GRADUATE STUDENTS' ASSOCIATION UNIVERSITY OF CALGARY

PEER BEYOND GRADUATE RESEARCH CONFERENCE
UCALGARY'S MULTI-DISCIPLINARY GRADUATE RESEARCH CONFERENCE
Impact starts here.

Bridging Emotional Understanding: A Multimodal Emotion Detection System for Neurodivergent Individuals

Wamika Jha, Zoe Kirsman, Mea Wang, Usman Alim
{wamika.jha, zoe.kirsman, meawang, ualim}@ucalgary.ca
Faculty of Science, University of Calgary

Abstract: Human communication is inherently tied to emotions, which play a critical role in guiding and enhancing social interactions. For neurodivergent individuals, particularly children, challenges often arise in expression and interpretation of emotions. Emotion detection technologies can therefore serve as powerful tools to aid in communication and to improve social interaction. However, emotional changes among neurodivergent individuals span a wider spectrum and exhibit greater subtle differences. Existing emotion detection models have been predominantly trained with data in single modality. Integrating data from multiple modalities provides a more comprehensive approach to understanding emotions, mirroring the way humans naturally perceive the world through all five senses. This study presents a Multimodal Emotion Detection System that leverages publicly available datasets to enhance recognition accuracy. By fusing diverse data sources, the proposed model captures subtle emotional cues more effectively than traditional methods. Experimental results confirm its robustness and suitability for real-world applications.

1. Introduction

Human communication is inherently emotional – we tend to use body language, facial expressions and vocal tones when we are communicating with each other. Human emotions form the bedrock of social interactions. However, neurodivergent individuals, particularly children, often encounter difficulties in expressing and interpreting emotions through conventional means. Li investigated various interpersonal differences and variations between autistic and non-autistic children by following their development over the period of two to three years; the results of their study indicated that autistic children faced more challenges in discriminating, identifying, and attributing emotions compared to their non-autistic peers [1]. A recent clinical study involving 113 children aged 5–9 years across three distinct cultures found that children with Autism Spectrum Disorder (ASD) experience significant challenges in recognizing emotions across multiple sensory modalities, including visual, auditory, and multimodal cues [2].

Emotion detection technologies can serve as valuable tools in aiding communication, social interaction, and overall well-being for neurodivergent users. Previous studies have explored the use of Virtual Reality, wearable devices, and Machine Learning algorithms for emotion detection in humans across various contexts, including e-learning environments, classroom settings, and healthcare applications [3, 4, 5, 6]. Automatic emotion detection involves using machine learning models to identify and interpret human emotions, often through facial expressions, voice tone, and other behavioral cues. However, there is very limited work done in this field, which is geared towards neurodivergent individuals, especially children. This brings a need to create supportive tools which help neurodivergent population to navigate these challenges and create a more inclusive environment for them.

This paper aims to address the gap in automatic emotion detection for neurodivergent individuals. We propose a multimodal emotion detection system designed to accommodate their unique emotional expressions and nuanced cues. Our preliminary findings indicate improved accuracy compared to existing studies in this field, demonstrating the effectiveness of our approach.

2. Related Work

Previous studies on emotion detection have primarily relied on single-modality approaches, with Convolutional Neural Networks (CNNs) widely used to analyze static images or extract frames from videos. One of the earliest studies by Pramerdorfer and Kampel [7] investigated the use of a standard CNN model for facial emotion recognition. However, in practical scenarios, a single modality may be compromised (e.g., poor lighting affecting facial recognition), making multimodal approaches more robust and accurate [8]. Previous research has explored audio-visual emotion recognition, demonstrating the effectiveness of multimodal approaches. For instance, Salas-Cáceres et al. introduced a model that analyzes both facial expressions and voice features, improving accuracy compared to methods that rely on just one type of data [9]. Their study also explored how tracking emotions over time can enhance recognition, which aligns with the goal of creating more adaptive and responsive systems. Similarly, Shahzad et al. demonstrated that combining facial and vocal data significantly improves

recognition accuracy [10]. Using deep learning-based fusion and classification techniques, their model outperformed unimodal approaches, achieving a 79.81% accuracy rate. However, the datasets utilized in these studies are primarily on neurotypical populations, limiting their applicability to neurodivergent individuals.

