Audience Sentiment Analysis for Real-Time Adaptation

Brenner Vanessa Noemi
Department of Computer Science
Babeș-Bolyai University
Cluj-Napoca, Romania
vanessabrenner23@gmail.com

Silivăstru Oana Maria
Department of Computer Science
Babeș-Bolyai University
Cluj-Napoca, Romania
oana.silivastru@stub.ubbcluj.ro

Sima Alin
Department of Computer Science
Babeș-Bolyai University
Cluj-Napoca, Romania
sima@gmail.com

Pasăre Vladuț
Department of Computer Science
Babeș-Bolyai University
Cluj-Napoca, Romania
vlad@gmail.com

Audience sentiment analysis is crucial for enhancing human-computer interaction and live entertainment. This study introduces a real-time emotion detection system combining Faster R-CNN for robust face detection in low-light conditions, trained on the DARKFACE dataset, with MobileNetV2 and a modified VGG Network for emotion recognition using AffectNet and FER2013 datasets. MobileNetV2 outperforms VGG16. A fusion model integrating MobileNetV2 and a geometry-based CNN further enhances classification accuracy by combining global and local facial features. This work demonstrates the effectiveness of advanced models for real-time emotion analysis in dynamic environments, with future improvements aimed at broader emotion categories, occlusion handling, and adaptive feedback.

Index Terms—Audience sentiment analysis, real-time emotion detection, Faster R-CNN, MobileNetV2, VGG Network, DARKFACE, AffectNet, FER2013, face detection, convolutional neural networks, geometric features

1.INTRODUCTION

Detecting human emotions poses a significant challenge for artificial intelligence due to the extraordinary complexity of the human brain. The brain processes emotional stimuli quickly and intuitively, seamlessly integrating visual and auditory cues, social context, personal history, and physiological states. In contrast, AI systems struggle with interpreting subtle facial expressions, vocal tones, and the nuances of emotional context.

The branch of AI dedicated to emotion detection from images relies on visual analysis and advanced algorithms to interpret non-verbal communication. These systems use machine learning models to analyze facial expressions, posture, and gaze. Despite significant advancements in machine learning and image processing, AI still falls short of the sophistication demonstrated by the human brain in identifying and understanding emotions.

Real-time emotion detection systems bring substantial benefits to live entertainment, enabling performers to adapt their acts based on audience reactions. A notable example of this technology was the "Pay per Laugh" system implemented by TeatreNeu in Barcelona in 2014. By placing tablets in front of each audience member, the system monitored their smiles in real time throughout the performance. At the end, attendees paid based on the number of smiles detected, turning emotional reactions into a direct metric for evaluating the performance.

While innovative, this system focused exclusively on detecting smiles, one of the easiest emotions to recognize due to its well-defined and relatively consistent facial characteristics. This limitation highlights the technical challenges in 2014 of detecting more complex emotions like sadness, anger, or surprise, which require a more sophisticated interpretation of facial features and context (source 1, source 2).

Recent advancements, including those demonstrated in our study, have expanded this approach by utilizing advanced machine learning models capable of detecting a wider range of emotions. This opens new

possibilities for providing nuanced, real-time feedback, transforming the interaction between audience and performer into a dynamic and adaptive experience.

Our contributions include:

A. Utilizing Faster R-CNN for robust face detection in dark environments using the DARKFACE dataset.
B. Comparing VGG Network and MobileNetV2 for emotion detection, trained on AffectNet and FER2013 datasets.
C. Demonstrating the superiority of MobileNetV2 combined with AffectNet for real-world applications.

II .RELATED WORK

A. Face Detection

Faster R-CNN is an object detection method proposed by Ren et al. in their paper (source). This model builds upon earlier architectures like R-CNN and Fast R-CNN, addressing one of their major limitations: the slow and computationally expensive generation of region proposals.

1. Evolution of Faster R-CNN (source)

R-CNN, introduced by Girshick et al. [5], brought the precision of convolutional neural networks (CNNs) from image classification to object detection. It operates in two stages: generating region proposals using methods like **Selective Search** [14] and transforming each proposal into a fixed region (e.g., 227 × 227 for AlexNet [9]), followed by classification and bounding box refinement via a regressor.

However, R-CNN is inefficient because it requires passing each proposal through a CNN independently, leading to significant computational overhead.

Fast R-CNN and **SPPnet** [6] were introduced to mitigate this issue. These approaches process the entire image through a CNN only once, projecting the proposals onto the convolutional feature maps. Fast R-CNN further simplifies this process by using RoI pooling, enabling end-to-end fine-tuning of a model pre-trained on ImageNet. While this results in improved performance over R-CNN, both R-CNN and Fast R-CNN still rely on manually generated region proposals (e.g., Selective Search or EdgeBox [2]), which are time-consuming and less efficient compared to deep learning-based methods.

