Emotional Facial Expression Detection using YOLOv8

Aadil Alshammari
Department of Information Systems, Faculty of Computing and Information Technology, Northern Border University, Arar, Saudi Arabia
aadil.al-shammari@nbu.edu.sa (corresponding author)

Muteb E. Alshammari
Department of Information Technology, Faculty of Computing and Information Technology, Northern Border University, Arar, Saudi Arabia
muteb.alshammari@nbu.edu.sa

Received: 18 July 2024 | Revised: 27 July 2024 and 31 July 2024 | Accepted: 4 August 2024

 DOI: https://doi.org/10.48084/etasr.8433

ABSTRACT

Emotional facial expression detection is a critical component with applications ranging from human-computer interaction to psychological research. This study presents an approach to emotion detection using the state-of-the-art YOLOv8 framework, a Convolutional Neural Network (CNN) designed for object detection tasks. This study utilizes a dataset comprising 2,353 images categorized into seven distinct emotional expressions: anger, contempt, disgust, fear, happiness, sadness, and surprise. The findings suggest that the YOLOv8 framework is a promising tool for emotional facial expression detection, with a potential for further enhancement through dataset augmentation. This research demonstrates the feasibility and effectiveness of using advanced CNN architectures for emotion recognition tasks.

Keywords-emotion detection; deep learning; YOLOv8

I. INTRODUCTION

Machine learning has revolutionized numerous fields by enabling systems to learn from data and make intelligent decisions. Among the various machine learning techniques, deep learning has shown remarkable success, particularly in tasks that involve image and video analysis. One of the most advanced and efficient deep learning architectures for object detection is YOLO (You Only Look Once) [1], with YOLOv8 offering significant improvements in speed and accuracy. The architecture of YOLOv8 is designed to perform real-time object detection, making it highly suitable for applications requiring quick and precise identification of objects within images.

Facial Expression Recognition (FER) is a critical application of machine learning, particularly in the domains of human-computer interaction, security, and psychological research. FER involves the automatic identification of human emotions based on facial expressions, which can provide valuable insights into human behavior and emotional states. Traditional FER methods often struggle with low accuracy and poor generalization, especially under varying conditions such as different lighting, occlusions, and facial makeup. Convolutional Neural Networks (CNNs) have emerged as a powerful tool for FER due to their ability to automatically learn and extract hierarchical features from raw image data. CNNs consist of multiple layers, including convolutional layers, pooling layers, and activation functions, which work together to capture spatial hierarchies in images. This capability makes CNNs particularly effective for tasks that involve image recognition and classification [2]. The application of CNNs to FER has shown promising results, with various studies demonstrating significant improvements in recognition accuracy. By leveraging the strengths of CNNs, various models have been proposed to accurately classify facial expressions into distinct emotional categories [3-9]. These models typically involve preprocessing steps, such as face detection, normalization, and augmentation, to enhance the quality and diversity of the training data.

This study explores the use of YOLOv8, a state-of-the-art object detection architecture, for FER. Integrating the robust feature extraction capabilities of CNNs with the real-time detection efficiency of YOLOv8 aims to develop a model that can accurately and swiftly recognize facial expressions.

II. RELATED WORKS

In [10], FER was explored using CNNs to address the limitations of traditional methods, such as low accuracy and weak generalization. The architecture of the proposed CNN included multiple convolution layers, pooling layers, and activation functions. Data preprocessing steps included the use of the Viola-Jones algorithm for face detection. The dataset was expanded through transformations, consisting of 35,887

facial images categorized into seven expressions. The CNN model achieved a recognition rate of more than 70% on the training set and more than 80% on the test set. This study concluded that CNNs are effective for FER, although further research is needed to handle extreme conditions such as makeup and occlusion, and emphasized the importance of large datasets to improve model generalization and reduce overfitting.

