UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Ciencias e Ingenierías

Application of Convolutional Neural Networks to Emotion Recognition for Robotic Arm Manipulation

Walter Marcelo Fuertes Encalada
Karen Madelein Hunter Ordóñez

Ingeniería en Electrónica y Automatización

Trabajo de fin de carrera presentado como requisito
para la obtención del título de
Ingeniería en Electrónica

Quito, 19 de mayo de 2023

UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Ciencias e Ingenierías

HOJA DE CALIFICACIÓN DE TRABAJO DE FIN DE CARRERA

Application of Convolutional Neural Networks to Emotion Recognition for Robotic Arm Manipulation

Walter Marcelo Fuertes Encalada

Karen Madelein Hunter Ordóñez

Nombre del profesor, Título académico **Diego Benítez, Ph.D.**

Quito, 19 de mayo de 2023

© DERECHOS DE AUTOR

Por medio del presente documento certifico que he leído todas las Políticas y Manuales de la Universidad San Francisco de Quito USFQ, incluyendo la Política de Propiedad Intelectual USFQ, y estoy de acuerdo con su contenido, por lo que los derechos de propiedad intelectual del presente trabajo quedan sujetos a lo dispuesto en esas Políticas.

Asimismo, autorizo a la USFQ para que realice la digitalización y publicación de este trabajo en el repositorio virtual, de conformidad a lo dispuesto en la Ley Orgánica de Educación Superior del Ecuador.

Nombres y apellidos: Walter Marcelo Fuertes Encalada

Código: 207095

Cédula de identidad: 1722816608

Lugar y fecha: Quito, 19 de mayo de 2023

Nombres y apellidos: Karen Madelein Hunter Ordóñez

Código: 206895

Cédula de identidad: 1726513243

Lugar y fecha: Quito, 19 de mayo de 2023

ACLARACIÓN PARA PUBLICACIÓN

Nota: El presente trabajo, en su totalidad o cualquiera de sus partes, no debe ser considerado como una publicación, incluso a pesar de estar disponible sin restricciones a través de un repositorio institucional. Esta declaración se alinea con las prácticas y recomendaciones presentadas por el Committee on Publication Ethics COPE descritas por Barbour et al. (2017) Discussion document on best practice for issues around theses publishing, disponible en http://bit.ly/COPETheses.

UNPUBLISHED DOCUMENT

Note: The following capstone project is available through Universidad San Francisco de Quito USFQ institutional repository. Nonetheless, this project – in whole or in part – should not be considered a publication. This statement follows the recommendations presented by the Committee on Publication Ethics COPE described by Barbour et al. (2017) Discussion document on best practice for issues around theses publishing available on http://bit.ly/COPETheses.

RESUMEN

Este documento presenta el desarrollo de un sistema que opera un brazo robótico para entregar un objeto basado en la expresión facial de una persona parada frente al robot, demostrando el reconocimiento de emociones en tiempo real para la interacción física humano-robot. Para lograr esto, se desarrolló un modelo basado en redes neuronales convolucionales para identificar emociones en tiempo real. El funcionamiento del brazo robótico se implementó utilizando una computadora integrada NVidia Jetson Nano, una cámara web y bibliotecas OpenCV, ROS y TensorFlow. Utilizando un conjunto de datos de 26,6k fotos de rostros de la base de datos de detección de emociones, el modelo de detección de emociones construido demostró una precisión del 93,5 % y un error del 6,5 % durante el entrenamiento y la validación. El prototipo final en tiempo real tuvo una precisión de prueba del 94 % con un error del 6 %. Esta prueba de concepto muestra que, en un futuro próximo, también se podrán crear aplicaciones más avanzadas que aprovechen las emociones de los usuarios.

Palabras clave: reconocimiento de emociones, redes neuronales convolucionales, control de brazo robótico

ABSTRACT

This paper presents the development of a system that operates a robotic arm to deliver an object based on the facial expression of a human standing in front of the robot, demonstrating real-time emotion recognition for physical Human-Robot Interaction. To achieve this, a convolutional neural network-based model was developed to identify emotions in real- time. The operation of the robotic arm was implemented using an embedded computer NVidia Jetson Nano, a web camera, and OpenCV, ROS, and TensorFlow libraries. Using a dataset of 26.6k face photos from the Emotion Detection database, the built emotion detection model demonstrated an accuracy of 93.5% and an error of 6.5% during training and validation. The final real-time prototype had a testing accuracy of 94% with a 6% error. This proof-of-concept shows that in the near future more advanced applications that harness user emotions may also be built.