Faster R-CNN was developed to eliminate this bottleneck by integrating a Region Proposal Network (RPN) directly into its architecture. RPN is a fully convolutional network that generates object proposals and passes them to the Fast R-CNN detector for refinement.

Both the **RPN** and **Fast R-CNN** share convolutional layers, allowing the entire image to be processed through the CNN in a single pass. This shared architecture enables the use of very deep networks, such as VGG16 [13], to generate high-quality region proposals.

Tests on the **FDDB** dataset [7] demonstrated that Faster R-CNN is significantly faster than both R-CNN and Fast R-CNN, using an NVIDIA Tesla K40c GPU and an Intel Xeon E5-2697 CPU.

Figure 1 Sample detection results on the FDDB dataset, where green bounding boxes are ground-truth annotations and red bounding boxes are detection results of the Faster R-CNN

2. Why Faster R-CNN?

Faster R-CNN has demonstrated superior performance in capturing relevant features from images with complex and dynamic scenes, including the detection of fine details such as subtle facial expressions. This makes it ideal for applications like real-time sentiment analysis, where detection accuracy plays a critical role.

In this study (source), Faster R-CNN, with specific modifications, outperformed YOLO_V3 by 6.6% mAP (source) compared to the standard version of Faster R-CNN. This highlights its potential for adaptation and optimization in specialized domains. Key enhancements included replacing the VGG16 backbone with ResNet101, adjusting anchor box dimensions using the K-means clustering algorithm, optimizing the Non-Maximum Suppression (NMS) algorithm (to reduce false positives and improve detection efficiency), and integrating candidate box positions with inherent text data.

With these recent improvements, Faster R-CNN has proven to be an exceptional model for applications requiring fine-grained detection and precision, such as audience sentiment analysis. Its enhanced ability to capture subtle details in complex and dynamic images makes it a top choice for real-time facial expression analysis, where accuracy and efficiency in detecting nuances are essential for accurately interpreting audience emotions.

Faster R-CNN and Face Detection in Low Light

In the research literature, Faster R-CNN has been successfully applied to face detection in images, including scenarios with low-light conditions.

The study (source) introduces a face detector based on Faster R-CNN, called DSFD (Different Scales Face Detector). This detector improves real-time face detection accuracy by addressing challenges related to the variability in face sizes across images. To achieve precise regions of interest (ROIs) for faces, DSFD employs an efficient multitask Region Proposal Network (RPN) that generates inhomogeneous anchors on higher-level feature maps.

Additionally, the paper (source) explores the application of Faster R-CNN to face detection, showcasing state-of-the-art results on datasets such as WIDER (source), FDDB (source), and IJB-A (source). These studies underline the effectiveness of Faster R-CNN in detecting faces under diverse and challenging conditions.

3. DARKFACE dataset for face detection

DARKFACE (source) is a dataset designed to improve facial detection under low-light conditions. It contains approximately 6,000 images captured in nocturnal environments, such as streets, parks, and buildings, annotated with bounding boxes for faces.

The primary purpose of DARKFACE is to support the development of algorithms capable of detecting faces in poorly lit settings, a significant challenge in facial recognition tasks. This dataset is highly beneficial for researchers aiming to build and evaluate facial detection systems that perform effectively in low-light environments.

Figure 2 Image samples from DARKFACE

B. Emotion Detection

1. VGG-networks style (source)

VGG Networks (Visual Geometry Group) are a family of convolutional neural network (CNN) architectures proposed by researchers from Oxford in 2014. These networks gained popularity for their simplicity and efficiency, relying solely on 3×3 convolutional layers and 2×2 max-pooling layers. This straightforward design makes VGG networks easy to implement and highly effective for image recognition tasks.

VGG is characterized by its deep architecture, utilizing a large number of layers (16–19) while maintaining a relatively simple structure. It does not employ advanced regularization techniques, such as dropout or batch normalization, yet it has achieved excellent performance in image recognition competitions like **ImageNet** (source).

There are several variants of the VGG architecture, including **VGG-16** and **VGG-19**, which differ only in the number of convolutional and fully connected (FC) layers. These architectures remain highly popular in deep learning and are frequently used as backbones for many transfer learning models.

2. MobileNetV2 (source)

MobileNetV2 is a convolutional neural network (CNN) architecture developed by Google, specifically designed to be resource-efficient, making it ideal for mobile devices and other resource-constrained applications, such as edge computing.

