Relational Edge-Node Graph Attention Network for Classification of Micro-Expressions

Ankith Jain Rakesh Kumar and Bir Bhanu
Department of Electrical and Computer Engineering
University of California, Riverside
arake001@ucr.edu, bhanu@ece.ucr.edu

Abstract

Facial micro-expressions (MEs) refer to subtle, transient, and involuntary muscle movements expressing a person's true feelings. This paper presents a novel two-stream relational edge-node graph attention network-based approach to classify MEs in a video by selecting the high-intensity frames and edge-node features that can provide valuable information about the relationship between nodes and structural information in a graph structure. The paper examines the impact of different edge-node features and their relationships on the graphs. The first step involves extracting high-intensity-emotion frames from the video using optical flow. Second, node feature embeddings are calculated using the node location coordinate features and the patch size information of the optical flow across each node location. Additionally, we obtain the global and local structural similarity score using the jaccard's similarity score and radial basis function as the edge features. Third, a self-attention graph pooling layer helps to remove the nodes with lower attention scores based on the top-k selection. As the final step, the network employs a two-stream edge-node graph attention network that focuses on finding correlations among the edge and node features, such as landmark coordinates, optical flow, and global and local edge features. A three-frame graph structure is designed to obtain spatio-temporal information. For 3 and 5 expression classes, the results are compared for SMIC and CASME II databases.

1. Introduction

The rapid progress in artificial intelligence, particularly in computer vision, machine learning, and deep learning, has made human-computer interaction an important topic of research, thereby creating a form of augmented reality. In human-computer interaction, facial expression analysis is a crucial area of research, among numerous other tasks.

Humans have various ways of expressing their emotions and thoughts. Verbal communication and facial expressions are two common ways of communicating between humans and machines. Facial expressions are a fundamental aspect of nonverbal communication, conveying a wide range of emotions and attitudes. However, not all facial expressions are easily observable or consciously controlled. Micro-expressions (MEs) are a type of facial expression that occurs involuntarily and last only a fraction of a second. The time frame of MEs is less than 0.6s [1], [2]. These expressions are often subtle and difficult to detect, but they can reveal important information about a person's inner emotional state, intentions, and attitudes. MEs are useful in various applications, including lie detection, online learning, security, healthcare, and online gaming. Therefore, developing a system to recognize and classify MEs is crucial.

Classifying MEs poses a significant challenge due to three primary characteristics of MEs: (i) they are subtle and brief in nature, (ii) they occur spontaneously, leading to ephemeral changes in facial muscle movements, and (iii) transient nature. Furthermore, a significant obstacle in ME spotting and classification tasks is the availability of sufficient and well-balanced training data samples.

In recent years, MEs spotting and classification have become increasingly significant within the computer vision community. To recognize MEs, researchers have employed hand-crafted approaches, such as LBP [3], Bi-WOOF [4], LBP-TOP [5], optical flow and optical strain, and 3DHOG [6] to extract the spatio-temporal information to recognize MEs. However, traditional techniques are not sufficient to capture the subtle changes on the human face that are characteristic of MEs. Recent advancements in computer vision and deep learning have allowed researchers to use CNNs and GNNs to extract subtle changes on the face as spatio-temporal features, improving the accuracy of micro-expression recognition (MER) tasks.

Our proposed solution aims to address the aforementioned critical issues by introducing a new approach called the Relational Edge-Node Graph Attention (ENGAT) Network with a self-attention graph pooling layer (SAGPOOL) to understand the patch, global, and local structural feature

information associated with the graph. The global edge feature information represents the structural similarity between the edges, while the local edge feature embedding represents the feature-based similarity between 2 nodes. Additionally, we use node location coordinate features and optical flow patch size information across each node location coordinate in a graph to enhance the global-local edge features. The relational edge-node featured graph attention network (ENGAT) takes into consideration node coordinate points, patch size, and global and local structural feature information to better understand the face-structured graph. We use a SAGPOOL to assign a confidence score for each node, which enables us to identify the top-k nodes to retain in the final graph structure. To calculate the spatio-temporal information, we employ a three-frame graph structure. We use a two-stream ENGAT network with a SAGPOOL model to extract the correlation between the essential edge-node features. To reduce the total count of low-intensity video frames, we utilize a frame selection approach using the optical flow method. Furthermore, we balance the training data samples by augmenting them using multiple amplification factors of the Eulerian Motion Magnification (EMM) [7] method for the category of expression with the least video data. We conducted an ablation analysis to assess the importance of our approach and performed cross-dataset experiments to validate its effectiveness.

The following is the structure of this paper: In Section 2, we present our contributions along with the related works. In Section 3, we provide a detailed explanation of our technical approach. In Section 4, we present the results of our qualitative and quantitative experiments, which include the results of our ablation study. Lastly, in Section 5, we summarize our findings and discuss future work.

2. Related Work and Contributions

2.1. Related Work

Over the past decade, there has been a growing interest in micro-expression recognition (MER) among computer vision researchers. Prior to meaningful attribute extraction, an initial processing stage must be completed, which includes tasks such as image resizing, registration, video amplification, and frame selection methods. These pre-processing techniques are essential for preparing the data to ensure that it is suitable for subsequent analysis.

There are several methods used for classifying MEs into different types of emotions. The first technique is handcrafted feature extraction, as shown in Table 1. This method involves manually extracting features from the data, such as LBP, Bi-WOOF, LBP-TOP, optical flow and optical strain, and 3DHOG. While this method has been widely used in the past, it has several limitations, such as low accuracy and the inability to capture subtle changes in the human face.

The second method for attribute extraction is to use CNNs, as shown in Table 2. The third method for attribute extraction is to use GNNs, as described in Table 3. In recent years, CNNs and GNNs have been increasingly used for the attribute extraction process of ME videos, as they are more decisive and outperform handcrafted approaches for classification tasks.

