IEICE TRANSACTIONS

on Information and Systems

DOI:10.1587/transinf.2024EDL8020

Publicized:2024/08/08

A PUBLICATION OF THE INFORMATION AND SYSTEMS SOCIETY

The Institute of Electronics, Information and Communication Engineers

Kikai-Shinko-Kaikan Bldg., 5-8, Shibakoen 3 chome, Minato-ku, TOKYO, 105-0011 JAPAN

LETTER

Multi-dimensional and Multi-task Facial Expression Recognition for Academic Outcomes Prediction

Yi Huo†, ***and*** **Yun Ge**††

SUMMARY Recent studies on facial expression recognition mainly employs discrete category labels to represent emotion states. However, current intelligent emotion interaction systems require more diverse and precise emotion representation metrics, which has been proposed as Valence, Arousal, Dominance (VAD) multi-dimensional continuous emotion parameters. But there are still very less datasets and methods for VAD analysis, making it difficult to meet the needs of large-scale and high-precision emotion cognition. In this letter, we build multi-dimensional facial expression recognition method by using multi-task learning to improve recognition performance through exploiting the consistency between dimensional and categorial emotions. The evaluation results show that the multi-task learning approach improves the prediction accuracy for VAD multi-dimensional emotion. Furthermore, it applies the method to academic outcomes prediction which verifies that introducing the VAD multi-dimensional and multi-task facial expression recognition is effective in predicting academic outcomes. The VAD recognition code is publicly available on github.com/YeeHoran/Multi-task-Emotion-Recognition .

key words: *Multi-dimensional emotion recognition, multi-task learning, VAD facial expression recognition, intelligent education.*

1. Introduction

It investigates multi-dimensional facial expression recognition to provide a more comprehensive and accurate emotion recognition tool. Traditional emotion annotation primarily involves emotion category labels, such as seven standard emotion categories: anger, contempt, fear, joy, sadness, surprise, and neutral [1]. Alternatively, a set of emotion category labels is designed for a specific application [2]. However, with the increasing demand for emotion intelligence, obtaining more comprehensive, refined, and accurate emotion states has become increasingly urgent. Therefore, this study proposes a multi-dimensional emotion recognition method. Furthermore, it utilizes the relations between the annotation information to conduct multi-task learning to further improve recognition performance. To show its effectiveness in practical scenarios, it introduces VAD emotions in academic outcomes prediction and verifies its advantages.

With increasing demand for comprehensive emotional representation models, multi-dimensional emotion space models have been recognized [3]. AffectNet[4] pioneered the establishment of facial expression recognition datasets annotated with (valence arousal VA) labels, and found that traditional emotion categories can be observed as points in the VA two-dimensional space. Based on this, VA emotion recognition methods have been developed.

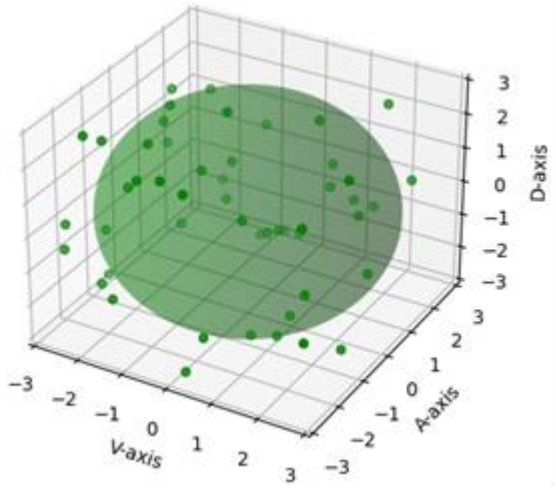

Fig.1 Discrete emotion categories are 3D coordinate points in the VAD three-dimensional space

However, methods that recognize D-dimensional information have not been supplemented, which is the first problem to be addressed in this study. Additionally, AffectNet's discovery that categories are points in VA two-dimensional space [4] inspired this research, suggesting categories are as well the points in VAD (Valence, Arousal, Dominance) three-dimensional space. Based on this hypothesis, it utilizes the correlation between discrete emotion categories and VAD multi-dimensional emotion parameters to establish consistency constraints for training models, thereby improving the accuracy of VAD and category emotion recognition.