Key words: emotion recognition, convolution neural networks, robotic arm control

TABLA DE CONTENIDO

II. INTRODUCTION 10

II. MATERIALS AND METHOD 10

A. Hardware 11

B. Emotion Detection database 11

C. Deep-Learning Model 11

D. Face Detection 11

E. Robotic Arm Control 11

III. EXPERIMENTATION AND RESULTS 11

A. Experimental Setup 11

1. Model configuration 11

2. Image preprocessing 13

3. Training and test partitions 13

4. Assessment metrics 13

5. Real-time test 13

B. Results and Discussion 13

III. CONCLUSIONS AND FUTURE WORK 14

ACKNOWLEDGMENT 14

REFERENCES 14

ÍNDICE DE TABLAS

TABLE I. Performance of the proposed model for binary emotion classification................. 14

ÍNDICE DE FIGURAS

Figure 1. Prototype setup 12

Figure 2. Example of the images of sadness and happiness expressions available within the emotion detection database 12

Figure 3. Deep-Learning Model 13

Figure 4. Example of an image of a detected person with a green frame drawn around their face, indicating the classified expression and its corresponding prediction confidence value 13

Figure 5. Experimental setup at the USFQ campus for real-time testing of the prototype 14

Figure 6. Example of Real-Time testing scenarios 14

Figure 7. Model performance in solving the binary classification during training and validation 14

Figure 8. Real-time prediction examples, a green frame is drawn around the face, indicating the classified expression and its corresponding prediction confidence value 15

Application of Convolutional Neural Networks to Emotion Recognition for Robotic Arm Manipulation

Walter Fuertes, Karen Hunter, Diego Benítez, Noel Peréz, Felipe Grijalva and Maria Baldeon-Calisto[†]
Colegio de Ciencias e Ingenierías “El Politécnico”,
[†]Ingeniería Industrial, CATENA-USFQ,
Universidad San Francisco de Quito USFQ, Quito 170157, Ecuador
email: {wmfuertes, kmhunter}@estud.usfq.edu.ec, {dbenitez, nperez, fgrijalva, mbaldeonc}@usfq.edu.ec

***Abstract*—This paper presents the development of a system that operates a robotic arm to deliver an object based on the facial expression of a human standing in front of the robot, demonstrating real-time emotion recognition for physical Human-Robot Interaction. To achieve this, a convolutional neural network-based model was developed to identify emotions in real-time. The operation of the robotic arm was implemented using an embedded computer NVidia Jetson Nano, a web camera, and OpenCV, ROS, and TensorFlow libraries. Using a dataset of 26.6k face photos from the Emotion Detection database, the built emotion detection model demonstrated an accuracy of 93.5% and an error of 6.5% during training and validation. The final real-time prototype had a testing accuracy of 94% with a 6% error. This proof-of-concept shows that in the near future more advanced applications that harness user emotions may also be built.**

***Keywords*—emotion recognition, convolution neural networks, robotic arm control**

I. Introduction

The evolution of technology in the modern age focuses on the interaction between humans and machines. These devices are intended to facilitate procedures and ensure human safety. The complexity level of a machine is determined by the sophistication of the task it can perform. With the advancement of technology, machines can now communicate more efficiently with humans, enabling them to identify human needs and provide a richer interaction. The applications of this technology range from machines that respond to voice commands to autonomous robots. Artificial Intelligence (AI) is a technology that makes this possible by allowing machines to mimic human behavior. Humanoid robots are anticipated to coexist with humans in the future as companions or coworkers.

Facial expression recognition (FER) holds significance in multiple interactive computing domains, including human-robot interaction, social robots, and satisfaction surveys. The term physical human-robot interaction (pHRI) is defined in [1] as an interaction in which the robot may physically act on a person without previous warning, which can occur in various circumstances. Although physical safety is of utmost importance in such scenarios, it is also necessary to consider the user’s psychological and mental states to ensure a comfortable experience in both working and domestic environments. For example, the automotive industry uses advanced driver assistance systems to aid drivers. These systems can support different functions for safe driving and evaluate drivers’ ability to drive stably and safely. Numerous studies have indicated that a driver’s emotions are crucial in managing their behavior, which can result in serious accidents. Hence, monitoring drivers’ emotions can help forecast their behavior and prevent accidents. An architecture for driver emotion detection is proposed in [2].

