EMMA:EMotion Mixing Algorithm for compound expression recognition using angle-based metric learning

Riku Yamamoto[1 [0009−0001−9468−7282]] and Noriko Takemura[1 [0000−0003−1977−4690]]

Kyushu Institute of Technology, 680-4 Kawazu, Iizuka, 820-8502, Japan
takemura@ai.kyutech.ac.jp

Abstract. Facial expression recognition (FER) is a key component in various AI-based systems and has been extensively studied. However, most FER research has focused on clear and simple basic emotions such as happiness and sadness, which are not suitable for real-world applications where numerous many ambiguous and complex emotions exist. Complex emotions are challenging to define and require ample data for each emotion to train a FER model. Moreover, due to their ambiguous nature, these emotions are difficult to annotate. Consequently, the difficulty in constructing comprehensive databases is a significant bottleneck in recognizing complex emotions. In this study, we propose complex emotion recognition method using only a database of basic emotions based through angle-base metric learning. This approach can mitigate the reduction in recognition accuracy caused by insufficient data and allows for the definition of new emotions in the future, unlike general FER tasks that require pre-definition of emotions.

Keywords: facial expression recognition · compound emotion · metric learning

1 Introduction

The development of facial expression recognition (FER) technology has enabled human-centric intelligent systems to provide adaptive services and support based on individual emotions and needs[19]. Specifically, it affects areas such as security[18][13], learning[17][21], and medical rehabilitation[25]. Facial expressions are a prominent expression of human emotions, which in fact, it is said that about 55% of human emotional communication is conveyed through facial expressions[2]. However, most previous studies have focused on Ekman's basic emotions[7], which are happiness, sadness, anger, surprise, fear, and disgust, and these emotions alone are insufficient to fully capture the diversity and complexity of human emotions. This is because human emotions go beyond these basic emotions and encompass more subtle and complex emotions that are important in real-world systems.

In recent years, FER has been dominated by methods based on convolutional neural networks (CNNs), which are type of neural networks, but while a CNN can achieve high recognition accuracy, it requires a large amount of data during training. It is possible to intentionally create facial expression data for basic emotions through acting; however, it is difficult to artificially create complex facial expressions because they are naturally expressed in daily life. In addition, complex emotions are challenging to define, and sufficient amount of data is required for each of the various emotions to be accurately recognized. Furthermore, complex emotions are often ambiguous, and it is not easy to annotate them correctly. The difficulty of constructing such a dataset is a bottleneck in complex emotion recognition.

By contrast, in the field of psychology, there is the idea that a wide variety of emotions can be expressed as compound emotions, which are combinations of basic emotions. Du et al. demonstrated that the Action Unit (AU) patterns of compound emotions reflect the Action Unit patterns of the basic emotions that constitute them[6][5]. AUs are facial muscle movements associated with specific emotions, serving as the basis for identifying these emotions[8]. This finding underscores the importance of basic emotions as fundamental building blocks in understanding more complex emotional states. Basic emotions are emotional categories that people generally tend to share. Explaining complex emotions on this basis greatly improves interpretability and allows for more detailed representations of subtle differences in human emotions. As an example, the complex emotion of nostalgia is expressed in terms of happy memories of the past (happiness) and the reality that these memories are in the past (sadness). In this way, it is possible to subdivide various emotions and analyze them in more detail.

In this study, we focus on the idea that complex emotions are expressed by combining basic emotions, and propose EMMA (EMotion Mixing Algorithm) for compound emotion recognition (CER) method using only the features related to basic emotions. In other words, we construct a CER method using only basic emotion data without considering compound emotion data. This method also solves the problem of constructing complex emotion datasets.

This method uses a CNN-based model and basic emotion data to train a feature extractor for basic emotions and a representative vector for the features of each basic emotion. In this training, we perform angle-based metric learning instead of the euclidean distance commonly used in general classification tasks. Angle-based metric learning does not consider the magnitude of vectors, thus preventing overfitting to factors such as the orientation of faces, occlusion, and gender in the training data. Therefore, it allows for the extraction of features that are more effectively separated by emotion. The compound emotion estimation is based on the compound emotion similarity calculated by linearly combining the basic emotion features of the image. In compound emotion estimation, each of compound emotion similarity is calculated as an unweighted average of the two basic emotion similarities that compose that.

The contributions of this study are as follows.

CER using only basic emotion data
This method does not require a dataset of compound emotions because it estimates compound emotions using only the features of the basic emotions. Conventional methods using compound emotion data have the problem that the amount of data for each emotion is not sufficient, and the bias in the amount of training data for each emotion significantly affects the accuracy of estimation.

An emotional space based on angles
Human facial images are diverse and include various elements. By introducing angle-based metric learning, it is expected to reduce the influence of elements other than expression information. Thus, it can lead to more isolated the features of basic emotion and improve the identification performance of compound emotions represented as linear combinations of basic emotions.

The ability to estimate a wide variety of emotions
Because the proposed method defines composite emotions as linear combinations of basic emotions, it is possible to freely set the composite emotions to be estimated. The estimated complex emotion need not be a human-definable emotion such as nostalgia or respect, as long as the proportion of the basic emotions that comprise the complex emotion is known. In other words, this method should be able to quantitatively express emotions that cannot be defined by humans, and hence advance our understanding of complex emotions in the field of psychology.

