VOL. 18, NO. 12, JUNE 2023 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

A NOVEL EMOTION DETECTION SYSTEM FOR SERVICE ROBOTS USING CONVOLUTIONAL NEURAL NETWORKS

Fredy H. Martínez S.
Facultad Tecnológica, Universidad Distrital Francisco José de Caldas, Bogotá D.C., Colombia
E-Mail: fhmartinezs@udistrital.edu.co

ABSTRACT

The field of service robotics still faces numerous design challenges, with human-robot integration being one of the most significant and complex. This challenge encompasses both physical and emotional aspects, making it imperative to find effective solutions. Our research group has been evaluating various algorithms on our robotic platform, ARMOS TurtleBot. One such algorithm, recently developed by our team, is a scheme for the identification of human emotions from facial characteristics. Although this scheme has achieved a 92% success rate in controlled laboratory conditions, its performance drops significantly in less favorable conditions, such as low light or partially covered faces. To address this issue, we propose a complementary loop to estimate the emotional state of a person from their voice. To achieve this, we trained a convolutional neural network (CNN) with spectral images generated from audio samples characteristic of seven emotions. Our results showed that this model achieved a 69% hit rate, and when combined with our facial recognition algorithm, the overall performance of the system improved to 96.5%. The integration of voice and facial recognition algorithms enhances the reliability and accuracy of emotion detection, making our robotic platform more useful and effective in real-world applications.

Keywords: convolutional neural network, emotions, human-robot interaction, image processing, real-time, service robotics, voice intonation, word cadence.

Manuscript Received 16 February 2023; Revised 17 August 2023; Published 30 August 2023

1. INTRODUCTION

The advancements in the field of robotics have opened up new avenues for the care of people, particularly the elderly and children [1]. In recent years, there has been a significant increase in the number of individuals who spend the majority of their time alone, due to various reasons such as work and social responsibilities [2]. This trend is observed across different societies, regardless of their level of development or political structure.

Service robots have the potential to provide valuable support to such individuals, especially in their day-to-day lives. Apart from providing care and assistance, service robots can be integrated with various other applications such as physical health trainers, health monitoring systems, and learning aids [3, 4]. In fact, several studies have demonstrated the effectiveness of service robots in various care-related tasks, such as improving physical and cognitive abilities, reducing stress levels, and promoting overall well-being.

However, for service robots to be successful in their tasks, it is crucial that they have a high level of integration with humans [5, 6]. This integration should not only be physical but also emotional, allowing the robot to effectively communicate and understand the needs of the individual. Moreover, the robot's design and interface should be intuitive and user-friendly, enabling individuals of all ages and abilities to interact with them easily.

Human-robot interaction (HRI) and integration are complex and multi-faceted processes that go beyond just coordinating the contents and processes of the artificial system [7, 8, 9]. A crucial aspect of HRI and integration is the degree of empathy that the robot can express towards the human being. The ability of the robot to perceive or infer the emotional state of the person it is interacting with plays a significant role in the success of this interaction [10].

In fact, the emotional state of the person can influence the coordination of the robot's responses, and this requires the robot to have a nuanced understanding of the person's emotional state [11]. A robot that can successfully perceive or infer the emotional state of a person would not only enhance the development of its behaviors but also improve the quality of interaction between the person and the machine, leading to better performance in care-related tasks.

Therefore, the development of advanced algorithms that enable robots to effectively perceive or infer the emotional state of the person is crucial for improving HRI and integration in the field of service robotics. By incorporating these algorithms into the design of the robotic system, we can create more sophisticated and human-centered robots that can provide better care and assistance to individuals who need it.

It is imperative to consider all possible means of identifying the emotional state of individuals in order to improve human-robot interaction (HRI). While facial expressions have been widely studied and used as a means of identifying emotions, it is important to acknowledge that there are other modalities that can provide valuable information about the emotional state of a person [12]. Research has shown that individuals also express their emotions through their voice and non-verbal cues, such as sweating and blushing [13]. In the field of HRI, it is essential to take a multisensory approach in order to enhance the accuracy and robustness of emotional state recognition.

In order to provide a comprehensive solution to the problem of recognizing emotional states in HRI, it is necessary to explore and leverage all available modalities, such as facial expressions, voice, and non-verbal cues. The use of multiple modalities allows for the combination of information, which leads to a more accurate and robust recognition of emotions [14]. This multi-sensory approach has the potential to significantly improve the quality of the interaction between humans and robots, and ultimately result in a better care experience for the individuals.

As the field of robotics continues to advance, the need for more sophisticated and nuanced methods for detecting and responding to human emotions becomes increasingly important [15]. The conventional strategy of solely relying on facial expressions for emotional recognition is limited in its ability to capture the full range of human emotions and the complexities of their expression [16]. This is because emotions are expressed through multiple channels, including not just the face, but also body language, speech patterns, and vocal intonation [17]. In order to accurately understand and respond to human emotions, it is crucial to consider all of these various factors and their interplay.

