

RESEARCH ARTICLE

Facial Expression Recognition Using Visible, IR, and MSX Images by Early and Late Fusion of Deep Learning Models

MUHAMMAD TAHIR NASEEM¹, (Member, IEEE), CHAN-SU LEE², (Member, IEEE), AND NA-HYUN KIM², (Student Member, IEEE)

¹Research Institute of Human Ecology, Yeungnam University, Gyeongsan-si 38541, Republic of Korea

²Department of Electronic Engineering, Yeungnam University, Gyeongsan-si 38541, Republic of Korea

Corresponding author: Chan-Su Lee (chansu@ynu.ac.kr)


This work was supported in part by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education under Grant 2021R1A6A1A03040177; and in part by the Ministry of Trade, Industry and Energy (MOTIE), South Korea, through the Project “Development of convergence lighting technology based on affective perception for autonomous vehicles.”

ABSTRACT Facial expression recognition (FER) is one of the best non-intrusive methods for understanding and tracking mood and mental states. In this study, we propose early and late fusion methods to recognize five facial expressions (angry, happy, neutral, sad, and surprised) using different combinations from a publicly available database (VIRI) with visible, infrared, and multispectral dynamic imaging (MSX) images and the (NVIE) database. A distinctive feature is the use of concatenation and combining techniques to combine ResNet-18 with transfer learning (TL) to create a model that is significantly more accurate than individual models. In the early fusion, we concatenated features from the modalities and classified facial expressions (FEs). In the late fusion, we combined the outputs of the modalities using weighted sums. For this purpose, we used different weighting factors depending on the accuracy of the individual models. The experimental results demonstrated that the proposed model outperformed the previous works by providing an accuracy of 83.33% when we trained the model (1-step training). Through further fine-tuning (3-step training), we obtained an improved performance of 84.44%. We conducted additional experiments by combining them with another modality (MSX) available in the database. By performing experiments with an additional modality (MSX), we obtained improved performance, which confirms that the additional modality combined with existing modalities can help improve the performance of fusion models for facial expression recognition. We also experimented by changing the backbones (Vgg-16, ShuffleNetv2, MobileNetv2, and GhostNet) in addition to ResNet-18 for visible and MSX data. ResNet-18 outperformed the other backbones in facial expression recognition for visible and MSX data.

INDEX TERMS Deep learning, early fusion, facial expressions, infrared, late fusion, MSX, transfer learning, visible, ResNet-18.

I. INTRODUCTION

Humans can only communicate approximately 55% of the information verbally, and they convey the remaining 45% nonverbally. In emotional communication, 55% of signals are delivered through facial expressions, 38% through a

The associate editor coordinating the review of this manuscript and approving it for publication was Bing Li .

paralanguage-like tone of voice, and only 7% through verbal language [1]. Therefore, it is important to understand facial expressions under different lighting conditions to interpret human behavior and monitor mood and mental states. Facial expression recognition (FER) is also attracting attention in industries including criminology, hologram, smart healthcare, security, and entertainment, as well as industries like robotics and entertainment and stress detection. FER is

regarded by the research community as an essential area of study owing to its variety of applications.

Recently, researchers have shown an interest in creating FER systems using machine learning (ML) and deep learning (DL). Because visible light cameras are readily available, both as standalone units and as attachments for inexpensive portable devices such as phones and tablets, they are typically employed to capture images that are needed to classify facial expressions. Although there are many studies on identifying facial expressions in photos captured by cameras that use regular visible light [2], [3], [4], [5], [6], daily occurrences such as shadows, reflections, and darkness (or low light) make it difficult to classify expressions. In particular, under low-light conditions, it is difficult to extract facial expressions from conventional visible-light cameras. Consequently, the temperature distribution in facial muscles sometimes offers improved facial expression classification when working with thermal picture aids.

Thermal imaging has several benefits that make it useful in a variety of real-world applications. The use of thermal imaging is essential for facilitating human-robot interactions [7]. A new perspective on the use of infrared imaging, which can be used to comprehend physiological signals and make them significant in social interactions, was presented in [8]. This method aids the identification of cognitive processes using unconscious facial expressions. By identifying facial expressions, the researchers in [9] demonstrated how infrared imaging can be used to measure the degree of arousal. In addition, it aids robots in recognizing human emotions when interacting with them.

In this study, we proposed early and late fusion methods using different FER combinations. This study uses transfer learning (TL) and DL with the modified architecture of the ResNet-18 model for modalities in a publicly available database (VIRI) to classify five expressions (angry, happy, neutral, sad, and surprised) and the (NVIE) database. For early fusion, we concatenated the visible and infrared features and trained the model using 1-step training. The model was further fine-tuned (3-step training). In 1-step training, we concatenate the features by loading the weights of the corresponding modalities from ResNet-18 and then train a fully connected layer for the final classification. In 3-step training, we concatenated the features after further training each corresponding modality using ResNet-18. For late fusion, we combined the corresponding modalities using a weighted summation. We used different weight factors chosen according to the accuracy of the already trained single models (visible, infrared, and MSX), which meant that the higher the accuracy, the higher the weight factor.

We performed additional experiments using the multi-spectral dynamic imaging (MSX) modality, which was not available at baseline. When we combined the additional modality (MSX) with the existing modalities, we achieved improved performance using early and late fusion methods. This clearly shows that the additional modality is very useful, especially when combined with existing modalities, and

can assist other researchers in improving the performance of their systems. We also evaluated other backbones (Vgg-16, ShuffleNetv2, MobileNetv2, and GhostNet) to combine visible with the MSX images.

The contributions of our proposed model are listed as follows:

- We proposed a FER system using early fusion by training the model in 1 step and 3 steps. We show that 3-step training can improve performance over 1-step training.
- We propose a FER system that uses late fusion by combining modalities using a weighted summation. We show that weight factors, which depend on the accuracy of single models, can improve the performance in late fusion.
- We performed additional experiments to combine another modality (MSX) with existing modalities. Our experimental results show that an additional modality combined with existing modalities can improve the FER performance. In addition, the experiment results show that the combination of all three modalities in facial expression does not show the best performance but the combination of visible and MSX image shows the best performance.

II. RELATED WORKS

We review FER systems according to the database used. The first part focuses on works for the visible dataset, the second part on the infrared dataset, and the third part focuses on combined (visible and infrared) datasets.

A. FER USING VISIBLE DATASET

Deep convolutional neural networks (CNNs) that can automatically and reliably understand the semantic data present in faces without manually creating feature descriptors were presented in [10] by applying different loss functions and training tricks. Another real-time system for emotion recognition based on labeled facial images, which was tested on publicly available datasets, was discussed in [11].

A unique real-time vectorized facial feature-based model using deep CNNs is also discussed in [12], which provided a classification accuracy of 84.33%. To increase the accuracy, a CNN with four convolutional layers and two hidden layers was suggested for expression recognition using the CK+ dataset [13].

A video-based emotion recognition CNN that can be expressed over a sequence of frames was presented with an accuracy of 97.56% [14]. Detecting the occurrence of facial expressions (AUs) as a subpart of FACS, which represents human emotions, is also important for this task [15]. The CK+ dataset was used for the evaluation and exhibited an accuracy of 92.81%.

Instead of focusing on the emotional content (such as anger) of facial expressions, quantitative descriptions of low-level image elements provide a better way to describe and understand the automatic reactions to these emotions [16].

With regard to video-based facial expression recognition (VFER), three-dimensional (3D-CNNs) and long short-term memory (LSTM) regularly exceed other methods. A CNN, which is a blend of a 3D-CNN and long short-term memory (LSTM) for capturing expressions from videos, has been presented by researchers to capture spatiotemporal information from video sequences of emotions [17]. A DL method based on an attentional convolutional network that can focus on key facial features and significantly outperforms earlier models on a variety of datasets, including FER-2013, CK+, FER2011, and JAFFE, was presented in [18]; it is able to find important facial regions to detect different emotions based on the classifier's output.

