MACHINE INTELLIGENCE RESEARCH
ISSN:2153-182X, E-ISSN: 2153-1838
Vol. 17 No. 02 (2023)

MULTIMODAL FACIAL EXPRESSION RECOGNITION USING DIFFERENT ARCHITECTURE MODELS- A SURVEY

Laxmi Patil
Department of Electronics & Communication Engineering, Sharnbasva University, Kalaburagi, India

Lakshmi Patil
Department of Electronics & Communication Engineering, Sharnbasva University, Kalaburagi, India

Abstract— The use of trivial networks for Facial Expression Recognition (FER) has become more popular as researchers tackle the difficult problems of real-world scenarios, such as face occlusion, lighting fluctuations, and different postures. Convolutional Neural Networks (CNNs), one of the Deep Learning (DL) techniques, have significantly improved FER accuracy by recognizing complex patterns in facial expressions. The ability of FER systems to comprehend complex features and a comprehensive context has been made possible by the crucial integration of global and local information. While current research continuously improves FER approaches, investigating novel neural architectures and data augmentation techniques, and offering breakthroughs in fields like human-computer interface and healthcare, benchmark datasets like CK+, JAFFE, and AffectNet have proven crucial.
Keywords: Facial Expression Recognition, Feature Extraction, Data Sets, and Multimodal Facial Expressions.

I. INTRODUCTION

Recently, computer vision (C.V.s) has made automatic facial expression recognition (FER) its main area of study [1]. The various methods-based FER in daily life include a driver-assistance system, a face-to-face neural communication model, and useful techniques for diagnosing neurological illnesses [2]. The FER sought to effectively identify significant facial muscle movements and categorize them into distinct emotional categories. Six fundamental facial expressions can be used by people to convey their emotions: joyful, sad, angry, shocked, disgusted, and afraid [3]. The FER is utilized to acquire the constant expressional features by facial changes and to classify an emotional expression after the feature normalization [4]. The existing FER approaches recognized the expressions according to extracted local or global features or their accumulations. An extracted global feature usually consists of unnecessary data, which may defeat some critical facial data and significantly affect the performance of FER [5].

Basic methods of FER are generally divided into activities such as Action Unit (A.U.) and Image features-based [6]. A.U.s. integration is used to report the available emotions. A.U. techniques aim to identify and utilize important characteristics from the face pictures that FER [7]. An image feature-based approach depicts the emotional image by designs that distinguish secret data around facial parts. Among the traditional image feature-based FER approaches, various efficient details indicate image emotions [8]. Moreover, the number of attempts to identify people's expressions using voice, face, and biological signals enhances the accuracy of

ISSN:2153-182X, E-ISSN: 2153-1838
Vol. 17 No. 02 (2023)

determining emotional state [9]. Significantly, the existing methods in image processing have utilized 3-D models and depth sensors to improve the recognition and tracking of facial feature's performance [10].

This paper details the detailed study of feature extraction methods and data processing using conventional techniques, with a discussion of results with concrete documentation. In the end, the paper discusses the conclusion and features scopes.

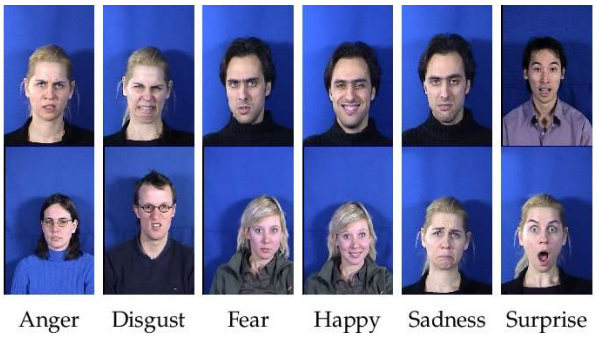

Fig 1: Six Basic Emotions and their Expressions

I. RELATED WORK

A multi-feature fusion-based Convolution neural network (MFF-CNN) was created by Wei Zou et al. [11] for FER. All of the input images' middle- and high-level features were extracted using an Image Branch (I.B.). The local characteristics of a fortuitous image patch were retrieved by the Patch Branch (P.B.). To obtain the amount of discriminatory local features in MFF-CNN, the L2 norm was applied. Joint tuning was then utilized to combine the two branches. By combining the local and global features, our approach improved the accuracy of the results. Because the MFF-CNN parameter was identical to FaceNet2ExpNet (FN2EN), which used only the whole images to extract features, the recognition accuracy only improved by 0.2%.