The lack of focus on emotion detection systems tailored for neurodivergent individuals, coupled with the lack of relevant datasets, presents a significant gap in current research. Addressing this limitation is crucial for developing specialized models that accurately capture the unique emotional expressions of neurodivergent populations. In this paper, we aim to address key limitations in existing research by:

- Developing an automatic emotion detection system specifically designed for neurodivergent individuals, an area that remains largely unexplored.
- Improving the accuracy of existing multimodal emotion detection systems by utilizing advanced deep learning techniques.

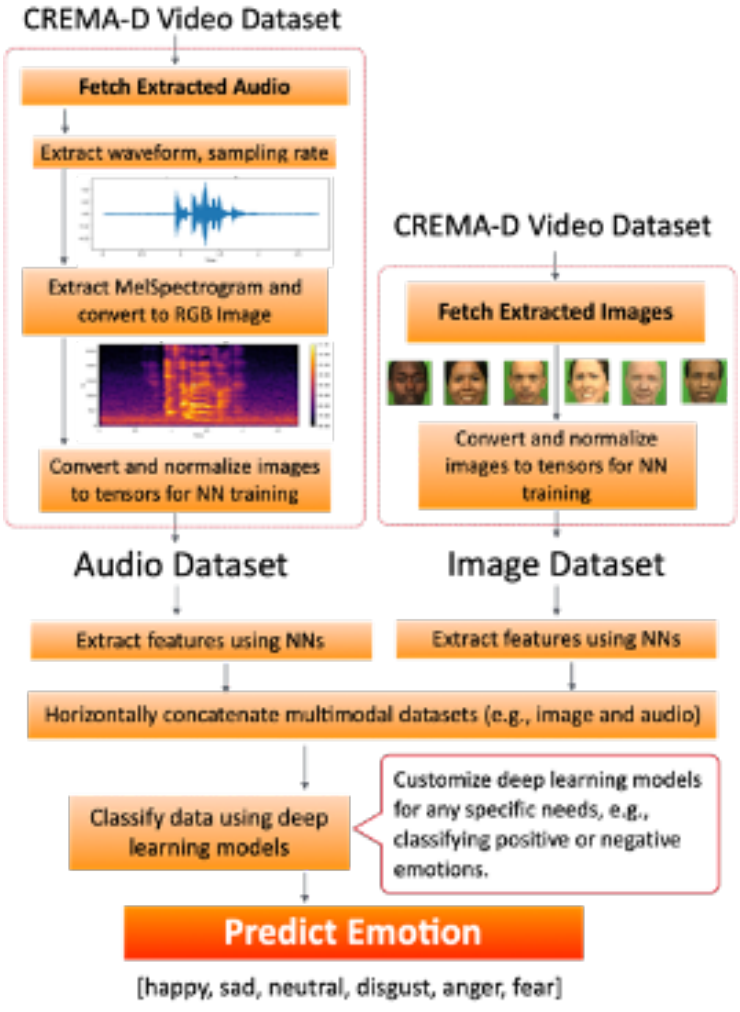
3. Multimodal Emotion Detection

A key challenge in working with multimodal data is aligning different data representations. Audio data is typically in waveform format, while visual data consists of images. To address this, we propose a streamlined workflow that converts audio signals into Mel spectrograms, which are further transformed into RGB images. These generated images are then used alongside facial expression features extracted from video frames.

Specifically, two separate pipelines were developed for audio and visual feature extraction. In the audio pipeline, audio signals are first extracted from the video and converted into waveforms. These waveforms are then transformed into Mel spectrograms, which are represented as RGB images. The images are preprocessed and prepared for training in the neural network. The visual pipeline involves extracting facial expressions from video frames as RGB images and processing them into a standardized format suitable for neural network input.