The three-dimensional CNN (3D-CNN) and LSTM have consistently outperformed many approaches in video-based facial expression recognition (VFER). Regarding this, the author in [17] presented a blend of 3D-CNN and ConvLSTM for VFER. The hybrid architecture captures spatiotemporal information from the video sequences of emotions. An automated framework for facial detection using a CNN is discussed with four convolution layers and two hidden layers with an improved accuracy [13].

Predictive models that can classify emotions in the context of active teaching, specifically a robotics workshop, which is more challenging were proposed in [19]. The two models, Inception-v3 and ResNet-34 were combined for real-time emotion prediction.

B. FER USING INFRARED DATASET

Infrared or thermal images were employed to make the facial portion visible under very low-light conditions. Thermal images only expose human skin [20]. Additionally, the heat distribution in the facial muscles was observed using thermal imaging. This feature makes it possible to classify FEs more accurately and clearly because it is not reliant on outside variables, such as poor lighting or human viewing with the naked eye. Due to the limited number of studies conducted in this field, there is more potential for research in this area. The researchers in [21] recognized the FEs based on their geometrical properties.

A video-based FER system was employed in [22] in which sequential features were first extracted using horizontal and vertical temperature differences, and the most pertinent features for the classification goal were then identified. An FER system was built by another researcher [23] based on IR measurements of changes in facial skin temperature. A non-contact methodology to classify human emotions through thermal images of the face was presented using histogram features obtained from facial patches fed to a support vector machine (SVM) [24]. Using constrained local models (CLMs), which iteratively fit a dense model to an unseen image, precise 3D shape information of the human face can be calculated [25]. Most conventional methods rely heavily on feature extraction and classification techniques with significant preprocessing.

CNN is a DL technique that can automatically identify and learn significant features from the raw data of images through numerous layers [26]. Another study [27] proposed a model for emotion identification using ResNet152 to predict facial expressions in the NVIE dataset. The concept of TL allows features learned from high-resolution images of enormous datasets to be applied to train a model of a relatively small dataset without losing the ability to generalize [28]. For FER from IR photos, a DL model was proposed in [29], which uses the residual and transformation units, two building components, to extract key aspects from the input photos. Focusing on important areas of a face rather than on a complete face is adequate to reduce processing while simultaneously increasing accuracy [30]. Keeping this in mind, researchers proposed a model in which the entire face is divided into four parts.

C. FER USING FUSION OF VISIBLE AND INFRARED DATASETS

Recently, visible and infrared image fusion has been employed to boost recognition performance by combining images or results. Using this technique, the most important information can be extracted from different modalities and combined for classification. Researchers utilized array equipment with two infrared and two visible lenses to take infrared and visible images concurrently. Thus, the fused image not only had the texture information of the visible image but also the contrast information of the infrared image [31].

Researchers [32] also combined visible and infrared images for FER in three phases. First, they used the active appearance model (AAM) and then defined three head motion features from the visible spectrum. Several thermal statistical features were extracted from the IR images. Second, the F-test was used for feature selection. Third, support vector machines (SVMs) and Bayesian network SVMs are suggested for feature- and decision-level fusion. The usefulness of the suggested techniques was demonstrated by experiments on a natural visible and infrared facial expression (NVIE) spontaneous database [33], which also demonstrated the additional value of thermal infrared images for recognizing visible FEs.

A multimodal facial emotion database including both natural spontaneous visible and thermal infrared videos was presented in [34]. The dataset also included data on emotional strength, with each feeling categorized into three degrees (low, medium, and high). They recorded seven spontaneous emotions in 30 participants. Modern ML models, such as CNN, ResNet50, YOLO, and a combination of multiple models, were used to validate the created database. The obtained findings are realistic and demonstrate the practical value of using this dataset. Another simultaneous dataset containing visible and infrared images was presented in [35], which combined the features from the visible and infrared images using TL and Alexnet. However, the approach for fusing visible and IR images is limited, and further investigation is required.

III. DATABASES USED AND OUR PROPOSED MODEL

This section describes the databases used in our study and the proposed models based on early and late fusion.

A. DATABASES USED

In this section, we describe the databases used in this study. First, we used the VIRI database, which was created and presented by the researchers in [35], containing 566 image pairs for visible infrared, and MSX, which were captured with an uncontrolled wild background.

The database contained five expressions (angry, happy, neutral, sad, and surprised), and 110 subjects participated in the experiments. The resolution of each image in the database is 500×500 pixels. A few images selected from the visible, infrared, and MSX datasets are shown in Fig. 1. It is clear from Fig. 1 that facial features such as the inner brow are very difficult to notice in a few infrared images than in visible and MSX images (a blend of visible and infrared images). Second, we used the NVIE spontaneous database contains both visible and infrared images of more than 100 subjects [32]. The database contains six basic facial expressions (happy, sad, surprise, fear, anger, and disgust). A few samples from the NVIE database are shown in the Fig. 2.

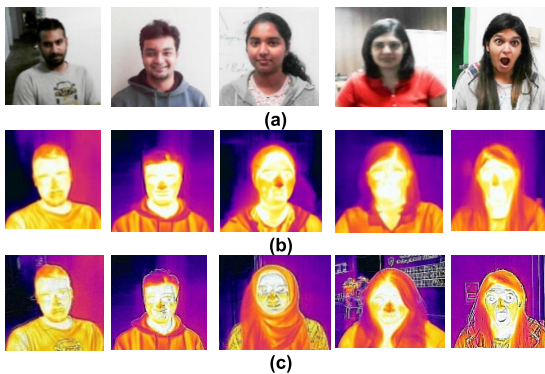


FIGURE 1. A few samples from the VIRI database left to right pairwise (angry, happy, neutral, sad, surprise): (a) visible (b) infrared (c) MSX.

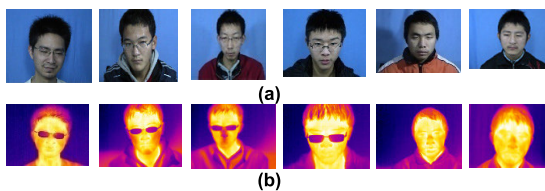


FIGURE 2. A few samples from the NVIE database left to right pairwise (happy, disgust, fear, surprise, anger, sad): (a) visible (b) infrared.

B. PROPOSED MODEL

In this section, we describe the proposed model used in this study. Here, we first propose single models, visible, infrared, and MSX for the classification of five expressions (angry, happy, neutral, sad, and surprised) using the modified architecture of the ResNet-18 model and three expressions

(happy, fear, and disgust) from the NVIE database. The motivation behind using the ResNet-18 is a kind of deep neural network using skip connections or short-cuts that jump over some layers which shows fast convergence.

Second, we propose early fusion methods that concatenate the features from (visible and infrared), (visible and MSX), and (visible, infrared, and MSX) using 1-step and 3-step training. It is likely that features from a single modality are insufficient for appropriately categorizing an image. To increase the classification accuracy under these circumstances, the concatenation function of the CNN can be utilized to integrate features from various modalities. In 1-step training, we combined the features from the corresponding modalities and then trained the model by loading weights from ResNet-18. In 3-step training, we loaded weights from our trained single models (visible, IR, and MSX). When we changed the training strategy from 1-step to 3-step, we obtained an improved performance.

Third, we proposed a late fusion method that combines the outputs from (visible and infrared), (visible and MSX), and (visible, infrared, and MSX) using a weighted summation. Different weight factors were used to combine the different modalities. The weights in the late fusion architecture were selected from validation dataset which were separated datasets from training datasets. In the last, we also evaluated the performance of other backbones (Vgg-16, ShuffleNetv2, MobileNetv2, and GhostNet) using the same architecture of the best performance of the RenNet-16.

