Decoupling Spatial and Temporal Modeling in 3D Facial Expression Recognition

Bouzid Hamza, Lahoucine Ballihi

Decoupling Spatial and Temporal Modeling in 3D Facial Expression Recognition

Hamza Bouzid *
LRIT-CNRST URAC 29
Mohammed V University in Rabat
Faculty Of Sciences, Morocco
hamza-bouzid@um5r.ac.ma

Lahoucine Ballihi
LRIT-CNRST URAC 29
Mohammed V University in Rabat
Faculty Of Sciences, Morocco
lahoucine.ballihi@fsr.um5.ac.ma

Abstract

Recent advancements in 3D facial expression recognition have revolutionized non-verbal human communication analysis. Our work focuses on dynamic facial expression recognition through 3D mesh sequences. Unlike conventional methods, which either rely on hand-designed feature descriptors or project the faces to 2D domain, we directly extract spatio-temporal information from these meshes using a spatial auto-encoder and a temporal transformer for classification. We evaluate our method on the MUG and BU-4DFE databases, showing promising results.

1 The proposed model

We propose a 3D FER model directly operating on facial meshes, unlike methods that rely on hand-designed descriptor-based, projection-based, and landmarks-based. The proposed method decouples spatial and temporal modeling, utilizing a spiral auto-encoder to capture task-free spatial information and a transformer to model the temporal context, leading to the final facial expression prediction.

- **Pre-processing:** We first convert the irregular unstructured input data (3D scans and 2D images) into meshes with a consistent vertex count and topology. To achieve this, we use FLAME model (Li et al., 2017).
- **Spiral Auto-Encoder:** We utilize spiral convolutions, proposed in (Bouritsas et al., 2019), as a building block for our fully differentiable auto-encoder. Spiral convolutions follow a spiral path, along the mesh surface, considering the connectivity information provided by the mesh structure to define a convolutional kernel that can capture local and global features on the mesh. This enables the encoder to generate hierarchical representations E_i, and the decoder to reconstruct a mesh $\bar{M}_i$ similar to the input. The auto-encoder is trained using the L1 loss.
- **Temporal Transformer:** After encoding all the meshes in the sequence, we use a transformer (Vaswani et al., 2017) that applies self-attention, inspired by its success in natural language processing and computer vision, to capture temporal context in mesh sequences. Self-attention enables parallel processing and long-range dependency modeling in sequence data.

1.1 Experimental Results

- **Databases:** MUG(Aifanti et al., 2010) and BU-4DFE(Zhang et al., 2013) databases.
- **Recognition rates Results:** In the MUG database, FLAME registration yields smooth transitions between expressions, achieving a 91.07% recognition accuracy. Notably, "happiness" and "disgust" expressions score high (100% and 97%), while "anger" lags slightly at 85%. In contrast, FLAME registration applied to the BU-4DFE dataset leads to substantial errors, resulting in a lower overall classification rate of 67.64%. (See Fig.1 for visual representation).
- **Memory Consumption results:** We demonstrate the efficiency of our model on variable-length sequences from the MUG dataset, showing consistent recognition rates with memory usage ranging from 1804+115MB for 16 frames to 1804+220MB for 128 frames (batch size = 16).

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

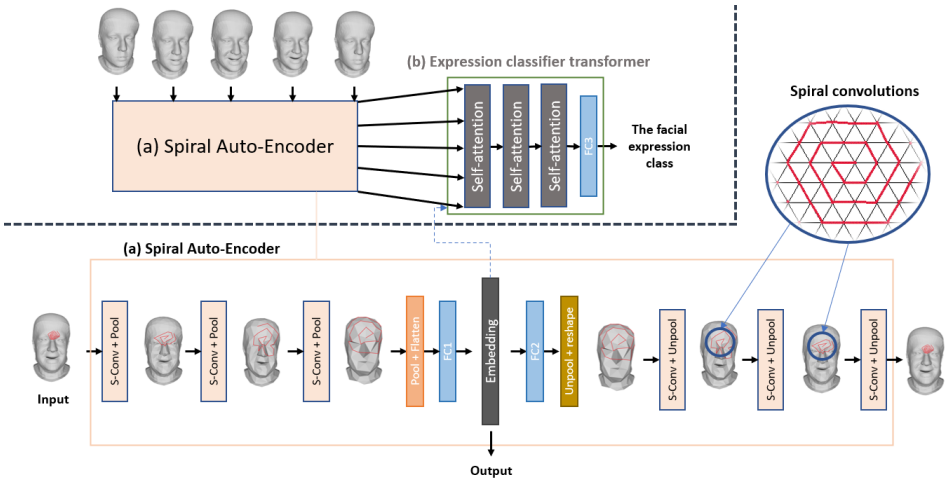

Figure 1: The overview of the proposed model for 3D facial expression recognition.

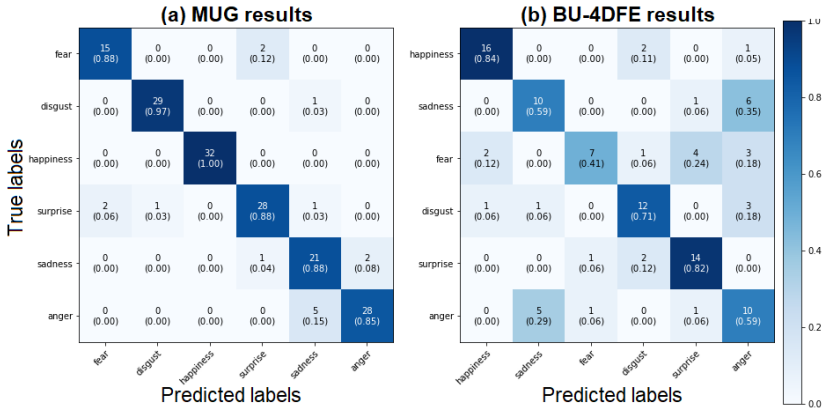

Figure 2: Confusion matrices for our model on the (a) MUG and (b) BU-4DFE test sets.

2 Conclusion

In conclusion, our model exhibits promise in 3D facial expression recognition, efficiently handling mesh data and recognizing expressions across varying frame counts. It offers efficient memory utilization and versatile feature extraction potential but is sensitive to data quality and preprocessing challenges, therefore accurate data registration is crucial for its effectiveness. As future work, we aim to enhance preprocessing to handle noisy data and minimize artifacts, and we plan to extend it to other recognition tasks, such as human action recognition.

References

Aifanti, N., Papachristou, C., and Delopoulos, A. (2010). The mug facial expression database. In *11th International Workshop on Image Analysis for Multimedia Interactive Services WIAMIS 10*, pages 1–4. IEEE.

Bouritsas, G., Bokhnyak, S., Ploumpis, S., Bronstein, M., and Zafeiriou, S. (2019). Neural 3d morphable models: Spiral convolutional networks for 3d shape representation learning and generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7213–7222.

Li, T., Bolkart, T., Black, M. J., Li, H., and Romero, J. (2017). Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6):194–1.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Zhang, X., Yin, L., Cohn, J. F., Canavan, S., Reale, M., Horowitz, A., and Liu, P. (2013). A high-resolution spontaneous 3d dynamic facial expression database. In *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, pages 1–6. IEEE.