions to the message space (EXPR and VA).

Kollias *et al*. [11] proposed the "co-annotation" method utilizing prior knowledge about the relationship between facial expressions and action units, which provides better robustness to data distribution shifts. They assigned annotations to missing labels based on the labels of other emotion descriptors. For example, "Happiness" is automatically assigned if AU12 (lip corner puller) and certain AUs are activated. They only considered relationships between discrete emotions and action units for co-annotation. A similar method was used in [12]. Our approach shares a similar motivation as the co-annotation method: exploiting prior knowledge for multiple emotion descriptors estimation. However, rather than using prior knowledge to fill in missing annotations, we utilize insights from psychological studies in designing the structure of our network. This captures the similarities and differences between the descriptors, but avoids the possible introduction of labeling errors due to a rigid rule-based assignment.

3. Methodology

3.1. Model Architecture

The architecture of our proposed Sign-and-Message EmotionNet is shown in Figure 1.

Sign Vehicle Space. We define the sign vehicle space as the metric space for inferring AUs. To learn representations for each AU, we use the Emotion Transformer proposed by Jacob *et al*. [7], who suggested that exploiting intra-AU attention and inter-AU correlations is the key component in AU prediction. The Emotion Transformer is shown in parts I and II of Figure 1.

In part I of Figure 1, the feature extractor is an InceptionV3 model. The size of the extracted feature map is 17×17 and it has 768 channels. The feature map is then fed into ROI attention modules that learn spatial attention maps for each region. Each attention map has the same size as the feature map. The element-wise product between the attention map and the feature map is fed into the ROI embedding module to generate the feature vector of each region. The general architecture of our part I is the same as Jacob *et al*. [7], but some details are different, *e.g*., the number of facial regions U and the feature dimension D.

In part II of Figure 1, the features of different facial regions are fed into the Transformer model [22]. The Transformer model can learn the correlations among the different regions, which has been proved highly effective in AU detection [7]. We define the output tensors of the Transformer model as the features in the sign vehicle space. The sign vehicle space is composed of 12 D-dimensional spaces, one for each of the AUs. Twelve fully connected (FC) layers

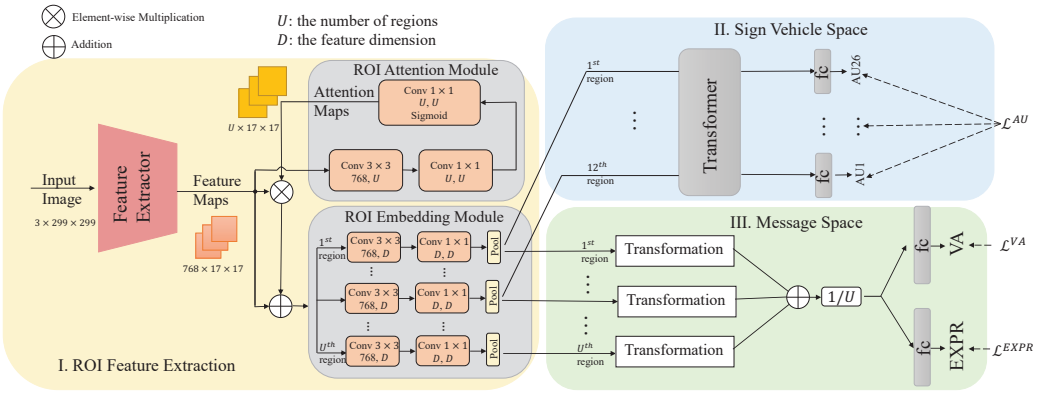


Figure 1. The architecture of our SMM-EmotionNet.

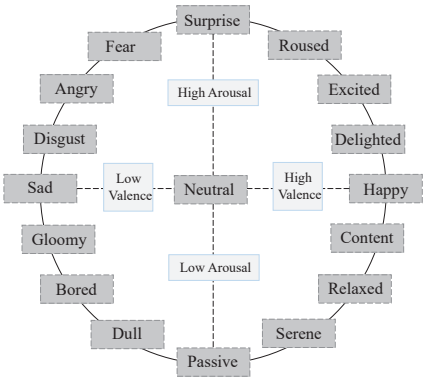


Figure 2. Russell's circumplex model.

with a sigmoidal activation function map the AU metric spaces to estimates of the probabilities of each AU's occurrence.