TL is the practice of using a DL network as a pre-trained model to perform a new task by tweaking some of its layers. This method is useful when a small amount of data is available for training because it can train the final network with even a small number of datasets using a pre-trained model. This procedure is typically quicker and more effective than networks formed from scratch for the same quantity of training data. The following steps comprise a typical CNN image classification TL process.

- Pick a pre-trained network.
- To adjust to the new dataset, replace the final layers.
- Increase the accuracy by adjusting the values of the hyper-parameters and training again.

A fully connected (fc) layer, a max pooling layer with a 3×3 filter size, and 17 convolutional layers collectively form a ResNet-18 model. The conventional ResNet-18 model has 33.16 million parameters and uses batch normalization (BN) and ReLU activation functions to cover the complete back of the convolutional layers in the basic block [36]. ResNet-18 can classify images into 1000 categories after training on more than a million images and is known as the ImageNet repository [37].

In the publicly available VIRI (visible, infrared, and MSX) database, 111 images for angry, 114 images for happy, 114 images for neutral, 114 images for neutral, and 113 images for surprise are available. The total number of images for each dataset in the VIRI database was divided into three parts: training (second column), validation, and test,

with 70%, 15%, and 15% split ratios, respectively, as listed in Table 1.

TABLE 1. Data distribution before and after augmentations for the VIRI database.

Classes	training images before augmentation	training images after augmentation	validation	test
Angry	77	2002	16	18
Happy	79	2054	17	18
Neutral	79	2054	17	18
Sad	79	2054	17	18
Surprise	79	2054	16	18
Total	393	10218	83	90

The number of images required to attain the appropriate level of accuracy was insufficient for training the CNN for expression recognition from the VIRI database. So, various image augmentations such as rotations, zooming, distortion, shear, and flipping were performed on each training image to multiply the images and thereby increasing the size of the datasets. For example, for angry expressions, the train images before augmentations were 77. When we applied rotations 5 times on each image, the total number of images we got was $(77 \times 5 = 385)$. Similarly, we again applied the other augmentations like zooming, distortion, shearing, flipping, etc. 5 times, 5 times, 5 times, and 6 times respectively and we got 385, 385, 385, and 482 images. In total, we got $(385 + 385 + 385 + 385 + 462 = 2002)$ images for the angry expression. In the same way, we increased the total number of images for other expressions. The training images before and after the augmentation for each expression are listed in Table 1 (second and third columns).

A similar procedure was applied to the visible and infrared datasets from the NVIE database to increase the size of the training set. During the training of the modified architecture of the ResNet-18 model, the validation set was used to tune the hyperparameters of the network, whereas the test set was used to evaluate the model only.

Fig. 3 shows the proposed single model for visible images using the CNN architecture ResNet-18. The last layer of ResNet-18 was modified according to the visible dataset, and thus, two dropouts, two FC layers, and one SoftMax layer were added (which applied the SoftMax function).

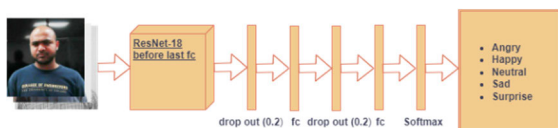


FIGURE 3. Proposed single model for visible images.

As shown in Fig. 1 (a), although the features were distinguishable by the naked eye, there was still slight overfitting. This is because of the small number of visible images in the available dataset. Augmentation was performed to expand

the dataset. To mitigate overfitting, parameter tuning was performed, which involved adjusting values while monitoring the loss and accuracy on both the training and validation sets; thus, the last layer of the ResNet-18 model was altered, as shown in Fig. 3.

Fig. 4 shows the proposed single-modal infrared images using the CNN architecture ResNet-18. The last layer of ResNet-18 was modified according to the infrared dataset, and thus, three dropouts, two activation functions, a batch normalization layer, three FC layers, and a SoftMax layer were added. Compared with the visible images, the features faded in numerous infrared images, as shown in Fig. 1 (b). Augmentation was performed to increase the dataset size. To minimize overfitting, we checked the loss and accuracy of both the training and validation sets during parameter training and altered the last layer of ResNet-18, as shown in Fig. 4.

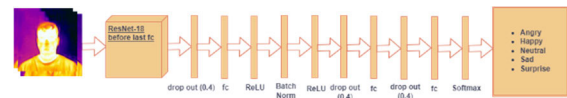


FIGURE 4. Proposed single model for infrared images.

Fig. 5 shows the proposed single model for MSX images using the CNN architecture ResNet-18. The last layer of ResNet-18 was modified according to the MSX dataset, and thus, two dropouts, two FC layers, and one SoftMax layer were added. These features were more distinguishable than those of infrared, as shown in Fig. 1 (c). A similar augmentation was used to expand the dataset. To reduce overfitting, we observed the loss and accuracy for both the training and validation sets to tune the parameters by tweaking the last layer of ResNet-18, as shown in Fig. 5.

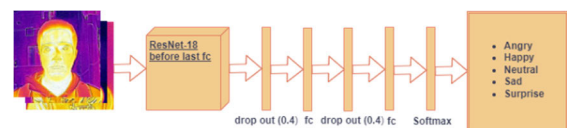


FIGURE 5. Proposed single model for MSX images.

Fig. 6 shows the proposed model using different combinations for concatenating features from the visible, IR, and MSX images using early fusion with 1-step and 3-step training. Feature extraction was performed to obtain useful information that can be exploited for image classification. For instance, feature extraction enables us to identify the eyes, nose, and mouth of an input human face, if we have an image of the human face. Using feature concatenation from different modalities, we can aggregate the performances of the combined models to increase the accuracy of individual models. The fusion architecture is important because it can significantly enhance system performance, robustness, and adaptability by integrating information from multiple modalities. To concatenate the features from the