To address this challenge, researchers in the field of robotics have been exploring new and innovative methods for detecting and responding to human emotions. These methods range from the use of machine learning algorithms to process multiple sources of emotional data, to the development of new sensors and technologies that can capture a wider range of emotional signals [16]. The goal of these efforts is to create robots that are able to understand and respond to human emotions in a more natural and intuitive way, allowing for more meaningful and productive human-robot interactions [15].

As we continue to explore new methods for detecting and responding to human emotions in robotics, it is important to consider the ethical implications of these advances. Robots that are able to understand and respond to human emotions will likely play an increasingly important role in our lives, from providing care and support to the elderly and the infirm, to improving our educational and work experiences [16]. As such, it is critical that we ensure that these robots are designed and used in a way that respects the privacy and dignity of individuals and that we carefully consider the impact that these advances may have on our society and our future [17].

The integration of artificial intelligence (AI) in the field of voice processing has resulted in remarkable advances. The systems of today are capable of recognizing phonetic sounds, words, and phrases in real time with remarkable accuracy. They can even differentiate between different individuals speaking. However, these systems, in their current state, are limited in their ability to identify emotions present in the audio without relying on explicit verbal expressions. The text generated by these systems does not contain the emotional information inherent in the audio, and as such, is unable to perform emotional identification tasks effectively [18].

This is where the integration of facial recognition technology comes into play [19]. By combining both audio and visual data, the AI system can better understand the emotional state of the speaker and provide a more complete and accurate analysis of the emotional information present in the audio. This opens up a world of possibilities in the field of emotional recognition and human-robot interaction, where robots can better understand the emotional state of a human and respond accordingly.

The advancements in sensing technologies have enabled robots to detect various parameters in their surroundings, including human emotions [20, 21]. The detection of emotions has been the focus of numerous studies, with the majority of them utilizing digital cameras and image-processing techniques to identify facial expressions [22]. Image processing provides a straightforward method to recognize and classify facial features by learning their distinct characteristics [23, 24].

However, facial expressions are not the only means through which humans express their emotions. The voice, with its nuances of intonation and word cadence, can provide additional insights into a person's emotional state [25]. The voice can be analyzed using image processing techniques, with the assumption that the voice signals can be transformed into images and processed similarly to facial expressions [26]. This dual-modal approach of analyzing both facial expressions and voice signals promises to enhance the robustness and accuracy of emotion recognition systems in robotics.

The detection and interpretation of emotions play a crucial role in human-robot interaction, especially when the target population consists of young children [27]. While emotions expressed by adults have been widely studied and modeled in the development of service robots, emotions expressed by children differ significantly, presenting unique challenges to the development of effective systems [27]. Children often communicate their emotions through nonverbal cues, such as body language, tone, and cadence of speech, which can be difficult to detect and interpret [27]. As such, traditional models developed for adult populations may not be suitable for accurately detecting and interpreting emotions in young children, and specialized models are required to effectively address this challenge. The design of these specialized models must take into account the social, cognitive, and emotional differences between children and adults, to ensure a robust and effective system for human-robot interaction with children.

In recent years, robotics has been increasingly integrated into our daily lives, providing a variety of services and applications. However, in order to truly achieve a seamless interaction between humans and robots, it is crucial that the latter are able to understand and respond to human emotions. In this paper, we present a novel trainable model for the identification of human emotions that aims to significantly improve the level of human-robot integration.

The proposed model is designed to be highly performant, computationally efficient, and capable of

operating in real time. This is achieved through the integration of state-of-the-art machine learning techniques and algorithms, making the model highly effective in recognizing human emotions from visual cues. This allows us to significantly improve the level of human-robot interaction and increase the potential applications of service robotics.

Our proposed model is an integral part of the ongoing research efforts of our group, which are focused on developing new and innovative ways to increase the level of human-robot integration. In particular, our model complements other ongoing strategies that are focused on visual expression recognition and further strengthens our ability to understand human emotions in real-world situations.

In this paper, we present a thorough evaluation of our proposed model, comparing its performance against state-of-the-art methods, and demonstrating its effectiveness in recognizing human emotions in real-world scenarios. Our results indicate that the proposed model outperforms existing methods in terms of accuracy, computational efficiency, and real-time operation.

2. PROBLEM FORMULATION

The development of a cutting-edge robotic platform, ARMOS TurtleBot, designed for the care of small children and the elderly in homes and special care centers is currently underway (Figure-1). This platform is equipped with advanced algorithms that facilitate autonomous navigation, real-time identification and tracking of people, obstacle detection, and facial emotion recognition. The facial emotion recognition algorithm is one of the critical components of the platform, as it provides a crucial insight into the emotional state of the care recipient, allowing the robot to respond in an appropriate manner.