Message Space. Russell's circumplex model [19], a psychological emotion model derived from large-scale crowdsourcing research and shown in Figure 2, is the motivation for our idea of learning the two emotion descriptors in a shared message space. Russell's circumplex model describes a 2-dimensional circular space. The horizontal axis is valence. The vertical axis is arousal. Discrete facial expressions can be mapped onto this 2D space, indicating their close relationship with valence and arousal. If this model is accurate for most emotion datasets, then we expect that learning a shared message space should regularize feature learning of one emotion descriptor when the annotation of the other is absent.

The message space is the metric space for facial expressions, valence, and arousal. Given the features on the message space, we feed them into two FC layers to estimate EXPR and VA.

Notations. Given a facial image x, we first extract the regions of interest features. We denote the u^{th} ROI feature vector as $f^{(u)}(x) \in \mathbb{R}^D$. D is the feature dimension. For all U ROI features, they are denoted as $F^{(U)}(x) = \{f^{(u)}(x)\}_{u=1}^{U}$.

The Transformer in part II (Figure 1) transforms the ROI features into the AU metric space. We denote the number of action units to be estimated as H. In this paper, $H = 12$. Note that when $U > H$, it means that we have more facial regions than the number of action units to be estimated. We simply feed the first H ROI features into the Transformer. The output of the Transformer is given by:

$$S^{(H)}(x) = \Phi(F^{(H)}(x) + PE), \tag{1}$$

where PE denotes the positional encoding vector. Φ denotes the Transformer function. $S^{(H)}(x) = \{S^{(h)}(x)\}_{h=1}^{H}$, where $S^{(h)}(x) \in \mathbb{R}^D$ is the feature vector on the h^{th} AU's metric space. We denote the weight matrix of the last FC layer for h^{th} AU as $W_h \in \mathbb{R}^D$. The output of this FC layer is given by:

$$y_h^{AU} = \langle W_h, S^{(h)}(x) \rangle + b_h, \tag{2}$$

where $\langle \cdot, \cdot \rangle$ is the inner product between the weight vector W_h and the feature vector $S^{(h)}(x)$. b_h is the bias term.

For the message space learning in part III (Figure 1), the transformation module is learned for each region through the back-propagation. Since we do not have assumptions on the transformation module, we consider the simplest case: it is a linear transformation matrix. We denote the weight matrix as $A^{(u)} \in \mathbb{R}^{D \times D}$ for the u^{th} region. The transformed vector on the message space can be denoted as:

$$M^{(u)}(x) = \langle A^{(u)}, f^{(u)}(x) \rangle. \tag{3}$$

Given multiple transformed vectors, we can obtain the consensus by taking the average over all of them: $\bar{M}^{(U)}(x) = \frac{1}{U} \sum_u^U M^{(u)}(x)$. Next, we can compute the

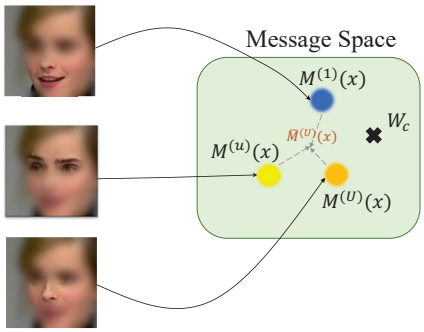


Figure 3. The process of averaging transformed vectors on the message space.

prediction of c^{th} class: $\hat{y}_c = \langle W_c, \bar{M}^{(U)}(x) \rangle + b_c$, where W_c is the weight vector for the c^{th} class. Alternatively, we can take the average over $\langle W_c, M^{(u)}(x) \rangle$ for all $u \leq U$. Because of the linearity of the last FC layer, the two approaches are equivalent. We use the first approach in this work. A diagram in Figure 3 illustrates the process of feature projection and averaging on the message space.