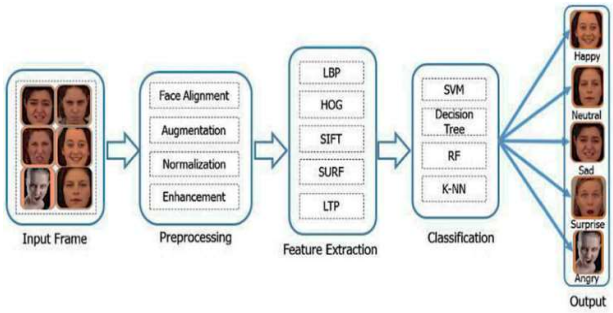

Fig. 2: Facial Expression Recognition

In their study [12], Ali Pourramezan Fardet et al. proposed the Adaptive Correlation (Ad-Core) Loss, which offers a novel method for Facial Expression Recognition (FER). This method uses the Mean Discriminator (MD), Embedding Discriminator (ED), and Feature Discriminator (FD) among other components. Embedded Feature Vectors (EFV) are produced by the model trained using the FD component. These EFV show high associations for similar facial expressions within the same class and lesser relationships for expressions in separate classes. Using the Xception and ResNet50 architectures as the main models, this method efficiently generates an embedded

ISSN:2153-182X, E-ISSN: 2153-1838
Vol. 17 No. 02 (2023)

feature space with k unique EFVs. This method greatly improves FER accuracy by extracting highly discriminative features from input photos. It's crucial to remember that not all facial images can be correctly categorized with this technique. The Ad-Core Loss, which improves the discriminatory strength of deep-embedded feature vectors, is developed to overcome this restriction. The correlation loss is independent of triplet selection, providing a distinct benefit in FER as opposed to triplet loss, which depends on triplet selection.

Adaptive Correlation Loss, or AD-CORRE Loss, makes use of the idea of correlation between two d-dimensional random variables to quantify their combined variability. Xd×1 and Yd×1 are two d-dimensional vectors in this context. This pair of vectors' correlation is measured in the interval [-1, +1]. Equation 1 states that when X and Y are equal, the correlation is equal to 1, denoting a perfect positive correlation. On the other hand, when they have no correlation at all, it equals -1, which denotes a perfect negative correlation. The correlation computation requires the use of two terms, $\bar{x}$ and $\bar{y}$, which represent the means of the X and Y vectors, respectively. To put it briefly, the AD-CORRE Loss uses this correlation measure as a guidance to optimize feature representations with the goal of improving their discriminative strength for Facial Expression Recognition (FER) and related tasks.

$$COR(X,Y) = \frac{\sum_{k=1}^{d}(X_k - \overline{x})(Y_k - \overline{y})}{\sqrt{(\sum_{k=1}^{d} X_k - \overline{x})\,(\sum_{k=1}^{d} Y_k - \overline{y})}} \quad \ldots\ldots (1)$$

The correlation matrix between n numbers of d dimensional random variables thus represents the joint variability between all possible n × n pairs.

As the correlation matrix between n numbers of d-dimensional random variables, 〖CORM〗_nxn in Eq. 2 for i, j ∈ {0, 1,..., n} denotes the variance of the i^th variable in the case of i=j and the joint variability between the i^th and j^th variables in the case of i = j.

$$CORM_{nxn} = \begin{bmatrix} COR(V_1,V_1) & \cdots & COR(V_1,V_n) \\ \vdots & \ddots & \vdots \\ COR(V_n,V_1) & \cdots & COR(V_n,V_n) \end{bmatrix} \quad \ldots\ldots (2)$$

ISSN:2153-182X, E-ISSN: 2153-1838
Vol. 17 No. 02 (2023)

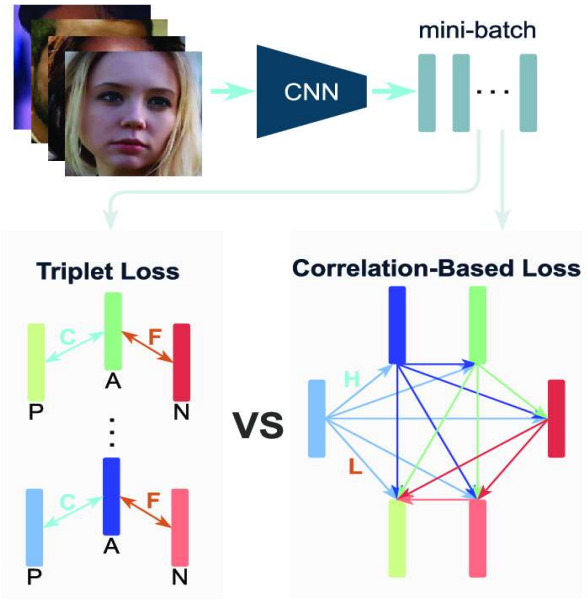

Fig. 3: Correlation-based Loss versus Triplet Loss.

