Journal of Artificial Intelligence and Data Science Techniques

ISSN: 3029-2794

Volume 01 Issue 03 (September) 2024

Global and local feature analysis for video-based facial expression identification

Tan Wei Ling[1] and Mohd Amirul Arif[2]

[1]Reseach Assistant, Department of Computer Science, Universiti Malaya, Malaysia
[2]Professor, Department of Artificial Intelligence, Universiti Teknologi Malaysia, Malaysia

Abstract:
In recent years, with the increasing demand for surveillance and security systems in public places and for understanding human behaviour, an automated expression understanding from video is receiving significant attention. Video-based facial expression analysis has significantly increased and is applied to different states of individuals in healthcare, security, and human-computer interaction. The potential of accurately identifying emotions from facial expressions in video motion can enhance communication and user experience. However, extracting facial expressions in dynamic motion, like raising an eyebrow, is a complex mechanism compared to actions in static positions. Thus, with the development of deep learning models, the research introduces a Global-Local feature-assisted Dynamic Convolutional Neural Network (Glo-DynaCNN) for understanding facial expression changes across dynamic video frames. This network model's main objective is to dynamically extract global features and temporal characteristics to analyze the expressions in different frames. At the same time, local features focus on critical facial regions, including mouth and eye position, which extract expression variations. The proposed model dynamically improves facial expression identification accuracy with these feature extractions. For this, the research employed publicly available video facial expression datasets that demonstrate, with annotations for happiness, sadness, anger, surprise, fear, and disgust expressions, achieving significant improvements over existing recognition models. The model's performance is evaluated using metrics like identification accuracy, precision, recall, F1-score, confusion matrix, and Mean Square Error (MSE) with an improved analysis rate

Keywords: Global Feature Analysis, Local Feature Analysis, Facial Expression Identification, Feature Extraction, Video Analysis.

1. Introduction

Video-based facial expression identification is essential for analyzing human emotional patterns. The primary emotions include happiness, sadness, anger, fear, disgust and surprise [1]. Facial expressions are associated with the global and local face regions with large no. of identities to extract independent expression through the reconstruction process in the earlier system [2]. Facial expression identification in the form of varying poses and facial occlusion helps understand the critical regions of facial features with multiscale analysis [3]. The safety video monitoring systems incorporate geometric pattern analysis of facial characteristics to identify emotional states in different environments [4]. The performance of six basis emotion identification from facial expression still involves a challenging process, applying principal component analysis for feature reduction and integrating with the linear embedding mechanism [5]. The textural image features are analyzed using local binary and ternary patterns to identify facial expressions with the trained CNN model [6] using various datasets.

The local and global feature learning with extraction process is based on the priority of core face features on discriminative assistance for symmetrical feature analysis in the

Journal of Artificial Intelligence and Data Science Techniques
ISSN: 3029-2794
Volume 01 Issue 03 (September) 2024

convolutional attention block [7]. These analyzed local and global perception views on facial expression are observed from regions of interest in the attention module at the decision level for robust identification [8]. The local spatial-temporal features globally learn temporal cues for recognizing emotional changes in each critical joint [9] in the convolutional model for different behaviours like depression and pain identification and dynamic facial encodings at multiple spatial dimension patterns [10].

For recognizing expressions, the features of mouth and eye variations influence facial expression's emotional representation of varying low- and high-level features [11,12]. Videos can promote a range of emotional responses among viewers with the challenge of low latency, with an accurate facial identification system for different emotions using a deep convolutional neural network [13]. The initial process begins with extracting the necessary local and global features with the histogram of oriented gradient descriptors to locate the region of facial emotion expressions [14]. The key highlights of this research are listed below:

1) To develop a novel Global-Local feature-assisted Dynamic Convolutional Neural Network (Glo-DynaCNN) for enhanced facial expression identification in video analysis

2) To demonstrate superior identification accuracy compared to existing models, particularly in handling expression cues of happiness, sadness, anger, surprise, disgust, and fear in video frames.

3) To enhance expression identification by capturing global and local facial movement features, improving model performance.

4) Conduct detailed performance comparisons using metrics like identification accuracy, precision, recall, f1-score, confusion matrix and MSE.

An outline of the research article is given below. Section 2 reviews recent articles to analyze previous facial expression recognition models in videos thoroughly. The implemented research model focuses on identifying dynamic features locally and globally. Features are selected through the CNN model with different video analyses discussed in Section 3. Expression Identification results with an improved performance metrics comparison are covered in Section 4. The conclusion of the research idea explores the summarized vital findings and the scope for future work.