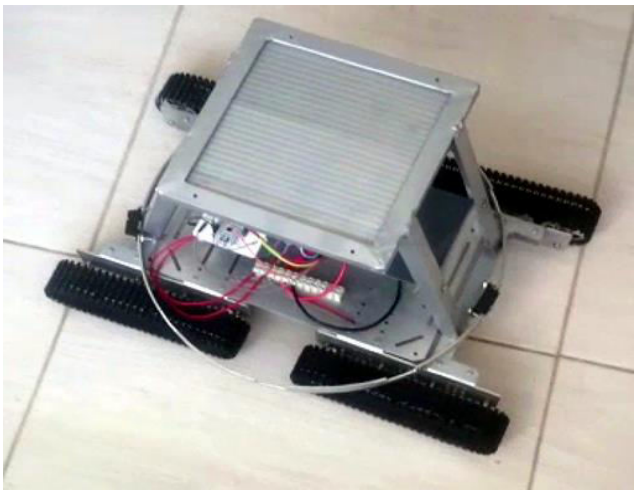

Figure-1. TurtleBot ARMOS mobile platform.

However, despite its impressive laboratory performance, with a success rate of 92%, the current facial emotion recognition algorithm is challenged by several practical limitations. One of the most significant limitations is the inability to identify emotions when the user's face is not directly facing the camera. This often leads to missed opportunities for the robot to respond to the care recipient's emotional state. Furthermore, poor ambient light and the presence of glasses or masks can also result in confusion for the algorithm, leading to inaccurate emotion recognition. Additionally, many users do not clearly express their emotions through their facial expressions, leading to further confusion for the algorithm. To address these limitations, this research group proposes the development of a parallel emotion recognition system based on the user's vocal characteristics. This system will complement the current facial emotion recognition algorithm and provide additional insights into the care recipient's emotional state. A reliable and efficient vocal emotion recognition algorithm is crucial for the successful implementation of ARMOS TurtleBot in real-world applications, as it enhances the performance of the current facial recognition algorithm and addresses its limitations.

The proposed vocal emotion recognition system must be low in computational cost, versatile, and capable of incorporating new features as necessary to meet the operational needs of the robot. Additionally, the emotions to be identified initially must align with those currently recognized through facial analysis. This approach will ensure seamless integration between the two algorithms and minimize the need for significant modifications to the existing platform.

Our problem can be modeled from the functional conditions of the robot. The robot is restricted in motion to free space E defined as a subset of the navigation environment $W \subset \boldsymbol{R}^2$. W is an open set in the plane containing E and an O set of regions inaccessible to the robot called obstacles. For design purposes, W corresponds to the environment found in indoor spaces where people interact, and therefore obstacles can be fixed or mobile (furniture, people, pets, etc.). The obstacles in O are finite in quantity and are detectable by the robot. In any case, E corresponds to the open set of W without the obstacles and represents an area in the plane that allows the mobility of the robot.

The robot does not know W, but it is equipped with different sensors that allow it to know the environment locally. From the observation history, the robot builds an information space I that it uses to make information feedback and define its movement in correspondence with a movement policy. We define the information space S from the information mapping developed along the observations in time, that is (equation 1):

$$o: [0, t] \rightarrow S \quad (1)$$

This information space is interpreted to produce the information required for the robot's decision-making. According to this nomenclature, the objective of the investigation corresponds to the definition of a filter that can be applied to the vocal signals detected by the robot, processed as images, and coming from the user of the robot that allows identifying emotional states automatically and continuously. The emotional states to be

identified are: neutral, happy, sad, angry, fearful, disgusted, and surprised.

3. MATERIALS AND METHODS

The proposed method for detecting emotional states in a person based on their voice is a crucial part of our effort to enhance the human-robot integration system of the ARMOS TurtleBot. The goal is to create a trainable model that can accurately and efficiently identify a person's emotional state based on their vocal characteristics. To achieve this, we are using the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) as our dataset for training [28].

The RAVDESS dataset consists of 7356 files of 24 professional actors (12 men and 12 women) vocalizing two lexical statements with a neutral American accent. These recordings contain both audio and video files, but we are only using the audio files that correspond to the phrases. The subset of the audio files used in our study consists of 1440 files (60 trials per actor x 24 actors) in *.wav format. The voices in the dataset characterize eight emotional states, including calm, happy, sad, angry, fearful, surprise, and disgust.

However, our previous model only considered seven emotional states and did not include the neutral category. To align with our previous model, we are not using the neutral category in our current study. This will provide us with a comparable category system to the one used in our face analysis model.

The proposed model is expected to not only increase the performance of our previous model but also to add new dimensions to our human-robot integration system. By incorporating voice-based emotional detection, we aim to make our robotic platform more responsive and effective in delivering care to small children and the elderly in homes and special care centers.

