Multi-Face Emotion Detection for Effective Human-Robot Interaction

Mohamed Ala Yahyaoui[1], Mouaad Oujabour[1], Leila Ben Letaifa[1][a] and Amine Bohi[2][b]
[1]*CESI LINEACT Laboratory, UR 7527, Vandoeuvre-lès-Nancy, 54500, France*
[2]*CESI LINEACT Laboratory, UR 7527, Dijon, 21800, France*
{*mayahyaoui, moujabour, lbenletaifa, abohi*}*@cesi.fr*

Keywords: Emotion Detection, Facial Expression Recognition, Human-Robot Interaction, Deep Learning, Graphical User Interface.

Abstract: The integration of dialogue interfaces in mobile devices has become ubiquitous, providing a wide array of services. As technology progresses, humanoid robots designed with human-like features to interact effectively with people are gaining prominence, and the use of advanced human-robot dialogue interfaces is continually expanding. In this context, emotion recognition plays a crucial role in enhancing human-robot interaction by enabling robots to understand human intentions. This research proposes a facial emotion detection interface integrated into a mobile humanoid robot, capable of displaying real-time emotions from multiple individuals on a user interface. To this end, various deep neural network models for facial expression recognition were developed and evaluated under consistent computer-based conditions, yielding promising results. Afterwards, a trade-off between accuracy and memory footprint was carefully considered to effectively implement this application on a mobile humanoid robot.

1 INTRODUCTION

The rapid advancement of technology in recent years has accelerated research in robotics, with a particular emphasis on humanoid robots. Designed to resemble humans in body, hands, and head, humanoid robots are increasingly capable of sophisticated interactions with people, including recognizing individuals and responding to commands. This human-like form and behavior make them particularly well-suited for applications in human-computer interaction, serving as effective platforms for studying and improving user engagement and interaction dynamics. Current examples of humanoid robots include Honda's ASIMO (Hirose and Ogawa, 2007), known for its advanced mobility and dexterity; Blue Frog Robotics' Buddy (Peltier and Fiorini, 2017), designed for social interaction and domestic assistance; and Aldebaran Robotics' NAO (Gouaillier et al., 2009), recognized for its versatility in research and educational settings. These robots showcase the diversity of roles humanoid robots can play, from companionship and entertainment to education and beyond.

Developing emotional intelligence in robots is relevant as they increasingly participate in social settings. Indeed, beyond performing physical tasks, enhancing robots' ability to perceive, interpret and respond to human needs is essential for effective social Human-Robot Interaction (HRI) and Human-Robot Collaboration (HRC).

In the realm of social robotics, integrating sensors such as microphone for ”mouth” or camera for ’eyes' into the humanoid robot, enables the robot to capture human emotions in real-time, and to adapt its response and behavior accordingly (Justo et al., 2020; Olaso et al., 2021; Palmero et al., 2023). This capability enhances their utility in various applications and facilitates engagement and intuitive interaction experiences between robots and humans. Detecting emotions from camera starts with face detection, which involves identifying and locating human faces within images or video frames. This process includes preprocessing images, extracting distinct facial features, classifying regions as faces or non-faces, refining detection accuracy, and handling variations in lighting, occlusions, poses and scales. Face emotion recognition (FER) employs computer vision and machine learning techniques to analyze human emotions from face.

Often, emotion recognition systems deals with only one user while he is communicating with a machine. However, multiple users can communicate si-

[a] https://orcid.org/0000-0002-0474-3229
[b] https://orcid.org/0000-0002-2435-3017

Yahyaoui, M. A., Oujabour, M., Ben Letaifa, L. and Bohi, A.
Multi-Face Emotion Detection for Effective Human-Robot Interaction.
DOI: 10.5220/0013170300003890
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 17th International Conference on Agents and Artificial Intelligence (ICAART 2025) - Volume 1*, pages 91-99
ISBN: 978-989-758-737-5; ISSN: 2184-433X

multaneously with it. Multi-face emotion recognition is particularly valuable across various scenarios. For instance, at a comedy club, it provides real-time feedback to comedians, manages lighting and sound, interacts with the audience, and detects disruption.

In this work, we present a complete facial emotion recognition interface and its deployment in a mobile humanoid robot. The proposed interface can display emotions from multiple individuals in real-time within an advanced user interface. To achieve this, several deep neural network models have been developed and evaluated under the same conditions. Then a tradeoff between system accuracy and model size have been considered in order to implement the optimal solution into a humanoid robot. The model's performance and its confidence interval also guided this choice of solution.

