Three Stream Graph Attention Network using Dynamic Patch Selection for the 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

To understand the genuine emotions expressed by humans during social interactions, it is necessary to recognize the subtle changes on the face (micro-expressions) demonstrated by an individual. Facial micro-expressions are brief, rapid, spontaneous gestures and non-voluntary facial muscle movements beneath the skin. Therefore, it is a challenging task to classify facial micro-expressions. This paper presents an end-to-end novel three-stream graph attention network model to capture the subtle changes on the face and recognize micro-expressions (MEs) by exploiting the relationship between optical flow magnitude, optical flow direction, and the node locations features. A facial graph representational structure is used to extract the spatial and temporal information using the three frames. The varying dynamic patch size of optical flow features is used to extract the local texture information across each landmark point. The network only utilizes the landmark points location features and optical flow information across these points and generates good results for the classification of MEs. A comprehensive evaluation of SAMM and the CASME II datasets demonstrates the high efficacy, efficiency, and generalizability of the proposed approach and achieves better results than the state-of-the-art methods.

1. Introduction

Human beings express their thoughts/emotions in various ways: (i) in the form of verbal communication, (ii) in facial expressions. Human emotions originate from the amygdala region, and these emotions last between 0.5 and 4.0 seconds. Facial expressions are a non-verbal mode of communication between people, and they reflect the instant fluctuations of human emotional states. These expressions convey the emotional state of an individual to observers, regardless of their culture, language, and personal background. Facial expressions are grouped into two categories:

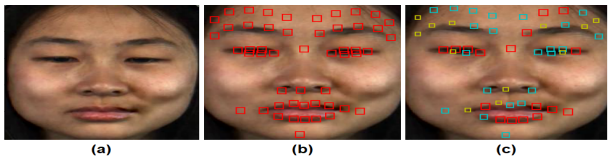

Figure 1. Illustration of why varying dynamic patch size selection is required. (a) high-intensity expression frame of a video, (b) high-intensity expression frames with a fixed patch size of 10×10 shown in red color for happy expression, (c) dynamic varying patch size for happy expression. The smaller patches are mostly selected on the eyebrow, forehead, eye and nose. The bigger patch size is selected across mouth region. Different patch sizes are shown in color: 6x6 - yellow, 8x8 - blue, and 10x10 - red.

facial macro-expressions and facial micro-expressions. Facial macro-expressions last between 0.6 to 4 seconds. These expressions have high intensity in expressing emotion, are prolonged, and are easily recognizable by humans and machines. It is easy to conceal or fake macro-expressions by humans. Facial micro-expressions (MEs) are subtle, rapid, brief, and involuntary facial muscle movements beneath the skin and last between 0.1 to 0.6 seconds. MEs occur in everyone and often without their knowledge. Therefore, MEs show the person's true emotions [1]. MEs cannot be concealed and faked as they happen in an instant. There are a variety of applications for micro-expressions in the fields of lie detection, online learning, ensuring security, and health care domain for depression recovery, therapies, and more, and online gaming. Therefore, it is essential to develop a system to recognize micro-expressions.

Micro-expression recognition has gained importance in the computer vision community in the last few years. Researchers have been recognizing MEs by using hand-crafted approaches such as Bi-Weighted Oriented Optical Flow (Bi-WOOF) [4], Local Binary Pattern with Three Orthogonal Planes (LBP-TOP) [45], and 3D Histogram of Oriented Gradient (3DHOG) [21] to extract the textural spatio-temporal information. But these techniques fail to capture

the subtle changes on the human face. Recent advances in the field of deep learning have made it possible for researchers to utilize convolutional neural networks (CNNs) and graph neural networks (GNNs) to extract the features for the classification of micro-expressions. These techniques have helped improve the precision of detecting and micro-expression recognition (MER) task.

The classification of facial micro-expressions is a difficult task for the following three characteristics of MEs: (i) subtle and brief behavior, (ii) ephemeral and spontaneous change in the facial muscle movements, and (iii) short time duration. Providing sufficient and balanced training data samples is another problem associated with facial micro-expression classification and spotting tasks.

To overcome the above significant issues, we propose a novel approach for *end-to-end training of a graph structure that uses three-stream Graph Attention Network with a self-attention graph pooling layer by exploiting the relationship between the landmark points location, optical flow magnitude and the optical flow direction*. We use three frames structure connections to exploit the spatio-temporal information. The varying patch size across each landmark point is dynamically selected based on the optical flow information. The patch size is selected dynamically to give prominence to the landmark points with higher intensity of facial muscle movement as shown in Fig. 1. This helps in removing the unwanted noise from the bigger and fixed size patch features. We choose the frames with high intensity of facial movements and remove the frames with low intensity of facial muscle movements. To address the unbalanced data samples issue, we use videos from the other datasets of the same class to increase the number of samples. Along with the above data augmentation method, we use various values of magnification factors in EMM [38] techniques to increase the number of data samples for classes with smaller numbers, thus balancing the dataset.

The rest of this paper is organized as follows. In section 2, we introduce the related works and our contributions. In section 3, we explain the technical approach. In Section 4, we present the qualitative and quantitative experimental results, including ablation study results. Finally, in section 5, we present conclusions and future work.