2 Related work

2.1 Basic emotion recognition (BER)

The six emotions of surprise, fear, disgust, happiness, sadness, and anger are called Ekman's basic emotions and are the basis of current FER. As the name suggests, they are the basic elements of human emotional expression and are common to different cultures and societies around the world, and hence recognizing them plays an important role in understanding human emotions. Therefore, BER has been the subject of active research.

Among the various BER techniques, deep learning, which has contributed greatly to the field of image recognition in recent years, has also had a significant impact on the field of FER, improving recognition accuracy. Wang et al. proposed the Region Attention Network (RAN) as a model for FER that is robust to occlusion and pose changes in facial images[23]. RAN aggregates various numbers of domain features generated by the backbone CNN and embeds them in a compact fixed-length representation. The Distract your Attention Network (DAN) proposed by Zhengyao et al. extracts higher-order features of facial expressions from various regions of an expression image by convolution, and encodes these interactions to achieve a comprehensive understanding of facial expressions.

2.2 Complex facial emotion recognition

As described above, deep learning has made a significant contribution to FER, but this method requires a large amount of annotated data for training. However, there is a lack of data for complex facial expressions, which poses a serious problem. This is due to the fact that complex emotions can be interpreted differently by different people, making it impossible to maintain consistent indices, and the annotation process becomes time-consuming as the number of labels increases[1]. Therefore, it is necessary to devise a way to achieve performance even with a small amount of data.

Through a multi-task learning approach that simultaneously performs Action Unit (AU) detection tasks, C-EXPR-NET[10] can efficiently learn from a small amount of complex facial expression data by exploiting the interaction of information obtained from each task. Because the middle layer of a CNN can be used as a feature extractor, there are also methods that pre-train on another task [4]and estimate complex emotions from the features. DLP-CNN[12] uses a model pre-trained with basic emotions as a feature extractor and classifies the features with a support vector machine to estimate compound emotions. Although this method uses features of the basic emotion to classify the compound emotion, the feature extractor for the basic emotions has not been designed with compound emotion estimation in mind.

2.3 Position of this study

As mentioned in the previous section, conventional studies have improved the accuracy of complex FER by devising models. However, these methods require labeled data, which is a major barrier to their incorporation into actual systems, because new data must be prepared each time the emotion to be estimated changes. In this study, we propose a method for inference without using complex emotion data. This method significantly reduces the cost and time associated with data collection and labeling, allowing machine learning systems to be applied to real-world problem solving more quickly and efficiently.

3 Proposed method

In this study, we propose EMotion Mixing Algorithm (EMMA) for CER that does not require data labeled with compound emotions. The specific procedure is as follows.

1) Pre-training on basic emotions

Through angle-based metric learning for basic emotions (six classes), we train a feature extractor that extracts features related to basic emotions from face images, and a representative vector for each basic emotion (basic EmoVec) in the feature space.

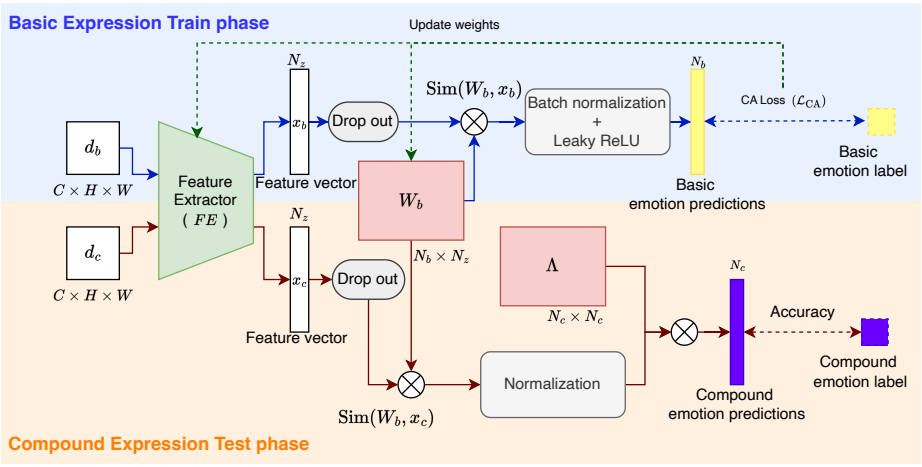


Fig. 1: **Outline of the proposed method.** The upper part shows the pre-training process for the basic emotions, and the lower part shows the estimation process of the compound emotions.In basic emotion train phase, Feature extractor and matrix of basic EmoVecs (W_b) are trained using the $\mathcal{L}_{\mathrm{CA}}$ and basic emotion datasets; in compound emotion test phase,the similarity of the basic emotions is combined in the same ratio to define the compound emotion intensity and estimate it. Λ is the coupling coefficient of basic EmoVecs.

