West Texas A&M University™

A Multimodal Facial Emotion Recognition Framework through the Fusion of Speech with Visible and Infrared Images

Mohammad Faridul Haque Siddiqui[1], PhD, Ahmad Y. Javaid, PhD[2];
[1]West Texas A&M University, ECSM, [2]University of Toledo, EECS;

ABSTRACT

The exigency of emotion recognition is pushing the envelope for meticulous strategies of discerning actual emotions through the use of superior multimodal techniques. This work presents a multimodal automatic emotion recognition (AER) framework capable of differentiating between expressed emotions with high accuracy. The contribution involves implementing an ensemble-based approach for the AER through the fusion of visible images and infrared (IR) images with speech. The framework is implemented in two layers, where the first layer detects emotions using single modalities while the second layer combines the modalities and classifies emotions. Convolutional Neural Networks (CNN), Fig 2, have been used for feature extraction and classification. A hybrid fusion approach comprising early (feature-level) and late (decision-level) fusion, was applied to combine the features and the decisions at different stages. The output of the CNN trained with voice samples of the RAVDESS database was combined with the image classifier's output using decision-level fusion to obtain the final decision. An accuracy of 86.36% and similar recall (0.86), precision (0.88), and f-measure (0.87) scores were obtained. A comparison with contemporary work endorsed the competitiveness of the framework with the rationale for exclusivity in attaining this accuracy in wild backgrounds and light-invariant conditions.

BACKGROUND

This paper presents a novel automatic emotion recognition (AER) framework that utilizes the fusion of visible images, infrared images, and speech to accurately identify emotions in individuals. The proposed framework follows an ensemble-based approach, combining different techniques' strengths to achieve the best results. In order to develop the framework, the authors have adhered to certain baselines. One of the baselines was to avoid the overwhelming use of sensors to eliminate any unwanted stress or anxiety induced due to monitoring. This was done to ensure that the framework does not cause any discomfort to the individuals being monitored. The framework relies on facial expressions as the principal modality for emotion recognition. The use of infrared images is proposed to counter the limitations posed in previous approaches, such as poor performance in low-light conditions. The speech is used as a supporting modality to refine the classification further. Feature and decision-level fusions were applied at different stages to make the framework light-invariant and suitable for use in real-world conditions. This allows the framework to work accurately in varying lighting conditions. The proposed framework offers a promising solution for accurate emotion recognition in real-world scenarios by combining multiple modalities' strengths for light-invariant AER in-the-wild conditions.