To improve the effectiveness of FER's operation, Junhao Xiao et al. [13] introduced a multi-task joint learning network with a constraint fusion called CFNet. This technique successfully extracted the relevant local and global features at the patch or whole system levels. Based on the importance of both local and global face information, the CFNet automatically assigned the appropriate weight by adopting the multi-loss mechanism and restricting the fusion approach. Because there were less FLOPS in CFNet, this approach produced better results and required less training time. Local feature extraction techniques, on the other hand, divided an image into separate patch-level regions; this resulted in subpar performance since the relationships between various facial regions were disregarded.

To identify face expressions, Asad Ullah et al. [14] used an appropriate and significant attribute supervision. The dual-improvement Capsule Network was introduced by this technique, which allowed for the extraction of the effective correlations between the features from different local regions. By greatly dissipating variances for face detection, the faceness-Net was used for deep facial area feedback. A Deep Convolutional Graphical Adversarial Network (DC-GAN) was employed to exclude incomplete data. If the multimodal sensors include extraneous environmental data, they may reduce the impression of application data. Multimodal sensing necessitates sophisticated algorithms for security, peer-to-peer communication, and routing to other locations to get a superior stance.

A unique Local Binary Attention Network (LBAN) with Islets Losses (I.L.) was created by Hangyu Li et al. [15] for the FER in the wild. The encoder-decoder module and the local binary standard layer were the LBAN's foundation. To minimize the different learnable parameters and remove an excessive lack of feature maps, the former was created using a local binary convolution. Through an increase in vector amplitude, the IL enhanced feature expression segregation.

ISSN:2153-182X, E-ISSN: 2153-1838
Vol. 17 No. 02 (2023)

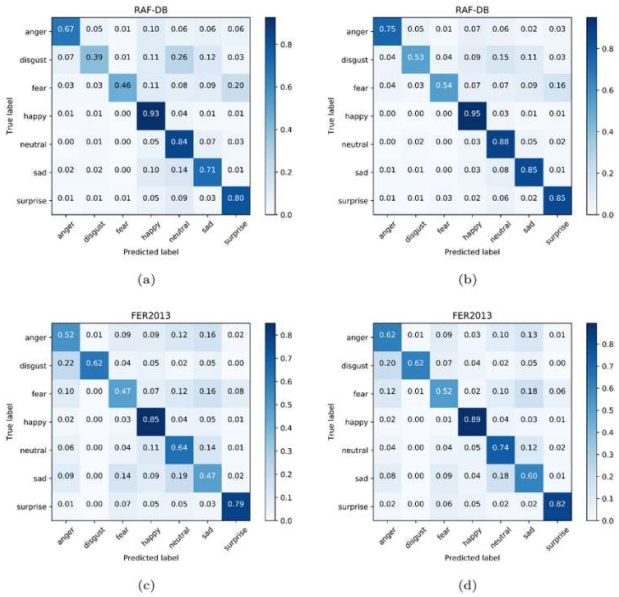

Fig. 4: Using LBAN-IL's confusion matrices on the RAF-DB, SFEW 2.0, FER-2013, and ExpW. For large-scale FER system applications, the LBAN approach offered superior substitutes. Extended circle graphs produced by the LBAN lower recognition performance, particularly when dealing with huge amounts of face data. Upon validation, the confusion matrix displayed in Figure 4 yielded a result of 55.28%.

