****

**Enhancing Facial Expression Detection through Multi-Source Data Fusion of Graph Convolutional Networks and Convolutional Neural Networks**

**Muladi Muhamad1, Felix Indra Kurniadi1\***

*1Computer Science Program, School of Computer Science, Bina Nusantara University, 11530, Indonesia*

\* Corresponding author’s Email: [felix.indra@binus.ac.id](mailto:felix.indra@binus.ac.id), [muladi.muhamad@binus.ac.id](mailto:muladi.muhamad@binus.ac.id)

**Abstract:** Due to the problems of partial occlusion and illumination changes in real-life scenarios, facial expression detection becomes a challenging task. To solve this problem, this research proposes the utilization of multiple sources of facial data, including facial images and facial landmarks. Sub-neural networks are employed to extract each data feature, utilizing pre-trained models such as CNN and graph convolution network. The proposed method consists of three essential parts. First, feature extraction is designed to obtain global features from the facial image using the pre-trained model VGG19. Then, a graph convolutional neural network is used to capture structural information from landmarks, which will serve as the local feature. Furthermore, the extracted data will be fused to enhance the expression information. Experimental results demonstrate that on the CK+ facial expression dataset, the proposed method can effectively recognize facial expressions with 97.56% accuracy. Moreover, it achieves a 75% accuracy rate for the occlusion dataset, specifically for the eyes and mouth of the CK+ dataset. In contrast, the pre-trained model itself only achieves 92.05% accuracy for the CK+ dataset and 49% accuracy for the occlusion dataset.

**Keywords:** Facial expression recognition, Facial landmark, Graph convolutional neural network, Delaunay Triangulation

1. **Introduction**

Human emotions are the responses or reactions of individuals to events, circumstances, or stimuli. These emotions can be revealed through facial expressions, which serve as non-verbal modes of communication. Facial expressions convey an individual's emotional state, providing insights into their feelings. The seven basic emotions recognized are contempt, disgust, anger, fear, sadness, happiness, and surprise. Various internal and environmental factors, such as thoughts, feelings, behaviors, and level of satisfaction, contribute to the experience of these emotions [1].

The widespread use of cameras, coupled with recent advancements in machine learning, biometric analysis, and pattern recognition, has profoundly influenced the development of Facial Expression Recognition (FER) technology. This intersection of disciplines also has attracted the attention of researchers in the field of computer vision and human-computer interaction due to the rich information conveyed by facial expressions. In fact, facial expressions provide approximately 55% of the information related to human emotion in everyday communication, surpassing other social signals [2].

The task of facial expression recognition itself has already utilized various traditional recognition methods. One notable approach was conducted by Ghimire et al., where they employed position-based geometric features extracted from facial landmarks. This involved calculating angles and Euclidean distances between pairs of identified landmarks. The classification task was then performed using a Support Vector Machine (SVM) classifier [3]. In another study by Lanitis et al., the authors aimed to calculate the relative positions of facial landmarks for expression recognition. They employed a geometric feature-based approach, relying on earlier manual feature extraction methods. However, this approach proved to be inadequate due to the involvement of human intervention [4].

For example, Khan et al. (2015) utilized facial landmarks as the primary source of information for facial expression recognition. They trained a multi-layer perceptron (MLP) using these important facial points [5]. In another study, Simonyan and Zisserman (2014) investigated the impact of convolutional network depth on accuracy in large-scale image detection [6]. Atanasov et al. (2017) utilized a pre-trained model for detecting facial emotions [7]. Additionally, Tang (2018) employed a combination of convolutional neural networks (CNN) and Support Vector Machines (SVM) for expression recognition, and achieved superior results compared to traditional methods [8].

However, research on landmark-based Facial Expression Recognition (FER) and also on occlusion scenarios has been limited recently. This can be due to the lack of a suitable deep learning model specifically designed to extract relevant information from facial landmarks and occlusion scenarios. Currently, neural networks developed for extracting image features have been directly applied to landmark features without proper analysis. For instance, Multi-Layer Perceptron (MLP) used to analyzing dense relations between feature points. Therefore, a new algorithm is required to effectively examine the arbitrary relations between landmark points and utilize it to address the challenges posed by occlusion scenarios.

We propose a feature fusion approach that combines a graph convolutional neural network (GCN) with the pre-trained VGG19 model to leverage multiple sources of facial data. This approach is particularly beneficial for handling occlusion scenarios, as the multiple sources of data can help in accurately recognizing human emotions. The utilization of facial landmarks proves advantageous in detecting parts of the face that are not affected by occlusion. The pre-trained VGG19 model is employed to extract global features from the entire facial image. On the other hand, each facial landmark captures the positional information of facial muscles, contributing to the extraction of local features. In our approach, the facial landmarks serve as nodes in the graph, and the edges are constructed using the Delaunay triangulation method. This methodology allows us to capture structural information by focusing on specific facial regions, including the eyes, nose, lips, and eyebrows, utilizing the representation provided by facial landmarks.