2) Compound emotion recognition (CER)
Using the Feature extractor and basic EmoVec trained in Step 1, we calculate similarity of each basic emotion in the input images. By calculating the average of these similarities, we obtain the similarity of compound emotion. The highest similarity is taken as the predicted value.

This method only performs pre-training for the basic emotions, it does not perform any new training for the compound emotions. Therefore, it is possible to estimate compound emotions using only the features of the basic emotions, without using compound-emotion data. The outline of the method is shown in Fig 1, and the following sections describe each step in detail.

3.1 Pre-training on basic emotions

Feature extractors and basic EmoVecs are obtained through pre-training on basic emotions. For a single emotion, the feature vectors output by the feature extractor are distributed in close proximity to each other in the feature space, and a basic EmoVec determines the center of each emotion distribution. Therefore, to ensure that each emotion is clearly distinguished, the distance between basic EmoVecs should be large. To obtain such a feature extractor and basic EmoVecs, this study implements as angle-based metric learning, rather than pre-training as a general classification task using cross entropy loss ($\mathcal{L}_{CE}$).

In general, a model trained for a classification task can only classify the trained class, but the metric learning method should increase the separability between known and unknown emotions because it trains the transformations to the feature values. Angle-based metric learning is used in several methods proposed for the face recognition task, including sphereFace[14], arcFace[3], and cosFace[22]. This method is effective in terms of the scale invariance of the feature vectors for face images, which have many intra-class variations, such as lighting conditions and face angle. Specifically, unlike Euclidean distance, it does not consider the magnitude of the vectors, but focuses only on the direction, thus mitigating the increase in within-class variance due to these variations. Because facial expression images also have the characteristic of large intra-class variations, we expect that angle-based metric learning will be effective.

As a specific learning procedure, first, basic expression data $d_b \in \mathbb{Z}^{C\times H\times W}$ are input to the feature extractor $FE(\cdot)$ to obtain a feature vector $x_b \in N_z$.

$$x_b = FE(d_b) \tag{1}$$

Where, C, H, and W are the number of channels, height, and width of the image, respectively, and FE is the transform of the feature extractor. In addition, N_z is the dimensionality of the feature vectors. Dropout is then applied to the feature vectors, and W_b is created by normalization in each row of the weight matrix $W_b' \in \mathbb{R}^{N_b\times N_z}$. Then, N_b is the number of basic emotions, which in this study is six (surprise, fear, disgust, happiness, sadness, and anger). Using this feature vector x_b and the normalized weight matrix W_b, the loss function $\mathcal{L}_{\mathrm{CA}}$ is calculated (Eq. 2).

$$\mathcal{L}_{\mathrm{CA}} = -\frac{1}{N}\frac{e^{\|x_b\| LReLU(\cos(w_{by},x_b))}}{\sum_i e^{\|x_b\| LReLU(\cos(w_{bi},x_b))}} \tag{2}$$

We design $\mathcal{L}_{\mathrm{CA}}$ to be the loss based on the modified softmax loss[14] with the constrained by Leaky ReLU, which we call constrained angle loss in this paper. In Eq. 2, N is the number of training samples, $\cos(w_{bi}, x_b)$ is the cosine similarity between feature vector x_b and the i-th row vector w_{bi} of weight matrix W_b, and $LReLU$ is the Leaky ReLU function. In this loss function, if each row vector of the weight matrix W_b is regarded as a representative vector of each basic emotion (basic EmoVec), it can be interpreted that learning occurs so that the cosine similarity is large between feature vector x_b and the correct basic EmoVec w_{by}, and small between feature vector x_b and the other basic EmoVecs. The modified softmax loss minimizes the similarity with all emotions other than the correct emotion, but the minimum of the cosine is -1, resulting in a negative correlation. If the negative correlation is too strong, the within-class variance of each emotion in the feature space becomes too small, and the diversity of the feature vectors for one emotion may be lost. Because this study uses this feature vector to estimate compound emotions, it is desirable for the feature vector to be diverse to some extent. By contrast, if excessive restrictions such as setting the minimum value of the cosine similarity to 0 are applied to eliminate negative correlations, the classification of the training data itself becomes difficult. Therefore, Leaky ReLU

is employed to moderate the negative correlation and achieve a balance between the diversity of feature vectors and separation performance. The accuracy of the Leaky ReLU parameters is evaluated in Section 4.3. The cosine similarity values are input to the batch normalization layer before being input to Leaky Relu to improve learning efficiency.

3.2 Compound emotion recognition (CER)

In compound emotion estimation, each of compound emotion similarity is calculated as an unweighted average of the two basic emotion similarities that compose that. The issue at hand is whether it is fair to combine basic emotions in a one-to-one ratio. Each emotion has a different distribution of similarity, which may prevent them from being evaluated on the same scale. Therefore, in this study, instead of using the similarity to basic emotions as is, we use the values after batch normalization layer. Batch normalization layer performs scaling by emotion similarity. This process is expected to enable fair treatment of each emotion.