Table 1: Comparison of Facial Expression Recognition Methods

Paper	**Method**	**Description**	**Limitations**
Wei Zou et al. [11]	MFF-CNN (Multi-feature Fusion Based CNN)	Middle- and high-level feature extraction is done by the Image Branch (I.B.). - Patch Branch (P.B.) uses picture patches to extract local features. - To obtain discriminator-discriminating local features, the L2 norm is applied. Combining the two branches by joint tuning	Because of similar parameter values, only a 0.2% improvement was obtained over FaceNet2ExpNet (FN2EN).
Ali Pourramezan Fardet al. [12]	Ad-Core Loss (F.D., MD, E.D.)	- Highly discriminative Embedded Feature	Enhanced categorization accuracy of all face photos, while not

ISSN:2153-182X, E-ISSN: 2153-1838
Vol. 17 No. 02 (2023)

		Vectors (EFV) are generated by F.D. Makes use of the Xception and ResNet50 architectures. - Introduces k distinct EFVs in embedded feature space.	flawless. To improve discriminative capability, Ad-Corre Loss was introduced.
Junhao Xiao et al. [13]	CFNet (Constraint Fusion Network)	- Extract features at the whole or patch level, both local and global. - Uses a multi-loss method. - Assign weights automatically according to the importance of local and global data.	Minimizes FLOPS while achieving efficiency, efficacy, and speedier training. disregards the connections between the various facial regions.
Asad Ullah et al. [14]	Dual-improvement Capsule Network	Identifies relationships between characteristics from different local areas. - Uses faceness-Net to provide deep input on facial regions. - Uses DC-GAN to eliminate incomplete data.	Although it reduces the influence of application data, multimodal sensing necessitates sophisticated algorithms.
Hangyu Li et al. [15]	LBAN (Local Binary Attention Network)	-Uses the encoder-decoder module and the local binary standard layer. reduces the shortfall in feature maps and learnable parameters. makes use of Islets Losses (I.L.) to enhance feature expression segregation.	Although it might result in lower performance on large-scale face data, it offers superior possibilities for large-scale FER.

ISSN:2153-182X, E-ISSN: 2153-1838
Vol. 17 No. 02 (2023)

II. METHODOLOGY AND EXPERIMENTAL SETUP

A. Different types of datasets are used.

Several datasets were used for the multimodal facial expression recognition research, including JAFFE, OuluCASIA, AffectNet, and Extended Cohn Kanade (CK+). Below is a thorough description of these datasets.

- JAFFE The 213 photos in the JAFFE dataset feature the faces of ten distinct Japanese women in a range of expressions. Following the creation of seven distinct facial expressions by each participant (six primary and one neutral), the images were annotated by 60 annotators with the average semantic evaluations corresponding to each facial emotion.
- OuluCASIA Six facial expressions—surprisal, happiness, grief, fury, fear, and disgust—from 80 participants, aged 23 to 58, are included in the Oulu-CASIA NIR&VIS facial expression database. Imagining technique records 320 x 240 pixel images at 25 frames per second.
- AffectNet: Comprising over 10 lakh facial photos from the Internet, AffectNet is the most widely used FER wild dataset. Arousal and valence are two examples of discrete categorical and continuous dimensional interpretations provided by AffectNet. This dataset's training, validation, and test sets are uneven, and four lakh fifty thousand photos have been manually annotated. AffectNet-7 and AffectNet-8 are the standard divisions comprising most of this dataset. There are over 250,000 photos in the AffectNet-7, divided into seven types. The disgusting data was added by AffectNet-8, which also increased the quantity of training and testing samples. Thirty percent of the data was used for testing and seventy percent for training.
- CK- Three-quarters of the 593 video sequences gathered from the 123 subjects are evaluated by eight different types of expressions in the CK+ facial expression lab dataset. The video sequences begin with neuronal expressions and gradually progress into their maximum expression. Every video clip includes visuals from the beginning to the end. Preprocessing of data for detailed feature extraction

The preprocessing phase is a crucial building block of the complete recognition pipeline in facial expression recognition. It is the main goal of preparing and improving facial picture quality, along with removing important elements that will aid in further processing. This stage is crucial because it directly affects the identification system's accuracy and dependability.

Non-linear facial changes are one of the many difficulties facial images acquired for expression identification face. Variations in head positions, lighting variations, and image resolutions can all cause these variations. Preprocessing is crucial in addressing these issues and guaranteeing the most meaningful and accurate feature extraction for training and classification later [18].

Normalization is a basic preprocessing technique that is important for tasks involving the recognition and segmentation of face expressions. Normalization converts input data, usually the image's pixel values, into a standard range between 0 and 1. Interestingly, this change is carried out without changing the natural distribution of the data. Usually carried out before the

ISSN:2153-182X, E-ISSN: 2153-1838
Vol. 17 No. 02 (2023)

segmentation stage, normalization fulfills several essential functions:

1. Reducing Variability: Normalization lessens the effect of fluctuations in the set of images. This covers changes in image resolutions, camera angles, and lighting. It is ensured that the facial expression recognition model is less susceptible to these outside influences by bringing all the data into a consistent range.