2. Existing Research Methodology

Yan et al. [15] integrated global and local information through Principal Component Analysis (PCA). The proposed Fusion of Multi-Feature Local Directional Ternary Pattern (FMF-LDTP) approach improves facial expression identification. The technique achieves a 96.5% identification rate on the more extensive Japanese Female Facial Emotions Database (JAFFE), which includes regions, by focusing on essential areas like the eyes and mouth and avoiding irrelevant features. Enhanced accuracy and robustness are

Journal of Artificial Intelligence and Data Science Techniques
ISSN: 3029-2794
Volume 01 Issue 03 (September) 2024

benefits, but the computational complexity that comes with multi-feature extraction is a possible downside.

Sheron et al. [16] presented a Projection-Dependent Input Processing (PDIP) approach to HRI object recognition that uses multi-dimensional projection analysis and labelled analysis to separate incorrectly identified indices. It guarantees strong identification by reaching a 96.4% recognition ratio in 630.36 ms with reduced complexity (5.93) and mistakes (0.605). Optimal performance is dependent on precise labeling and computing overhead is an issue.

Jiang et al. [17] presented a new approach to face expression identification using a convolutional block attention module (CBAM) and multi-feature fusion. It also includes a loss function for local feature clustering that improves class separation during training. By emphasizing local expression features, the strategy outperformed state-of-the-art approaches in recognition tests conducted on the RAF and CK+ datasets. The benefits include enhanced local identification of features and differentiation between classes, while the downsides could include more complicated training.

Chen et al. [18] applied temporal-modeling adapters and facial landmark-aware prompters with the Static-to-Dynamic (S2D) model to improve dynamic recognition of facial expressions by combining knowledge of static recognition with dynamic facial characteristics. A model that improves performance with little parameter increase is tested on static and dynamic datasets. It uses emotional anchors-based self-distillation losses to reduce ambiguity in emotion labels. The result demonstrated a higher recognition accuracy and practical model implication, but it also has certain drawbacks, such as the possibility of not adequately capturing complicated dynamic expressions.

Li et al. [19] proposed an Adaptive Multiscale Correlation Module (AMSCM) to extract small and large face movements from RGB frames video-based facial expression analysis. This method detects little facial movements Without optical flow or explicit motion labelling. Results on the various datasets show that it performs at the cutting edge using CNNs, which may still have difficulty with highly nuanced expressions. It is a possible downside, but rapid extraction of various motion features is advantageous.

Singh et al. [20] suggested a hybrid neural network combining 3D-CNN and Convolutional Long-Short Term Memory (ConvLSTM) for video-based facial expression analysis. This network can capture time-related and spatial information, preserving critical spatial aspects. The model is optimized for real-time applications on platforms with limited resources, as it achieves a higher identification rate with improved accuracy while being more efficient and lightweight. It has been tested on three different datasets. Although it has fewer parameters and faster execution, it may not be able to compete with more complicated state-of-the-art models on more extensive datasets.

3. Proposed methodology

a. Dataset Study

The dataset observed from [21] contains video samples. Each ZIP file contains 50 annotated videos with highly accurate per-frame annotations and varying levels of valence and arousal based on per-frame annotations. This includes 68 facial landmarks with 600 challenging video clips. This dataset contains approximately 1809 videos of

Journal of Artificial Intelligence and Data Science Techniques
ISSN: 3029-2794
Volume 01 Issue 03 (September) 2024

facial expressions, including videos of individuals expressing six basic emotions listed in Table 1: happiness, sadness, anger, surprise, fear, and disgust.

TABLE 1: *DATA SOURCE DESCRIPTION*

Emotion Categories	Video Counts
Happiness	571
Sadness	527
Anger	485
Surprise	465
Fear	215
Disgust	256

Facial expression identification using the suggested Glo-DynaCNN architecture begins with video input and extracting relevant frames for analysis. Feature analysis includes two aspects: one captures the entire face structure, called a global representation of facial expression analysis, and the other focuses on specific regions like the eyes and mouth positions, like raising and spurring, to detect subtle expression cues. Initially, the data preparation procedures, such as normalization, improve the input quality, and then the videos are cropped into frames for analysis. Adaptive feature fusion is used to merge these features into a comprehensive set. The expression identification model uses this enhanced data to categorize emotions, including surprise, anger, sadness, fear, happiness, and astonishment. As a last step, we compare the model's performance to that of existing methods and use these metrics to conclude how well the proposed approach recognizes facial expressions depicted in Figure 1.

