An Efficient Multi-modal Facial Gesture-based Ensemble Classification and Reaction to Sound Framework for Large Video Sequences

SaiTeja Chopparapu
Department of EECE, GITAM (Deemed to be University), India
saiteja.chopparapu960@gmail.com (corresponding author)

Joseph Beatrice Seventline
HoD, Department of EECE, GITAM (Deemed to be University), India
sjoesph@gitam.edu

Received: 3 June 2023 | Revised: 11 June 2023 and 12 June 2023 | Accepted: 13 June 2023

 DOI: https://doi.org/10.48084/etasr.6087

ABSTRACT

Machine learning-based feature extraction and classification models play a vital role in evaluating and detecting patterns in multivariate facial expressions. Most conventional feature extraction and multi-modal pattern detection models are independent of filters for multi-class classification problems. In traditional multi-modal facial feature extraction models, it is difficult to detect the dependent correlated feature sets and use ensemble classification processes. This study used advanced feature filtering, feature extraction measures, and ensemble multi-class expression prediction to optimize the efficiency of feature classification. A filter-based multi-feature ranking-based voting framework was implemented on different multiple-based classifiers. Experimental results were evaluated on different multi-modal facial features for the automatic emotions listener using a speech synthesis library. The evaluation results showed that the proposed model had better feature classification, feature selection, prediction, and runtime than traditional approaches on heterogeneous facial databases.

Keywords-multi-modal facial features; feature ranking; multi-modal outlier component

I. INTRODUCTION

Automatic facial expression analysis in an image has two parts, extracting features and partitioning them into groups [1]. Feature extraction is the process of getting information from an image or a series of images, and is usually performed in two steps: turning the input image into a feature vector and then reducing its number of dimensions. The need for real-time systems poses even more strict limits on how well these tasks can be performed. In general, the methods for feature extraction can be partitioned into two groups: model-based and pixel-based. Model-based techniques, such as active appearance and active contour models, need some information and assumptions about the shape of facial features to model [2-4]. Model-based methods are not affected by changes in lighting or head pose but take a long time to fit the model as they use iterative steps. Pixel-based techniques usually work by mapping information about pixels to information about features without any modeling or iteration. Pixel-based methods use pixels instead of making wrong assumptions about the shape of facial features, as model-based methods do.

Pixel-based methods get information about the face and take less time to process but have the drawback of being sensitive to changes in lighting and head position. Facial expression analysis can be used to help with human-computer interaction, entertainment, medical applications (e.g. pain recognition), interactive videos, behavior monitoring, and figuring out when someone is lying. People's desire to interact with computers not only with their hands but also with their facial expressions is another important source of motivation. The progress in computer vision and social intelligence allows human-centered designs, instead of computer-centered ones, that only take into account the user's intentional input and ignore the majority of information communicated by affective states. Deformable Parts-Models (DPMs), also known as pictorial structure modeling, are one of the most common ways to make generic object detectors. Recently, DPMs have been used to find faces, and the results are state-of-the-art. Estimating the DPM parameters is performed in two ways: weakly and strongly supervised [5-7]. Generally, the SVM classification scheme is based on the characteristics of the statistical learning mechanism [8]. Minimizing structural risk is supported by the ability of the SVM classifier to perform the entire classification process smoothly and effectively. Additionally, the SVM technique is quite effective in the case of small training data.

II. LITERATURE REVIEW

In a weakly supervised setting, the bounding boxes of the face and non-face images are used as examples of what to do and what not to do. The mixtures and the location of the part are known ahead of time but are kept secret in the training set. In this case, only facial bounding-box annotations are needed, so thousands of annotations can be made in just a few hours. In a strongly supervised setting, parts of the mixture are labeled in the training database, which also has facial images with facial landmarks that have been labeled. This method is not practical as marking important landmarks is time-consuming and takes hundreds of hours for a few thousand facial images. Different arguments have been pointed out on how well weakly and strongly supervised annotations work. In [3], facial DPMs trained with strongly supervised annotations performed better than those trained with weakly supervised annotations. In [4], the weakly supervised DPM with fixed a priori components performed better than the current best method. Researchers are also concerned that DPMs are hard to compute, making it hard to use them in real-time systems. Recently, scientists have been trying to obtain real-time performance by simplifying computations. In [5], a branch-and-bound framework and a dual-tree data structure [6] were used to speed up star-shaped DPMs. Other approaches to speed up DPMs are DPM fitting by computing filter correlations using the Fourier transform [7], building a cascade of classifiers from DPMs [9], and using a fast neighborhood-aware cascade strategy [10]. The DPMs perform great in recognizing faces but cannot extract landmark points, which is a requirement for many facial expression analysis systems. An Active Appearance Model (AAM) and its variations [11] were used to get the best performance at the moment in landmark localization.

AAM variants work better because they use a fully connected matrix to model the shape, while DPMs are part-based texture models. Feature extraction turns an image into a feature vector so that classification can be done quickly. The need for real-time applications puts even more strict limits on how well this task can be performed. In [12], facial features, used to figure out how someone is feeling, were split into two main groups: primary and secondary features. Primary features are the most basic and important ones that can be used to identify the six basic emotions [13]. Secondary features are extra features that make emotion recognition more accurate and help distinguish features that are the same based on primary features. Some secondary features are the length of the mouth, the side wrinkles on the nose, the presence of teeth, the angle of the eyebrows, and the thickness of the lips. Geometric and appearance feature extraction methods are the two main types. Researchers have created hybrid features by putting these two features together. Face parts such as the nose, eyes, eyelids, and lips are used as landmarks to create geometric features.