1. **Dataset**

The CK+ dataset, also known as the Extended Cohn-Kanade dataset, is a widely utilized dataset for facial expression detection in the fields of computer vision and affective computing. It serves as an extension to the original Cohn-Kanade dataset, offering a larger collection of labeled facial expression data. The CK+ dataset comprises image sequences capturing various individuals' facial expressions. These expressions include anger, contempt, disgust, fear, happiness, sadness, and surprise. Each sequence starts from a neutral expression and progresses to the target expression. This dataset is valuable as it includes both posed and spontaneous expressions, allowing researchers to investigate both controlled and naturalistic expressions. In the CK+ dataset, only the last frame of each video capturing the peak expression is analyzed for facial expression recognition. Table 1 provides an example image for each expression category and displays the data distribution for each category within the dataset. Figure 1 illustrates the distribution of each class. It is worth nothing that each emotion class in the dataset consists of image data from different subjects, resulting in a total of 327 images.

Table 1. Distribution of the CK+ Dataset

|  |  |
| --- | --- |
| Emotion | Frequency |
| Anger | 45 |
| Contempt | 18 |
| Disgust | 59 |
| Fear | 25 |
| Happy | 69 |
| Sadness | 28 |
| Surprise | 84 |

Figure. 1 Distribution of data for each class

1. **Methodology**

This section of the research paper describes the facial expression recognition proposed methodology and architecture. The overall architecture of the proposed method is shown in Figure 2. In our approach, the facial expression recognition process involves several steps. Firstly, we acquire an image for each frame of the video and detect the location of the face. We then crop the image to the desired size for further processing (as described in Section 3.1).

Given the utilization of multiple sources of data, the input image undergoes different feature representation techniques. For the facial landmark feature-based approach, we extract landmarks from each frame using a landmark extractor (as described in Section 3.2). Subsequently, a graph structure is constructed based on the locations of these landmarks. To capture information among distant nodes, we introduce a master node in the graph structure (as discussed in Section 3.3). Finally, we employ a convolutional graph neural network to analyze the graph structure and extract relevant features (as outlined in Section 3.4). Regarding the image features extracted using the pre-trained VGG19 model, we modify the model by removing unnecessary parts such as the classifier layer (as outlined in Section 3.5). Additionally, we add a global average pooling layer to reduce the dimensionality of the extracted features. After completing the feature extraction process, the next step is to fuse the features using concatenation (as described in Section 3.6). Finally, we employ a fully connected layer to predict the emotion based on the fused features.

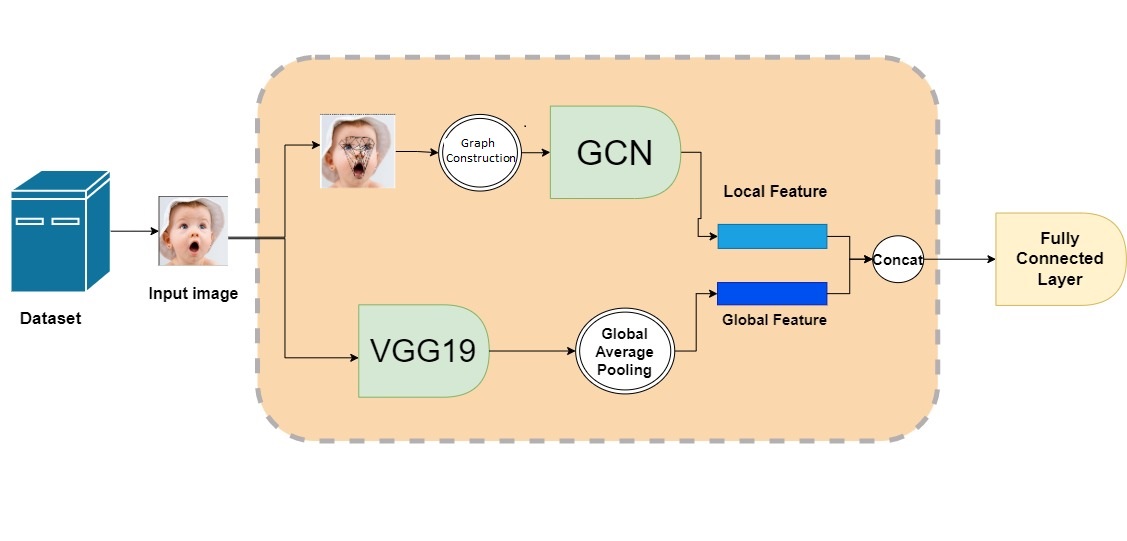


Figure. 2 Overview of the proposed method

* 1. Image Pre-Processing

To begin, the input image undergoes image pre-processing to eliminate any undesired noise. A common library used for face detection is Dlib [9], which offers a highly regarded face detection model. Dlib's face detector employs the Histogram of Oriented Gradients (HOG) feature descriptor along with a linear SVM classifier, known for its accuracy and robustness in face detection tasks. Once the face is detected using Dlib, the image is cropped based on the detected face region. To ensure compatibility with the VGG19 pre-trained model, the cropped image is resized to the recommended size of 227x227 pixels. This size ensures compatibility with the VGG19 model's input requirements.