Part of the MobileNet family, MobileNetV2 enhances its predecessors by introducing a novel block type called **inverted residual block**. These blocks are computationally efficient as they allow for better separation between convolution and activation operations. Additionally, MobileNetV2 employs **depthwise separable convolutions**, which significantly reduce the number of parameters and computations without compromising network performance.

Key features of MobileNetV2 include:

- **Inverted Residuals**: Enable greater efficiency in reducing dimensionality while improving accuracy.

- **Depthwise Separable Convolutions**: Reduce the number of parameters and computational costs.
- **Linear Bottleneck**: Utilizes a linear connection between activation blocks, simplifying operations and enhancing efficiency.

MobileNetV2 is widely used in applications requiring high processing speeds and computational efficiency, such as object recognition in images, face detection, and real-time video analysis on mobile or embedded devices.

3. *AffectNet and FER2013 datasets*

a. *AffectNet (source)*

AffectNet is an extensive dataset containing approximately 1 million images labeled with eight primary emotions: happiness, sadness, fear, anger, disgust, surprise, contempt, and neutral. Sourced from online platforms, the dataset includes high-resolution images (typically 500×500 pixels or larger), capturing a wide range of facial expression. AffectNet is widely used in research and development for facial emotion recognition models, and it is particularly favored for applications in areas like behavior analysis and human-computer interaction. The diversity and complexity of the dataset enable models to learn detailed facial features and recognize emotions under various lighting conditions and environmental settings..

Figure 3 Image samples from AffectNet.

b. *FER2013 (source)*

FER2013 is another dataset used for facial emotion recognition, but with a smaller number of images—approximately 35,000. This dataset includes seven primary emotions (excluding contempt) and consists of smaller grayscale images with dimensions of 48×48 pixels. The images in FER2013 are sourced from standardized datasets, making it particularly useful for quick research and algorithm evaluation in emotion recognition tasks. Due to its compact size, FER2013 is often employed for benchmarking purposes and academic studies, allowing for the testing of machine learning models on simpler and faster-to-process datasets.

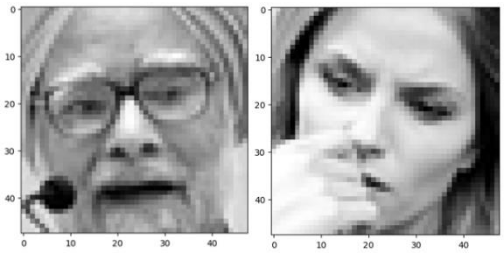

Figure 4 Image samples from FER2013.

c. *Differences between AffectNet and FER2013*

AffectNet and **FER-2013** are two widely used datasets for facial emotion recognition, but they differ significantly in size, complexity, and features. AffectNet is substantially larger, containing approximately 1 million images, whereas FER2013 consists of only 35,000 images. AffectNet includes annotations for **eight emotions** (including contempt), while FER2013 covers only **seven emotions** (excluding contempt).The image dimensions in AffectNet are larger (typically 500×500 pixels), enabling more detailed facial feature capture, compared to the smaller, standardized 48×48 grayscale images in FER2013. Additionally, AffectNet offers greater diversity in image sources and lighting conditions, making it more complex and realistic. In contrast, FER2013 is more standardized and suited for quick benchmarking and controlled experiments.

4. *Differences Between MobileNetV2 and VGG Network on the Proposed Datasets*

The study referenced (source) evaluates the performance of a modified **VGG16** architecture and **MobileNetV2** for emotion recognition using the FER2013 and AffectNet datasets. The modified VGG16 incorporates additional depth, an extra fully connected layer, more neurons, dropout regularization, and a learning rate scheduler.

Experimental results highlight key distinctions between the two models. MobileNetV2, designed for efficiency, employs depthwise separable convolutions and inverted residual blocks, making it better suited for resource-constrained environments. On the other hand, the VGG16 architecture, with its depth and large number of parameters, excels in extracting hierarchical features but is computationally expensive.

Figures *5* and *6* illustrate the comparative performance of these models on *FER2013* and *AffectNet*.

Model	Accuracy (%)	Precision (%)
CNN	66.20	66.27
Modified VGG16	67.43	67.60

Figure 5 Accuracy and precision of CNN and Modified VGG16 models with the FER2013 Dataset.

Model	Accuracy (%)	Precision (%)
CNN	41.43	51.34
Modified VGG16	42.86	56.70

Figure 6 Accuracy and precision of CNN and Modified VGG16 models with the AffectNet Dataset.

In the article (source), the authors developed a **real-time facial expression recognition system** using two CNN architectures: **MobileNetV2** and a custom-designed CNN. Both architectures were selected for their efficiency in real-time applications, aiming to evaluate and compare their performance in emotion recognition tasks.