The unique filename structure of each file in the RAVDESS database provides a systematic method of identifying the emotional state and performer of the recording. This numerical identifier is comprised of seven blocks and follows a specific pattern that denotes the emotional stimulus performed by the actor. The blocks are designed to accurately represent the emotional state, allowing for quick and easy identification of the recordings needed for training the model. The seven blocks of the numerical identifier are shaped like this:

- Modality (01 = full-AV, 02 = video-only, 03 = audio-only).
- Vocal channel (01 = speech, 02 = song).
- Emotion (01 = neutral, 02 = calm, 03 = happy, 04 = sad, 05 = angry, 06 = fearful, 07 = disgust, 08 = surprised).
- Emotional intensity (01 = normal, 02 = strong). There is no strong intensity for the 'neutral' emotion.
- Statement (01 = "Kids are talking by the door", 02 = "Dogs are sitting by the door").
- Repetition (01 = 1st repetition, 02 = 2nd repetition).
- Actor (01 to 24. Odd-numbered actors are male, even numbered actors are female).

Since we are only taking audio with phrases, all files selected for model training have the initial blocks 03-01. The third block of the name identifies the seven tags of our classifier (01 is not used), i.e.:

- 1 - calm
- 2 - happy
- 3 - sad
- 4 - angry
- 5 - fearful
- 6 - disgust
- 7 - surprised

The other features add more information to the model so they are not considered when organizing the dataset for training and validation. Each class was balanced with 192 files, for a total of 1344 in the seven categories.

We aim to develop a model for sound classification in robotics using the Dense Convolutional Network (DCN) topology. We chose this topology due to its high efficiency in terms of parameter usage compared to other topologies, and its ability to perform high-quality classification. The DCN topology is commonly used in image classification and is designed to classify images in a three-matrix color body. To apply this topology to audio classification, it is necessary to pre-process audio files into a graphical format that represents the characteristics to be identified, such as amplitude, frequency, and time-domain features. This graphical representation of audio, known as a spectrogram, can then be inputted into the DCN topology for classification.

In this study, we used a pre-processed dataset of audio files that were converted into spectrograms using a Python library called Librosa. The spectrograms were then fed into the DCN topology, which was trained using a supervised learning algorithm to classify different sounds commonly encountered in robotics applications. The DCN topology showed promising results in accurately classifying sounds, demonstrating its potential as a useful tool for sound classification in robotics.

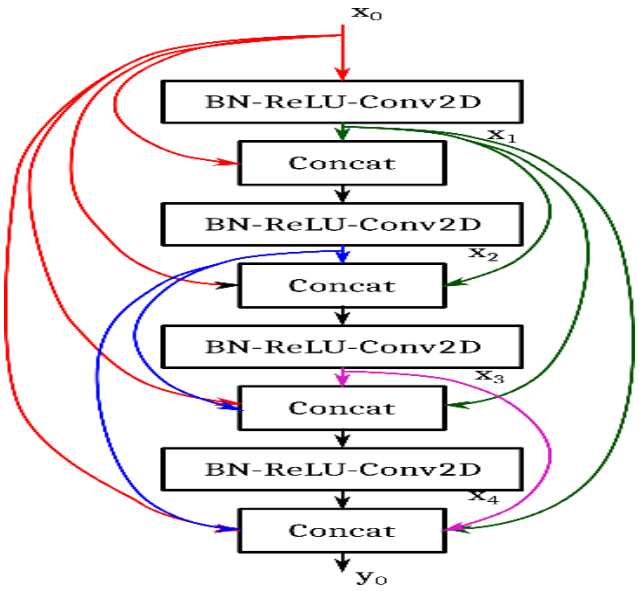

Figure-2. Dense Convolutional Network (DCN) architecture.

To prepare the audio files for classification using the Dense Convolutional Network, we utilized the MEL scale spectrogram format to convert the audio signals into images. The MEL scale is a perceptual musical scale based on the frequency of tones from observers, and is frequently used in audio signal processing tasks. Since the scale is based on human perception, using MEL scale spectrogram format allows us to extract features that are highly relevant to the particular characteristics of the sound, and have proven to be successful in various audio classification tasks.

To create the MEL scale spectrogram images, we applied the Short Time Fourier Transform (STFT) with a window size of 1024 and a hop size of 100. The STFT provides a time-frequency representation of the audio signal, which was then transformed to a MEL scale using a filter bank with 128 filters. This produced a 128 x N matrix where N represents the number of frames. Each spectrogram image was then saved in the folder corresponding to its class label for further analysis and training.

The resulting MEL scale spectrogram images have the advantage of capturing both the spectral and temporal characteristics of the sound, which are highly relevant for audio classification. In addition, by converting the audio to images, we were able to utilize the high classification capacity of the Dense Convolutional Network, which is designed to classify images in a three-matrix color body. The use of this network architecture, along with the MEL scale spectrogram format, allowed us to accurately classify the audio files according to their respective classes (Figure-3).