Next, the specific CER process is described. First, compound expression data $d_c \in \mathbb{Z}^{C\times H\times W}$ are input to the feature extractor $FE(\cdot)$ trained in basic emotion to obtain a feature vector $x_c \in N_z$ (Eq. 3).

$$x_c = FE(d_c) \tag{3}$$

The cosine similarity of each basic emotion is calculated by taking the product of matrix of basic EmoVecs $W_b \in \mathbb{R}^{N_z\times N_b}$ obtained by pre-training and the feature vector x_c (Eq. 4).

$$Sim(W_b, x_c) = W_b^T x_c = \begin{bmatrix} w_{b0}^T x_c \\ w_{b1}^T x_c \\ \vdots \\ w_{bN_c}^T x_c \end{bmatrix} = \|x_c\| \begin{bmatrix} \cos(w_{b0}, x_c) \\ \cos(w_{b1}, x_c) \\ \vdots \\ \cos(w_{b1}, x_c) \end{bmatrix} \tag{4}$$

Where, each column of W_b are basic EmoVec and are normalized. That is, $\|w_{bi}\| = 1, (0 \le i \le N_b - 1, i \in \mathbb{Z})$. Finally, after inputting this similarity into Batch normalization layer ($BN(\cdot)$), a linear combination is performed to arrive at compound emotion. The largest compound emotion intensity created is then used as the predictive value y_c.

$$y_c = argmax(\Lambda^T BN(Sim(W_b, x_c))) \tag{5}$$

Where, $\Lambda \in \mathbb{R}^{N_b\times N_c}$ is a coupling matrix for combining the basic EmoVecs. In the case of a compound emotion that contains two basic emotions in a ratio of 1:1, the weight is 0.5 if the basic emotion constitutes the compound emotion and 0 otherwise.

4 Experiments and discussion

To verify the usefulness of the framework proposed in this study for CER using only basic emotion data and the loss function $\mathcal{L}_{CA}$ for angle-base metric learning using Leaky ReLU, we conducted evaluation experiments using existing data sets on compound emotions.

4.1 Experimental setup

Dataset In this study, we use RAF-DB[12][11] as the dataset for basic and compound emotions. This dataset consists of natural facial images obtained from the Internet and includes diverse elements in terms of subject's age, gender, ethnicity, head pose, lighting conditions, occlusion (e.g., glasses or facial hair), and post-processing (various filters and special effects). They were also labeled by 40 crowd-sourced annotators. The data includes seven classes of basic emotions (including neutral) and eleven classes of compound emotions (composed of two basic emotions). The actual types of compound emotions and the number of data are shown in Table 2 below. Neutral represents the absence of emotion and is not considered to contribute much to the estimation of the compound emotion, so neutral images are excluded. In this dataset, the training data and test data are provided, but in this study, 20% of the training data were randomly selected to act as validation data. The input images were cropped to a square around the face area 224×224 pixels in size. However, in an experiment using DDRAMFN as the feature extractor has an image size of 112×112 pixels.

Hyperparameters In this study, the batch size was 64 and the initial value of the learning rate was 0.001. Stochastic gradient descent (SGD) was used as the optimization method. In addition, early stopping was used to improve learning efficiency. Therefore, the number of training epochs varied from model to model. The patience of early stopping was set to 5, and learning was terminated when the value of the loss function at the time of validation rises five times in a row. Verification was performed at the end of each learning batch. Furthermore, negative slope of $\mathcal{L}_{CA}$ (Eq. 2), was set to 0.1.

4.2 Comparison with general classification methods

Table 1 presents the results obtained under various conditions in the CER model. These conditions include the type of backbone used as a feature extractor, the loss function, and predict method. The models used as feature extractors were ResNet50[9], EffectNet B2[20], DDAMFN[26], and DAN[24]. For the loss function, the cross entropy loss ($\mathcal{L}_{CE}$) used in general classification task models and the constrained angle loss ($\mathcal{L}_{CA}$) proposed in this study were compared and evaluated. As for the prediction method of compound emotion, we compared a general classification method, the EMMA method proposed in this study and

Table 1: **Result of CER.** General classification (CLN) trained on the compound emotion dataset;EMMA is the proposed method and uses the basic emotion estimation model and no training data for compound emotions; Transfer is pre-trained on the basic emotion and then additionally trained on the compound emotion (this is described in Sec.5).

FE	Loss	predict method					
		General CLN		EMMA(ours)		Transfer	
		Acc.[-]	UAR[-]	Acc.[-]	UAR[-]	Acc.[-]	UAR[-]
ResNet50[9]	$\mathcal{L}_{\mathrm{CE}}$	0.434	0.271	0.436	0.407	0.484	0.325
EffectNet B2[20]	$\mathcal{L}_{\mathrm{CE}}$	0.505	0.328	0.447	0.419	0.563	0.398
DAN[24]	$\mathcal{L}_{\mathrm{CE}}$	0.606	0.440	0.539	0.502	0.621	0.495
DDRAMFN[26]	$\mathcal{L}_{\mathrm{CE}}$	0.646	0.492	0.615	0.530	0.653	0.528
ResNet50	$\mathcal{L}_{\mathrm{CA}}$		-	0.429	0.391	0.487	0.329
EffectNet B2	$\mathcal{L}_{\mathrm{CA}}$		-	0.434	0.422	0.578	0.408
DAN	$\mathcal{L}_{\mathrm{CA}}$		-	0.578	0.527	0.659	0.542
DDRAMFN	$\mathcal{L}_{\mathrm{CA}}$		-	**0.617**	**0.559**	**0.659**	**0.559**

transfer learning. General classification method is trained on a compound emotion dataset and directly predicts. EMMA uses models trained on basic emotion to predict compound emotions. Transfer learning is pre-trained on basic emotions and then undergoes additional training on the compound emotion dataset. Parameters are updated only for compound EmoVec. The transfer learning method is examined in detail in Section 5.