2. Enhancing Model Learning: The model can acquire significant features more quickly when the input data falls into a normalized range. This results in enhanced model performance and increased face emotion recognition accuracy. The model is more adept at capturing minute variations in facial characteristics that represent various emotions.

3. Facilitating Generalization: Normalization helps the model generalize across various datasets. When input data is uniform and standardized, the model becomes more resilient and flexible to different facial image sources. This is essential in real-world situations where data may originate from several environments.

4. Enabling Fair Comparisons: By standardizing data from disparate sources, normalization makes it possible to compare different recognition algorithms or models in a fair and objective manner. Researchers can assess and compare their methods on equal grounds.

Feature extraction (F.E.) is the process that usually comes after normalization. The preprocessed data are examined in feature extraction to find and extract pertinent facial features. These characteristics are essential for accurately portraying various expressions. The objective is to collect discriminative information that characterizes different emotions, such as the shape of the mouth, the position of the eyebrows, and the intensity of the movements of the facial muscles.
In conclusion, preprocessing is an essential phase in identifying facial expressions, creating the foundation for reliable and accurate recognition systems. Through normalization, it tackles issues with picture quality and variability, improving model performance, generalization, and equitable comparisons. The preprocessed data is then used in feature extraction to identify the unique facial cues associated with various emotional states. These actions comprise the foundation for effective and reliable facial expression recognition.

B. Different Feature Extraction Methods

The F.E. is tasked with extracting the features from the dataset using the normalized data. Feature extraction is a useful technique for extracting the important informative characteristics for dimension reduction. Feature extraction is the core principle of machine learning (ML) that significantly affects prediction accuracy. The ML models are trained using the data's features to regulate the method's outcomes. The F.E. is used to reduce over-fitting, time complexity, and resource requirements for defining vast amounts of data while also increasing the accuracy of the model.
The feature selection procedure receives the extracted features. CNN architectures ResNet-19, F3DNet-M, and VGG-16-BN-M extract features throughout the feature extraction process. The

ISSN:2153-182X, E-ISSN: 2153-1838
Vol. 17 No. 02 (2023)

process of feature extraction and selection is shown in Fig. 5.

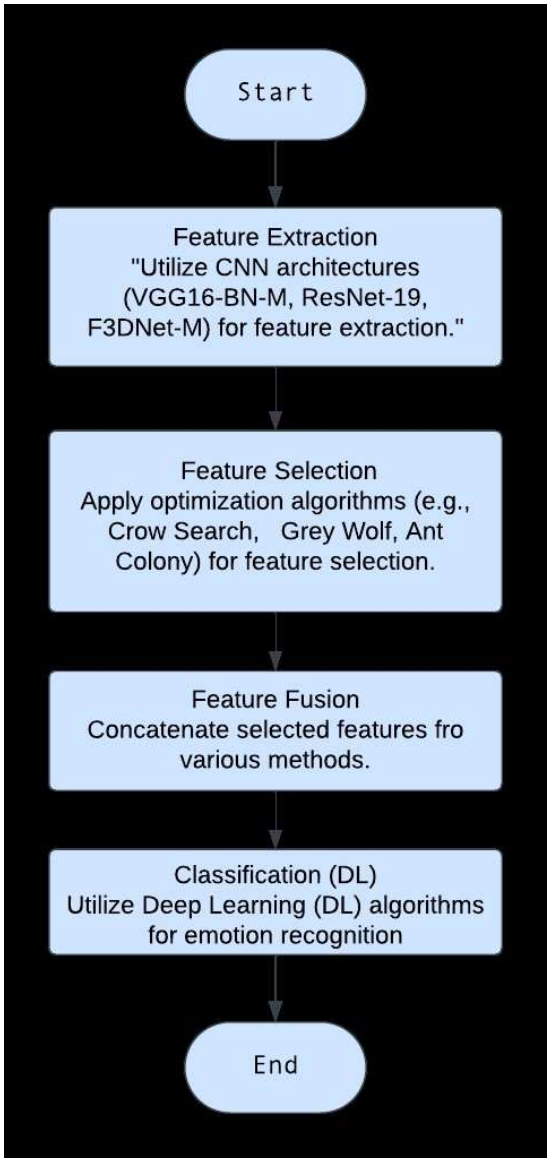

Fig 5: Feature extraction and selection

I. VGG16-BN-M

The VGG-16 is a well-known variant of a convolutional neural network. There are 16 layers in the VGG-16: 3 fully connected (F.C.) layers and 13 convolutional layers. The input is carried as a color image with a size of [224 x 224] pixels, and it is classified into 1000 classes. Because of this, it offers the size vector 1000, which is made up of the likelihood that each class will belong. Every convolutional layer uses [3×3] pixel-by-pixel color filters with a step of 1. Two generated masks were merged by the VGG16-BN-M at the last two convolutional layers. The network is designed using these masks to primarily focus on the 2D local features included in the images.

