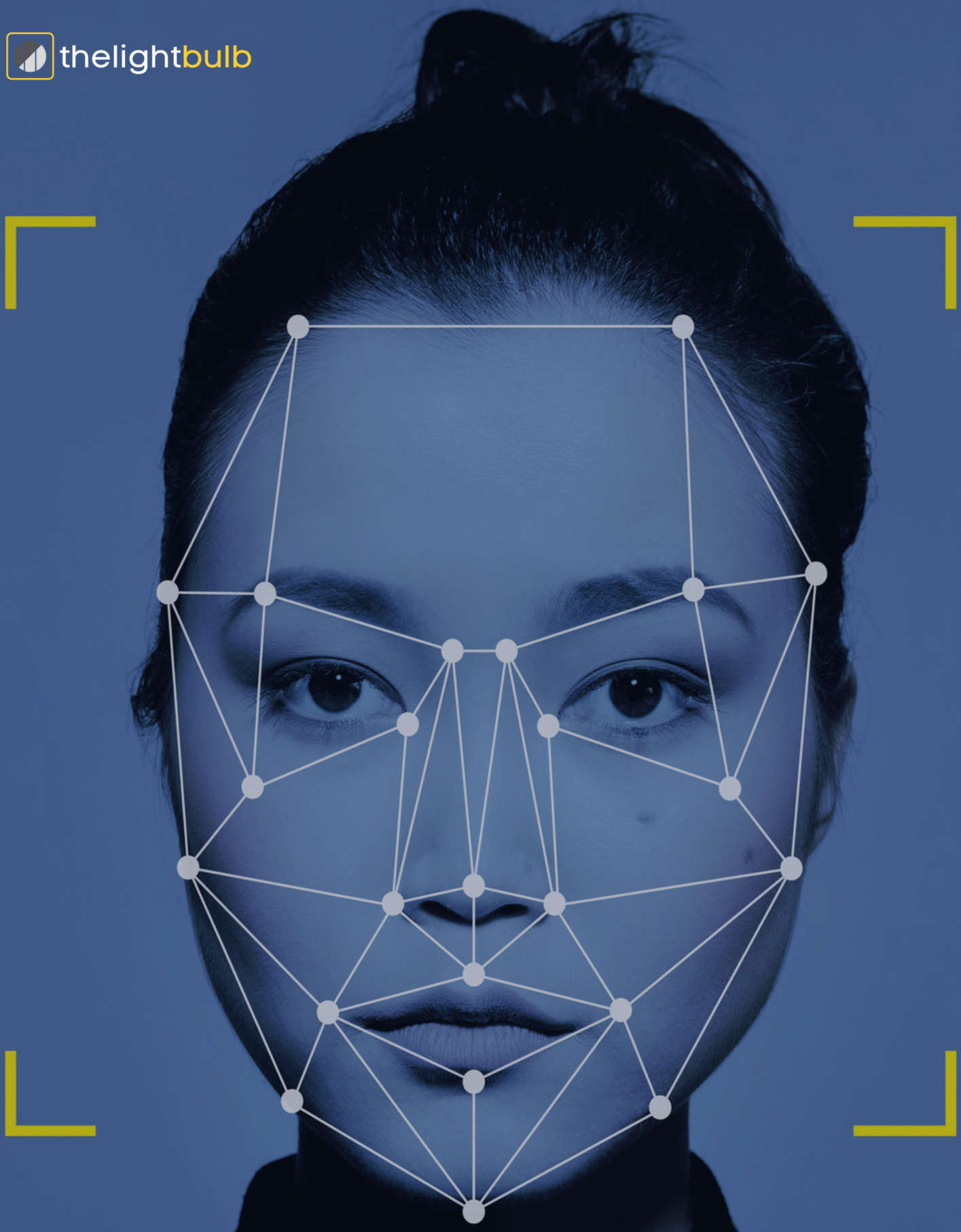
thelightbulb



Facial Coding

Accuracy Document

thelightbulb

Table of **Contents**

Table of **Contents (Contd.)**

1. Introduction

Computer vision is a field of study that enables computers to extract valuable information from digital images and videos. It involves developing algorithms to interpret visual data, mimicking human perception.
One vital application is face detection, which identifies and localizes human faces within images or videos by analyzing visual features like edges and patterns. Face detection is used in various domains like security systems, surveillance, biometrics, and photography.

Emotion detection, a subfield of computer vision, analyzes facial expressions to recognize and interpret emotions using the Facial Action Coding System (FACS). FACS assigns specific action units to facial muscle groups, allowing the inference of underlying emotions. Emotion detection with FACS has applications in market research, psychology, human-computer interaction, and entertainment.

2. Background

2.1 Machine Learning

Machine learning applied to facial emotion classification is an intriguing field that harnesses advanced algorithms and techniques to analyze and interpret human emotions from facial expressions. On training models on extensive datasets of labeled facial images, machine learning algorithms acquire the ability to detect and classify emotions like happiness, sadness, anger, fear, and surprise, leveraging distinct patterns and features in each expression.

This technology boasts diverse applications, spanning improved human-computer interaction, augmented emotional intelligence in artificial systems, and domains like healthcare, customer service, and entertainment. Now, if you unlock the capacity to comprehend and respond to human emotions, machine learning for facial emotion classification paves the way for more empathetic and intuitive interactions between humans and machines.

2.2 Convolutional Neural Networks

Convolutional neural networks (CNNs) are powerful tools for facial emotion classification, extracting meaningful features from images automatically. They excel at analyzing facial expressions. In this context, CNNs operate on grayscale or color face images, identifying patterns and structures associated with different emotions.

A CNN architecture consists of convolutional, pooling, and fully connected layers. Convolutional layers employ filters to capture local features, while pooling layers down sample feature maps, preserving important information. Fully connected layers perform classification based on extracted high-level features.

During training, a labeled dataset of facial images is used to adjust the CNN's parameters through techniques like backpropagation and stochastic gradient descent, minimizing the difference between predicted and actual emotion labels.