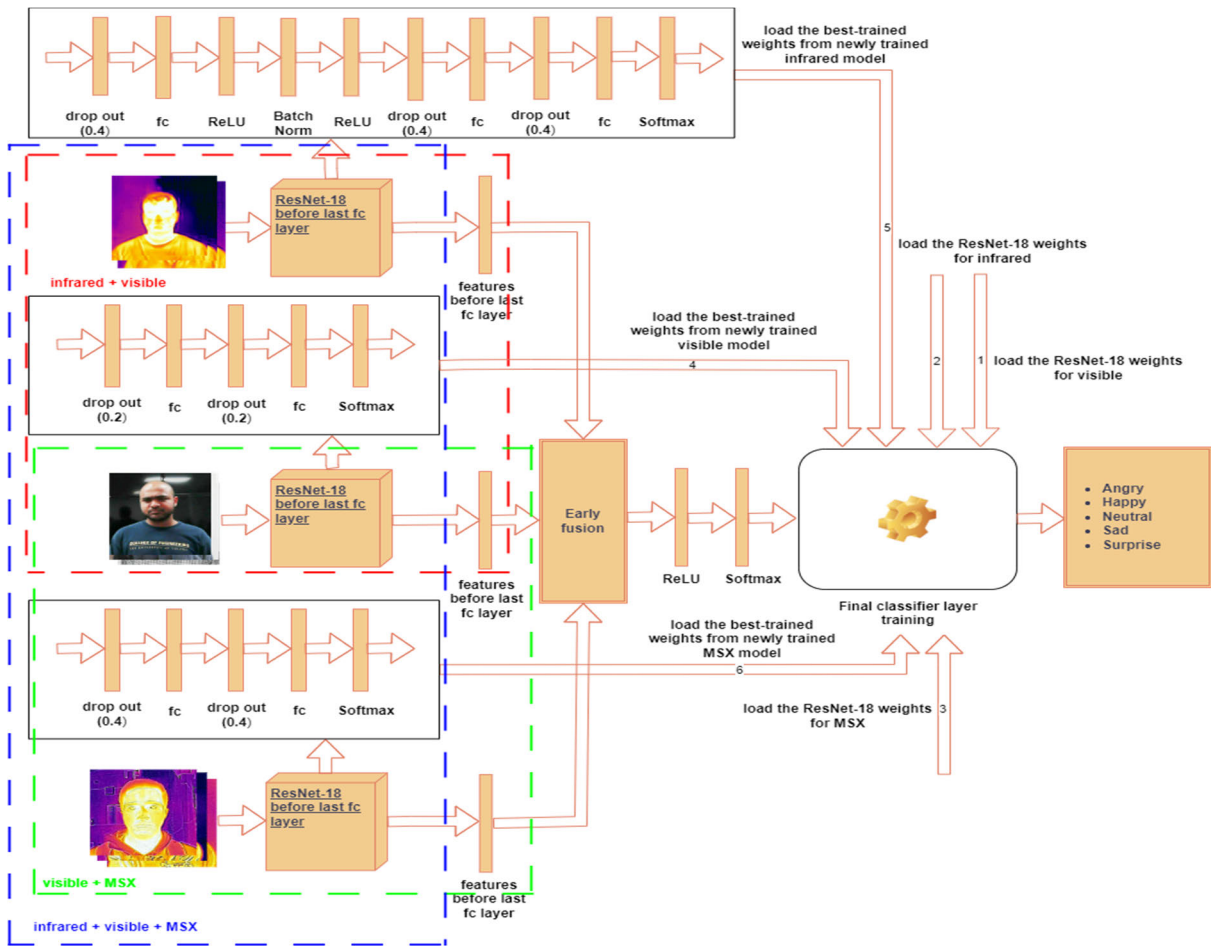


FIGURE 6. The proposed model for early fusion using different combinations for concatenating features from visible, infrared, and MSX datasets.

corresponding modalities, we used the modified architectures of the ResNet-18 models (Figs. 3, 4, and 5).

We developed the fusion models by combining different modalities. First, we combined the features from the visible and infrared images (shown by the bounding box with red dotted lines in Fig. 6) before the last FC layer and then passed them through a nonlinear activation function ReLU. A SoftMax layer was applied, and the features were trained using 1-step training. In 1-step training, we loaded the ResNet-18 weights (shown by lines 1, 2, and 3 in Fig. 6) to train the concatenated features for the classification of expressions. In 3-step training, we loaded the weights from newly pre-trained single models using the modified architecture of ResNet-18 (as shown in Figs. 3, 4, and 5) with updated features (shown by lines 4, 5, and 6 in Fig. 6).

Second, in a similar way, we combined the features from the visible and MSX images (shown by the bounding box with green dotted lines in Fig. 6) and again passed them through ReLU and SoftMax and trained them using 1-step (shown

by lines 1, 2, and 3 in Fig. 6) and 3-step training (shown by lines 4, 5, and 6 in Fig. 6).

Finally, in a similar fashion, we combined the features from the visible, infrared, and MSX datasets (shown by the bounding box with blue dotted lines in Fig. 6), and the SoftMax layer was applied after passing through ReLU. We then trained the classifier using 1-step (shown as lines 1, 2, and 3 in Fig. 6) and 3-step training (shown as lines 4, 5, and 6 in Fig. 6).

Fig. 7 shows the proposed late fusion model using different combinations of visible, infrared, and MSX models. We combined the outputs from the different modalities using different weighted factors. For combining the corresponding modalities, we loaded the corresponding validation datasets and chose the best combination of weight factors with the increment of 0.1, and then used the already chosen combinations to find the test accuracy. Here, we did not train the model again but combined the outputs from the corresponding modalities using the best-trained weights from the newly trained models

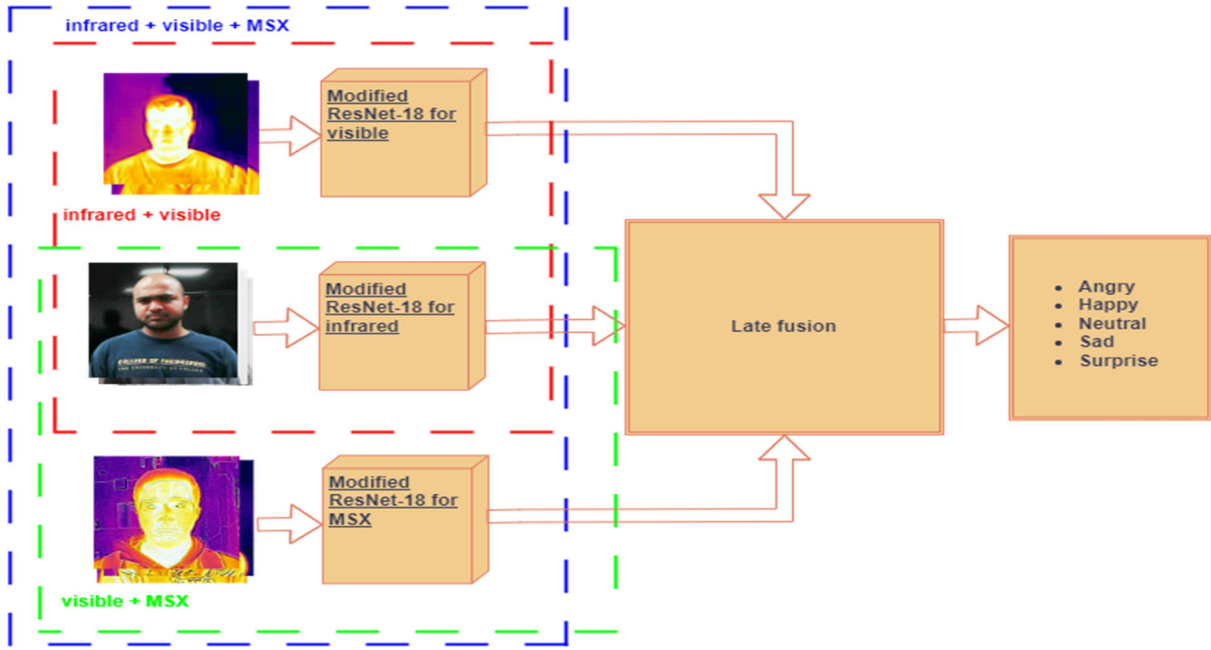


FIGURE 7. The proposed model for late fusion using different combinations for combining outputs from visible, infrared, and MSX datasets.

(using the modified architecture of ResNet-18, as shown in Figs. 3, 4, and 5). We evaluated the late fusion models using different weight factors.

Similar to the early fusion, we first combined the outputs from the visible and infrared datasets (shown by the bounding box with red dotted lines in Fig. 7). We used the weight factors of 0.7 and 0.3, respectively for the visible and infrared.

Second, we combined the outputs from the visible and MSX datasets (shown by the bounding box with green dotted lines in Fig. 7). Here, we used weight factors of 0.5 and 0.5 for visible and MSX, respectively.

Finally, we combined the outputs from the visible, infrared, and MSX images (shown by the bounding box with blue dotted lines in Fig. 7). We used weight factors of 0.5, 0.2, and 0.3 for the visible, infrared, and MSX datasets, respectively.

IV. RESULTS AND DISCUSSIONS

This section describes the experimental results and discusses our findings. For the NVIE database, as discussed in [32] we chose the samples according to the three conditions. First, the average intensity associated with a sample label must be larger than 1. Second, as three expressions (happiness, fear, and disgust) were successfully induced in most cases when the NVIE spontaneous database was constructed, the sample label should be one of happiness, fear, and disgust. Third, the sample should consist of both visible and infrared images.

Using the above conditions, we obtained 518 samples from 123 subjects under different illuminations. We optimized our models by choosing different hyperparameters like batch size and optimizers (shown in Table 2) to maximize our

TABLE 2. Different batch sizes and optimizers for our models.