3.2. Losses

For the AU prediction task, the inference loss is the binary cross entropy loss. Given the input x, the AU prediction is denoted by $\hat{y}^{AU}$. The ground truth label for AU is denoted by y^{AU}. The inference loss of the AU task is given by:

$$\mathcal{L}^{AU}(\hat{y}^{AU}, y^{AU}) = -\frac{1}{U}\sum_{i}^{U} P_i^{AU} y_i^{AU} \log(\sigma(\hat{y}_i^{AU})) + (1 - y_i^{AU}) \log(1 - \sigma(\hat{y}_i^{AU})), \tag{4}$$

where P_i^{AU} is the weight of each AU for data balancing. It is computed from the training set data distribution. P_i^{AU} equals to the number of negative samples divided by the number of positive samples for each AU. $\sigma(\cdot)$ represents the sigmoidal function.

For facial expression (EXPR) classification, we use a cross-entropy loss as the inference loss shown in Equation 5.

$$\mathcal{L}^{EXPR}(\hat{y}^{EXPR}, y^{EXPR}) = -\sum_{i}^{C} P_i^{EXPR} y_i^{EXPR} \log(\rho_i(\hat{y}^{EXPR})), \tag{5}$$

where $\rho_i(\hat{y}^{EXPR}) = \frac{exp(y_i^{EXPR})}{\sum exp(y_i^{EXPR})}$ denotes the Soft-max function. P_i^{EXPR} is the re-weighting factor, which is computed from the training set data distribution.

Finally, for valence and arousal (VA) prediction, we use the negative Concordance Correlation Coefficient (CCC) as the inference loss.

$$\mathcal{L}^{VA}(\hat{y}^{VA}, y^{VA}) = 1 - CCC^V + 1 - CCC^A. \tag{6}$$

To learn multiple tasks, we use a unweighted sum to combine different inference losses:

$$L = \mathcal{L}^{AU} + \mathcal{L}^{EXPR} + \mathcal{L}^{VA}. \tag{7}$$

4. Experiments

4.1. Datasets

The datasets provided by the ABAW3 challenge can be divided into two categories: the uni-task video datasets and the multi-task learning (MTL) static datasets.

The videos of the uni-task datasets are the same as the videos from the Aff-wild2 database [14], including the AU subset, the EXPR subset, and the VA subset. Each subset is annotated with one corresponding emotion descriptor. What is different is that, in addition to 7 facial expressions (six basic emotions plus neutral) annotated in the Aff-wild2 EXPR subset, the ABAW3 challenge organizers annotated another facial expression: "other". The category "other" indicates facial expressions which cannot be classified into six basic emotions, nor the neutral category. For example, the "bored" expression shown in Figure 2 is one of the "other" expressions. In total, there are 548 videos of around 2.7 million frames in the uni-task video datasets.

The MTL static dataset contains only a subset of frames from the Aff-wild2 dataset. Each frame is labeled with complete annotations: 8 facial expressions, 12 action units, valence, and arousal. In total, there are around $175,000$ images in the MTL static dataset.

In our experiments, we did not use the MTL static dataset which has complete emotion annotations. We only used the video data from the three uni-task datasets. There are mainly two reasons. Firstly, we notice that the MTL static dataset has overlapped videos with the three uni-task datasets in both the training and the validation sets. It is problematic to use the three uni-task datasets and the MTL static dataset simultaneously because of the data leakage problem. Secondly, our approach is proposed to alleviate the problem of lack of complete annotations by regularizing feature learning with prior knowledge. It is very flexible about incomplete annotations. Therefore, we ignored the MTL static dataset and only used the uni-task datasets for training and validation. However, for the performance evaluation on the test set , we submitted our predictions on the test set of the MTL static dataset. This is part of the competition requirements. Besides, we aim to evaluate the

Model	F1-AU	F1-EXPR	CCC-V	CCC-A
Single-task Baseline	0.39	0.23	0.31	0.17
Multi-task Baseline	0.536	0.489	0.441	0.485
Ours	**0.548**	**0.518**	**0.447**	**0.499**

Table 1. The comparison between our static model and two baseline models (also static). The results are evaluated on validation sets of the AU, EXPR and VA subset from the Aff-wild2 dataset. F1-AU stands for the unweighted F1 score of 12 action units. F1-EXPR stands for the macro F1 score of eight facial expressions. CCC-V stands for the CCC value of valence. CCC-A stands for the CCC value of arousal.

overall performance with the annotations of three emotion descriptors given that our model were trained with incomplete annotations.