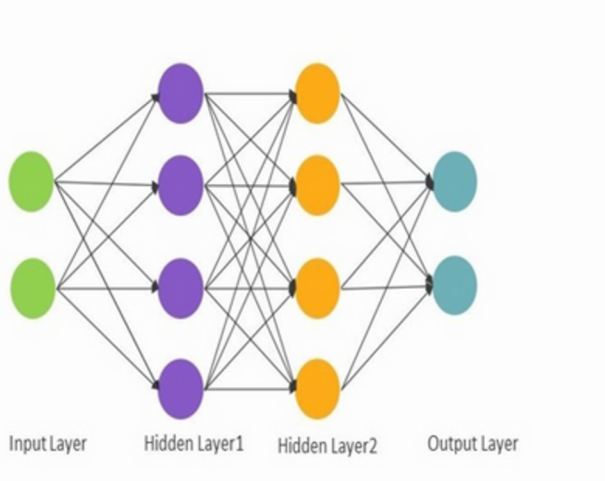


Fig. 2.1 Simple neural network

Once trained, the CNN can classify facial emotions by processing an input image and producing a probability distribution over different emotion categories. Higher probabilities indicate recognized emotions, enabling accurate classification. CNNs have achieved impressive results in facial emotion classification, setting new benchmarks. Their ability to learn discriminative features from facial images makes them valuable tools for understanding human emotions through visual cues.

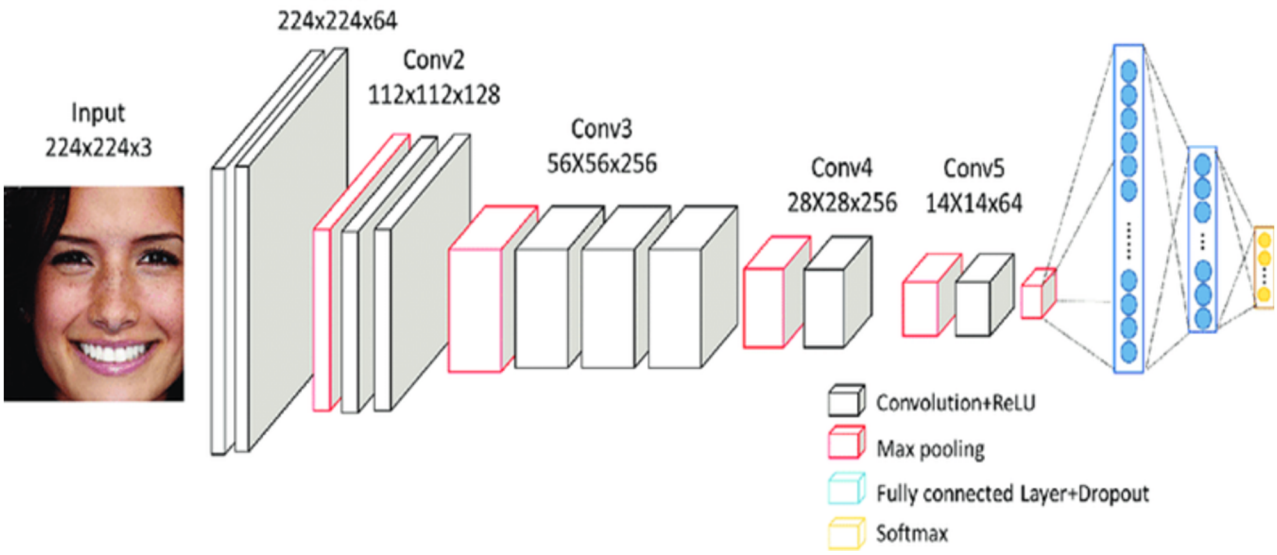


Fig. 2.2 Function of a CNN for image classification.

2.3 Facial Emotion Analysis

Facial emotion recognition plays a vital role in customer behavioral research, offering valuable insights into emotional responses and preferences. By analyzing customers' facial expressions, researchers can shape marketing strategies and enhance the overall customer experience. It finds applications in Advertisement Testing, Product Testing, UX Testing, Customer Satisfaction Analysis, and Market Research. This non-intrusive and objective approach provides real-time data, enabling businesses to create personalized experiences and make informed decisions in their marketing and customer engagement strategies.

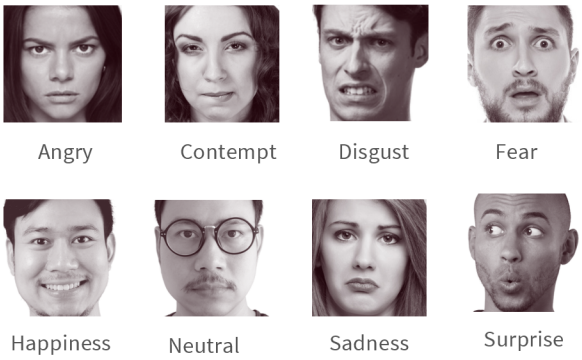


Fig. 2.3 Faces representing various emotions

2.4 Facial Expression Recognition (FER) Dataset

The FER2013 dataset is a widely used dataset for facial emotion recognition. It contains grayscale face images with emotion labels, widely used for facial emotion recognition. Here is the metadata information for the FER2013 dataset:

Dataset Name: FER2013 Source: The dataset was created by Pierre-Luc Carrier and Aaron Courville at the University of Montreal.

Data format: The dataset is in CSV format, with each row representing an image and having two columns: The "emotion" column holds the emotion label (0-6), while the "pixels" column stores the image's pixel values as a space-separated string.

2.5 Traditional Vs TheLightBulb's approach

Traditional emotional analysis algorithms are limited by CNN models. Our network combines base models (VGGNET, Xception, etc.) with additional CNN and dense layers, optimizing performance and enabling excellent generalization and real-time data processing.

	Traditional Approach	LB Approach
Transfer learning	No	Yes
Datasets for validation	Limited	Multiple
Generalization on new data	Limited	Good

Table 2.1 Traditional Vs TheLightBulb's approach

3. Implementation

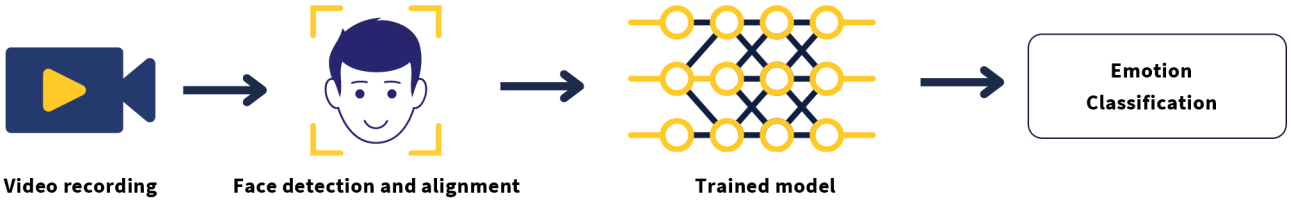


Fig. 3.1 An overview of the implementation of Facial Encoding product:
1. Data acquisition 2. Data-preprocessing 3. Model Prediction 4. Classification.

