HetEmotionNet: Two-Stream Heterogeneous Graph Recurrent Neural Network for Multi-modal Emotion Recognition

Ziyu Jia, Youfang Lin, Jing Wang, Zhiyang Feng, Xiangheng Xie, and Caijie Chen

Introduction

Emotion:

- Emotion is a mental and physiological state which results from many senses and thoughts;
- Emotion recognition plays an increasingly important role in multiple areas;
- Multimedia materials stimulate participants and induce emotions.

Emotion recognition based on external behavior:

- People can disguise their facial expression, sound, and so on;
- External behavior cannot reflect a real emotional state.

Emotion recognition based on physiological signals:

- Physiological signals reflect emotion objectively;
- Multi-modal physiological signals contain more emotion information.

Related Work

· ***Traditional machine learning methods:***

- Require a lot of prior knowledge, such as SVM[1].

· ***Multi-domain features extraction:***

- Extract spatial-temporal domain features or spatial-spectral domain features, such as STRNN[2];
- Ignore the complementarity among spatial-spectral-temporal domain features.

· ***Multi-modal physiological signals:***

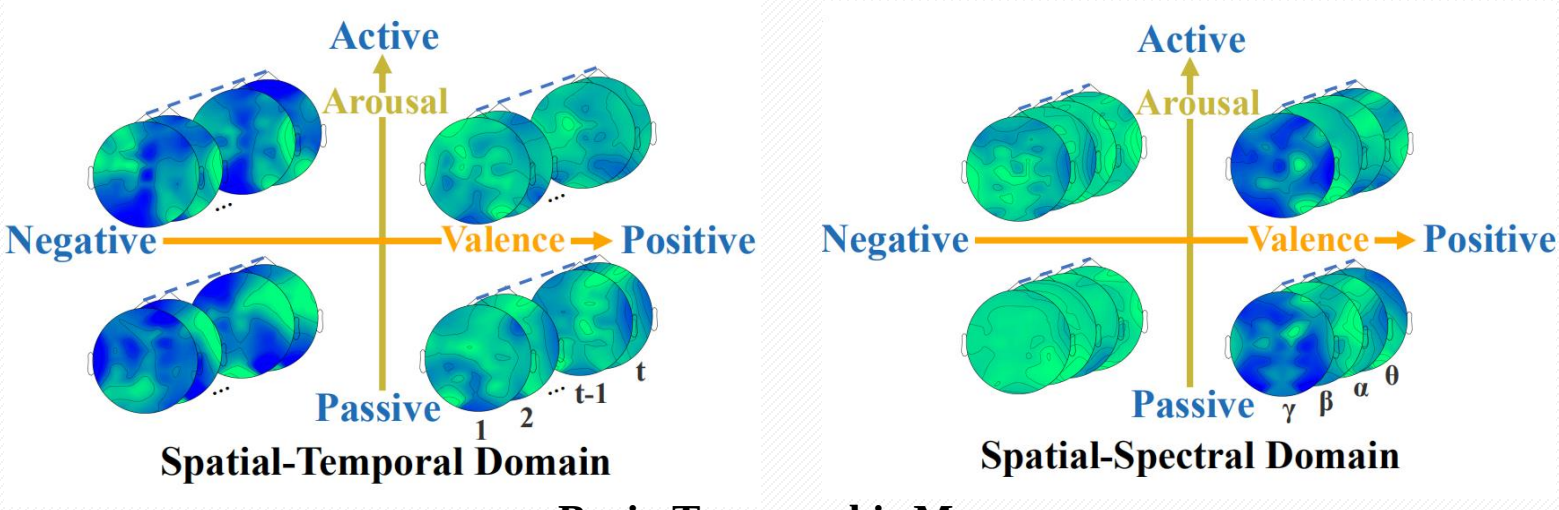
- Only model one of the heterogeneity or correlation of multi-modal physiological signals, such as MMResLSTM[3].

· ***SST-EmotionNet[4]:***

- Extract spatial-spectral-temporal domain features;
- Image-like maps ignore the functional connectivity of the brain and may introduce noise;
- Does not use multi-modal physiological signals.