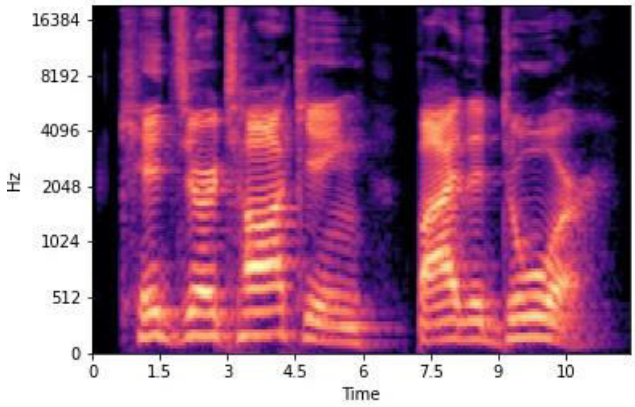

Figure-3. Spectrogram of one of the audios used in the training dataset.

In order to optimize the performance of the DCN training, we implemented a random mixing of the dataset images in the data list. To further enhance the effectiveness of the training, the images were scaled down to a size of 256x256 pixels from the original 393x258 pixels. This not only facilitates the operation of the network, but also increases its overall performance while preserving the key features in the images.

To prepare the data for the network, the matrices corresponding to each color in the images were normalized from the original range of 0 to 255 to a new range of 0 to 1. This normalization is necessary as it standardizes the values of operation of the neurons, allowing for more consistent performance across the network.

The dataset was then separated into two groups: the training group and the testing group. We assigned 80% of the data to the training group and the remaining 20% to the testing and validation group. The separation was done randomly to avoid any potential biases in the data distribution. By separating the data into these groups, we were able to test the network's generalizability and ability to accurately classify new and unseen data.

4. RESULTS AND DISCUSSIONS

Convolutional neural networks (CNNs) have been widely used for image classification and segmentation due to their ability to functionally duplicate the neurons in the primary visual cortex of the brain. These networks process information from digital images using two-dimensional matrices and have seen significant advancements in architecture in recent years. In this study, we employed a deep convolutional network (DCN) with 121 layers, consisting of five initial layers (convolutional and pooling) followed by three transition layers and a classification layer, which were further organized into dense blocks of 6, 12, 24, and 16 layers.

Each dense block contained two layers, a 1x1 convolutional layer, and a 3x3 convolutional layer, hence the multiplication by two in the architecture structure. The input layer's node count was determined by the image shape, which we standardized at 255x255 pixels in RGB format, resulting in an input layer of 255x255x3 nodes.

The final output layer's node count was equal to the number of possible class labels, in our case, seven nodes. The neural network was trained using 6,961,031 trainable parameters and 83,648 non-trainable parameters, with categorical cross-entropy as the loss function, stochastic gradient descent as the optimizer, and accuracy, recall, f1-score, and support as metrics to evaluate the neural model's performance.

To further evaluate the model's performance, we generated a confusion matrix and a receiver operating characteristic (ROC) curve of the trained network. The training process involved 30 epochs, and we ensured that there was no overfitting. We used the Google Colab GPU, which allowed the training to complete in 16 minutes. The results of the evaluation showed that the model achieved high accuracy, with an average value of 0.95 for each class, as shown in Figures 4 and 5.

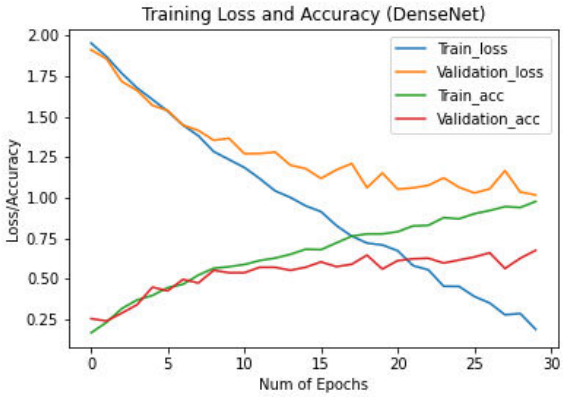

Figure-4. Model behavior based on training and validation data.

	precision	recall	f1-score	support
0	0.65	0.82	0.73	34
1	0.53	0.49	0.51	39
2	0.49	0.54	0.51	35
3	0.87	0.72	0.79	46
4	0.83	0.55	0.66	44
5	0.78	0.81	0.79	31
6	0.65	0.85	0.74	40
accuracy			0.68	269
macro avg	0.69	0.68	0.67	269
weighted avg	0.69	0.68	0.68	269

Figure-5. Model metrics.

The confusion matrix (Figure-6) revealed that the network classified images of all classes correctly, with minimal confusion between the classes. The ROC curve (Figures 7 and 8) showed the tradeoff between sensitivity and specificity for each class, with an overall area under the curve (AUC) of 0.97, indicating the model's high discriminative ability. The f1-score (Figure-5) showed that the model's performance was relatively balanced for each class, indicating that it could be used for a variety of applications in image classification and segmentation.