2.2. Contributions

The contributions of this work are:

- We present an automatic landmark-aided two-stream Relational Edge-Node Graph Attention Network (ENGAT) with a self-attention graph pooling, that incorporates both edge and node features.
- We select the global and local structural similarity edge feature embedding which provides additional information about the relationship between nodes and structural information in a graph and enhances our ability to comprehend and differentiate between various expressions.
- We conduct a thorough evaluation of our approach using two publicly available datasets, SMIC, and CASME II, for 3 and 5 categories of MEs. We conducted cross-dataset experiments to assess the generalization of our method.

3. Technical Approach

The proposed method for classifying MEs is depicted in Figure 1. Initially, we amplified the input video using Eulerian Motion Magnification (EMM) [7] and extracted the magnified input videos. To identify high-intensity emotion frames, we employed a threshold value based on the optical flow magnitude and excluded the low-intensity emotion frames using the method proposed in [24]. Next, we used the dlib software [25] to obtain 51 landmark points on the face. To effectively capture subtle changes in the optical flow magnitude components, we chose a patch size for each landmark coordinate point. To improve our understanding and ability to distinguish between different expressions, we developed a two-stream edge-node graph attention network (ENGAT). This network takes into account several important features, including landmark coordinate points, fixed patch size of optical flow magnitude, and global and local edge structural similarity score edge features. By integrating these features, the network can better capture the relationship between the edge-node features in a given video. This approach allows us to classify MEs into different categories. We used a three-frame graph structure to capture the spatio-temporal features from the video. Ultimately, we classified the MEs into different types of emotions using a relational two-stream ENGAT and a SAGPOOL layer.

Table 1. Research studies focusing on the use of handcrafted features for classifying MEs

Author	Video/ Image Frames	Attributes Extractor	Classifier
Huang *et al.* [8]	Video	STLBP-IP	SVM
Saeed *et al.* [9]	Video	LGBP + LBP-TOP	SVM
Lu *et. al.* [10]	Video	Delaunay-based temporal coding model	RF, SVM
Guo *et al.* [11]	Video	CBP-TOP	ELM
Liong *et al.* [6]	Video	Optical strain	SVM
Donia *et al.* [12]	Video	HOG	SVM
Oh *et al.* [13]	Video	Riesz wavelet transform	SVM

Table 2. Research studies focusing on the use of CNN features for classifying MEs

Author	Video/ Image Frame	Attributes Extractor	Classifier
Gan *et al.* [14]	Onset + Apex	Optical Flow + CNN	MLP
Choi *et al.* [15]	Video	CNN-LSTM	MLP
Khor *et al.* [16]	Video	Optical flow + CNN-LSTM	SVM
Kumar *et al.* [17]	Video	CNN, CNN-LSTM, 3DHOG	SVM, MLP
Khor *et al.* [18]	Video	2S-CNN	MLP
Song *et al.* [19]	Onset, Apex and Offset	3S-CNN	MLP
Wang *et al.* [20]	Video	Optical flow + Contrastive Learning	MLP
Guo *et al.* [21]	Video	3DCNN + Multi-scale Local Transformer	MLP
Thuseethan *et al.* [22]	Video	3DCNN + ANN	MLP
Yang *et al.* [23]	Video	AU + Optical flow + CNN	MLP

3.1. Edge-Node Graph Attention Network

The Graph Attention Networks (GAT) proposed in [26] utilizes all node attributes and shares them with neighboring nodes. However, in our proposed relational edge-node graph attention network (ENGAT) model, we use both node feature and edge feature embeddings. This allows for a better correlation between the nodes and edges, which indeed helps in obtaining a good attention score for the nodes and edges and helps in understanding the structured graph of the face for the task of ME classification.

Consider a graph with N number of nodes, with node features, $\mathbf{h} = \{\vec{h}_1, \vec{h}_2, \vec{h}_3, ..., \vec{h}_N\}$, $\vec{h}_i \in R^T$, where T is the total count of node attributes in each node, and edge attribute features are denoted by $\vec{f}_{ij}$. A graph convolutional layer then computes a set of new node attributes as its output $\mathbf{h'} = \{\vec{h}'_1, \vec{h}'_2, \vec{h}'_3, ..., \vec{h}'_N\}$.

Initially, the graph convolutional layer applies a learnable linear transformation with a parameterized weight matrix $\mathbf{W}$ to each node and edge to obtain a higher-level transformation of node features and edge features. Then, a self-attention mechanism is applied to the node and edge using an attentional mechanism a, which is shared to calculate attention coefficients through equation (1).

$$e_{ij} = a\left(\mathbf{W}\vec{h}_i, \mathbf{W}\vec{h}_j, \mathbf{W}\vec{f}_{ij}\right) \tag{1}$$

which describes the significance of the node j's attributes to node i, and the importance of each edge features $\vec{f}_{ij}$. e_{ij} are calculated only for the nodes having neighborhood nodes and edges. To model coefficients comparable across neighborhood nodes and edges, the softmax function is used to normalize them, shown in equation (2).

$$\alpha_{ij} = \text{softmax}_j\left(e_{ij}\right) = \frac{\exp\left(e_{ij}\right)}{\sum_{k\in\mathcal{N}_i}\exp\left(e_{ik}\right)} \tag{2}$$

where, $\mathcal{N}_i$ represents the neighborhood of node i.

The attention mechanism a is a one-layer feed-forward network, which is parameterized by a weight vector $\vec{a}$. After calculating the normalized attention coefficients (α_{ij}), we apply the LeakyReLU non-linear activation function with a negative slope of 0.2, which expands the coefficients computed by the attention mechanism. The reasons for using LeakyReLU [27] are twofold: (1) it addresses the dying ReLU problem, and (2) it can aid in faster training. This is detailed in equation (3).

$$\alpha_{ij} = \frac{\exp\left(\text{LeakyReLU}\left(\vec{\mathbf{a}}^T\left[\mathbf{W}\vec{h}_i \| \mathbf{W}\vec{h}_j \| \mathbf{W}\vec{f}_{ij}\right]\right)\right)}{\sum_{k\in\mathcal{N}_i}\exp\left(\text{LeakyReLU}\left(\vec{\mathbf{a}}^T\left[\mathbf{W}\vec{h}_i \| \mathbf{W}\vec{h}_k \| \mathbf{W}\vec{f}_{ik}\right]\right)\right)} \tag{3}$$

where T represents the transpose, and $||$ is the concatenation operator.

