Is Facial Geometric-Encoding Crucial to Understanding Facial Data in Video Sequences?

Emmanuel K. Baah[1,2], James Ben Hayfron-Acquah[1], Dominic Asamoah[1], Kwabena Owusu-Agyemang[1]

[1]*Kwame Nkrumah University of Science and Technology (KNUST), Ghana*
College of Science, Computer Science Department
[2]*Lancaster University, Ghana, Department of Computer Science*

Introduction

Understanding video sequences based on facial expressions is strongly linked with NLP. Both are sequential data. Additionally, as these long-range attention models from NLP work will help understand a word in a sentence in the context of its usage, similarly, the subtle minute changes in video sequences extracted in segments would require contextualization with the entire sequence to yield a full understanding of the video sequences. Thus, with facial data in video sequences, adapting these long-range self-attention models from NLP would yield better results when used in modelling video sequences based on facial data. In the video domain, several works have focused on 3D and 2D convolutions, which work well with spatiotemporal features extracted from video tasks [8], [9], [10] and the recent evolution of the convolution-free video architecture based on divided space-time attention [11]. However, TimeSformer is a dedicated transformer architecture built on ViT [12] is only useful for action recognition and not facial related tasks.

Problem Statement

The sensitivity to subtle facial changes requires a dedicated attention model that models the geometric layout of faces in video sequences, enabling the divided space-time attention architecture to understand the structural and dynamic aspects of facial data for engagement monitoring. Thus, our architecture explicitly integrates facial landmarks as geometric structural priors, providing an unambiguous signal about the configuration and shape of the face, thereby forcing the model to learn from patches with painstaking precision. Our work aims to investigate whether introducing a facial geometry encoding module can enhance the understanding of facial data in a video sequence.

Literature Review

Our design was inspired by recent studies on video action recognition, which utilize both convolution-based and transformer-based models to understand video sequences.

With models based on the former, the Deep Facial Spatio-Temporal Network [13] utilizes the pre-trained SE-ResNet-50 to extract spatial facial features, while integrating Long-Short Term Memory (LSTM) with global attention to generate an attention hidden state for engagement detection.

Huang et al. [14] combine a Bidirectional LSTM with an attention mechanism to extract facial features in detecting engagements with the proposed Deep Engagement Recognition Network (DERN). Beyond the use of LSTM and BiLSTM, Abedi and Khan [15] employ an end-to-end residual network (ResNet) combined with a temporal convolutional network (TCN).

While the TCN analyses the temporal differences in the video frames for detection, the 2D ResNet is dedicated to extracting spatial features across the sequential video frames.

Our method is closely related to TimeSformer, a unique approach that enables the direct learning of spatiotemporal features from video sequences at the frame-level patch level [17], [18], [19]. Since these works focused on video-based applications of transformers without a dedicated focus on geometric encoding of the facial data, to modulate the TimeSformer to perform better with video sequences based on facial data, they must restrict the scope of the self-attention first to extract the frames and their corresponding landmarks related to them, and generate the mesh-aware encoding [20]. Narayan et al. [21] utilize FaceXFormer as a single, unified framework to analyse facial data, without an explicit focus on long-range video sequences.

Methodology

The proposed architecture, FMeshformer, builds upon the TimeSformer as its base architecture and incorporates computer vision and geometric facial data to facilitate video understanding for facial expressions and their related applications. This is expressed in Figure 1.

FMeshformer: The generic spatial attention overlooks the identity-invariant geometry of facial-expression-related videos and the temporal consistency of these relations. Thus, the introduction of the MeshPositionalEncoder in the FMeshformer is crucial for capturing the relative facial structure using detected facial landmarks.

Methodology

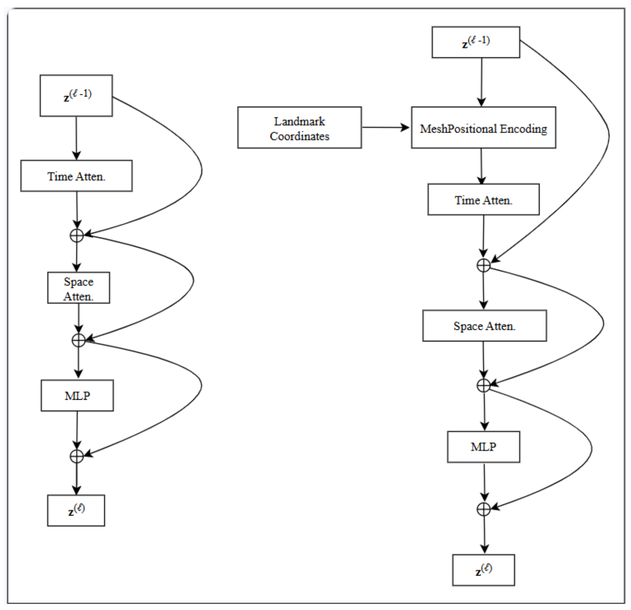

Figure 1. Proposed Architecture (FMeshfomer)