3.1 Data Pre-processing

- **Data Preprocessing with RetinaFace**

Data preprocessing in facial emotion analysis can be enhanced by employing RetinaFace, a state-of-the-art (SOTA) face detection and alignment algorithm. RetinaFace provides robust and accurate face detection and landmark localization, making it a preferred choice for modern facial emotion analysis pipelines.

- **Why RetinaFace is SOTA**

RetinaFace surpasses other algorithms, including MTCNN, in several ways:

1. High Precision: RetinaFace achieves higher accuracy in detecting small faces and faces in challenging conditions such as extreme poses, occlusions, and low-light scenarios.

2. Single-Stage Detector: Unlike MTCNN's multi-stage pipeline, RetinaFace is a single-stage detector, making it faster and more efficient.

3. Dense Regression: RetinaFace employs dense regression for five facial landmarks, providing precise face alignment critical for downstream tasks like emotion recognition.

4. Deep Supervision: It incorporates auxiliary supervision (e.g., pixel-wise face annotations) to enhance its detection capabilities.

5. Robust Backbone: RetinaFace uses powerful backbones such as ResNet and MobileNet, which are pre-trained on large-scale datasets, contributing to superior generalization.

**Source: CMC-COMPUTERS MATERIALS& CONTINUA, vol. 67.

- **RetinaFace-Based Preprocessing Pipeline**

1. Face Detection:

- RetinaFace detects bounding boxes for faces in the input image.
- The algorithm is particularly effective in complex scenarios, handling variations in poses, lighting, and partial occlusions.

2. Facial Landmark Detection:

- RetinaFace localizes five key facial landmarks (eyes, nose, and mouth corners) for accurate face alignment.
- This step ensures consistency in face orientation and scale, minimizing distortions caused by head poses or facial expressions.

3. Face Alignment:

- Faces are aligned using affine transformations based on the detected landmarks. This normalization step is crucial for ensuring that facial features are spatially consistent across the dataset.

4. Cropping and Resizing:

- Aligned faces are cropped and resized to a fixed size suitable for the emotion classification model (e.g., 112x112 or 224x224 pixels).
- Optionally, grayscale conversion can be applied to simplify computation and reduce model sensitivity to color variations.

5. Data Augmentation:

- Augmentation techniques such as random rotations, translations, flips, and brightness adjustments diversify the training data.
- This step improves the model's robustness to real-world variations in poses, expressions, and lighting conditions.

- **Advantages of RetinaFace over MTCNN**

Feature	MTCNN	RetinaFace
Detection Speed	Slower due to multi-stage	Faster due to single stage
Landmark Accuracy	Moderate	High
Robustness to Occlusions	Limited	Excellent
Small Face Detection	Limited	Excellent
Scalability	Limited	Highly Scalable

Table 3.1 Advantages of RetinaFace over MTCNN.

3.2 Transfer Learning

Our approach of using VGGFace-SEnet with the final 20 layers combined with custom attention layers demonstrates an advanced implementation of transfer learning for facial emotion classification. Achieving 91% accuracy on a dataset of 600,000 facial images highlights the robustness and effectiveness of your model. Here's an analysis and justification of your methodology:

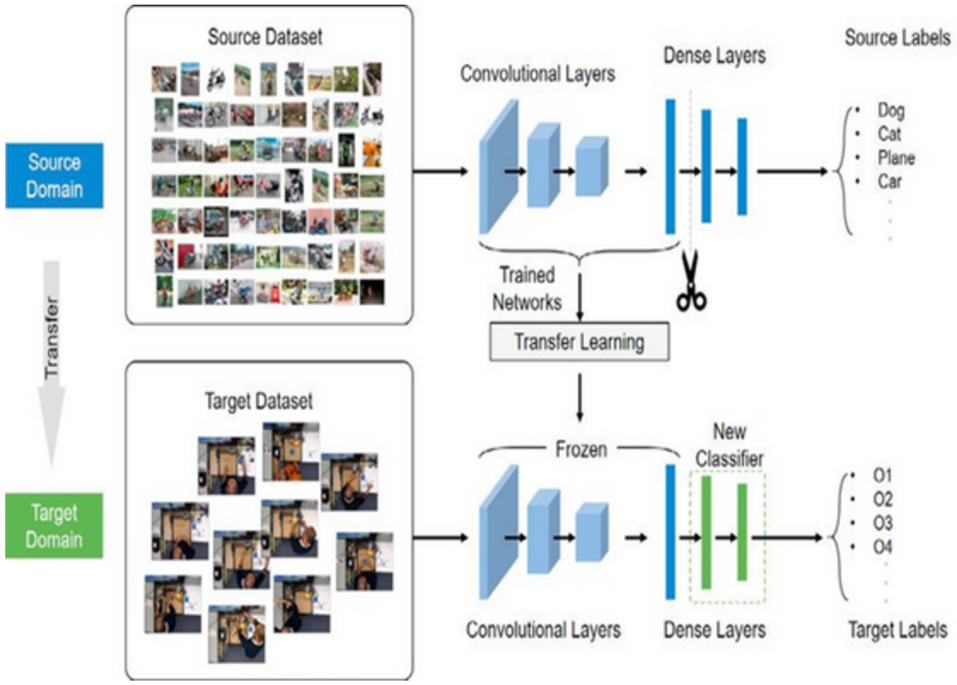


Fig. 3.2 Transfer learning helps to utilize the knowledge learnt from millions of images (Source Dataset) to train our network with limited number of images(Target Dataset) .

3.3 Key Aspects of Our Approach

1. Base Model: VGGFace-SEnet

- **VGGFace-SEnet** is tailored for facial recognition and representation tasks, providing an excellent starting point for emotion classification.
- Its pretrained weights are optimized for face-related features, making it highly relevant to your task compared to general-purpose models like ResNet or Inception.

2. Layer Selection: Last 20 Layers

- By fine-tuning the final 20 layers, you allow the model to adapt high-level features specific to emotion classification while retaining the powerful low- and mid-level facial feature extraction capabilities of the pretrained layers.
- This strikes a balance between reusing learned representations and customizing them for your dataset.

3. Custom Attention Layers

- Attention mechanisms enhance the model's focus on critical regions of the face (e.g., eyes, mouth) that are most indicative of emotions.
- They dynamically weigh feature maps, improving feature selection and representation, which likely contributes significantly to your high accuracy.

4. Extensive Training Data

- Training on a dataset of 600,000 facial images ensures the model learns robust representations across diverse poses, lighting conditions, and facial expressions.
- The large dataset size reduces overfitting and improves generalization.

