Representation Learning for Topology-adaptive Micro-gesture Recognition and Analysis

Atif Shah[1], Haoyu Chen[1] and Guoying Zhao[1,*]

[1]Center for Machine Vision and Signal Analysis (CMVS), University of Oulu, Finland

Abstract

Human-to-human communication is greatly influenced by micro-gestures. The actions of a person inherently reveal information about their true sentiments and potential intentions. Micro-gestures are non-verbal cues that indicate a person's true feelings and intentions; however, they become more challenging to recognize than normal gestures because micro-gestures are subtle and appear for milliseconds. In this work, we propose a graph-encoding convolutional network to extract intrinsic joint representations from skeletons using a self-attention graph convolution module in the spatial domain. The multi-scale temporal convolution module extracts the temporal representation in the time domain and sends it to the classification module to recognize micro-gestures. We evaluate the proposed framework using two micro-gesture datasets, SMG and iMiGUE, and achieve state-of-the-art results.

Keywords

Micro-gestures (MG), Skeleton-based-method, Graph convolutional networks, Emotion recognition

1. Introduction

Micro-gestures (MGs) play an important role in human-to-human communication [1, 2, 3]. The person's acts aid in understanding by naturally revealing information such as actual feelings and possible intentions. MGs are subtle movements that reveal a person's actual emotions and intentions. Recently, there has been a lot of interest in empowering intelligent machines with the same capabilities for understanding human behaviors [4, 5], which is essential for natural human-computer interaction and many other useful applications [6].

In recent times, modern sensor technology and human position estimation algorithms have made it considerably simpler to extract human 2D and 3D skeletons [7, 8]. Skeletons are tempting for MG recognition tasks because they are small in size, robust, and resistant to changes in viewpoint and cluttered backgrounds [9, 10]. Skeletons are commonly used for action and MG recognition via Graph convolutional networks (GCNs) [11, 12]. GCNs are an ideal technique for extracting topological data from skeletons due to their lightweight nature and the fact that joints and bones naturally form graphs in the human body.

However, there are still some limitations with skeleton data. One notable issue is the lack of critical interactive elements and contextual information within the skeleton representation, which makes it difficult to distinguish between comparable activities.

MiGA@IJCAI23: International IJCAI Workshop on Micro-gesture Analysis for Hidden Emotion Understanding, August 21, 2023, Macao, China.

*Corresponding author.

atif.shah@oulu.fi (A. Shah); chen.haoyu@oulu.fi (H. Chen); guoying.zhao@oulu.fi (G. Zhao)

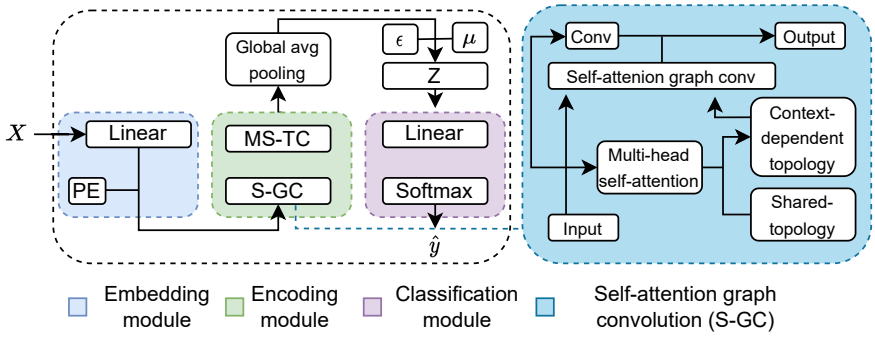


Figure 1: The proposed GE-CN framework architecture

To address the aforementioned issue, we present a Graph-Encoding Convolutional Network (GE-CN) with an attention layer that can learn spatial-temporal contextual representations to recognize MGs. The proposed framework consists of the embedding, encoding and classification module as shown in Figure 1. The embedding module takes input skeleton sequences, integrates them with the position embedding of joints and passes the skeleton representation for feature extraction to the embedding module. The embedding module consists of a self-attention module that extracts spatial representation and captures intrinsic relationships among joints in the spatial domain and a multi-scale temporal module to extract temporal features. The learned latent representations are forwarded to the classification module to recognize the learned gesture categories. In summary, we have the following contributions:

- We proposed a Graph-Encoding Convolutional Network (GE-CN) framework for various skeleton topologies using a self-attention graph convolution.
- We analyze the insight of latent space encoding by using the distribution visualization to find the semantic features.
- We achieved state-of-the-art results on the two datasets.

