Proceedings of the 8th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/2026.35825

Research on a Fine-Granularity Text-Image Emotional Recognition Algorithm Optimized Through Cross-Modality Feature Fusion

Guanhui Wang

Department of Software Engineering, Qufu Normal University, Qufu, China
1787367218@qfnu.edu.cn

Abstract. Emotion recognition is key tech for human-computer interaction and mental health check, it got big value in affective computing and smart healthcare. Old single-modal way only use one data source, so noise mess it up easy and it hard to catch full emotional meaning. But multimodal method mix visual, voice and text data together, use how they help each other to make result much better. Still, it have some trouble: feature talk bad between modes, meaning not match well, and too much noise stay around. The LRD method split weights and fused modal features side by side, which cut down model params a lot. Tests on CMU-MOSI and CMU-MOSEI datasets show the new model boost accuracy in multimodal emotion recognition and also keep model simpler.

Keywords: Multimodal emotion recognition, feature fusion, attention mechanism, Transformer, Information-Assisted Learning

1. Introduction

Emotions play a super important role in human life. They deeply shape how people make choices, talk daily, and understand others. As a natural response to outside stuff, emotions closely link to thinking, actions, and body state. They affect personal mood and health in mind level, and also matter a lot in social interaction.To break through these limits, multimodal emotion recognition become a new hot direction that researchers pay more and more attention to. It combine voice, text, face expression and other modal info to figure out people's emotional state, use how different modes help each other, so it can catch individual emotion more fully and more right [1].

2. Technologies related to multimodal emotion recognition

Emotion is a key trait of human intelligence, it grow deep in each person's brain as subjective experience, and always shape how people think and decide. Emotion types are diverse, include joy, anger, sadness and happiness shifts, which show it is fuzzy, subjective, complex and change over time. Now the main emotion description models fall into two kinds: discrete emotion description model and dimensional emotion description model.(As shown in Table 1)

DOI: 10.54254/2755-2721/2026.35825

Table 1. Different discrete emotion description models

presenter	Emotional Category
Amold (1960) Gray (1982)	Anger, disgust, bravery, frustration, longing, despair, fear, dislike, hope, affection, rage, terror, anxiety, happiness
Izard (1971)	Anger, contempt, disgust, sorrow, fear, guilt, amusement, happiness, shame, surprise
Ekman et.al.(1982)	Anger, disgust, fear, joy, sadness, surprise
Fridja (1986)	Desire, happiness, love, surprise, astonishment, regret
James (1890)	Fear, sadness, love, rage

Convolutional layer is the heart of CNN. This layer use learnable kernel move on input image receptive field to catch certain features [2]. As Figure 1 show. First it slide the kernel across image width and height with set stride, then at each spot it do dot product between kernel and pixel values in that receptive field, and after many rounds it scan the whole image.

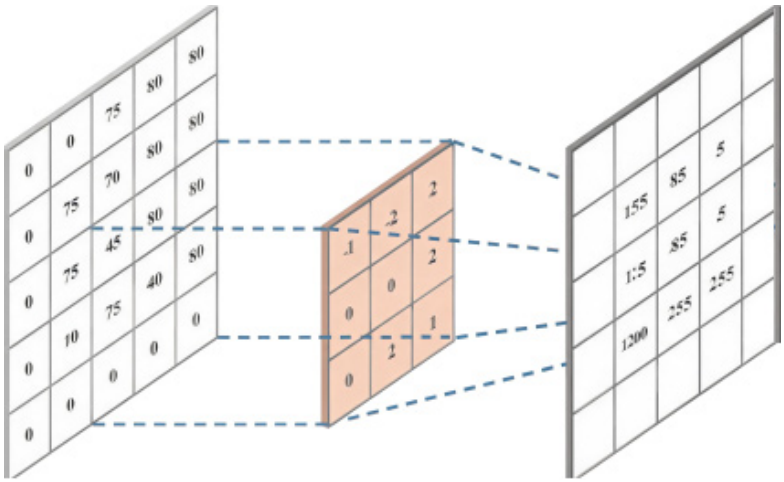

Figure 1. Schematic diagram of the convolution process

A typical CNN usually stack many convolution layers, through this layered build it form feature show from low level to high level. Shallow layers pull out basic stuff like edge, texture and color, while deep layers take feature maps from before as input, and further abstract high-level semantic features like whole outline [3]. Sigmoid function is one of the common activation functions used in neural networks, its curve is S-shaped, the mathematical expression is as Formula 1 show:

$$\sigma(x)=\frac{1}{1+e^{-x}} \tag{1}$$

When neural network do back propagation, neurons close to 0 or 1 have gradient near 0 too, weights basically stop update, so gradient vanish problem easy happen, which make network hard to converge or train slow.

DOI: 10.54254/2755-2721/2026.35825

3. Spatiotemporal fusion based on bidirectional long-term and short-term memory networks

BiLSTM is a special kind of RNN. It can handle sequence data, but unlike other sequence models, it also keep long-term memory.As Figure 2 show. The difference from traditional RNN is that BiLSTM look at both past and future info at same time, so it can catch context relation in sequence data better.