The remainder of this paper is structured as follows. Section 2 reviews the state of the art related to emotion detection for Human-Robot Interaction (HRI) and Facial Emotion Recognition (FER) systems. Section 3 presents the design and implementation of the proposed emotional interface, detailing the multi-face detection, emotion recognition system, and the graphical user interface. Section 4 describes the integration of the facial emotion recognition system into the Tiago++ humanoid robot, highlighting the processes of face tracking and real-time emotion detection. Section 5 outlines the experimental setup and presents the results, including performance metrics, model comparisons, and user interaction analysis. Finally, Section 6 concludes the paper with a discussion of the findings, limitations of the current approach, and potential directions for future work.

2 RELATED WORK

Although emotions have been investigated in the context of HRI, it remains a significant challenge. In this section, we report recent research in HRI as well as FER systems.

2.1 Emotion Detection for HRI

In social robotics, emotion detection is mimicked by robots to interact naturally and harmoniously with humans. Several studies have focused on implementing facial emotion recognition in robots. For instance, the study (Zhao et al., 2020) applied facial emotion recognition on three datasets: FER2013, FERPLUS and FERFIN. The system was implemented on a NAO robot, which responds with actions based on the detected emotions. However, this study has some limitations, as it does not provide details on the robot's implementation. Additionally, the research (Dwijayanti et al., 2022) integrated a facial detection system with a facial emotion recognition system and implemented it in a robot. They also explored automatic detection of the distance between the camera of the robot and the person. One drawback is that the robot is stationary, so mobility is not considered. The study (Spezialetti et al., 2020) serves as a survey of emotion recognition research for human-robot interaction. It reviews emotion recognition models, datasets, and modalities, with a particular emphasis on facial emotion recognition. However, it does not include any research utilizing deep learning models for facial emotion recognition.

2.2 Facial Emotion Recognition

Deep learning has revolutionized computer vision tasks, including Facial Emotion Recognition (FER), with numerous studies proposing various methodologies to achieve high classification accuracy using well known benchmark datasets (Farhat et al., ; Goodfellow et al., 2013; Letaifa et al., 2019; Mollahosseini et al., 2017; Justo et al., 2021; Lucey et al., 2010).

Several recent studies have proposed innovative approaches for FER. Farzaneh et al. (Farzaneh and Qi, 2021) introduced the Deep Attentive Center Loss (DACL) method, which integrates an attention mechanism to enhance feature discrimination, showing superior performance on RAF-DB and AffectNet datasets. Similarly, Pecoraro et al. (Pecoraro et al., 2022) proposed the LHC-Net architecture, which employs a multi-head self-attention module tailored for FER tasks, achieving state-of-the-art results on FER2013 with lower computational complexity. In another work, Han et al. (Han et al., 2022) presented a triple-structure network model based on MobileNet V1, which captures inter-class and intra-class diversity features, demonstrating strong results on KDEF, MMI, and CK+ datasets. Fard et al. (Fard and Mahoor, 2022) introduced the Adaptive Correlation (Ad-Corre) Loss, which improved performance on AffectNet, RAF-DB, and FER2013 datasets when applied to Xception and ResNet50 models. Other notable contributions include the Segmentation VGG-19 model (Vignesh et al., 2023), which enhanced FER on FER2013 using segmentation-inspired blocks, and the DDAMFN network by Zhang et al. (Zhang et al., 2023), which incorporated dual-direction attention to achieve excellent results on AffectNet and FERPlus. Lastly, in our recent work, we introduced EmoNeXt (El Boudouri and Bohi, 2023), a deep learning framework that has set new state-of-the-art benchmarks on the FER2013 dataset. EmoNeXt integrates a Spatial Transformer Network (STN) for handling facial alignment variations, along with Squeeze-and-Excitation (SE) blocks for channel-wise feature recalibration. Additionally, a self-attention regularization term was introduced to enhance compact feature generation, further improving accuracy.

This brief review shows that many FER models have focused exclusively on improving accuracy. As a result, today's leading models can reach memory sizes in the order of gigabytes, which poses challenges for deployment in memory-constrained environments, such as the robots.

3 THE EMOTIONAL INTERFACE

One of the challenges in the domain of emotion detection for HRI, is the simultaneous detection of emotions from multiple faces, which is useful where robots interact with groups of people.

3.1 Multi-Face Detection

We choose the Haarcascade classifier, proposed by Paul Viola and Michael Jones in their seminal paper (Viola and Jones, 2001), as a highly effective method for face detection. Other notable methods include the Histogram of Oriented Gradients (HOG) combined with Support Vector Machines (SVM) and deep learning approaches such as the Multi-task Cascaded Convolutional Networks (MTCNN). While these methods have shown promising results in various applications, the Haarcascade classifier is particularly advantageous for real-time scenarios.