2. Related work

Recently many researchers have employed skeleton-based methods for action and gesture recognition, which are classified as unsupervised and supervised methods.

2.1. Skeleton-based unsupervised methods

Zheng et al. [13] utilized unsupervised representation learning to capture global motion dynamics. The generative adversarial network (GAN) is used as an encoder-decoder to model motion dynamics and acquire discriminative features for action recognition. Similarly, Su et al. [14] leverage encoder-decoder recurrent neural networks (RNN) for learning features for action recognition. Lin et al. [15] used an unsupervised Bidirectional-Gated Recurrent Unit (Bi-GRU) encoder to learn more generalized representations by combining jigsaw puzzles,

motion prediction, and contrastive learning. Li et al. [16] utilized data without labels for learning view-invariant action and predicting 3D motion by combining RGB and depth images. Many researchers have also utilized contrastive learning; likewise, Zhou et al. [17] added a contrastive learning module to the framework to distinguish between confident and ambiguous action samples. Lin et al. [18] Proposed actionlet-based contrastive learning method where a motion-adaptive transformation strategy was designed to learn semantic consistency in actionlet regions.

2.2. Skeleton-based supervised methods

In skeleton data, joints and bones naturally form graphs in the human body; therefore, most researchers adopted GCNs. Yan et al. [11] presented a spatial-temporal GCN (STGCN) to capture complex features from a human skeleton. Long short-term memory (LSTM) was used by Si et al. [19] with the conjunction of convolutional networks to improve performance by employing more discriminatory spatial and temporal features. Liu et al. [20] proposed a multiscale spatial graph convolutional operator (MSG3D) to disentangle the skeleton, which removes redundant features from the skeleton and aggregates effective graph relationships. Shi et al. [21] proposed a two-stream adaptive graph convolutional network (2s-AGCN) for action recognition. Chi et al. [22] presented the Infogcn framework, which consists of self-attention-based graph convolution and representation learning to capture complementary features for action recognition.

3. Methodology

3.1. Framework Architecture

The GE-CN framework contains an embedding module followed by a stack of encoding modules and a global average pooling layer and a classification module. The embedding module takes a sequence of skeleton representation, converts it to initial joint representation and forwards it to the encoder module, to capture complex spatio-temporal features. A reparametrization step is employed that is commonly used in variational autoencoders [23]. A random noise (ϵ) is introduced, followed by a normal distribution. The mean (μ) value is then added to the product of a diagonal covariance matrix to sample a variable named Z. This approach enables us to find unbiased gradients and efficiently tune the model via gradient-based optimization strategies. A classifier is added at the end with a single linear layer and soft-max function to convert the learned features into class distribution categories.

3.2. Embedding module

The skeleton is represented as a graph $G(V, E)$, with N joints denoted as V and bones as edges E. Edges are represented as $N \times N$ dimensional adjacency matrix A, where element $A_{i,j}$ reflects the link between joints i and j, $A_{i,j}$ value is 1 if i and j joints are connected, otherwise 0. The combination of skeletons makes a sequence of skeleton graphs, which is represented as a joint feature $X \in R^{T \times N \times C}$, where T shows the number of frames and C denotes the feature dimensions.

The input joint representations are linearly transformed by the embedding module into $D^{(0)}$ dimensional vectors while adding the positional embedding (PE) to accommodate the joint position information. The PE is transferable and learnable across many temporal channels.

$$\psi_t^{(0)} = Lin(X_t) + PE \tag{1}$$

where $\psi_t^{(0)}$ represents a hidden layer, $PE \in R^{N \times D^{(0)}}$, $Lin(.)$ is a linear function and t is time index.

3.3. Encoding module

The encoder module consists of two modules: Self-attention Graph Convolution (S-GC) to extract spatial representations and Multi-Scale Temporal Convolution (MS-TC) to extract temporal features from the skeleton. The input joints are sequentially encoded via S-GC and MS-TC, followed by normalization. The graph convolution updates the hidden representation via the following rules:

$$\psi_t^{(l+1)} = \Theta(\bar{A}\psi_t^{(l)}W^{(l)}) \tag{2}$$

where $\bar{A}$ is the normalized adjacency matrix, W is the learned matrix and $\Theta(.)$ is the nonlinear activation function. The S-GC uses the self-attention of joint features to determine the intrinsic topology and employs the topology as a source of neighborhood vertex information for graph convolution. The self-attention is a mechanism that links several joints of the body. S-GC takes into account all the possible relations and estimates positive and bounded weights which are known as self-attention maps that indicate how strong the connection is between joints. To get self-attention maps the following mathematical definition is used:

$$S(\psi_t) = softmax(\frac{\psi_t W_k (\psi_t W_Q)^T}{\sqrt{D'}}) \tag{3}$$

where W_K, W_Q are learnable matrics of $D^{'}$ dimensions. Aside from the self-attention map, the S-GC learns a topology $\ddot{A}$ shared over time. The self-attention map and shared topology employ M multi-heads to enable the model to attend multiple representation subspaces at the same time. To achieve intrinsic topology, the shared topology and self-attention maps are integrated.