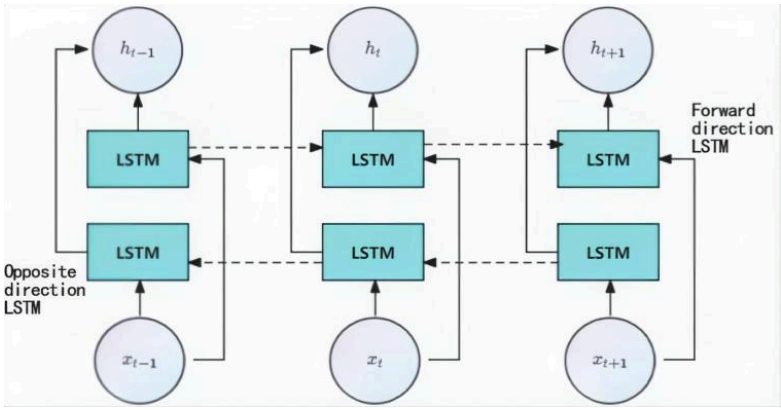

Figure 2. Structure diagram of a bidirectional memory network

At the specific speaker stage, a 3D face model guided attention network is proposed to make video tailored for different people [4]. It take animation params as input, and use attention mask to control facial expression change of input person. Also, to build real match between visual motion (that is, face expression change and head movement) and audio, text-driven face generation methods often use high-precision motion capture dataset, instead of rely on long video of specific person. In this paper, to make lip-audio sync and mouth in generated video clear, LSGAN loss is applied for mouth training as formula 2 show:

$$\mathrm{L}_1^{\mathrm{mou}}=\frac{1}{\mathrm{T}}\sum_{\mathrm{i}=1}^{\mathrm{T}}||\mathrm{m}_{\mathrm{i}}^{\mathrm{mou}}-\widehat{\mathrm{m}}_{\mathrm{i}}^{\mathrm{mou}}||_1 \tag{2}$$

In the formula, m and $\mathrm{m}_{\mathrm{i}}^{\mathrm{mou}}$ is the real and generated mouth feature vector of the i-th frame respectively.

MMD is calculated as Formula 3 show, during training, MMD loss decrease to narrow source domain and target domain in feature space, which help predict target domain better.

$$MMD_{X^S X^r}=\left|\left|\frac{1}{N^S\sum_{i=1}^{NS}\phi\left(x_i^S\right)}-\frac{1}{N^T\sum_{j=1}^{NT}\phi\left(x_i^T\right)}\right|\right| \tag{3}$$

This method build separate feature extractor and classifier for each source domain, and finally do simple average on target domain prediction results, which preliminarily solve the challenge of multi-source heterogeneity.

DOI: 10.54254/2755-2721/2026.35825

4. Space transformation network module

To make the proposed model more robust to geometric changes in face images,a STN module is added at the input layer.This module can do dynamic spatial transform on input face images,so it can handle geometric changes caused by shooting angle,pose,expression and other factors [5]. As Figure 3 show,the STN module include three main parts.

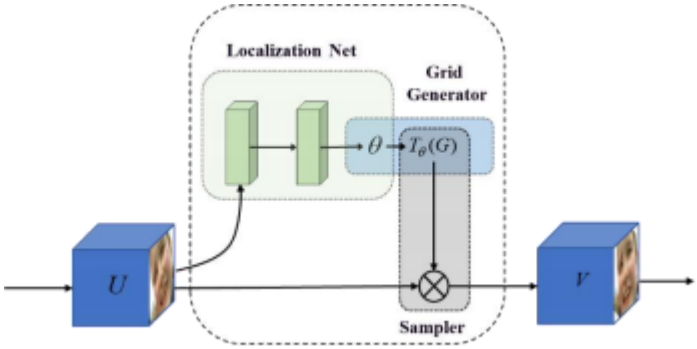

Figure 3. Space transformation network module

The output feature map is $V \in R^{H \times W \times C}$, H and W is the height and width of grid respectively, C is channel number, as Formula 4 show:

$$\begin{pmatrix} x_i^t \\ y_i^t \end{pmatrix} = T_\theta(G_i) \tag{4}$$

Sampler sample input feature map according to sampling grid, use bilinear interpolation to generate output after spatial transform, the calculation formula is as follow:

$$V_i^c = \sum_n^H \sum_m^W U_{nm}^c \max(0,1-|x_i^s - m|)\max(0,1-|y_i^s - n|) \tag{5}$$

Where, U_{nm}^c is the output at position (m,n) in input channel C, V_i^c is the output at position (x_i^s, y_i^s).

Features go through linear layer with GELU activation function then input to SW-MSA, so it can do cross-window info interaction and boost model's global feature modeling ability. During W-MSA and SW-MSA calculation, relative position bias (RPB) and residual connection is added at same time, finally the extracted global features is input to MPF [6]. The process is as Formula 6 and 7 show:

$$g_i = f^{1x1}(W-MSA(LNG_{i-1})) + G_{i-1} \tag{6}$$

$$G_i = f^{1x1}(SW-MSA(LNg_{i-1})) + g_{i-1} \tag{7}$$

DOI: 10.54254/2755-2721/2026.35825

Where, g_i is the intermediate feature after W-MSA, G_i is the output feature of global feature block.