MobileNetV2 was chosen for its resource efficiency, particularly for mobile devices and low-power environments. The custom CNN architecture was tailored to provide an alternative specifically optimized for facial expression recognition. The experiments demonstrated that both models are well-suited for real-time applications, offering high accuracy and optimal processing times, making them viable for practical implementations.

For our case study, we will display only the results obtained in the aforementioned article, as shown in **Figure 7**.

Dataset	***Accuracy%***
AffectNet	*94.2%*
FER2013	*77.8%*

Figure 7 Accuracy of FER MobileNetV2 Model.

The results compared across the two studies clearly demonstrate that **MobileNetV2 significantly outperforms the VGG16 architecture** in emotion recognition on both datasets. On **AffectNet**, MobileNetV2 achieves an accuracy of **94.2%**, compared to just **42.86%** for VGG16. Similarly, on the **FER2013** dataset, MobileNetV2 attains an accuracy of **77.8%**, surpassing the **67.43%** achieved by VGG16. This highlights the **superiority of MobileNetV2** in diverse scenarios, thanks to its better performance and **computational efficiency**, making it a more suitable choice for resource-constrained environments and practical applications.

III. METHODOLOGY

A. Face Detection

1. Model Selection

Faster R-CNN was chosen for this project due to its high precision and ability to detect objects in challenging conditions, such as low-light environments. Its integrated Regional Proposal Network (RPN) efficiently identifies potential facial regions, making it highly suitable for real-time applications. Unlike traditional face detection methods, Faster R-CNN excels in scenarios where other models may struggle, such as detecting partially obscured faces or faces with low contrast in poorly lit settings. This capability aligns perfectly with our goal of analyzing audience expressions in theater conditions, where lighting is often dim.

To adapt Faster R-CNN for face detection in low-light conditions, the model was trained and fine-tuned using the DARKFACE dataset. DARKFACE is a specialized dataset designed for object detection in nighttime scenarios. It includes a wide variety of images captured in real-world low-light environments, with variations in light intensity, noise, and face visibility.

2. Preprocessing data

To prepare the DARKFACE dataset for training, a series of preprocessing steps were applied to enhance robustness.

Brightness adjustments modified pixel intensity using

$$Ibright = I \cdot \delta, \delta \in [1 - \Delta b, 1 + \Delta b]$$

where we represents the input image and Δb controls brightness variation. Contrast was enhanced using:

$$Icontrast = \frac{I - \mu}{\sigma}\delta c, \delta c \in [1 - \Delta c, 1 + \Delta c]$$

where μ and σ denote the mean and standard deviation of pixel intensities, and Δc determines contrast variation. Hue and saturation modifications were

applied in the HSV color space to increase color diversity.

Additionaly, all images were resized to 640×640 pixels for consistency. Random Horizontal Flipping $p_{flip} = 0.5$ and anchor-based sampling were applied to enhance spatial diversity and address varying object scales.

3. Implementation Details

The Faster R-CNN architecture was employed with a ResNet-50 backbone pre-trained on ImageNet. To enhance object detection across varying scales, the architecture was extended with a Feature Pyramid Network (FPN), which aggregates multi-scale feature maps. This integration improved the detection of small objects, a common challenge in object recognition tasks.

The Region Proposal Network (RPN) was customized to generate region proposals with anchors tailored to the dataset. Anchor dimensions and aspect ratios were adjusted to better match the size distribution of objects within the data. Non-Maximum Suppression (NMS) was used in the RPN to refine region proposals by removing redundant overlaps.

For the Region of Interest (RoI) pooling stage, RoIAlign was employed to ensure precise spatial alignment of features. This technique mitigates the inaccuracies introduced by quantization in traditional RoI pooling. The classifier head and bounding box regressor utilized fully connected layers to output object class probabilities and bounding box refinements, respectively.

Quantitative results and performance comparisons will be detailed in the subsequent section.

B. Emotion Detection

1. Model Comparison

In this study, we compared two convolutional neural network (CNN) architectures: VGG Network and MobileNetV2, each offering distinct advantages depending on the requirements of our emotion detection application. Both models were optimized for emotion detection, and their performance comparison across the two datasets revealed significant differences in terms of accuracy and computational efficiency, depending on the type and complexity of the input data.

Preprocessing data

The **FER2013** dataset comprises **48×48 pixel grayscale facial images**, with faces automatically aligned and centered to ensure uniform positioning. Each image is categorized into one of **seven emotion classes (0=Angry, 1=Disgust, 2=Fear, 3=Happy, 4=Sad, 5=Surprise, 6=Neutral)**. The dataset was divided into training and validation subsets, with 80% of the data used for training and 20% reserved for validation.