Another crucial consideration is the distribution of data, particularly the imbalance observed in the distribution of each class, as illustrated in Figure 2. To address this issue and enhance the performance and results of the models, a useful technique called data augmentation is employed. Data augmentation involves generating new and diverse examples to augment the training dataset. The data augmentation technique implemented in this study is based on the method proposed by Jung et al. [13]. It involves applying various transformations and modifications to the existing data samples, such as rotation, flipping, and noise addition which is gaussian distribution.

In the context of facial landmarks, the data augmentation technique is also applied. The facial landmarks obtained from the images are subjected to various transformations, including rotation and flipping. This ensures that the facial landmarks capture the variations in facial expressions across different orientations and viewpoints. Additionally, the technique involves applying Gaussian distribution to the landmark points. This is done to generate a probability distribution around each landmark point, representing the uncertainty or variability associated with its location. By incorporating the Gaussian distribution, the model becomes more robust to slight variations in the exact positioning of the landmarks, allowing it to better generalize to unseen data.

Overall, the combination of rotation, flipping, and Gaussian distribution augmentation techniques helps to enhance the dataset's diversity and robustness, enabling the model to effectively capture the variations in facial expressions and improve its performance in facial landmark detection tasks.

* 1. Facial Landmark

Facial landmarks play a crucial role in various facial recognition tasks, including facial detection and emotion recognition. These landmarks encompass important facial features such as the mouth, nostrils, eyes, and eyebrows. To extract these features accurately, the utilization of a library is necessary. One popular library for landmark prediction is Dlib [9], which has proven to deliver excellent results in real-time applications.

The Dlib facial landmark detection algorithm employs machine learning techniques, specifically Active Shape Models (ASM), to precisely locate the facial landmarks. A pre-trained shape predictor model, trained on a large dataset, is utilized to detect 68 specific facial landmarks on a given face. These landmarks serve as key reference points for various facial analysis tasks, including face alignment, expression recognition, and estimation of facial attributes. This approach provides accurate identification of facial landmarks and has gained widespread acceptance in the field of computer vision.

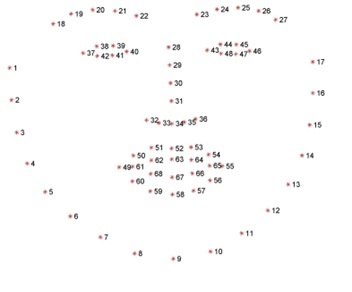


Figure. 3 Facial Landmark with 68 points

According to Figure 3, a face is divided into 68 landmark locations, each represented by coordinates (x, y). These landmark points correspond to specific regions of the face, including the left eye, right eye, nose, eyebrows, mouth, and jawline. However, for this particular study, only 51 landmark points are used as landmark features [10]. This selection is based on the significance and importance of these specific points in capturing the facial expressions and features necessary for the study's objectives.

|  |  |
| --- | --- |
|  | (1) |
|  |  |

where LM is the set of feature landmarks, () is the coordinates of x and y at the landmark point on the face and denotes the number of feature landmarks.

* 1. Graph Construction

Each unique landmark point on the face is significantly connected to other specific landmarks due to the underlying muscular connections. Consequently, the locations of these landmarks can be utilized to construct a graph structure that conveys valuable structural information. The initial step in building this graph structure involves translating the x and y coordinates of each landmark into node (V) attributes. This can be accomplished by following these steps:

|  |  |
| --- | --- |
|  | (2) |
|  |  |
|  | (3) |
|  |  |

In this method, the nodes in the graph represent the landmark features, where V denotes the node and v\_t represents the number of nodes. This strategy has been utilized in various studies where its effectiveness has been demonstrated [10]. Once the landmark features have been identified as nodes, the next step is to establish relationships between these nodes, specifically by creating edges.