Multi-task learning is a machine learning paradigm in which multiple tasks are learned simultaneously, often with the expectation that learning tasks jointly can improve generalization performance compared with learning them independently. There are several main research orientations and approaches for multi-task learning, e.g., shared representation learning [5][6], transfer learning [7], task relationship modeling [8][9].

This research falls into the third category modeling the relationships between multiple tasks. However, designing

† The author is with Educational Information Technology Department, Beijing Union University, Beijing, China.
†† The author is with Department of Computer Teaching and Research, University of Chinese Academy of Social Sciences, Beijing, China,

and maintaining a trainable model for multiple tasks is challenging because changes in the training data, loss function, or hyperparameters of one task can affect other tasks. Integrating different architectures into a single model is challenging. Therefore, this study aims to identify the relationships between multiple tasks and incorporate them into learning process to enhance the capabilities of all tasks.

Thus, this study proposes a multi-dimensional multi-task facial expression recognition method. It builds a VAD three-dimensional regression model to predict emotion states in VAD three-dimension space and builds a joint multi-task training model according to the correlations between emotion categories and VAD three-dimensional. At last, it presents its effectiveness in smart education scenarios by introducing VAD emotion parameters for academic achievement predictions.

2. Methods

2.1 Relationships between VAD and categories

Inspired by AffectNet's perspective [4], this study further incorporates the dominance dimension (D) into emotion representation model and proposes that the emotion categories are positioned within the three-dimensional VAD (Valence-Arousal-Dominance) emotional space, as shown in Fig.1.

The VAD three-dimensional emotional space contains green points representing various emotion categories defined manually. These are essentially discrete points within the three-dimensional emotional space. Based on this hypothesis, this paper establishes consistency loss for the training model by leveraging the correlation between discrete emotion categories and VAD multi-dimensional emotional parameters, aiming to improve the performance of emotion recognition model.

However, datasets containing simultaneous VAD (Valence-Arousal-Dominance) facial emotion samples are currently scarce. Therefore, this study began by building a dataset of VAD 3D facial emotion samples to provide foundational support for VAD 3D emotion recognition [8], which has already added VAD three-dimensional parameters to FER2013, thereby making it a dataset that simultaneously possesses annotations for seven standard emotion categories and VAD three-dimensional emotion annotations. This study utilized this newly created dataset for experiments.

Table1 presents the positions of seven standard emotions in Valence-Arousal-Dominance (VAD) space. It can be observed that for Happiness, $V > 0$, $A > 0$, and $D > 0$ are reasonable; for Disgust, $V < 0$, $A > 0$, $D > 0$ are justifiable; for Anger, $V < 0$, $A > 0$ are acceptable, but no matter $D > 0$ (due to personal reasons) or $D < 0$ (due to external events) are all fair; for Fear, $V < 0$, $A < 0$, and $D < 0$ are acceptable; for Sadness, $V < 0$, $A < 0$ are appropriate, and $D > 0$ (due to personal reasons) or $D < 0$ (due to external influences) are coherent; for Surprise, $A>0$, $D < 0$ are right, but V may be > 0 (if pleasantly surprised) or < 0 (if startled). Finally, for Neutral, all the V, A and D are equal to 0.

According to Table1, it produces the consistency loss definition in accordance with the relationships between emotion categories and VAD parameters in Table 2. To simplify the model, it only assigns 0 or 1 to consistency value, i.e., if the polarity VAD are consistent with categories in Table 1, it is set to 0, otherwise it is 1.