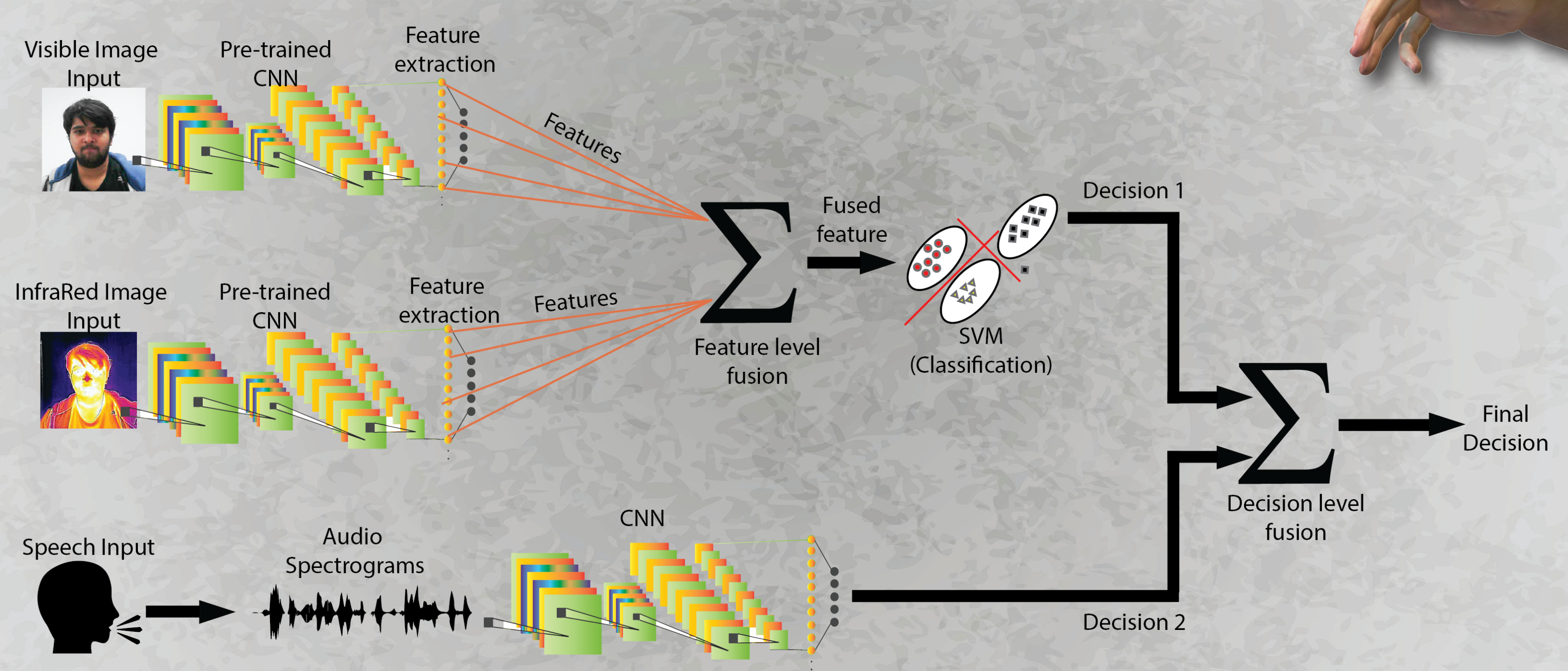


FIG 1: PROPOSED MODEL

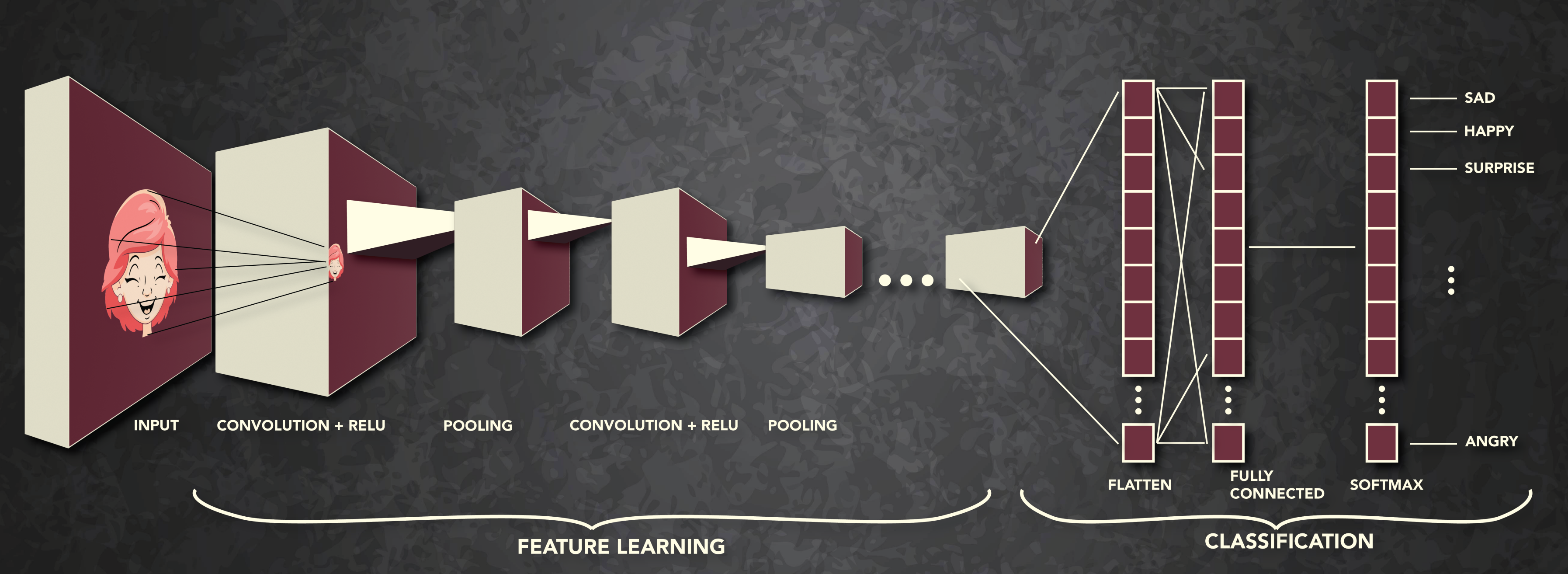


FIG 2: CONVOLUTIONAL NEURAL NETWORK

METHODOLOGY

The proposed method for emotion recognition uses an Artificial Neural Network (ANN) for deep learning and ensemble-based classification. The model consists of two layers of detection, with the first layer involving training two CNNs using visible and infrared images individually. Transfer learning is incorporated to extract the features of the images, and a feature-level fusion is then applied, with the fused feature being fed to an SVM for classification. Additionally, a third CNN is deployed in the first layer to learn emotions from speech by incorporating audio spectrograms for training the ANN.

In the first stage of the model, AER with visible and IR images, transfer learning is used to adapt a pre-trained CNN, the AlexNet, to the VIRI DB. The AlexNet has been trained on over a million images and can classify them into 1000 categories. The original 1440 X 1080 pixel images in the database are down-sampled to 227 X 227 to feed to the input layer of the AlexNet. The last three layers are replaced with new layers to adapt to the new dataset, resulting in a 27-layer CNN. The features generated at the last fully connected layer are extracted and used for the feature-level fusion.