In [11], a FER approach was introduced by combining CNNs with Scale Invariant Feature Transform (SIFT) features, specifically Dense SIFT. This hybrid method leveraged the high accuracy of CNNs and the robustness of SIFT, particularly in scenarios with limited training data to improve performance. The proposed model was tested on the FER-2013 and CK+ datasets, achieving 73.4% accuracy on FER-2013 and 99.1% on CK+. This study showed that the combination of Dense SIFT and CNN features significantly improves facial expression recognition accuracy, outperforming traditional CNN and other hybrid models. In [12], a deep learning-based approach was introduced, comprising two main components. The first component extracted local features from facial images using a local gravitational force descriptor. These features were then input into a Deep Convolutional Neural Network (DCNN) model, which had two branches. The first branch focused on geometric features, such as edges, curves, and lines, while the second branch extracted holistic features. The final classification score was calculated using a score-level fusion technique. This method was tested on five benchmark datasets including seven emotions, and was compared to existing approaches using accuracy, precision, recall, and F1-score.

In [13], efficient CNN structures and image preprocessing methods were investigated to enhance FER. The main contributions of this paper were twofold: the identification of an efficient CNN structure and the determination of the most effective image preprocessing method for FER. Four different CNN architectures and several preprocessing techniques were evaluated across five datasets (FER2013, SFEW2.0, CK+, KDEF, Jaffe). The results showed that Histogram equalization (Hist-eq) consistently achieved the highest performance across all network models. Among the CNN structures, Tang's network, when combined with Hist-eq images, achieved the highest accuracy while maintaining lower computational complexity compared to other models such as Caffe-ImageNet. In [14], a two-channel CNN architecture was introduced for FER, improving previous Multi-channel Convolutional Neural Networks (MCCNN). This model replaced hard-coded Sobel feature extractors with a combination of a standard CNN channel and a channel trained as a Convolutional Autoencoder (CAE) to learn Gabor-like filters. These filters were pre-trained on the Kyoto natural images dataset and remained fixed during the supervised training phase. This architecture was evaluated using the JAFFE dataset, which includes images of Japanese women displaying seven different facial expressions. The study employed two evaluation methods: leave-one-out and ten-fold cross-validation. The results showed that the proposed model achieved recognition rates with an average accuracy of 95.8% in the leave-one-out experiment and 94.1% in the ten-fold cross-validation, significantly outperforming previous models. This study highlighted the proposed model's improved classification performance and faster training times, attributing these benefits to the use of CAE-based filters and the reduced complexity of the architecture.

In [15], a CNN-based solution was proposed for FER, involving multiple distinct subnets, each representing a compact CNN model that was trained independently. These subnets were then integrated to form the complete network. The model was trained and tested using the FER2013 dataset. The highest-performing individual subnet achieved an accuracy of 62.44%, while the entire assembled model achieved 65.03%, placing it 9th and 5th, respectively, among all considered models. In [16], a CNN-based method was proposed for real-time FER. The proposed CNN-based model was capable of classifying human facial expressions into seven universal categories in real-time using a webcam. The model was trained on the FER2013 dataset, which includes 35,887 labeled images with variations in viewpoint, lighting, and scale. This method encompassed preprocessing steps such as normalization, gray-scaling, and resizing, followed by face detection using Haar cascades, and emotion classification using a CNN with multiple convolutional layers, max pooling, fully connected layers, and activation functions such as ReLU and Softmax. The model achieved a training accuracy of 79.89% and a test accuracy of 60.12%. This study also compared the proposed system with other related works, highlighting its performance.

In [17], a CNN enhanced with an attention mechanism and Local Binary Pattern (LBP) features was proposed for FER. The proposed architecture comprised four modules: feature extraction, attention, reconstruction, and classification. By integrating LBP features, which capture fine texture details and subtle facial changes, with an attention mechanism, the network effectively focused on crucial facial features, such as eyes and mouth, improving recognition accuracy. This study also introduced a new dataset, the Nanchang University Facial Expression (NCUFE), containing RGB and depth images of seven expressions from 35 subjects. Extensive experiments on multiple datasets, including CK+, JAFFE, FER2013, Oulu-CASIA, and NCUFE, showed that this method outperformed other algorithms, achieving recognition rates of 75.82% on FER2013, 98.68% on CK+, 98.52% on JAFFE, 94.63% on Oulu-CASIA, and 94.33% on the newly introduced NCUFE dataset. In [18], a CNN architecture was proposed, which included two convolutional layers, two max-pooling layers, and one fully connected layer. The training process involved setting parameters such as batch size, epoch number, and learning rate. This study merged three datasets (JAFFE, KDEF, and a custom dataset) to train the CNN to classify seven different emotions. The proposed model achieved a training accuracy of 96.43% and a validation accuracy of 91.81%. This study highlights the effectiveness of the LeNet model in real-time emotion recognition, particularly excelling in predicting emotions such as surprise, fear, and neutrality, while being less accurate in predicting sadness.