5. Multimodal gating fusion module

This paper propose a multimodal gate control module, which is improved from GCT (Gated Channel Transformation) network to get efficient and accurate context info model. As Figure 4 show. The model use normalization method, through simple normalization it achieve efficient calculation [7]. Second, a training architecture based on normalized component is carefully designed, and gate control is used to adapt to context.

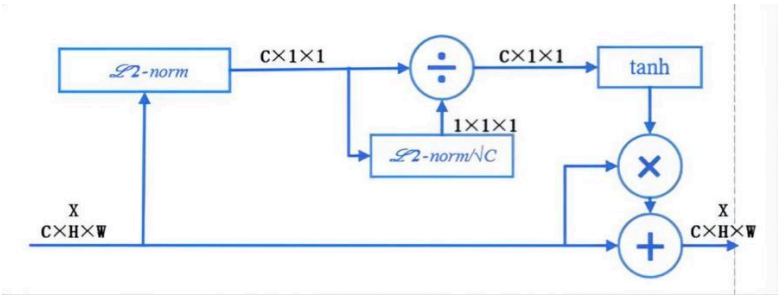

Figure 4. Structural diagram of the multimodal gate control module

$$x=\mathrm{cancat}(X_s,X_T)+\mathrm{cancat}(Y_s,Y_T) \tag{8}$$

However, when s value close to 0 or 1, gradient vanish easy happen in training. So this paper introduce gate control self-adaption, use 1 + tanh(x) for activation, which make training process more stable.

Given the obtained representation x, first apply a fully connected layer with rectified linear units (ReLUs) for nonlinear transform g, and use a softmax output layer to get y for utterance emotion classification.

$$g=\mathrm{ReLu}(W_g x) \tag{9}$$

$$y=\mathrm{softmax}(W_e g+b) \tag{10}$$

Further, emotion weight w and validity threshold t~ is calculated according to global loss function, parameters is learned by gradient descent method, where yi is predicted emotion category, yi is actual emotion category, N is total sample number.

6. Conclusion

Multimodal emotion recognition combine visual, voice, text and other modal info, which can catch and understand human emotional state more fully and accurately, and provide strong support for affective computing, human-computer interaction and mental health analysis. To improve emotion

DOI: 10.54254/2755-2721/2026.35825

recognition accuracy and efficiency, this paper study multimodal fusion based emotion recognition technology. Aiming at current challenges, from feature fusion, modal alignment and modal translation angle, effective methods is proposed to enhance emotion recognition accuracy. Multimodal emotion recognition is used for mental disease auxiliary treatment. Through real-time monitor and analyze patient emotional state, it can provide personalized treatment plan and auxiliary support for mental disease patients. In future work, application of multimodal emotion recognition in mental disease early screening, condition monitoring and treatment effect evaluation will be explored.

References

[1] Wang, Y., Cui, Z., Liu, M., et al. (2026). Incomplete multimodal probability flow recovery for emotion recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, PP*. DOI:10.1109/TPAMI.2026.3713383.

[2] Xiao, K., Yang, W., Zhu, G., et al. (2026). Improved evidence theory based on information fusion for multimodal emotion recognition. *The Journal of Supercomputing, 82*(10), 546-546. DOI:10.1007/S11227-026-08685-1.

[3] Liu, Y., Liu, S., Zhang, J., et al. (2026). Hinting the unknown: Effective open-set multimodal emotion recognition with a hierarchical cross-modal emotion-interactive prompting approach. *Information Fusion, 136*, 104530-104530. DOI:10.1016/J.INFFUS.2026.104530.

[4] Rishu, & Kukreja, V. (2026). A triangular fuzzy multimodal framework for comic emotion recognition: Integrating onomatopoeia, characters, and environmental cues. *Entertainment Computing, 58*, 101158-101158. DOI:10.1016/J.ENTCOM.2026.101158.

[5] Zhang, J., Zhao, P., Wang, Q., et al. (2026). Senti: Semantic enhancement and relation propagation network for multimodal emotion recognition in conversations. *Pattern Recognition, 180*(PB), 114019–114019. https://doi.org/10.1016/J.PATCOG.2026.114019.

[6] Ping, X., Huang, W., & Zheng, Y. (2026). Mamba-based enhanced multimodal emotion recognition with EEG guidance. *IEEE Journal of Biomedical and Health Informatics, PP*. https://doi.org/10.1109/JBHI.2026.3702978.

[7] Nguyen, A. P., Le, S. T., Le, T. D., et al. (2026). Leveraging self-paced curriculum learning for enhanced modality balance in multimodal conversational emotion recognition. *Neural Computing and Applications, 38*(11), 459–459. https://doi.org/10.1007/S00521-026-12160-6.