A feature-level fusion technique based on Canonical Correlation Analysis (CCA) is applied to combine the features extracted from the two CNNs in the first layer. The fused feature vector is more discriminative than any of the individual input vectors and is fed to an SVM classifier for the intermediate classification of AER. The SVM and the third CNN's output are then fed to a decision-level fusion in the second layer, with the final classification obtained as the second layer's output. The fusion of visible and infrared images classified emotions depicted with an accuracy of 82.26%. Speech samples from the RAVDESS multimodal dataset resulted in detection accuracy of 73.28%. The decisions from the images and speech were fused using decision templates (a decision level fusion technique) to achieve an overall accuracy of 86.36%. A comparison of the accuracy with the recent work in multimodal emotion detection proves the framework's superiority

Modality	Accuracy	Recall	Precision	F-Measure
V	71.19%	0.71	0.78	0.75
IR	77.34%	0.77	0.78	0.78
V+IR	82.26%	0.82	0.85	0.83
S	73.28%	0.73	0.72	0.72
V+IR+S	86.36%	0.86	0.88	0.87

FIG 3: Comparison of the AER metrics by different modalities (Visible: V, Infrared: IR, Speech: S).

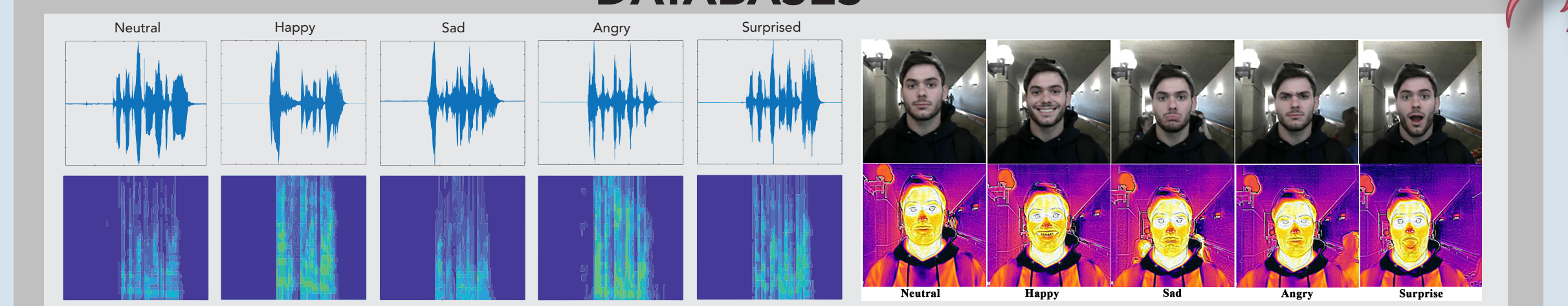
REFERENCES

[1] Siddiqui, Mohammad Faridul Haque, and Ahmad Y. Javaid. "A multimodal facial emotion recognition framework through the fusion of speech with visible and infrared images." Multimodal Technologies and Interaction 4.3 (2020): 46.