Many different methods have been used to find landmarks by modeling facial features [14]. Active contours are geometric models whose initial coordinates are set by the model's parameters. Parametric shape models, on the other hand, are set by explicit shape equations. Active contour models get closer to the truth more quickly, but they depend a lot on the initial parameters. Active Shape Model (ASM) was suggested as a way to solve this problem, as it is a shape-restricted iterative modeling method that uses a statistical model, called Point Distribution Model (PDM), to get information about a shape that has already been used. A function that tries to save energy is used to fit the model to different images. The functions that try to minimize energy work either in the spatial [15] or the parametric domain [8]. In [16], a Gabor filter bank was used with 8 orientations and 5 spatial frequencies to get Gabor magnitudes from the whole face, and the AdaBoost method was then used to choose the subset of features. SVM used the results of the filters chosen by AdaBoost to sort seven facial expressions based on how they make people feel. Gabor features are used to extract features not only in the spatial but also in the temporal domain. In [17], Gabor features were used to analyze the behavior of both faces at the same time. In [18], the use of spatiotemporal Gabor filters was proposed to obtain information about space and time when recognizing subtle expressions. In [19], Gabor features were used around landmark points to continuously estimate the intensity of FACS data.

A Local Binary Pattern (LBP) was suggested for texture analysis in [20]. Due to its small neighborhood, the basic LBP cannot find large-scale structures that are the most important. To solve this problem, a circular neighborhood was chosen that was centered on the pixel that needed to be labeled. Several different versions of LBP have been suggested [21]. HOG features were first used to find people, and they are made by counting how many times a gradient orientation shows up in certain parts of an image. The idea behind HOG features is that the shape and look of facial features can be described by how intensity gradients are spread out. To find the HOG features, the image is broken up into small, connected areas called cells, and a histogram is made for each cell. Normalizing the histogram compensates for differences in light and dark. Feature classification is used in several systems to help with recognition, verification, and identification. A facial expression analysis system needs to be able to recognize both static and changing patterns. Dynamic patterns are the hardest to recognize because they change in space and time. This makes them the most difficult to classify. There are supervised and unsupervised methods to train facial expression analysis systems. In supervised learning, the event labels are already set in the training data, while in unsupervised learning, the training data are not labeled and event categories are figured out on their own [22]. There are two main types of facial expression recognition systems. In frame-based expression recognition, each video frame is examined on its own, while in sequence-based expression recognition, the movement between frames is examined, which depends on other frames in the video. In [23], a nearest neighbor classifier was used for the parameters that were calculated by fitting local parametric motion models to different parts of the face.

Facial expression recognition is another area where neural networks have been used [24]. In [25], Fischer linear discriminant analysis was used to recognize expressions in the encrypted domain. In [26], a fuzzy relational approach was proposed to recognize human emotions from facial expressions, using three different fuzzy sets: high, low, and moderate. In [27], a genetic algorithm was used to adjust the

parameters of the membership function and make fuzzy inference systems work better. In [28], after the system was able to recognize faces at the frame level, facial features were modeled over time by adding the movement of facial expressions. A common way to divide expressions by time is to use HMMs to find a match between the start, peak, and end of an action and an underlying latent state. In [29], HMM was used to recognize natural emotions. In [28], a DBN was used to model two types of temporal dependencies: how each AU changes over time and how different AUs change over time.

In [31], a method was proposed and evaluated on several benchmark datasets, achieving high accuracy compared to previous methods, but had some potential drawbacks as the multi-region approach can increase computational cost and complexity. In [32], the proposed method leveraged the self-attention mechanism of the transformers to learn feature representations of facial images and improve the accuracy of facial expression recognition. The widely adopted VGG-Face architecture was used for feature extraction, introducing a spatial and channel-wise attention mechanism to weigh feature maps. In [33], the first branch used a CNN with a local attention mechanism to extract local features, while the second used a CNN with a global attention mechanism to capture contextual global information. In [34], a method was proposed that extracted high-order statistical features, including the histogram of oriented gradients and local binary patterns, from the input facial images.

III. PROPOSED FRAMEWORK

This study designed a novel facial gesture-based sign detection method using multilevel facial features and an annotated sign database. This framework used sign text annotations, text audio library, and multi-model features to find and map the facial gestures to their corresponding sign annotation. Each gesture and its pair of signs were fused for the training data. This model used different multi-modal feature ranking measures to classify the multi-modal features along with the sign annotations. Finally, a hybrid ensemble heterogeneous multi-modal sign classification framework was implemented in the training data, using novel base classifiers such as linear feature extraction measures, non-linear feature extraction measures, and hybrid multidirectional local and global gradient descriptors, as shown in Figure 1.

The proposed framework included a novel metaheuristic-based feature ranking and classification approach on a real-time facial dataset. Initially, facial data were taken as input for the feature extraction, ranking, and classification process using a 1-sec interval video sequence. The video sequences with facial features were stored in a real-time cloud server for data analysis. The statistical feature anomaly detection approach was used to filter outliers in the input data. In the next step, the ensemble feature ranking approach was used to find the essential multi-modal features for the segmentation process. Finally, an advanced ensemble learning model was used to optimize the prediction rate in the segmented classes.