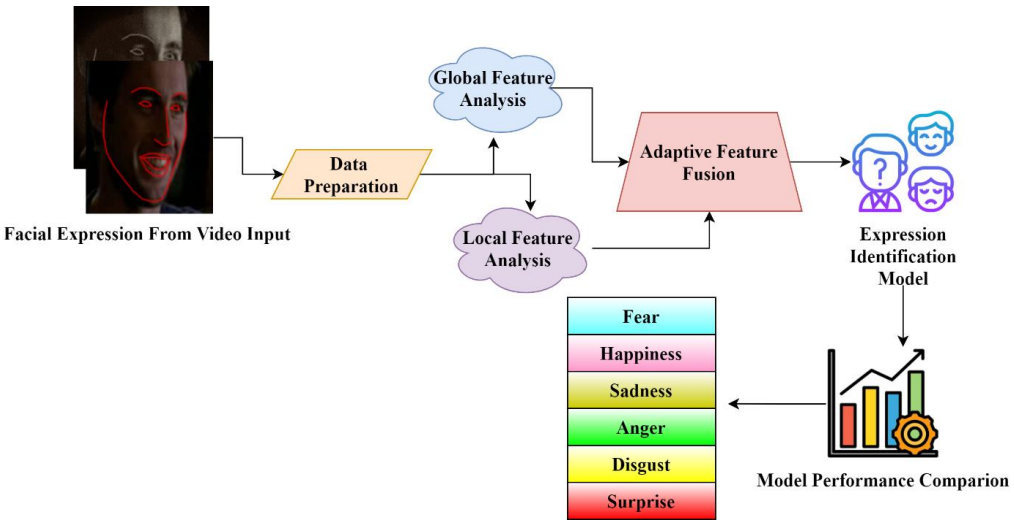

Figure 1: Overall Proposed Architecture

Journal of Artificial Intelligence and Data Science Techniques
ISSN: 3029-2794
Volume 01 Issue 03 (September) 2024

b. Data Preparation

For input video resizing and cropping in a training stage, the input samples are resized a sample video to 128 × 128 pixels and randomly cropped to 112 × 112 pixels. This representation is then transformed to standardize the input size for the model with the batch size fixed as 32.

c. Analyzing Global Features

Geometrical Feature: The distance between key points includes the overall shape and facial structure regarding the eyes, nose, and mouth. These distances can be mathematically computed using equation (1)

$$D_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \quad (1)$$

Where (x_i, y_i) and (x_j, y_j) are the coordinates representing the facial expressions i and j . Distance provides insights into the overall facial structure, such as the shape of the regions like eyes, nose, and mouth.

Texture Pattern Analysis: The general structure of the skin that the lighting conditions can impact. The local binary pattern (LBP) is applied and calculated using equation (2) for this texture pattern analysis.

$$LBP(x, y) = \sum_{n=0}^{N-1} s(g_n - g_c) \cdot 2^n \quad (2)$$

Where g_n indicates the pixel value of the neighbouring pixels, g_c indicates the center pixel value, s represents the step function that returns one if the condition is met, and N is the no. of neighbouring pixels considered for facial expression identification.

Color Analysis: The distribution of colours varies across the face for different expression states. For instance, redness indicates the anger emotion. The distribution of colours can be computed using the RGB model or transformed into HSV color space, where redness can be described using an equation (3) given below:

$$R = \frac{R_{pixel}}{R_{pixel} + G_{pixel} + B_{pixel}} \quad (3)$$

The variable R shows the normalized redness value of the pixel, ranging from 0 to 1. The parameter R_{pixel} represents the intensity of the red component of a specific pixel in the frame cropped and aligned from a video clip. The parameter G_{pixel} represents the intensity of the green component of the same pixel, whereas B_{pixel} indicates the intensity of the blue component of the same pixel. A higher value represents a more substantial presence of red that may correlate with anger.

Journal of Artificial Intelligence and Data Science Techniques
ISSN: 3029-2794
Volume 01 Issue 03 (September) 2024

