2024 3rd International Conference on Artificial Intelligence and Computer Information Technology (AICIT) | 979-8-3503-6544-3/24/$31.00 ©2024 IEEE | DOI: 10.1109/AICIT62434.2024.10730531

Engagement Prediction Method Based on Spatiotemporal Facial Feature Redistribution

Weiwei Zhu*
School of Electrical and Information Engineering
Wuhan Institute of Technology
Wuhan, China
22203010049@stu.wit.edu.cn
*Corresponding author

Xi Li
School of Electrical and Information Engineering
Wuhan Institute of Technology
Wuhan, China
lixi@wit.edu.cn

***Abstract*—Automated detection of learners' engagement levels enhances online teaching by enabling timely feedback and personalized adjustments based on student needs, thereby improving teaching effectiveness. Traditional methods rely heavily on single-frame facial spatial data, often neglecting temporal emotional and behavioral features and being sensitive to pose variations. Additionally, convolutional padding can degrade feature maps, impacting feature extraction quality. To tackle these challenges, we propose the Spatiotemporal Facial Feature Redistribution and Temporal Convolutional Network (SFRNet). This hybrid neural network architecture incorporates three key components: utilizing the spatial attention mechanism LKA to capture local patches and mitigate pose effects, employing the FOWD module to redistribute feature weights, eliminate irrelevant features, and enhance facial representation, and analyzing temporal changes using the ModernTCN module to detect engagement levels. We curated a near-infrared engagement video dataset (NEVD) to validate SFRNet's efficiency, demonstrating superior performance in engagement video analysis on both NEVD and publicly available DAiSEE datasets through extensive experimentation and comprehensive studies.**

Keywords—engagement prediction, spatiotemporal network, re-distributing facial featrues, temporal convolutional network.

I. Introduction

With the rise of the internet, online education has become a hot topic. Although schools have transitioned to online teaching, assessing student engagement remains challenging, especially with large student numbers. Developing automated engagement predictions is crucial for quantifying online education quality and improving learning effectiveness. Automatic detection of student engagement involves various modalities such as images[1], videos[2], audio, and electrocardiograms (ECGs), primarily using video equipment to assess engagement in online classrooms. Despite advancements in attention recognition, challenges persist in data collection, annotation complexity, and the spatiotemporal aspects of engagement prediction. Image-based methods are affected by lighting, pose, and facial expression variations, limiting their ability to fully capture students' spatiotemporal emotional behaviors. While video-based approaches require less annotation, they face challenges in accurate facial feature extraction, sample availability, and high computational demands, hindering real-time online deployment. Additionally, convolutional neural networks exhibit inherent characteristics that limit their performance and application. Convolution kernels often focus on the central regions of images, neglecting edge information—a phenomenon known as perception bias. As network depth increases, padding to balance pixel perception can lead to sparse edge pixels, making it challenging for deeper layers to accurately capture edge details.

This paper introduces SFRNet, a novel hybrid neural network architecture combining residual networks and temporal convolutional networks for engagement prediction. Facial images are extracted using MTCNN [3] from attention video datasets. The spatial module employs ResNet-18 [4] for feature extraction up to the GAP layer. The redistributing facial features module integrates LKA [5] spatial attention and FOWD to mitigate feature erosion. Dynamic facial feature changes are categorized into common and unique features to simplify extraction. High-quality feature vectors from consecutive frames input into ModernTCN [6] simulate temporal information. A SoftMax-activated projection layer maps final representations for classification. Existing attention detection datasets focus on visible light, sensitive to lighting changes. Near-infrared (NIR) imaging offers a reliable alternative. NEVD validates SFRNet, achieving 90.8% accuracy on NEVD and 61.2% on DAiSEE [7], outperforming most other competing methods.

II. Methods

A. Data preprocessing

First, we preprocess the faces in the input video using the multitask convolutional neural network (MTCNN), extracting high-quality facial images of size C×112×112 from consecutive frames. The MTCNN consists of three cascaded lightweight CNN models: the proposal network (P-Net), the refine network (R-Net), and the output network (O-Net), progressively extracting and refining candidate boxes.