In the uni-task datasets we used, the AU subset contains around 1.8 million frames; the EXPR subset contains around 0.8 million frames; the VA subset contains around 1.8 million frames. To reduce the training time, we downsampled the number of frames in the training set by 8 times, while keeping the frames in the validation sets intact.

4.2. Training Details

The cropped and aligned faces are provided by the ABAW3 challenge. We resized the face images to the size of $299 \times 299 \times 3$ in pixels. For image augmentation, we used random cropping, random horizontal flipping, and random color jitter. Since the InceptionV3 model were firstly pretrained on ImageNet [5] and then finetuned on the training set, we used the mean and standard deviation of ImageNet to normalize face images.

The dimension of each ROI embedding is $D = 16$. To find the best number of facial regions, we performed a grid search of $U \in [12, 17, 27]$. 12 is the number of AUs with annotations in the AU subset. 17 is the number of AUs related to emotions found in [18]. 27 is the number of major AUs in the facial action coding system [6]. The three values were selected given prior knowledge about action units, because there existed a correspondence between some facial regions and the action units. From the grid search results, we found that the multitask performance when $U = 17$ was better than the performance when $U = 12$, but very close to the performance when $U = 27$. Therefore, we chose the number of regions as $U = 17$. Since $U = 17$ is larger than the number of AUs with annotations (*i.e.*, 12), five ROI embedding vectors are not fed into the Transformer model to learn the sign vehicle space. They only provide certain degrees of freedom to message space learning.

The optimizer we used is SGD. The momentum of SGD is equal to 0.9. The initial learning rate is 10^{-3}. A cosine annealing learning rate schedule is used to improve convergence. The total number of training iterations is 3×10^5. The batch size is 72, where for each task, 24 images are sampled from one of three uni-task datasets in the same batch.

4.3. Evaluation Metrics

For the AU detection task, we used the averaged F1 score of 12 AUs to evaluate performance. For discrete emotion classification, we used the averaged macro F1 score of eight facial expressions (six basic emotions, neutral and others) as the evaluation metric. For continuous emotion prediction, we used CCC as the evaluation metric.

In the multitask learning (MTL) challenge, the performance metric is defined as the weighted sum of the evaluation metric of each task.

$$P = \frac{1}{2}(CCC^V + CCC^A) + \frac{1}{8}\sum_{i}^{8} F_1^{EXPR_i} + \frac{1}{12}\sum_{j}^{12} F_1^{AU_j}. \quad (8)$$

5. Results

In this section, we introduce the evaluation results of two recognizers: a static model, which is the model introduced in Section 3, and a temporal model, which applies temporal smoothing to the features learned by the static approach in an attempt to exploit the relationship between adjacent frames.

5.1. Static Approach

We first compared our static approach with two baseline models. In the ABAW3 challenge, the official baseline model provided by the challenge organizer is a single-task CNN model (ResNet50 model for VA; VGG16 model for EXPR and AU) [16]. They altered the size of the last FC layer to adapt to the dimension of emotion descriptors. We also compared our SMM-EmotionNet with a multi-task baseline model: an InceptionV3 feature extractor followed by parallel branches of FC layers corresponding to different tasks. In each branch, the sizes of consecutive FC layers are: $768 \times 16 \rightarrow 16 \times C$. C is the dimension of the corresponding emotion descriptor. This multi-task baseline model is trained with the same hyper-parameters (*e.g.*, batch size) as our proposed model.

The evaluation results are shown in Table 1. Compared with the single-task and multitask baseline models, our model shows superior performance on every task. Since the architecture of the feature extractor and training hyper-parameters are the same in our model and the multi-task

Fold ID	F1-AU	F1-EXPR	CCC-VA (Averaged)
1	0.463	0.428	0.542
2	0.586	0.552	0.577
3	0.576	0.450	0.558

Table 2. The three-fold cross-validation results given the optimal μ on each of the validation sets in the AU, EXPR and VA subset. CCC-VA stands for averaged CCC values of valence and arousal.