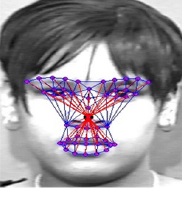
The Delaunay triangulation method is utilized in this study to establish connections between nodes, creating triangle meshes at key facial landmarks and forming a more intricate graph structure. However, this complex structure can make data transfer between distant nodes more challenging. To address this, a master node is introduced to facilitate efficient communication throughout the graph [11]. The master node is connected to all other nodes in the graph, enabling efficient information exchange through shortcut connections. In this study, the middle of the nose is selected as the location for the master node, resulting in a graph structure that effectively represents facial emotion data. Experimental results demonstrate that the proposed technique achieves exceptional performance.

Figure. 4 The red edges represent connections established based on the master node, while the blue edges indicate connections formed using the Delaunay method. This combination of edges enables efficient communication and information exchange within the graph structure, enhancing the representation of facial emotion data.

The relationships between nodes (edges) are established using the Delaunay triangulation, as depicted in Figure 4. The resulting triangulation is represented as an adjacency matrix (A), which captures the connections and relationships between nodes in the graph structure. This adjacency matrix is instrumental in visualizing and understanding the interconnections among the facial landmark nodes. The adjacency matrix is crucial in graph neural networks as it defines the neighborhood relationships and influences the propagation of information across the graph. It allows the model to capture and leverage the structural information present in the graph to make predictions or perform tasks such as facial expression recognition.

|  |  |
| --- | --- |
|  | (4) |

where denotes the adjajency matrix and denotes as delaunay triangulation method. However, a mesh structure only indicates whether edges are connected or not. Therefore, we using euclidan distance to represent the edge strength, as follows:

|  |  |
| --- | --- |
|  | (5) |
|  | (6) |

where denotes the euclidan distance between nodes. The relationship between nodes in the graph structure is represented by directed edges, where each edge has a source node and a target node. The value of each edge is determined based on the Euclidean distance between the corresponding nodes. This distance represents the spatial proximity between nodes in the graph.

The graph structure is composed of these features, including the nodes, directed edges, and their associated values. It captures the spatial relationships and connectivity patterns among the facial landmarks, allowing for the representation of complex structural information in facial expressions.

By utilizing this graph structure, advanced techniques such as graph convolutional neural networks (GCNs) can be applied to effectively analyze and process the information within the graph. These models leverage the connectivity and spatial relationships encoded in the graph structure to extract meaningful features and make accurate predictions in tasks such as facial expression recognition. In addition, these features compose a graph structure, which is defined as:

|  |  |
| --- | --- |
|  | (7) |

As a result, contains geometric information of facial emotions.

* 1. Graph Convolutional Neural Network

Graph Convolutional Networks (GCNs) have been developed to enable iterative learning of graph-level predictions by leveraging neighbor information. Recent research has focused on reducing the computational complexity associated with this process. GCNs leverage the principles of convolutional networks to extract information from both a node and its neighboring nodes within a graph. The process involves three main steps: message passing, aggregation, and update.

During message passing step in a GCN, the goal is to update the node features by incorporating information from their neighboring nodes. It can be represented as:

(8)

Where represents output feature matrix of GCN, is a activation function applied element-wise to the output , is the diagonal degree matrix of the adjajency matrix of , computed as D = ΣA , is the weight of matrix and lastly **X** as the input feature matrix. The next step is aggregation the neighbouring nodes with the source node, the aggregation that used in this research is the “add” aggregation method that simply sums up the messages from neighboring nodes. At the end of the step, the source node is updated based on the aggregated information, enabling GCNs to learn the structural information of the graph. This is particularly advantageous in facial expression recognition, as it allows the model to consider not only individual facial landmarks but also their relationships and contextual information.

In the context of facial expression recognition, the use of GCNs can be especially valuable when dealing with occlusion scenarios or capturing fine-grained details related to specific facial regions. By incorporating the graph structure built from facial landmarks, GCNs can effectively model the interactions between different facial muscles and capture the positional dependencies among landmarks.

In a Graph Convolutional Network (GCN), each layer represents the maximum travel distance for node features. Let's consider a one-layer GCN as an example, where each node can only gather information from its immediate neighbors. The process of accumulating information happens independently and simultaneously for all nodes. When we add another layer on top of the first layer, the information gathering process is repeated. However, this time, the neighboring nodes have already acquired information from their own neighbors. This iterative process of gathering information can be performed for multiple layers.

The number of layers to be used in a GCN is determined by the distance a node needs to traverse in order to obtain relevant information from the network. It is important to find a balance because going too far in the graph can be unnecessary. Typically, using 3 to 6 layers allows for comprehensive coverage of the entire graph, rendering further aggregation redundant [11].