All images were normalized to the $[0,1]$ range and resized to $48x48$ pixels, ensuring consistency and numerical stability during model training. Random horizontal flipping $p_{flip} = 0.5$, rotations$[-15 \circ, +15 \circ]$ and zooming were employed to simulate real-world variability.

AffectNet is a large-scale facial expression dataset containing approximately 400,000 images manually labeled for eight emotion categories: neutral, happy, angry, sad, fear, surprise, disgust, and contempt. The dataset also includes annotations for valence and arousal intensity, allowing for affective analysis.

For computational efficiency and to address memory constraints, the images were resized to a uniform resolution of 96×96 pixels. This preprocessing ensures consistency and compatibility with model training pipelines. The dataset was split into 70% (280,000 images) for training, 15% (60,000 images) for validation, and 15% (60,000 images) for testing., maintaining a balance between training and evaluation.

These preprocessing steps for the datasets were applied to both models.

2. Proposed algorithms

2.1. VGG-network style

The **VGG network** is a convolutional neural network architecture introduced by Simonyan and Zisserman in 2014, known for its simplicity and efficiency in feature extraction. It replaces large convolutional filters (e.g.,7x7) with stacks of 3x3 filters, reducing parameters while maintaining a larger effective receptive field.

The architecture is composed of repeated blocks of 3x3 convolutions with a stride of 1 and padding to preserve spatial dimensions. Each block is followed by

2x2 max-pooling layer with stride 2, halving the spatial dimensions. This hierarchical design enables the extraction of features from simple edges to complex patterns as depth increases.

VGG networks, such as VGG-16 and VGG-19, are named based on their depth. For example, VGG-16 includes 13 convolutional and 3 fully connected layers. After each convolution, a ReLU activation function introduces non-linearity and mitigates vanishing gradients. The final softmax layer outputs class probabilities:

$$p(y_i|x) = \frac{e^{z_i}}{\sum_j e^{z_j}}$$

where z_i represents the logits for class *i.*

2.2. MobileNetV2

This model introduces a highly efficient convolutional neural network architecture optimized for resource-constrained environments. The key innovations are the incorporation of depthwise separable convolutions, linear bottlenecks, and inverted residual blocks, which significantly reduce computational costs while maintaining competitive accuracy.

The model begins with a standard $3x3$ with a stride of $s = 2$, reducing the spatial dimensions of the input image and projecting the basic features into a higher-dimensional space. This initial step enables the model to extract primary features such as contours and textures. Inverted residual blocks form the core of the architecture. Each block consists of three main components: the expansion layer, which performs $1x1$ to increase the depth of the input tensor by an expansion factor *t* (default $t = 6$).

The expansion layer uses the ReLU6 activation, chosen to prevent saturation under low-precision computation conditions. ReLU6 keeps values within the range [0,6], ensuring numerical stability and minimizing the risk of information loss on resource-constrained devices. The computational cost formula for this layer is: $h \cdot w \cdot di \cdot t$, where h and w are the spatial dimensions of the tensor, di represents the initial depth, iar t is the expansion factor. **The Depthwise Convolution** applies a kxk filter independently to each channel, enabling the extraction of spatial. This method significantly reduces computational complexity compared to standard convolutions.

The computational cost for this layer is: $h \cdot w \cdot di \cdot k^2$, where k is the kernel size of the convolution. The **projection block** concludes each residual unit with a 1x1 convolution without activation, reducing the expanded tensor's dimensions back to the initial size. Removing non-linear activations, such as ReLU, prevents irreversible information loss and preserves the manifold structure. **Inverted residual connections** directly link the projection layer to the block input, bypassing the expansion layer. This strategy optimizes gradient flow during training, reduces memory requirements, and improves information propagation between blocks. The architecture ends with a **global pooling layer**, which aggregates global information by compressing the spatial dimensions to 1x1. The **softmax activation** in the output layer transforms these values into probabilities, enabling efficient and accurate classification.

A. Geometric Expression Recognition

The CNN model based on geometric features analyzes the movements of key facial landmarks. A total of 68 landmarks are utilized, with the essential ones for facial expressions (eyebrows, eyes, mouth) being selected. The differences between the coordinates of landmarks in neutral and emotional images are represented as a geometric vector:

$$Ve = [x_0^e - x_0^n, y_0^e - y_0^n, \dots, x_{35}^e - x_{35}^n, y_{35}^e - y_{35}^n]$$

The CNN architecture consists of convolutional layers (e.g., 3x1) and max-pooling operations (2×1), which compress and transform the features. The convolutional layers gradually increase the depth of the features, reaching up to 256 feature maps. At the end, the features are flattened and connected to a fully connected layer with 500 nodes, followed by a softmax layer that classifies the emotions. The ReLU activation function and dropout are employed to prevent overfitting.