Furthermore, accuracy and unweighted average recall (UAR) evaluation metrics were used. Accuracy is the rate of correct answers out of all test data. UAR is the average recall of each class, which means it corresponds to the average of the diagonal elements of the confusion matrix. This indicator allows us to measure whether we are reasoning in a balanced manner.

First, we compare the results of the general classification method (Table 1, red background), which uses $\mathcal{L}_{CE}$ and training data of compound emotions, and EMMA method (blue background), which uses $\mathcal{L}_{CA}$ and only the basic emotion data to construct the CER model. Despite not using a training dataset of compound emotions, EMMA achieves performance comparable to the general classification method across all models. In particular, UAR shows a significant improvement when compared to general method, which means that it can predict all emotions comprehensively. The confusion matrix of the proposed method is shown in Figure 2 and the one when trained as a general classification task is shown in Figure 3(b). Comparing these two confusion matrices, it can be seen that the proposed method has a high recognition rate without bias by class. We consider this to be a significant advantage, because it does not use composite emotion data with a disproportionate number of data.

In fact, to examine the relationship between the amount of data and recognition accuracy, the True Positive Rate (TPR) for each emotion is shown in Table 2. The numbers in parentheses in this table indicate the amount of training data

Table 2: **TPR for the EMMA and general classification method.**The numbers in parentheses indicate the number of training data. Bold number is the higher TPR for the two methods.

Label	TPR[-]	
	EMMA(ours)	General classification
Happily Surprised	**0.81**	0.78 (438)
Happily Disgusted	**0.70**	0.40 (176)
Sadly Fearful	**0.64**	0.23 (85)
Sadly Angry	**0.27**	0.15 (110)
Sadly Surprised	**0.39**	0.11 (55)
Sadly Disgusted	0.67	**0.70 (483)**
Fearfully Angry	0.58	**0.61 (98)**
Fearfully Surprised	0.58	**0.73 (356)**
Angrily Surprised	**0.37**	0.24 (116)
Angrily Disgusted	0.45	**0.72 (521)**
Disgustedly Surprised	**0.34**	0.17 (91)

for each emotion. The backbone was DAN. The results clearly show that the breakdown of the TPR is different between the EMMA and the general method. Specifically, the TPR of the general method is high for the happily surprised and sadly disgusted emotions, which have sufficient data, whereas it is low for the sadly surprised and disgustedly surprised emotions, which have insufficient data. In other words, the general method is highly dependent on data imbalance. By contrast, the proposed method does not have such a problem and can recognize the data in a relatively balanced manner.

Next, comparing the performance obtained when using $\mathcal{L}_{CE}$ (green background) and when using $\mathcal{L}_{CA}$ (blue background) in this framework, we find that the proposed loss function improves UAR, indicating the effectiveness of angle-based metric learning. The confusion matrix in Figure 3(c) uses Cross Entropy loss as the loss function. A comparison of this result with that of the proposed method (Figure 2) shows that the overall trend is the same, but the actual recognition rate is improved by the proposed method. In other words, changing the loss function from Cross Entropy loss ($\mathcal{L}_ce$) to $\mathcal{L}_caproposed$ in this study improves the overall recognition rate. In other words, changing the loss function from Cross Entropy loss ($\mathcal{L}_{ce}$) to Constrained Angle loss ($\mathcal{L}_{ca}$) proposed in this study improves the overall recognition rate.

4.3 Verification of the Leaky ReLU gradient

In this section, we examine the effectiveness of Leaky ReLU in the loss function $\mathcal{L}_{CA}$(Eq. 2) newly introduced in this study. Fig. 4 shows the recognition accuracy of compound emotions when the hyperparameter of the Leaky ReLU function, negative slope (NS), was varied, and the shape of the cosine similarity in this case is shown in Fig. 5. The backbone used DAN, which resulted in the largest changes. When NS=0, the result is consistent with the ReLU function, and