2. Related Work and Contributions

2.1. Related Works

In the last decade, micro-expression recognition (MER) has gained a lot of interest from computer vision researchers. In MER, the pre-processing stage includes all processes such as image resizing, alignment, motion magnification, and frame selection approaches that must be completed before meaningful feature extraction can begin. The approaches used to classify MEs into different categories of expressions are based on various feature extraction techniques, namely: i) handcrafted feature extraction, as shown in Table. 1. ii) convolutional neural networks (CNNs), as shown in Table. 2, and iii) graph neural networks (GNNs), as shown in Table. 3. In recent years, researchers use CNNs and GNNs for the feature extraction process of facial micro-expression video clips as they are more reliable and perform better than the handcrafted approaches for the classification.

2.2. Contributions

The contributions of this work are:

- We propose an end-to-end landmark-assisted three-stream Graph Attention Network with a self-attention graph pooling, which integrates optical flow magnitude, optical flow direction and the landmark points location features.
- We propose a dynamic selection of varying patch size across each landmark points to capture the change in optical flow magnitude and direction features.
- We provide a comprehensive evaluation of the proposed approach on two datasets for 3 and 5 classes of micro-expressions. We also evaluate our approach on cross-datasets to generalize our approach.

3. Technical Approach

The overall architecture of our proposed method for the classification of MEs is shown in Fig. 2. Using Eulerian Motion Magnification (EMM) [38], we amplified the input signals and extracted the magnified input videos. As a next step, we apply a threshold value using the optical flow magnitude to identify high-intensity expression frames and remove the low-intensity expression frames same as mentioned in the paper [11]. We use dlib software [10] to obtain the landmark points on the face. The patch size across each landmark point is dynamically selected to capture the subtle change in the optical flow magnitude and direction components. We constructed a three-stream graph attention network for the features such as landmark points location features, dynamic patch size of optical flow magnitude features and the direction features. Finally, we classify the MEs into different classes of expressions using a three-stream Graph Attention Network with a self-attention graph pooling layer.

3.1. Landmark Points Detection and Dynamic Patch Selection (DPS)

We use the dlib [10] software to obtain 68 facial landmark points. Out of these 68 points, we use only 37 points and remove the rest of the points from the contour region of a face, a few points on the nose region, and the inner points of the mouth region. In order to capture the subtle variation

Table 1. Related works for the classification of MEs using handcrafted features.

Author	Video/ Key Frames	Features	Classifier
Wang *et al.* [37]	Video	LBP-TOP+ EVM	KNN, SVM
Zhang *et al.* [44]	Video	(HIGO-TOP, LBP-TOP, HOG-TOP)+Relief	SVM
Davison *et. al.* [4]	Video	3DHOG	SMO+SVM
Wang *et al.* [36]	Video	LBP-TOP + Optical flow	SVM
Liong *et al.* [21]	Apex	Optical flow + Bi-WOOF	SVM
Liong *et al.* [20]	Video	Optical flow + Optical Strain +LBP	SVM
Liu *et al.* [26]	Video	Optical flow features + affine transform	SVM
Li *et al.* [18]	Video	HIGO+EVM+TIM	SVM

Table 2. Related works for the classification of MEs using CNN features.

Author	Video/ Key Frame	Features	Classifier
Liong *et al.* [24]	Onset + Apex	Optical Flow + CNN	MLP
Kumar *et al.* [13]	Energy Avatar Image	CNN	MLP
Khor*et al.* [9]	Video	Optical flow + CNN-LSTM	SVM
Kumar *et al.* [12]	Video	CNN, CNN-LSTM, 3DHOG	SVM, MLP
Peng *et al.* [30]	Video	2S-3D CNN	MLP
Khor *et al.* [7]	Video	2S-CNN	MLP
Song *et al.* [33]	Onset, Apex and Offset	3S-CNN	MLP
Xia *et al.* [39]	Video	CNN+GAN + Transfer Learning	MLP
Gan *et al.* [5]	Apex	Optical flow + CNN	MLP
Li *et al.* [19]	Apex	CNN + Attention	MLP
Jia *et al.* [31]	Video	CNN + Transfer Learning	MLP

on the forehead and the cheek region of the face, we add 10 reference points at a minimum distance of 20 pixels above the eyebrow and 4 points near the mouth, [11]. These reference landmark points are included using the onset frame.

The graph is constructed using the 51 landmark points. The points are connected based on the human facial structure. The landmark point locations are the node features for the first stream of the network.