[2] Siddiqui, Mohammad Faridul Haque, et al. "A Survey on Databases for Multimodal Emotion Recognition and an Introduction to the VIRI (Visible and InfraRed Image) Database." Multimodal Technologies and Interaction 6.6 (2022): 47.

[3] Mehta, Dhwani, Mohammad Faridul Haque Siddiqui, and Ahmad Y. Javaid. "Facial emotion recognition: A survey and real-world user experiences in mixed reality." Sensors 18.2 (2018): 416.

DATABASES



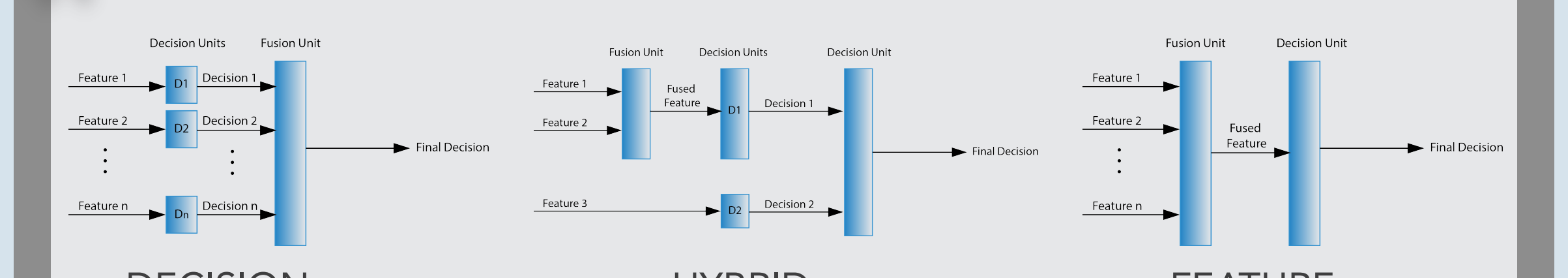
Spectograms of RAVDESS DB

VIRI DB

The VIRI database is a facial image database developed at The University of Toledo, which includes both visible and infrared images of spontaneous facial expressions in uncontrolled wild backgrounds. The database comprises pictures from on-campus students, with five expressions captured (happy, sad, angry, surprised, and neutral) from 110 subjects, resulting in 550 images in a radiometric JPEG format. The radiometric JPEG comprises three formats, visible, infrared, and MSX format, and the database contains all three formats. The images were preprocessed before use, with subjects being brought to the center and darkness reduced, and each image was reduced to 227 X 227 pixels. Image augmentation techniques were also applied to increase the number of images and bring more heterogeneity for training a CNN for AER by IR and visible images.

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) is an audio-visual emotional database developed at the SMART Lab of Ryerson University, Canada. It contains validated emotional speech and songs recorded by 24 professional actors in a North American accent. The database covers seven emotional states in speech and five emotions in songs, and it has two emotional levels (normal and strong). It has 2452 audio recordings and is available for download under a creative commons license. The audio data was pre-processed by trimming the length of the files to make them uniform, eradicating any issues that might arise during the creation of audio spectrograms and the CNN training.

FUSION



DECISION LEVEL FUSION

HYBRID FUSION

FEATURE LEVEL FUSION

Fusion is a technique used to improve the performance of a machine learning model by integrating different modalities to form a coherent combination of input. There are two main types of fusion: feature level fusion and decision level fusion, and hybrid approach, which combines both feature level and decision level fusion techniques. Each method has its own advantages and disadvantages and can be used in different situations depending on the problem and the data. In this work, a feature-level fusion technique based on canonical Correlation Analysis (CCA) was applied to combine the features extracted from the two CNNs in the first layer of the framework. The fused feature vector was more discriminative than any of the individual input vectors and was fed to an SVM classifier for the intermediate classification of AER.