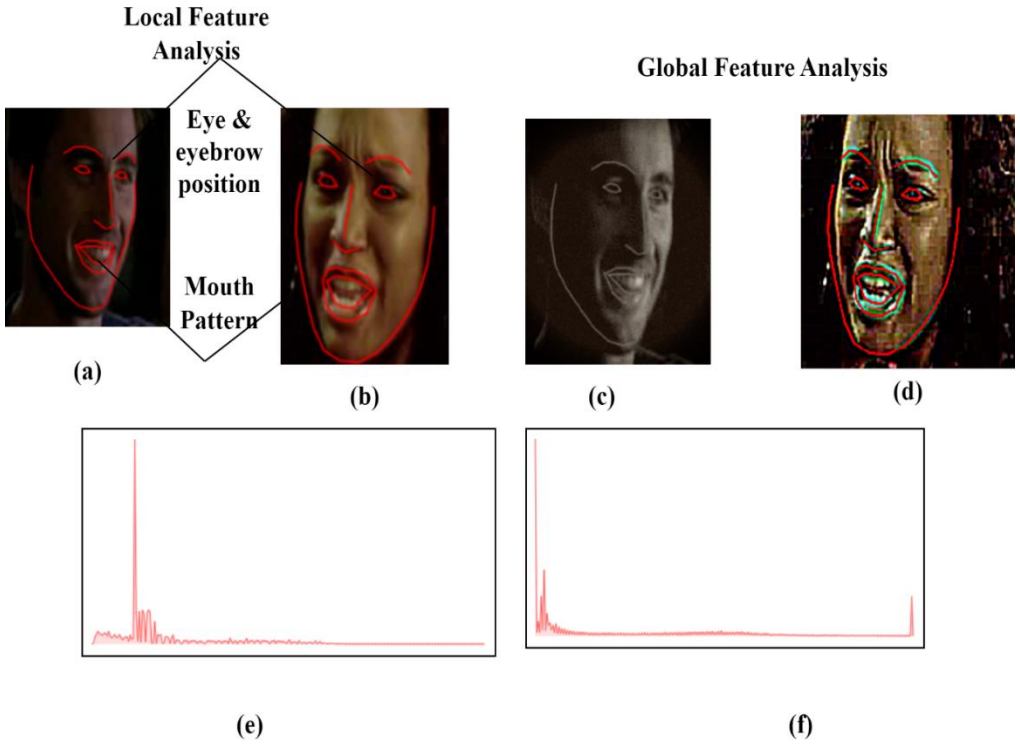

Figure 2: (a) and (b) Local Feature Analysis (c) and (d) Global Feature Analysis (e) and (f) Image Histogram of Feature Samples

Figure 2 illustrates the images of individuals with different facial expressions, including eye and eyebrow positions, along with mouth patterns in Figures 2(a) and 2(b). The expressions are conveyed based on the positioning of eyes, and eyebrows play a significant role; for instance, raised eyebrows indicate surprise and furrowed brows signify anger. Likewise, lip movement and position variation convey expressions like happiness and sadness. By analyzing these local features, the model better recognizes and classifies specific emotions based on facial movements in Figures 2(c) and 2(d). The comparison of histograms of different expressions depicted in Figures 2(e) and 2(f) captures the overall contours and shapes of the face.

d. Local Features:

A particular region of the face focuses on identifying the emotion identification that uniquely analyzes the expression. The eye region plays a significant role in identifying emotions like happiness, sadness, and surprise based on the eyebrow's position and eye openness, which impact expression recognition. Mouth region: The mouth is essential for expressing joy and anger. It is used to analyze lip movements to identify an emotional state quickly.

e. Adaptive Feature Fusion

An adaptive feature fusion process incorporates an attention mechanism to weigh the importance of global and local features dynamically based on the cropped current frame in the annotated video. The attention weights α_t for representing the global features and β_t for the local features at the time, t is computed as follows

$$\alpha_t = \sigma(W_\alpha * T_t + b_\alpha) \quad (4)$$

Journal of Artificial Intelligence and Data Science Techniques
ISSN: 3029-2794
Volume 01 Issue 03 (September) 2024

$$\beta_t = \sigma(W_\beta * T_t + b_\beta) \quad (5)$$

Where σ represents the sigmoid function that constrains the weights between 0 and 1. The other parameters from equations (4) and (5) like W_α and W_β indicates learnable weight matrices specific to global and local features in addition to the bias terms b_α and b_β. The fused features F_t at time t can be computed using an equation (6)

$$F_t = \alpha_t \odot G_t + \beta_t \odot L_t \quad (6)$$

The operator $\odot$ indicates element-wise multiplication, allowing the model to emphasize the most relevant features based on the learned attention weights observed from the facial expression. Through this adaptive fusing procedure, the network can improve the classification's resilience by prioritizing features contributing more to face expression recognition.

f. Model Training

Input Layer: This layer receives the input from fused global and local features as input. Convolutional Layer: In this layer, convolutional operations to the input features are performed, extracting hierarchical features that help understand patterns like textures, edges and facial shapes. Each convolution layer uses the operation using an equation (7)

$$f(x) = \sum(w_i \cdot x_i + b) \quad (7)$$

Where w_i represents learnable weights and the input feature map is termed as x_i, with the bias term b. An activation function, the Rectified Linear Unit (ReLU), is applied to introduce non-linearity, which allows the model to capture complex facial expressions. Dynamic Feature Extraction: The model can capture temporal dependencies and introduce recurrent connections between frames. The hidden state h_t at time t can be expressed using an equation (8)

$$h_t = \tanh(W_h * [F_t, h_{t-1}] + b_h) \quad (8)$$

The weight matrix is given as W_h that integrates both the current fused features F_t and the previous hidden state h_{t-1} with the bias term. This recurrent computation model allows the proposed Glo-DynaCNN to represent the facial expressions as they evolve across varying frames. Fully Connected Layers: After performing convolution and temporal feature extraction, fully connected layers combine all features and feed them for final expression classification among one of the expressions happiness, sadness, anger, surprise, fear, and disgust.Output Layer: This layer outputs derived in equation (11) show the final predicted facial expression using the softmax activation function that can convert the output scores into probabilities for each expression.