Challenge

C1: **How to utilize the complementarity among spatial-spectral-temporal domain information efficiently.**



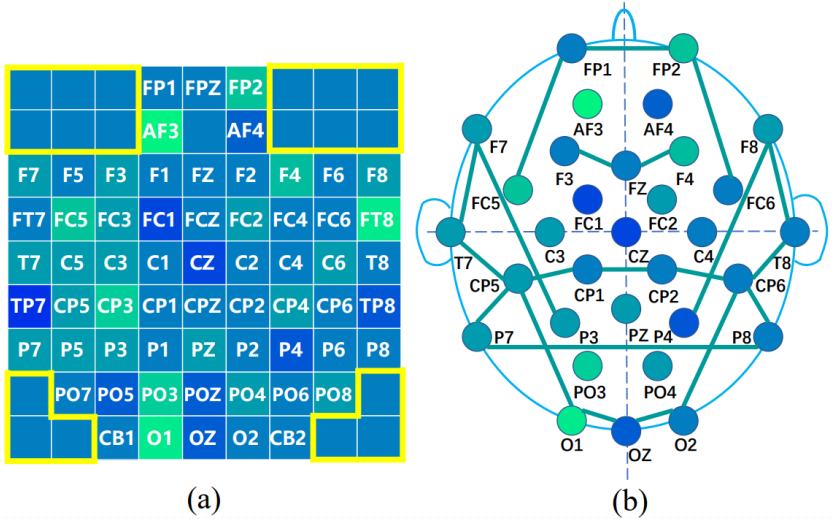
Brain Topographic Maps

- Different activation degree;
- Spatial-temporal domain information and spatial-spectral domain information are complementary.

Challenge

C1: **How to utilize the complementarity among spatial-spectral-temporal domain information efficiently.**

- ***Image-like maps:***
  - Extract spatial-spectral-temporal domain features;
  - May introduce noise;
  - Can not reflect functional connectivity.
- ***Brain graph representation:***
  - Reflect the topological relationship of the brain;
  - No noise introduced.



Different EEG Representation

Challenge

C2: **How to model heterogeneity and correlation among different modalities simultaneously.**

· ***Heterogeneity:***

- Differences among the attributes of various signals collected from different organs;
- Existing works use different feature extractors to capture the heterogeneity.