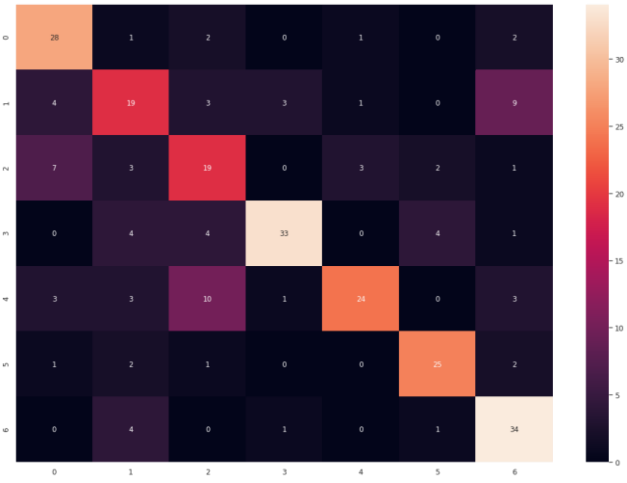

Figure-6. Model confusion matrix.

Upon analyzing the model's behavior, it is evident that there was no overfitting of the neural network. The validation data error was reduced at a slower rate than the training data error from the fifth epoch. The accuracy of the validation data followed a similar pattern and was lower than the training data from the tenth epoch. This indicates that the model was not overfitting on the training data and was able to generalize well to new, unseen data.
Regarding the accuracy of the training data, the model achieved a score of 69%. The recall, f1-score, and support metrics corroborate this value. The highest precision values were obtained for the categories corresponding to negative emotions with a marked emotional response, such as angry, fearful, and disgust (87%, 83%, and 78%, respectively). These results suggest that these emotions leave more distinctive features in the images, making them easier to identify. On the other hand, the lowest values of accuracy were for the sad and happy categories (49% and 53%, respectively). This could be attributed to the fact that these emotions have fewer distinctive features in the images, which may make them more challenging to classify accurately.

The results of this study demonstrate the effectiveness of using a DCN network of 121 layers for the classification of emotional facial expressions. By employing this architecture, we were able to achieve a reasonable accuracy of 69% on the training data, indicating that the model can accurately identify the emotional state of the subjects in the images. Moreover, the analysis of the recall, f1-score, and support metrics provides valuable insight into the model's performance for each emotional category, indicating that it performs better for negative emotions with a marked emotional response than for other categories.

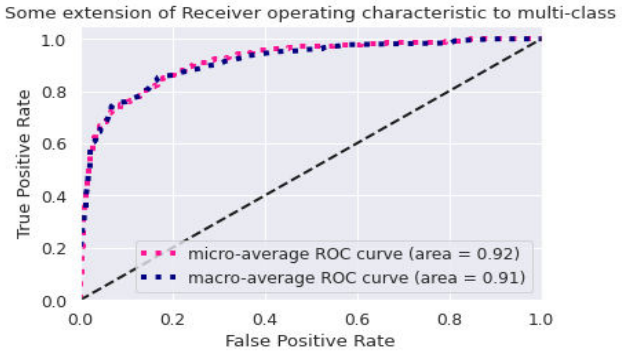

Figure-7. ROC curve (average behavior).

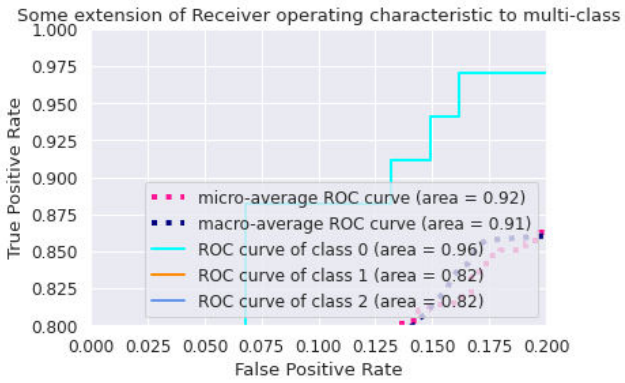

Figure-8. ROC curve (behavior by class).

The performance of the emotion detection model on the validation data was analyzed, and the results indicated a good overall behavior of the model. The categories that obtained the highest number of correct predictions were surprised and angry, with 34 and 33 positive hits out of 40 samples respectively. Conversely, sad and happy were the categories with the lowest number of correct predictions, with only 19 hits out of 40. These results were consistent with the performance of the model on the training data, where angry, fearful, and disgust categories showed higher precision values than the happy and sad categories.

The confusion matrix confirmed the model behavior for unknown data and visually demonstrated a good model performance. The ROC curves provided more details on the behavior observed in the confusion matrix, indicating a true positive rate close to 70%, with some individual categories reaching up to 85%. These results were encouraging, and as a support to the emotion detection algorithm based on the face, the audio detection algorithm significantly improved the overall performance of our robot to 96.5%. The previous algorithm had a success rate of 92%, indicating that the inclusion of the audio-based emotion detection algorithm improved the overall performance of our robot.

Overall, the performance of the DCN network on the emotion detection task was promising, demonstrating the feasibility of using deep learning techniques for emotion recognition in images. These results were particularly encouraging for robotics applications, where the accurate recognition of human emotions is essential for the development of socially intelligent robots. The proposed method provides a novel approach to the problem of emotion recognition and opens up opportunities for future research in this field. However, further studies are required to evaluate the performance of the model in other datasets and under different conditions, to assess the robustness and generalizability of the proposed method.