Model	Val sets (uni-task)	Test set (MTL)
ours (static)	1.539	1.104
ours (temporal)	**1.664**	**1.113**
IMLAB	-	0.953
HSE-NN	-	0.809
NFVH	1.480	0.675

Table 3. The multitask evaluation metric (Equation 8) on the validation sets of uni-task subsets and the test set of the MTL static dataset.

baseline, the advantage of our model seems to result from our use of distinct architectures for the sign vehicle space and the message space.

5.2. Temporal Smoothing

We considered a simple temporal smoothing method to filter high-frequency noise in a sequence of feature vectors. We did not re-train our static model, but used it to extract features for the sign vehicles or the emotional messages. After temporal smoothing, these features were fed into the last FC layer for classification or regression.

For the feature vector $\epsilon_t = M^{(u)}(x_t)$ or $\epsilon_t = S^{(u)}(x_t)$ at the time step t, we smoothed it with the this function:

$$\Lambda_t = \frac{1}{1+\mu}(\epsilon_t + \mu\Lambda_{t-1}). \tag{9}$$

Λ_t is the feature vector at the time step t after smoothing. μ is a hyper-parameter which determines the smoothness. Larger μ indicates higher dependency of the current variable on its previous values.

We found the best μ by grid search with three-fold cross-validation on the validation sets of the uni-task datasets. The search region is $\mu = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$. Given the cross-validation results, we chose the optimal μ for the AU task to be 7, and the optimal μ for EXPR and VA to be 9. The cross-validation results given the optimal values of μ are shown in Table 2.

5.3. Multi-task Learning Performance

We submitted our model to the MTL (Multitask Learning) challenge track.

Using the Equation 8 as the evaluation metric, we evaluated the performance of our static model and temporal model on the validation sets of the three uni-task datasets and the test set of the MTL static dataset. The results are shown in Table 3. The temporal smoothing method slightly increased overall multitask performance compared with the static approach. The improvement was around 0.8% on the test set of the MTL static dataset.

We also compared our models with the methods submitted by other teams in the same challenge, which we refer to by team names:

- IMLAB. Jeong *et al*. [8] proposed an audio-visual model with two streams to extract visual features from facial image sequences and audio features from audio signals. They then concatenated the visual and audio features and fed them to task-specific FC layers for emotion prediction.
- HSE-NN. Savchenko *et al*. [20] proposed a multitask model to predict facial attributes, such as age and gender, in addition to the emotions. The architecture they used was a CNN feature extractor followed by several branches for different tasks.
- Netease Fuxi Virtual Human (NFVH). Zhang *et al*. [26] proposed a multi-modal framework consisting of four streams: one vision stream extracts features from single visual frames, while the other three streams extract features from input sequences, *i.e*., the visual frame sequences, the audio signals, and the word embeddings of transcripts.

Table 3 compares our and other teams' approaches. Although we proposed a uni-modal approach, it outperformed all of the multi-modal approaches. The performance of multi-modal approaches' might have been affected by the qualities of the modalities. For example, facial images may be occluded. Audio signals may consist of background noise. The transcripts generated by out-of-the-box speech recognition models may contain errors. Previous papers on multi-modal emotion prediction [3,4] have shown the visual modality contributed the most to the performance of emotion prediction, while the audio and text modalities provided only supplementary information.

6. Ablation Study

This ablation study seeks to evaluate the extent to which using a shared space for EXPR and VA regularizes the learning of both emotion descriptors when using partially annotated datasets, *i.e*., the annotations of one emotion descriptor are absent. We compare our approach with a variant model that uses separate metric spaces for EXPR and VA, the message-space-separated (MSS) model.

Figure 4 shows the message space learning of the MSS model. Unlike the SMM-EmotionNet, the MSS model

Model	F1-AU	F1-EXPR	CCC-V	CCC-A
MSS	**0.562**	0.487	0.439	0.442
SMM	0.548	**0.518**	**0.447**	**0.499**

Table 4. The comparison between our SMM-EmotionNet with the variant model, MSS model, on the validation sets of uni-task datasets.

learns two separate spaces for EXPR and VA. In particular, the transformation modules (matrices) are not shared by EXPR and VA.