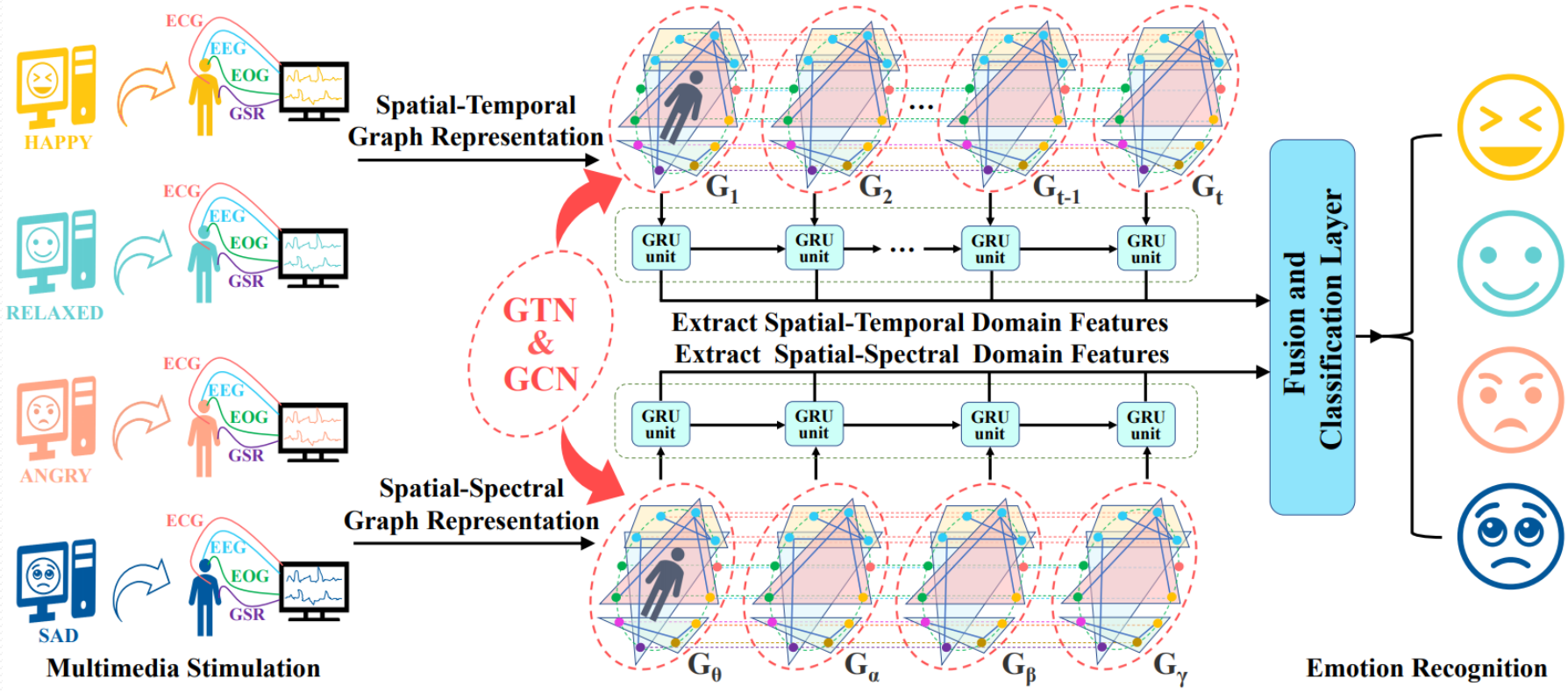
· ***Correlation:***

- The relationship among channels in the same modality or in different modalities;
- Existing works usually feed a new data representation to a deep neural network to capture the correlation, such as GSCCA[5].

There are differences in the waveform and amplitude between EEG and ECG signals.

Methods

The whole process for multi-modal emotion recognition.

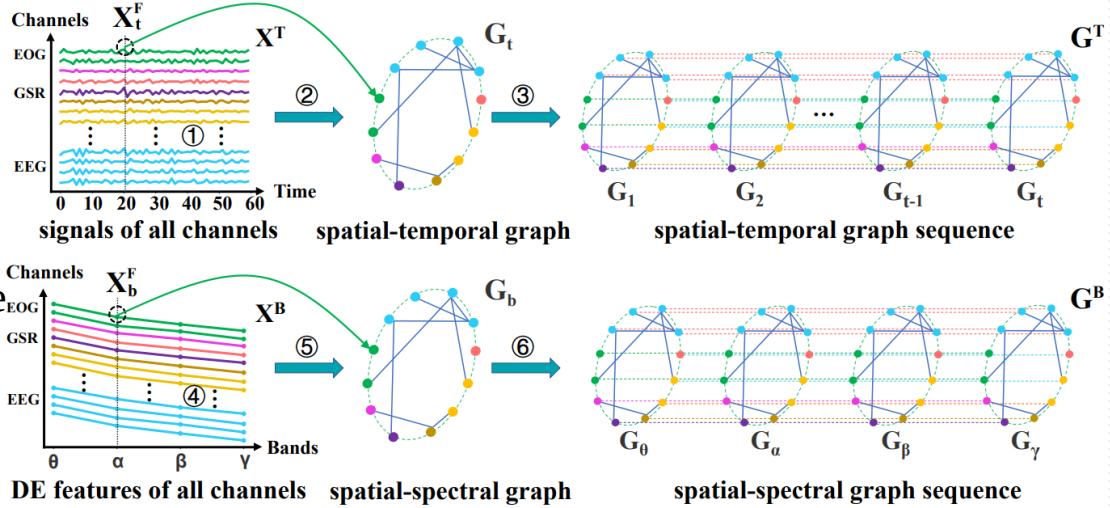


Methods

C1: **How to utilize the complementarity among spatial-spectral-temporal domain information efficiently?**

S1: **We construct spatial-temporal graph sequence and spatial-spectral graph sequence to extract the spatial-spectral-temporal domain features of physiological signals.**

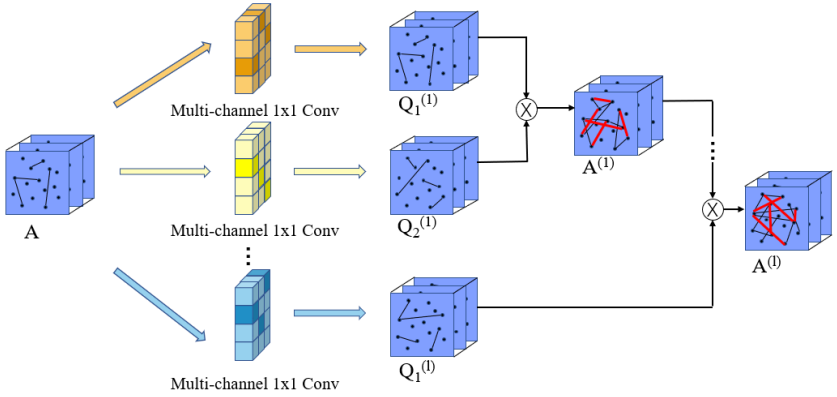
- Calculate the mutual information between channels to get the adjacency matrix;
- Construct spatial-temporal graph and spatial-spectral graph;
- Construct spatial-temporal graph sequence and spatial-spectral graph sequence;
- Use GRU to extract time-domain and frequency-domain dependencies.