The fixed patch size across the landmark points captures equal information across each point and does not pay importance locally to the region-of-interest, which leads to unwanted information and noise added to the patch. Therefore, it is necessary to select different patch sizes based on the subtle changes that occur locally in the *region-of-interest*. In order to select the different patch size of the optical flow magnitude and direction components for each landmark point as the node features, we first calculate the optical flow magnitude component of patch size equal to 10×10 at the respective landmark location as shown in Fig. 3 for the entire video. Then for each landmark point in an image, the sum of the optical flow magnitude (10×10 patch size) is calculated and repeated for the entire video. We select 10×10 patch size in the beginning to understand the changes that occur at each point, and we do not want to miss any changes in the facial muscle movement near the landmark points. The other reason is that we have motion magnified the video. Now, after calculating the sum of the optical flow magnitude for each point in a video, we calculate the *(max-min)* of the optical flow magnitude for each landmark point. At this point, we have 51 points with a *(max-min)* value calculated for an entire video. Next, we calculate the percentile score component of these 51 points of the optical flow magnitude. The patch size is selected based on the equation 1.

$$p = \begin{cases} 6 \text{ x } 6\,, & \text{if } percentile < 0.34 \\ 8 \text{ x } 8\,, & \text{if } 0.34 \leq percentile < 0.67 \\ 10 \text{ x } 10 & \text{otherwise} \end{cases} \quad (1)$$

We select a dynamic patch size in our approach based on the above algorithm to capture the subtle changes of micro-expressions across each landmark point. The varying dynamic patch size across each landmark point are shown in Fig. 1 and 4. This helps in improving the performance of classification tasks (F1-score) for an individual class of micro-expressions, as shown in Table. 9 and 10. The optical flow feature matrix of size (NxN) is computed, where N is the patch size selected. After computation of the optical flow feature matrix across each landmark, we zero pad the feature matrix to 10×10 patch size to make computation easier. We will investigate the use of incremental change in the patch size for the optical flow features based on our current approach. The feature matrix is flattened to a *1D* vector of the feature vector as shown in Fig. 3. The optical flow magnitude feature vector is an input to the second stream, and the optical flow direction feature vector is an input to the third stream of the graph network.

Table 3. Related works for the classification of MEs using GNN features.

Author	Video/ Key Frame	Features	Classifier
Lo *et al.* [28]	Video	AU features + 3D CNN + GNN	MLP
Lei *et al.* [17]	Onset + Apex	Landmark points + 2S-GCN	MLP
Xie *et al.* [42]	Video	AU features + GCN	MLP
Kumar *et al.* [11]	Important Frames	Landmark points + Optical flow magnitude + 2S-GAT	MLP
Zhou *et al.* [46]	Onset + Apex	Optical flow + AU features + GCN	MLP
Lei *et al.* [16]	Video	Landmark points + CNN + Graph transformer	MLP

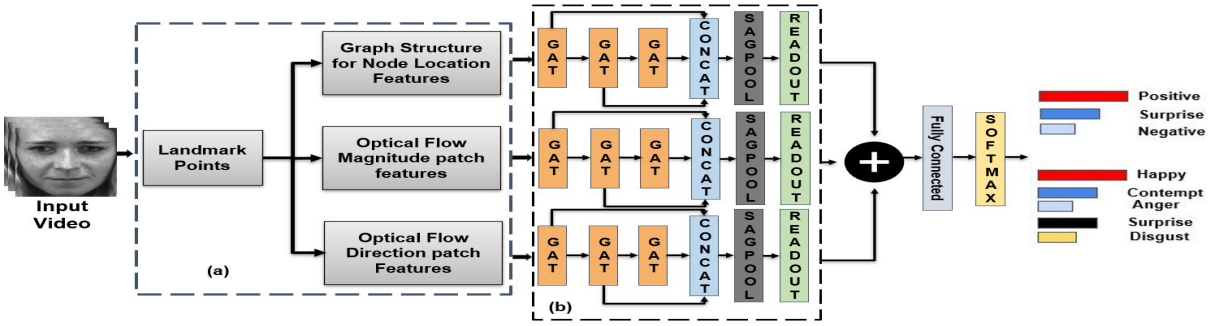

Figure 2. Overall architecture of our proposed approach. (a) landmark points are detected on the face and based on these points, the landmark point location and dynamic patch size features for the optical flow magnitude and direction features are extracted. (b) graph attention network (GAT) with a self-attention graph pooling (SAGPOOL) layer for training the graph representation is used and, finally, the fusion of the 3 streams for the classification of MEs for different categories of expressions is based on the datasets.

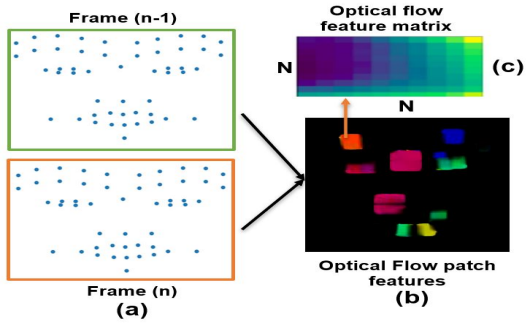

Figure 3. The process of obtaining the optical flow patch information. (a) input frames of a video, (b) N×N optical flow feature patch around each landmark point, (c) N×N optical flow feature matrix is the patch around each landmark point.

3.2. Graph Attention Network

Graph Attention Networks [35] are used to influence the self-attention layer and employ the self-attention of node features. The graph attention layer shares the node features with all its neighboring nodes. The network learns the attention weights between two connected nodes on the assumption that neighboring nodes do not contribute the same weights to the central nodes as the Graph Convolutional Network (GCN) model does. The attention weights indicate the importance of node features of one node to another node. The graph structure is introduced and the shared attention mechanism is only performed between the nodes and their corresponding edges.