B. Redistributing facial features Module

1) LKA attention mechanism: In Fig. 1,for a given facial image, CNNs extract feature mappings. We then use multiple spatial attention mechanisms to capture local patches automatically. However, ensuring comprehensive recognition of facial regions, especially under pose variations or strong occlusions, remains challenging and can affect attention recognition performance. Traditional methods treat images as

This work was supported by the 15th graduate education innovation fund of Wuhan Institute of Technology (Grant No.CX2023551) and the scientific research foundation of Nanchang Institute of Science and Technology (No. NGRCZX-20-06).

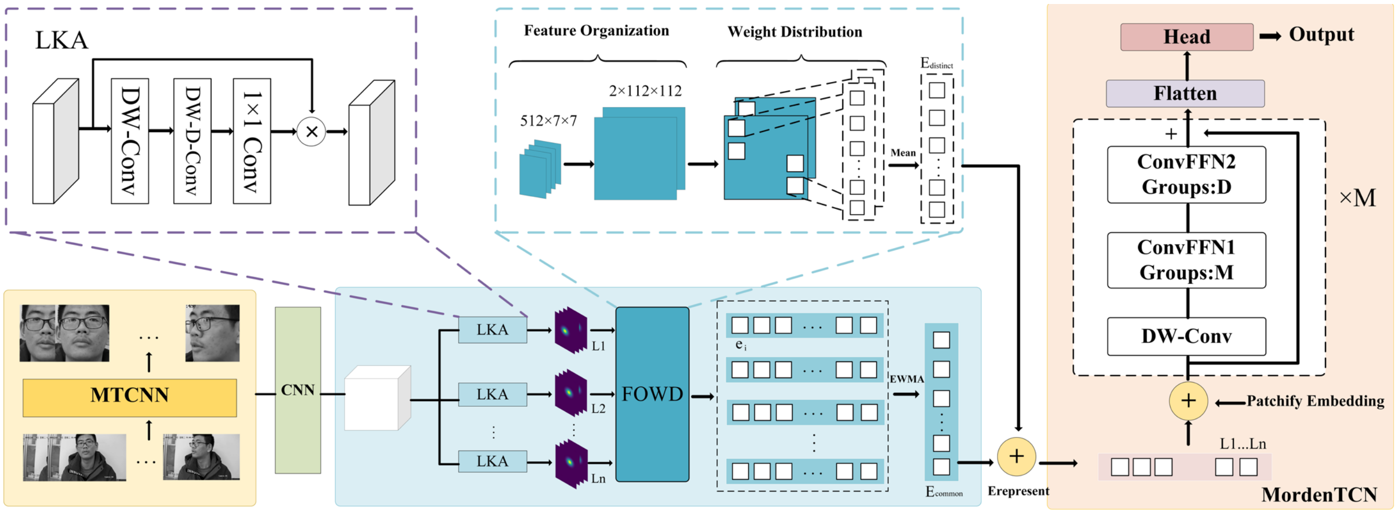

Fig. 1. Architecture of the SFRNet model

one-dimensional sequences, lacking channel adatability. To improve this, we employ LKA, combining convolution and self-attention for better adaptation to local context and long-term dependencies. LKA convolution is divided into three parts: spatial local convolution (DW-Conv), spatial remote convolution (DW-D-Conv), and channel convolution (1×1Conv).

$$Attention = Conv_{1\times1}\left(DW\text{-}D\text{-}Conv\left(DW\text{-}Conv\left(DW\text{-}Conv\left(F\right)\right)\right)\right) \quad (1)$$

$$Output = Attention \otimes F \quad (2)$$

where F represents the input features $F \in \mathbb{R}^{C\times H\times W}$, $Attention \in \mathbb{R}^{C\times H\times W}$ represents the attention map, where different values denote the importance of different features, and $\otimes$ is the elementwise product.

2) FOWD module: Due to convolution's nature, padding maintains feature map size and enhances network performance by reducing information loss. However, excessive padding introduces irrelevant white features, which traditional pooling layers can't fully address. To counter this, we propose the FOWD module, with Feature Organization (FO) amplifying features and Weight Distribution (WD) managing their weights. Padding results in white pixels at edges, which disrupt precise information processing. We need a method to aggregate white pixels while maintaining their relative positional distribution, concentrating the most eroded pixels at the periphery of the feature map, and reconstructing the feature map, as shown in Fig. 1. Additionally, we use unpadded blocks to weaken the weight of edge whitening information, counteracting the adverse effects of padding. We adopted the effective subpixel convolution layer method to reconstruct the feature map, as shown in Fig. 2. In this process, only the absolute position of the feature points changes, while their relative position remains unchanged. The channel is scaled down to 4, corresponding to the original feature point positions and those in the feature cluster. We designed a dedicated convolutional layer to mitigate the impact of peripheral pixels, cleverly utilizing perceptual bias, as shown in Fig. 3. By adjusting the size and stride of the convolutional kernel, we can filter out white features while retaining crucial parts.