$$softmax(x_i) = \frac{exp(x_i)}{\sum exp(x_j)} \quad (9)$$

This model can classify the input as one of the predefined annotated expression categories. The dynamic CNN is trained end-to-end using backpropagation through time (BPTT) to minimize the loss function, which is typically the cross-entropy loss for classification tasks, as derived in equation (10).

$$L = -\sum\left(y_{true} \cdot log\left(y_{pred}\right)\right) \quad (10)$$

Facial expressions captured from video input are processed using two aspects of the feature evaluation approach, global and local analysis, based on frames per video. This approach efficiently compares the model's effectiveness across different expression

Journal of Artificial Intelligence and Data Science Techniques
ISSN: 3029-2794
Volume 01 Issue 03 (September) 2024

categories and improves emotion categorization with adaptive feature fusion and dynamic CNN.

4. Results and Discussion

In this research, the existing algorithms such as FMF-LDTP [15], CBAM [17], AMSCM [19], and the proposed Glo-DynaCNN algorithm are taken for comparison study. For the following metrics for performance evaluation like identification accuracy, precision, recall, F1-score, confusion matrix, and Mean Square Error (MSE) on varying frames limited up to 10 counts per video and for varying six different expressions of happiness, sadness, anger, surprise, fear, and disgust.

a. Identification accuracy

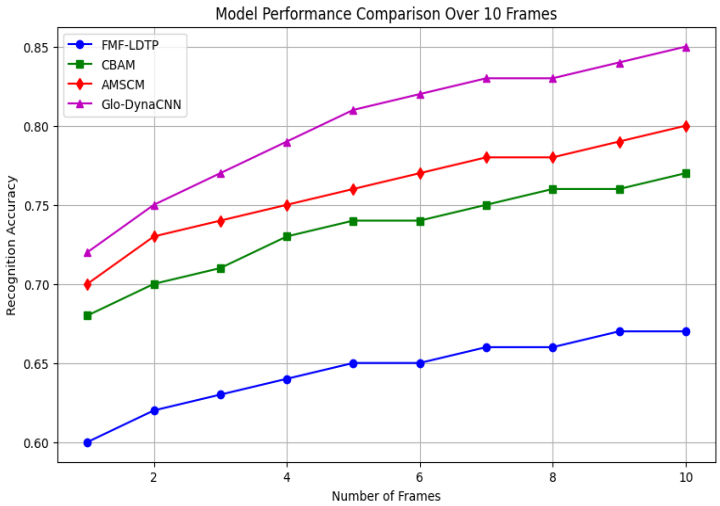

Figure 3: Recognition Accuracy Analysis

Figure 3 compares the performance of several emotion recognition models during a 10-frame period. The planned Glo-DynaCNN, FMF-LDTP, CBAM, and AMSCM are each shown as independent lines. The number of frames utilized in the facial expression identification is shown on the x-axis, while the recognition accuracy is shown on the y-axis. Glo-DynaCNN regularly outperforms the competition in terms of accuracy, especially when dealing with video data containing subtle emotional cues; this is true even when the no. of frames grows. While CBAM and AMSC have comparable trends, their accuracy levels are lower, indicating that the other models perform worse. The research goal is to improve the accuracy of expression identification from different facial patterns with varying feature detection using advanced modelling techniques, and this comparison shows that the suggested model is better at identifying expression tasks.

b. Precision and Recall

By calculating these performance metrics using ground truth input facial expressions and projected labels and then visualizing the results in a bar chart as shown in Figure 4, this algorithm compares the recall and precision of four models with an existing algorithm like FMF-LDTP, CBAM, AMSCM, and the proposed Glo-DynaCNN.

Journal of Artificial Intelligence and Data Science Techniques
ISSN: 3029-2794
Volume 01 Issue 03 (September) 2024