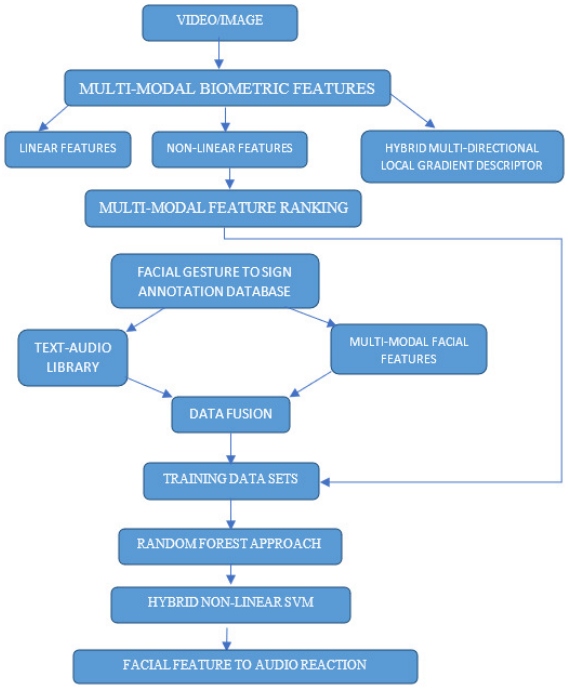

Fig. 1. Proposed multi-modal facial sign annotation and classification framework.

The proposed model was implemented in 2 phases:

- Phase 1: Multi-modal feature extraction measures and filtering.
- Phase 2: Proposed feature segmentation-based ensemble classification.

A. *Phase 1: Feature Extraction Measures and Filtering*

In this phase, different statistical measures such as Log Cosine and Exponential Inverse Differential Moment (LCEIDM), Max-Chi-Correlation based Inertia, facial curvatures, standard deviation, kurtosis, skewness, and RMS were used to find the essential multi-modal features in each frame of the training dataset. Filter-based feature ranking used the ant bee colony approach.

- Step 1: The missing values filling algorithm was:

```
for each feature F in dataset D
  for each instance value I(j) in the
F(i) do
    if (instance is null) then
        Ins(j) = max{|Ins(i) − μ_F(Ins(j))|)/
(Max_F(Ins(j)) − Min_F(Ins(j)))}
    endif
  endfor
endfor
```

- Step 2: Initialization of ant parameters.

$$\phi_i = max\{\mu.\eta_j^r(1-\eta_j^r), \frac{1}{\sqrt{2\pi\sigma_D}} e^{-\frac{(X-\mu_D)^2}{\sigma_X{}^2}}\}$$

$$\eta_j^k \in (0,1)$$

- Step 3: Employing bee phase: In this phase, new candidate solutions are generated for each employed bee. Initially, each value of the employed bee was initialized to the new candidate solution $c(i)=x(i)$.
- Step 4: The fitness value of the candidate solutions can be computed using:

$fit_i = \frac{1}{1+Kernel_i}$; if $Kernel_i \geq 0$

fit_i=1+abs($Kernel_i$) ; otherwise

B. Phase 2: The Proposed Cluster-based Ensemble Classification Framework

This phase implemented a hybrid signal clustering approach to find the classes on the training data. Here, cluster labels are used to predict the variation of the multi-modal gesture segmentation process.

1) Feature Segmentation Framework

Two steps were implemented using the likelihood probability estimation functions. In the first step, referred to as the expectation step, model variables are estimated and the probability of each instance is predicted for the cluster assignment.

Let $\hat{\theta}$ represent the novel posterior estimation parameter used to predict the occurrence of each feature in the given large number of training samples.

$Prob(\hat{\theta}/c)$ is the occurrence of new patterns in the category of classes. In the expectation phase, the maximization of the probability estimation is performed on the training data as:

$$Prob(\hat{\theta}/c_m) = \frac{max\{Corr(F_i, C_m), P(F_i . C_m)\}.\sum_{i=1}^{N} P(c_m/F_i)}{|N|.P(c_m/F)}$$

In the maximization phase, model parameters are estimated using the signal features and their class patterns. In the maximization step, the probability of data occurrence in the given class is given as:

$$Prob(c/F_i, \hat{\theta}) = \frac{P(F_i/c_i).max\{P(c_m); m=0..n\}.log(P(\hat{\theta}))}{P(c/D_i)}$$

These two phases are repeated until the number of maximum iterations or no change in the error rate.

2) Proposed Ensemble Classification Learning Model

A hybrid feature ranking-based decision tree model was constructed using the following equation:

$$PFRDT(F, C_m) = \sqrt[3]{\left(\sum_{j=1}^{|f|}(\sqrt[3]{F_i/|F_i|} - \sqrt[3]{\frac{F_j}{chi(F_j}})^2\right) \times max\{GainRatio(F_i, C_m), ConditGainRatio(F_j, C_m)\}}$$

where F_i is the i-th cluster class and F_j is the j-th cluster class.

3) Mutual Information-based Feature Ranking

The ensemble feature ranking measure based on mutual information was given as:

$$MIChisquare(D_i, D_j) = Max\{Prob(D_i) \times log(D_i/D), \sum\sum(|D_i - D_o| - \frac{0.5)^2}{D_o}\}$$

$$Prestimation = -Pestimation(D_i).log(Pestimatino(D_i))$$

$$EntCondest(D) = \sum_i^{\sum} Prestimation$$

The hybrid Hellinger feature ranking measure is given as:

Math.cbrt(*entropy*(*data*)×*total*×*GHDSplitCriterion.computeHellinger*(*data*)) ×*Pr*/ (*chiVal*(*data*))

4) Non-linear SVM Kernel Function

An advanced non-linear classification approach was used to find the essential multi-modal classification process. The proposed classification approach was used to detect the level of facial patterns based on the selected features. Different filters were used in the framework to find the essential key features for the classification process. This approach was composed of the following steps:

- Step 1: Feature Space *FS*= *F*(*Filter*(*Images*))
- Step 2: For each feature map in the layer, compute the feature correlation among the different features space as:

$$e_k = \frac{1}{2} correlation\left(F_0(r)\right).\sum_a^m (MI_0(r) - O_o(r))^2$$

$$O_o(r) = \sum_k^L MI_{ho}(r)\psi_{h0}$$

$$N_{ho}(r) = x\left(N_{ho}(r)\right).g(N_{ih}(r)$$

$$N_{ih}(r) = \sum^m w_{ih}x_i(r) + a_h$$

- Step 3: Apply the non-linear probabilistic optimization function as:

$$Min\{O_1\} = \omega.\sum_{i=1}^{k}\sum_{j=1}^{|RD|} e^{-\log|P_{ij}|}.D(g_i, Tr[j])$$

where D is the Euclidean distance.

$$min\, J = \frac{1}{2} \parallel sign(w^* \varphi(x)_i + k) \parallel^2 . + Max\{O_1\} = \omega.\sum_{i=1}^{k}\sum_{j=1}^{|RD|} e^{-\log|P_{ij}|}.\log|P_{ij}| + \varepsilon.v_t\sum_{i=1}^{n}(\gamma_i + \gamma_i)$$

$$P_{ij} = \frac{Prob(\frac{g_i}{Tr[j]}.e^{-(\frac{Dg_{i,Tr[j]}}{\min(|G_t|,|RD|)})}}{\sum_{m=1}^{k} e^{-(\frac{Dg_{i,Tr[j]}}{m*T_{rc}})}}$$

where $g_i \epsilon G_t, Tr[i] \epsilon RD$, and k is the number of segments.

$$T_{rc} = \sum_{i=1}^{|Tr|} g_i.Tr[i] / \sum_{j=1}^{|Tr|} g_j$$

$$v_t = \frac{1}{2}\ln(\frac{1-\gamma_t}{\gamma_t})$$

$$\varepsilon = arg_{tel,M1} \min v_t$$

$$w.\varphi(x) + k - y_i - v_t \leq \varepsilon + \gamma_i^*$$

$$(y_i - w.\varphi(x) - k + v_t \leq \varepsilon + \gamma_i$$

$$\gamma_i, v_t, \gamma_i \geq 0, i = 1,2,..n$$

$$B_f = Unique\ CV(D); //Unique\ column\ values$$

$$HB_f = Histobins[\,] = histogrambin(D)$$

$$GaussianKernel: GK(\phi, \theta) = e^{-\theta^2}/(2 * \phi^2)$$

$$\psi = gkv = GK\left(\sum HB_f, \sum B_f\right)$$

$$Kernel\ Probability = KP(D) = |HB_f/(\sum \psi * HB_f)|$$

$$GaussianEntropy: GE(d_i) = -GK(\sum_i d_i . \log(d_i), \mu_d)$$

In the above equations, the probabilistic Gaussian estimation function was used to find the ranked features on each multi-modal facial feature space.

$$PE = e^{-D^2}/(2 * D_1^2) * |HB_{D_1}/(\sum \psi * HB_{D_1})| + e^{-D^2}/(2 * D_2^2) * HB_{D_2}/(\sum \psi * HB_{D_2})$$

if (*PE*>0) then:

$$S = Classify\left((D_i, D_j)\right)$$

else continue

Let *I* be the input image and *S* be the input hyperplane with inter and intra-block regions intensities as *M* and *N*. Then the objective function of the proposed model is given as:

$$\max\{[\sum_{i-1}^{N}(\eta . \min\{wgt_2 I_n\} + p_i S_i^T S_i)]^{-1}\ [\sum_{i-1}^{N} q_i S_j^T S_j]R\}$$

$$\eta . \min\{wgt, I_n\} \geq 0$$

$$q_i > 0$$

Proof: In the following formulation, *N*(*M*, *N*) is used to find the multi-class objects in the given classification problem. The non-linear equation of the proposed model is solved using the following formulation, where *wgt*, *S*, *M*, and *N* are the input parameters, and the derivation of the proposed non-linear kernel equation is as follows:

$$N(M,N) = \sum_{j=}^{N} \| \ \eta S_j - \min\{wgt, I_n\} S_j M . N^T \|_F^2 + \eta . wgt \| N \|_F^2$$

where *S* are the segmented regions, *I* is the block intensity, *M* and *N* are the block size, η is the scaling factor, and *wgt* is the weight of the block.

$$= \sum_{j=1}^{N} Tr[\left(\eta S_j - \min\{wgt, I_n\} S_j M . N^T\right)^T . \left(\eta S_j - \min\{wgt, I_n\} S_j M . N^T\right) + \eta . Tr\{(wgt \ \| N \|)^T . (wgt \| N \|)\}$$

$$= \sum_{j=1}^{N} Tr[\left(\eta S_j^T - N . M^T . \min\{wgt, I_n\} S_j^T\right) . \left(\left(\eta S_j - \min\{wgt, I_n\} S_j M . N^T\right) + \eta . Tr\{(wgt \ \| N \|)^T . (wgt \| N \|)\}$$

$$= \sum_{j=1}^{N} Tr\{\eta S_j^T . \eta S_j - N . M^T . \min\{wgt, I_n\} S_j^T . \eta S_j - \eta S_j^T . \min\{wgt, I_n\} S_j M . N^T + N . M^T . \min\{wgt, I_n\} S_j^T . \min\{wgt, I_n\} S_j M . N^T\} + \eta . Tr\{(wgt \ \| N \|)^T . (wgt \| N \|)\}$$

$$= \sum_{j=1}^{N} Tr\{\eta S_j^T . \eta S_j) - Tr(N . M^T . \min\{wgt, I_n\} S_j^T . \eta S_j) - Tr\left(\eta S_j^T . \min\{wgt, I_n\} S_j M . N^T\right) + Tr(N . M^T . \min\{wgt, I_n\} S_j^T . \min\{wgt, I_n\} S_j M . N^T)\} + \eta . Tr\{(wgt \ \| N \|)^T . (wgt \| N \|)\}$$