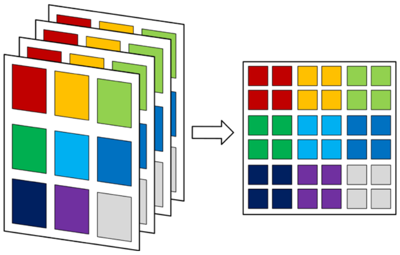

Fig. 2. Example of our feature map rearrangement method

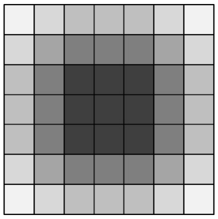

Fig. 3. A refill-free convolution over a 7×7 pixel image

3) Facial Affinity Enhancement: In facial attention detection, minimal changes occur between adjacent frames. We divide facial features into common and unique types. Learning common features simplifies extraction and differentiates individuals without starting from scratch. We use Exponentially Weighted Moving Average (EWMA) to update average features and adapt to varying conditions.

$$\begin{cases} E_{represent} = E_{common} + E_{distinct} \\ E_{batch}, E_{init} = \dfrac{\sum_{i=1}^{N} e_i}{N} \\ E_{common} = \lambda E_{batch} + (1-\lambda) E'_{common} \end{cases} \quad (3)$$

C. Timing feature extraction module

Feature vectors from consecutive frames serve as multidimensional inputs for ModernTCN, simulating temporal video data. ModernTCN uses large kernels to expand the receptive field, enhancing global awareness and capturing variable correlations. Before the backbone network, channel features are embedded into d-dimensional vectors using a patchwise variable-agnostic method to manage complex dependencies. After appropriately padding a time series $x_{in} \in R^{K \times L}$ consisting of K variables (each of length L), it is further divided into N patches of size P, spaced apart by a step size of S to ensure nonoverlapping. Then, these patches are embedded into vectors.

$$x_{emb} = Embbdding\left(X_{in}\right) \tag{4}$$

$x_{emb} \in \mathbb{R}^{K \times D \times N}$ is the embedded input. Then, these patches are embedded into d-dimensional embedding vectors using an equivalent fully convolutional approach. After reshaping the shape to $x_{in} \in R^{K \times 1 \times L}$, the padded input is mapped to a vector with D output channels through a 1D convolutional layer. The backbone network consists of two stacked ModernTCN blocks, applying residual formulas for classification.

III. EXPREIMENTAL VALIDATION

A. Datasets

1) NEVD: The experiment utilized the self-collected Near Infrared Engagement Dataset (NEVD) and the online education emotion dataset DAiSEE [7] to train and validate our approach. Videos of 25 university students were recorded using Hikvision network cameras, with each student contributing four four-minute videos covering various engagement levels. Videos were captured at 1920×1080 pixels and 30 frames per second. Consistency between engagement processing methods in NEVD and DAiSEE ensured experimental reliability and comparability. Fig. 4 shows examples of videos from the NEVD.

2) DAiSEE: DAiSEE (Database for Affect in Situations of Elicitation) is a significant dataset for engagement detection, containing 9068 videos from 112 students in online courses. These videos assess emotional states like boredom, confusion, engagement, and frustration. Each 10 second video is categorized into four engagement levels: 0 (very low), 1 (low), 2 (high), and 3 (very high), recorded at 30 fps with a resolution of 640 × 480 pixels.

B. Experimental details

Our method uses the SGD optimizer with L2 loss, an initial learning rate of 10-4, and a momentum of 0.9 for training, which includes 100 epochs. Videos are downsampled to obtain a tensor of size 16 × C × 112 × 112 as input. ResNet-18 extracts 512× 7× 7 feature vectors from consecutive frames. After redistributing facial features, the spatial area of the feature map increases to 112 × 112 pixels. No padding is used, making the 32×32×1 pixel convolution kernel, with a stride of 8, crucial. For facial affinity enhancement, the λ gain hyperparameter is set to 0.2 and updated through gradients. During patch embedding, the patch size and stride are set to P = 1 and S = 1. All networks are implemented in PyTorch, and experiments are conducted on an NVIDIA A100 40 GB GPU.