Figure 1

Automatic Multimodal Emotion Detection Pipeline



Once the audio and visual datasets are prepared, features are extracted using a popular deep learning architecture – EfficientNet [11]. This architecture is used in multiple deep learning applications like image classification, medical image analysis and object detection [13][14][15][16]. The extracted features from the audio and visual data are then concatenated to form a unified feature set, ensuring that both sources of information contribute to the final prediction. This combined feature set is then passed through a custom classification layer designed to map the extracted patterns to specific emotions. The final output of this step is the detected emotion, categorized into one of the predefined emotion classes. This approach ensures that both audio and visual cues are effectively utilized, improving the model's ability to recognize emotional expressions accurately. This final classification step can be adapted based on the desired number of emotion categories. For instance, it can be configured to distinguish between broad emotional states, such as positive and negative emotions. The final structured pipeline is illustrated in Figure 1.

4. Evaluation and Preliminary Results

We utilized the CREMA-D dataset [12] which comprises 7,441 video recordings from 91 actors (48 male and 43 female), aged between 20 and 74, representing a diverse range of races and ethnicities. The sentences were performed with one of six distinct emotional expressions: Anger, Disgust, Fear, Happiness, Neutral, and Sadness.

Table 1 compares the accuracy of leading models in this field with the proposed model. The selected papers for comparison were carefully chosen to ensure a fair and meaningful evaluation. These studies also focus on multimodal emotion recognition, specifically utilizing audio and image data, and have been conducted on the same dataset used in this work. This selection ensures consistent benchmarking and an objective performance assessment.

Table 1
Comparison of SOTA models and proposed model

Models	Accuracy (%)
Salas-Cáceres et al. (Audio + Image) [9]	80.27
Radoi and Cioroiu (Audio + Image) [17]	74.20
Shahzad et al (Audio + Image) [10]	79.80
Ghaleb et al. (Audio + Image) [18]	74.00
Proposed Model - Just Image	68.82
Proposed Model - Just Audio	60.69
Proposed Model - Audio + Image	**83.73**

As demonstrated in the table, the proposed model achieves a performance improvement of approximately 4% over the best performing model. This enhancement in accuracy may be attributed to two key factors: (a) representing audio signals as RGB images, which facilitates better integration with visual features, and (b) employing a simplified model architecture, which mitigates the risk of overfitting and enhances generalization compared to more complex models.

5. Conclusions/Future Work

This study introduces a multimodal emotion detection system that combines audio and visual data to improve the accuracy and reliability of emotion recognition. The results demonstrate that converting audio signals into RGB images and using a simplified deep learning model enhances classification performance, outperforming existing approaches.

Despite these promising results, several challenges remain. Future research will focus on expanding the dataset to include a wider range of neurodivergent individuals to improve the model's ability to generalize across different populations. Since emotional cues in neurodivergent individuals may differ significantly from those in neurotypical individuals, further refinement of feature extraction techniques is necessary. However, the promising performance of our model suggests potential for meaningful advancements in this area.

Another important direction for future work is real-time implementation. This would enable practical applications in VR-based assistive technologies, supporting education, therapy, and social communication for neurodivergent individuals.