The general principle of the Haarcascade approach is illustrated in Figure 1. This machine learning-based method involves training a cascade function using a large dataset of positive (face) and negative (non-face) images. The classifier relies on Haar features, which are similar to convolutional kernels, to extract distinguishing characteristics from images. Each Haar feature is a single value calculated by subtracting the sum of pixels under a white rectangle from the sum of pixels under a black rectangle. To efficiently compute these features, the concept of integral images is utilized, reducing the calculation to an operation involving just four pixels, regardless of the feature's size.

During training, all possible sizes and positions of these features are applied to the training images, resulting in over 160,000 potential features. To select the most relevant features, the AdaBoost algorithm is utilized, which iteratively adjusts the weights of misclassified images and selects features with the lowest error rates, thereby creating a strong classifier from a combination of weak classifiers. Despite the high initial number of features, this process narrows it down significantly (e.g., from 160,000 to around 6,000).

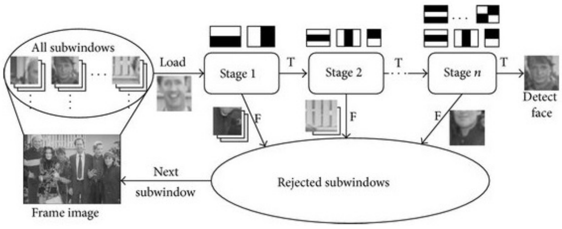


Figure 1: Cascade structure for Haar classifiers (Kim et al., 2015).

For detection, the image is scanned with a 24x24 pixel window, applying these selected features. To enhance efficiency, the authors introduced a cascade of classifiers. This means that features are grouped into stages, and if a window fails at any stage, it is immediately discarded as a non-face region. This hierarchical approach ensures that only potential face regions undergo the full, more complex evaluation process, allowing for real-time face detection with high accuracy.

3.2 Emotion Recognition System

Pretrained deep learning models have demonstrated exceptional effectiveness for feature extraction across various domains (Palmero et al., 2023). In our emotion recognition system (Figure. 2), we leverage a pretrained convolutional neural network (CNN) model to apply transfer learning using the FER2013 dataset (Goodfellow et al., 2013). Specifically, we utilize pretrained CNN models, initially trained on the ImageNet dataset which encompass millions of images from various categories (Deng et al., 2009). This extensive training enables these models to extract highly relevant and general visual features through their convolutional layers. These layers detect fundamental elements such as edges, textures, and shapes, which are essential for understanding facial structures. We utilize these convolutional layers to process our input images, leaving out the top portion of the model, specifically the fully connected layers initially designed for the ImageNet classification tasks. Instead, by passing our facial images through the pretrained model's convolutional layers, we generate a feature stack that encapsulates essential visual information. This feature stack, representing a rich set of features extracted from the images, is then flattened into a format suitable for further processing. Subsequently, we introduce additional fully connected layers tailored to the FER2013 dataset to recognize and classify seven distinct emotions: anger, disgust, fear, happiness, sadness, surprise, and neutrality.

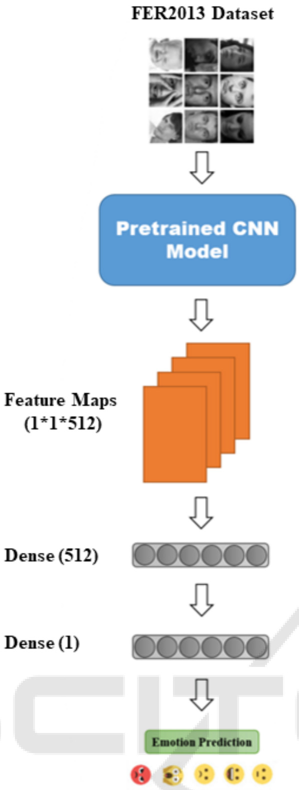


Figure 2: The architecture of the emotion recognition system using transfer learning on the FER2013 dataset.

These newly added layers are trained to fine-tune the model specifically for emotion recognition, leveraging the robust feature extraction capabilities of the pretrained model's convolutional layers.

3.3 Graphical Interface

The graphical interface of our emotion recognition system integrates multiple advanced technologies to provide a seamless and responsive user experience. Upon launching the application, the interface is built using the Tkinter library [1], creating a user-friendly graphical environment. The system activates the webcam through the OpenCV library [2], capturing a live video feed for real-time analysis. Captured video frames undergo face detection using the HaarCascade classifier, a robust method for identifying faces under various lighting conditions and angles (see description in subsection 3.1).