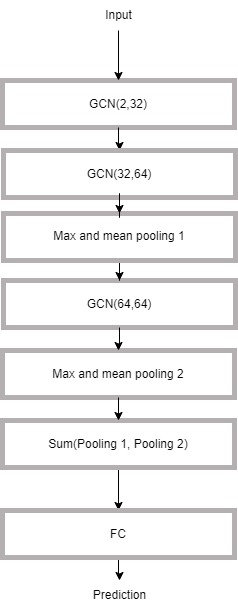


Figure. 5 Architecture of GCN

Figure 5 illustrates the architecture of the proposed Graph Convolutional Neural Network (GCN) model. The model comprises three GCN layers, where the first layer has a feature size of two nodes. To further enhance the model's representation and capture global information, a pooling layer is incorporated. The pooling layer used in this study combines global mean pooling and global max pooling. Global mean pooling calculates the average node feature across all node dimensions, resulting in a graph-level output that captures the overall representation of the graph. On the other hand, global max pooling selects the maximum value of the node feature across all node dimensions, emphasizing the most prominent features in the graph. The final layer of the model combines the outputs of each layer's pooling, effectively aggregating the global representations from multiple layers. This step contributes to the generation of a comprehensive and informative global representation of the graph.

* 1. Pre-trained VGG19

The pre-trained VGG19 model in this study is modified to match the input image size of 224 x 224. The process involves passing the image through multiple layers, resulting in a 7 x 7 x 512 array. However, this can lead to a high number of parameters. To address this issue, the previous classifier layer, which is not necessary for feature extraction, is removed. Instead, the global average pooling layer is utilized to reduce the number of learning parameters. This layer calculates the average of each feature map individually, consolidating the information into a single value per channel.

As the subsequent features will be combined with the features generated by the GCN, the resulting features are then mapped into a feature map with a size of 128. These modifications, including the removal of the classifier layer and the incorporation of the global average pooling layer, optimize parameter utilization and facilitate the integration of VGG19 features with those generated by the GCN.

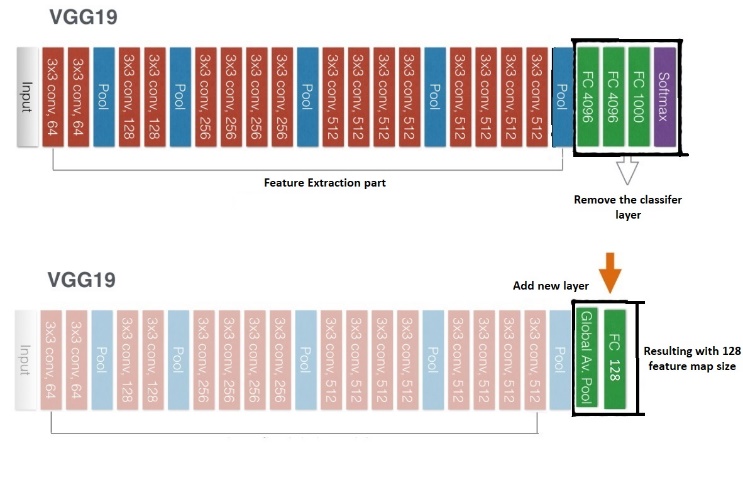


Figure. 6 Modification on VGG19 architecture

* 1. Feature Fusion

Feature fusion is an approach employed to combine multiple types of extracted features, with the objective of capturing complementary information from diverse feature representations and leveraging their respective strengths to enhance overall performance. There exist several fusion methods for combining different types of features, one of which is early fusion. In early fusion, the feature vectors from different modalities are concatenated or stacked together, forming a unified representation in the form of a single feature vector. This combined feature vector is subsequently utilized for classification tasks.

The neural network itself requires a total of two multi sources of data, which are images, and facial landmark, and denotes them as **I** , **L** respectively. When the neural network receives the data, it will first to use the sub networks to extract features as desribed on previous section.

(9)

(**L**) (10)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Method** | **Accuracy** | **F1-Score** | **Training Time (Epoch)** | **Type** |
| VGG19 [6] | 92.05 | 86 | 12-13s | Image |
| FerGCN [15] | 93 | 90 | 0.8-1s | Landmark |
| XGBoost[16] | 72 | 67 | - | Landmark |
| Ngoc et al [10] | 86 | 77 | 0.8-1s | Landmark |
| **Proposed Method** | **97.56** | **96.61** | **12-13s** | **Image+**  **landmark** |

Where F1, F2 respectively denote the feature map extracted by two of the sub networks GCN and VGG19. Then both of feature map will be concatenated and passed through to a fully connected layer to make an prediction that can be seen in figures 2.

1. **Experimental Result**
   1. Training Configuration

Prior to presenting the results, it is important to provide an overview of the training process parameters. The model underwent training for a total of 30 epochs, using a batch size of 64 and a learning rate of 0.001. To optimize the learning process, an Exponential LR learning scheduler was implemented with a gamma value of 0.80. The Adam optimizer was utilized with a weight decay of 0.01. In order to assess the model's performance and ensure its generalizability, a five-fold cross-validation procedure was used.