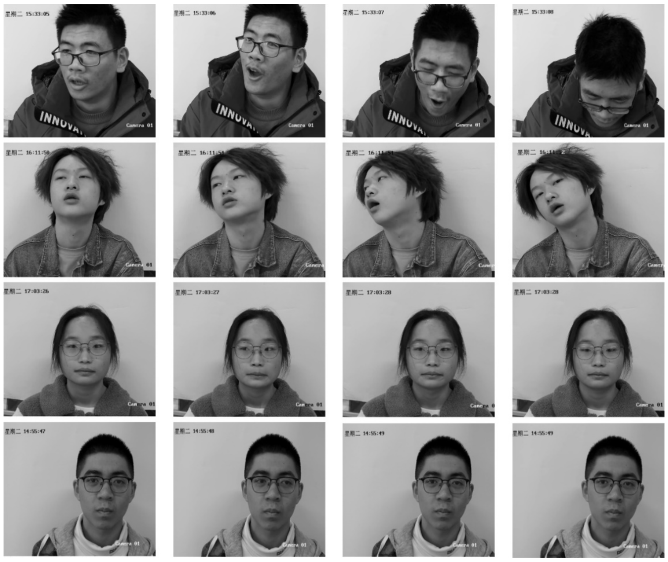

Fig. 4. Examples of videos from the NEVD

C. Comparison with state-of-the-art methods

Our method was tested on the NEVD, and TABLE I compares our results with methods from GitHub repositories, including I3D, C3D, S3D, ResNet+LSTM, ResNet+TCN, and Swin-B[8, 9, 10-12]. Our method outperforms traditional end-to-end methods like I3D, C3D, and S3D, achieving 90.8% accuracy. Compared to state-of-the-art Swin-B, our method shows a 1.6% improvement. The third column indicates video sequence frames, and the fourth and fifth columns show "FLOPS" and "Params" computational complexities. Traditional methods face challenges of high computational cost and memory usage, while ours performs better. We achieve higher accuracy with fewer frames and lower computational cost and memory usage compared to Swin-B.

To validate our method, we conducted experiments on DAiSEE. TABLE II shows engagement intensity prediction results. Our method outperforms most end-to-end and feature-based methods. Compared to SE-GA-RES-LSTM, our method achieves a 1% accuracy increase, using fewer attention modules to capture local patterns effectively. However, our method slightly lags behind the fusion feature-based algorithm in [20], likely due to its larger backbone integrating multiple features. The fourth column lists video sequence frames, and despite using fewer frames, we achieved excellent results.

D. Ablation study

1) Effectiveness of the proposed modules： To validate the SFRNet modules, we conducted ablation experiments on NEVD and DAiSEE in TABLE III. The baseline used RES18 feature maps with ModernTCN, omitting LKA and FOWD. Integrating LKA improved performance by 1.5% and 1.1%, capturing both local and long-range dependencies. Adding FOWD further enhanced performance by 1.6% and 1.2%, optimizing weight distribution to focus on relevant features. The FAE module streamlined feature extraction, adapting to various engagement states, and achieved top performances of 90.8% and 61.2%. The combined modules effectively improved results on NEVD and DAiSEE.

TABLE I. Comparison with other methods on NEVD

Method	Acc	Frames	FLOPS	Param
I3D	83.4	64	37.2	14.4
C3D	85.7	64	38.6	63.2
S3D	87.7	64	24.4	10.3
ResNet-LSTM	87.8	32	9.4	11.2
Swin-B	89.2	32	147.5	88.1
ResNET-TCN	89.5	16	6.3	12.9
Ours	**90.8**	16	19.8	46.2

TABLE II. Comparison with other methods on the DAiSEE

Method	Year	Frames	Acc
I3D [14]	2019	64	52.4
C3D [15]	2019	64	57.6
S3D [16]	2018	64	59.7
DFSTN [17]	2021	20	58.8
DERN [18]	2019	64	60.0
Swin-B [12]	2022	32	60.2
SA-GA-RES-LSTM [19]	2023	16	60.2
SE-ResNET-TCN [20]	2022	10	61.4
SFRNet	2024	16	61.2