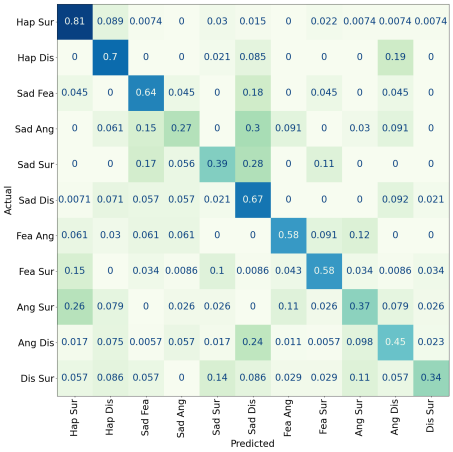
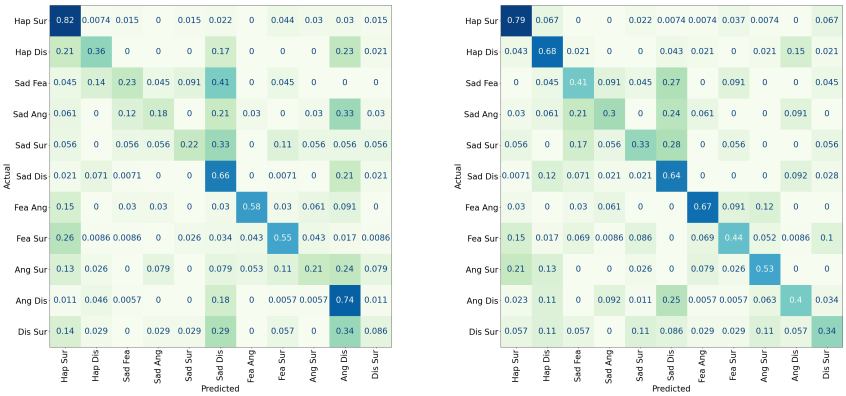


Fig. 2: **Confusion matrix when predicting compound emotions using the proposed method.** DAN is used as a feature extractor. This model trained basic emotional features using Constrained Angle loss.

when the cosine similarity is negative, the value is zero. When NS=1, the input is directly output, so nothing is introduced (the cosine similarity ranges from -1 to 1). When $0 < NS < 1$, the value is slightly larger when the cosine similarity is negative. The results of Fig. 4 reveal that the highest accuracy was obtained when $NS = 0.1$ was used to constrain the results with Leaky ReLU.

Next, we consider the change in the distribution of feature vectors due to this constraint. The distribution of feature vectors for basic emotions was visualized using t-SNE, as shown in Fig 6. This visualization was performed separately for three different groups: those employing $\mathcal{L}_{CE}$ as the loss function, those incorporating Leaky ReLU within $\mathcal{L}_{CA}$, and those not utilizing Leaky ReLU. Table 3 shows the intra-class variance for each emotion and inter-class in these three conditions. Comparing t-SNE of $\mathcal{L}_{CE}$ and $\mathcal{L}_{CA}$, we can see that $\mathcal{L}_{CA}$ has a better separation performance, especially for fear and disgust, because the distribution of $\mathcal{L}_{CA}$ is more coherent. Actually, Table 3 shows that the inter-class variance is smaller for $\mathcal{L}_{CA}$. By contrast, comparing the results with and without Leaky ReLU in $\mathcal{L}_{CA}$, we can see that the intra-class variance is larger and the feature vector diversity is higher when Leaky ReLU is applied. From these facts, the application of Leaky ReLU can improve the diversity within a class while providing separability between classes. Due to these characteristics, we consider that even a basic emotion feature extractor can capture the characteristics of compound emotions.



(a) DAN+general classification (b) DAN+$\mathcal{L}_{CE}$

Fig. 3: **Confusion matrix when predicting compound emotions.** DAN is used as a feature extractor. (a) a model trained for the general classification task, i.e., using an 11-class compound emotion dataset, (b) a model trained basic emotional features using Cross Entropy loss.

5 Compound emotion recognition using transfer learning

When compound emotion data are available, the recognition accuracy of the proposed model trained on the basic emotion data should be improved if additional training on the compound emotion data is conducted. In this section, we use transfer learning to learn about compound emotions. Specifically, the parameters of the feature extractor (FE) pre-trained with the basic emotion shown in Figure 1 are fixed and matrix of basic EmoVecs (W_b) is excluded. Instead, we introduce the matrix of the representative vectors of the compound emotion (W_c) as the training parameter. This method makes it possible to estimate compound emotions using basic emotion features. The loss function is the constrained angle

Table 3: Intra and inter-class variance of feature vector x_b transformed by t-SNE.

Label	Variance[-]		
	$\mathcal{L}_{CE}$	$\mathcal{L}_{CA}$ (w/o $LReLU$)	$\mathcal{L}_{CA}$ (w/ $LReLU$)
Surprise	639.9	603.9	618.3
Fear	481.3	214.7	468.5
Disgust	250.4	124.3	253.2
Happiness	552.7	268.5	549.3
Sadness	493.5	508.2	492.7
Anger	162.1	165.2	163.7
inter-class	433.8	327.2	353.1

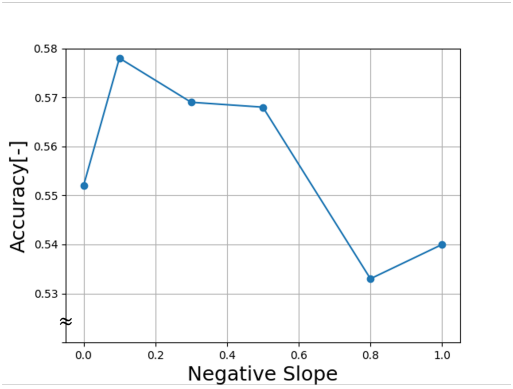
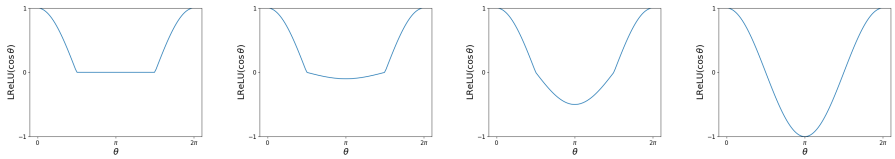