III. DATASET AND METHODS

A. Dataset

The dataset used in this study is publicly available in Roboflow [19]. As shown in Table I, the dataset consists of a

total of 2353 images, categorized into 7 classes: anger, contempt, disgust, fear, happy, sadness, and surprise.

TABLE I. DATASET STATS

Category	Number of images
Training dataset	2058
Validation dataset	196
Testing dataset	99
Total	2353

The dataset was divided into three subsets: training (87%), validation (8%), and testing (4%). Specifically, 2058 images were allocated for training, 196 images for validation, and 99 images for testing. This structured division ensures a robust evaluation of the model's performance across different stages of development.

B. Model

This study is based on the state-of-the-art YOLOv8 architecture, which is anchored on a 53-layer convolutional neural network that leverages cross-stage partial connections to enhance information flow and performance. In particular, it incorporates a feature pyramid network to adeptly detect objects of varying sizes. The backbone, akin to YOLOv5's design but with improvements, extracts multilevel features and employs the C2f module, which is inspired by YOLOv7's ELAN concept, for efficient training and rich feature representation, while sidelining mosaic data augmentation in later training phases. The neck acts as a conduit, refining features from the backbone using FPNs and C2f modules for robust multiscale integration before passing them to the head. The head component independently performs classification and regression, avoiding the combined approach of previous iterations for improved accuracy. It employs an anchor-free strategy, directly predicting bounding boxes to accommodate diverse object sizes and shapes, and incorporates a self-attention mechanism to focus on pertinent image features, culminating in precise object detection. This innovative design, with its focus on independent task processing and attention to detail, allows YOLOv8 to achieve superior detection outcomes.

IV. RESULTS AND DISCUSSION

The results are evaluated using a confusion matrix, precision, recall, mean Average Precision (mAP), and other performance metrics. Figure 1 shows the accuracy and performance of the model through precision-recall and precision-confidence curves. The curves show the overall accuracy of the model for each class: anger (0.874), contempt (0.927), disgust (0.985), fear (0.610), happy (0.987), sadness (0.489), and surprise (0.986). The mAP at 0.5 (mAP50) for all classes is 0.837, indicating a high level of precision and recall for most emotion classes, with room for improvement in detecting fear and sadness. These results highlight the effectiveness of YOLOv8 in emotional facial expression detection, while also identifying areas for further refinement, particularly in the detection of emotions with lower accuracy rates.

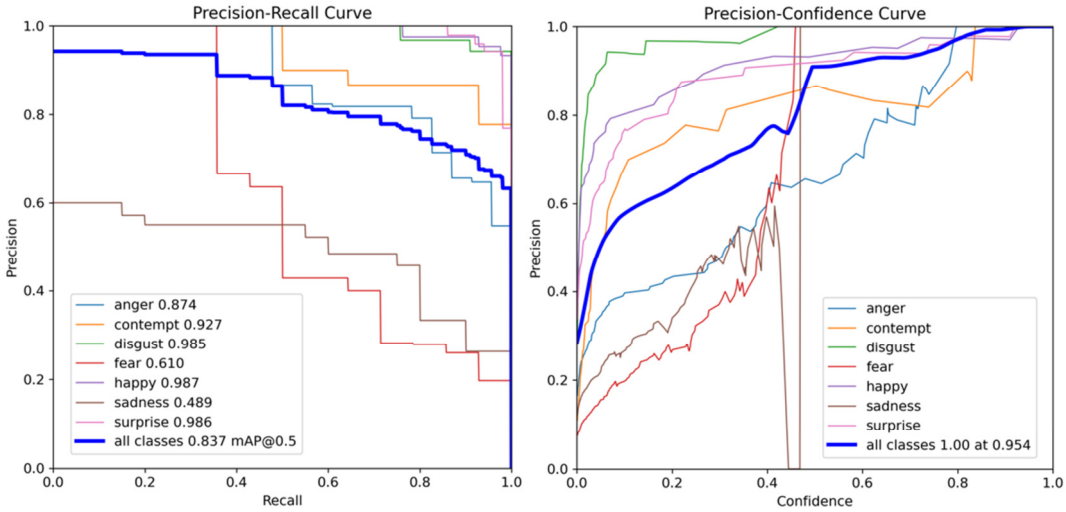

Fig. 1. Precision-recall and precision-confidence curves.