Emotion recognition in HRI can also be observed in social robots. In [3], this is examined under the heading of multimodal expressions, although the recognition process still has some limitations. This is frequently encountered in social robotics, owing to robots’ vast range of sensory capabilities. Consequently, more adaptable models are necessary to improve performance. Furthermore, connecting social robots to the Internet of Things (IoT) and cloud services is possible, resulting in a comprehensive hybrid-face affective robotic system capable of displaying human-like facial emotions, as described in [4].

Recently, numerous studies have been proposed to improve the robustness and accuracy of FER [5]. However, there is still significant scope for enhancing the performance and robustness of FER techniques. One of the most challenging research areas is recognizing facial expressions in the wild. In order to achieve high-level performance in FER, a considerable number of well-aligned and high-resolution face images are required. Nevertheless, compared to face images collected in a controlled environment, in-the-wild face images exhibit significant variations, such as diverse head poses and illumination.

Regarding neural network models, Convolutional Neural Networks are advised (CNN) for artificial vision [6]. In [7], human-robot interactions are examined by directing a robotic arm to deliver an object based on the gender identity (male or female) of a human standing in front of the robot. The forecast utilizes data captured by a webcam in image format and a pre-trained CNN with three layers is employed to extract facial features, and two layers are used for classification.

This study explores the interaction between humans and robots using a pre-trained deep learning-based model to control a robotic arm to deliver an object based on a human’s detected emotion (sadness/neutral or happiness) in front of the robot. In order to achieve this, a CNN is utilized to analyze the image captured by a webcam. The CNN is implemented in a Jetson Nano-embedded computer integrated into the robotic arm. This results in a responsive machine that can process information in real-time and take appropriate actions based on the analyzed data. This emotion detection-based application is the first step towards developing more sophisticated applications, where machines can be programmed to perform different actions based on the user's emotions.

II. Materials and Method

A. Hardware

The proposed system prototype consists of several hardware components, as depicted in Fig. 1. One of the main components is a 6-axis robotic arm, specifically the Yahboom DOFBOT AI Vision Robotic Arm [8]. This arm is constructed primarily from an aluminum alloy, which provides strength without adding too much weight. In order to control the six servo motors that power the arm, the kit comes with a custom expansion board that is interfaced with an NVIDIA Jetson Nano board [9]. The Jetson Nano executes the trained CNN model for the system. In addition, the kit includes a USB camera, which can be used with OpenCV [10] to detect objects. However, a Logitech Brio 4K Webcam [11] was used instead of the original camera for this application. The webcam offers auto light correction and a wider field of view, which benefits the system. To control the six motors of the robotic arm, the Robot Operating System (ROS) [12] was utilized as an open-source robotics middleware suite.

B. Emotion Detection database

Using the Emotion Detection - FER2013 dataset [13] from Kaggle, an online data science platform, we created a subset to train and evaluate our emotion recognition model. The original dataset contains approximately 35.9k facial images depicting seven emotions, including various races, genders, and ages. However, we narrowed our dataset to 26.6k images featuring only three emotional states: happiness, sadness, and neutral expressions. Fig. 2 shows a sample of the facial images included in our dataset.

C. Deep-Learning Model

The proposed custom CNN architecture is shown in Fig. 3, implemented in Python using TensorFlow libraries within Google Colab. Our model is composed of three convolutional blocks of 32, 64 and 128 filters of size 3x3, respectively. Each convolutional block contains a convolutional layer, a batch normalization layer that normalizes the input data for each mini-batch, a rectified linear activation function (ReLU) [14], a max pooling layer and a dropout layer with dropout rate of 0.25. The dropout layer deactivates some neurons during training to avoid overfitting. The flattened output of the last convolutional block is passed to a 1024-nodes dense layer and finally to a 2-nodes classification layer. The classification layer provides an output of two classes, 0 for sadness (sad or neutral expression) and 1 for happiness (happy expression).