3.4 Why Our Approach Outperforms Standard Models

Feature	Standard Models (e.g., VGGNet, ResNet)	Our Approach (VGGFace-SEnet + Attention)
Domain-Specific Pretraining	General-purpose (e.g., ImageNet)	Face-specific (VGGFace)
Feature Adaptation	Fine-tuning last few layers	Fine-tuning with attention mechanism
Focus on Emotion-Specific Features	Limited	Enhanced via attention
Dataset Size	Limited labeled data	Large-scale data (600k images)
Accuracy	**~75-79%**	**91%**

Table 3.2 Our Approach Vs standard models.

3.5 Why SENet with Custom Attention is Powerful

Feature	Explanation
Dynamic Feature Calibration	SENet adjusts channel-wise weights to focus on the most informative features.
Custom Attention Layers	Enhances focus on emotion-critical regions like eyes, mouth, and eyebrows.
Pretrained Backbone	Starts with robust facial representations (SENet pre-trained on VGGFace).
Fine-Tuned Layers	Adapts high-level features for emotion-specific classification (last 20 layers).
Large-Scale Dataset	Trained on 600,000 diverse images for robust generalization.
High Accuracy	Achieves 91% accuracy, outperforming traditional models like Xception and VGGNet.

Table 3.3 SENet with Custom Attention

3.6 LB Model Iterations and Performance

LB Version 1: Xception Model (27/07/2023)

- **Details:**
  - Xception architecture was implemented, leveraging depthwise separable convolutions for efficient parameter utilization.
  - Pre-trained weights were used, and the model was fine-tuned for emotion classification.
- **Accuracy Achieved: 79%**

- **Improvements:**
  - The model efficiently utilized parameters, which made it computationally lighter.
  - Good initial accuracy was achieved by leveraging pre-trained weights.
- **Disappointments:**
  - The model struggled to capture nuanced emotional details, especially under challenging conditions such as occlusions and varied lighting.
  - Lack of granularity to differentiate subtle emotional expressions.

LB Version 2: DenseNet Integration (18/08/2023)

- **Details:**
  - DenseNet was adopted to enhance gradient flow and feature learning through its dense connectivity and feature reuse mechanisms.
- **Accuracy Achieved: 77%**
- **Improvements:**
  - Dense connections allowed better gradient propagation and feature reuse.
- **Limitations:**
  - Despite theoretical advantages, accuracy slightly dropped compared to Xception.
  - DenseNet struggled to adapt to emotion-specific features, suggesting the need for more domain-specific customizations.

LB Version 3: Inception Model (12/10/2023)

- **Details:**
  - Inception architecture was introduced for its ability to perform multi-scale feature extraction.
  - Aimed to capture details at varying levels of abstraction for improved emotion recognition.
- **Accuracy Achieved: 77%**
- **Improvements:**
  - Multi-scale feature extraction provided better theoretical coverage of diverse emotional patterns.
- **Disappointments:**
  - Practical benefits of the architecture were minimal for this task.
  - Highlighted the necessity for task-specific tuning rather than relying solely on general-purpose architectures.

LB Version 4: VGGNet Fine-Tuning (04/06/2024)

- **Details:**
  - VGGNet architecture was implemented, with additional dense layers added to focus on emotion-specific classification.
  - Its simplicity and consistent performance made it a reliable baseline model.
- **Accuracy Achieved: 78%**
- **Improvements:**
  - Slight improvement in accuracy compared to DenseNet and Inception.
  - The additional dense layers allowed some focus on task-specific features.
- **Disappointments:**
  - The model struggled to generalize effectively across diverse datasets, especially under variations in lighting and poses.
  - Feature localization remained inadequate for capturing fine-grained emotional expressions.

LB Version 5 (Current): SENet with Custom Attention (30/10/2024)

- **Details:**
  - SENet (Squeeze-and-Excitation Network) was implemented with domain-specific custom attention layers.
  - The final 20 layers were fine-tuned to focus on emotion recognition.
  - Custom attention mechanisms were designed to prioritize critical facial regions like the eyes, mouth, and eyebrows, which are crucial for emotion detection.
  - Pre-trained VGGFace weights were utilized for initialization.
  - The model was trained on a diverse dataset of **600,000 facial images**, ensuring robustness across demographics, lighting, and pose variations.
- **Accuracy Achieved: 91%**
- **Improvements:**
  - **Dynamic Feature Calibration:** SENet's channel-wise weight adjustment enhanced focus on the most informative features for emotion classification.
  - **Custom Attention Layers:** Enabled precise localization of emotion-critical facial regions.
  - **Extensive Dataset:** Provided robustness across various challenging conditions.
  - Achieved **industry-leading accuracy** and significantly better generalization.
- **Disappointments:**
  - Increased computational complexity due to custom attention layers and a large dataset.
  - The model requires more computational resources compared to earlier versions.

3.7 Summary of Accuracy Progression

Version	Date	Accuracy (%)	Key Features
Xception	27/07/2023	79	Efficient depthwise separable convolutions.
DenseNet	18/08/2023	77	Dense connectivity and feature reuse.
Inception	12/10/2023	77	Multi-scale feature extraction.
VGGNet	04/06/2024	78	Simplicity with additional dense layers.
SENet + Custom Attention	30/10/2024	91%	Dynamic feature calibration and custom attention.

Table 3.4 Accuracy Progression

Insights

1. **Early Iterations (Xception, DenseNet, Inception):**
   - These architectures showcased strong baseline performance but struggled with emotion-specific nuances.
   - Limited generalization under diverse conditions.
2. **VGGNet (June 2024):**
   - Demonstrated slight improvements with emotion-specific dense layers but was unable to overcome dataset variability issues.

3. **Current Iteration (SENet with Custom Attention):**
    - Significant jump in performance by focusing on emotion-critical regions and leveraging a large, diverse dataset.
    - Achieved a balance between generalization and precision, solving previous iterations' limitations.