In the second stage of the framework, the classifications from speech and images were combined using a decision-level fusion. A late fusion technique called weighted decision templates (WDT) was applied to integrate the decisions emanating from SVM (for images) and CNN (for speech). The WDT algorithm assigns weights to each classifier based on its performance and output. It uses the fusion rule of the weighted sum to evaluate the final probability of a result after fusion belonging to a specific emotion.

MODALITIES

SPEECH VISIBLE IMAGES INFRARED IMAGES

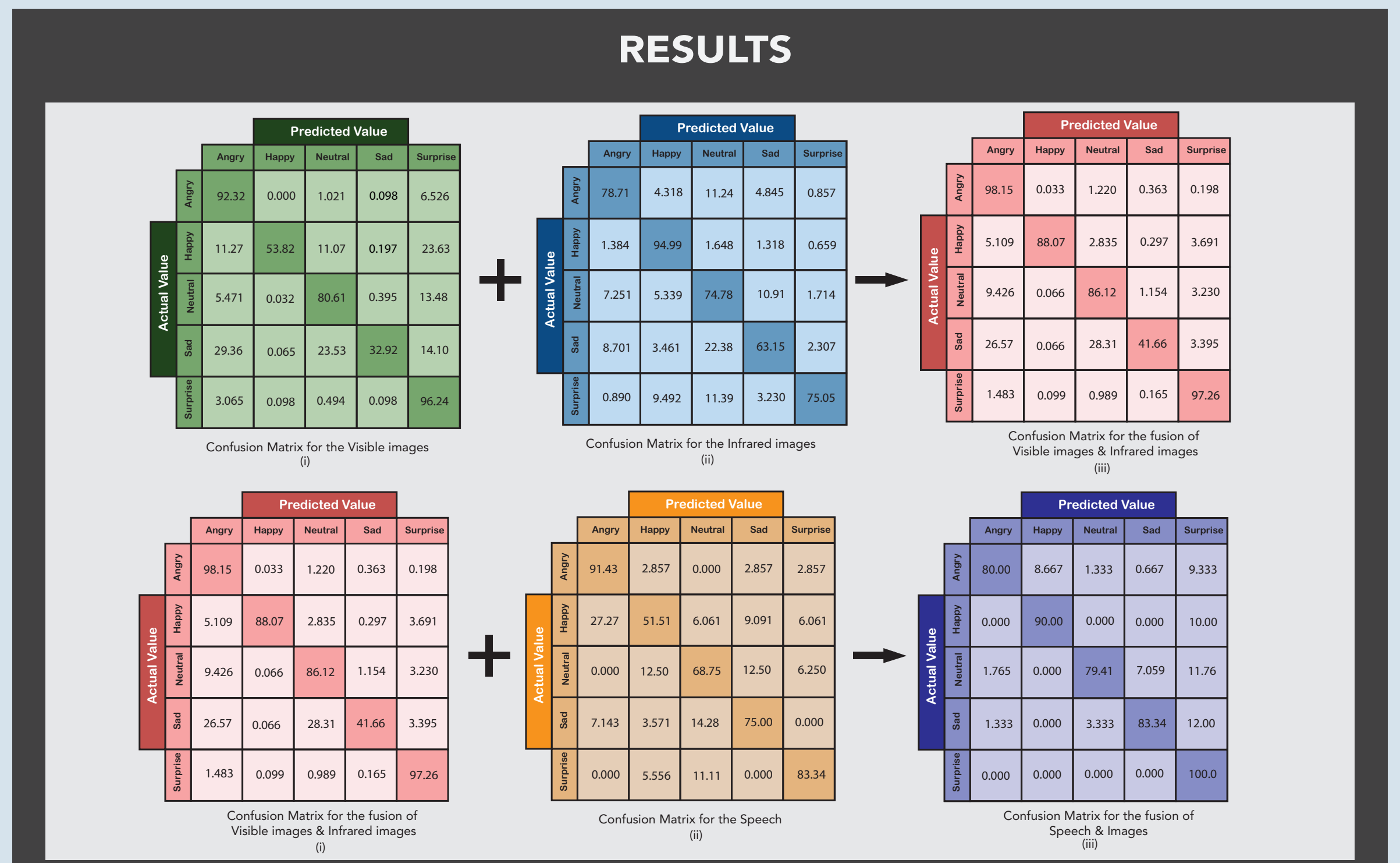
Three modalities were used for AER in this work:

The prime modality in this work for emotion recognition is visible images. This modality captures the emotional expressions of an individual through the use of cameras and image processing techniques. Visible images can capture the subtle changes in facial expressions when a person experiences an emotion. These changes include changes in the shape of the eyes, mouth, and eyebrows, as well as changes in other features of the face. Visible images train the CNN models by transfer learning to recognize and classify different emotional states.