To get the final attributes for each node, a graph convolutional operator is used for embedding node features and edge features from the neighborhood. After applying the non-linearity function to it, they are aggregated to satisfy the node localization property, as shown in equation (4).

$$\vec{h}'_i = \sigma\left(\sum_{j\in\mathcal{N}_i}\alpha_{ij}\mathbf{W}\vec{h}_j\right) \tag{4}$$

where σ is an activation function. $\vec{h}'_i$ represents the final output feature for every node.

Table 3. Research studies focusing on the use of GNN features for classifying MEs

Author	Video/ Image Frame	Attributes Extractor	Classifier
Lo *et al.* [28]	Video	AU + 3D CNN + GNN	MLP
Xie *et al.* [29]	Video	AU + GCN	MLP
Kumar *et al.* [24]	Frames	Landmark points + Optical flow magnitude + GAT	MLP
Zhou *et al.* [30]	Onset + Apex	Optical flow + AU + GCN	MLP
Kumar *et al.* [31]	High-Intensity Frames	Landmark points + Optical flow magnitude and direction + GAT	MLP

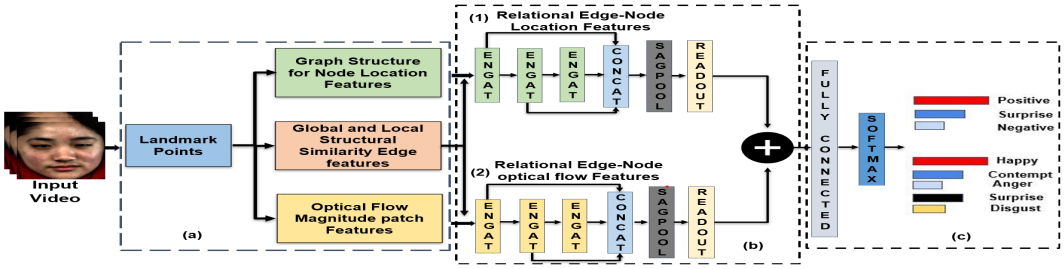


Figure 1. The architecture of our presented technique. (a) First, landmark points are identified, and node location features and the summed magnitude of optical flow node attributes are extracted based on these points. Additionally, global and local structural similarity edge features are extracted, which are shared between both streams. (b) In the first stream, relational edge-node location features are used, while in the second stream, relational edge-node optical flow features are employed with the help of an edge-node graph attention network (ENGAT) and a self-attention graph pooling (SAGPOOL) layer to train the graph structure. (c) Ultimately, the two streams are fused, and the resulting graph representation is used to classify MEs based on the available datasets.

3.1.1 Selection of Node features

The first stream of our graph network uses feature embeddings of location coordinate points as node features. The size of the node attribute vector is 2, equivalent to the x and y coordinate positions across each node. On the other hand, for the second stream, we compute the optical flow information by considering a 10x10 patch size across the respective landmark location coordinates. The node attribute vector size for the second stream is 100.

3.1.2 Selection of Edge features

The process of selecting edge features for a video involves analyzing the face graph structure and considering both the global and local structural similarity scores for each edge. By utilizing both types of scores, the resulting edge features are able to effectively enhance and illustrate the relationships and correlations between various edges within the face graph structure, while also providing important structural information about different parts of the face that are essential for distinguishing between different classes of expressions. Additionally, these edge features enable the calculation of more accurate attention scores for both nodes and edges. The vector size of edge feature embedding is equivalent to the global structural similarity score (GSSS), and the local structural similarity score (LSSS) is shown in the equation 5

$$\vec{f}_{ij} = (GSSS, LSSS) \tag{5}$$

3.1.2.1 Global Structural Similarity Score (GSSS): The Jaccard similarity score index is utilized to calculate the global structural similarity score (GSSS). This index is beneficial in acquiring a comprehensive understanding of the overall graph structure of the face, including how the edges are interconnected and the degree of significance associated with each edge. After the score is obtained, it is normalized using the sigmoid function to ensure it is scaled between 0 and 1. The calculation for GSSS is represented by the equation 6.

$$\text{GSSS} = \text{sigmoid}\left(\frac{N_G(i) \cap N_G(j)}{N_G(i) \cup N_G(j)}\right) \tag{6}$$

where $N_G(i)$ and $N_G(j)$ represent the vectors associated with node i and node j that includes the neighbors of node i and node j. $N_G(i) \cap N_G(j)$ represents the intersection of the common node neighbors and $N_G(i) \cup N_G(j)$ represents the union of neighbors between node i and node j.

3.1.2.2 Local Structural Similarity Score (LSSS): Another significant aspect of the edge feature selection process is the local structural similarity score, which is based on feature similarities. This score is calculated by evaluating the similarity between the features of two nodes, providing insight into the importance of each edge in the graph structure. To calculate the local structural similarity score, a radial basis function is utilized, as shown in equation 7.

$$LSSS = sigmoid\left(\exp\left(-\frac{\left\|\vec{h}_i - \vec{h}_j\right\|^2}{2\gamma^2}\right)\right) \tag{7}$$

where γ represents the learnable parameter.

By analyzing both global and local structural similarity scores, the edge features are able to more effectively enhance the correlation and relationship between different edges within the face graph structure. This ultimately contributes to a more accurate and reliable analysis of the video's content, supporting the identification and classification of different expressions with greater precision.

3.2. Self-Attention Graph Pooling

Graph pooling is a technique that reduces the number of parameters in a network and helps prevent overfitting by retaining only a subset of the input graph nodes. SAGPOOL, as described in the [32] paper, utilizes the GNN network to obtain attention scores that guide the pooling process. A pooling ratio, denoted by $k \in (0, 1]$, is used to determine the number of nodes to be retained in the final graph structure. The SAGPOOL layer first calculates attention scores from the graph attention layer and then selects the top-k nodes based on their attention scores and the selected ratio k. Subsequently, a new feature matrix and adjacency matrix are constructed based on the remaining node ids and their connections, forming a new graph structure.