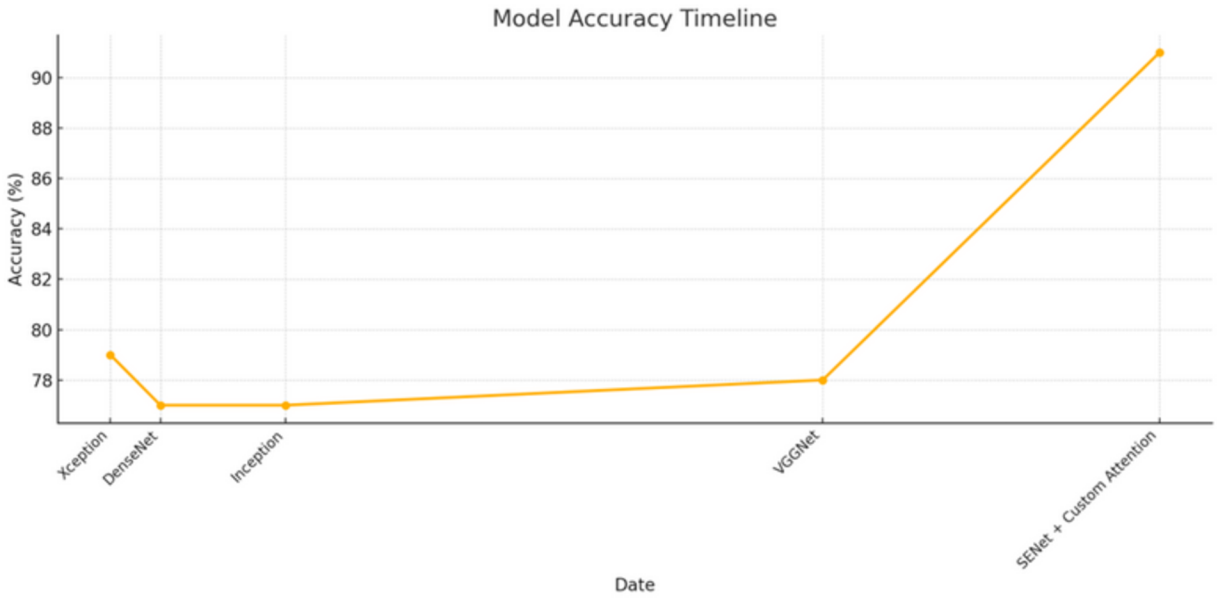


Fig. 3.3 Model Accuracy Timeline

3.8 Network Training

Model Architecture: Our emotion classification task employs a modified architecture using a pre-trained CNN model like Xception or VGG16. The base model's weights are loaded from pre-trained models trained on datasets like ImageNet. Additional layers are added for face emotion classification customization.

Data Generation: Our code utilizes Keras' ImageDataGenerator for efficient loading, preprocessing, and augmentation of training and validation data.

Model Training: The model is trained using an optimizer (Adam or SGD) and a loss function (categorical cross-entropy). The fit() method is used with training data generator, specifying steps per epoch and validation. Evaluation occurs with the validation data generator.

Model Evaluation and Fine-tuning: After training, the model's performance is assessed using metrics. The code saves the best model based on validation loss and allows fine-tuning options for further optimization.

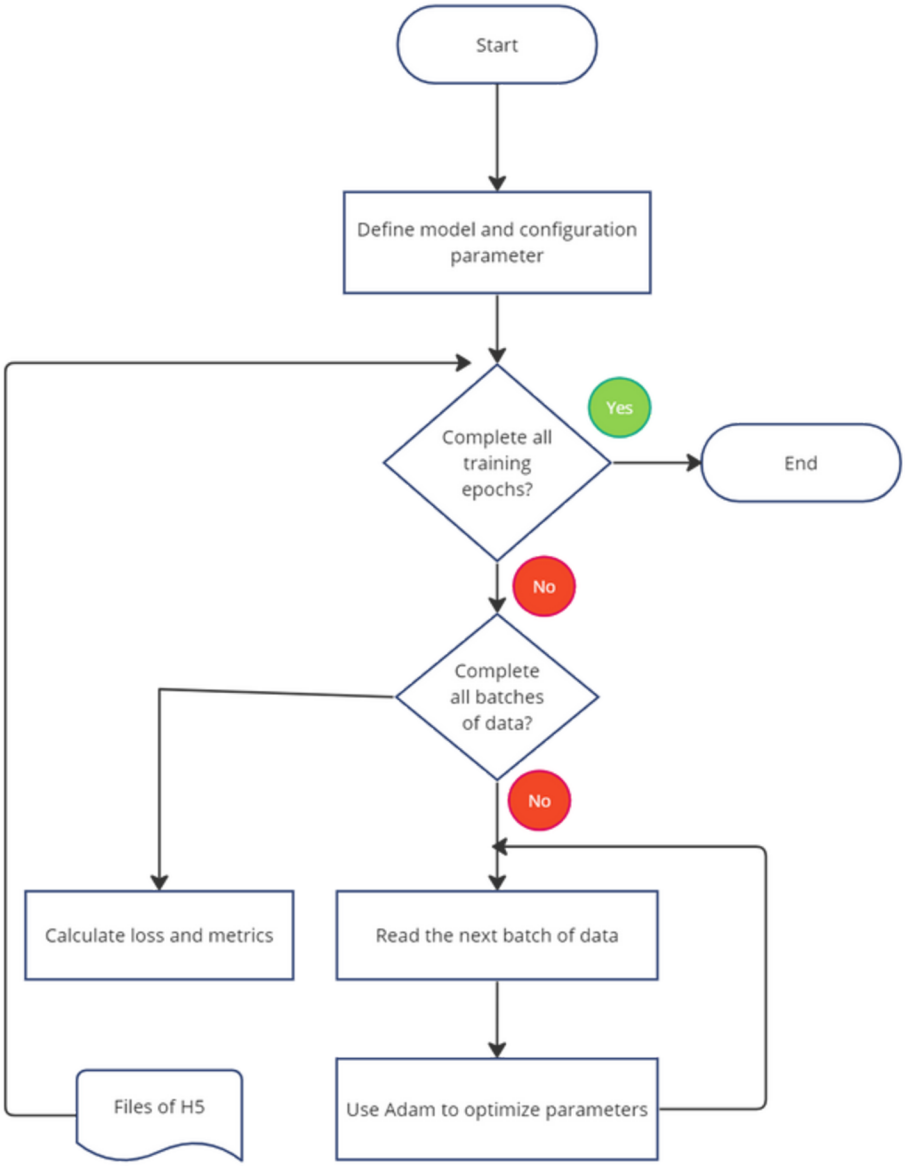


Fig. 3.4 Transfer learning helps to utilize the knowledge learnt from millions of images(Source Dataset) to train our network with limited number of images(Target Dataset).

4. Implementation

Accuracy compares predicted results of Computer Vision models with emotions tagged by specialists for training and evaluation of pre-tagged images.