3.3. Self-Attention Pooling Layer

Graph pooling reduces the number of parameters from the network and retains a portion of input graph nodes, preventing overfitting. Self-attention graph pooling [15] uses any GNN networks to obtain the attention score for the process of pooling. To select only the necessary nodes, a pooling ratio of $k \in (0, 1]$ which determines the number of nodes to remain in the final graph structure. First, the self-attention graph pooling layer calculates the attention scores from the graph attention layer. Later, it selects the top-k nodes to remain in the graph based on the attention score determined from the graph attention layer for the nodes and also based on the ratio k selected. Finally, based on the ids of the nodes remaining and their connections between the nodes, a new feature matrix and the new adjacency matrix are created to form a new graph structure, respectively.

Finally, the output from the self-attention graph pooling layer is passed through the readout layer [2] (global average pooling and global max pooling) to construct a fixed size node feature representation.

3.4. Three-Stream Graph Attention Network

We developed a novel Three-stream Graph Attention Network that extracts temporal features from the video, as shown in Fig. 2. We construct a graph representation us-

ing a three-frame structure, where the entire video is transformed into a single graph.

We use graph attention layers to design our graph network as shown in Fig. 2. We use 3 graph attention layers with the ReLU activation function after each layer. We use 32 hidden channels, the concatenation operation is off, and the number of heads = 1 for the graph attention layer. For the first stream of the graph network, the node feature vector is the x and y location coordinates of the landmark points. The node location features help in understanding the change in the movement of each landmark point w.r.t to its previous position. For the second and the third stream of our network, we use the varying patch size of the optical flow magnitude features and the optical flow direction features. The 3-stream network helps in extracting the relationship between the different node features to the full extent. The optical flow magnitude and direction component capture the spatio-temporal information of the MEs along with the three frames 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 to remove the less important nodes based on the attention scores of the nodes and the ratio of *k* value in the *top-k* selection process.

The output of the self-attention graph pooling layer is passed on to the readout layer to get a fixed size representation of the output layer. At the end of the readout layer of the three-stream graph networks, the results are concatenated for the graph representation of the three streams. Finally, the output is passed through the fully connected layer and softmax layer for classification.

4. Experimental Results

In this section we describe the experimental results such as the datasets used, experimental setup, and experimental details used for the classification of MEs. We conducted a comprehensive study to evaluate the performance of our proposed approach by removing each of the components of our approach, to fully understand the overall method. For the classification of micro-expressions, we conducted cross-dataset evaluations of the approach to verify the robustness of our approach and its generalizability to different environments and subjects.

4.1. Experimental Setup

We conduct experiments on two publicly available datasets CASME II [43] and SAMM [3] datasets for the evaluation on 3 and 5 classes of facial micro-expressions. We evaluate our results using *leave-one-subject-out* (LOSO-CV) cross validation approach. The experiments run on a workstation that has Ubuntu 16.04 OS with 16GB RAM and two NVIDIA GeForce GTX 1080Ti GPUs.

4.2. Datasets and Preprocessing

The *two* publicly available datasets are: CASME II [43] and SAMM [3]. We are interested in classifying the micro-expressions into 3 and 5 categories of facial micro-expressions. LOSO-CV is a subject-independent cross-validation method, which can avoid subject bias and evaluate the generalization ability of various algorithms. Table. 4 and 5 shows the dataset distributions for each expression class for CASME II and SAMM 3 and 5 class categories.

CASME II dataset has 247 ME videos consisting of 26 subjects, categorized into five classes of facial micro-expressions. The dataset suffers from class imbalance, and the subjects are limited to only one ethnicity. The video are in RGB format and the mean age group of subjects is 22.03 years. SAMM dataset has 159 ME samples from 32 subjects which are categorized into 8 classes. The dataset has gray-scale video samples with 13 different ethnicities. The mean age group of participants is 33.24 years.

We have aligned each image with a reference image (onset frame) and resized the image frames to 256x256. To solve the issue of data imbalance, we used the data (Happy and Surprise) from the other dataset (CASME II/SAMM) while training for the class having a lower number of data samples of videos to improve the training accuracy depending on the dataset used for classification. Also, we used different values of magnification factor (α) (1, 2, 3, 4, and 5) to increase the data samples to overcome the class imbalance of the datasets during training. For the testing purpose, magnification factor (α) 4 is used. For self-attention graph pooling layer, we use k = 0.75 as the ratio to calculate the number of nodes to remain in the graph structure after the self-attention graph pooling layer. The value of k = 0.75 is chosen so that we do not eliminate important nodes from the graph structure and still have plenty of nodes in the graph structure for the classification process of MEs. We use an Adam optimizer with a learning rate equal to 0.001. The learning rate decreases by half every 100 epochs.

Table 4. Summary of the data distributions for CASME II and SAMM for 3 classes.

Expression Class	CASME II	SAMM
Negative	88	92
Positive	32	26
Surprise	25	15

4.3. Evaluation Metrics

The class distributions of the two datasets are unbalanced with respect to the number of classes. Therefore, we cannot use accuracy as the only performance metric to gauge our approach for the classification of MEs. We use the unweighted F1 (UF1) score and accuracy as the performance metrics to evaluate the recognition performance.

Table 5. Summary of the data distributions for CASME II and SAMM for 5 classes.