$$\ddot{A}_m \otimes S_m(\psi_t) \in R^{T \times N \times N} \tag{4}$$

where $\otimes$ is the broadcasted element-wise product. S-GC utilizes the equation 4 as neighborhood information, and the overall joint representation update rules are formulated as [22]

$$\psi_t^{(l+1)} = \Theta\left(\sum_m^M (\ddot{A}_m^{(l)} \otimes S_m(\psi_t^{(l)}))\psi_t^{(l)} W_m^{(l)}\right) \tag{5}$$

After extracting the spatial intrinsic features the MS-TC block extracts the temporal features using three parallel convolution branches with various kernel sizes.

Table 1
The top-1% and top-5% accuracy of SMG and iMiGUE datasets with baseline results

Method	Self-attention	Dataset	Accuracy	
			Top-1%	Top-5%
Baseline	✗	iMiGUE	53.98	90.5
		SMG	62.89	94.95
GE-CN	✓	iMiGUE	56.12	90.01
		SMG	64.26	95.23

4. Experiments

The proposed GE-CN framework is evaluated using two MG datasets. Both datasets contain skeleton data, which is utilized in this work. In this work, we used an SGD optimizer with a 0.9 moment coefficient, a frame size of 64 and a batch size of 32. We used maximum-mean discrepancy loss borrowed from [22] and used the same experimental settings.

4.1. Datasets

Spontaneous Micro-Gesture (SMG) dataset [3, 1] contains 3,692 samples with 17 MGs. The dataset is obtained from 40 individuals using Kinect [8] and 25 3D joints are collected while they are narrating a fake and true story.

Micro-Gesture Understanding and Emotion Analysis (iMiGUE) dataset [2] contains 32 MGs extracted from the post-match press conference videos. The sample is collected in RGB and skeleton modalities using OpenPose [7]. A total of 137 key points are collected, including 70 face points, 42 hand points, and 25 body joints. Following the protocol of [11, 20, 24], we use the 3rd dimension of the OpenPose joint as a pseudo dimension.

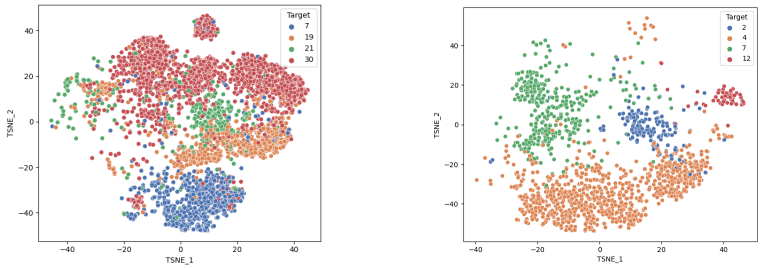
4.2. Results

We evaluated the proposed framework using two MG datasets and the results are shown in Table 1. We compare the results with the baseline method without a self-attention layer and use accuracy as a metric of evaluation, with top-1% and top-5% accuracy. The first row of Table 1 shows the baseline results on both datasets without a self-attention layer. Using the baseline, the iMiGUE dataset achieved a top-1 accuracy of 53.98% and a top-5 accuracy of 90.5%. Similarly, the SMG achieved 62.89% and 94.95% accuracies of top-1 and top-5, respectively. The second row shows the proposed method with a self-attention layer, where the iMiGUE dataset reached top-1 accuracy of 56.12 and top-5 accuracy of 90.01%, which is a significant improvement as compared to the baseline. Likewise, the SMG also improved its accuracy and reached 64.26 and 95.23, top-1 and top-5 accuracies, respectively. Both datasets show that the GE-CN method improved the results notably. One of the reasons SMG results show better performance is because most of the person's action skeleton data is available with full-body joints; however, for the iMiGUE dataset, only the upper body joints are extracted because almost all of those skeletons are extracted while the person is sitting on a chair.