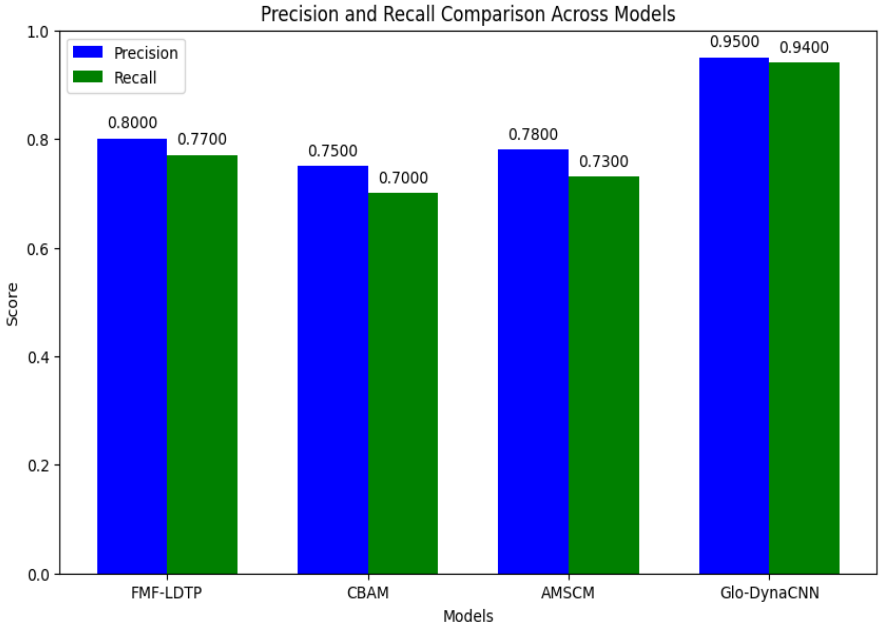

Figure 4: Precision-Recall Comparison Analysis

Regarding recall and precision, the results reveal that Glo-DynaCNN is the most effective and accurate model for classification tasks. In light of this, the suggested usage of Glo-DynaCNN is justified since it has proven to be an effective model for enhancing decision-making in applications like facial expression identification and performs feature extraction in both global and local ways of each facial region.

c. F1-Score

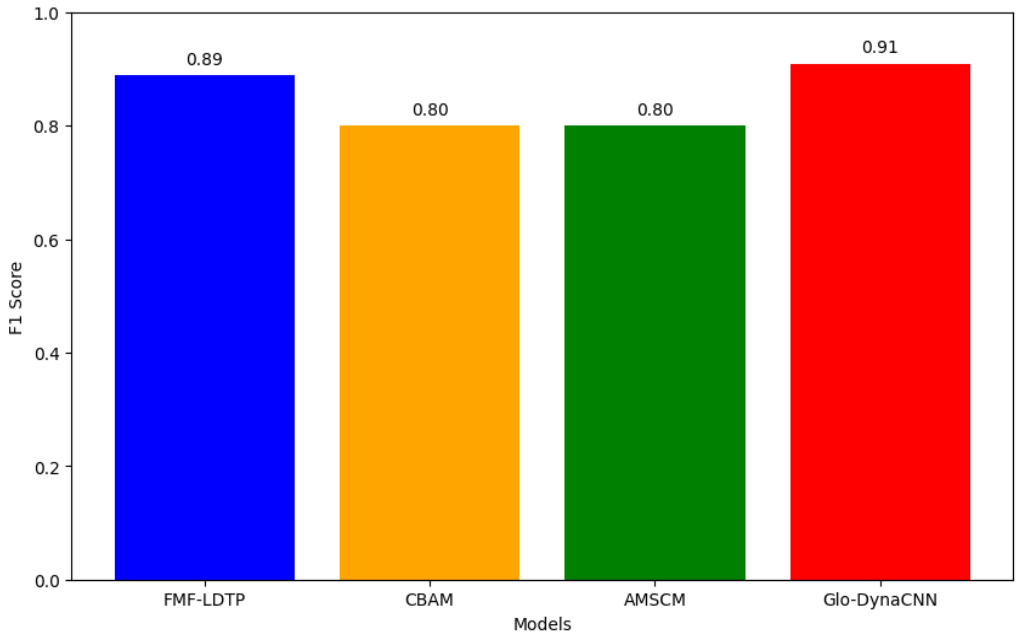

Figure 5: F1-Score Analysis

Journal of Artificial Intelligence and Data Science Techniques
ISSN: 3029-2794
Volume 01 Issue 03 (September) 2024

Figure 5 depicts simulated ground truth labels across ten frames; the plot compares the F1 scores of four models: FMF-LDTP, CBAM, AMSCM, and the proposed Glo-DynaCNN. Glo-DynaCNN has the best performance in accurately categorizing emotions, as seen by its highest F1 score. This finding demonstrates the model's efficacy and promise for enhancing expression recognition tasks, contributing significantly to the field's future facial expression applications.