Facial Geometry Representation: The 2D landmark coordinates, K, in each frame t, after normalising the frames to patch grid coordinates, it is expressed as:

$$\mathcal{L}_t = \{\ell_{t,k} \in \mathbb{R}^2 \mid k = 1, \dots, K\}$$

We build a geometric (triangular) mesh $\mathcal{M}_t$ over these landmarks. The mesh-aware feature vector, $\mathbf{g}_{t,x}$ is defined for each flattened patch $x_{(p,t)}$ with centre $(u_{x_{(p,t)}}, v_{x_{(p,t)}})$ is given as:

$$\mathbf{g}_{t,x} = \left[\phi_1\left(u_{x_{(p,t)}}, v_{x_{(p,t)}}, \mathcal{M}_t\right), \dots, \phi_m\left(u_{x_{(p,t)}}, v_{x_{(p,t)}}, \mathcal{M}_t\right)\right],$$

where ϕ_i can include the geodesic distances along the mesh, the Euclidean distances to key landmarks (mouth corners, eyes), and barycentric coordinates w.r.t. nearest facial triangle. This produces the geometry descriptor, $\mathbf{g}_{t,x} \in \mathbb{R}^m$.

MeshPositionalEncoder: A learnable MLP denoted, $f_\theta \longrightarrow \mathbb{R}^d$ maps the geometry descriptor, $\mathbf{g}_{t,x}$, into the embedding space:

$$x^{mesh}_{(p,x)} = f_\theta(\mathbf{g}_{t,x})$$

Thus, the mesh-aware token becomes:

$$z'^{(0)}_{(p,x)} = z^{(0)}_{(p,x)} + x^{mesh}_{(p,x)} + x^{time}_{(t)}$$

Figure 1 shows that the integration of the mesh-aware token occurs after the patch embedding, before the temporal attention, and these are encoded with the generic spatial encoding.

By adapting the TimeSformer, the FMeshformer becomes more efficient, as it employs divided space-time attention. The positioning of the MeshPositionalEncoder before the temporal attention is crucial, as it ensures that the temporal attention "sees" tokens whose spatial representation already encodes the geometry of the face. Our experiments indicate that integrating the MeshPositionalEncoder with the TimeSformer for video sequences based on facial data is not only efficient but also yields a slight increase in accuracy compared to the traditional "Divided Space-Time Attention".

Experimental Results

We evaluate the FMeshformer on the popular facial-expression related video dataset: DAiSEE [25]. We adopt the TimeSformer architecture [11], which was also adopted from the "Base" ViT architecture [12] pretrained on either the ImageNet-21K or ImageNet-1K [26]. We use clips of size 224 × 224, with a sample of 16 frames. The patch size is maintained as 16 × 16 pixels. We apply *k-fold* cross-validation with 5 folds. Thus, the results were averaged from the 5 folds and evaluated with accuracy and receiver operating characteristic (ROC) curve shown in Figure 2. We adopt the classification level of two (2), being Engaged and Disengaged, based on the work by Malekshahi et al. [27].

Ablation Studies

We experimented with different patch sizes, *P*. We observed that increasing the patch size to 32 results in a decline in performance, a phenomenon also observed in TimeSformer [1]. We therefore trained the model on $P = 16$ and did not train the model on *P* values lower than 16, as these would be computationally intensive. We also experimented with frame sizes larger than 224x224, which affected the computational cost; therefore, we maintained the original configuration with the number of frames for each video clip set to 16.

Experimental Results

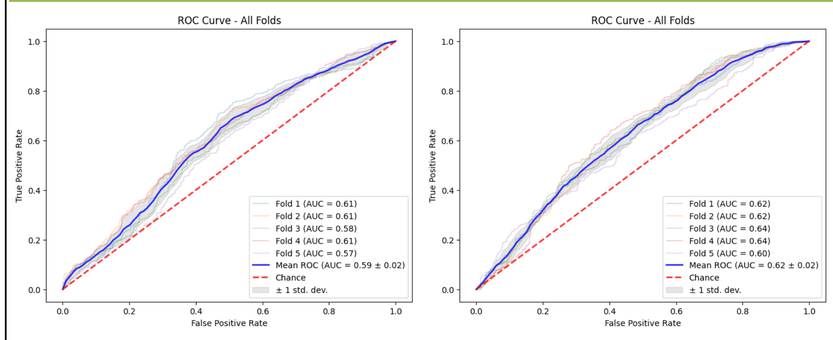

Figure 2. ROC Comparison of TimeSformer with FMeshformer.