$$= \eta^2 \sum_{j=1}^{N} Tr\left(S_j^T . S_j\right) - \eta \min\{wgt, I_n\} \sum_{j=1}^{N} Tr\left(N . M^T . S_j^T . S_j\right) - \sum_{j=1}^{N} \eta . \mathrm{mi}$$
$$\sum_{j=1}^{N} \min\{\mathrm{wgt}, \mathrm{I_n}\}^2 . \mathrm{Tr}(\mathrm{N} . \mathrm{M^T} . S_j^T . S_j M . N^T)\} + \eta . wgt^2 Tr(\| N \|)^T . (\| N \|)\}$$

$$Z > \eta^2 \sum_{j=1}^{N}\left(S_j\right) - \eta \min\{wgt, I_n\} \sum_{j=1}^{N}\left(N . M^T . S_j\right) - \eta . \min\{wgt, I_n\} . \sum_{j=1}^{N}\left(S_j M . N^T\right) + \min wgt, I_n\}^2 . \sum_{j=1}^{N} N . S_j M)\} + \eta . wgt^2 (\| N \|)\} + \delta$$

Differentiating *wgt*, *N* we get:

$$Z^* = -\eta \min\{wgt, I_n\} \sum_{j=1}^{N}\left(M^T . S_j\right) - \eta . \min\{wgt, I_n\} . \sum_{j=1}^{N}\left(S_j M . N^T\right) + \min\{wgt, I_n\}^2 . \sum_{j=1}^{N}\left(N . S_j M\right) + \eta\ wgt^2\ (\| N \|)\} + \delta$$

This Z^* estimates the feature class classification process.

IV. EXPERIMENTAL RESULTS

The proposed framework was implemented in Python with third-party audio annotation libraries. In the initial phase, all the facial features were extracted using multi-modal feature extraction and ranking measures. The extraction process used the Apache Math and Weka libraries to improve the feature ranking process. These collected features were annotated using the training sign annotations. In this study, OpenCV was used to implement a novel ensemble multi-modal facial feature classification process. Figures 2-5, represent the test data predictions for the neutral, happy, angry, and surprise classes using the different multimodal feature classes. In each figure, the respective class has a higher probability estimation compared to the other multi-modal feature classes using the ensemble classification process.

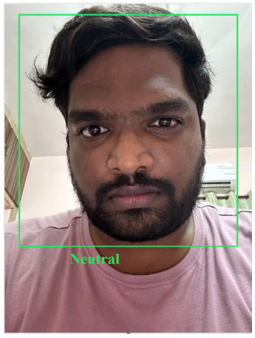

Fig. 2. Experimental result of neutral facial class on the multi-modal different classes.

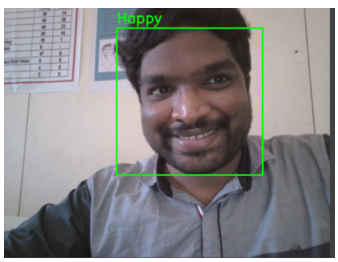

Fig. 3. Experimental result of the happy facial class.

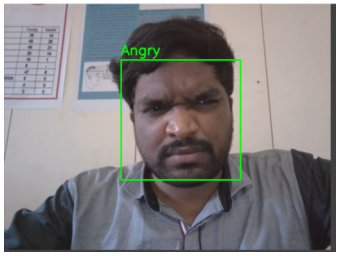

Fig. 4. Experimental result of the angry facial class.

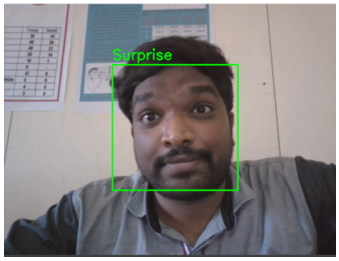

Fig. 5. Experimental result of the surprise facial class.

Table I displays the total count of features extracted using the proposed feature extraction measures for each multi-modal feature in the input data. This table provides information about the number of features extracted from different modalities or types of data. The multi-modal features undergo a filtering process to remove noisy features in each frame, ensuring that only relevant and meaningful features are retained for further analysis. Noisy features, which may introduce errors or irrelevant variations, are eliminated to improve the quality and usefulness of the extracted features. The resulting filtered multi-modal features are then used for subsequent processing or analysis tasks.

TABLE I. PERFORMANCE ANALYSIS OF THE PROPOSED MULTI-MODAL FILTERED FEATURES COUNT AND COMPARISON TO TRADITIONAL MODELS

Multi Modal FER	MI+CNN	Chisquare +CNN	Correlation +CNN	KernelFS +CNN	Proposed FEM
#100	82	83	72	55	32
#200	86	79	64	64	34
#300	81	75	63	65	33
#400	89	86	72	59	30
#500	82	90	66	60	38
#600	88	92	71	63	31
#700	86	79	66	58	33
#800	82	92	68	66	30
#900	79	86	71	56	32
#1000	89	76	74	63	36

Figure 6 represents the total feature count of the proposed feature extraction measures on each multi-modal feature on the input data. The results show that the proposed multi-modal features are filtered after removing the noisy features in each frame. Table II presents the test classification accuracy of the proposed multimodal facial features using an ensemble learning framework, presenting the accuracy scores obtained for different facial expression evaluations. The proposed filtered-based classification approach outperformed traditional models in terms of accuracy, indicating that the filtering process, which removed noisy features from the multimodal facial data, led to improved classification results. The ensemble learning framework combined with the filtered multimodal features contributed to higher accuracy in evaluating facial expressions. The results highlight the effectiveness of the proposed approach in achieving superior performance compared to traditional models. Figure 7 shows the accuracy of the proposed multimodal facial feature ensemble learning framework. The results show that the proposed filtered-based classification had better accuracy than traditional models on different facial expression evaluations.