B. Fusioning the models

MobileNetV2 excels at extracting global features from the entire image, such as contours and textures, but may struggle to capture fine details or subtle facial movements. The geometry-based network complements this limitation by focusing on the dynamic changes in facial landmarks that reflect emotions.

Thus, MobileNetV2 processes the full facial image and extracts a global feature vector (V_{global}) with dimension $d1$.

Geometric Expression Recognition generates a geometric vector V_{geo} based on the differences in

landmark points (V_e) between neutral and emotional facial images, with a dimension of *d2*.

The extracted vectors are concatenated into a single fused vector: $V_{fused} = [V_{global}; V_{geo}]$, V_{fused} has a dimension of *d1* + *d2*.

IV. TRAINING, RESULTS AND DISCUSSION

1. Faster R-CNN

Loss Function: consists of two components: Region Proposal Network (RPN) loss and detection head loss:

$$L_{RPN} = L_{cls}(p, p^*) + \lambda L_{reg}(t, t^*)$$

where:

- L_{cls} is the binary cross-entropy loss between predicted objectness scores p and ground thruth p^*.
- L_{reg} is the smooth L1 loss for bounding box regression, calculated only for positive anchors.
- λ balances the two terms.

$$L_{det} = L_{cls}(p_k, p_k^*) + \lambda L_{reg}(t_k, t_k^*)$$

where:

- L_{cls} is the cross-entropy loss for multi-class classification
- L_{reg} is the smooth L1 loss for bounding box adjustments.
- p_k and t_k represent predicted class probabilities and bounding box coordinates for class *k*

Total Loss:

$$L = L_{RPN} + L_{det}$$

Adam Optimizer: The model was trained using the Adam optimizer with a learning rate of 0.001, decayed by 0.1 after 10 epochs without improvement, *β1=0.9*, *β2=0.999*, and a weight decay of 10^{-4}. Adam's adaptive learning rates ensured stable convergence across classification and regression tasks.

Metrics: We evaluated performance using **accuracy** for classification and **Intersection over Union (IoU)** to measure the overlap between predicted and ground truth bounding boxes for localization.

Epoch Start	Epoch End	DarkFace Accuracy	IoU
1	12	40.25%	0.42
13	25	69.16%	0.63
26	38	85.77%	0.79
38	50	88.95%	0.84

Figure 8 Training and Validation Metrics

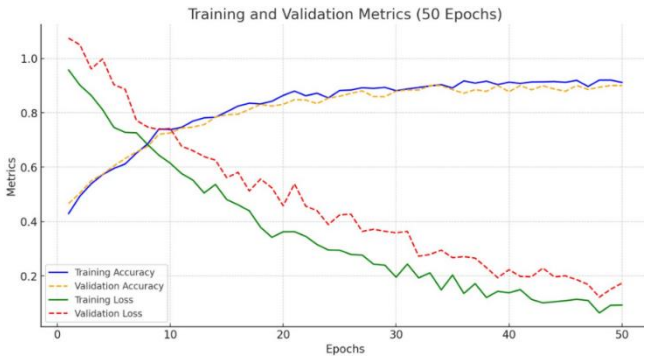

Figure 9 Training and Validation Metrics

The model demonstrates significant improvement in both accuracy and IoU across training epochs, indicating successful learning and optimization. The DarkFace accuracy increases from 40.25% to 88.95%, while IoU rises from 0.42 to 0.84, reflecting enhanced object detection precision and localization performance.

Initially, we experimented with VGG-16 and MobileNetV2 architectures for emotion recognition using the FER2013 and AffectNet datasets. Both models were fine-tuned with pre-trained ImageNet weights, using a learning rate of 0.001, a batch size of 32, and Adam optimizer. While these architectures provided reasonable results, they did not fully meet performance expectations, as reflected in their accuracy values (FER2013: 69% and 72.3%, AffectNet: 52% and 88.3%, respectively).

Accuracy	Fer2013	AffectNet
VGGStyle CNN	69%	52%
MobileNetV2	72.3%	88.3%

Figure 10 Accuracy of VGG and MobileNet on Fer2013 and AffectNet dataset

VGG-16, known for its deep architecture with 16 layers, excels at capturing hierarchical features but is computationally expensive and lacks efficiency for lightweight applications.