Expressions	CASME II	Expressions	SAMM
Disgust	63	Anger	57
Happy	32	Happy	26
Surprise	25	Surprise	15
Repression	27	Contempt	12
Other	99	Other	26

4.3.1 Unweighted F1 score (UF1)

F1-score provides equal importance to each class of the datasets. From the confusion matrix, we compute the True Positives (TP), False Positives (FP), and False Negatives (FP) for each class c. The final balanced F1 score is computed by taking the average for each class F1 scores shown in equation 2 and 3.

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

$$UF1 = \frac{F1_c}{C}, \tag{3}$$

where, $F1_c$ is F1-score for each individual class, C is the number of classes.

4.3.2 Accuracy

The accuracy is calculated using the equation 4.

$$Acc = \frac{P}{N} \times 100\% \tag{4}$$

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

4.4. Experimental Results

Table 6 shows the comparison results between the state-of-the-art methods and our proposed approach for two publicly available datasets: CASME II and SAMM datasets for three categories of expressions using the LOSO-CV. In the LOSO-CV technique, we use a K-fold cross-validation technique, with K equal to N number of subjects, which means we repeat the experiment N times for the classification of MEs with N-1 subjects data for the training process and the remaining 1 subject for the testing process.

Our three-stream graph attention network approach outperforms the state-of-the-art methods as shown in Table 6. For CASME II dataset (3 classes), our accuracy is lower by *0.69*% as compared to Kumar *et al.* [11] and F1-Score is lower by *0.59*% as compared to Gan *et al.* [5]. Similarly, for the SAMM dataset our approach achieves higher accuracy by *2.26*% and F1-Score is higher by *3.45*% compared to the best state-of-the-art method.

The comparison results (5 classes) for the CASME II dataset using the state-of-the-art approaches and our approach are shown in Table 7. The recent paper from Nie *et al.* [29] had the best F1-Score result for CASME II datasets for five classes until recent times.The paper from Kumar *et al.* [11] had the best accuracy result for CASME II for the classification of MEs (5 classes) until recent times. When compared to Nie *et al.* [29], our proposed approach improves the accuracy by *7.32*%, and the F1 score is higher than their method by *1.63*%. Similarly, when compared to Kumar *et al.* [11], our approach increases precision by *1.22*%, and F1-score is higher by *4.27*%.

Table 8 shows the comparative results (5 classes) using the state-of-the-art approaches and our approach for the SAMM dataset. The recent paper from Kumar *et al.* [11] proposed a 2-stream graph attention network which had the best results in terms of both F1-score and accuracy for SAMM dataset (5 classes). When compared to [11], our proposed approach improves the accuracy by *1.47*% and the F1 score is higher by *0.86*%.

When the number of categories of expressions increases from 3 to 5 the performance can be explained from the Tables 6, 7 and 8. For the CASME II dataset, as the number of expression categories increases from 3 to 5, the accuracy and F1 score decrease as shown in Table 6 and 7. This decrease is due to the following reasons. 1) scalability problem arises as to the similarities among the classes increase, and 2) the number of class samples in each category is highly imbalanced, and it is difficult to balance the dataset, for example, for a repression class of expression other datasets are not available for augmentation except for CASME II. For the SAMM dataset, the accuracy and F1 score for three and five categories of classification of MEs are almost the same as shown in Tables 6 and 8. The reason for at par accuracy and F1-score even after increasing the number of categories from 3 to 5 is because balancing the dataset is easy as the expression categories are universal, and the similarities among regions-of-interest among the 5 classes of expressions is low.

The approximate computation time of our approach is 4.5s for happy expression, 3.5s for repression expression, 4.8s for surprise expression, and 4s for disgust and other expressions. The approach by Kumar *et al.* [11], takes 4s for happy expression, 3s for repression, 4.5s for surprise expression, 3.7s for disgust, and other expressions. The reasons for our approach taking a longer time than [11] are due to the fact that our network is a 3 stream network, and it consists of a self-attention graph pooling layer, and multiple fully-connected layers. Our method improves the accuracy and F1 score by having a trade-off with the computation time.

Table 6. Comparison with the state-of-the-art approaches for CASME II and SAMM datasets for 3 categories of expressions.

Method	Feature Extraction Approach	CASME II		SAMM	
		Accuracy	F1 Score	Accuracy	F1 Score
Ngo *et al.* [14]	Handcrafted	0.4900	0.5100	0.5900	0.364
Wang *et al.* [37]	Handcrafted	0.4650	0.4480	0.4150	0.4060
Liong *et al.* [21]	Handcrafted	0.5880	0.6100	0.5830	0.3970
Huang *et al.* [6]	CNN	0.6400	0.6380	0.6380	0.6110
Khor *et al.* [7]	CNN	0.7080	0.7300	0.5740	0.4640
Gan *et al.* [5]	CNN	0.8828	**0.8697**	0.6818	0.5423
Kumar *et al.* [13]	CNN	0.8621	0.8280	0.8195	0.7056
Liong *et al.* [23]	CNN	0.8741	0.8382	0.7744	0.6588
Xia *et al.* [41]	CNN	0.8030	0.7470	0.7860	0.7410
Lo *et al.* [28]	Graph based	0.5440	0.3030	0.5340	0.2830
Xie *et al.* [42]	Graph based	0.7120	0.3550	0.5230	0.3570
Kumar *et al.* [11]	Graph based	**0.8966**	0.8695	0.8872	0.8118
Ours	Graph based	0.8897	0.8638	**0.9098**	**0.8463**

Table 7. Comparison with the state-of-the-art approaches for CASME II datasets for 5 categories of expressions.