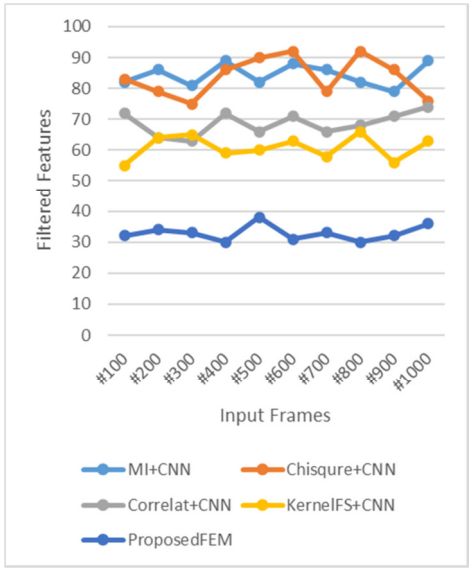

Fig. 6. Performance analysis of the multi-modal filtered feature count of the proposed and traditional models.

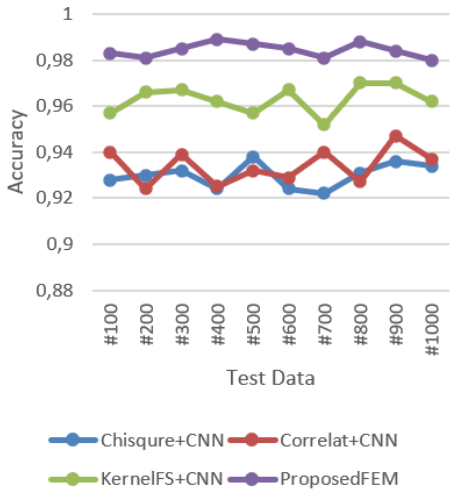

Fig. 7. Performance analysis of the proposed multi-modal class recall filter-based ensemble learning model with traditional models.

Table III and Figure 8 illustrate the test classification precision of the proposed multimodal facial feature model using an ensemble learning framework. The table shows the

precision scores achieved for various facial expression evaluations. The proposed model showed higher precision compared to traditional models. This suggests that the filtering process, which removes noisy features from multimodal facial data, improved precision in classifying facial expressions. The combination of the ensemble learning framework and the filtered multimodal features improved precision. These results emphasize the superiority of the proposed approach over the traditional models in accurately evaluating facial expressions.

TABLE II. PERFORMANCE COMPARISON OF THE PROPOSED MULTI-MODAL FILTER-BASED ENSEMBLE LEARNING MODEL WITH TRADITIONAL MODELS

Multi ModalFER	Chisqure+ CNN	Correlat+ CNN	KernelFS+ CNN	Proposed FEM
#100	0.928	0.94	0.957	0.983
#200	0.93	0.924	0.966	0.981
#300	0.932	0.939	0.967	0.985
#400	0.924	0.925	0.962	0.989
#500	0.938	0.932	0.957	0.987
#600	0.924	0.929	0.967	0.985
#700	0.922	0.94	0.952	0.981
#800	0.931	0.927	0.97	0.988
#900	0.936	0.947	0.97	0.984
#1000	0.934	0.937	0.962	0.98

TABLE III. PERFORMANCE ANALYSIS OF PROPOSED MULTI-MODAL CLASS PRECISION FILTER-BASED ENSEMBLE LEARNING AND TRADITIONAL MODELS

Multi Modal FER	Chisqure+ CNN	Correlat+ CNN	KernelFS+ CNN	Proposed FEM
#100	0.92	0.937	0.955	0.983
#200	0.94	0.941	0.96	0.988
#300	0.926	0.925	0.964	0.985
#400	0.938	0.947	0.969	0.988
#500	0.937	0.923	0.963	0.981
#600	0.928	0.948	0.957	0.983
#700	0.938	0.927	0.957	0.984
#800	0.93	0.945	0.958	0.984
#900	0.923	0.933	0.962	0.985
#1000	0.924	0.944	0.958	0.98

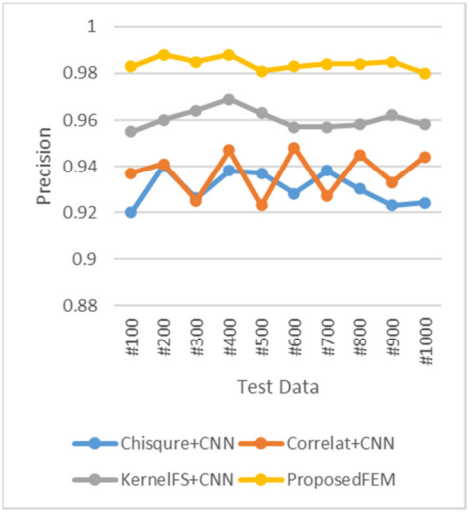

Fig. 8. Performance comparison of the proposed multi-modal class filter-based ensemble learning model with traditional models.

Overall, the proposed approach offers superior precision in facial expression evaluation, showcasing its potential for more accurate and reliable results compared to traditional models.