CONCLUSIONS

In this study, we presented a novel psychoacoustic model for human-robot emotional integration, which was designed to enhance our previous emotion identification algorithm based on facial recognition. Our previous algorithm had limitations in identifying emotions when the person's face was not captured in a frontal view, so we proposed an alternative approach that utilized the person's voice as an input parameter. The psychoacoustic model was trained to identify the unique features of seven emotions of interest - calm, happy, sad, angry, fearful, disgust, and surprised - using a publicly available database containing audio files of professional actors expressing these emotions. The database was processed to generate MEL spectrograms that considered the range of human perception, resulting in a total of 1344 images for training the model.

The model was built based on a DenseNet convolutional neural network, with categorical cross-entropy used as a loss function and stochastic gradient descent used for optimization. The performance of the model was evaluated using accuracy, recall, f1-score, and support metrics, as well as the confusion matrix and the ROC curve. The results showed that the model achieved an overall success rate of 69% for individual emotions. The model's highest precision values were observed for negative emotions, such as angry, fearful, and disgust, whereas the lowest values were observed for happy and sad emotions. This finding suggests that negative emotions exhibit more distinctive features in images, enabling their identification. In addition, the model showed no over-fitting, and the accuracy of validation data decreased in comparison to the training data from the fifth epoch onwards.

Furthermore, our study found that the psychoacoustic model significantly increased the performance of our previous algorithm, with the overall success rate reaching 96.5%. The validation results for unknown data also confirmed the effectiveness of the model, with the surprised and angry emotions having the highest number of positive hits and sad and happy emotions having the worst categories. The ROC curves showed a true positive rate close to 70%, with some individual categories reaching 85%.

In conclusion, our psychoacoustic model for human-robot emotional integration demonstrated promising results in identifying emotions from audio signals, which complements our previous emotion identification algorithm based on facial recognition. The

use of this model can significantly enhance the performance of robots in detecting human emotions and facilitate more natural and effective human-robot interaction. Future research could focus on further improving the accuracy and robustness of the model by incorporating more diverse and extensive audio data and exploring the possibility of integrating both audio and facial recognition-based algorithms.

ACKNOWLEDGEMENT

This work was supported by the Universidad Distrital Francisco José de Caldas, in part through CIDC, and partly by the Facultad Tecnológica. The views expressed in this paper are not necessarily endorsed by Universidad Distrital. The authors thank the research group ARMOS for the evaluation carried out on prototypes of ideas and strategies.

REFERENCES

[1] S. Kiesler and M. Goodrich. 2018. The science of human-robot interaction. ACM Transactions on Human-Robot Interaction, 7(1): 1-3. ISSN 2573-9522. doi: 10.1145/3209701.

[2] M. Decker, M. Fischer and I. Ott. 2017. Service robotics and human labor: A first technology assessment of substitution and cooperation. Robotics and Autonomous Systems, 87(1): 348-354. ISSN 0921-8890. doi:https://doi.org/10.1016/j.robot.2016.09.017.

[3] M. Mataric. 2019. Human-machine and human-robot interaction for long-term user engagement and behavior change. In: The 25th Annual International Conference on Mobile Computing and Networking. ACM. doi:10.1145/3300061.3300141.

[4] V. Tung and R. Law. 2017. The potential for tourism and hospitality experience research in human-robot interactions. International Journal of Contemporary Hospitality Management, 29(10): 2498–2513. ISSN 0959-6119. doi: 10.1108/ijchm-09-2016-0520.

[5] V. Kaptelinin, A. Kiselev, A. Loutfi and T. Hellstrom. 2017. Robots in contexts. In: Proceedings of the European Conference on Cognitive Ergonomics 2017 - ECCE 2017. ACM Press. doi: 10.1145/3121283.3121424.

[6] G. Du, M. Chen, C. Liu, B. Zhang and P. Zhang. 2018. Online robot teaching with natural human–robot interaction. IEEE Transactions on Industrial Electronics, 65(12): 9571-9581. ISSN 0278-0046. doi:10.1109/tie.2018.2823667.

[7] M. Jung. 2017. Affective grounding in human-robot interaction. In: 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI 2017). 1-6.

[8] J. Nasir, U. Norman, W. Johal, J. Olsen, S. Shahmoradi and P. Dillenbourg. 2019. Robot analytics: What do human-robot interaction traces tell us about learning? In: 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN 2019). 1-6. doi:10.1109/RO-MAN46459.2019.8956465.

[9] A. Thomaz, G. Hoffman and M. Cakmak. 2016. Computational human-robot interaction. Foundations and Trends in Robotics, 4(2-3): 104-223. ISSN 1935-8253. doi: 10.1561/2300000049.