Table 1: Result Comparison with State-of-the-art Models

Reference	Accuracy	Method
Liao et al. [13]	58.84	SE-ResNet50 + LSTM with Global Attention
Ma et al. [28]	61.3	Neural Turing Machine
Abedi and Khan [15]	61.15	ResNet and LSTM
Abedi and Khan [15]	63.9	ResNet and TCN
Hu et al. [29]	63.9	ShuffleNet v2
Selim et al. [30]	66.39	Bi-LSTM and TCN
Selim et al. [30]	67.48	EfficientNet B7 and LSTM
TimeSformer	**71.00**	**TimeSformer**
Proposed Model	**73.00**	**FMeshformer**

Conclusions / Recommendations

We introduced the FMeshformer, a domain-specific transformer for video modelling based on mesh positional encoding , utilising facial landmark coordinates as the geometric layout to influence the modelling, and compared our work with the TimeSformer and other convolution-based video networks. We demonstrate that mesh positional encoding based on facial landmark coordinates is efficient for modelling video sequences based on facial expressions. Our method is (1) geometric-aware, (2) achieves state-of-the-art results on engagement detection benchmarks, and (3) requires minimal training and computational cost. We plan to expand our method to other video datasets based on facial expression analysis, specifically those with balanced data.

References

[8] Z. Teed and J. Deng, "RAFT: recurrent all-pairs field transforms for optical flow," in Computer Vision - ECCV 2020 - 16th European Conference, Proceedings, Part II, Glasgow, UK, 2020.

[9] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019.*

[10] G. Bertasius and L. Torresani, "Classifying, segmenting, and tracking object instances in video with mask propagation," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* Jun. 2020.

[11] G. Bertasius, H. Wang, and L. Torresani, "Is Space-Time Attention All You Need for Video Understanding?," Jun. 2021, [Online]. Available: http://arxiv.org/abs/2102.05095

[12] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," Oct. 2020, [Online]. Available: http://arxiv.org/abs/2010.11929

[13] 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, Oct. 2021, doi: 10.1007/s10489-020-02139-8.

[15] A. Abedi and S. S. Khan, "Improving state-of-the-art in Detecting Student Engagement with Resnet and TCN Hybrid Network," *ArXiv*, 2021, [Online]. Available: http://arxiv.org/abs/2104.10122

[16] S. Mandia, K. Singh, and R. Mitharwal, "Recognition of student engagement in classroom from affective states," *Int J Multimed Inf Retr*, vol. 12, no. 2, Dec. 2023, doi: 10.1007/s13735-023-00284-7.

[17] S. Mandia, K. Singh, R. Mitharwal, F. Mushtaq, and D. Janu, "Transformer-Driven Modeling of Variable Frequency Features for Classifying Student Engagement in Online Learning," Feb. 2025, [Online]. Available: http://arxiv.org/abs/2502.10813

[18] S. Mandia, K. Singh, and R. Mitharwal, "Vision Transformer for Automatic Student Engagement Estimation," in *5th IEEE International Image Processing, Applications and Systems Conference, IPAS 2022, Institute of Electrical and Electronics Engineers Inc., 2022*. doi: 10.1109/IPAS55744.2022.10052945.

[19] O. Moutik et al., "Convolutional Neural Networks or Vision Transformers: Who Will Win the Race for Action Recognitions in Visual Data?," Jan. 01, 2023, MDPI. doi: 10.3390/s23020734.

[20] Z. Liu et al., "Video Swin Transformer," Jun. 2021, [Online]. Available: http://arxiv.org/abs/2106.13230

[21] R. Girdhar, J. Carreira, C. Doersch, and A. Zisserman, "Video action transformer network," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019*.

[24] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "ViViT: A Video Vision Transformer," Nov. 2021, [*Online*]. Available: http://arxiv.org/abs/2103.15691

[25] H. Fan et al., "Multiscale Vision Transformers," Apr. 2021, [Online]. Available: http://arxiv.org/abs/2104.11227

[26] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," 2018.

[27] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer Normalization," Jul. 2016, [Online]. Available: http://arxiv.org/abs/1607.06450

[28] A. Gupta, A. D. ' Cunha, K. Awasthi, and V. Balasubramanian, "DAISEE: Dataset for Affective States in E-Learning Environments DAiSEE: Towards User Engagement Recognition in the Wild," 2016, doi: 10.48550/arXiv.1609.01885.

[29] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255. doi: 10.1109/CVPR.2009.5206848.

[30] S. Malekshahi, J. M. Kheyridoost, and O. Fatemi, "A General Model for Detecting Learner Engagement: Implementation and Evaluation," May 2024, [Online]. Available: http://arxiv.org/abs/2405.04251