Once a face is detected, the region of interest is extracted and subjected to preprocessing to ensure compatibility with the model's input size. The processed image is then fed into a pretrained CNN model that have been fine-tuned on the FER2013 dataset. This model analyzes the facial image to predict the user's emotional state, categorizing it into distinct emotions such as anger, fear, disgust, happiness, sadness, surprise, and neutrality. The predicted emotion is then displayed on the graphical interface, providing immediate feedback to the user. All these steps are illustrated by Fig. 3.

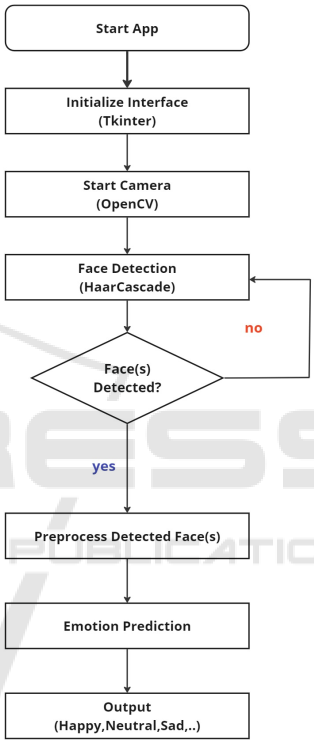

Start App
Initialize Interface (Tkinter)
Start Camera (OpenCV)
Face Detection (HaarCascade)
Face(s) Detected?
no
yes
Preprocess Detected Face(s)
Emotion Prediction
Output (Happy,Neutral,Sad,..)

Figure 3: Global architecture of our real-time multi-face emotion recognition user interface.

4 THE HUMANOID ROBOT

The Tiago robot, developed by PAL Robotics (Pages et al.,), is a humanoid mobile robot [3]. Its modular design allows for customization to meet specific needs. In this section, we outline our approach to equipping the Tiago++ model of the robot with face emotion recognition capabilities. By using the Robot Operating System (ROS)[4] for communication and processing, and integrating a Tkinter-based GUI for real-time visualization, we enhance the ability of the robot to

[1] https://docs.python.org/3/library/tkinter.html
[2] https://docs.opencv.org/4.x/
[3] https://pal-robotics.com/robots/tiago/
[4] https://wiki.ros.org

interact with humans. This implementation is divided into two primary tasks: face tracking and emotion detection, each described in the following subsections.

4.1 Face Tracking Integration on Tiago++ Robot

We implemented a face tracking module on the Tiago robot by integrating ROS with a Tkinter-based GUI application. The process begins with initializing a ROS node named `Tiago_FER` and setting up essential publishers and subscribers to facilitate communication between the robot and the software. We use the `CvBridge` [5] library to convert images from ROS format to OpenCV format. Meanwhile, the `MediaPipeRos` instance processes these images to detect regions of interest (ROI) for face tracking. The application's main loop receives images from the robot's camera through the `/xtion/rgb/image` ROS topic, processes these images to detect faces, and generates commands to adjust the robot's yaw and pitch. These commands, which control head movements, are published to the `head_controller/increment/goal` topic using the `IncrementActionGoal` message type, enabling the robot to track the detected faces. These steps are outlined in the diagram generated by ROS, as shown in Figure 4.

4.2 Emotion Detection and GUI Display on Tiago++ Screen

Following face tracking, the processed images are analyzed to predict emotions. The detected emotions are displayed on a Tkinter GUI, which features a canvas for image display and progress bars to visualize emotion scores. The processed images and emotion data are published back to the `/imagesBack` ROS topic. Additionally, incremental commands for torso movements are sent to the `/torso_controller/safe_command` topic using the `JointTrajectory` message type, allowing the robot to dynamically respond to detected emotions (see Figure 4.

5 EXPERIMENTS AND RESULTS

Developing a human-robot interface for FER involves detecting faces and emotions, implementing the user interface, and integrating it into the robot platform. The robot's camera captures images of individuals interacting with it, processes these images to detect emotions, and then displays the detected emotions on the user interface. This interface is visible on the tablet mounted on the robot's chest. Several challenges are to be addressed, particularly focusing on the accuracy of the models and the feasibility of implementing them on the robot.

5.1 Face Emotion Detection

In this work, we fine-tuned several pretrained models from the Keras library[6], initially trained on the ImageNet 1000K dataset. These models were selected based on their strong performance in the ImageNet classification task and their ability to generalize well for FER tasks. We applied transfer learning, as explained in subsection 3.2, to the following models: MobileNet (Howard et al., 2017), DenseNet201 (Huang et al., 2017), ResNet152V2 (He et al., 2016b), ResNet101 (He et al., 2016a), Xception (Chollet, 2017), EfficientNetV2-B0 (Tan and Le, 2021), InceptionResNetV2 and InceptionV3 (Szegedy et al., 2017), VGG16 and VGG19 (Karen, 2014), and ConvNeXt (from Tiny to XLarge version) (Liu et al.,).