Fig. 4: Accuracy when changing *NS* in Leaky ReLU

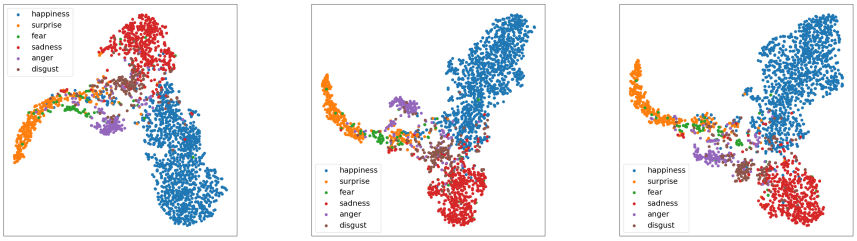


(a) *NS*=0.0 (b) *NS*=0.1 (c) *NS*=0.5 (d) *NS*=1.0

Fig. 5: **Graph of *LReLU*$(\cos\theta)$.** *NS* indicates the negative slope of Leaky ReLU. When $NS = 1$, it coincides with the general $\cos\theta$. The smaller the *NS*, the less training in the negative direction.

loss (NS=1), which is the loss function excluding Leaky ReLU from constrained angle loss (Eq. 2). The reason for excluding Leaky ReLU is that, unlike basic emotion learning, diversity in the feature vectors is not required.

Next, the verification results are presented. The same backbone used in the previous section is used for the feature extractor. The results of the method using transfer learning are shown in Table 1, which reveals that the accuracy is improved by transfer learning. However, as with the conventional method (red background), the UAR is low relative to accuracy, and the results for resNet50 are instead significantly reduced by the transfer learning. This is also due to the imbalance in the data. By contrast, the DAN+$\mathcal{L}_{CA}$ results show that the UAR is improved by transfer learning, and this is a result not seen in the case of DAN+$\mathcal{L}_{CE}$. DDRAMFN has a similarly improved UAR. Common to all UAR-improved backbones is the use of complex attention mechanisms. We therefore infer that $\mathcal{L}_{CA}$ has a high affinity with this mechanism and contributes to the improvement of UAR through synergistic effects.



(a) $\mathcal{L}_{CE}$ (b) $\mathcal{L}_{CA}$ (w/o Leaky ReLU) (c) $\mathcal{L}_{CA}$ (w/ Leaky ReLU)

Fig. 6: **t-SNE visualization of feature vector distribution.** The feature extractor is DAN.

6 Conclusion

In this study, we proposed a EMotion Mixing Algorithm (EMMA) for CER using basic emotional features. By introducing metric learning into the training of basic emotion features, it is possible to improve the performance of class separation in the feature space. The greatest advantage of this method is that it does not use compound emotion data for estimating compound emotions. It is difficult to annotate and collect data for complex emotions, including compound emotions, and large datasets are not available. The recognition accuracy of conventional methods is highly dependent on data size, making it difficult to recognize emotions that are difficult to collect. By contrast, the proposed method does not use compound emotion data at all, which avoids this problem.

In addition, after learning the features of basic emotions, we also tested a method for learning the additional features of compound emotions using transfer learning. Most CNN architectures are affected by data imbalance, but models that use complex attention mechanisms can solve this problem while improving recognition accuracy.

One of our future tasks is to examine more specifically what kind of dataset is appropriate for pre-learning basic emotions. For example, this study used a dataset consisting of natural facial expressions, but it will be necessary to verify whether the features learned using this dataset are sufficient to recognize natural complex emotions in other datasets such as those created by actors in a laboratory (e.g., CK+[15] or JAFFE[16]). We also believe that there is room for improvement in the way the representative vectors of complex emotions are defined. In the method proposed in this study, they were defined by a simple linear combination, but the relationships among basic emotions are more complex, and a new definition that takes these into account is needed.

Acknowledgments. This work was supported by JSPS Grant-in-Aid for Scientific Research (B) Grant Number JP23H03486.