TABLE III. Results of ablation experiments on the NEVD and DAiSEE

ALK	FOWD	FAE	NEVD	DAiSEE
			87.4	58.7
✓			88.9	59.8
✓	✓		90.5	61.0
✓	✓	✓	90.8	61.2

IV. Conclusion

This paper presents SFRNet, an end-to-end hybrid neural network combining residual and temporal convolutional networks for engagement prediction. By integrating facial spatiotemporal information, SFRNet captures subtle changes in student engagement. We introduce the redistributing facial features module to address whitening erosion from padding convolutions, optimizing feature extraction. Evaluated on the NEVD and DAiSEE datasets, our approach shows high detection accuracy and strong performance.

References

[1] O. Mohamad Nezami, M. Dras, L. Hamey, D. Richards, S. Wan, and C. Paris, "Automatic Recognition of Student Engagement Using Deep Learning and Facial Expression," Cham, 2020: Springer International Publishing, in Machine Learning and Knowledge Discovery in Databases, pp. 273-289.

[2] P. Guhan, M. Agarwal, N. Awasthi, G. Reeves, D. Manocha, and A. Bera, "ABC-Net: Semi-supervised multimodal GAN-based engagement detection using an affective, behavioral and cognitive model," arXiv preprint arXiv:2011.08690, 2020.

[3] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," IEEE signal processing letters, vol. 23, no. 10, pp. 1499-1503, 2016.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770-778.

[5] M.-H. Guo, C.-Z. Lu, Z.-N. Liu, M.-M. Cheng, and S.-M. Hu, "Visual attention network," Computational Visual Media, vol. 9, no. 4, pp. 733-752, 2023.

[6] D. Luo and X. Wang, "Moderntcn: A modern pure convolution structure for general time series analysis," in The Twelfth International Conference on Learning Representations, 2024.

[7] A. Gupta, A. D'Cunha, K. Awasthi, and V. Balasubramanian, "Daisee: Towards user engagement recognition in the wild. arXiv," arXiv preprint arXiv:1609.01885, 2016.

[8] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2818-2826.

[9] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in Proceedings of the IEEE international conference on computer vision, 2015, pp. 4489-4497.

[10] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy, "Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification," in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 305-321.

[11] A. Abedi and S. S. Khan, "Improving state-of-the-art in detecting student engagement with resnet and tcn hybrid network," in 2021 18th Conference on Robots and Vision (CRV), 2021: IEEE, pp. 151-157.

[12] Z. Liu et al., "Video swin transformer," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 3202-3211.

[13] Y. Lu, Y. Zhan, Z. Yang, and X. Li, "Student engagement recognition network integrating facial appearance and multi-behavior features," Computer Science and Application, vol. 12, p. 1163, 2022.

[14] H. Zhang, X. Xiao, T. Huang, S. Liu, Y. Xia, and J. Li, "An novel end-to-end network for automatic student engagement recognition," in 2019 IEEE 9th International Conference on Electronics Information and Emergency Communication (ICEIEC), 2019: IEEE, pp. 342-345.

[15] L. Geng, M. Xu, Z. Wei, and X. Zhou, "Learning deep spatiotemporal feature for engagement recognition of online courses," in 2019 IEEE symposium series on computational intelligence (SSCI), 2019: IEEE, pp. 442-447.

[16] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy, "Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification," in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 305-321.

[17] J. Liao, Y. Liang, and J. Pan, "Deep facial spatiotemporal network for engagement prediction in online learning," Applied Intelligence, vol. 51, no. 10, pp. 6609-6621, 2021.

[18] T. Huang, Y. Mei, H. Zhang, S. Liu, and H. Yang, "Fine-grained engagement recognition in online learning environment," in 2019 IEEE 9th international conference on electronics information and emergency communication (ICEIEC), 2019: IEEE, pp. 338-341.

[19] Y. Lu, Y. Zhan, Z. Yang, and X. Li, "Student engagement recognition network integrating facial appearance and multi-behavior features," Computer Science and Application, vol. 12, p. 1163, 2022.

[20] Y. Liang, Z. Zhou, W. Huang, and Z. Guo, "Attention detection in online education based on spatiotemporal attention mechanism," Software Guide, vol. 23, no. 01, pp. 150-155, 2024.