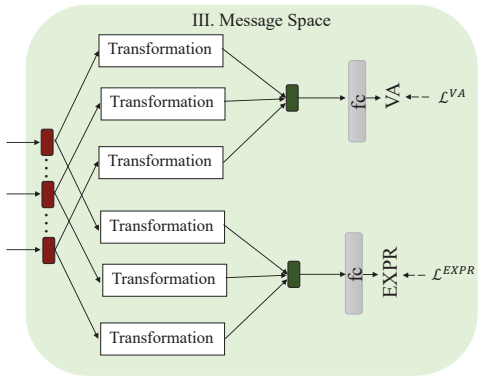


Figure 4. The message space learning of the message-space-separated (MSS) model.

We compared the SMM-EmotionNet with the MSS model on the validation sets of the uni-task datasets. The results are shown in Table 4. Using separate spaces for EXPR and VA degraded the performance metrics of EXPR and VA. The EXPR F1 was reduced by around 6%. The average CCC of valence and arousal was reduced by 7.5%. Although the AU F1 score was better for the MSS model, the overall performance evaluated using Equation 8 was reduced by around 4%. This suggests that without the mutual influence between EXPR and VA ensured by constraining to the shared space, the prior knowledge that EXPR and VA are closely related cannot be fully utilized.

7. Visualization

In this section, we visualized the features learned by our static model to verify whether they were consistent with prior knowledge. We used the validation sets of the three uni-task datasets for visualization.

Using t-SNE [21] to reduce the feature dimension to 2, we plot the distribution of the learned features in the sign vehicle space in Figure 5. There are 12 AUs, corresponding to the 12 spaces. The F1 score of each AU is given in its title. Some AUs' F1 scores are relatively low, for example, AU23 and AU24. We think due to the rare occurrence of these AUs in the training set, our model failed to learn very discriminative features for them. For other AUs, the differences between samples with AU presence and samples with AU absence are mostly angular differences. This is because we use a single FC layer for AU prediction. The predicted AU probability mainly depends on the inner product between the sign vehicle feature and the weight vector of the corresponding AU. Although the bias term will also affect the output, it shifts decision boundary for every sample regardless of the input.

Figure 6 shows the learned features in the message space after dimension reduction. Figure 6 (a) shows message space features, where the ground truth labels of eight facial expressions indicated by different colors. For each facial expression, we draw the covariance ellipse estimated assuming the features follow Gaussian class conditional densities to better visualize their distribution. The distributions of some facial expressions, *e.g.*, neutral, happiness, surprise, and fear, are consistent with Russell's circumplex model in Figure 6 (b). Sadness, anger, and disgust, although they are close to each other, they do not follow the exact relations depicted in Russell's circumplex model. The features of the facial expression, "other", are scattered over the plane, which indicates the learned features are not discriminative enough for the "other" category. This also explains the classification performance of the "other" expression. Treating each facial expression as a binary classification task, we notice that the F1 score of the "other" expression is only 0.32, while the F1 score of the "neutral" expression is 0.68.

Figures 6 (c) and (d) visualize the features of the message space with respect to valence/arousal scores. Lighter color indicates higher value. To find the the direction in which valence/arousal increases most quickly, we fitted a bi-variate linear model to the 2-dimensional features reduced by t-SNE. We show plot gradient of this linear model with an arrow. The length of the arrow indicates the gradient magnitude. Since Figures 6 (c) and (d) are generated from the same data, the directions of their arrows can be directly compared. Although Russell's circumplex model assumes that the arousal and valence dimensions are orthogonal, our feature visualization shows that they are not orthogonal, but correlated. This is consistent with the symmetric V-shaped relation between valence and arousal, which has been found in multiple cultures, such as Canada, Korea and Spain [17].