We compare the results with previous methods, as shown in Table 2. We achieved the best

Table 2
Comparison with other methods using iMiGUE and SMG datasets

Method	iMiGUE dataset		SMG dataset	
	Top-1%	Top-5%	Top-1%	Top-5%
ST-GCN [11]	46.97	84.09	41.4	86.07
2S-GCN [21]	47.78	88.43	43.11	86.90
Shift-GCN [25]	51.51	88.18	55.31	87.34
GCN-NAS [26]	53.90	89.21	58.85	85.08
MS-G3D [20]	54.91	89.98	**64.75**	91.48
TRN [27]	55.24	89.17	-	-
TRN [28]	-	-	59.51	88.53
GE-CN (our)	**56.12**	**90.01**	64.26	**95.23**



(a) iMiGUE dataset (b) SMG dataset

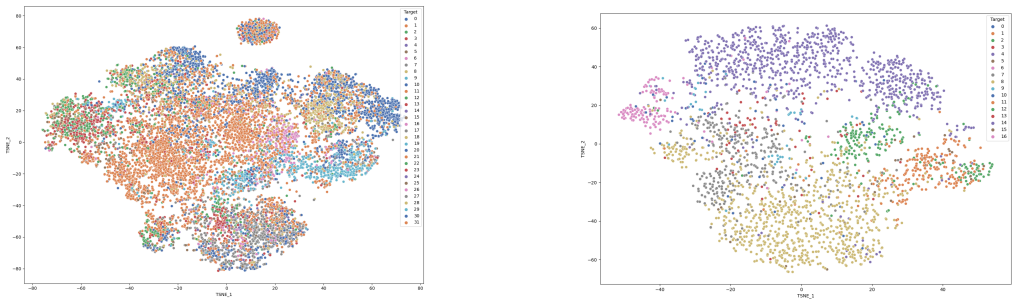
Figure 2: Latent features visualization of four categories for each dataset

results both on the iMiGUE and SMG datasets. If we look at the MG-G3D method in the 5^{th} row, we improved the performance significantly using the iMiGUE dataset and the top-5% accuracy using the SMG dataset; however, we didn't achieve the best results in the top-1% accuracy on the SMG dataset and placed slightly below the MS-G3D method.

We visualized the latent features of all categories of iMiGUE and SMG datasets, as shown in Figure 3, but for better visualization, we only visualized four categories from each dataset in Figure 2. Both images show the clusters of classes in different color schemes. The target legend reflects the number of classes in a dataset. The visualization was done using the T-SNE tool. We transformed the high-dimensional latent features to two dimensions and used one epoch samples for better visualization.

5. Conclusion

In this work, we introduced a graph-encoding convolutional network that extracts intrinsic joint representation from skeleton data using a self-attention module in the spatial domain. The extracted features are forwarded to the multi-scale temporal convolution module to extract the temporal join relationship. The learned features are fed to the classification module for micro-gesture classification. The proposed framework was evaluated on two micro-gesture



(a) iMiGUE dataset (b) SMG dataset

Figure 3: Latent features visualization of all categories of each dataset

datasets, SMG and iMiGUE, which achieved state-of-the-art results with top-1% accuracy of 64.26 and 56.12, respectively.

Acknowledgments

This work was supported by the Academy of Finland for Academy Professor project EmotionAI (grants 336116, 345122), the University of Oulu & The Academy of Finland Profi 7 (grant 352788), by the Ministry of Education and Culture of Finland for AI forum project.

The authors also wish to acknowledge CSC – IT Center for Science, Finland, for computational resources.

References

[1] H. Chen, H. Shi, X. Liu, X. Li, G. Zhao, Smg: A micro-gesture dataset towards spontaneous body gestures for emotional stress state analysis, International Journal of Computer Vision 131 (2023) 1346–1366.

[2] X. Liu, H. Shi, H. Chen, Z. Yu, X. Li, G. Zhao, imigue: An identity-free video dataset for micro-gesture understanding and emotion analysis, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 10631–10642.

[3] H. Chen, X. Liu, X. Li, H. Shi, G. Zhao, Analyze spontaneous gestures for emotional stress state recognition: A micro-gesture dataset and analysis with deep learning, in: 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), IEEE, 2019, pp. 1–8.

[4] H. Shi, W. Peng, H. Chen, X. Liu, G. Zhao, Multiscale 3d-shift graph convolution network for emotion recognition from human actions, IEEE Intelligent Systems 37 (2022) 103–110.