Figure 2 shows the normalized confusion matrix used to evaluate the model's performance. The values in the matrix indicate the proportion of predictions for each category, ranging from 0 to 1, where 1 suggests a perfect match between the predicted and the true value. The model was exceptionally accurate in predicting anger with a perfect score of 1.00, indicating that whenever anger was the true emotion, the model identified it correctly. Similarly, happy emotions were predicted with high accuracy at 0.95, suggesting that the model had little difficulty recognizing happiness. Contempt was also predicted with a high accuracy of 0.93, while disgust was correctly identified 70% of the time. Surprise was also detected with relatively high accuracy, achieving an 88% accuracy, although sometimes it was misclassified as fear.

The fear emotion appeared to be challenging for the model, with a low correct prediction rate of 0.36, often being mistaken for sadness and surprise. Sadness was the least accurately predicted emotion at 25% and was most often confused with fear, with a 29% rate of misclassification. The performance of

the model in correctly predicting sadness and fear could be improved with a more diverse and larger set of training images. The fact that sadness was represented by only 75 images and fear by 84 images in the training set likely contributed to their lower prediction accuracy. In contrast, happy, with a high accuracy of 95%, was represented by 207 images in the training set, and anger by 135 images, which provided the model with more examples to learn from, enhancing its ability to accurately recognize these emotions. More comprehensive training data for sadness and fear could help the model distinguish these emotions more effectively.

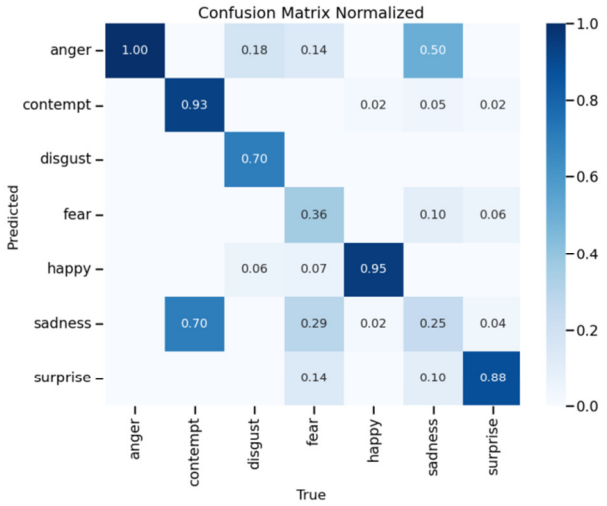

Fig. 2. Normalized confusion matrix.

Figure 3, shows the number of instances for all the emotions. The y-axis represents the number of instances, and the x-axis lists the emotions. The tallest bar is for surprise, indicating that it had the most instances in the dataset.

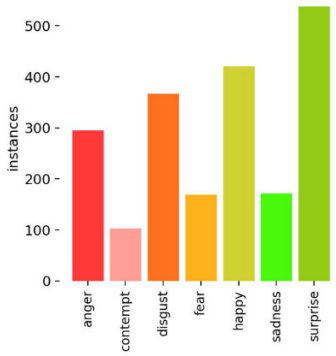

Fig. 3. Instances of emotions in the dataset.

Figure 4 contains four graphs, each representing different evaluation metrics used to assess the performance of the model. The graphs represent the following metrics: precision, recall, mAP50, and mAP50-95.

- The precision graph shows considerable variability in precision values across batches, but the trend line indicates a general increase in precision as training progresses, suggesting that the model is gradually learning to make more accurate predictions.
- The recall graph also exhibits fluctuations, but the trend line points to an overall improvement, suggesting that the model is getting better at identifying all relevant instances over time.
- The mAP50 graph displays a clear upward trend, with the metric values increasing steadily with each batch. This implies that the model's ability to detect emotions with at least 50% accuracy is consistently improving.

The consistent upward trends in all metrics suggest that the model is effectively learning and improving its detection capabilities for emotional facial expressions as it processes more training data.