Methods	Descriptors	Accuracy	F1-Score
Khor *et al.* [8]	LBP-TOP	0.3968	0.3589
Khor *et al.* [8]	Alexnet	0.6296	0.6675
Liong *et al.* [22]	Bi-WOOF	0.6255	0.6500
Liu *et al.* [27]	MDMO	0.6695	0.6911
Li *et al.* [18]	HIGO-Mag	0.6721	N/A
Peng *et al.* [32]	ME-Booster	0.7085	N/A
Khor *et al.* [8]	DSSN	0.7078	0.7297
Li *et al.* [19]	CNN+Att.	0.6502	0.6400
Khor *et al.* [8]	SSSN	0.7119	0.7151
Nie *et al.* [29]	2S-CNN+ML	0.7520	0.7354
Liu *et al.* [25]	CNN	0.6463	0.6349
Lei *et al.* [17]	Graph TCN	0.7398	0.7246
Lei *et al.* [16]	Graph-AU	0.7427	0.7047
Kumar *et al.* [11]	GACNN	0.8130	0.7090
Ours	3 Stream Graph	**0.8252**	**0.7517**

Table 8. Comparison with the state-of-the-art approaches for SAMM datasets for 5 categories of expressions.

Method	Descriptors	Accuracy	F1-Score
Khor *et al.* [8]	LBP-TOP	0.3968	0.3589
Khor *et al.* [8]	CNN	0.5294	0.4260
Khor *et al.* [8]	SSSN	0.5662	0.4513
Khor *et al.* [8]	DSSN	0.5735	0.4644
Li *et al.* [19]	CNN+Att.	0.4090	0.3400
Su *et al.* [34]	CNN+Att.	0.6324	0.5709
Xia *et al.* [40]	CNN+GAN	0.7410	0.7360
Nie *et al.* [29]	CNN+ML	0.5588	0.4538
Song *et al.* [33]	CNN	0.7176	0.6942
Lei. *et al.* [17]	Graph TCN	0.7500	0.6985
Lei *et al.* [16]	Graph-AU	0.7426	0.7045
Kumar *et al.* [11]	GACNN	0.8824	0.8279
Ours	3 Stream Graph	**0.8971**	**0.8365**

4.5. Ablation Study Results

A comprehensive study is performed to evaluate our proposed method and understand the performance of our approach by removing each component of our method to fully comprehend the overall approach. Table. 9 and 10 shows the significance of having the varying and dynamic patch size selection for capturing the optical flow magnitude and optical flow direction features.

Table 9 shows the ablation study results (3 classes) for two publicly available datasets: CASME II and SAMM, respectively. We get the baseline results of our approach with a constant patch size of optical flow magnitude and optical flow direction features. The baseline results for the CASME II datasets accuracy and F1-score are *88.26*% and *85.02*%. The baseline results for the SAMM datasets accuracy and F1-score are *87.97*% and *78.69*%. When we use the complete network with the varying patch size features of optical flow and direction features, accuracy increases by *0.71*% and *1.36*% in the F1-score for the CASME II dataset, and accuracy increases by *3.01*% and *5.94*% in F1-score for the SAMM datasets, respectively.

Table 10 shows the ablation study results (5 classes) for two datasets: CASME II and SAMM, respectively. We get the baseline results of our approach with a constant patch size of optical flow magnitude and optical flow direction features. The baseline results for the CASME II datasets accuracy and F1-score are *82.11*% and *71.43*%. The baseline results for the SAMM datasets accuracy and F1-score are *87.50*% and *80.80*%. When we use the complete network with the varying patch size features of optical flow and direction features, accuracy increases by *0.41*% and *3.74*% in F1-score for the CASME II dataset, and accuracy increases by *2.21*% and *2.85*% in F1-score for the SAMM dataset.

4.6. Cross-Dataset Evaluation Results

We perform a cross-dataset evaluation on the two publicly available micro-expressions datasets to verify the robustness of our approach and its generalizability to learning features from different environments and subjects irrespective of their gender, race, age, and other aspects. The

Table 9. Ablation study for CASME II and SAMM dataset for 3 classes of expressions.

Method	CASME II (3 classes)		SAMM (3 classes)	
	Accuracy	F1-Score	Accuracy	F1-Score
3 stream network (constant patch size selection)	0.8826	0.8502	0.8797	0.7869
3 stream network (with dynamic patch size selection)	**0.8897**	**0.8638**	**0.9098**	**0.8463**

Table 10. Ablation study for CASME II and SAMM dataset for 5 classes of expressions.