Methods

C2: **How to model heterogeneity and correlation among different modalities simultaneously?**

S2: **We use GTN to model the heterogeneity of multimodal data, and GCN to model the correlation.**



· ***Model the heterogeneity:***

- Use GTN to extract multiple meta-paths automatically.

· ***Model the correlation:***

- Use GCN to aggregate neighborhood nodes.

Experiments-Dataset

DEAP dataset

- A total of 32 subjects;
- Each participant needs to undergo 40 trials;
- Collect 60s signals for each trial;
- 32-channel EEG signals and 8-channel peripheral physiological signals (PPS);
- PPS include EOG, EMG, GSR, BVP, respiration, and temperature;
- Music videos are rated on valence and arousal from 1 to 9.

MAHNOB-HCI dataset

- A total of 27 subjects;
- Each participant needs to undergo 20 trials;
- The length of video clips is between 34.9s and 117s;
- 32-channel EEG signals and 6-channel PPS;
- PPS include ECG, GSR, respiration, and temperature;
- Video clips are rated on valence and arousal from 1 to 9.

Experiments-Baselines

- **MLP[6]:** Multilayer perceptron is a classical artificial neural network.
- **SVM[7]:** Support vector machine classifier with radial basis function kernel.
- **GCN[8]:** Graph convolutional network extracts spatial domain information from signals through spectral graph convolution.
- **DGCNN[9]:** Dynamical graph convolutional neural network can dynamically learn correlation among channels.
- **MM-ResLSTM[3]:** Multi-modal residual LSTM can capture the temporal domain information and spatial domain information of multi-modal signals by sharing the LSTM weight and residual network.
- **ACRNN[10]:** Attention based convolutional recurrent neural network uses CNN with channel-wise attention mechanism to capture spatial domain information, and utilizes LSTM with self-attention mechanism to capture temporal domain information.
- **SST-EmotionNet[4]:** Spatial-spectral-temporal based attention 3D dense network integrates various features in a unified network framework, and uses 3D attention mechanism to capture local pattens in EEG signals.

Experiments-Results

Comparison with the state-of-the-art models

Table 1: The performance on the DEAP dataset.

Model	Valence (%)	Arousal (%)
MLP [30]	74.31±4.56	76.23±5.12
SVM [2]	83.14±3.66	84.50±4.43
GCN [13]	89.17±2.90	90.33±3.59
DGCNN [35]	90.44±3.01	91.70±3.46
MM-ResLSTM [25]	92.30±1.55	92.87±2.11
ACRNN [36]	93.72±3.21	93.38±3.73
SST-EmotionNet [10]	95.54±2.54	95.97±2.86
HetEmotionNet	**97.66±1.54**	**97.30±1.65**

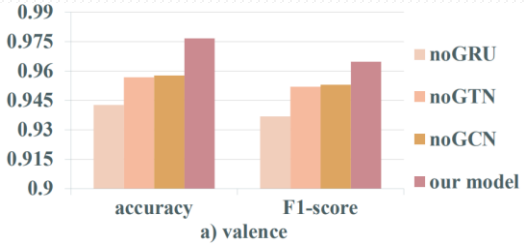
Table 2: The performance on the MAHNOB-HCI dataset.

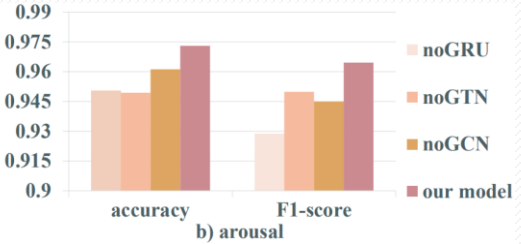
Model	Valence (%)	Arousal (%)
MLP [40]	72.84±6.88	73.61±10.75
SVM [2]	77.31±6.77	77.16±9.14
GCN [20]	81.31±6.09	83.43±7.50
DGCNN [45]	86.21±9.35	86.07±10.53
MM-ResLSTM [35]	89.46±8.64	89.66±8.09
ACRNN [46]	88.10±5.49	89.90±5.96
SST-EmotionNet [12]	90.06±4.80	88.37±7.19
HetEmotionNet	**93.95±3.38**	**93.90±3.04**

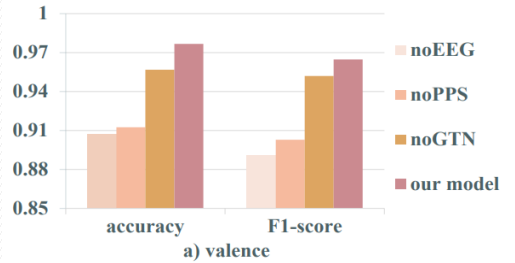
HetEmotionNet achieves the best performance in DEAP and MAHNOB-HCI datasets.