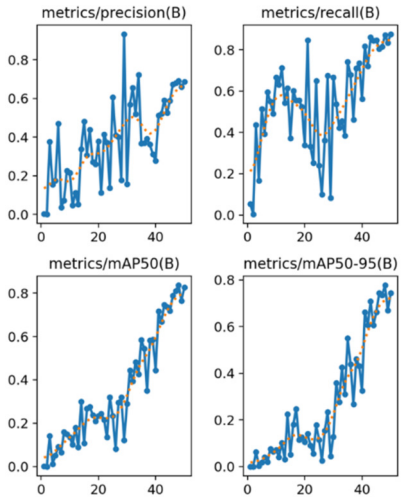

Fig. 4. Model's performance.

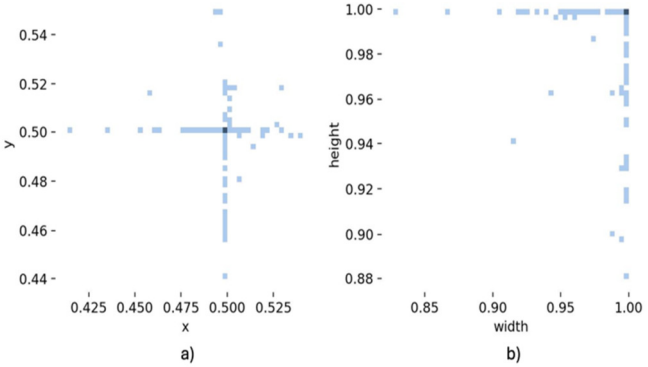

Fig. 5. Scatter plots of detected expressions.

The scatter plot in Figure 5(a) plots the normalized x and y coordinates of the detected facial expressions. Most data points cluster around the center (0.5, 0.5), suggesting that the model frequently detects facial expressions around the center of the image, which is the case. The scatter plot in Figure 5(b) shows the distribution of the width and height of the detected facial expressions. The plot indicates a concentration of detection

with widths and heights close to 1, suggesting that the most detected facial expressions occupy a significant portion of the image. This can be confirmed in the images of the dataset [19].

V. CONCLUSION

This study demonstrated the effectiveness of the YOLOv8 framework in detecting emotional facial expressions, achieving high accuracy rates across several emotion classes. The model's performance indicates that YOLOv8 is a robust tool for emotion recognition tasks. The mean Average Precision (mAP@0.5) of 0.837 for all classes underscores the model's ability to accurately detect and classify emotions from facial expressions. The results highlight the model's strengths in recognizing emotions including anger, contempt, disgust, happiness, surprise, fear, and sadness. The YOLOv8 framework was proven to be effective in the detection of emotional facial expressions, with potential for further refinement. However, future work could focus on augmenting the dataset to address imbalance and improve the detection of challenging emotions. The presence of more instances for certain emotions could indicate a bias in the dataset, potentially affecting the model's ability to generalize across less represented emotions. Furthermore, the scatter plots suggest that the model may be better at detecting larger, centrally located facial expressions, which could affect its performance on smaller or off-center detection.

ACKNOWLEDGEMENT

The authors extend their appreciation to the Deanship of Scientific Research at Northern Border University, Arar, KSA, for funding this research work through project number NBU-FFR-2024-1196-02.

REFERENCES

[1] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics YOLO." Jan. 2023, [Online]. Available: https://github.com/ultralytics/ultralytics.

[2] A. M. Islam, "Exploring Convolutional Neural Networks for Facial Expression Recognition: A Comprehensive Survey," *Global Mainstream Journal of Innovation, Engineering & Emerging Technology*, vol. 3, no. 02, pp. 14–26, Mar. 2024, https://doi.org/10.62304/jieet.v3i02.87.

[3] B. Ma *et al.*, "Distracted Driving Behavior and Driver's Emotion Detection Based on Improved YOLOv8 With Attention Mechanism," *IEEE Access*, vol. 12, pp. 37983–37994, 2024, https://doi.org/10.1109/ACCESS.2024.3374726.

[4] P. Agarwal, "Real-time facial emotion recognition web application," Ph.D. dissertation, International Institute of Information Technology, Hyderabad, India, 2024.

[5] H. Ma, S. Lei, T. Celik, and H. C. Li, "FER-YOLO-Mamba: Facial Expression Detection and Classification Based on Selective State Space." arXiv, May 09, 2024, https://doi.org/10.48550/arXiv.2405.01828.