After the self-attention graph pooling layer selects the necessary nodes and creates a new graph structure, the resulting output is fed into the readout layer [33]. This layer uses global average pooling and global max pooling to generate a fixed-size node feature representation. These pooling techniques aggregate the features of all nodes in the graph and produce a condensed feature vector that can be used for downstream tasks.

3.3. Two-Stream Graph Attention Network

We developed a novel Two-stream Edge-Node Graph Attention (ENGAT) Network that extracts temporal features from the video, as shown in Fig. 1. In our proposed method, we extract node location features, optical flow magnitude attributes, and global and local structural similarity score edge features from the video frames and connect them to form a single graph using a three-frame graph structure.

Our graph network is designed using edge-node graph attention (ENGAT) layers, as shown in Fig. 1. We use three graph attention layers with ReLU activation functions after each layer, 32 hidden channels (hidden channels refer to the intermediate representations or features learned by the network between input and output layers.), and one head for the graph attention layer. In the first stream of the graph network, we use the x and y location coordinates of the landmark points as the node feature vector. This helps capture the change in the movement of each landmark point relative to its previous position. For the second stream, we use a fixed patch size for the optical flow magnitude features. The same edge features are used in both streams of the graph network to provide additional information about the relationship between nodes in the graph, which can improve the accuracy and generalization of graph neural networks (GNNs). The optical flow magnitude component captures spatio-temporal information about the MEs, along with the three-frame graph structure used in our network. The outputs from the three graph attention layers are concatenated and propagated to the self-attention graph pooling layer, which removes less important nodes based on their attention scores and a ratio of k in the *top-k* selection process.

To obtain a fixed-size representation of the output layer, the output of the self-attention graph pooling layer is fed into the readout layer. After passing through the readout layer of each of the two graph networks, the results are concatenated to form the graph representation of the two streams. The output is then fed into a fully connected layer and a softmax layer for classification.

4. Experimental Results

This section presents the results of our experiments, including the datasets used, experimental setup, and details of our approach for classifying micro-expressions (MEs). To thoroughly evaluate our proposed method, we performed a comprehensive study where we removed each component of our approach to assess its impact on the overall performance. To test the robustness and generalizability of our approach, we conducted cross-dataset evaluations for ME classification, ensuring that our method can perform well in different environments and with different subjects.

4.1. Experimental Setup

To evaluate our approach for ME classification, we conducted experiments on two publicly available datasets: CASME II [34] and SMIC [35]. We evaluated our approach on both 3-class and 5-class ME classification tasks using the *leave-one-subject-out* (LOSO-CV) cross-validation approach. All experiments were conducted on a workstation running Ubuntu 20.04 OS with 16GB RAM and 4 NVIDIA GeForce GTX 1080Ti GPUs.

4.2. Datasets and Preprocessing

The *two* datasets used for classifying MEs are: CASME II [34] and SMIC [35]. We are interested in classifying the MEs into 3 and 5 types. LOSO-CV, a subject-independent cross-validation method, was utilized to eliminate subject bias and apply the approach to assess the universal applicability of different methods. The CASME II and SMIC datasets video distributions for the 3 classes are Negative (88 and 70), Positive (32 and 51), and Surprise (25 and 43) videos, respectively. Likewise, for the CASME II 5 classes, the video distributions are Disgust (63), Happy (32), Surprise (25), Repression (27), and Other (99).

We registered each image with an onset image (source frame). The registered images were resized to 256x256. To address the case of unbalanced video data and to further improve the training accuracy, we used data samples from

other databases of the same class to augment the training samples. Additionally, we augmented the video samples with different magnitudes of motion amplification factors ranging from 1 to 5 during training. We used a magnification factor of 4 for testing. The SAGPOOL layer retained 75% of the nodes in the graph structure by using k = 0.75 as the ratio. This helped to maintain important nodes and ensure enough nodes were present in the graph. We used an optimizer called Adam with a learning rate of 0.001, which was decreased by half after every 100 epochs. We used the cross-entropy the loss function.

4.3. Evaluation Metrics

The total number of classes in the two available databases is unbalanced, therefore, to overcome this issue, we use both the unweighted F1 (UF1) score and accuracy as evaluation metrics.

4.3.1 Unweighted F1 score (UF1)

The UF1-score is used for evaluating classification performance on imbalanced datasets because it gives equal importance to each class regardless of their frequencies in the dataset. During the testing phase, this metric is utilized. Using the confusion matrix, we extracted the results, including both correct and error values, and calculated the True Positives (TPc), False Positives (FPc), and False Negatives (FN$_c$) for each class c. To calculate the UF1, we first obtain the F1 score for each class using equation 8, and then take the average across all classes, as shown in equations 9.

$$F1_c = \frac{2 \times TP_c}{2 \times TP_c + FP_c + FN_c} \tag{8}$$

$$UF1 = \frac{F1_c}{C}, \tag{9}$$

where F1$_c$ is F1-score for each individual class, and C is the number of classes.

4.3.2 Accuracy

The accuracy is calculated using the equation 10.

$$Acc = \frac{P}{N} \times 100\% \tag{10}$$

where P is the total number of correct predictions and N is the number of video samples.

4.4. Experimental Results

The results of our proposed approach and state-of-the-art methods for 3 expression categories on CASME II and SMIC datasets are presented in Table 4. We use the LOSO-CV technique, wherein we repeat the experiment N times, each time using data from N-1 subjects for training and the remaining 1 subject for testing.