For training, we consistently used data augmentation techniques such as rotation, shift, zoom, horizontal flip and adjustments in brightness and contrast to improve the model's robustness. Additionally, Random Erasing was used to simulate occlusions, while resizing and recropping variations improved robustness to differences in face positioning. The models were optimized using Adam with a learning rate of 0.0001, combined with strategies like EarlyStopping and ReduceLROnPlateau to prevent overfitting and dynamically adjust the learning rate.

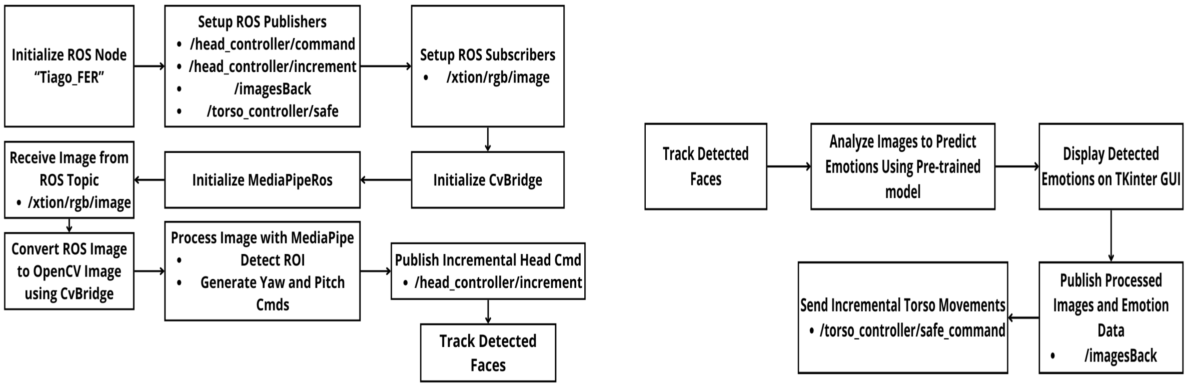
The accuracy and memory footprint of each fine-tuned model on the FER2013 dataset are reported in Table 1. While ConvNeXt XLarge achieved the highest accuracy at 72.27%, it comes with a significantly larger memory footprint than the other models.

5.2 Confidence Interval

Accuracy is an estimate of the performance of a system, and its reliability depends on the number of tests conducted that is in our case the number of emotions to be recognized. The measurement of the confidence interval is introduced to assess the trustability of our recognition rate. In (Zouari, 2007), the successes are modeled by a binomial distribution. If N is the number of tests and P is the recognition rate, then the confidence interval [P-, P+] at x% is:

[5] https://wiki.ros.org/cv_bridge/Tutorials

[6] https://keras.io/api/applications/


Initialize ROS Node "Tiago_FER"
Setup ROS Publishers
• /head_controller/command
• /head_controller/increment
• /imagesBack
• /torso_controller/safe
Setup ROS Subscribers
• /xtion/rgb/image
Receive Image from ROS Topic
• /xtion/rgb/image
Initialize MediaPipeRos
Initialize CvBridge
Convert ROS Image to OpenCV Image using CvBridge
Process Image with MediaPipe
• Detect ROI
• Generate Yaw and Pitch Cmds
Publish Incremental Head Cmd
• /head_controller/increment
Track Detected Faces
Track Detected Faces
Analyze Images to Predict Emotions Using Pre-trained model
Display Detected Emotions on TKinter GUI
Send Incremental Torso Movements
• /torso_controller/safe_command
Publish Processed Images and Emotion Data
• /imagesBack

(a) Face tracking integration on Tiago++ robot.

(b) Emotion detection and GUI display on Tiago++ robot.

Figure 4: ROS-based Tiago++ face emotion recognition integration process: the diagram in the left (a) depicts the steps involved in face tracking integration, while the diagram in the right (b) shows the emotion detection and GUI display process.

Table 1: Pretrained models fine-tuned on the FER2013 dataset: accuracy (%) and memory footprint (Megabytes).