Method	CASME II (5 classes)		SAMM (5 classes)	
	Accuracy	F1-Score	Accuracy	F1-Score
3 stream network (constant patch size selection)	0.8211	0.7143	0.8750	0.8080
3 stream network (with dynamic patch size selection)	**0.8252**	**0.7517**	**0.8971**	**0.8365**

Figure 4. *6×6* patch size: yellow, *8×8*: blue and *10×10*: red. The first image in a row is the high-intensity expression frame, the second image in a row is with fixed patch size, and the third image in a row is of varying patch size. The first row is where both fixed patch size and our approach succeeded, whereas for 2nd and 3rd rows, our approach classified correctly and the fixed patch size failed to classify, and final row is the video frame where our varying patch size approach failed. (a) happy expression, (b) surprise expression, (c) disgust expression, and (d) surprise expression. (b) and (d) expressions are the same but with the different subject.

cross-dataset evaluation is conducted only on three classes of MEs. The reason for not performing cross-dataset evaluation on other categories of expressions is due to having different classes of expressions present in the SAMM and CASME II datasets. Table 11 shows the robustness of our approach for the classification of MEs on 3 classes for the cross-dataset evaluation procedure. We use the same proposed method as mentioned in the technical approach in section 3. When trained on the CASME II dataset and tested on the SAMM dataset, we achieved an accuracy of *82.71*% and *67.01*% F1-Score, respectively. Similarly, when trained on the SAMM dataset and tested on the CASME II dataset, we achieved an accuracy of *75.17*% and *63.91*% F1-Score, respectively. These results are better than the state-of-the-art graph-based approaches [28] [42], as shown in Table. 6 for CASME II and SAMM datasets (3 classes).

Table 11. Cross-Dataset Evaluation for two publicly available Facial Micro-Expression Datasets (3 classes).

Training Dataset	Testing Dataset			
	CASME II		SAMM	
	Accuracy	UF1	Accuracy	UF1
CASME II	-	-	0.8271	0.6701
SAMM	0.7517	0.6391	-	-

5. Conclusions and Future Work

In this paper, we proposed a Three-stream Graph Attention Network for the node location features, optical flow magnitude, and optical flow direction features with the help of three frames structures to extract the spatio-temporal information. We designed an algorithm to dynamically select the varying patch size across each landmark point for the optical flow features to be extracted. Our dynamic patch size approach helped in improving the accuracy and F1 score for the CASME II and SAMM datasets when compared to the fixed patch size approach as shown in Table. 9 and 10. We conducted a comprehensive evaluation of the CASME II and SAMM datasets for 3 and 5 classes of expressions. Our proposed approach outperforms the state-of-the-art methods by *1.22*% in terms of accuracy for the CASME II dataset (5 classes). For the SAMM dataset, our approach improves the accuracy results from the current approaches by *2.26*% and *1.47*% for the 3 and 5 categories of expressions. We conducted an ablation study for CASME II and SAMM datasets. We also performed a cross-dataset experiment to evaluate the generalization capability of our approach for 3 categories of expressions. In the future, we will work on generating ME videos to overcome the imbalance of the classes and increase the data samples for the training purpose.

6. Acknowledgment

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

References

[1] How many licensed clinical psychologists are there in the usa? https://www.apa.org/monitor/2014/06/datapoint. 1

[2] C. Cangea, P. Veličković, N. Jovanović, T. Kipf, and P. Liò. Towards sparse hierarchical graph classifiers, 2018. 4

[3] A. K. Davison, C. Lansley, N. Costen, K. Tan, and M. H. Yap. Samm: A spontaneous micro-facial movement dataset. *IEEE Transactions on Affective Computing*, 9(1):116–129, Jan 2018. 5

[4] A. K. Davison, W. Merghani, and M. H. Yap. Objective classes for micro-facial expression recognition. *Journal of Imaging*, 4(10), 2018. 1, 3

[5] Y.S. 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*, 74:129 – 139, 2019. 3, 6, 7

[6] X. Huang, G. Zhao, X. Hong, W. Zheng, and M. Pietikäinen. Spontaneous facial micro-expression analysis using spatiotemporal completed local quantized patterns. *Neurocomputing*, 175:564–578, 2016. 7

[7] H. Khor, J. See, S. Liong, R. C. W. Phan, and W. Lin. Dual-stream shallow networks for facial micro-expression recognition. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 36–40, 2019. 3, 7

[8] H. Khor, J. See, S. Liong, R. C. W. Phan, and W. Lin. Dual-stream shallow networks for facial micro-expression recognition. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 36–40, 2019. 7

[9] H. Khor, J. See, R. C. W. Phan, and W. Lin. Enriched long-term recurrent convolutional network for facial micro-expression recognition. In *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, pages 667–674, May 2018. 3

[10] D. E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10(60):1755–1758, 2009. 2

[11] 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*, pages 1511–1520, June 2021. 2, 3, 4, 6, 7

[12] A. J. R. Kumar, B. Bhanu, C. Casey, S. Grace Cheung, and A. Seitz. Depth videos for the classification of micro-expressions. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 5278–5285, 2021. 3

[13] 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*, June 2019. 3, 7

[14] A. C. Le Ngo, J. See, and R. C. . Phan. Sparsity in dynamics of spontaneous subtle emotions: Analysis and application. *IEEE Transactions on Affective Computing*, 8(3):396–411, 2017. 7

[15] J. Lee, I. Lee, and K. Jaewoo. Self-attention graph pooling. *CoRR*, abs/1904.08082, 2019. 4

[16] 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*, pages 1571–1580, June 2021. 4, 7