d. Confusion Matrix

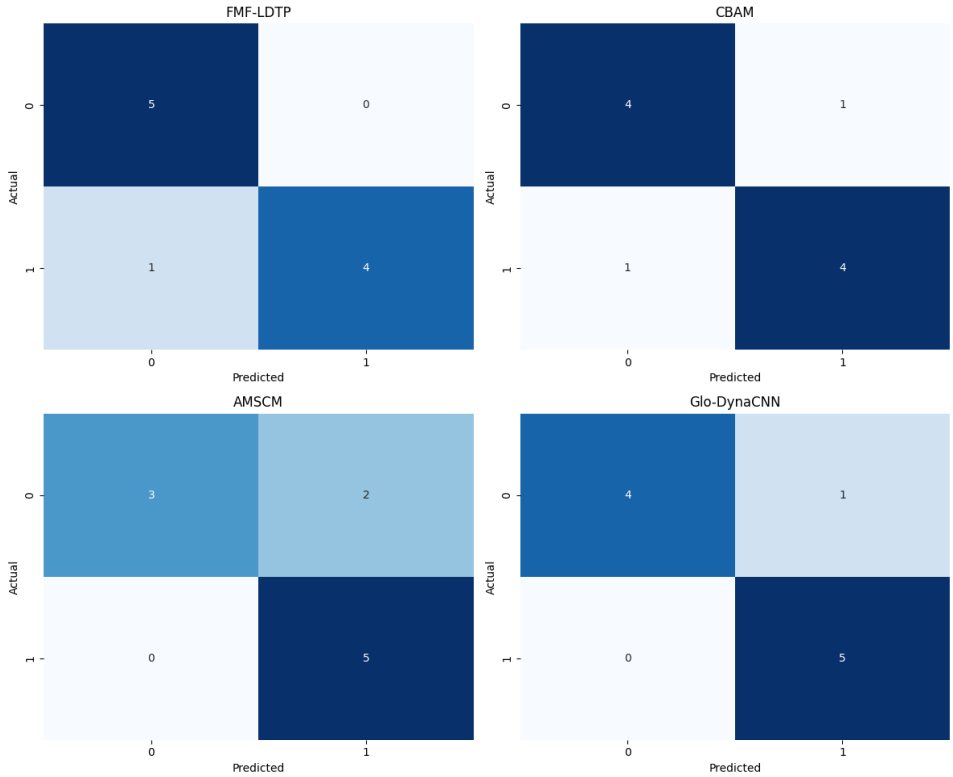

Figure 6: Confusion Matrix Comparison

Figure 6 illustrates the confusion matrices for four models tested for facial expression classification: FMF-LDTP, CBAM, AMSCM, and Glo-DynaCNN. The confusion matrices visually show the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) for each model to compare the models' performance in emotion classification tasks. Glo-DynaCNN and AMSCM reached five true positives, showing they are good at detecting positive emotional expressions. Nevertheless, Glo-DynaCNN surpasses AMSCM due to its lower occurrence of false positives, which lowers the probability of incorrectly labelling negative cases as positive. In this particular situation, Glo-DynaCNN stands out for its dependability and accuracy, making it the ideal choice for expression classification of facial movements.

e. Mean Square Error

Figure 7 compares the MSE of four expression identification models, including FMF-LDTP, CBAM, AMSC, and the suggested Glo-DynaCNN for analysis. The six expressions of joy, sadness, anger, surprise, fear, and disgust are used to

Journal of Artificial Intelligence and Data Science Techniques
ISSN: 3029-2794
Volume 01 Issue 03 (September) 2024

measure the performance of each model. While the current models have much larger MSE values, which means they aren't very good at making predictions, Glo-DynaCNN has much lower MSE values, which means it's much better at identifying emotions. This visualization showcases how the proposed Glo-DynaCNN model outperforms previous methods, supporting its promise for improved emotion recognition in real-world scenarios.

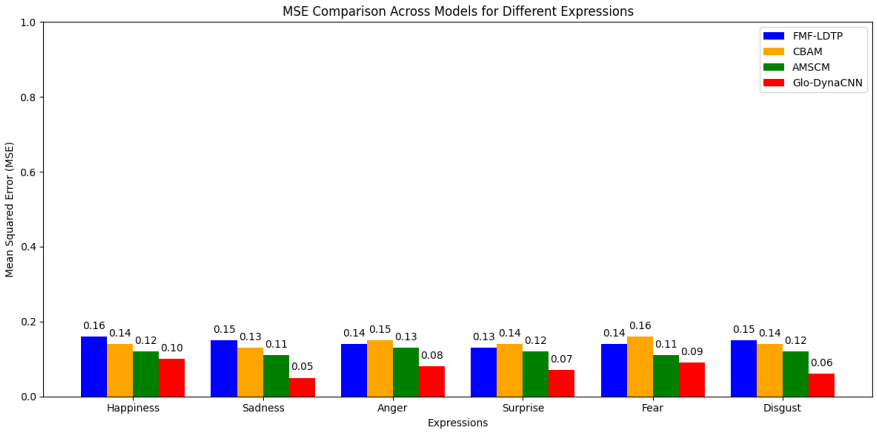

Figure 7: MSE Calculation

5. Conclusion

The suggested Glo-DynaCNN model for identifying facial expressions in video analysis performed better on the following metrics such as recognition accuracy, recall, precision, F1-score, and MSE. Extensive video clips from publicly available datasets allow the frames to be processed using data preparation in several frames. Improved recognition performance resulted from the applied dynamic CNN model's capacity to extract global and local information across different face parts efficiently. The extracted global and local features are combined using adaptive feature fusion. The recurrent nature emphasizes the temporal analysis of facial expressions over time. In the future, hybrid models with attention mechanisms will be applied, and deeper convolutional network designs will be integrated. Potential avenues for further investigation include optimizing processing in real-time and expanding the model's applicability to other expression domains, like mental health analysis and human-computer interaction.