Model name	Accuracy	Model size
MobileNet	66.11	14.5
ResNet152V2	67.28	611.3
DenseNet201	67.84	221.0
InceptionV3	68.43	268.6
Xception	68.93	346.9
ConvNeXt Tiny	69.43	362
EfficientNetV2-B0	70.00	139.0
ConvNeXt Small	70.15	566
InceptionResNetV2	70.29	648.2
ConvNeXt Base	70.32	1120
VGG16	71.18	171.0
ResNet101	71.30	549.8
VGG19	71.46	262.5
ConvNeXt Large	71.57	2733
ConvNeXt XLarge	**72.27**	**3900**

$$P\pm = \frac{P + \frac{z_x^2}{N} \pm z_x\sqrt{\frac{P(1-P)}{N} + \frac{z_x^2}{4N^2}}}{1 + \frac{z_x^2}{N}}$$

with z95% = 1.96 and z98% = 2.33. This means that there is a x% chance that the rate falls within the interval [P-, P+].

The FER2013 dataset consists in 35,887 grayscale images, divided into training (80%), test (10%) and validation (10%). Hence, using each model 3589 samples have been evaluated on the test set. We compute the confidence interval with z98 for all models and report the results in Figure 5. We notice that several models, including VGG16, InceptionResNetV2, ConvNeXt Base, EfficientNetV2-B0, and VGG19, show overlapping results. In terms of precision, these models demonstrate similar efficiency. However, there is a notable difference in their sizes, with EfficientNetV2-B0 being the most compact. Due to its smaller size, EfficientNetV2-B0 has been chosen for implementation on the robot.

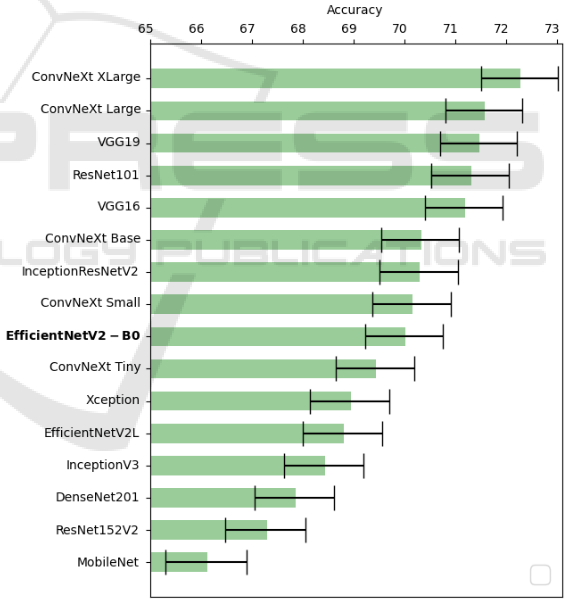


Figure 5: Accuracy and confidence intervals of the models.

5.3 User Interface Development

Our emotion recognition application features an intuitive and user-friendly graphical interface designed for both single-face and multi-face emotion detection. The interface allows users to utilize their device's camera to capture live video streams, which are then processed in real-time to detect and classify facial expressions. For single-face emotion recognition, the application highlights the detected face and displays the identified emotion with corresponding confidence levels. In multi-face scenarios, the interface efficiently detects multiple faces within the same frame, assigning emotions to each detected face individually. The results are visually presented using bounding boxes and emotion labels directly on the video feed, providing clear and immediate feedback.

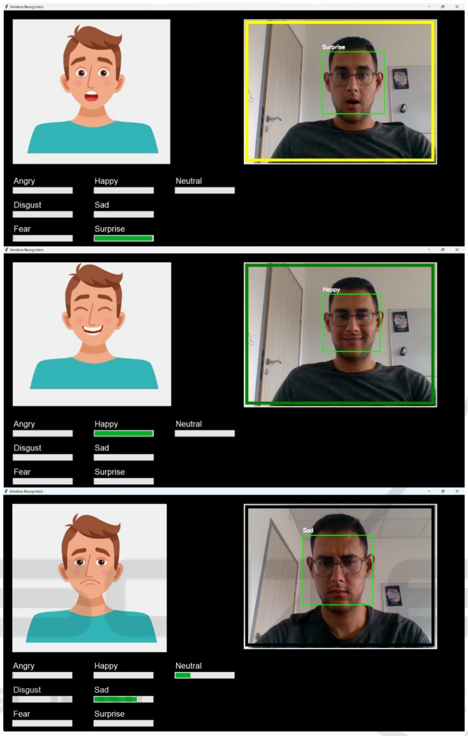


Figure 6: The user interface displays face and emotion detection for a single person. Progress bars indicate the confidence score for each recognized emotion.

Additionally, the interface includes progress bars for the detected emotion, visually representing the confidence level of each prediction. An avatar further enhances user interaction by imitating the predicted emotion in real-time, offering an engaging and dynamic way to understand the results. This comprehensive and interactive interface ensures that users can easily interpret the emotion detection outcomes, making the application practical for various real-world settings, including human-robot interaction and affective computing. Figure. 6 and Figure. 7 show some examples of the user interface applied to single and multi-face emotion detection.