D. Face Detection

A face detector algorithm using Haar's cascade algorithm [15] from the OpenCV library was used to detect the faces of persons in a scene. The algorithm compensates for lighting changes by returning the input image after mean subtraction, normalization, and channel swapping. After real-time detection, a green rectangular frame around the face was also created to highlight the area of the face. Later over this rectangle, the emotion detected and its corresponding prediction confidence value will be displayed above the face, as shown in Fig. 4.

E. Robotic Arm Control

The Dofbot robotic arm is controlled using a set of preloaded ROS commands. The arm's movement depends on the emotion detected, so several positions are defined in the six servos. Three basic commands are used for this purpose. The first command changes the angles of the six servos and controls their speed. The second command allows a servo to be selected, its angle changed, and its speed controlled. The third command allows the arm to reach a specific position before moving to the next. Similarly, three groups of actions are defined to control the arm movements: the first move the arm's position to focus the camera on the person for detection, while the second and third groups of movements deliver to a central position either a yellow or red cube located to the right or left of the robot, respectively.

In order to search for a person, the arm is initially moved to the first location using the commands of the first group. When a sad/neutral facial expression is detected, the arm moves to the right using the second series of commands or to the left using the third group of commands when a happy facial expression is detected. Nevertheless, during testing, it was observed that a minor movement of the subject during the prediction affects the classification, as the system only analyzes the most recent forecast value. Thus, a voting system was implemented to make a move. Nine consecutive real-time predictions must be made, eliminating the chance of a tie. Each time a neutral/sad facial expression is recognized, the counter raises, and if the counter equals or surpasses five, the forecast is sadness. Otherwise, it is happiness. After the arm moves, this counter is reset.

III. Experimentation and Results

A. Experimental Setup

1) Model configuration: The model was designed to receive grayscale images with input dimensions of 48 pixels in width and height, and output classes of either sadness or happiness. Due to previous studies indicating that it is more challenging to differentiate between sadness and neutral facial expressions [16]–[20], particularly when people do not show a marked difference between these two emotions, we decided to combine sad and neutral expressions into a single class labeled as "sadness". The Adam optimizer, which is an extended version of the stochastic gradient descent algorithm, was used to train the model with a learning rate of 0.001 and a decay rate set

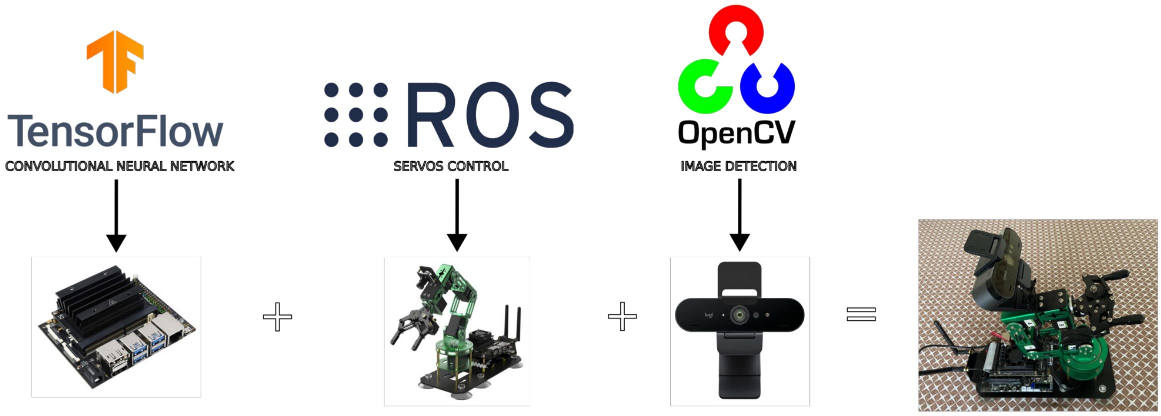

Fig. 1: Prototype setup.