Another modality (that was secondary) for emotion recognition is infrared images. This modality captures the thermal radiation emitted by an individual's face to detect changes in temperature and blood flow that occur when a person experiences an emotion. Infrared images can capture changes in temperature in different regions of the face, such as the eyes, cheeks, and forehead, that are not visible in visible images. Infrared images train the CNN models by transfer learning to recognize and classify different emotional states.

A third modality for emotion recognition that was secondary is speech. This modality captures an individual's emotional state by analyzing their speech patterns and prosody. Speech patterns include changes in pitch, volume, and rhythm that occur when a person experiences an emotion. Speech signals were converted into their respective spectrograms. They were used to train the CNN models to recognize and classify different emotional states, such as happy, surprise, sad, angry, and neutral.

RESULTS



Predicted Value

Actual Value

	Angry	Happy	Neutral	Sad	Surprise
Angry	92.32	0.000	1.021	0.098	6.526
Happy	11.27	53.82	11.07	0.197	23.63
Neutral	5.471	0.032	80.61	0.395	13.48
Sad	29.36	0.065	23.53	32.92	14.10
Surprise	3.065	0.098	0.494	0.098	96.24

Confusion Matrix for the Visible images
(i)

	Angry	Happy	Neutral	Sad	Surprise
Angry	78.71	4.318	11.24	4.845	0.857
Happy	1.384	94.99	1.648	1.318	0.659
Neutral	7.251	5.339	74.78	10.91	1.714
Sad	8.701	3.461	22.38	63.15	2.307
Surprise	0.890	9.492	11.39	3.230	75.05

Confusion Matrix for the Infrared images
(ii)

	Angry	Happy	Neutral	Sad	Surprise
Angry	98.15	0.033	1.220	0.363	0.198
Happy	5.109	88.07	2.835	0.297	3.691
Neutral	9.426	0.066	86.12	1.154	3.230
Sad	26.57	0.066	28.31	41.66	3.395
Surprise	1.483	0.099	0.989	0.165	97.26

Confusion Matrix for the fusion of Visible images & Infrared images
(iii)

	Angry	Happy	Neutral	Sad	Surprise
Angry	98.15	0.033	1.220	0.363	0.198
Happy	5.109	88.07	2.835	0.297	3.691
Neutral	9.426	0.066	86.12	1.154	3.230
Sad	26.57	0.066	28.31	41.66	3.395
Surprise	1.483	0.099	0.989	0.165	97.26

Confusion Matrix for the fusion of Visible images & Infrared images
(i)

	Angry	Happy	Neutral	Sad	Surprise
Angry	91.43	2.857	0.000	2.857	2.857
Happy	27.27	51.51	6.061	9.091	6.061
Neutral	0.000	12.50	68.75	12.50	6.250
Sad	7.143	3.571	14.28	75.00	0.000
Surprise	0.000	5.556	11.11	0.000	83.34

Confusion Matrix for the Speech
(ii)

	Angry	Happy	Neutral	Sad	Surprise
Angry	80.00	8.667	1.333	0.667	9.333
Happy	0.000	90.00	0.000	0.000	10.00
Neutral	1.765	0.000	79.41	7.059	11.76
Sad	1.333	0.000	3.333	83.34	12.00
Surprise	0.000	0.000	0.000	0.000	100.0

Confusion Matrix for the fusion of Speech & Images
(iii)

The framework was trained over the VIRI and RAVDESS datasets. The results were combined using a feature-level fusion and a decision-level fusion at different stages to achieve adequate accuracy. This section provides an overview of the results at every stage of the framework and compares individual and fused input modalities' accuracy. The framework accuracy is augmented with each layer of an integrated modality, and the trend is illustrated in the results obtained. The results are presented in the form of confusion matrices.