Models	Batch Sizes	Optimizers
ResNet-18	128 for visible, infrared, 1-step & 3-step	Adam
Vgg-16	128 for visible & MSX 64 for 1-step & 3-step	SGD
ShuffleNetv2	128 for visible, MSX, 1-step & 3-step	Adam
MobileNetv2	128 for visible, MSX, 1-step & 3-step	SGD
GhostNet	128 for visible, MSX, 1-step & 3-step	Adam

performance measures. We used the ReduceLRonPlateau scheduler for all the models and modalities except 1-step and 3-step using ResNet-18 for the NVIE database where we used the CosineAnnealingLR scheduler.

Since, cross-entropy is a popular choice because it penalizes confident, which can lead to better convergence during training. It is also known as log loss, commonly used as a loss function in classification problems. It measures the performance of a classification model whose output is a probability value between 0 and 1. Cross-entropy loss increases as the predicted probability diverges from the actual label. we used cross-entropy as a loss function for all the modalities and models. All the experiments were conducted for 100 epochs. All models were implemented with Pytorch using an NVIDIA GeForce RTX 3030 GPU with 32GB of RAM. We evaluated all the models for accuracy, recall, precision, and F1-score.

A. FOR THE VISIBLE DATASET

Here, we trained the model for the visible dataset by modifying the last layer of ResNet-18, as previously described. We obtained an overall accuracy of 81.11%, as shown in the confusion matrix in Fig. 8. For the angry class, 72.22% of images were correctly classified, whereas 11.11% of images were classified as neutral, 5.56% as sad, and 11.11% as surprised. Similarly, 94.44% of happy images were correctly classified. In addition, for neutral expressions, we obtained 83.33% of images that were correctly classified, and for sad expressions, we obtained 72.22% of images that were correctly classified. Finally, 83.33% of images were correctly classified for the surprised expression.

	Angry	Happy	Neutral	Sad	Surprise
Actual Angry	72.22%	0.00%	11.11%	5.56%	11.11%
Actual Happy	0.00%	94.44%	5.56%	0.00%	0.00%
Actual Neutral	0.00%	0.00%	83.33%	16.67%	0.00%
Actual Sad	5.56%	0.00%	22.22%	72.22%	0.00%
Actual Surprise	0.00%	0.00%	16.67%	0.00%	83.33%

Predicted

FIGURE 8. Confusion matrix for the visible dataset using proposed ResNet-18.

B. FOR THE INFRARED DATASET

As discussed previously, we trained the model for the infrared dataset by modifying the last layer of ResNet-18. We obtained an overall accuracy of 50%, as shown in the confusion matrix in Fig. 9.

	Angry	Happy	Neutral	Sad	Surprise
Actual Angry	44.44%	5.56%	22.22%	27.78%	0.00%
Actual Happy	5.56%	72.22%	0.00%	22.22%	0.00%
Actual Neutral	16.67%	5.56%	44.44%	22.22%	11.11%
Actual Sad	16.67%	5.56%	22.22%	50.00%	5.56%
Actual Surprise	16.67%	0.00%	22.22%	22.22%	38.89%

Predicted

FIGURE 9. Confusion matrix for the infrared dataset using proposed ResNet-18.

C. FOR THE MSX DATASET

We trained the model for MSX images by modifying the last layer of ResNet-18, as previously discussed. We obtained an overall accuracy of 74.44%, as shown in the confusion matrix in Fig. 10.

	Angry	Happy	Neutral	Sad	Surprise
Actual Angry	61.11%	11.11%	11.11%	5.56%	11.11%
Actual Happy	0.00%	88.89%	0.00%	5.56%	5.56%
Actual Neutral	16.67%	0.00%	72.22%	11.11%	0.00%
Actual Sad	0.00%	0.00%	16.67%	72.22%	11.11%
Actual Surprise	11.11%	0.00%	0.00%	11.11%	77.78%

Predicted

FIGURE 10. Confusion matrix for the MSX dataset using proposed ResNet-18.

D. EARLY FUSION USING DIFFERENT COMBINATIONS BY CONCATENATING FEATURES FROM VISIBLE, INFRARED, AND MSX DATASETS

Using the proposed Resnet-18 model, when we concatenated the features using different combinations from the visible, infrared, and MSX datasets using early fusion, their confusion matrices are shown in Fig. 11.

When we combined the visible and infrared features using 1-step training, we obtained an overall accuracy of 83.33%, as indicated by the confusion matrix in Fig. 11 (a). Similarly, for 3-step training, we obtained an improved accuracy of 84.44%, as indicated by the confusion matrix in Fig. 11 (b).

When we concatenated the features from the visible and MSX using 1-step training, we obtained an overall accuracy of 85.56%, as shown by the confusion matrix in Fig. 11 (c). Similarly, for 3-step training, we obtained an improved accuracy of 87.78%, as indicated by the confusion matrix in Fig. 11 (d). Finally, when we combined the visible, infrared, and MSX features using 1-step training, we obtained an overall accuracy of 86.67%, as shown by the confusion matrix in Fig. 11 (e). Similarly, for 3-step training, we obtained an overall accuracy of 86.67%, as indicated by the confusion matrix in Fig. 11 (f).

In conclusion, when we concatenated the features from the visible and infrared regions, there was an improvement in accuracy from 1-step to 3-step training (83.33% to 84.44%). Similarly, when we concatenated the visible and MSX features, there was an improvement from 1-step to 3-step training (85.56% to 87.78%). Finally, when the visible, IR, and MSX features were concatenated, there was no improvement from 1-step to 3-step training (86.67% to 86.67%).