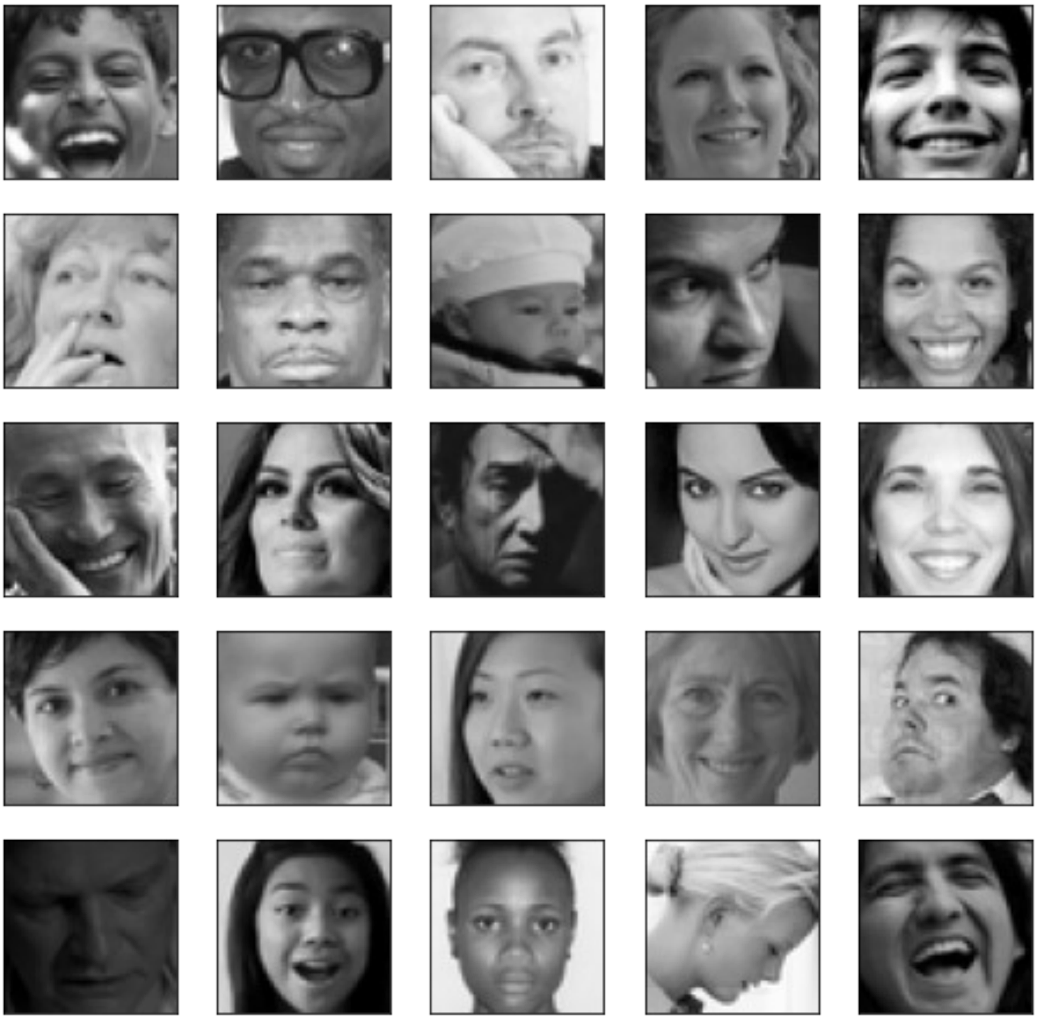

Fig. 2: Example of the images of sadness and happiness expressions available within the emotion detection database.

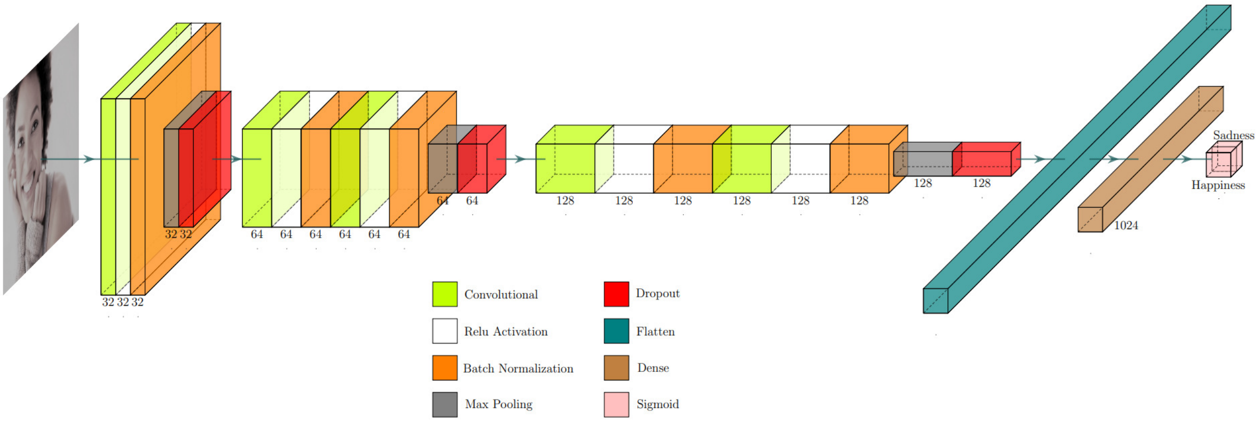

Fig. 3: Deep-Learning Model.