Table 1 VAD 3D Emotion Values for 7 Standard Emotions

	Valence	Arousal	Dominance
Happy	>0	>0	>0
Disgust	<0	>0	>0
Angry	<0	>0	>0 or <0
Fear	<0	<0	<0
Sad	<0	<0	>0 or <0
Surprise	>0 or <0	>0	<0
Neutral	=0	=0	=0

Table 2 Consistency Loss Definition

	Valence		Arousal		Dominance	
	>0	<0	>0	<0	>0	<0
Happy	0	1	0	1	0	1
Disgust	1	0	0	1	0	1
Angry	1	0	0	1	0	0
Fear	1	0	1	0	1	0
Sad	1	0	1	0	0	0
Surprise	0	0	0	1	1	0
Neutral	1	1	1	1	1	1

2.2 The Framework of Multi-Task Emotion Recognition

A model is established by integrating Valence-Arousal-Dominance (VAD) continuous emotion analysis and discrete emotion classification. They share a dataset of facial images that are annotated by the project [8] with VAD values and emotion categories. The input is facial images and is fed into two recognition tasks, i.e., 7 Basic Emotion Classification VAD Emotion Regression, separately. Each one generates its own supervision loss, representing the loss for the emotion category and regression analysis.

To conduct joint emotion recognition, it expresses the relationship between the two tasks through consistency loss, which represents the logical or physical relationship between the two emotion representation models. As shown in Table 2, the consistency loss between them is utilized to express the logical relationship between VAD emotion information and emotion category.

Thus, the overall loss function includes emotion

classification loss, emotion regression loss, and consistency loss. The model is trained to minimize the sum of these three losses. For the network backbone, both of emotion category classification and VAD regression are built on ResNet-18. To further improve recognition accuracies, their convolution layers are orthogonalized to extract more diverse features. The loss function of VAD emotion regression is MSE (Mean Squared Error), while the loss function for emotion classification is cross-entropy. The formal definition of the model is provided in the following sections.

2.3 Formulating Multi-Task Emotion Recognition

First, given sample *i*, the supervised losses defined for each individual task are denoted as $L^i_{classify}$, L^i_{VReg}, L^i_{AReg}, and L^i_{DReg} Then, the total classification and VAD regression losses from each independent task are defined in (1):

$$L^i_{class-VAD}(\theta) = L^i_{classify}(\theta) + L^i_{VReg}(\theta) + L^i_{AReg}(\theta) + L^i_{DReg}(\theta) \quad (1)$$

Where Θ is the model parameters to be learned. Then, the consistency losses between classification and *VAD* parameters are defined as $L^i_{Vconsist}$, $L^i_{Aconsist}$, and $L^i_{Dconsist}$ are formulated as (2):

$$L^i_{consist}(\theta) = L^i_{Vconsist}(\theta) + L^i_{Aconsist}(\theta) + L^i_{Dconsist}(\theta) \quad (2)$$

Where $L^i_{Vconsist}$, $L^i_{Aconsist}$, and $L^i_{Dconsist}$ are defined in Table 2. Then, the joint learning of classification, *VAD* regression and consistency are optimized in (3):

$$\min_{\theta} \sum_{i=0}^{n} \left[L^i_{class-VAD}(\theta) + \lambda L^i_{consist}(\theta) \right] \quad (3)$$

Where θ is learned by minimizing (3), λ is used to adjust proportional distribution between the loss of normal emotion classification combined with VAD emotion prediction, and their correlation constraint loss.The loss function for VAD regression is MSE, as defined in (4) to (6), The loss function for classification is cross entropy shown in (7), and the consistency loss is defined as Table 2.

$$L_{VReg} = \frac{1}{n}\sum_{i=0}^{n}(V_i - \hat{V}_i)^2 \quad (4)$$

$$L_{AReg} = \frac{1}{n}\sum_{i=0}^{n}(A_i - \hat{A}_i)^2 \quad (5)$$

$$L_{DReg} = \frac{1}{n}\sum_{i=0}^{n}(D_i - \hat{D}_i)^2 \quad (6)$$

Where n is the number of samples, V_i, A_i *and* D_i are actual values, and $\hat{V}_i$, $\hat{A}_i$ *and* $\hat{D}_i$ are predicted values.

$$L_{classify} = -\sum_{i=0}^{n}(Real_i \times \log(Output_i)) \quad (7)$$

Where $Real_i$ represents the true emotion category distribution for the i^{th} image, and $Output_i$ represents the emotional category distribution generated by the emotion classification network.