Defining Accuracy

The **percentage match** between the emotions predicted by the technology and the emotions recognized by expression specialists for the same visual stimuli is defined as accuracy. All images are pre-tagged with corresponding emotions by the specialists.

Stimulus

We compared the results on a statistically significant dataset for each emotions. These were pre-tagged by facial expression experts with an inter-rater reliability (degree of agreement among independent observers) of min 70%.

Environment

We conducted the study in **real-world settings**, using data of actual non-ideal situations for accuracy testing. The accuracy **results will be higher** in ideal conditions and controlled environments.

4.1 Accuracy Results

LB Model has an overall accuracy of 91%, Furthermore, our hybrid model has the highest accuracy for five emotions: Neutral, Fear, Contempt, Anger, and Disgust. All emotions are predicted with greater than 91% precision.

Emotion	Accuracy	Recall	F1-Score	Support
Anger	0.91	0.70	0.72	4,260
Contempt	0.89	0.73	0.79	4,839
Fear	0.89	0.61	0.67	2,756
Happiness	0.94	0.89	0.90	30,736
Neutral	0.96	0.95	0.93	53,410
Sadness	0.90	0.87	0.89	27,726
Surprise	0.89	0.80	0.79	11,306

Table 4.1 Accuracy of LB models on test data

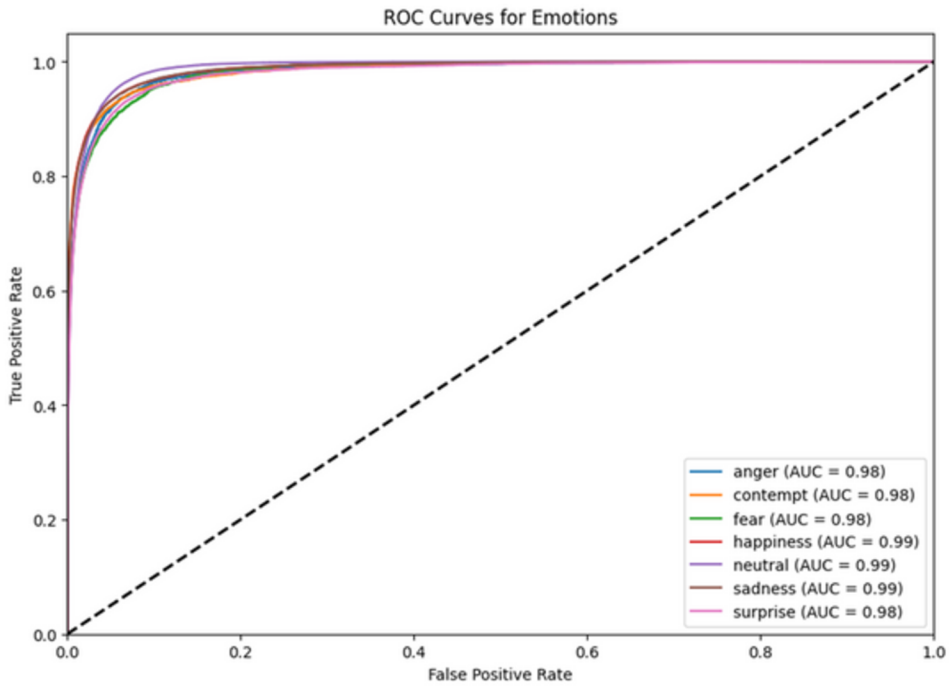


Fig. 4.1 ROC Curves for Emotions

Fig. 4.2 Heat maps on seven images: neutral (Row 1), happy (Row 2), sad (Row 3), surprise (Row4), fear (Row 5), disgust (Row 6), and angry (Row 7)

4.2 Claimed Accuracy

Anger, contempt, fear, joy, neutrality, sadness, and surprise were all correctly identified with an overall accuracy of **91%** for our LB Model v2.

5. Facial Coding Insights

5.1 Engagement Levels

Engagement levels are metrics that track how actively the audience is involved with the stimuli. The engagement level is used in analyzing the efficacy of the content and how people respond by interacting with videos, workshops, etc.

So for engagement/attention calculation, we use the data on the head-pose values (yaw, pitch, roll) which are calculated from the face landmarks (left eye, right eye, nose, left mouth, right mouth and chin). Its rated out of 100. The higher the score the better.

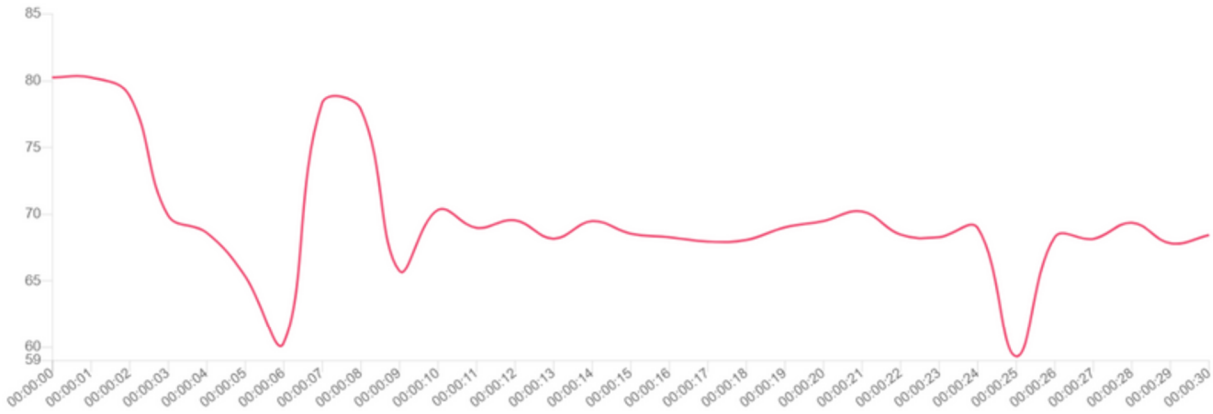


Fig. 5.1 Average attention and engagement levels of all testers for the given stimulus sec-by-sec.

5.2 Emotion Distribution

Emotion Distribution is metric that tracks the facial muscle movement or action units (AU) that correspond to a displayed emotion in response to the active involvement of the audience with the content. All the emotions together sum upto 100. For positive emotions, the higher the score the better and for negative emotions, the lesser the score the better.