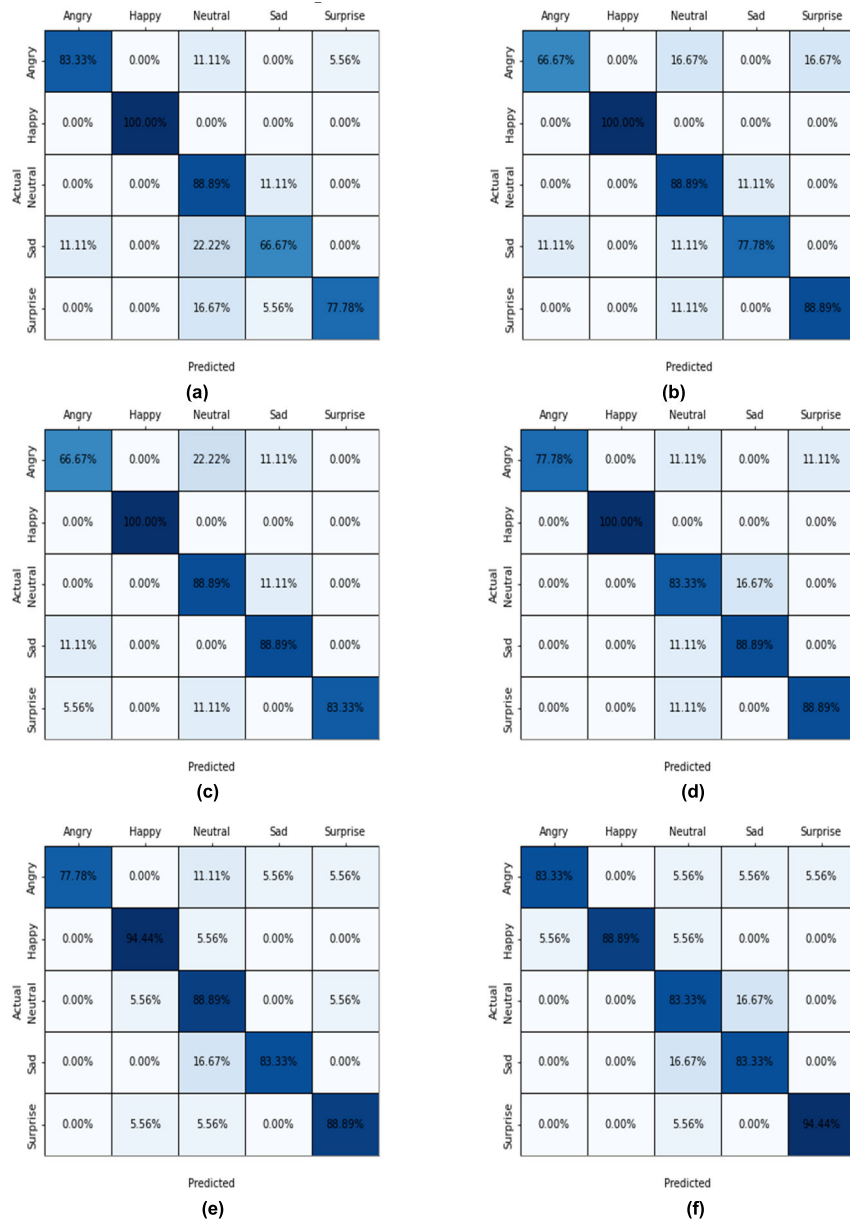


FIGURE 11. The confusion matrix for early fusion using different combinations by concatenating features from visible, Infrared, and MSX datasets using proposed ResNet-18: (a) 1-step training for visible and infrared (b) 3-step training for visible and infrared (c) 1-step training for visible and MSX (d) 3-step training for visible and MSX (e) 1-step training for visible, infrared and MSX (f) 3-step training for visible, infrared and MSX.

E. LATE FUSION USING DIFFERENT COMBINATIONS BY COMBINING OUTPUTS FROM VISIBLE, INFRARED, AND MSX DATASETS

Using the proposed Resnet-18 model, the confusion matrices for combining the outputs from the visible, infrared, and MSX datasets using different combinations for late fusion are shown in Fig. 12. When we combined the visible and infrared outputs, we obtained an overall accuracy of 82.22%, as indicated by the confusion matrix in Fig. 12 (a). Similarly, when we combined the visible and MSX outputs, we obtained an overall accuracy of 83.33%, as shown by the confusion matrix

in Fig. 12(b). Similarly, when we combined the outputs from the visible, infrared, and MSX inputs, we obtained an overall accuracy of 84.44%, as shown by the confusion matrix in Fig. 12 (c).

F. HEAT MAP ANALYSIS

Class activation mapping (CAM) was used to evaluate activation area in different expression classes. Fig. 13 shows the heat maps for the disgust, fear, and happy expressions for the NVIE database using the proposed ResNet-18 model. For the disgust expression in the first row, wrinkling of the nose area

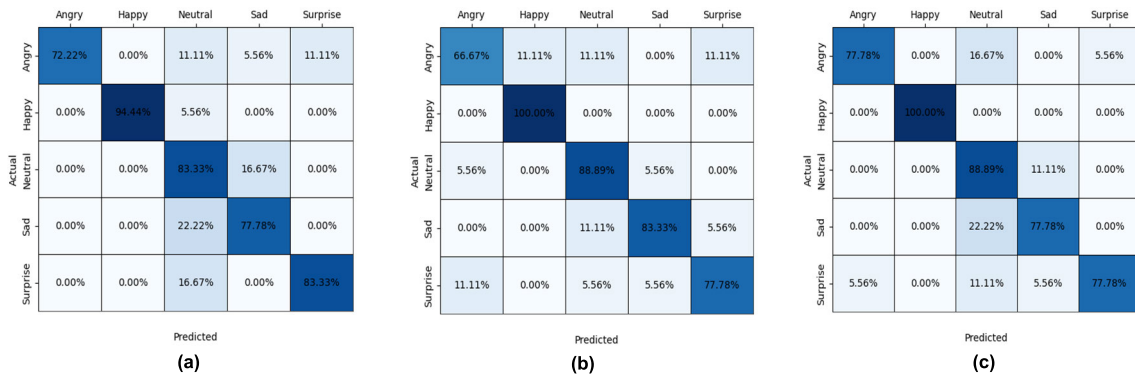


FIGURE 12. The confusion matrix for late fusion using different combinations by combining outputs from visible, infrared, and MSX datasets using proposed ResNet-18 (a) visible and infrared (b) visible and MSX (c) visible, infrared, and MSX.

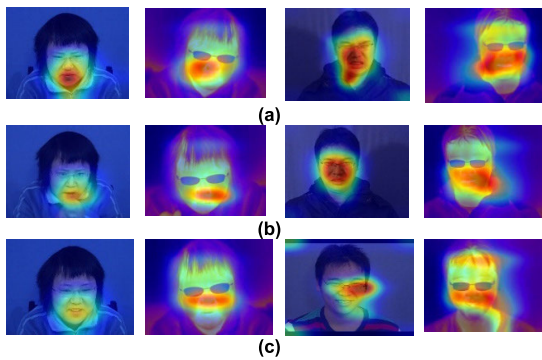


FIGURE 13. The heat map analysis for visible and infrared models for different expressions from the NVIE database using the proposed ResNet-18 model: first and third columns for visible while the second and fourth columns are corresponding infrared (a) disgust (b) fear (c) happy.

is activated and the mouth is tightened with a raised upper lip. For the fear expression in the second row, the mouth is slightly opened with pulled-back corners and the eyes are widened with raised eyebrows. Similarly, for the happy expression in the third row, the mouth is upturned with the exposing of teeth, and the cheeks are raised due to the contraction of the zygomaticus major muscle. The CAM results show that the proposed model activate different areas in different expressions.

G. COMPARISON WITH PREVIOUS WORKS

Table 3 presents a comparison of our proposed ResNet-18 model with the VIRI [35] and NVIE [32] databases. The VIRI database in [35] presented an accuracy of 71.19% for the visible images, recall of 0.71, precision of 0.78, and F1-score of 0.75 while the NVIE database in [32] presents an accuracy of 76.82%. Our proposed model presents an improved accuracy of 81.11% for the VIRI and 82.71% for the NVIE database. Similarly, other performance measures were also improved. For infrared images, the VIRI database in [35] presented an accuracy of 77.34%, a recall of 0.77, a precision of 0.78, and an F1-score of 0.78 while the NVIE database presented an accuracy of 52.90%. Our proposed model provides an

accuracy of 50% for the NVIE database. This accuracy is lower than that of the baseline because we encountered a large amount of overfitting during parameter tuning for loss and accuracy in the limited details of the infrared images. For the NVIE database, our model improved the accuracy of the infrared dataset.

Using 1-step training, the VIRI database in [35] presented an accuracy of 82.26%, recall of 0.82, precision of 0.85, and F1-score of 0.83 while the NVIE database in [32] presented an accuracy of 76.82%. Our proposed model yielded an accuracy of 83.33% and similarly, other performance measures are also improved. For 3-step training, the accuracy was increased to 84.44%, which was higher than the VIRI as well as NVIE. Similarly, the recall, precision, and F1-score were also improved. Our model also improved by combining the outputs from the visible and infrared using the weighted sum.

Table 4 presents our additional experiments when we combined the MSX with the visible and infrared. For MSX, the accuracy comes to 74.44% but when we combine the outputs from the visible with MSX, the accuracy was 83.33%. For 1-step and 3-step training, the accuracy was 85.56% and 87.78%.

In the same way, when we combined the outputs from the visible, infrared, and MSX, the accuracy was 84.44%. Similarly, for 1-step and 3-step training, the accuracy was 86.67%.