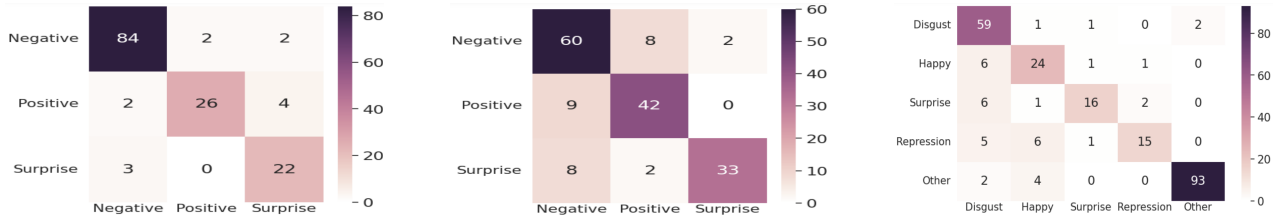
During the training for recognizing micro-expressions (ME), scarcity of ME videos and imbalanced datasets are common issues that are addressed using data augmentation techniques. To tackle these challenges, various techniques have been utilized in previous studies. For instance, [28], [38], [44], [45], and [46] have employed the Temporal Interpolation Model (TIM) [52], while [39], and [51] have utilized multiple motion amplification factors for data augmentation methods during the training process. In addition to utilizing multiple motion amplification factors as a data augmentation method, previous studies such as [24] and [31] have employed videos from other datasets as an augmentation approach during the training process. Moreover, [40] used transfer domain knowledge from macro-expression datasets to micro-expression datasets during training. Similarly, [29] balanced the datasets using the AU-GAN model to augment the data. In another study, [48] used rotation, multi-scaling, and translation as data augmentation techniques to balance the datasets during the training.

Table 4 and 5 display the results obtained from various types of data augmentation during the training process. Specifically, we used motion magnification (mm-aug) in combination with videos from other datasets (mm-oth-aug) to augment the data.

- *CASME II Dataset (3 classes):* As presented in Table 4, our proposed method, which incorporates a relational edge-node graph attention (ENGAT) network with the SAGPOOL layer, outperforms all existing techniques across all databases. Specifically, for the CASME II database, our approach achieves at least *1.37*% higher accuracy and a minimum of *0.22*% higher UF1-Score compared to state-of-the-art methods such as [14], [18], [24], [28], [42], [31], and others. The accuracy improvement for the CASME II database ranges from *1.37*% to *44.53*%, while UF1-Score improvement ranges from *0.22*% to *43.40*%. Additionally, the confusion matrix for the CASME II database, depicting the classification results for 3 categories, is illustrated in Fig. 2(a).
- *SMIC Dataset (3 classes):* Similarly, for the SMIC database, our proposed method achieved a minimum of *4.88*% improvement in accuracy and a minimum of *6.54*% improvement in UF1-Score, outperforming recent techniques such as [14], [18], [24], [42], [37], [38], and others, as indicated in Table 4. The accuracy improvement for the SMIC database ranges from *4.88*% to *20.73*%, while the UF1-Score improvement ranges from *6.54*% to *25.33*%. The confusion matrix for the SMIC database, representing the classification results for 3 types of expressions, is displayed in Fig. 2(b).
- *CASME II Dataset (5 classes):* The performance analysis results for the CASME II database, comparing our technique with the current methods, are presented in Table 5. Our proposed approach outperforms the existing methods, such as [18], [24], [42], [31], [44], [45], [47], [50], [51],and others, with a minimum improvement of *1.63*%

Table 4. Comparative performance analysis of current techniques for CASME II and SMIC databases for 3 types of emotions (Positive, Negative, and Surprise). ["-"] indicates the results are not reported. (mm-aug) refers to motion magnification augmentation and (mm-oth-aug) refers to motion magnification and videos from other datasets for data augmentation.

Approaches	Feature Extraction	CASME II		SMIC	
		Accuracy	UF1	Accuracy	UF1
Liong *et al.* [6]	Handcrafted	0.8069	0.7805	0.6159	0.5727
Khor *et al.* [18]	CNN	0.7080	0.7300	0.6341	0.6462
Gan *et al.* [14]	CNN	0.8828	0.8697	0.6817	0.6709
Zhou *et al.* [36]	CNN	0.8758	0.8621	0.6585	0.6645
Kumar *et al.* [37]	CNN	0.8621	0.8280	0.7744	0.7451
Liong *et al.* [38]	CNN	0.8741	0.8382	0.6829	0.6801
Xia *et al.* [39]	CNN	0.8030	0.7470	0.7230	0.6950
Liu *et al.* [40]	CNN	-	0.8293	-	0.7461
Xia *et al.* [41]	CNN	-	0.8090	-	0.5980
Kumar *et al.* [24]	Graph	0.8966	0.8695	0.7622	0.7606
Lo *et al.* [28]	Graph	0.5440	0.303	-	-
Kumar *et al.* [31]	Graph	0.8897	0.8638	-	-
Xie *et al.* [29]	Graph	0.7120	0.3550	-	-
Lei *et al.* [42]	Graph	-	0.8798	-	0.7192
Ours with mm-aug	Graph	0.8897	0.8686	**0.8171**	**0.8143**
Ours with mm-oth-aug	Graph	**0.9103**	**0.8820**	**0.8232**	**0.8260**
Range of Improvement in %		**1.37 to 44.53**	**0.22 to 43.40**	**4.88 to 20.73**	**6.54 to 25.33**



(a) CASME II datasets (3 classes) (b) SMIC datasets (3 classes) (c) CASME II datasets (5 classes)

Figure 2. Confusion matrices correspond to our evaluations on 2 databases for classifying MEs for 3 and 5 classes

in accuracy and a minimum improvement of *2.74*% in UF1-Score, as displayed in Table 5. The accuracy improvement for the CASME II database ranges from *1.63*% to *44.47*%, and the UF1-Score improvement ranges from *2.74*% to *42.02*%. The confusion matrix for the CASME II database, representing the classification results for 5 expressions, is shown in Fig. 2(c).

4.5. Ablation Study Results

We conducted a thorough investigation to assess the effectiveness of our proposed method by analyzing the impact of each component. We removed each part of our method to interpret the overall methods performance, and the results are presented in Tables 6 and 7. The significance of having various types of features (node features and edge features) and the selection of different types of networks such as GCN and GAT are evaluated in these tables.