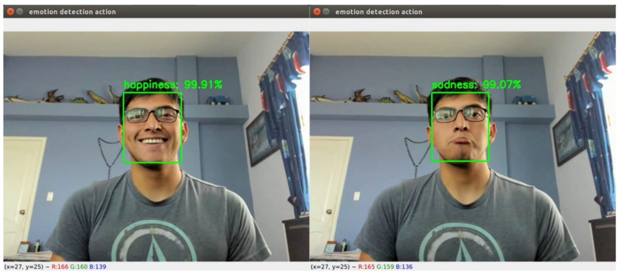

Fig. 4: Example of an image of a detected person with a green frame drawn around their face, indicating the classified expression and its corresponding prediction confidence value.

to the learning rate divided by the number of epochs. The model was trained with a batch size of 64 and 150 epochs. At the end of each epoch, loss and accuracy were calculated, it was observed that the achieved values were similar during both training and validation, indicating that our model was not overfitting. After training, a final loss of 10% and an accuracy of 90% were achieved, indicating that the model is suitable for real-time use.

2) Image preprocessing: A total of 10,308 neutral, 7,528 happy, and 3,514 sad grayscale images of size 48x48 pixels were used in the study. To ensure that images of sadness/neutrality and happiness were interspersed, the images were randomly shuffled and normalized. We performed data augmentation to increase efficiency by enabling the model to receive new variations of the images at each epoch during the training process. Random transformations such as displacement of the image in width and height, cropping portions of the image, zooming to a range of 0.2, horizontal flips, and filling new pixels using the nearest pixel, crop, or offset, rotations at an angle of 25° were applied to the images in the training set.

These random modifications provide a greater variety of images, preventing the algorithm from memorizing the training set and leading to overfitting. Data augmentation adds flexibility to forecasting, improves efficiency, and reduces losses. The training model was fed with a total of 150x17,080 image variations (i.e., 150 epochs, 17,080 training image data)

3) Training and test partitions: The image dataset was split into 80% training and 20% validation sets. As mentioned earlier, the distribution of the emotions in the dataset is unbalanced due to the relatively greater complexity in detecting sad or neutral expressions compared to happy expressions. However, it is worth noting that the degree of imbalance is not severe.

4) Assessment metrics: The performance of the model was monitored using accuracy (ACC) [21] and error metrics during the training and testing phases. These metrics evaluated how well the model could predict the correct emotion from the input image. In addition, since the prototype was implemented for real-time use, the same metrics were computed for real-time testing to ensure that the model could accurately predict emotions in a real-world scenario.

5) Real-time test: Combining the face detection vision module, the trained model, and the robotic arm control produced a real-time prototype. One hundred random volunteers participated in a real-time experiment on the Universidad San Francisco de Quito (USFQ) campus. As seen in Fig. 5, the robot arm was positioned on a table (80 cm high from the ground). Participants were briefed on the working operation of the prototype. The participants were then instructed to grin at the robot. Later, they were instructed to alter their facial expression to sad ones. Before the test, participants were instructed to remove accouterments such as glasses, hats, and face masks, among others, to eliminate any artifacts. The robot produced a prediction as shown in Fig. 6 and it delivered either a yellow cube in response to sadness/neutral expression recognition or a red cube in response to joyful expression detection. During testing, the performance of the model was tracked using the accuracy and error metrics.

B. Results and Discussion

The model was trained and tested on our emotion dataset, achieving an accuracy of 93.5% and an error of 6.5%, as shown in Fig. 7. The graph depicts four curves, with the

Fig. 5: Experimental setup at the USFQ campus for real-time testing of the prototype.