MobileNetV2 introduces depthwise separable convolutions, linear bottlenecks, and inverted residual blocks, which reduce computational costs while maintaining competitive accuracy, making it ideal for resource-constrained environments. However, both

models faced challenges in capturing subtle facial dynamics essential for emotion recognition.

To address these challenges, we trained the fused model described. This hybrid approach leverages the strengths of both networks: MobileNetV2 extracts global features like contours and textures, while the geometric model focuses on subtle, localized changes in facial landmarks. The resulting fused vector integrates these complementary features, achieving improved performance and accuracy across both datasets.

2. MobileNetV2 fusioned with Geometric Expression Recognition CNN

Loss Function: Categorical Crossentropy

Categorical Crossentropy is widely used for multi-class classification tasks where the output is a probability distribution across multiple classes. It measures the difference between the true class distribution and the predicted probability distribution output by the model. The mathematical formulation is:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} y_{ij}\ log(\hat{y}_{ij})$$

where:

- N is the number of samples,
- C is the number of classes,
- y_{ij} is a binary indicator (0 or 1) representing whether class j is the correct class for sample iii,
- $\hat{y}_{ij}$ is the predicted probability for class j of sample i.

This function ensures that the model maximizes the probability of the correct class while penalizing incorrect predictions. It works effectively with softmax activation in the output layer, which normalizes the predicted logits into probabilities.

Adam Optimizer: The **Adam** optimizer (Adaptive Moment Estimation) combines the advantages of two popular optimization methods: **Momentum** and **RMSprop**.

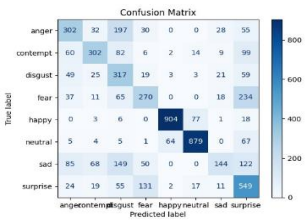

Figure 13 Confusion Matrix on Fer2013

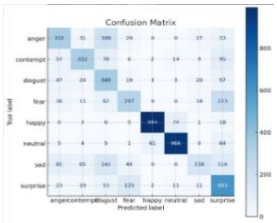

Figure 12 Confusion Matrix on AffectNet

Metrics: We used accuracy as the primary evaluation metric, calculated as the ratio of correct predictions to the total number of predictions. This metric is widely recognized for classification tasks, enabling straightforward comparison with other models. Additionally, the confusion matrix provided detailed insights into class-specific performance, highlighting areas where the model required improvement.

Results:

Our model was trained for 50 epochs. Training time for an epoch was approximately 10 minutes on an L40S GPU Tensor Core with 48 GB VRAM and 128 GB RAM.

Epoch Start	Epoch End	Fer Accuracy	AffectNet Accuracy
1	25	50.40%	68.20%
26	50	65.29%	75.14%
51	75	72.81%	85.36%
76	100	77.46%	90.27%

The comparison shows FER starts at a lower accuracy (50.40%) and reaches 77.46% by epoch 100, reflecting its simpler structure. AffectNet starts higher at 68.20% and achieves 90.27%, demonstrating its richer diversity and better generalization potential despite greater complexity.

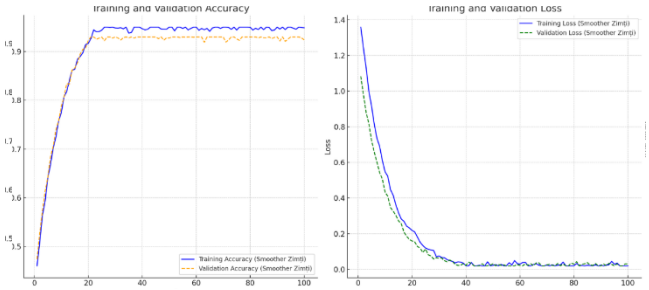

Figure 11 Training and validation accuracy and loss AffectNet

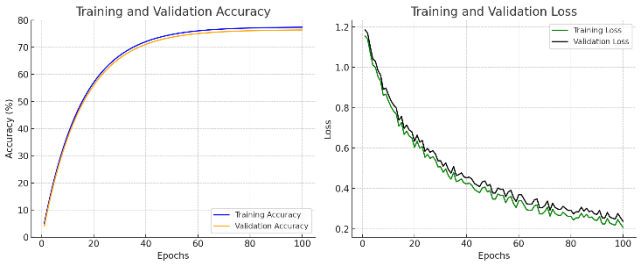

Figure 12 Training and validation accuracy and loss Fer2013

From the images 11 and 12, AffectNet shows slower convergence and lower peak accuracy compared to FER, reflecting its greater complexity. The loss on AffectNet is more fluctuating and higher throughout, indicating the model struggles more with its features. FER converges faster with smoother loss and higher accuracy, suggesting it is easier for the model to handle.