8. Conclusion

In this paper, we proposed a novel model for multiple emotion descriptors recognition and feature learning. We designed distinct architectures for learning the feature space of sign vehicles (*e.g.*, facial action units) and learning the feature space of emotional messages (*e.g.*, facial expressions, valence, and arousal), exploiting knowledge about

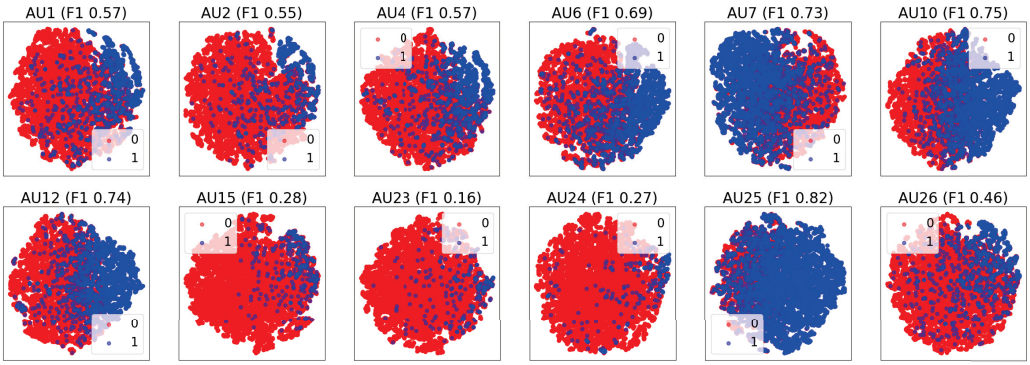
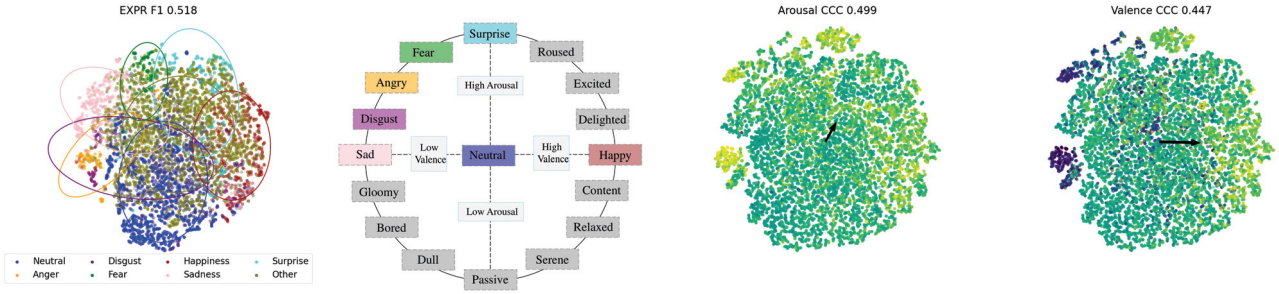


Figure 5. The t-SNE visualization of features on the sign vehicle space. The presence (1) and absence (0) of each AU are indicated by different colors.



(a) (b) (c) (d)

Figure 6. The t-SNE visualization of features on the message space. (a) The eight facial expressions. For each facial expression, we draw the ellipse of their co-variance matrix. (b) Russell's circumplex model with colors indicating different facial expressions. (c) Arousal. Lighter color indicates higher arousal. (d) Valence. Lighter color indicates higher valence. Arrow indicates the ascent direction of a linear fit. Different colors indicate the values of ground truth labels, not predictions.

their properties gleaned from related psychological studies. Our ablation study supports the advantage of sharing the feature space for facial expressions, valence and arousal, which corresponds to Russell's circumplex model.

Acknowledgements

This work is supported in part by the Hong Kong Innovation and Technology Fund via Project No. ITS/210/20. We would like to thank the Turing AI Computing Cloud (TACC) [23] and the HKUST iSING Lab for providing computational resources on their platform.