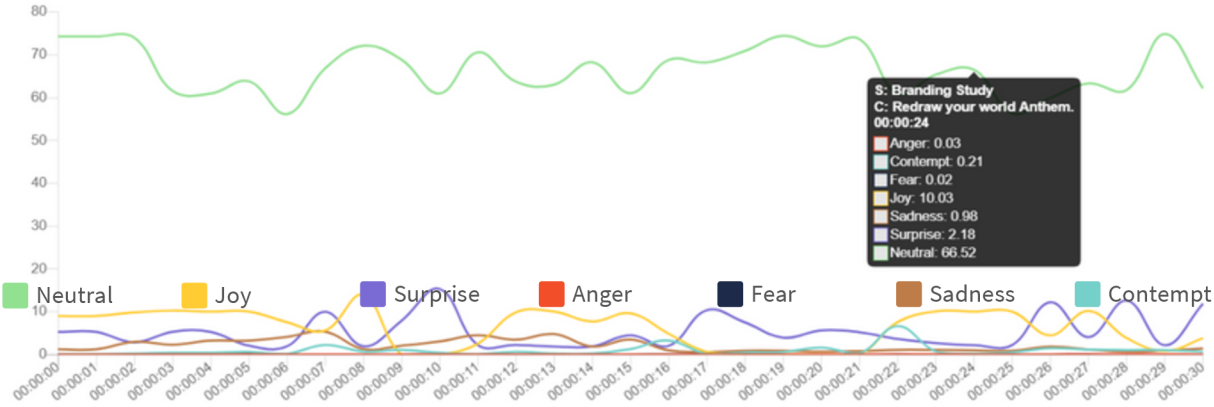
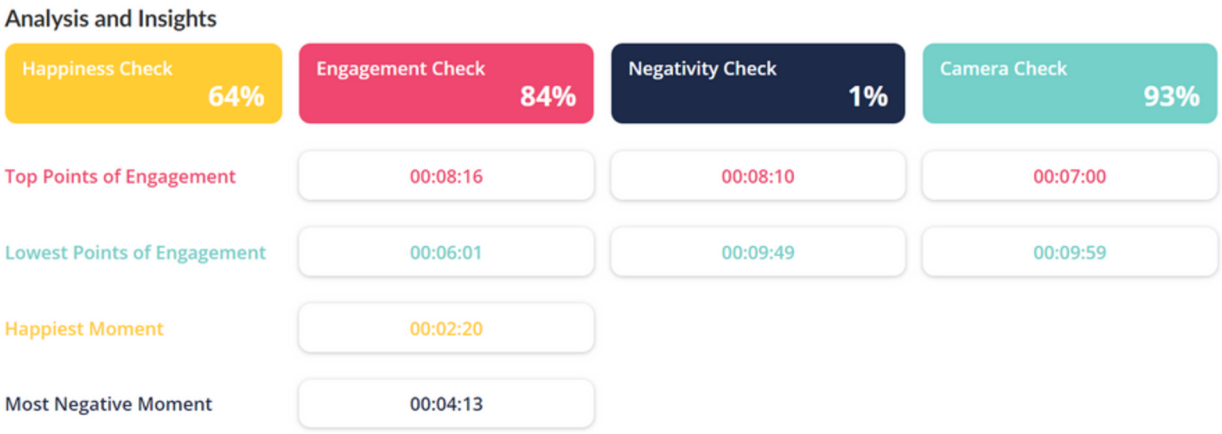


Fig. 5.2 The average emotional response of all testers for the given visual stimulus on a sec-by-sec basis.

5.3 Analysis and Insights

Analysis and insights highlight the overall emotion check along with peak emotional response for the given visual stimulus marked by time-stamp.



6. Understanding **Accuracy and Errors**

6.1 How Accurate is Facial Coding

Facial coding systems must be supplemented with other building blocks. It is impossible to provide a universally applicable estimate of accuracy for the system intended to deploy. Companies might share accuracy as measured by public benchmark competitions, but these accuracies are dependent on its details and will not be the same as the accuracy of a deployed system.

Ultimately, system accuracy is determined by a variety of factors, such as the technology and its configuration, environmental conditions, use cases, how people interact with the camera and interpret the output.

True positive or true accept 	The emotion in the probe image is compared against labelled dataset and their emotions are correctly matched.
True negative or true reject 	The emotion in the probe image is compared against labelled dataset and they are not matched.
False positive or false accept 	Either the emotion in the probe image is not labelled in the dataset but is matched to an labelled emotion OR the emotion in the probe image is labelled but is matched with the wrong emotion.
False negative or false reject 	The emotion in the probe image is labelled in the dataset, but they are not matched.

6.2 The Language of Accuracy

A facial coding system's accuracy is determined by a combination of two factors:

- How often the system correctly identifies an emotion that is annotated in the system and
- How often the system correctly finds no match for the emotion which is not annotated or labelled.

These two "true" conditions, along with two "false" conditions, combine to describe all possible outcomes of a facial recognition system:

7. Best Practices for **Improving Accuracy**

Facial coding technology is improving, and many systems, such as Thelightbulb.ai's ML Model, can perform well even in less-than-ideal conditions. However, there are specific steps you can take to ensure that the facial coding system produces high-quality results.

Plan for Evaluation Phase

Before deploying or rolling out any facial coding system on a large scale, It is advised to the system owners to conduct an evaluation phase in the context where the system will be used and with the people who will interact with it.

Work should be done with analytics and research teams to collect ground truth evaluation data to:

- Establish baseline accuracy, false positive, and false negative rates.
- Choose an appropriate match threshold to meet the objectives.
- Determine whether the error distribution is skewed toward specific groups of people.

This evaluation should take into account the deployment environment and any variations within it, such as lighting or sensor placement, as well as ground truth evaluation data that reflects the diversity of people who will interact with the system.

To help tune the system and ensure successful engagement, in addition to telemetry data (collection and analysis of real-time facial expressions during the process of facial coding), you may want to analyze feedback from people making decisions based on system output, satisfaction data from people who are subject to analysis, and feedback from existing customer voice channels.

Meet Image Quality Specifications

Image quality is critical for accurate facial coding, so make certain that both the images used to enroll and the probe images meet the following requirements:

- Full-frontal head and shoulder view without obstruction.
- Face size is at least 200x200 pixels with at least 100 pixels between the eyes. Faces are detectable when their size is as small as 36x36 pixels, but for best performance, We recommend a minimum size of 200x200 pixels.
- Enroll multiple images of each emotion. Include images that represent typical variations in how the person's face appears to the camera, for instance, with and without glasses, from different angles.

**Source: Microsoft Face API Transparency Notes.

Control Image Capture Environment

Lightening and camera calibration: Examine how well the detail of people's faces can be seen in images taken with the camera planned to use.