To address the challenge of imbalanced data, a focal loss function was utilized instead of the traditional cross-entropy loss. The focal loss function helps to address the issue of class imbalance by downweighting the contribution of easy examples during training and focusing the model's attention on misclassified and more challenging examples. This approach helps the model to learn more effectively from difficult instances, thus improving its ability to recognize facial expressions accurately. By employing these strategies, such as cross-validation, focal loss function, and parameter optimization, the aim is to enhance the model's performance in facial expression recognition, particularly in challenging environments.

* 1. Experimental Result

In this section, we present the experimental results on the CK+ dataset, which is widely used for emotion recognition research. The dataset consists of 327 video sequences, each labeled with one of seven emotional categories: anger, contempt, disgust, fear, happiness, sadness, and surprise. The videos in the dataset contain frontal face images without significant variations in illumination or different faces, making it suitable for our landmark extractor and the GCN model.

To evaluate the performance of our proposed method, we split the CK+ dataset into a training dataset (70%) and a test dataset (30%). The experimental results are shown in Table 1, where we compare our approach with state-of-the-art methods such as [6], [10], [15], and [16]. It is important to note that their specific methodology and approach may differ from ours. It is worth mentioning that by comparing our results with other state-of-the-art methods, we can evaluate the effectiveness and superiority of our proposed method in terms of its performance and accuracy in facial expression recognition tasks.

Table 2. Comparison of other methods on the CK+ dataset

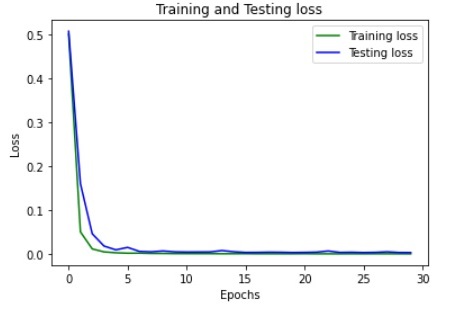


Figure. 6 Plot of training and testing loss on proposed method

Based on table 2, the experimental result shown that our proposed method provided a perfromance of 97.56% , outperforming another methods. Figure 6 depicts the training and testing loss, which reveals a the slight difference between training and testing loss indicates that the model is neither overfit nor underfit. In order to prevent overfitting, it is essential to consider an early stoppage during model training. Even though the results presented in Figure 6 do not involve early halting, this method is commonly employed in practice.

Table 3. Precision,recall, F1-score for each emotion

|  |  |  |  |
| --- | --- | --- | --- |
| Emotion | Precision | Recall | F1-Score |
| Anger | 0.94 | 0.96 | 0.95 |
| Contempt | 0.92 | 0.93 | 0.92 |
| Disgust | 0.97 | 0.99 | 0.98 |
| Fear | 0.96 | 0.95 | 0.96 |
| Happy | 0.98 | 0.99 | 0.99 |
| Sadness | 0.97 | 0.91 | 0.94 |
| Surprise | 0.99 | 0.99 | 0.99 |

Based on the precision, recall, and F1-score presented in Table 3. Different feature representations may capture different aspects or modalities of the data. By fusing these features, we can leverage the strengths of each representation and capture more comprehensive and diverse information. This can lead to a more robust and informative representation that enhances the model's ability to understand and discriminate between different classes or categories by provide a more discriminative and expressive feature representation. It also can help reduce redundancy in the feature space by selecting and combining the most relevant and informative features from different modalities. This can prevent overfitting and improve generalization performance.

It is evident that contempt and sadness present the greatest challenge in predicting emotions within the entire emotion model. This difficulty arises due to the similarity in facial landmarks associated with these emotions, such as the curvature of the eyebrows and lips. However, despite this challenge, the model demonstrates good performance in distinguishing and accurately classifying these emotions based on the learned feature representations. The model is capable of identifying the most suitable emotion based on the extracted features, thereby showcasing its effectiveness in capturing the subtle differences between emotions.

To further evaluate the performance of the proposed model, experiments were conducted to predict emotions from partially obscured facial images. Figure 7 illustrates how certain facial images in the dataset were intentionally obscured, simulating real-world scenarios. The objective of these experiments was to assess the model's capability to recognize emotions even when facial features are partially occluded, a common challenge in real-life situations. The CK+ dataset was modified to include two types of occlusion: one covering the eye area and another covering both the eye and oral areas. These occlusions were introduced to mimic scenarios where facial expressions may be partially obscured. By incorporating such occlusions in the dataset, researchers can assess the effectiveness of facial expression recognition algorithms in challenging conditions.