5.4 FER Deployment on Tiago++ Robot

The Tiago++ is a humanoid mobile robot with constrained resources (CPU, memory, and storage). Besides interacting with humans, the robot must concurrently perform critical tasks such as navigation and detection, which are also resource-intensive. Consequently, for deploying our application on the Tiago++ robot, it is essential to select a model not only based on its test accuracy but also on the memory footprint of the model. The Tiago++ robot has a maximum capacity of about 150 MB for model files to ensure real-time inference without disrupting other processes running on the robot. According to Table 1 and the previous subsection, EfficientNetV2-B0 stands out with a good balance between accuracy (70.00%) and model size (139 MB), meeting the robot's constraints.

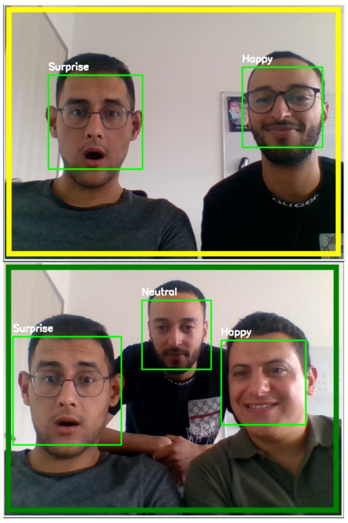


Figure 7: Face detection is followed by emotion detection for multiple individuals present in the same image.

To illustrate the system's effectiveness, we conducted two sets of experiments. In the first set, a single participant interacted with the robot, displaying a range of emotions. The system's ability to accurately detect the face and classify the emotional state of the participant in real-time was meticulously observed and documented. In the second set, two participants were present simultaneously, engaging in various interactions with the robot. This scenario tested the system's robustness in detecting multiple faces and correctly identifying each individual's emotional state in real-time. The results of these experiments are depicted through a series of images captured during the interactions on Figure. 8.

6 CONCLUSIONS

In this paper, we presented a facial emotion detection interface implemented on a mobile humanoid robot. This interface is capable of displaying emotions from multiple individuals in real-time video. To achieve

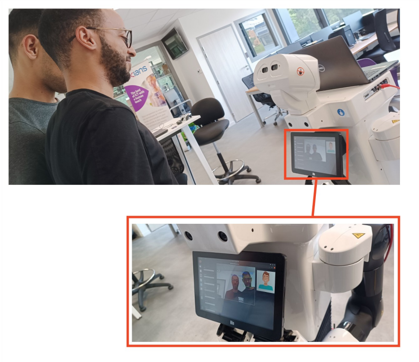


Figure 8: Multi face emotion detection deployed on robot

this, we developed and evaluated several deep neural network models under consistent conditions, carefully considering factors such as model size and accuracy to ensure compatibility with both personal computers and mobile robots like the Tiago++.

While our system demonstrates strong performance, it is important to note the limitations of relying solely on facial expressions for emotion detection, particularly in contexts where communication may be impaired. Emotions are complex and multifaceted, often requiring the integration of multiple modalities for more accurate recognition. Therefore, future work will focus on incorporating additional modalities, such as voice, text, gestures, and biosignals, to enhance the performance and reliability of emotion recognition systems. Additionally, we will focus on optimizing large models used in FER tasks to ensure their efficiency for deployment on the Tiago++ robot, considering the balance between model size and accuracy.

REFERENCES

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255.

Dwijayanti, S., Iqbal, M., and Suprapto, B. Y. (2022). Real-time implementation of face recognition and emotion recognition in a humanoid robot using a convolutional neural network. *IEEE Access*, 10:89876–89886.

El Boudouri, Y. and Bohi, A. (2023). Emonext: an adapted convnext for facial emotion recognition. In *2023 IEEE 25th International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6. IEEE.

Fard, A. P. and Mahoor, M. H. (2022). Ad-corre: Adaptive correlation-based loss for facial expression recognition in the wild. *IEEE Access*, 10:26756–26768.

Farhat, N., Bohi, A., Letaifa, L. B., and Slama, R. Cg-mer: a card game-based multimodal dataset for emotion recognition. In *Sixteenth International Conference on Machine Vision (ICMV 2023)*.

Farzaneh, A. H. and Qi, X. (2021). Facial expression recognition in the wild via deep attentive center loss. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2402–2411.

Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., Lee, D.-H., et al. (2013). Challenges in representation learning: A report on three machine learning contests. In *Neural information processing. 20th international conference, ICONIP*. Springer.