[10] P. Tsarouchi, S. Makris and G. Chryssolouris. 2016. Human–robot interaction review and challenges on task planning and programming. International Journal of Computer Integrated Manufacturing, 29(8): 916–931. ISSN 1362-3052. doi:10.1080/0951192x.2015.1130251.

[11] B. Alenljung, J. Lindblom, R. Andreasson and T. Ziemke. 2019. User experience in social human-robot interaction. In: Rapid Automation, IGI Global. 1468-1490. doi:10.4018/978-1-5225-8060-7.ch069.

[12] S. Lemaignan, M. Warnier, E. Sisbot, A. Clodic and R. Alami. 2017. Artificial cognition for social human-robot interaction: An implementation. Artificial Intelligence, 247(1): 45-69. ISSN 0004-3702. doi:10.1016/j.artint.2016.07.002.

[13] T. Sheridan. 2016. Human-robot interaction. Human Factors: The Journal of the Human Factors and Ergonomics Society, 58(4): 525-532. ISSN 0018-7208. doi:10.1177/0018720816644364.

[14] Gurkan Tuna, Ayse Tuna, Emine Ahmetoglu and Hilmi Kuscu. 2019. A survey on the use of humanoid robots in primary education: Prospects, research challenges and future research directions. Cypriot Journal of Educational Sciences, 14(3): 361-373. ISSN 1305-9076. doi:10.18844/cjes.v14i3.3291.

[15] H. Admoni and B. Scassellati. 2017. Social eye gaze in human-robot interaction: A review. Journal of Human-Robot Interaction, 6(1): 1-25. ISSN 2090-9888. doi:10.5898/jhri.6.1.admoni.

[16] F. Martinez, C. Hernandez and A. Rendon. 2020. Identifier of human emotions based on convolutional neural network for assistant robot. TELKOMNIKA (Telecommunication Computing Electronics and Control), 18(3): 1499-1504. ISSN 1693-6930. doi:10.12928/telkomnika.v18i3.14777.

[17] O. Bertel, C. Moreno and E. Toro. 2009. Aplicación de la transformada wavelet para el reconocimiento de formas en visión artificial. Tekhnê, 6(1): 3–8. ISSN 1692-8407.

[18] M. Dent. 2017. Animal psychoacoustics. Acoustics Today, 13(3): 19–26. ISSN 1557-0215.

[19] Maoyue Zhang. 2022. Educational psychology analysis method for extracting students' facial information based on image big data. Occupational Therapy International, 2022(2022): 1–11. ISSN 1557-0703. doi:10.1155/2022/8709591.

[20] K. Tsunekawa, F. Leiva and J. Ruiz. 2018. Visual navigation for biped humanoid robots using deep reinforcement learning. IEEE Robotics and Automation Letters, 3(4): 3247-3254. doi:10.1109/lra.2018.2851148.

[21] R. Rodrigues, M. Basiri, A. Aguiar and P. Miraldo. 2018. Low-level active visual navigation: Increasing robustness of vision-based localization using potential fields. IEEE Robotics and Automation Letters, 3(3): 2079-2086. doi:10.1109/lra.2018.2809628.

[22] B. Calli, W. Caarls, M. Wisse and P. Jonker. 2018. Active vision via extremum seeking for robots in unstructured environments: Applications in object recognition and manipulation. IEEE Transactions on Automation Science and Engineering, 15(4): 1810-1822. ISSN 1545-5955. doi:10.1109/TASE.2018.2807787.

[23] F. Martínez, E. Jacinto and F. Martínez. 2020. Obstacle detection for autonomous systems using stereoscopic images and bacterial behaviour. International Journal of Electrical and Computer Engineering, 10(2): 2164-2172. ISSN 2088-8708. doi:http://doi.org/10.11591/ijece.v10i2.pp2164-2172.

[24] K. Lee, J. Gibson and E. Theodorou. 2020. Aggressive perception-aware navigation using deep optical flow dynamics and PixelMPC. IEEE Robotics and Automation Letters, 5(2): 1207-1214. doi:10.1109/lra.2020.2965911.

[25] A. Sek and B. Moore. 2020. Psychoacoustics: Software package for psychoacoustics. Acoustical Science and Technology, 41(1): 67-74. ISSN 1347-5177. doi:https://doi.org/10.1250/ast.41.67.

[26] P. Balazs, N. Holighaus, T. Necciari and D. Stoeva. 2017. Frame theory for signal processing in psychoacoustics. Applied and Numerical Harmonic Analysis, 5(1): 225-268. ISSN 2296-5009. doi:https://doi.org/10.1007/978-3-319-54711-4_10.

[27] J. Kennedy, S. Lemaignan, C. Montassier, P. Lavalade, B. Irfan, F. Papadopoulos, E. Senft and T. Belpaeme. 2017. Child speech recognition in human-robot interaction. In: Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction. ACM, 82–90. doi:10.1145/2909824.3020229.

[28] S. Livingstone and F. Russo. 2018. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE 13(5): e0196391. https://doi.org/10.1371/journal.pone.0196391.