6. References

[1] Liu, Y., Wang, W., Feng, C., Zhang, H., Chen, Z., & Zhan, Y. (2023). Expression snippet transformer for robust video-based facial expression recognition. *Pattern Recognition*, *138*, 109368.

[2] Zhang, W., Li, L., Ding, Y., Chen, W., Deng, Z., & Yu, X. (2023). Detecting facial action units from global-local fine-grained expressions. *IEEE Transactions on Circuits and Systems for Video Technology*, *34*(2), 983-994.

[3] Xu, R., Huang, A., Hu, Y., & Feng, X. (2023). GFFT: Global-local feature fusion transformers for facial expression recognition in the wild. *Image and Vision Computing*, *139*, 104824.

Journal of Artificial Intelligence and Data Science Techniques

ISSN: 3029-2794

Volume 01 Issue 03 (September) 2024

[4] Kalyta, O., Barmak, O., Radiuk, P., & Krak, I. (2023). Facial emotion recognition for photo and video surveillance based on machine learning and visual analytics. *Applied Sciences*, *13*(17), 9890.

[5] Yaddaden, Y. (2023). An efficient facial expression recognition system with appearance-based fused descriptors. *Intelligent Systems with Applications*, *17*, 200166.

[6] Mukhopadhyay, M., Dey, A., & Kahali, S. (2023). A deep-learning-based facial expression recognition method using textural features. *Neural Computing and Applications*, *35*(9), 6499-6514.

[7] Zhang, Z., Tian, X., Zhang, Y., Guo, K., & Xu, X. (2023). Enhanced discriminative global-local feature learning with priority for facial expression recognition. *Information Sciences*, *630*, 370-384.

[8] He, Z., Meng, B., Wang, L., Jeon, G., Liu, Z., & Yang, X. (2023). Global and local fusion ensemble network for facial expression recognition. *Multimedia Tools and Applications*, *82*(4), 5473-5494.

[9] Wei, J., Hu, G., Yang, X., Luu, A. T., & Dong, Y. (2024). Learning facial expression and body gesture visual information for video emotion recognition. *Expert Systems with Applications*, *237*, 121419.

[10] De Melo, W. C., Granger, E., & Lopez, M. B. (2024). Facial expression analysis using decomposed multiscale spatiotemporal networks. *Expert Systems with Applications*, *236*, 121276.

[11] Abbas, G. A Sequential Pattern Mining Method for the Individualized Detection of Online Banking Fraudulent Transactions.

[12] Karnati, M., Seal, A., Jaworek-Korjakowska, J., & Krejcar, O. (2023). Facial expression recognition in-the-wild using blended feature attention network. *IEEE Transactions on Instrumentation and Measurement*.

[13] Lee, J. R., Wang, L., & Wong, A. (2021). Emotionnet nano: An efficient deep convolutional neural network design for real-time facial expression recognition. *Frontiers in Artificial Intelligence*, *3*, 609673.

[14] Saeed, V. A. (2024). A framework for recognition of facial expression using HOG features. *International Journal of Mathematics, Statistics, and Computer Science*, *2*, 1-8.

[15] Yan, L., Shi, Y., Wei, M., & Wu, Y. (2023). Multi-feature fusing local directional ternary pattern for facial expressions signal recognition based on video communication system. *Alexandria Engineering Journal*, *63*, 307-320.

[16] Sheron, P. F., Sridhar, K. P., Baskar, S., & Shakeel, P. M. (2021). Projection-dependent input processing for 3D object recognition in human robot interaction systems. Image and Vision Computing, 106, 104089.

[17] Jiang, M., & Yin, S. (2023). Facial expression recognition based on convolutional block attention module and multi-feature fusion. *International journal of computational vision and robotics*, *13*(1), 21-37.

[18] Chen, Y., Li, J., Shan, S., Wang, M., & Hong, R. (2024). From static to dynamic: Adapting landmark-aware image models for facial expression recognition in videos. *IEEE Transactions on Affective Computing*.

[19] Li, T., Chan, K. L., & Tjahjadi, T. (2023). Multi-Scale correlation module for video-based facial expression recognition in the wild. *Pattern Recognition*, *142*, 109691.

[20] Singh, R., Saurav, S., Kumar, T., Saini, R., Vohra, A., & Singh, S. (2023). Facial expression recognition in videos using hybrid CNN & ConvLSTM. *International Journal of Information Technology*, *15*(4), 1819-1830.

[21]https://ibug.doc.ic.ac.uk/resources/afew-va-database/