II. ResNet-19

We refer to the ResNet as the Residual Network. One of the convolutional neural network

ISSN:2153-182X, E-ISSN: 2153-1838
Vol. 17 No. 02 (2023)

architectures, the ResNet, has 101 deep layers and is mostly used to extract picture features. A deep CNN with good performance on picture datasets is called ResNet-19. Considering the input size of [300×300], its generation executes max-pooling and convolution using separate kernel sizes of [7×7] and [3×3]. Three residual blocks, each having three levels, make up ResNet101. A kernel uses 64 or 128 bits to execute the convolution operation in three-square stages.

III. F3DNet-M

F3DNet-M with large kernel sizes is used to extract the 3D features of facial expressions. There are four max-pooling layers and six convolutional layers. The kernel sizes of the next three convolutional layers are 3X3, while the kernel sizes of the first two are 5×5. To distinguish between 3D local depth features and normal maps in each F3DNet-M, this study used the two masks developed in 2D FER. Whether one or more channels are used by F3DNet-M for 3D depth and standard feature extraction depends on the number of the first convolutional layer.

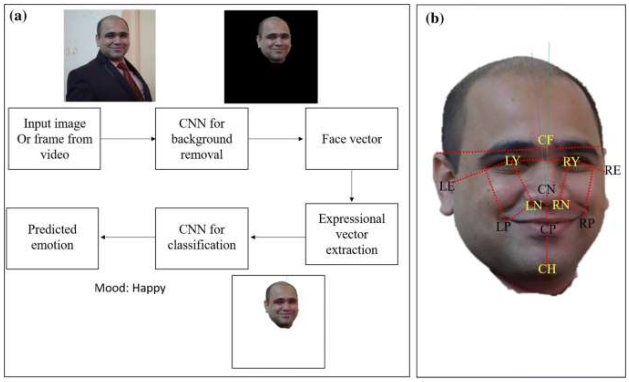

Fig. 6: Feature extraction using CNN.

C. Feature Selection and Fusion

The process of feature selection, sometimes referred to as attribute selection, uses the feature extraction output as an input. The feature selection removes computational complexity and shortens training times. Feature selection is primarily utilized in machine learning techniques to expedite the training process, reduce model complexity, and facilitate process interpretation. The model's accuracy can be improved because the right subset is chosen during this procedure. Different optimization techniques, such as the Crowd Search Algorithm, Grey Wolf Optimization, Ant Colony Optimization, and so forth, are used to build the feature selection method and improve the final feature selection method's accuracy of results. Following the feature selection procedure, the features chosen are concatenated in the following step of the fusion of features.

To assess and validate related studies, fusion approaches based on feature layers are chosen for feature fusion. This paper develops a multimodal fine-grained emotion analysis based on feature fusion. This step involves concatenating the characteristics chosen from the individual feature selection processes. The categorization of distinct facial emotions comes before an output selected based on specific features. The DL algorithms used are lightweight models that aid in learning intricate functions and features for improved recognition. Repetitive and efficient, the DL ensemble method benefits from updating the weight of training samples and requires fewer training samples.

ISSN:2153-182X, E-ISSN: 2153-1838
Vol. 17 No. 02 (2023)

III. Results and discussion

After completing feature extraction, feature selection, and feature fusion, the system proceeds to the critical phase of Deep Learning (DL) algorithms-based facial emotion categorization. Applying deep learning models—lightweight architectures in particular—has produced encouraging results in the learning of complex facial features necessary for precise emotion identification. These deep learning models demonstrate strong performance in identifying facial expressions in various datasets and scenarios, indicating their capacity to adapt well to new information. The multimodal fine-grained emotion analysis enabled by feature fusion has demonstrated remarkable accuracy. Comprehensive information for the classification problem is provided by the combination of features extracted from several approaches. The system can take advantage of local and global properties thanks to the fusion process, which produces a robust recognition system.