2.4 Emotion based Academic Outcomes Prediction

This study conducted academic performance prediction (final examination grades) based solely on learning data and using both learning and VAD emotions simultaneously. It uses ADA_RF_EXP as the prediction model.

The experimental data includes learning process data and emotional state data. The former comes from actual classroom teaching, including predictive test results, which are the GPA scores of all students before joining the classroom, and the test scores of 12 classes, which are defined as learning behavior data. The latter comes from facial expression images of all students and is labeled as a VAD emotional state. It performs two experiments with the first using learning data only and the second using both learning and VAD emotion data, which will be elaborated in detail in the following section.

3. Evaluation

3.1 Multi-task Facial Expression Recognition

It evaluates the prediction of V, A, D. The dataset is based on FER2013, for which the VAD parameters are manually annotated. It performs 40 epochs with a batch-size of 16. The ablation tests were conducted between using consistency (the multi-task learning proposed in this study) and not using consistency. The evaluation results are shown in Table 3.

Table 3. The Average Loss for Predicting VAD

	V	A	D	Classify
use consistency	**0.000**	**0.0940657**	**0.107480**	64.252%
not use consistency	0.000	0.0940676	0.107483	**65.561**%

From Table 3, it is obvious that for predicting V, A and D, the accuracies by using consistency constraints are higher than which without it because for A and D, the average losses by using consistency are less than without it; for V, both of their average losses are 0. Thus, it could conclude that using consistency is effective in promoting VAD prediction accuracies. However, for emotion classification, the accuracy of using consistency is less than not using it. Nevertheless, the confusion matrix is produced in Fig.2.

From Fig2, it shows that the best recognition categories are “Happy” and “Surprise” with 0.85 and 0.80 respectively; the least accurate categories are “Disgust” and “Fear” with 0.11 and 0.32 respectively; “Angry”, “Sad”, and “Neutral” performs from 0.58 to 0.62 in accuracies which are acceptable.

In addition, it finds an interesting phenomenon that the most confused emotions are “Disgust” to “Angry” reaching 0.68, and the least confused are “Angry” to “Disgust” which is 0.00. This shows that “Disgust” is usually predicted as “Angry” but “Angry” isn’t recognized at “Disgust” at all within the dataset. It also indicates that “Disgust” is difficult to be detected which also exists in common, i.e., a person's

emotion of disgust is very subtle, difficult to identify, and easily mistaken for anger, whereas anger is usually an obvious and perceivable emotion.

3.2 Academic Outcomes Prediction: Ablation Test

It performs ablation test for predicting the final examination grades as introduced in Section 2. The mean square error (MSE) of learning data only is lower than that using both learning and VAD emotion data, as shown in Fig. 3. Similarly, it illustrates that the standard deviation (std error) for learning only data is lower than which using both learning and VAD emotions as well.