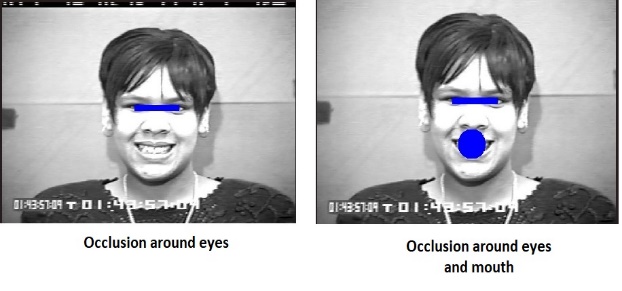


Figure. 7 An example of an image with an occlusion

Table 4. Result using occlusion eyes

|  |  |  |
| --- | --- | --- |
| **Method** | **Accuracy** | F1-Score |
| VGG19[15] | 85 | 76 |
| FerGCN [7] | 84 | 79 |
| XGBoost[14] | 64 | 42 |
| Ngoc et al [6] | 83 | 76 |
| **Proposed Method** | **91** | **89** |

Table 5. Result using occlusion eyes and mouth

|  |  |  |
| --- | --- | --- |
| **Method** | **Accuracy** | F1-Score |
| VGG19[15] | 49.52 | 27 |
| FerGCN [7] | 67 | 58 |
| XGBoost[14] | 59 | 37 |
| Ngoc et al [6] | 68 | 61 |
| **Proposed Method** | **75** | **72** |

The results of various methods for recognizing facial expressions in the presence of occlusions on the eyes and eyes and mouth regions are presented in Tables 4 and 5, respectively. The proposed method obtained the highest accuracy of 91 percent and the highest F1-Score of 89 percent for occlusion detection on eyes, outperforming the other methods. Also for occlusion detection on eyes and mouth obtained highest accuracy of 75 percent and the highest F1-Score of 72 percent. Overall, the proposed method demonstrates promising results in recognizing facial expressions in the presence of occlusions, which can be useful in real-world applications where partial facial obstructions are prevalent.

1. **Conclusion**

This research introduces a method for recognizing facial expressions that overcomes challenges like partial occlusion. The method combines facial images and facial landmarks to achieve accurate and reliable expression detection. The process involves several steps. First, we extract overall facial features from images using a pre-trained model called VGG19. This captures the general facial information and provides context for expression recognition. Next, we locate key facial features using the Dlib library, creating a graph structure that represents the face's structure. To analyze the graph structure effectively, we use a graph convolutional neural network (GCN). The GCN extracts local features from the facial landmarks, considering their spatial relationships and connections. By combining the global features from the facial image with the local features from the landmarks, we enhance the expression information.

Experimental results show that our method is effective. On the CK+ facial expression dataset, it achieves a recognition accuracy of 97.56%. It also performs well on an occlusion dataset, specifically for the eyes and mouth regions. Comparatively, the pre-trained VGG19 model alone achieves lower accuracy rates on both datasets, highlighting the importance of using facial landmarks and the graph structure. Future work can focus on refining the proposed method further and exploring additional techniques for combining and learning features. Additionally, applying the method to larger datasets and real-world scenarios will help assess its performance and applicability. Overall, this research paves the way for advancements in facial expression recognition and the development of more accurate emotion detection systems.

Conflicts of Interest

The author declare no conflict of interest

Author Contributions

Conceptualization, MM and FIK; methodology, MM; validation, MM; formal analysis, MM; resources, MM; data curation, MM; writing—original draft preparation, MM; writing—review and editing, FIK; visualization, JS, KB, and A; supervision, FIK;

References

[1] Liu, Y., Sourina, O., & Nguyen, M. K. (2010). Real-time EEG-based human emotion recognition and visualization. 2010 International Conference on Cyberworlds, 262–269. https://doi.org/10.1109/CW.2010.37

[2] Z. Ming, A. Bugeau, J.-L. Rouas, and T. Shochi, “Facial Action Units intensity estimation by the fusion of features with multi-kernel Support Vector Machine,” in Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on, 2015, vol. 6, pp. 1–6.

[3] D. Ghimire and J. Lee, “Geometric Feature-Based Facial Expression Recognition in Image Sequences Using Multi-Class AdaBoost and Support Vector Machines,” Sensors, vol. 13, no. 6, pp. 7714–7734, Jun. 2013, doi: 10.3390/s130607714.

[4] Lanitis A, Taylor C J, Cootes T F. Automatic interpretation and coding of face images using flexible models[J]. IEEE Transactions on Pattern Analysis and machine intelligence, 1997, 19(7): 743-756.