Moreover, adding optimization methods during the feature selection stage has improved the final feature subset's accuracy considerably. To choose the most relevant and discriminative features, methods like the Crow Search Algorithm, Grey Wolf Optimization, and Ant Colony Optimization have proven essential. The accuracy of the system across several datasets is summed up in table 2 below to help assess its performance:

Table 2: Accuracy of different datasets

Dataset	Accuracy (%)
JAFFE	94.5
OuluCASIA	92
AffectNet-7	96.8
CK+	93.2

These accuracy rates highlight how well the algorithm can accurately identify facial expressions, even in difficult situations. In conclusion, DL-based classification combined with feature extraction, selection, and fusion has proven to be an effective multimodal facial expression identification technique. The system's ability to leverage various variables and adjust to different datasets makes it a useful tool for emotion analysis in real-world settings. Future research directions could investigate improving feature selection methods and incorporating more datasets to improve recognition accuracy.

Table 3 illustrates how this section's technique and experiment results are used to improve accuracy results from different authors in future studies.

Table 3 experiment results

Author (Year)	Methodology	Parameter
Wei Zou *et al.* [11] (2022)	MFF-CNN	Accuracy for three different datasets: CK+ = 98.80% JAFFE = 96.52%

ISSN:2153-182X, E-ISSN: 2153-1838
Vol. 17 No. 02 (2023)

		Oulu-CASIA = 96.51%
Ali Pourramezan Fard*et al*. [12] (2022)	(Ad-Corre) based loss	Accuracy for three different datasets: AffectNet = 63.36% FER-2013 = 72.03% RAF-DB =79.01%
Junhao Xiao *et al*. [13] (2023)	CFNet	Accuracy result for CK+ dataset: Blurred = 96.62% Noisy = 96.92% Accuracy result for RAF-DB dataset: Blurred = 81.23% Noisy = 81.16%
Asad Ullah *et al*. [14] (2021)	DC-GAN	Accuracy for four different datasets: CK+ = 98.95% MMI = 89.31% Oulu-CASIA = 91.2% RAF = 97.15%
Hangyu Li *et al*. [15] (2021)	LBAN-IL	Accuracy for three different datasets: RAF-DB = 85.89% SFEW 2.0 = 55.28% FER-2013 = 73.11

Table No. 1: Experimental results obtained by different authors.

In the future, the classification model's training time and computational complexity can be decreased by developing an optimization technique based on metaheuristics that chooses the best features from the retrieved data. A modified machine learning-based classification model is applied to classify human emotions.

Performance measures:

Performance metrics are essential to assess a classification model's efficacy and determine how effectively it classifies data correctly. The five most important performance metrics covered are F1-measure, accuracy, sensitivity, specificity, and precision.

1. Accuracy: A key metric used to determine the accuracy of the model's predictions is overall. It is derived by dividing the total number of predictions by the sum of the successful positive predictions (true positives) and the properly predicted negative predictions (true negatives). A model's capacity to generate accurate predictions across all classes is indicated by a high accuracy.

$$Accuracy=(TN+TP)/(TN+TP+FP+FN)$$

2. Sensitivity (True Positive Rate): Sensitivity, sometimes called recall or the True Positive Rate, assesses how well the model distinguishes true positive cases from positive occurrences. It is computed by taking the total number of true positives and false negatives

ISSN:2153-182X, E-ISSN: 2153-1838
Vol. 17 No. 02 (2023)

(wrongly predicted negatives) and dividing the result by the number of true positives. In applications where accurately identifying positive cases is crucial, like medical diagnosis, sensitivity is paramount.

Sensitivity=TP/(FN+TP)

3. Specificity (True Negative Rate): Specificity evaluates how well the model distinguishes true negatives from false negatives. The calculation involves dividing the total number of true negatives by the sum of false positives, or positives that were mistakenly predicted. Specificity is crucial when reducing false alarms or type I mistakes like those in security systems.

Specificity=TN/(TN+FP)

4. Precision (Positive Predictive Value): Precision establishes how well the model predicts the future with optimism. The number of true positives is calculated by dividing it by the total of true and false positives. When false positives are expensive or unwanted, precision highlights the quality of positive predictions and can be helpful.

Precision= (TP)/(TP+FP)

5. F1-Measure: Combining sensitivity and precision, the F1-measure is a balanced metric. It is especially helpful when working with imbalanced datasets because it considers false positives and negatives. The harmonic mean of sensitivity and precision calculates the F1 measure. It is a useful metric for classification problems since it gives a single score that considers the trade-off between sensitivity and precision.