In the case of FER (Figure 13), similar trends are observed, with high accuracy for "happy" and "neutral" labels. However, the diagonal values are generally lower compared to AffectNet, indicating slightly weaker overall performance. Additionally, the model exhibits more frequent misclassifications, such as confusing "anger" with "disgust" or "fear." The class "surprise" appears particularly challenging, suggesting a less balanced performance across categories.

For AffectNet (Figure 14), the model performs better on dominant classes such as "happy" and "neutral," which is evident from the higher diagonal values associated with these labels. However, there are noticeable confusions between closely related emotions, such as "fear" and "sad," or "anger" and "disgust," indicating challenges in distinguishing subtle emotional nuances. The overall distribution reflects the complexity of the AffectNet dataset, which may introduce a bias toward certain dominant classes.

VI. CONCLUSION AND FUTURE WORK

Our study successfully implemented a **real-time audience sentiment analysis system** that recognizes 7 primary emotions (anger, disgust, fear, happiness, sadness, surprise, neutral) and provides **feedback on audience emotions**. Faster R-CNN achieved reliable face detection in low-light conditions with 88.95% accuracy and an IoU of 0.84 on the DARKFACE dataset. For emotion recognition, MobileNetV2 outperformed VGG-16 on AffectNet (90.27%) and FER2013 (77.46%). The integration of MobileNetV2 with a geometric CNN captured both global and localized facial features, enabling accurate and dynamic emotion recognition in real-world scenarios.

Future improvements will target better **occlusion handling**, optimizing for **low-light** and extending the emotion detection system to include **subtle and complex emotions**. Additionally, we aim to enhance the **real-time feedback mechanism** for seamless use in live applications, enabling performers or systems to adapt dynamically to audience emotions.

REFERENCES

[1] Sudhakar Mishra and Uma Shanker Tiwary. "*A Cognition-Affect Integrated Model of Emotion.*" Indian Institute of Information Technology, Allahabad, Center for Cognitive Computing, Allahabad, 211012, India. (2019)

[2] Siddique Latif, Hafiz Shehbaz Ali, Muhammad Usama, Rajib Rana1, Björn Schuller, and Junaid Qadir. "*AI-Based Emotion Recognition: Promise, Peril, and Prescriptions for Prosocial Path*" (2022)

[3] Becker, D. V., Anderson, U. S., Mortensen, C. R., Neufeld, S. L., & Neel, R. "The face in the crowd effect unconfounded: Happy faces, not angry faces, are more efficiently detected in single- and multiple-target visual search tasks". *Journal of Experimental Psychology: General, 140*, 637–659 (2011)

[4] Maria Avgitidis. "*How to Easily Read Faces and Facial Expressions*" (2024)

[5] Ali Sharifara, Mohd Shafry Mohd Rahim, Yasaman Anisi. "*A general review of human face detection including a study of neural networks and Haar feature-based cascade classifier in face detection*". International Symposium on Biometrics and Security Technologies (2014)

[6] Ashu Kumar, Amandeep Kaur, Munish Kumar3 "*Face detection techniques: a review*" Springer Nature B.V. (2018)

[7] Jian Li, Yabiao Wang, Changan Wang, Ying Tai, Jianjun Qian, Jian Yang, Chengjie Wang, Jilin Li, Feiyue Huang. "*DSFD: Dual Shot Face Detector*". Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019)

[8] Ji-hae Kim, Byung-gyu Kim, Partha Pratim Roy, Da-Mi Jeong. "*Efficient Facial Expression Recognition Algorithm Based on Hierarchical Deep Neural Network Structure*" (2019).

[9] Min Wang, Baoyuan Liu, and Hassan Foroosh. "Design of efficient convolutional layers using single intra-channel convolution, topological subdivisioning and spatial 'bottleneck' structure." CoRR, abs/1608.04337, 2016.

[10] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V. Le. "Learning transferable architectures for scalable image recognition." CoRR, abs/1707.07012, 2017.

[11] Lingxi Xie and Alan L. Yuille. "Genetic CNN." CoRR, abs/1703.01513, 2017.

[12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." In Bartlett et al. [47], pages 1106–1114.

[13] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. "MobileNetV2: Inverted Residuals and Linear Bottlenecks." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[14] Lalita Mishra, Shekhar Verma, and Shirshu Varma. "Hybrid Model using Feature Extraction and Non-linear SVM for Brain Tumor Classification." *arXiv preprint arXiv:2212.02794*, 2022.

[15] Ali Mollahosseini, David Chan, and Mohammad H. Mahoor. "AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild."

[16] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[17] Ross Girshick. "Fast R-CNN." *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.