[5] Khan F, Facial Expression Recognition using Facial Landmark Detection and Feature Extraction via Neural Networks. 2020.https://doi.org/10.48550/arXiv.1812.04510

[6] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. https://doi.org/10.48550/ARXIV.1409.1556

[7] A. Atanassov and D. Pilev, "Pre-trained Deep Learning Models for Facial Emotions Recognition," 2020 International Conference Automatics and Informatics (ICAI), Varna, Bulgaria, 2020, pp. 1-6, doi: 10.1109/ICAI50593.2020.9311334.

[8] Tang Y. Deep learning using support vector machines[J]. CoRR, abs/1306.0239, 2013.

[9] D. E. King, “Dlib-Ml: A Machine Learning Toolkit,” J Mach Learn Res, vol. 10, pp. 1755–1758, Dec. 2009.

[10] Q. T. Ngoc, S. Lee, and B. C. Song, “Facial Landmark-Based Emotion Recognition via Directed Graph Neural Network,” Electronics, vol. 9, no. 5, p. 764, May 2020, doi: 10.3390/electronics9050764.

[11] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, “Neural Message Passing for Quantum Chemistry,” in Proceedings of the 34th International Conference on Machine Learning - Volume 70, in ICML’17. JMLR.org, 2017, pp. 1263–1272.

[12] Markan, R., Kaur, G. (2013). Literature survey on elliptic curve encryption techniques. International Journal of Advanced Research in Computer Science and Software Engineering, 3, 908.

[13] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, “Joint Fine-Tuning in Deep Neural Networks for Facial Expression Recognition,” in 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile: IEEE, Dec. 2015, pp. 2983–2991. doi: 10.1109/ICCV.2015.341.

[14] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, “Focal Loss for Dense Object Detection,” in 2017 IEEE International Conference on Computer Vision (ICCV), Venice: IEEE, Oct. 2017, pp. 2999–3007. doi: 10.1109/ICCV.2017.324.

[15] L. Liao, Y. Zhu, B. Zheng, X. Jiang, and J. Lin, “FERGCN: facial expression recognition based on graph convolution network,” Mach. Vis. Appl., vol. 33, no. 3, p. 40, May 2022, doi: 10.1007/s00138-022-01288-9.

[16] A. V. Divekar and D. C. Gharpure, “Low-compute facial expression recognition using fiducial feature-sets,” IOP Conf. Ser. Mater. Sci. Eng., vol. 1187, no. 1, p. 012025, Sep. 2021, doi: 10.1088/1757-899X/1187/1/012025.

[17] Kavita and R. S. Chhillar, “Face Recognition Challenges and Solutions using Machine Learning,” Int. J. Intell. Syst. Appl. Eng., vol. 10, no. 3, pp. 471–476, 2022.

[18] J. Wan et al., “Robust facial landmark detection by cross-order cross-semantic deep network,” Neural Netw., vol. 136, pp. 233–243, Apr. 2021, doi: 10.1016/j.neunet.2020.11.001.

[19] K. Rohr, H. S. Stiehl, R. Sprengel, T. M. Buzug, J. Weese, and M. H. Kuhn, “Landmark-based elastic registration using approximating thin-plate splines,” IEEE Trans. Med. Imaging, vol. 20, no. 6, pp. 526–534, Jun. 2001, doi: 10.1109/42.929618.

[20] K. Krishnaveni and G. R. Priyadharsini, “Facial Expression Recognition using Low Level Histogram Features,” in 2020 Fourth International Conference on Inventive Systems and Control (ICISC), Coimbatore, India: IEEE, Jan. 2020, pp. 1–7, doi: 10.1109/ICISC47916.2020.9171050.

[21] T. N. Kipf and M. Welling, “Semi-Supervised Classification with Graph Convolutional Networks,” in 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings, OpenReview.net, 2017. [Online]. Available: https://openreview.net/forum?id=SJU4ayYgl

[22] Hong, S. (n.d.). An introduction to graph neural network (GNN) for analyzing structured data. https://towardsdatascience.com/an-introduction-to-graph-neural-network-gnn-for-analysing-structured-data-afce79f4cfdc (accessed: 08.11.2022).

[23] Khan, F. (2018a). Facial expression recognition using facial landmark detection and feature extraction via neural networks. CoRR, abs/1812.04510. http://arxiv.org/abs/1812.04510.

[24] Khan, M. A., Ashraf, I., Alhaisoni, M., Damaˇseviˇcius, R., Scherer, R., Rehman, A., & Bukhari, S. A. C. (2020). Multimodal brain tumor classification using deep learning and robust feature selection: A machine learning application for radiologists. Diagnostics, 10(8). https://doi.org/10.3390/diagnostics10080565.

[25] Li, J., Mi, Y., Yu, J., & Ju, Z. (2019). A novel convolutional neural network for facial expression recognition. https://doi.org/10.1007/978-981-13-7986-4\_28.