V. CONCLUSION

In this study, a hybrid approach combining multiple feature extraction and filtering techniques was designed and implemented for the ensemble classification of different multi-modal classes. Traditional models often suffer from high false positive rates and mean error rates when applied to various multi-modal facial classes. To address this issue, a filter-based multi-feature ranking system was integrated into a voting framework, which was implemented on different classifiers designed for multi-modal data. The experimental results were evaluated using various multi-modal facial features within an automatic emotion listener system utilizing a speech synthesis library. The proposed multi-modal facial gesture-based ensemble learner outperformed the traditional approaches on heterogeneous facial databases in terms of feature ranking, feature selection, prediction accuracy, and runtime performance. This paper showcased the effectiveness of the designed approach in improving the classification and prediction capabilities for multi-modal facial data, leading to improved performance compared to conventional methods.

VI. DECLARATIONS

The facial images showing the output results were of the corresponding author, who has no objections to their use in this research paper.

The authors declare that there is no conflict of interest associated with this study and they did not receive any financial support or funding that could influence the findings or conclusions of this study.

Data sharing does not apply to this article as no datasets were generated or analyzed during this study.

REFERENCES

[1] B. Zou, Y. Wang, X. Zhang, X. Lyu, and H. Ma, "Concordance between facial micro-expressions and physiological signals under emotion elicitation," *Pattern Recognition Letters*, vol. 164, pp. 200–209, Dec. 2022, https://doi.org/10.1016/j.patrec.2022.11.001.

[2] Y. Zhu, T. Peng, S. Su, and C. Li, "Neighbor-consistent multi-modal canonical correlations for feature fusion," *Infrared Physics & Technology*, vol. 123, Jun. 2022, Art. no. 104057, https://doi.org/10.1016/j.infrared.2022.104057.

[3] Y. Zhang, Y. Chen, and C. Gao, "Deep unsupervised multi-modal fusion network for detecting driver distraction," *Neurocomputing*, vol. 421, pp. 26–38, Jan. 2021, https://doi.org/10.1016/j.neucom.2020.09.023.

[4] L. Zhang and X. Wu, "Multi-task framework based on feature separation and reconstruction for cross-modal retrieval," *Pattern Recognition*, vol. 122, Feb. 2022, Art. no. 108217, https://doi.org/10.1016/j.patcog.2021.108217.

[5] J. Zhang, L. Xing, Z. Tan, H. Wang, and K. Wang, "Multi-head attention fusion networks for multi-modal speech emotion recognition," *Computers & Industrial Engineering*, vol. 168, Jun. 2022, Art. no. 108078, https://doi.org/10.1016/j.cie.2022.108078.

[6] D. Zeng, S. Zhao, J. Zhang, H. Liu, and K. Li, "Expression-tailored talking face generation with adaptive cross-modal weighting," *Neurocomputing*, vol. 511, pp. 117–130, Oct. 2022, https://doi.org/10.1016/j.neucom.2022.09.025.

[7] W. Yu and H. Xu, "Co-attentive multi-task convolutional neural network for facial expression recognition," *Pattern Recognition*, vol. 123, Mar. 2022, Art. no. 108401, https://doi.org/10.1016/j.patcog.2021.108401.

[8] S. Wang, Z. Wu, G. He, S. Wang, H. Sun, and F. Fan, "Semi-supervised classification-aware cross-modal deep adversarial data augmentation," *Future Generation Computer Systems*, vol. 125, pp. 194–205, Dec. 2021, https://doi.org/10.1016/j.future.2021.05.029.

[9] J. Yu, Y. Feng, R. Li, and Y. Gao, "Part-facial relational and modality-style attention networks for heterogeneous face recognition," *Neurocomputing*, vol. 494, pp. 1–12, Jul. 2022, https://doi.org/10.1016/j.neucom.2022.04.093.

[10] Y. Yaddaden, "An efficient facial expression recognition system with appearance-based fused descriptors," *Intelligent Systems with Applications*, vol. 17, Feb. 2023, Art. no. 200166, https://doi.org/10.1016/j.iswa.2022.200166.

[11] Z. Xing and Y. He, "Multi-modal information analysis for fault diagnosis with time-series data from power transformer," *International Journal of Electrical Power & Energy Systems*, vol. 144, Jan. 2023, Art. no. 108567, https://doi.org/10.1016/j.ijepes.2022.108567.

[12] W. Xiaohua, P. Muzi, P. Lijuan, H. Min, J. Chunhua, and R. Fuji, "Two-level attention with two-stage multi-task learning for facial emotion recognition," *Journal of Visual Communication and Image Representation*, vol. 62, pp. 217–225, Jul. 2019, https://doi.org/10.1016/j.jvcir.2019.05.009.

[13] A. B. S. Salamh and H. I. Akyüz, "A Novel Feature Extraction Descriptor for Face Recognition," *Engineering, Technology & Applied Science Research*, vol. 12, no. 1, pp. 8033–8038, Feb. 2022, https://doi.org/10.48084/etasr.4624.

[14] A. Alsheikhy, Y. Said, and M. Barr, "Logo Recognition with the Use of Deep Convolutional Neural Networks," *Engineering, Technology & Applied Science Research*, vol. 10, no. 5, pp. 6191–6194, Oct. 2020, https://doi.org/10.48084/etasr.3734.

[15] H. Wen, S. You, and Y. Fu, "Cross-modal dynamic convolution for multi-modal emotion recognition," *Journal of Visual Communication and Image Representation*, vol. 78, Jul. 2021, Art. no. 103178, https://doi.org/10.1016/j.jvcir.2021.103178.

[16] Q. Wang, M. Wang, Y. Yang, and X. Zhang, "Multi-modal emotion recognition using EEG and speech signals," *Computers in Biology and Medicine*, vol. 149, Oct. 2022, Art. no. 105907, https://doi.org/10.1016/j.compbiomed.2022.105907.