References

[1] Jeffrey F Cohn, Zara Ambadar, and Paul Ekman. Observer-based measurement of facial expression with the facial action coding system. *The handbook of emotion elicitation and assessment*, 1(3):203–221, 2007. 1, 2

[2] Didan Deng, Zhaokang Chen, and Bertram E Shi. Multitask emotion recognition with incomplete labels. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 592–599. IEEE, 2020. 2

[3] Didan Deng, Liang Wu, and Bertram E Shi. Iterative distillation for better uncertainty estimates in multitask emotion recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3557–3566, 2021. 2, 6

[4] Didan Deng, Yuqian Zhou, Jimin Pi, and Bertram E Shi. Multimodal utterance-level affect analysis using visual, audio and text features. *arXiv preprint arXiv:1805.00625*, 2018. 6

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5

[6] Wallace V Friesen, Paul Ekman, et al. Emfacs-7: Emotional facial action coding system. *Unpublished manuscript, University of California at San Francisco*, 2(36):1, 1983. 5

[7] Geethu Miriam Jacob and Bjorn Stenger. Facial action unit detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7680–7689, 2021. 2

[8] Euiseok Jeong, Geesung Oh, and Sejoon Lim. Multitask emotion recognition model with knowledge distillation and task discriminator. *arXiv preprint arXiv:2203.13072*, 2022. 6

[9] Dimitrios Kollias. Abaw: Valence-arousal estimation, expression recognition, action unit detection & multi-task learning challenges. *arXiv preprint arXiv:2202.10659*, 2022. 1

[10] D Kollias, A Schulc, E Hajiyev, and S Zafeiriou. Analysing affective behavior in the first abaw 2020 competition. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)(FG)*, pages 794–800, 2020. 2

[11] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Face behavior a la carte: Expressions, affect and action units in a single network. *arXiv preprint arXiv:1910.11111*, 2019. 2

[12] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Distribution matching for heterogeneous multitask learning: a large-scale face study. *arXiv preprint arXiv:2105.03790*, 2021. 2

[13] Dimitrios Kollias, Panagiotis Tzirakis, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, Björn Schuller, Irene Kotsia, and Stefanos Zafeiriou. Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. *International Journal of Computer Vision*, pages 1–23, 2019. 2

[14] Dimitrios Kollias and Stefanos Zafeiriou. Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface. *arXiv preprint arXiv:1910.04855*, 2019. 2, 4

[15] Dimitrios Kollias and Stefanos Zafeiriou. Affect analysis in-the-wild: Valence-arousal, expressions, action units and a unified framework. *arXiv preprint arXiv:2103.15792*, 2021. 2

[16] Dimitrios Kollias and Stefanos Zafeiriou. Analysing affective behavior in the second abaw2 competition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3652–3660, 2021. 5

[17] Peter Kuppens, Francis Tuerlinckx, Michelle Yik, Peter Koval, Joachim Coosemans, Kevin J Zeng, and James A Russell. The relation between valence and arousal in subjective experience varies with personality and culture. *Journal of personality*, 85(4):530–542, 2017. 7

[18] Shigeo Morishima and Hiroshi Harashima. Emotion space for analysis and synthesis of facial expression. In *Proceedings of 1993 2nd IEEE International Workshop on Robot and Human Communication*, pages 188–193. IEEE, 1993. 5

[19] James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980. 3

[20] Andrey V Savchenko. Facial expression and attributes recognition based on multi-task learning of lightweight neural networks. In *2021 IEEE 19th International Symposium on Intelligent Systems and Informatics (SISY)*, pages 119–124. IEEE, 2021. 6

[21] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 7

[22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[23] Kaiqiang Xu, Xinchen Wan, Hao Wang, Zhenghang Ren, Xudong Liao, Decang Sun, Chaoliang Zeng, and Kai Chen. Tacc: A full-stack cloud computing infrastructure for machine learning tasks, 2021. 8

[24] Stefanos Zafeiriou, Dimitrios Kollias, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, and Irene Kotsia. Aff-wild: Valence and arousal 'in-the-wild'challenge. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*, pages 1980–1987. IEEE, 2017. 2

[25] Wei Zhang, Zunhu Guo, Keyu Chen, Lincheng Li, Zhimeng Zhang, and Yu Ding. Prior aided streaming network for multi-task affective recognitionat the 2nd abaw2 competition. *arXiv preprint arXiv:2107.03708*, 2021. 2

[26] Wei Zhang, Zhimeng Zhang, Feng Qiu, Suzhen Wang, Bowen Ma, Hao Zeng, Rudong An, and Yu Ding. Transformer-based multimodal information fusion for facial expression analysis. *arXiv preprint arXiv:2203.12367*, 2022. 6