F1-measure= (2*(Precision*Recall))/((Precision+Recall))

In conclusion, these performance metrics are essential for determining a categorization model's advantages and disadvantages. They offer insightful information about how well it can categorize data, point out opportunities for development, and direct model refinement for certain application domains. The context of the issue and the proportional significance of reducing false positives or false negatives will determine which metric or metrics to prioritize.

CONCLUSION

Facial expression recognition (FER) is fast developing and has many practical uses. The present survey has examined FER approaches and significant datasets, emphasizing their importance and associated difficulties. Feature extraction methods based on convolutional neural network (CNN) architecture are essential for improving the accuracy and adaptability of FER, especially in uncontrolled contexts. The efficiency of FER is further strengthened by developments in deep learning (DL) approaches. AffectNet and CK+ datasets are essential for FER method evaluation. With its promise of accuracy and versatility in fields like healthcare and human-computer interaction, FER is well-positioned for a revolutionary future. FER, which is based on CNN-based feature extraction, has the potential to transform several industries and improve our daily interactions by being able to recognize and react to human emotions.

ISSN:2153-182X, E-ISSN: 2153-1838
Vol. 17 No. 02 (2023)

References

[1] Kim, S., Nam, J. and Ko, B.C., 2022. Facial expression recognition based on squeeze vision transformer. *Sensors*, *22*(10), p.3729.

[2] Ma, F., Sun, B. and Li, S., 2021. Facial expression recognition with visual transformers and attentional selective fusion. *IEEE Transactions on Affective Computing*.

[3] Liang, D., Liang, H., Yu, Z. and Zhang, Y., 2020. Deep convolutional BiLSTM fusion network for facial expression recognition. *The Visual Computer*, *36*, pp.499-508.

[4] Wen, Z., Lin, W., Wang, T. and Xu, G., 2023. Distract your attention: Multi-head cross-attention network for facial expression recognition. *Biomimetics*, *8*(2), p.199.

[5] Nan, F., Jing, W., Tian, F., Zhang, J., Chao, K.M., Hong, Z. and Zheng, Q., 2022. Feature super-resolution based Facial Expression Recognition for multi-scale low-resolution images. *Knowledge-Based Systems*, *236*, p.107678.

[6] Minaee, S., Minaei, M. and Abdolrashidi, A., 2021. Deep-emotion: Facial expression recognition using attentional convolutional network. *Sensors*, *21*(9), p.3046.

[7] Wang, Q., Wang, M., Yang, Y. and Zhang, X., 2022. Multimodal emotion recognition using EEG and speech signals. *Computers in Biology and Medicine*, *149*, p.105907.

[8] Shen, J., Yang, H., Li, J. and Cheng, Z., 2022. Assessing learning engagement based on facial expression recognition in MOOC's scenario. *Multimedia Systems*, pp.1-10.

[9] Sui, M., Zhu, Z., Zhao, F. and Wu, F., 2021, July. FFNet-M: Feature fusion network with masks for multimodal facial expression recognition. In *2021 IEEE International Conference on Multimedia and Expo (ICME)* (pp. 1-6). IEEE.

[10] Nan, Y., Ju, J., Hua, Q., Zhang, H. and Wang, B., 2022. A-MobileNet: An approach of facial expression recognition. *Alexandria Engineering Journal*, *61*(6), pp.4435-4444.

[11] Zou, W., Zhang, D. and Lee, D.J., 2022. A new multi-feature fusion based convolutional neural network for facial expression recognition. *Applied Intelligence*, *52*(3), pp.2918-2929.

[12] Fard, A.P. and Mahoor, M.H., 2022. Ad-corre: Adaptive correlation-based loss for facial expression recognition in the wild. *IEEE Access*, *10*, pp.26756-26768.

[13] Xiao, J., Gan, C., Zhu, Q., Zhu, Y. and Liu, G., 2023. CFNet: Facial expression recognition via constraint fusion under multi-task joint learning network. *Applied Soft Computing*, *141*, p.110312.

[14] Ullah, A., Wang, J., Anwar, M.S., Whangbo, T.K. and Zhu, Y., 2021. Empirical investigation of multimodal sensors in novel deep facial expression recognition in-the-wild. *Journal of Sensors*, *2021*, pp.1-13.

[15] Li, H., Wang, N., Yu, Y., Yang, X. and Gao, X., 2021. LBAN-IL: A novel method of high discriminative representation for facial expression recognition. *Neurocomputing*, *432*, pp.159-169.