Fig. 6: Example of Real-Time testing scenarios.

upper plots illustrating the training and validation accuracy and the lower plots illustrating training and validation losses. A stable validation curve approaching 1 indicates that the model is performing well. The losses tend towards 0 and remain stable, indicating that the model is ready for testing. The final real-time prototype performed similarly to the training and validation results, with a precision percentage shown in Table I. The results were consistent, with slightly higher errors in detecting happy expressions. Fig. 8 shows some examples of the real-time predictions produced by the system.

TABLE I: Performance of the proposed model for binary emotion classification.

Test result	**Training values**	**Real-time test**	
		Happiness	***Sadness/neutral***
Accuracy (%)	93.5	94	91
Error (%)	6.5	6	9

IV. Conclusions and Future Work

This study illustrated the feasibility of utilizing a CNN-based emotion identification model to operate a robotic arm and deliver an object according to the facial expression of an individual in front of the robot. The real-time prototype of the system worked effectively and produced an acceptable level of accuracy. Nevertheless, the current arm movement is position controlled, so if the delivery object is not in a predetermined place, the arm will not deliver it. For enhancing the system, future work aims to include other emotions, such as surprise or anger, and enable the robotic arm to explore its range of motion and vision to search for particular objects.

Acknowledgment

The authors thank XX for their enthusiastic support and valuable insights throughout the development of this project.

References

[1] Y. Hu, N. Abe, M. Benallegue, N. Yamanobe, G. Venture, and E. Yoshida, “Toward active physical human–robot interaction: Quantifying the human state during interactions,” *IEEE Transactions on Human-Machine Systems*, vol. 52, no. 3, pp. 367–378, 2022.

[2] S. B. Sukhavasi, S. B. Sukhavasi, K. Elleithy, A. El-Sayed, and A. Elleithy, “A hybrid model for driver emotion detection using feature fusion approach,” *International journal of environmental research and public health*, vol. 19, no. 5, p. 3085, 2022.

[3] J. Heredia, E. Lopes-Silva, Y. Cardinale, J. Diaz-Amado, I. Dongo, W. Graterol, and A. Aguilera, “Adaptive multimodal emotion detection architecture for social robots,” *IEEE Access*, vol. 10, pp. 20 727–20 744, 2022.

[4] M. Wairagkar, M. R. Lima, D. Bazo, R. Craig, H. Weissbart, A. C. Etoundi, T. Reichenbach, P. Iyengar, S. Vaswani, C. James *et al.*, “Emotive response to a hybrid-face robot and translation to consumer social robots,” *IEEE Internet of Things Journal*, vol. 9, no. 5, pp. 3174–3188, 2021.

[5] J.-Y. Jeong, Y.-G. Hong, J. Oh, S. Hong, J.-W. Jeong, and Y. Jung, “Learning from synthetic data: Facial expression classification based on ensemble of multi-task networks,” *arXiv preprint arXiv:2207.10025*, 2022.

[6] N. Mehendale, “Facial emotion recognition using convolutional neural networks (ferc),” *SN Applied Sciences*, vol. 2, no. 3, p. 446, 2020.

[7] L. Miranda, D. Jiménez, D. Benítez, N. Peréz, D. Riofrío, and R. F. Moyano, “Robotic arm handling based on real-time gender recognition using convolutional neural networks,” in *2022 IEEE International Autumn Meeting on Power, Electronics and Computing (ROPEC)*, vol. 6. IEEE, 2022, pp. 1–6.

[8] Yahboom, “Dofbot AI vision robotic arm.” [Online]. Available: http://www.yahboom.net/study/Dofbot-Jetson_nano

[9] F. N. Uzun, M. Kayrici, and B. Akkuzu, “Nvidia jetson nano development kit,” *Programmable Smart Microcontroller Cards*, p. 82, 2021.

[10] J. Howse and J. Minichino, *Learning OpenCV 4 Computer Vision with Python 3: Get to grips with tools, techniques, and algorithms for computer vision and machine learning*. Packt Publishing Ltd, 2020.