Table 5. shows the experimental results for other backbones like Vgg-16, Shufflenetv2, MobileNetv2, and GhostNet for the VIRI database. For these backbones, we used the same output layers of the corresponding modalities, we used for the modified ResNet-18 model. Since for 1-step and 3-step training, using the visible and MSX, we got the higher accuracy (shown in Table 4) that’s why we only considered these modalities for testing other backbones. Using Vgg-16, for the visible dataset, the accuracy was 76.67% accuracy while for MSX, the accuracy was 60%. Similarly, for 1-step and 3-step training, the accuracies were 77.78% and 78.89%, respectively. For Shufflenetv2, our accuracies were 77.78% and 61.11%, respectively for the visible and MSX. For 1-step and 3-step training, our accuracy was 81.11%

TABLE 3. Comparison of the proposed ResNet-18 (for visible and infrared) with the VIRI database [35] and NVIE database [32].

Modality	Accuracy	Recall	Precision	F1-Score
Visible in [35] for VIRI	71.19%	0.71	0.78	0.75
Visible in [32] for NVIE	76.82%	-	-	-
Proposed visible for VIRI	81.11%	0.81	0.84	0.82
Proposed visible for NVIE	82.71%	-	-	-
infrared in [35] for VIRI	77.34%	0.77	0.78	0.78
infrared in [32] for NVIE	52.90%	-	-	-
Proposed infrared for VIRI	50%	0.50	0.54	0.51
Proposed infrared for NVIE	55.56%	-	-	-
Visible+infrared (1-step training) in [35] for VIRI	82.26%	0.82	0.85	0.83
Visible + infrared in [32] for NVIE	76.82%	-	-	-
Proposed visible + infrared (1-step training) for VIRI	83.33%	0.83	0.85	0.84
Proposed visible + infrared (3-step training) for VIRI	84.44%	0.84	0.85	0.84
Proposed visible + infrared (1-step training) for NVIE	84%	-	-	-
Proposed visible + infrared (3-step training) for NVIE	84%	-	-	-
Proposed visible + infrared (combine outputs) for VIRI	82.22%	0.82	0.85	0.83

and 82.22%, respectively. For MobileNetv2, our accuracies were 80% and 58.89%, respectively for the visible and MSX and for 1-step and 3-step training, our accuracies were 80%

TABLE 4. Additional experiments using the proposed ResNet-18 for combining MSX with visible and infrared for the VIRI database [35].

Modality	Accuracy	Recall	Precision	F1-Score
Proposed MSX	74.44%	0.74	0.74	0.74
Proposed visible + MSX (combine outputs)	83.33%	0.83	0.83	0.83
Proposed visible + MSX (1-step training)	85.56%	0.86	0.87	0.86
Proposed visible + MSX (3-step training)	87.78%	0.88	0.89	0.88
Proposed visible + IR + MSX (combine outputs)	84.44%	0.84	0.87	0.85
Proposed visible + IR + MSX (1-step training)	86.67%	0.87	0.88	0.87
Proposed visible + IR + MSX (3-step training)	86.67%	0.87	0.88	0.87

TABLE 5. The experimental results of other backbones for the visible and MSX for the VIRI database [35].

Models	visible	MSX	1-step training	3-step training
Vgg-16	76.67%	60%	77.78%	78.89%
ShuffleNetv2	77.78%	61.11%	81%	82.22%
MobileNetv2	80%	58.89%	80%	8333%
GhostNet	74.44%	62.22%	75.55%	77.78%

and 83.33%. For GhostNet, the accuracy for the visible was 74.44% and for the MSX, the accuracy was 62.22%. For 1-step and 3-step training, the accuracies were 75.55% and 77.78%, respectively. If we compare the accuracies of other backbones with the ResNet-18 model (Table 3, 4 and Table 5), still the performance of ResNet-18 was better for the visible and MSX as well as for 1-step and 3-step training.

V. CONCLUSION

This paper proposes a novel multimodal method using early and late fusion to classify facial expressions using different combinations of visible, infrared, and MSX modalities. This method uses transfer learning with a modified architecture of the ResNet-18 model for this purpose. In the early fusion, we concatenated the features from both modalities and then trained the models using 1-step and 3-step training. The results show that 3-step training improves the performance of the visible and IR modalities. For late fusion, we combined the outputs from different modalities using weighted sum-matio; for this purpose, we used different weighted factors.

The results were also compared with baseline values, which clearly showed improved performance. We also performed additional experiments by combining the other modality (MSX) available at baseline with existing modalities to gain additional insight for researchers, which might be helpful in their research. The experimental results obtained by combining the additional modality (MSX) with existing modalities improved the performance, confirming that it can assist other researchers in improving the performance of their systems. We also checked the performance of other models (Vgg-16, ShuffleNetv2, MobileNetv2, and GhostNet) to combine the MSX with the visible but the ResNet-18 outperformed the other backbones.

There is a shortcoming of our proposed work. The datasets we used were the VIRI and the NVIE, which were very small. Although we used augmentations to increase the size of the datasets but still, we faced overfitting, especially for infrared.