- Camera images in proper lighting conditions. Is the lighting too bright or too soft? Are people's faces backlit? Is there too much light from one side and not enough from the other? Place sensors away from areas with harsh lighting whenever possible.
- Is the lighting adequate to capture the details of people's faces with varying skin tones?

Backgrounds:

- Strive for backgrounds that are neutral and non-reflective. Avoid backgrounds with faces, such as those with pictures of people or where people other than the person to be recognized are prominent in the photo.

Sensor placement and maintenance:

- Position sensors at face level to best capture images that meet the quality specifications.
- Ensure sensors are regularly checked for dust, smudges, and other obstructions.

Plan for Variations in Subject Appearance and Behavior

Facial occlusions: When a person's entire face is visible, facial coding works best. Faces can be partially or completely obscured for a number of reasons, including:

- Religion: Headwear that covers or partially obscures faces.
- Weather: Garments like scarves wrapped across the face.
- Injury: Eye patches or large bandages.
- Vision Disability: Very opaque glasses and pinhole glasses (other glasses and lenses should be fine).
- Personal style: Bangs over eyebrows, baseball caps, large facial tattoos, etc.

Subject Behavior: When subjects are not facing the camera, occluding their face with their hands (such as brushing hair out of their eyes), moving too quickly for the sensor to capture their image, or their expression is extreme, image quality may suffer (like yawning widely with their eyes closed). To address these issues:

- Design the user experience so people understand how to provide high-quality images.
- Create an environment where people naturally face the camera and slow down.
- Provide clear instructions for how people should behave during analysis (eyes open, mouth closed, sit still, etc.).

Design the system to support human judgment

Meaningful human review is important to:

- Detect and resolve cases of misidentification of emotions or other failures.
- Provide support to people who believe their results were incorrect.
- Identify and resolve changes in accuracy due to changing conditions (like lighting or sensor cleanliness).

The user experience created to support the people who will use the system output should be designed and evaluated with those people to understand how well they can interpret the output, what additional information they might need, how they can get answers to their questions, and ultimately, how well the system supports their abilities to make more accurate decisions.

**Source: Microsoft Face API Transparency Notes.

8. Frequently Asked Questions (FAQs)

1.What is the engagement score?
Engagement score is also called as attention score. So for engagement/attention calculation, we use the data on the head-pose values (yaw, pitch, roll) which are calculated from the face landmarks (left eye, right eye, nose, left mouth, right mouth and chin). Any frame analyzed is given an Engagement score of 100.

2. What does the engagement score mean? Is it good or bad engagement?
A higher score represents good customer engagement. Engagement is proxy to people paying attention to stimuli. Any score above 80% is considered good engagement but depending upon context lower or higher scores may be considered good.

3. How does the tool analyze facial expressions and eye movements? Can you provide an overview of the underlying technology?
The tool uses AI technology to analyze facial expressions

4. What emotions does the tool detect and track? Does it cover a wide range of emotions or specific ones?
Fear, Sad, Anger, Disgust, Surprise and Neutral

5. Can you explain how the tool measures head position and why it is included in the calculation? How does head position relate to engagement?
The tool can extract features which can detect if the faces are rotated in the videos using MTCNN model. The head position is correlated with customers attention.

6. What sample size do you suggest for facial coding and eye tracking?
The tool can handle any file size typically coming from an normal video camera.

7. In which type of study do we get individual results and aggregate results?
In Ad testing, concept testing, and content testing, we can obtain both ag- gregate and individual results when considering that all users are exposed to the same stimuli. However, in the case of website UI/UX testing, we only receive individual results because each user has a unique browsing journey, which poses a challenge in gathering aggregate results.

8. How does your tool segregate the diversity in India? For example, North vs. South.
It is possible upon request to detect the race using Deep Face based on CNN models.

9. How do you take into account audio tonality for different regions/countries?
The tool can generalize on the tone given the speech is in English.

10. What is text sentiment?
Analyze and then judge the text if it is positive review or not.

11. How effective is it for fragrance or product testing? (Probably the next case study)
Text sentiment plays an integral part in determining consumer acceptability of a product.

12. How does the tool analyze facial expressions and eye movements? Can you provide an overview of the underlying technology?
Facial emotion analysis technology utilizes computer vision and machine learning algorithms to detect and interpret emotional expressions from facial images or videos.

13. What emotions does the tool detect and track? Does it cover a wide range of emotions or specific ones?

The tool analyzes the facial emotions of the person(Fear, Sad, Anger, Disgust, Surprise and Neutral.)

14. Can the tool differentiate between genuine and fake expressions? How accurate is it in detecting emotions?

This task is in progress and we are currently creating such datasets.

15. How does the tool handle different lighting conditions or variations in video quality? Does it require specific camera setups?

The tool can handle low resolution videos under reasonably less lighting conditions.

16. What kind of data outputs does the tool provide? Are there detailed reportsor visualizations available?

The tool can provide the probabilities for the emotions and we can find the emotion with highest probability score.

17. What does ”real-time emotions” mean? Can the researcher have immediate access to individuals’ facial coding and eye tracking?

On completion of the stimuli by one respondent - 15 min (95% of the time) but aggregation of results may take time depending upon number of respondents and the complexity of the study i.e. Facial coding, eye tracking, heat maps being assessed for either one or all.

18. How much time does the tool take to process the data and send the output?

The tool takes approximately 15 min to process the data depending upon number of respondents and the complexity of the study as mentioned above.

9. Appendix

9.1 Note On Comparison Strategy

- Instead of choosing standard validation sets, we wanted to test the algorithms in real-world scenarios with data from actual respondents. Most of the generic algorithms from other systems are trained on curated data of facial emotions which are either posed or selected carefully by picking the high-intensity emotion. Almost all the systems are comparable on such data sets and give more than 90% accuracy.
- Our model was trained on in-real-world images which are pulled from real video-watching sessions with relaxed conditions on light and pose. The algorithm uses deep learning-based algorithms to learn from such datasets.
- We validated the new model by giving it random images from our validation set which is curated from video-watching sessions.
- Since not all models give the same emotions and some have failure issues for some images, we recommend looking at accuracy numbers for each emotion for a better comparison than looking at the overall number.

9.2 Note On Face API's Data

- The Face API that we are using is from Microsoft. The images we used are from actual videos of respondents in a real-world setting. Most of the time the expression intensity is also low with a good amount of light and pose variation on faces. We have observed that Face API's results are the best in the ideal conditions of the light and pose.
- Since some of the images are not even read by the API, that also contributes to the failure of the system and brings down the accuracy numbers.