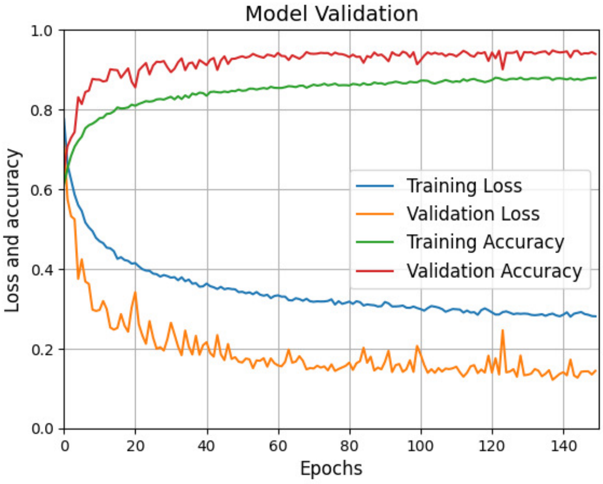

Fig. 7: Model performance in solving the binary classification during training and validation.

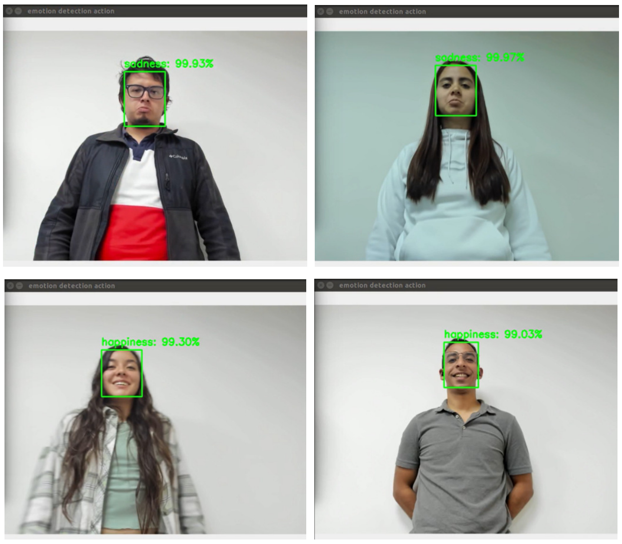

Fig. 8: Real-time prediction examples, a green frame is drawn around the face, indicating the classified expression and its corresponding prediction confidence value.

[11] Logitech, "Logitech brio 4k webcam." [Online]. Available: https://www.logitech.com/en-us/products/webcams/brio-4k-hdr-webcam.960-001105.html

[12] A. Koubaa, *Robot Operating System (ROS): The Complete Reference (Volume 7)*. Springer Nature, 2023, vol. 1051.

[13] Y. Khaireddin and Z. Chen, "Facial emotion recognition: State of the art performance on fer2013," *arXiv preprint arXiv:2105.03588*, 2021.

[14] N. Samadiani, G. Huang, Y. Hu, and X. Li, "Happy emotion recognition from unconstrained videos using 3d hybrid deep features," *IEEE access*, vol. 9, pp. 35 524–35 538, 2021.

[15] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*, vol. 1. IEEE, 2001, pp. I–I.

[16] J. Guo, Z. Lei, J. Wan, E. Avots, N. Hajarolasvadi, B. Knyazev, A. Kuharenko, J. C. S. J. Junior, X. Baro, H. Demirel *et al.*, "Dominant and complementary emotion recognition from still images of faces," *IEEE Access*, vol. 6, pp. 26 391–26 403, 2018.

[17] M. G. Calvo, A. Fernández-Martín, and L. Nummenmaa, "Facial expression recognition in peripheral versus central vision: Role of the eyes and the mouth," *Psychological research*, vol. 78, pp. 180–195, 2014.

[18] M. Guarnera, Z. Hichy, M. I. Cascio, and S. Carrubba, "Facial expressions and ability to recognize emotions from eyes or mouth in children," *Europe's journal of psychology*, vol. 11, no. 2, p. 183, 2015.

[19] G. Mancini, R. Biolcati, S. Agnoli, F. Andrei, and E. Trombini, "Recognition of facial emotional expressions among italian pre-adolescents, and their affective reactions," *Frontiers in psychology*, vol. 9, p. 1303, 2018.

[20] J. N. Schneider, M. Matyjek, A. Weigand, I. Dziobek, and T. R. Brick, "Subjective and objective difficulty of emotional facial expression perception from dynamic stimuli," *Plos one*, vol. 17, no. 6, p. e0269156, 2022.

[21] S. Dwijayanti, M. Iqbal, and B. Y. Suprapto, "Real-time implementation of face recognition and emotion recognition in a humanoid robot using a convolutional neural network," *IEEE Access*, vol. 10, pp. 89 876–89 886, 2022.