[6] W. Ismaiel, A. Alhalangy, A. O. Y. Mohamed, and A. I. A. Musa, "Deep Learning, Ensemble and Supervised Machine Learning for Arabic Speech Emotion Recognition," *Engineering, Technology & Applied Science Research*, vol. 14, no. 2, pp. 13757–13764, Apr. 2024, https://doi.org/10.48084/etasr.7134.

[7] H. M. Al-Dabbas, R. A. Azeez, and A. E. Ali, "Two Proposed Models for Face Recognition: Achieving High Accuracy and Speed with Artificial Intelligence," *Engineering, Technology & Applied Science Research*, vol. 14, no. 2, pp. 13706–13713, Apr. 2024, https://doi.org/10.48084/etasr.7002.

[8] M. A. Hasan and A. H. Lazem, "Facial human emotion recognition by using YOLO faces detection algorithm," *Central Asian Journal of Mathematical Theory and Computer Sciences*, vol. 4, no. 10, pp. 1–11, Oct. 2023.

[9] T. Saidani, "Deep Learning Approach: YOLOv5-based Custom Object Detection," *Engineering, Technology & Applied Science Research*, vol. 13, no. 6, pp. 12158–12163, Dec. 2023, https://doi.org/10.48084/etasr.6397.

[10] M. Wang, P. Tan, X. Zhang, Y. Kang, C. Jin, and J. Cao, "Facial expression recognition based on CNN," *Journal of Physics: Conference Series*, vol. 1601, no. 5, Dec. 2020, Art. no. 052027, https://doi.org/10.1088/1742-6596/1601/5/052027.

[11] T. Connie, M. Al-Shabi, W. P. Cheah, and M. Goh, "Facial Expression Recognition Using a Hybrid CNN–SIFT Aggregator," in *Multi-disciplinary Trends in Artificial Intelligence*, Gadong, Brunei, 2017, pp. 139–149, https://doi.org/10.1007/978-3-319-69456-6_12.

[12] K. Mohan, A. Seal, O. Krejcar, and A. Yazidi, "Facial Expression Recognition Using Local Gravitational Force Descriptor-Based Deep Convolution Neural Networks," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–12, 2021, https://doi.org/10.1109/TIM.2020.3031835.

[13] M. Shin, M. Kim, and D. S. Kwon, "Baseline CNN structure analysis for facial expression recognition," in *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, New York, NY, USA, Dec. 2016, pp. 724–729, https://doi.org/10.1109/ROMAN.2016.7745199.

[14] D. Hamester, P. Barros, and S. Wermter, "Face expression recognition with a 2-channel Convolutional Neural Network," in *2015 International Joint Conference on Neural Networks (IJCNN)*, Killarney, Ireland, Jul. 2015, pp. 1–8, https://doi.org/10.1109/IJCNN.2015.7280539.

[15] K. Liu, M. Zhang, and Z. Pan, "Facial Expression Recognition with CNN Ensemble," in *2016 International Conference on Cyberworlds (CW)*, Chongqing, China, Sep. 2016, pp. 163–166, https://doi.org/10.1109/CW.2016.34.

[16] I. Talegaonkar, K. Joshi, S. Valunj, R. Kohok, and A. Kulkarni, "Real Time Facial Expression Recognition using Deep Learning," in *Proceedings of International Conference on Communication and Information Processing (ICCIP)*, May 2019, https://doi.org/10.2139/ssrn.3421486.

[17] J. Li, K. Jin, D. Zhou, N. Kubota, and Z. Ju, "Attention mechanism-based CNN for facial expression recognition," *Neurocomputing*, vol. 411, pp. 340–350, Oct. 2020, https://doi.org/10.1016/j.neucom.2020.06.014.

[18] M. A. Ozdemir, B. Elagoz, A. Alaybeyoglu, R. Sadighzadeh, and A. Akan, "Real Time Emotion Recognition from Facial Expressions Using CNN Architecture," in *2019 Medical Technologies Congress (TIPTEKNO)*, Izmir, Turkey, Jul. 2019, pp. 1–4, https://doi.org/10.1109/TIPTEKNO.2019.8895215.

[19] "Emotion Detection Dataset." Roboflow Universe, 2022, [Online]. Available: https://universe.roboflow.com/lyanhvini/emotion-detection-a5i5h.