Gouaillier, D., Hugel, V., Blazevic, P., and Kilner, C. (2009). Mechatronic design of nao humanoid. In *IEEE International Conference on Robotics and Automation ICRA*.

Han, B., Hu, M., Wang, X., and Ren, F. (2022). A triple-structure network model based upon mobilenet v1 and multi-loss function for facial expression recognition. *Symmetry*, 14(10):2055.

He, K., Zhang, X., Ren, S., and Sun, J. (2016a). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

He, K., Zhang, X., Ren, S., and Sun, J. (2016b). Identity mappings in deep residual networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 630–645. Springer.

Hirose, M. and Ogawa, K. (2007). Honda humanoid robots development. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 365(1850):11–19.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Justo, R., Letaifa, L. B., Olaso, J. M., López-Zorrilla, A., Develasco, M., Vázquez, A., and Torres, M. I. (2021). A spanish corpus for talking to the elderly. *Conversational Dialogue Systems for the Next Decade*.

Justo, R., Letaifa, L. B., Palmero, C., Fraile, E. G., Johansen, A., Vazquez, A., Cordasco, G., Schlogl, S., Ruanova, B. F., Silva, M., Escalera, S., Velasco, M. D., Laranga, J. T., Esposito, A., Kornes, M., and Torres, M. I. (2020). Analysis of the interaction between elderly people and a simulated virtual coach. *Journal of Ambient Intelligence and Humanized Computing*, 11:6125–6140.

Karen, S. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv: 1409.1556*.

Kim, M., Lee, D., and Kim, K.-Y. (2015). System architecture for real-time face detection on analog video camera. *International Journal of Distributed Sensor Networks*, 11(5):251386.

Letaifa, L. B., Develasco, M., Justo, R., and Torres, M. I. (2019). First steps to develop a corpus of interactions between elderly and virtual agents in spanish with emotion. In *International Conference on Statistical Language and Speech Processing*.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., and Xie, S. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., and Matthews, I. (2010). The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 ieee computer society conference on computer vision and pattern recognition-workshops*, pages 94–101. IEEE.

Mollahosseini, A., Hasani, B., and Mahoor, M. H. (2017). Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31.

Olaso, J., Vázquez, A., Letaifa, L. B., de Velasco, M., Mtibaa, A., Hmani, M. A., Petrovska-Delacrétaz, D., Chollet, G., Montenegro, C., López-Zorrilla, A., et al. (2021). The empathic virtual coach: a demo. In *The 2021 International Conference on Multimodal Interaction (ICMI'21)*, pages 848–851. ACM.

Pages, J., Marchionni, L., and Ferro, F. Tiago: the modular robot that adapts to different research needs. In *International workshop on robot modularity, IROS*.

Palmero, C., DeVelasco, M., Hmani, A., Mtibaa, M. A., Letaifa, L. B., et al. (2023). Exploring emotion expression recognition in older adults interacting with a virtual coach. *arXiv preprint arXiv:2311.05567*.

Pecoraro, R., Basile, V., and Bono, V. (2022). Local multi-head channel self-attention for facial expression recognition. *Information*, 13(9):419.

Peltier, A. and Fiorini, L. (2017). Buddy: A companion robot for living assistance. *Journal of Robotics and Automation*, 3(2):75–81.

Spezialetti, M., Placidi, G., and Rossi, S. (2020). Emotion recognition for human-robot interaction: Recent advances and future perspectives. *Frontiers in Robotics and AI*, 7:145.

Szegedy, C., Ioffe, S., Vanhoucke, V., and Alemi, A. (2017). Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Tan, M. and Le, Q. (2021). Efficientnetv2: Smaller models and faster training. In *International conference on machine learning*, pages 10096–10106. PMLR.

Vignesh, S., Savithadevi, M., Sridevi, M., and Sridhar, R. (2023). A novel facial emotion recognition model using segmentation vgg-19 architecture. *International Journal of Information Technology*, pages 1–11.

Viola, P. and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Proceedings of the IEEE computer society conference on computer vision and pattern recognition. CVPR*.

Zhang, S., Zhang, Y., Zhang, Y., Wang, Y., and Song, Z. (2023). A dual-direction attention mixed feature network for facial expression recognition. *Electronics*, 12(17):3595.

Zhao, G., Yang, H., Tao, Y., Zhang, L., and and, C. Z. (2020). Lightweight cnn-based expression recognition on humanoid robot. *KSII Transactions on Internet and Information Systems*, 14(3):1188–1203.

Zouari, L. (2007). *Vers le temps réel en transcription automatique de la parole grand vocabulaire*. PhD thesis, Télécom ParisTech.