References

1. Bryant, D., Deng, S., Sephus, N., Xia, W., Perona, P.: Multi-dimensional, nuanced and subjective – measuring the perception of facial expressions. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 20900–20909 (2022). https://doi.org/10.1109/CVPR52688.2022.02026
2. Darwin, C.: The Expression of Emotions in Man and Animals. Oxford University Press (1872)
3. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
4. Donahue, J., Jia, Y., Vinyals, O., Hoffman, Tzeng, E., Darrell, T.: Decaf: A deep convolutional activation feature for generic visual recognition. In: Xing, E.P., Jebara, T. (eds.) Proceedings of the 31st International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 32, pp. 647–655. PMLR, Bejing, China (22–24 Jun 2014), https://proceedings.mlr.press/v32/donahue14.html
5. Du, S., Martinez, A.M.: Compound facial expressions of emotion: from basic research to clinical applications. Dialogues in Clinical Neuroscience **17**(4), 443–455 (dec 2015). https://doi.org/10.31887/dcns.2015.17.4/sdu, http://dx.doi.org/10.31887/DCNS.2015.17.4/sdu
6. Du, S., Tao, Y., Martinez, A.M.: Compound facial expressions of emotion. Proceedings of the National Academy of Sciences **111**(15), E1454–E1462 (2014). https://doi.org/10.1073/pnas.1322355111, https://www.pnas.org/doi/abs/10.1073/pnas.1322355111
7. Ekman, P.: An argument for basic emotions. Cognition & emotion **6**(3-4), 169–200 (1992)
8. Ekman, P., Friesen, W.V.: Facial action coding system. Environmental Psychology & Nonverbal Behavior (1978)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90
10. Kollias, D.: Multi-label compound expression recognition: C-expr database & network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5589–5598 (June 2023)
11. Li, S., Deng, W.: Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. IEEE Transactions on Image Processing **28**(1), 356–370 (2019)
12. Li, S., Deng, W., Du, J.: Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2584–2593. IEEE (2017)
13. Li, Z., Zhang, T., Jing, X., Wang, Y.: Facial expression-based analysis on emotion correlations, hotspots, and potential occurrence of urban crimes. Alexandria Engineering Journal **60**(1), 1411–1420 (2021)
14. Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., Song, L.: Sphereface: Deep hypersphere embedding for face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
15. Lucey, P., Cohn, J.F., Kanade, T., Saragih, Matthews, I.: The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops. pp. 94–101 (2010). https://doi.org/10.1109/CVPRW.2010.5543262

16. Lyons: The japanese female facial expression (jaffe) dataset (sep 2019). https://doi.org/10.5281/zenodo.3451524, https://doi.org/10.5281/zenodo.3451524
17. Mannepalli, K., Sastry, P.N., Suman, M.: A novel adaptive fractional deep belief networks for speaker emotion recognition. Alexandria Engineering Journal **56**(4), 485–497 (2017). https://doi.org/https://doi.org/10.1016/j.aej.2016.09.002, https://www.sciencedirect.com/science/article/pii/S1110016816302484
18. Nan, Y., Ju, J., Hua, Q., Zhang, H., Wang, B.: A-mobilenet: An approach of facial expression recognition. Alexandria Engineering Journal **61**(6), 4435–4444 (2022)
19. Sajjad, M., Ullah, F.U.M., Ullah, M., Christodoulou, G., Alaya Cheikh, F., Hijji, M., Muhammad, K., Rodrigues, J.J.: A comprehensive survey on deep facial expression recognition: challenges, applications, and future guidelines. Alexandria Engineering Journal **68**, 817–840 (2023). https://doi.org/https://doi.org/10.1016/j.aej.2023.01.017, https://www.sciencedirect.com/science/article/pii/S1110016823000327
20. Tan, M., Le, Q.: EfficientNet: Rethinking model scaling for convolutional neural networks. In: Chaudhuri, K., Salakhutdinov, R. (eds.) Proceedings of the 36th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 97, pp. 6105–6114. PMLR (09–15 Jun 2019), https://proceedings.mlr.press/v97/tan19a.html
21. Tonguç, G., Ozaydın Ozkara, B.: Automatic recognition of student emotions from facial expressions during a lecture. Computers & Education **148**, 103797 (2020). https://doi.org/https://doi.org/10.1016/j.compedu.2019.103797, https://www.sciencedirect.com/science/article/pii/S0360131519303471
22. Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, Li, Z., Liu, W.: Cosface: Large margin cosine loss for deep face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
23. Wang, K., Peng, X., Yang, J., Meng, D., Qiao, Y.: Region attention networks for pose and occlusion robust facial expression recognition. IEEE Transactions on Image Processing **29**, 4057–4069 (2020). https://doi.org/10.1109/TIP.2019.2956143
24. Wen, Z., Lin, W., Wang, T., Xu, G.: Distract your attention: Multi-head cross attention network for facial expression recognition. Biomimetics **8**(2) (2023). https://doi.org/10.3390/biomimetics8020199, https://www.mdpi.com/2313-7673/8/2/199
25. Yun, S.S., Choi, J., Park, S.K., Bong: Social skills training for children with autism spectrum disorder using a robotic behavioral intervention system. Autism Research **10**(7), 1306–1323 (2017). https://doi.org/https://doi.org/10.1002/aur.1778, https://onlinelibrary.wiley.com/doi/abs/10.1002/aur.1778
26. Zhang, S., Zhang, Y., Zhang, Y., Wang, Y., Song, Z.: A dual-direction attention mixed feature network for facial expression recognition. Electronics **12**(17) (2023). https://doi.org/10.3390/electronics12173595, https://www.mdpi.com/2079-9292/12/17/3595