The results of the ablation study on SMIC and CASME II, are presented in Table 6, for the 3-class classification problem. Initially, we obtain the baseline results by using node features and the GCN layer. Next, we replace the GCN layer with the GAT layer and observe an improvement in accuracy by 13.80% and 2.44%, and an increased UF1-score of 20.47% and 2.77% for CASME II and SMIC datasets, respectively. Finally, we incorporate edge features with node features and the GAT layer, which leads to further improvement of 1.37% and 6.10% in accuracy, and an enhanced UF1-score of 1.25% and 6.54% for the CASME II and SMIC databases.

The results of the ablation study for CASME II datasets are presented in Table 6, for the 5-class classification problem. Initially, we obtain the baseline results by using node features and the GCN layer. Next, we replace the GCN layer with the GAT layer and observe an improvement in accuracy by 4.06% and an increased UF1-score of 8.09%. Finally, we incorporate edge features with node features and the GAT layer, which leads to further improvement of 2.85% in accuracy, and an enhanced UF1-score of 7.01%.

4.6. Cross-Dataset Evaluation Results on 3 classes

To verify the robustness and generalizability of our approach, we conducted a cross-dataset evaluation on 2

Table 5. Comparative performance analysis of current methods for CASME II database for 5 classes. (mm-aug) refers to motion magnification augmentation and (mm-oth-aug) refers to motion magnification and videos from other datasets for data augmentation.

Approaches	Feat. Extract	Accuracy	UF1
Khor *et al.* [18]	Handcraft	0.3968	0.3589
Liong *et al.* [43]	Handcraft	0.6255	0.6500
Liu *et al.* [44]	Handcraft	0.6695	0.6911
Li *et al.* [45]	Handcraft	0.6721	N/A
Huang *et al.* [46]	Handcraft	0.6478	N/A
Peng *et al.* [47]	Handcraft	0.7085	N/A
Kim *et al.* [48]	CNN	0.6098	N/A
Khor *et al.* [16]	CNN	0.5244	0.5000
Zong *et al.* [49]	CNN	0.6397	0.6125
Khor *et al.* [18]	CNN	0.7078	0.7297
Li *et al.* [50]	CNN	0.6502	0.6400
Khor *et al.* [18]	CNN	0.7119	0.7151
Kumar *et al.* [24]	Graph	0.8130	0.7090
Lei *et al.* [51]	Graph	0.7398	0.7246
Lei *et al.* [42]	Graph	0.7427	0.7047
Kumar *et al.* [31]	Graph	0.8252	0.7517
Ours (mm-aug)	Graph	**0.8374**	0.7483
Ours (mm-oth-aug)	Graph	**0.8415**	**0.7791**
Range of Improvement in %		**1.63 to 44.47**	**2.74 to 42.02**

Table 6. Ablation study results for CASME II and SMIC databases for 3 types of emotions. [N]: Node features and [E]: Edge features

(N)(E)(GCN)(GAT)	CASME II (3 classes)		SMIC (3 classes)	
	Accuracy	UF1	Accuracy	UF1
(✓) (✗) (✓) (✗)	0.7586	0.6648	0.7378	0.7329
(✓) (✗) (✗) (✓)	0.8966	0.8695	0.7622	0.7606
(✓) (✓) (✗) (✓)	**0.9103**	**0.8820**	**0.8232**	**0.8260**

Table 7. Ablation study results for CASME II database for 5 types of emotions. [N]: Node features and [E]: Edge features

(N) (E) (GCN) (GAT)	CASME II (5 classes)	
	Accuracy	UF1
(✓) (✗) (✓) (✗)	0.7724	0.6281
(✓) (✗) (✗) (✓)	0.8130	0.7090
(✓) (✓) (✗) (✓)	**0.8415**	**0.7791**

Table 8. Cross dataset examination on two Facial Micro-Expression Databases (3 types of expressions).

Training Database	Evaluating Database			
	CASME II		SMIC	
	Accuracy	UF1	Accuracy	UF1
Baseline	0.7586	0.6648	0.7378	0.7329
CASME II	-	-	0.7683	0.7543
SMIC	0.8552	0.8053	-	-

databases. We evaluated our approach only on 3 classes of MEs since SMIC has only 3 classes of expressions. During the training process, we employed multiple motion magnification videos as a data augmentation technique to balance the dataset. We did not use any videos from other datasets for augmentation during the cross-validation approach. Our method, as described in section 3, was used for this evaluation. The results of the cross-dataset evaluation on three classes of MEs are presented in Table 8. We achieved an accuracy of *76.83*% and *75.43*% UF1-Score on the SMIC database when trained on the CASME II database. Likewise, when trained on the SMIC dataset and evaluated on the CASME II database, an accuracy of *85.52*% was achieved and *80.53*% UF1 Score. Our results outperformed state-of-the-art approaches [6], [18], [28], [29], [39], [40], and [41] and were comparable to [14], [36], and [37], as displayed in Table 4 for SMIC and CASME II databases.

5. Conclusions and Future Work

In this paper, we proposed a Two-stream Relational Edge-Node Graph Attention (ENGAT) Network for global and local structural similarity edge features, node coordinate features, and magnitude of optical flow attributes. We used three frame graph structure to extract spatial and temporal data. We selected global and local structural similarity score edge features based on the Jaccard similarity score index and radial basis function. This effectively enhances the relationships between various edges within the face graph structure, providing important structural information about different parts of the face that are essential in distinguishing between different classes of expressions. Additionally, these edge features, along with the node features, improved the relationship between the nodes in the graph, resulting in better attention scores for both nodes and edges. The accuracy and UF1-score of the SMIC and CASME II databases were enhanced with the implementation of our method, as demonstrated in Tables 6 and 7. A thorough evaluation of the SMIC and CASME II databases was carried out for 3 and 5 classes of expressions. Our design outperformed the current methods by *1.37*% in accuracy and *0.22*% in UF1-Score and *1.63*% in accuracy and *2.74*% in UF1-score for the CASME II database for 3 and 5 types, respectively. Likewise, for the SMIC database, our method outperformed the current methods by *4.88*% in accuracy and *6.54*% in UF1-Score. To evaluate the effectiveness of our approach, an ablation study experiment was conducted using the SMIC and CASME II databases. Furthermore, we carried out a cross-dataset experiment to assess the universal applicability of our method. In the future, we will concentrate on exploring different edge features and find out the weights of each edge and node feature in the graph structure.