References

[1] Boya Li, Understanding, expressing, and interacting: The development of emotional functioning in young children with autism, Ph.D. dissertation, Inst. of Psychology, Faculty of Social and Behavioural Sciences, Leiden Univ., Leiden, Netherlands, 2021. [Online]. Available: https://scholarlypublications.universiteitleiden.nl/handle/1887/3239138
[2] Shimrit Fridenson-Hayo, Steve Berggren, Amandine Lassalle, Shahar Tal, Delia Pigat, Sven Bölte, Simon Baron-Cohen, and Ofer Golan. 2016. Basic and complex emotion recognition in children with autism: cross-cultural findings. *Molecular autism* 7 (2016), 1–11.
[3] Ognjen Rudovic, Jaeryoung Lee, Miles Dai, Bj¨orn Schuller, and Rosalind W Picard. 2018. Personalized machine learning for robot perception of affect and engagement in autism therapy. *Science Robotics 3*, 19 (2018), eaao6760.
[4] Hui-Chuan Chu, William Wei-Jen Tsai, Min-Ju Liao, and Yuh-Min Chen. 2018. Facial emotion recognition with transitiondetection for students with high-functioning autism in adaptive e-learning. *Soft Computing 22* (2018), 2973–2999.
[5] Giuseppe Riva and Eleonora Riva. 2020. DE-ENIGMA: Multimodal Human–Robot Interaction for Teaching and Expanding Social Imagination in Autistic Children. *CyberPsychology, Behavior & Social Networking 23*, 11 (2020).
[6] Esubalew Bekele, Joshua Wade, Dayi Bian, Jing Fan, Amy Swanson, Zachary Warren, and Nilanjan Sarkar. 2016. Multimodal adaptive social interaction in virtual environment (MASI-VR) for children with Autism spectrum disorders (ASD). *In 2016 IEEE virtual reality (VR)*. IEEE, 121–130.
[7] Christopher Pramerdorfer and Martin Kampel. 2016. Facial expression recognition using convolutional neural networks: state of the art. arXiv preprint arXiv:1612.02903 (2016).
[8] Kaouther Ezzameli and Hela Mahersia. 2023. Emotion recognition from unimodal to multimodal analysis: A review. *Information Fusion* 99 (2023), 101847.
[9] José Salas-Cáceres, Javier Lorenzo-Navarro, David Freire-Obreg´on, and Modesto Castrill´on-Santana. 2024. Multimodal emotion recognition based on a fusion of audiovisual information with temporal dynamics. *Multimedia Tools and Applications (2024)*, 1–17.
[10] HM Shahzad, Sohail Masood Bhatti, Arfan Jaffar, Muhammad Rashid, and Sheeraz Akram. 2023. Multi-modal CNN features fusion for emotion recognition: A modified xception model. *IEEE Access* 11 (2023), 94281–94289.
[11] Mingxing Tan and Quoc Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*. PMLR, 6105–6114
[121] Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. 2014. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing* 5, 4 (2014), 377–390.
[13] Haleem Farman, Jamil Ahmad, Bilal Jan, Yasir Shahzad, Muhammad Abdullah, and Atta Ullah. 2022. Efficientnet-based robust recognition of peach plant diseases in field images. *Comput. Mater. Contin* 71, 1 (2022), 2073–2089.
[14] Gonc¸alo Marques, Deevyankar Agarwal, and Isabel De la Torre D´ıez. 2020. Automated medical diagnosis of COVID-19 through EfficientNet convolutional neural network. *Applied soft computing* 96 (2020), 106691.
[15] Oluwatosin Seyi Oyebanji, Akinkunmi Rasheed Apampa, PI Idoko, Akeem Babalola, Onuh Matthew Ijiga, Olugesun Afolabi, and Comfort Idongesit Michael. 2024. Enhancing breast cancer detection accuracy through transfer learning: A case study using efficient net. *World Journal of Advanced Engineering Technology and Sciences* 13, 01 (2024), 285–318.
[16]Pooja Yadav, Neeraj Menon, Vinayakumar Ravi, Sowmya Vishvanathan, and Tuan D Pham. 2022. EfficientNet convolutional neural networks-based Android malware detection. *Computers & Security* 115 (2022), 102622.
[17] Anamaria Radoi and George Cioroiu. 2024. Uncertainty-based Learning of a Lightweight Model for Multimodal Emotion Recognition. *IEEE Access* (2024)
[18] Esam Ghaleb, Mirela Popa, and Stylianos Asteriadis. 2019. Multimodal and temporal perception of audio-visual cues for emotion recognition. *In 2019 8th international conference on affective computing and intelligent interaction (ACII)*. IEEE, 552–558