REFERENCES

- [1] A. Mehrabian, "Communication without words," *Psychol Today*, vol. 2, no. 4, pp. 193–200, 1968.
- [2] H. Harashima, "3-D model-based synthesis of facial expressions and shape deformation," in *Proc. Human Interface*, 1989, pp. 157–166.
- [3] K. Mase, "An application of optical flow-extraction of facial expression," in *Proc. IAPR Workshop Mach. Vis. Appl.*, Nov. 1990, pp. 195–198. [Online]. Available: file:///C:/Users/Owner/Downloads/An_Application_of_Optical_Flow_-_Extraction_of_Fac.pdf
- [4] K. Mase, "Recognition of facial expression from optical flow," *IEICE Trans. Inf. Syst.*, vol. 74, no. 1, pp. 3474–3483, 1991.
- [5] K. Matsuno, "Recognition of facial expressions using potential net and KL expansion," *IEICE Trans.*, vol. 77, no. 8, pp. 1591–1600, 1994.
- [6] H. Kobayashi and F. Hara, "Analysis of neural network recognition characteristics of 6 basic facial expressions," *Trans. Jpn. Soc. Mech. Engineers Ser. C*, vol. 61, no. 582, pp. 678–685, 1995.
- [7] C. Filippini, D. Perpetuini, D. Cardone, A. M. Chiarelli, and A. Merla, "Thermal infrared imaging-based affective computing and its application to facilitate human robot interaction: A review," *Appl. Sci.*, vol. 10, no. 8, p. 2924, Apr. 2020.
- [8] C. Goulart, C. Valadao, D. Delisle-Rodriguez, E. Caldeira, and T. Bastos, "Emotion analysis in children through facial emissivity of infrared thermal imaging," *PLoS ONE*, vol. 14, no. 3, Mar. 2019, Art. no. e0212928.
- [9] J. Clay-Warner and D. T. Robinson, "Infrared thermography as a measure of emotion response," *Emotion Rev.*, vol. 7, no. 2, pp. 157–162, Apr. 2015.
- [10] D. V. Sang and N. Van Dat, "Facial expression recognition using deep convolutional neural networks," in *Proc. 9th Int. Conf. Knowl. Syst. Eng. (KSE)*, Oct. 2017, pp. 130–135. [Online]. Available: <https://ieeexplore.ieee.org/document/8119447>
- [11] C.-D. Wu and L.-H. Chen, "Facial emotion recognition using deep learning," pp. 1–5, 2019, *arXiv:1910.11113*.
- [12] G. Yang, "Emotion recognition using deep neural network with vectorized facial features," in *Proc. IEEE Int. Conf. Electro/Inf. Technol. (EIT)*, May 2018, pp. 0318–0322. [Online]. Available: <https://ieeexplore.ieee.org/document/8500080>
- [13] S. Saeed, A. A. Shah, M. K. Ehsan, M. R. Amirzadeh, A. Mahmood, and T. Mezgebo, "Automated facial expression recognition framework using deep learning," *J. Healthcare Eng.*, vol. 2022, pp. 1–11, Mar. 2022.
- [14] J. Haddad, "3D-CNN for facial emotion recognition in videos," in *Proc. Adv. Vis. Comput., 15th Int. Symp.*, San Diego, CA, USA, 2020, pp. 298–309.
- [15] D. Y. Liliana, "Emotion recognition from facial expression using deep convolutional neural network," *J. Physics: Conf. Ser.*, vol. 1193, Apr. 2019, Art. no. 012004.
- [16] S. M. Stuit, T. M. Kootstra, D. Terburg, C. van den Boomen, M. J. van der Smagt, J. L. Kenemans, and S. Van der Stigchel, "The image features of emotional faces that predict the initial eye movement to a face," *Sci. Rep.*, vol. 11, no. 1, p. 8287, Apr. 2021.
- [17] R. Singh, S. Saurav, T. Kumar, R. Saini, A. Vohra, and S. Singh, "Facial expression recognition in videos using hybrid CNN & ConvLSTM," *Int. J. Inf. Technol.*, vol. 15, no. 4, pp. 1819–1830, Apr. 2023.
- [18] S. Minaee, M. Minaei, and A. Abdolrashidi, "Deep-emotion: Facial expression recognition using attentional convolutional network," *Sensors*, vol. 21, no. 9, p. 3046, Apr. 2021.
- [19] D. Dukić and A. S. Krzic, "Real-time facial expression recognition using deep learning with application in the active classroom environment," *Electronics*, vol. 11, no. 8, p. 1240, 2022.
- [20] Y. Yoshitomi, T. Miyaura, S. Tomita, and S. Kimura, "Face identification using thermal image processing," in *Proc. 6th IEEE Int. Workshop Robot Hum. Commun. RO-MAN SENDAI*, Oct. 1997, pp. 374–379.
- [21] Y. Yoshitomi, N. Miyawaki, S. Tomita, and S. Kimura, "Facial expression recognition using thermal image processing and neural network," in *Proc. 6th IEEE Int. Workshop Robot Hum. Commun. RO-MAN97 SENDAI*, vol. 46, 2002, pp. 542–567.
- [22] P. Shen, S. Wang and Z. Liu, "Facial expression recognition from infrared thermal videos," in *Proc. Intell. Auton. Syst. 12th Int. Conf. IAS*, vol. 2, Jeju Island, South Korea, 2013, pp. 323–333.
- [23] M. M. Khan, M. Ingleby, and R. D. Ward, "Automated facial expression classification and affect interpretation using infrared measurement of facial skin temperature variations," *ACM Trans. Auto. Adapt. Syst.*, vol. 1, no. 1, pp. 91–113, Sep. 2006.
- [24] A. Basu. (2015). *Human Emotion Recognition From Facial Thermal Image Using Histogram Based Features and Multi-class Support Vector Machine*. [Online]. Available: <https://www.ndt.net/article/qirt2015/papers/CP0055.pdf>
- [25] L. A. Jeni, H. Hashimoto, and T. Kubota, "Robust facial expression recognition using near infrared cameras," *J. Adv. Comput. Intell. Intell. Informat.*, vol. 16, no. 2, pp. 341–348, Mar. 2012.
- [26] Y. M. Elbarawy, N. I. Ghali, and R. S. El-Sayed, "Facial expressions recognition in thermal images based on deep learning techniques," *Int. J. Image, Graph. Signal Process.*, vol. 11, no. 10, pp. 1–7, Oct. 2019.
- [27] A. Prabhakaran, "Thermal facial expression recognition using modified resnet152," in *Advances in Computing and Network Communications*. Singapore: Springer, 2021, pp. 389–396.
- [28] C. C. Atabansi, T. Chen, R. Cao, and X. Xu, "Transfer learning technique with VGG-16 for near-infrared facial expression recognition," *J. Phys., Conf. Ser.*, vol. 1873, no. 1, Apr. 2021, Art. no. 012033.
- [29] A. Bhattacharyya, S. Chatterjee, S. Sen, A. Sinitca, D. Kaplun, and R. Sarkar, "A deep learning model for classifying human facial expressions from infrared thermal images," *Sci. Rep.*, vol. 11, no. 1, p. 20696, Oct. 2021.
- [30] B. Assiri and M. A. Hossain, "Face emotion recognition based on infrared thermal imagery by applying machine learning and parallelism," *Math. Biosciences Eng.*, vol. 20, no. 1, pp. 913–929, 2022.
- [31] J. Zou, J. Li, J. Wei, Z. Li, and X. Yang, "Facial expression recognition based on the fusion of infrared and visible image," *J. Artif. Intell.*, vol. 3, no. 3, pp. 123–134, 2021.
- [32] S. Wang, S. He, Y. Wu, M. He, and Q. Ji, "Fusion of visible and thermal images for facial expression recognition," *Frontiers Comput. Sci.*, vol. 8, no. 2, pp. 232–242, Apr. 2014.
- [33] S. Wang, Z. Liu, S. Lv, Y. Lv, G. Wu, P. Peng, F. Chen, and X. Wang, "A natural visible and infrared facial expression database for expression recognition and emotion inference," *IEEE Trans. Multimedia*, vol. 12, no. 7, pp. 682–691, Nov. 2010.
- [34] H. Nguyen, N. Tran, H. D. Nguyen, L. Nguyen, and K. Kotani, "KTFFv2: Multimodal facial emotion database and its analysis," *IEEE Access*, vol. 11, pp. 17811–17822, 2023.
- [35] M. F. H. Siddiqui and A. Y. Javadi, "A multimodal facial emotion recognition framework through the fusion of speech with visible and infrared images," *Multimodal Technol. Interact.*, vol. 4, no. 3, p. 46, Aug. 2020.
- [36] D. Sarwinda, R. H. Paradisa, A. Bustamam, and P. Anggia, "Deep learning in image classification using residual network (ResNet) variants for detection of colorectal cancer," *Proc. Comput. Sci.*, vol. 179, pp. 423–431, Jan. 2021.
- [37] J. Deng, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Miami, FL, USA, 2009, pp. 22–24.



MUHAMMAD TAHIR NASEEM (Member, IEEE) was born in Sargodha, Pakistan, in 1983. He received the B.S. degree (Hons.) in computer science from the University of the Punjab, Lahore, Pakistan, in 2005, the M.S. degree in electronic engineering from International Islamic University, Islamabad, Pakistan, in 2008, and the Ph.D. degree in electronic engineering from Isra University, Hyderabad, Pakistan, in 2015. From 2017 to 2021, he was a Faculty Member of the Riphah School of Computing and Innovation (RSCI), Riphah International University, Lahore. Since 2022, he has been a Research Professor with the Research Institute of Human Ecology, Yeungnam University, South Korea. His research interests include computer vision, facial expressions, infrared thermography, sensor fusion, gaits, and machine and deep learning.



CHAN-SU LEE (Member, IEEE) received the B.S. degree in electronics engineering from Yonsei University, in 1995, the M.S. degree in electrical engineering from the Korea Advanced Institute of Science and Technology, in 1997, and the Ph.D. degree in computer science from Rutgers, The State University of New Jersey, in May 2007. From 1997 to 2001, he was a Member Research Engineer with the Electronics and Telecommunications Research Institute (ETRI). He is currently a Professor with the Department of Electronic Engineering, Yeungnam

University, South Korea. His research interests include computer vision, pattern recognition, machine learning, biometrics, gesture and facial expression analysis, smart lighting control, and human visual perception. He is a member of the IEEE Computer Society.



NA-HYUN KIM (Student Member, IEEE) received the B.S. degree in electronics engineering from Yeungnam University, in 2022, where she is currently pursuing the master's degree with the Department of Electronics Engineering. Her research interests include human action recognition, facial expression recognition, and sequence modeling of dynamic human motions.

...