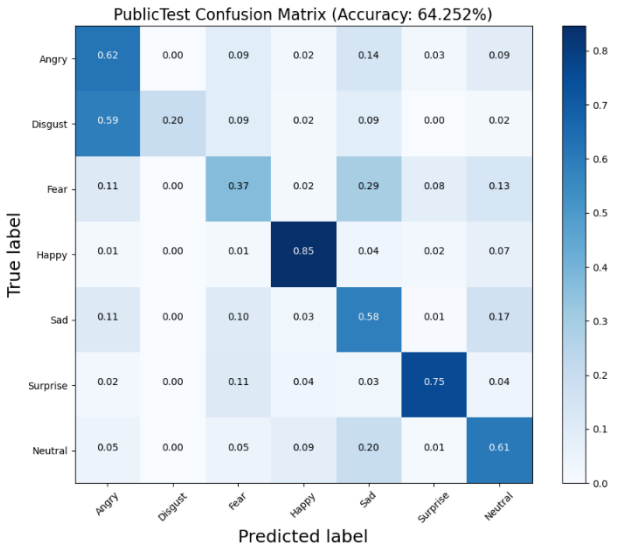

Fig. 2 The Confusion Matrix of Emotion Classification for Evaluation

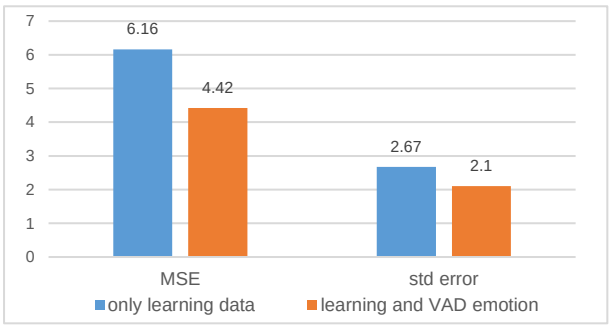

Fig**. 3 Predicted Academic Examination Grades**

Therefore, it verifies that introducing VAD emotion data is effective in advancing academic performance prediction. Thus, it could be a useful model in accessing students' learning performance in future study and applications.

4. Conclusion

In this research, the Valence, Arousal, Dominance (VAD) three-dimensional facial expression recognition method obtains a more complete and detailed interpretation than traditional emotion classification or Valence, Arousal (VA) recognition. It then shows that emotion categories are the points in VAD three-dimensional space and whose characteristic is exploited as the relation conditions in a multi-task learning model that combines VAD regression and emotion classification. A comprehensive experiment and results verified its effectiveness. However, it should be noted that combining multiple regression networks with a classification network in one cannot improve the performance compared to regressing them individually with classification without additional new constraints.

For its application, since facial expression is a very genuine and accurate parameter to display feelings, and is also convenient to obtain, it applies VAD facial recognition method in academic outcomes prediction at an early stage to find failure risk students in smart education and verifies its effectiveness.

Future research will be performed on a more extensive VAD dataset to provide a more accurate baseline for VAD emotion regression with a larger range of values quantifying the consistency cost. Another research will be conducted to explore more relations among VAD and category annotations, or to transform the structures of this multi-task network to improve the performance continuously.

Acknowledgments

It is funded by China Ministry of Education, Humanities, And Social Science Research Projects(23YJE880001); Beijing Higher Education Association Project 2023 General Project (MS2023138).

References

[1] Pekrun Reinhard. Progress and open problems in educational emotion research [J]. Learning & Instruction, 2005, 15(5): 497-506.
[2] Zhao Shuyuan. Study on the Academic Emotions of College Students Based on the Control-Value Theory [D]. Central South University, 2013.
[3] Mollahosseini A, Hasani B, Mahoor M H. AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild[J]. IEEE Transactions on Affective Computing, 1949:1-1.
[4] A. Mollahosseini, B. Hasani and M. H. Mahoor, "AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild," in IEEE Transactions on Affective Computing, vol. 10, no. 1, pp. 18-31,1 Jan.-March 2019.
[5] Bachmann, Roman, et al. "MultiMAE: Multi-modal multi-task masked autoencoders." European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022.
[6] M. Pirvu, A. Marcu, A. Dobrescu, N. Belbachir and M. Leordeanu, "Multi-Task Hypergraphs for Semi-supervised Learning using Earth Observations," in 2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Paris, France, 2023.
[7] Kollias, Dimitrios, Viktoriia Sharmanska, and Stefanos Zafeiriou. "Distribution Matching for Multi-Task Learning of Clas-sification Tasks: a Large-Scale Study on Faces & Beyond.". AAAI 2024.
[8] Y. Huo, Y. Ge., "VAD-Net: Multidimensional Emotion Recognition from Facial Expression Images", IJCNN2024.
[9] Kollias, Dimitrios, Viktoriia Sharmanska, and Stefanos Zafeiriou. "Distribution Matching for Multi-Task Learning of Classification Tasks: a Large-Scale Study on Faces & Beyond." . AAAI 2024.
[10] Goodfellow et al. Challenges in Representation Learning: A report on three machine learning contests. 1 Jul 2013, ICML 2013 Workshop.