[5] H. Chen, H. Tang, Z. Yu, N. Sebe, G. Zhao, Geometry-contrastive transformer for generalized 3d pose transfer, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2022, pp. 258–266.

[6] H. Chen, E. Tan, Y. Lee, S. Praharaj, M. Specht, G. Zhao, Developing ai into explanatory supporting models: An explanation-visualized deep learning prototype, in: 14th International Conference of the Learning Sciences (ICLS) 2020, 2020.
[7] Z. Cao, T. Simon, S.-E. Wei, Y. Sheikh, Realtime multi-person 2d pose estimation using part affinity fields, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 7291–7299.
[8] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, A. Blake, Real-time human pose recognition in parts from single depth images, in: CVPR 2011, Ieee, 2011, pp. 1297–1304.
[9] H. Chen, X. Liu, G. Zhao, Temporal hierarchical dictionary with hmm for fast gesture recognition, in: 2018 24th international conference on pattern recognition (ICPR), IEEE, 2018, pp. 3378–3383.
[10] H. Chen, X. Liu, J. Shi, G. Zhao, Temporal hierarchical dictionary guided decoding for online gesture segmentation and recognition, IEEE Transactions on Image Processing 29 (2020) 9689–9702.
[11] S. Yan, Y. Xiong, D. Lin, Spatial temporal graph convolutional networks for skeleton-based action recognition, in: Proceedings of the AAAI conference on artificial intelligence, volume 32, 2018.
[12] Y. Liu, X. Zhang, Y. Li, J. Zhou, X. Li, G. Zhao, Graph-based facial affect analysis: A review, IEEE Transactions on Affective Computing (2022).
[13] N. Zheng, J. Wen, R. Liu, L. Long, J. Dai, Z. Gong, Unsupervised representation learning with long-term dynamics for skeleton based action recognition, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 32, 2018.
[14] K. Su, X. Liu, E. Shlizerman, Predict & cluster: Unsupervised skeleton based action recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 9631–9640.
[15] L. Lin, S. Song, W. Yang, J. Liu, Ms2l: Multi-task self-supervised learning for skeleton based action recognition, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 2490–2498.
[16] J. Li, Y. Wong, Q. Zhao, M. S. Kankanhalli, Unsupervised learning of view-invariant action representations, Advances in neural information processing systems 31 (2018).
[17] H. Zhou, Q. Liu, Y. Wang, Learning discriminative representations for skeleton based action recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 10608–10617.
[18] L. Lin, J. Zhang, J. Liu, Actionlet-dependent contrastive learning for unsupervised skeleton-based action recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 2363–2372.
[19] C. Si, W. Chen, W. Wang, L. Wang, T. Tan, An attention enhanced graph convolutional lstm network for skeleton-based action recognition, in: proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 1227–1236.
[20] Z. Liu, H. Zhang, Z. Chen, Z. Wang, W. Ouyang, Disentangling and unifying graph convolutions for skeleton-based action recognition, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 143–152.
[21] L. Shi, Y. Zhang, J. Cheng, H. Lu, Two-stream adaptive graph convolutional networks

for skeleton-based action recognition, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 12026–12035.
[22] H.-g. Chi, M. H. Ha, S. Chi, S. W. Lee, Q. Huang, K. Ramani, Infogcn: Representation learning for human skeleton-based action recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 20186–20196.
[23] D. P. Kingma, M. Welling, Auto-encoding variational bayes, arXiv preprint arXiv:1312.6114 (2013).
[24] A. Shah, H. Chen, H. Shi, G. Zhao, Efficient dense-graph convolutional network with inductive prior augmentations for unsupervised micro-gesture recognition, in: 2022 26th International Conference on Pattern Recognition (ICPR), IEEE, 2022, pp. 2686–2692.
[25] K. Cheng, Y. Zhang, X. He, W. Chen, J. Cheng, H. Lu, Skeleton-based action recognition with shift graph convolutional network, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 183–192.
[26] W. Peng, X. Hong, H. Chen, G. Zhao, Learning graph convolutional network for skeleton-based human action recognition by neural searching, in: Proceedings of the AAAI conference on artificial intelligence, volume 34, 2020, pp. 2669–2676.
[27] B. Zhou, A. Andonian, A. Oliva, A. Torralba, Temporal relational reasoning in videos, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 803–818.
[28] M. Xu, M. Gao, Y.-T. Chen, L. S. Davis, D. J. Crandall, Temporal recurrent networks for online action detection, in: Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 5532–5541.