6. Acknowledgments

This material is based upon work supported by the National Science Foundation under grant number 1911197.

References

[1] M. Shreve, S. Godavarthy, D. Goldgof, and S. Sarkar, "Macro-and micro-expression spotting in long videos using spatio-temporal strain," in *IEEE Face and Gesture*, 2011. 1

[2] W. J. Yan, Q. Wu, J. Liang, Y. H. Chen, and X. Fu, "How fast are the leaked facial expressions: The duration of micro-expressions," *Journal of Nonverbal Behavior*, vol. 37, pp. 217–230, Dec 2013. 1

[3] D. C. He and L. Wang, "Texture unit, texture spectrum, and texture analysis," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 28, no. 4, pp. 509–512, 1990. 1

[4] A. K. Davison, W. Merghani, and M. H. Yap, "Objective classes for micro-facial expression recognition," *Journal of Imaging*, vol. 4, 2018. 1

[5] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, pp. 915–928, 2007. 1

[6] S. Liong, J. See, K. Wong, and R. C. W. Phan, "Less is more: Micro-expression recognition from video using apex frame," *Signal Processing: Image Communication*, vol. 62, 2018. 1, 3, 7, 8

[7] H. Y. Wu, M. Rubinstein, E. Shih, J. Guttag, F. Durand, and W. T. Freeman, "Eulerian video magnification for revealing subtle changes in the world," *ACM Transactions on Graphics (SIGGRAPH)*, vol. 31, 2012. 2

[8] Y. Wang, J. See, Y.-H. Oh, R. C.-W. Phan, Y. Rahulamathavan, H.-C. Ling, S.-W. Tan, and X. Li, "Effective recognition of facial micro-expressions with video motion magnification," *Multimedia Tools Appl.*, vol. 76, Oct. 2017. 3

[9] U. Saeed, "Facial micro-expressions as a soft biometric for person recognition," *Pattern Recognition Letters*, vol. 143, pp. 95–103, 2021. 3

[10] Z. Lu, Z. Luo, H. Zheng, J. Chen, and W. Li, "A delaunay-based temporal coding model for micro-expression recognition," in *Computer Vision - ACCV Workshops*, C. Jawahar and S. Shan, Eds. Cham: Springer International Publishing, 2015, pp. 698–711. 3

[11] Y. Guo, C. Xue, Y. Wang, and M. Yu, "Micro-expression recognition based on cbp-top feature with elm," *Optik*, vol. 126, no. 23, pp. 4446–4451, 2015. 3

[12] M. M. Donia, A. A. Youssif, and A. Hashad, "Spontaneous facial expression recognition based on histogram of oriented gradients descriptor." *Comput. Inf. Sci.*, vol. 7, no. 3, pp. 31–37, 2014. 3

[13] Y.-H. Oh, A. C. Le Ngo, J. See, S.-T. Liong, R. C.-W. Phan, and H.-C. Ling, "Monogenic riesz wavelet representation for micro-expression recognition," in *IEEE International Conference on Digital Signal Processing (DSP)*, 2015, pp. 1237–1241. 3

[14] Y. Gan, S. T. Liong, W. C. Yau, Y. C. Huang, and L. K. Tan, "OFF-ApexNet on micro-expression recognition system," *Signal Processing: Image Communication*, vol. 74, pp. 129 – 139, 2019. 3, 6, 7, 8

[15] D. Y. Choi, D. H. Kim, and B. C. Song, "Recognizing fine facial micro-expressions using two-dimensional landmark feature," in *25th IEEE International Conference on Image Processing (ICIP)*, 2018, pp. 1962–1966. 3

[16] H. Khor, J. See, R. C. W. Phan, and W. Lin, "Enriched long-term recurrent convolutional network for facial micro-expression recognition," in *13th IEEE International Conference on Automatic Face Gesture Recognition*, May 2018, pp. 667–674. 3, 8

[17] A. J. Rakesh Kumar, B. Bhanu, C. Casey, S. G. Cheung, and A. Seitz, "Depth videos for the classification of micro-expressions," in *25th International Conference on Pattern Recognition (ICPR)*, 2021. 3

[18] H. Khor, J. See, S. Liong, R. C. W. Phan, and W. Lin, "Dual-stream shallow networks for facial micro-expression recognition," in *IEEE International Conference on Image Processing (ICIP)*, 2019. 3, 6, 7, 8

[19] B. Song, K. Li, Y. Zong, J. Zhu, W. Zheng, J. Shi, and L. Zhao, "Recognizing spontaneous micro-expression using a three-stream convolutional neural network," *IEEE Access*, vol. 7, pp. 184 537–184 551, 2019. 3

[20] T. Wang and L. Shang, "Temporal augmented contrastive learning for micro-expression recognition," *Pattern Recognition Letters*, vol. 167, pp. 122–131, 2023. 3

[21] X. Guo, X. Zhang, L. Li, and Z. Xia, "Micro-expression spotting with multi-scale local transformer in long videos," *Pattern Recognition Letters*, 2023. 3

[22] S. Thuseethan, S. Rajasegarar, and J. Yearwood, "Deep3dcann: A deep 3dcnn-ann framework for spontaneous micro-expression recognition," *Information Sciences*, vol. 630, pp. 341–355, 2023. 3

[23] B. Yang, J. Wu, K. Ikeda, G. Hattori, M. Sugano, Y. Iwasawa, and Y. Matsuo, "Deep learning pipeline for spotting macro- and micro-expressions in long video sequences based on action units and optical flow," *Pattern Recognition Letters*, vol. 165, pp. 63–74, 2023. 3