[17] M. Wang, Z. Huang, Y. Li, L. Dong, and H. Pan, "Maximum weight multi-modal information fusion algorithm of electroencephalographs and face images for emotion recognition," *Computers & Electrical Engineering*, vol. 94, Sep. 2021, Art. no. 107319, https://doi.org/10.1016/j.compeleceng.2021.107319.

[18] L. C. O. Tiong, S. T. Kim, and Y. M. Ro, "Multimodal facial biometrics recognition: Dual-stream convolutional neural networks with multi-feature fusion layers," *Image and Vision Computing*, vol. 102, Oct. 2020, Art. no. 103977, https://doi.org/10.1016/j.imavis.2020.103977.

[19] Y. Tian, S. Sun, Z. Qi, Y. Liu, and Z. Wang, "Non-tumorous facial pigmentation classification based on multi-view convolutional neural network with attention mechanism," *Neurocomputing*, vol. 483, pp. 370–385, Apr. 2022, https://doi.org/10.1016/j.neucom.2022.01.011.

[20] C. Suman, S. Saha, A. Gupta, S. K. Pandey, and P. Bhattacharyya, "A multi-modal personality prediction system," *Knowledge-Based Systems*, vol. 236, Jan. 2022, Art. no. 107715, https://doi.org/10.1016/j.knosys.2021.107715.

[21] Z. Shen, A. Elibol, and N. Y. Chong, "Multi-modal feature fusion for better understanding of human personality traits in social human–robot interaction," *Robotics and Autonomous Systems*, vol. 146, Dec. 2021, Art. no. 103874, https://doi.org/10.1016/j.robot.2021.103874.

[22] Y. Said, M. Barr, and H. E. Ahmed, "Design of a Face Recognition System based on Convolutional Neural Network (CNN)," *Engineering, Technology & Applied Science Research*, vol. 10, no. 3, pp. 5608–5612, Jun. 2020, https://doi.org/10.48084/etasr.3490.

[23] S. Saxena, S. Tripathi, and T. S. B. Sudarshan, "An intelligent facial expression recognition system with emotion intensity classification," *Cognitive Systems Research*, vol. 74, pp. 39–52, Aug. 2022, https://doi.org/10.1016/j.cogsys.2022.04.001.

[24] N. Sankaran, D. D. Mohan, N. N. Lakshminarayana, S. Setlur, and V. Govindaraju, "Domain adaptive representation learning for facial action unit recognition," *Pattern Recognition*, vol. 102, Jun. 2020, Art. no. 107127, https://doi.org/10.1016/j.patcog.2019.107127.

[25] E. S. Salama, R. A. El-Khoribi, M. E. Shoman, and M. A. Wahby Shalaby, "A 3D-convolutional neural network framework with ensemble learning techniques for multi-modal emotion recognition," *Egyptian Informatics Journal*, vol. 22, no. 2, pp. 167–176, Jul. 2021, https://doi.org/10.1016/j.eij.2020.07.005.

[26] U. Saeed, "Facial micro-expressions as a soft biometric for person recognition," *Pattern Recognition Letters*, vol. 143, pp. 95–103, Mar. 2021, https://doi.org/10.1016/j.patrec.2020.12.021.

[27] M. Ren, W. Nie, A. Liu, and Y. Su, "Multi-modal Correlated Network for emotion recognition in speech," *Visual Informatics*, vol. 3, no. 3, pp. 150–155, Sep. 2019, https://doi.org/10.1016/j.visinf.2019.10.003.

[28] N. Rathour, R. Singh, A. Gehlot, S. Vaseem Akram, A. Kumar Thakur, and A. Kumar, "The decadal perspective of facial emotion processing and Recognition: A survey," *Displays*, vol. 75, Dec. 2022, Art. no. 102330, https://doi.org/10.1016/j.displa.2022.102330.

[29] D. G. Nair, J. J. Nair, K. Jaideep Reddy, and C. V. Aswartha Narayana, "A privacy preserving diagnostic collaboration framework for facial paralysis using federated learning," *Engineering Applications of Artificial Intelligence*, vol. 116, Nov. 2022, Art. no. 105476, https://doi.org/10.1016/j.engappai.2022.105476.

[30] R. K. Mishra, S. Urolagin, J. A. Arul Jothi, and P. Gaur, "Deep hybrid learning for facial expression binary classifications and predictions," *Image and Vision Computing*, vol. 128, Dec. 2022, Art. no. 104573, https://doi.org/10.1016/j.imavis.2022.104573.

[31] C. SaiTeja and J. B. Seventline, "A hybrid learning framework for multi-modal facial prediction and recognition using improvised non-linear SVM classifier," *AIP Advances*, vol. 13, no. 2, Feb. 2023, Art. no. 025316, https://doi.org/10.1063/5.0136623.

[32] J. Liao, Y. Lin, T. Ma, S. He, X. Liu, and G. He, "Facial Expression Recognition Methods in the Wild Based on Fusion Feature of Attention Mechanism and LBP," *Sensors*, vol. 23, no. 9, Jan. 2023, Art. no. 4204, https://doi.org/10.3390/s23094204.

[33] J. Zhong, T. Chen, and L. Yi, "Face expression recognition based on NGO-BILSTM model," *Frontiers in Neurorobotics*, vol. 17, 2023, https://doi.org/10.3389/fnbot.2023.1155038.

[34] D. Mamieva, A. B. Abdusalomov, M. Mukhiddinov, and T. K. Whangbo, "Improved Face Detection Method via Learning Small Faces on Hard Images Based on a Deep Learning Approach," *Sensors*, vol. 23, no. 1, Jan. 2023, Art. no. 502, https://doi.org/10.3390/s23010502.