[17] 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*, MM '20, page 2237–2245, New York, NY, USA, 2020. Association for Computing Machinery. 4, 7

[18] 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*, 9(4):563–577, 2018. 3, 7

[19] 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*, 30:249–263, 2021. 3, 7

[20] S.T. Liong, J. See, R. C.W. Phan, Y.H. Oh, A. C. Le Ngo, K. Wong, and S.W. Tan. Spontaneous subtle expression detection and recognition based on facial strain. *CoRR*, abs/1606.02792, 2016. 3

[21] S.T. 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*, 62:82 – 92, 2018. 1, 3, 7

[22] S. Liong and K. Wong. Micro-expression recognition using apex frame with phase information. In *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 534–537, 2017. 7

[23] S. T. Liong, Y. S. Gan, J. See, and H. Q. Khor. A shallow triple stream three-dimensional CNN (ststnet) for micro-expression recognition system. *CoRR*, abs/1902.03634, 2019. 7

[24] S. T. Liong, Y. S. Gan, W. C. Yau, Y. C. Huang, and T. L. Ken. Off-apexnet on micro-expression recognition system. *CoRR*, abs/1805.08699, 2018. 3

[25] N. Liu, X. Liu, Z. Zhang, X. Xu, and T. Chen. Offset or onset frame: A multi-stream convolutional neural network with capsulenet module for micro-expression recognition. In *2020 5th International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS)*, pages 236–240, 2020. 7

[26] Y. Liu, J. Zhang, W. Yan, S. Wang, G. Zhao, and X. Fu. A main directional mean optical flow feature for spontaneous micro-expression recognition. *IEEE Transactions on Affective Computing*, 7(4):299–310, 2016. 3

[27] 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*, 12(1):254–261, 2021. 7

[28] L. Lo, H. Xie, H. Shuai, and W. Cheng. Mer-gcn: Micro-expression recognition based on relation modeling with graph convolutional networks. In *2020 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*,

pages 79–84, Los Alamitos, CA, USA, aug 2020. IEEE Computer Society. 4, 7, 8

[29] X. Nie, M. A. Takalkar, M. Duan, H. Zhang, and M. Xu. Geme: Dual-stream multi-task gender-based micro-expression recognition. *Neurocomputing*, 427:13–28, 2021. 6, 7

[30] M. Peng, C. Wang, T. Chen, G. Liu, and X. Fu. Dual temporal scale convolutional neural network for micro-expression recognition. *Frontiers in Psychology*, 8:1745, 2017. 3

[31] Min Peng, Zhan Wu, Zhihao Zhang, and Tong Chen. From macro to micro expression recognition: Deep learning on small datasets using transfer learning. In *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, pages 657–661, 2018. 3

[32] W. Peng, X. Hong, Y. Xu, and G. Zhao. A boost in revealing subtle facial expressions: A consolidated eulerian framework. In *2019 14th IEEE International Conference on Automatic Face Gesture Recognition (FG 2019)*, pages 1–5, 2019. 7

[33] 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*, 7:184537–184551, 2019. 3, 7

[34] Y. Su, J. Zhang, J. Liu, and G. Zhai. Key facial components guided micro-expression recognition based on first amp; second-order motion. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2021. 7

[35] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio. Graph attention networks, 2018. 4

[36] L. Wang, H. Xiao, S. Luo, J. Zhang, and X. Liu. A weighted feature extraction method based on temporal accumulation of optical flow for micro-expression recognition. *Signal Processing: Image Communication*, 78:246–253, 2019. 3

[37] 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.*, 76(20):21665–21690, Oct. 2017. 3, 7

[38] 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 (Proc. SIGGRAPH 2012)*, 31(4), 2012. 2

[39] B. Xia, W. Wang, S. Wang, and E. Chen. Learning from macro-expression: A micro-expression recognition framework. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 2936–2944, New York, NY, USA, 2020. Association for Computing Machinery. 3

[40] B. Xia, W. Wang, S. Wang, and E. Chen. *Learning from Macro-Expression: A Micro-Expression Recognition Framework*, page 2936–2944. Association for Computing Machinery, New York, NY, USA, 2020. 7

[41] Z. Xia, X. Hong, X. Gao, X. Feng, and G. Zhao. Spatiotemporal recurrent convolutional networks for recognizing spontaneous micro-expressions. *CoRR*, abs/1901.04656, 2019. 7

[42] H. X. Xie, L. Lo, H. H. Shuai, and W. H. Cheng. Au-assisted graph attention convolutional network for micro-expression recognition. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 2871–2880, New York, NY, USA, 2020. Association for Computing Machinery. 4, 7, 8

[43] 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*, 9(1):1–8, 01 2014. 5

[44] Y. Zhang, H. Jiang, X. Li, B. Lu, K. M. Rabie, and A. U. Rehman. A new framework combining local-region division and feature selection for micro-expressions recognition. *IEEE Access*, 8:94499–94509, 2020. 3

[45] 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*, 29(6):915–928, June 2007. 1

[46] L. Zhou, Q. Mao, and M. Dong. Objective class-based micro-expression recognition through simultaneous action unit detection and feature aggregation. *CoRR*, abs/2012.13148, 2020. 4