[24] A. J. R. Kumar and B. Bhanu, "Micro-expression classification based on landmark relations with graph attention convolutional network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2021. 2, 4, 6, 7, 8

[25] D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009. 2

[26] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," 2018. 3

[27] B. Xu, N. Wang, T. Chen, and M. Li, "Empirical evaluation of rectified activations in convolutional network," *CoRR*, vol. abs/1505.00853, 2015. [Online]. Available: http://arxiv.org/abs/1505.00853 3

[28] L. Lo, H. Xie, H. Shuai, and W. Cheng, "MER-GCN: Micro-expression recognition based on relation modeling with graph convolutional networks," in *IEEE Conference on Multimedia Information Processing and Retrieval (MIPR),*, 2020. 4, 6, 7, 8

[29] H.-X. Xie, L. Lo, H.-H. Shuai, and W.-H. Cheng, "AU-assisted graph attention convolutional network for micro-expression recognition," in *28th ACM International Conference on Multimedia*, 2020. 4, 6, 7, 8

[30] L. Zhou, Q. Mao, and M. Dong, "Objective class-based micro-expression recognition through simultaneous action unit detection and feature aggregation," *CoRR*, 2020. 4

[31] A. J. Rakesh Kumar and B. Bhanu, "Three stream graph attention network using dynamic patch selection for the classification of micro-expressions," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2022, pp. 2475–2484. 4, 6, 7, 8

[32] J. Lee, I. Lee, and K. Jaewoo, "Self-attention graph pooling," *CoRR*, 2019. 5

[33] C. Cangea, P. Veličković, N. Jovanović, T. Kipf, and P. Liò, "Towards sparse hierarchical graph classifiers," 2018. 5

[34] W. J. Yan, X. Li, S. J. Wang, G. Zhao, Y. J. Liu, Y. H. Chen, and X. Fu, "CASME II: An improved spontaneous micro-expression database and the baseline evaluation," *PLOS ONE*, vol. 9, 01 2014. 5

[35] X. Li, T. Pfister, X. Huang, G. Zhao, and M. Pietikäinen, "A spontaneous micro-expression database: Inducement, collection and baseline," in *10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, 2013, pp. 1–6. 5

[36] L. Zhou, Q. Mao, and L. Xue, "Dual-inception network for cross-database micro-expression recognition," in *14th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2019)*, 2019, pp. 1–5. 7, 8

[37] A. J. R. Kumar, R. Theagarajan, O. Peraza, and B. Bhanu, "Classification of facial micro-expressions using motion magnified emotion avatar images," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2019. 6, 7, 8

[38] S. Liong, Y. S. Gan, J. See, and H. Khor, "A shallow triple stream three-dimensional CNN (STSTNet) for micro-expression recognition system," *CoRR*, 2019. 6, 7

[39] Z. Xia, X. Hong, X. Gao, X. Feng, and G. Zhao, "Spatiotemporal recurrent convolutional networks for recognizing spontaneous micro-expressions," *CoRR*, 2019. 6, 7, 8

[40] Y. Liu, H. Du, L. Zheng, and T. Gedeon, "A neural micro-expression recognizer," in *14th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2019)*, 2019, pp. 1–4. 6, 7, 8

[41] Z. Xia, W. Peng, H. Khor, X. Feng, and G. Zhao, "Revealing the invisible with model and data shrinking for composite-database micro-expression recognition," 2020. 7, 8

[42] L. Lei, T. Chen, S. Li, and J. Li, "Micro-expression recognition based on facial graph representation learning and facial action unit fusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2021. 6, 7, 8

[43] S. Liong and K. Wong, "Micro-expression recognition using apex frame with phase information," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2017. 8

[44] Y. J. Liu, B. J. Li, and Y. K. Lai, "Sparse mdmo: Learning a discriminative feature for micro-expression recognition," *IEEE Transactions on Affective Computing*, vol. 12, pp. 254–261, 2021. 6, 8

[45] X. Li, X. Hong, A. Moilanen, X. Huang, T. Pfister, G. Zhao, and M. Pietikäinen, "Towards reading hidden emotions: A comparative study of spontaneous micro-expression spotting and recognition methods," *IEEE Transactions on Affective Computing*, vol. 9, pp. 563–577, 2018. 6, 8

[46] X. Huang, S. Wang, X. Liu, G. Zhao, X. Feng, and M. Pietikäinen, "Discriminative spatiotemporal local binary pattern with revisited integral projection for spontaneous facial micro-expression recognition," *IEEE Transactions on Affective Computing*, vol. 10, pp. 32–47, 2019. 6, 8

[47] W. Peng, X. Hong, Y. Xu, and G. Zhao, "A boost in revealing subtle facial expressions: A consolidated eulerian framework," in *14th IEEE International Conference on Automatic Face Gesture Recognition*, 2019. 6, 8

[48] D. H. Kim, W. J. Baddar, and Y. M. Ro, "Micro-expression recognition with expression-state constrained spatio-temporal feature representations," in *Proceedings of the 24th ACM International Conference on Multimedia*, 2016. 6, 8

[49] Y. Zong, X. Huang, W. Zheng, Z. Cui, and G. Zhao, "Learning from hierarchical spatiotemporal descriptors for micro-expression recognition," *IEEE Transactions on Multimedia*, vol. 20, pp. 3160–3172, 2018. 8

[50] Y. Li, X. Huang, and G. Zhao, "Joint local and global information learning with single apex frame detection for micro-expression recognition," *IEEE Transactions on Image Processing*, vol. 30, pp. 249–263, 2021. 6, 8

[51] L. Lei, J. Li, T. Chen, and S. Li, "A Novel Graph-TCN with a graph structured representation for micro-expression recognition," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, p. 2237–2245. 6, 8

[52] Z. Zhou, G. Zhao, Y. Guo, and M. Pietikainen, "An image-based visual speech animation system," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 10, pp. 1420–1432, 2012. 6