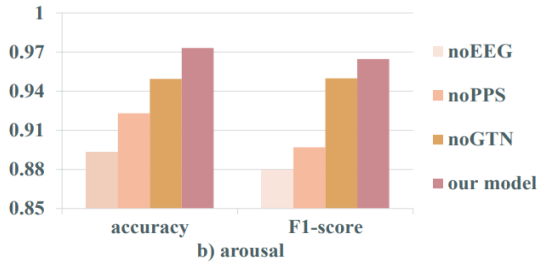
Experiments-Ablation studies in DEAP dataset

①Ablation studies on different components



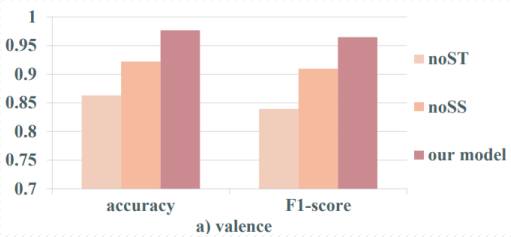






②Ablation studies on fusing different modalities

③Ablation studies on two-stream structure



Conclusion

· *Contributions:*

- We propose a novel *graph-based two-stream structure* composed of the *spatial-temporal stream* and the *spatial-spectral stream* which can *simultaneously* fuse *spatial-spectral-temporal domain features* of physiological signals in a unified deep neural network framework.
- Each stream consists of a *GTN* for modeling the *heterogeneity*, a *GCN* for modeling the *correlation*, and a *GRU* for capturing the temporal or spectral *dependency*.
- *Extensive* experiments are conducted on *two benchmark datasets* to evaluate the performance of the proposed model. The results indicate the proposed model *outperforms all the state-of-the-art models*.

· *Prospects:*

- The proposed model is *a general-framework* for multivariate physiological time series.
- It can be applied to time series classification, prediction, and other related fields.

References

[1] Yi-Lin Lin and Gang Wei. Speech emotion recognition based on hmm and svm. In 2005 international conference on machine learning and cybernetics, volume 8, pages 4898–4901. IEEE, 2005.

[2] Tong Zhang, Wenming Zheng, Zhen Cui, Yuan Zong, and Yang Li. Spatial–temporal recurrent neural network for emotion recognition. IEEE transactions on cybernetics, 49(3):839–847, 2018.

[3] Jiaxin Ma, Hao Tang, Wei-Long Zheng, and Bao-Liang Lu. Emotion recognition using multimodal residual lstm network. In Proceedings of the 27th ACM International Conference on Multimedia, pages 176–183, 2019.

[4] Ziyu Jia, Youfang Lin, Xiyang Cai, Haobin Chen, Haijun Gou, and Jing Wang. Sst-emotionnet: Spatial-spectral-temporal based attention 3d dense network for eeg emotion recognition. In Proceedings of the 28th ACM International Conference on Multimedia, pages 2909–2917, 2020.

[5] Xiaowei Zhang, Jing Pan, Jian Shen, Zia Ud Din, Junlei Li, Dawei Lu, Manxi Wu, and Bin Hu. Fusing of electroencephalogram and eye movement with group sparse canonical correlation analysis for anxiety detection. IEEE Transactions on Affective Computing, 2020.

[6] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

[7] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. A training algorithm for optimal margin classifiers. In Proceedings of the fifth annual workshop on Computational learning theory, pages 144–152, 1992.

[8] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907, 2016.

[9] Tengfei Song, Wenming Zheng, Peng Song, and Zhen Cui. Eeg emotion recognition using dynamical graph convolutional neural networks. IEEE Transactions on Affective Computing, 2018.

[10] Wei Tao, Chang Li, Rencheng Song, Juan Cheng, Yu Liu, Feng Wan, and Xun Chen. Eeg-based emotion recognition via channel-wise attention and self attention. IEEE Transactions on Affective Computing, 2020.

Thanks!