Implementing Emotion Recognition in a Social Robot

Alexander Vatamidis Norrstam, Marcus Lindelöf, Ruben Täpptorp, Simon Rosengren

This report concerns the integration of the DeepFace framework for automatic facial expression recognition into the social robot Epi. The project aimed to implement this system and evaluate its utility in providing participant emotional state data within experimental settings. A pilot study was conducted to investigate the inter-observer reliability of the robot compared to human observers. Results indicated that the robot's analysis significantly deviated from human perception, suggesting a need for further system improvement. Consequently, it is argued, that future developments should consider multimodal approaches, such as integrating vocal intonation or body posture, to enhance the reliability of emotion detection in social robotics.

1 Introduction

Social robotics is a subset of robotics in the context of Human-Robot Interaction (HRI), which contrasts automation robotics, remote-controlled robotics, and automated vehicles (Sheridan, 2016). While there is no precise and commonly accepted definition of what constitutes a social robot, the distinguishing feature that differentiates social robots from other types is that among these, social robots are embodied and designed with the intent of communicating with humans (Akalin & Loutfi, 2021). These robots can be deployed in a wide array of contexts, such as healthcare (Ragno et al., 2023), education (Woo et al., 2021) and service operations (Vishwakarma et al., 2024).

A subset of social robotics is designed as being anthropomorphised, which is to say they are humanoid in that they display and mimic physical features of humans, such as human bodily structures and facial layout. The anthropomorphising of social robot design is partly assumed to facilitate building functional robots since their embodiment will be apt to engage in a world designed for humans. A second assumption is that it facilitates HRI by allowing humans to engage with the robots in more intuitive ways. A comprehensive meta-analysis of 78 studies showed this to be true for social robots, but not other forms of robots (Roesler et al., 2021). Given this, and that the global market for social robots is projected to reach $19 billion by the end of 2025 (World Economic Forum, 2019), further research into which anthropomorphic features facilitate HRI and in which contexts is warranted, including behavioural anthropomorphisms.

This project utilises Epi, a type of humanoid social robot developed for research purposes by the Lund University Cognitive Science division (LUCS). The quintessential physical features of Epi are the humanoid head and adjustable pupils. Most versions of Epi come in the form of a bust; a head mounted on shoulders, while one embodiment of Epi includes a torso with functional arms. Epi is operated by the open-source software Ikaros, also developed at LUCS. Through Ikaros, Epi can be utilised in Wizard of Oz experiments (WoZ), where Epi can be controlled remotely in real-time as it interacts with others. As such, present Epi has sparse cognition. More comprehensive descriptions of Epi can be found here (Johansson et al., 2020), and Ikaros here (Balkenius et al., 2010; Balkenius et al., 2020).

Facial recognition (FR) is a subfield of computer vision focused on detecting, identifying and classifying human faces, along with extracting related attributes. The field has evolved significantly since its start in the 1960's and has come to use techniques such as Deep Neural Networks (DNNs). Despite exponential technological advances the consistent theme of using vectors in feature space since the 1960's remains (Turk and Pentland, 1991). Although the fundamental computer vision architecture has remained the same the tools have expanded their scope from face identification to the detection of categorical information such as gender, approximate age, ethnicity, and emotional state (Serengil, n.d.).

Although there has been a significant advancement in FR it is not widely implemented in social robots, suggesting that the domain might be an unexplored area of research (Keizer et al., 2014; Dwijayanti et al., 2022). The implementation of facial recognition in embodied agents has shown to consist of certain challenges when utilised in non-controlled environments. Most FR models are developed using datasets featuring optimal conditions such as consistent viewing angles and ideal lighting (Serengil, 2024). That however might not be feasible to expect in the real-world in which robots need to operate in dynamic and non-ideal conditions. This is the case for Epi's visual system —akin to a myriad of robotic systems— since Epi must contend with multiple data limitations (Johansson et al., 2020). Understanding how FR systems perform under these real-world constraints is crucial for advancing the field of social robotics (Roesler et al., 2021).

The relationship between automated and human emotion recognition in social robotics by implementing the DeepFace framework in the Epi is at large part what has been explored in this project (Serengil, 2024; Balkenius et al., 2020). Through controlled experimental conditions, we examine whether automated systems can reliably detect and classify emotional states in ways that correspond to human observations. The implementation leverages Epi's existing capabilities in Wizard of Oz experiments, where the robot engages in naturalistic interactions while being remotely operated (Johansson et al., 2020). By comparing automated emotion detection with human observer assessments, this project aims to validate the potential for using facial expression analysis as a tool for data collection in human-robot interaction research. The findings could inform future developments in social robotics, particularly in contexts requiring accurate emotional state detection (Ragno et al., 2023; Woo et al., 2021).

2 Implementation

Choice of framework: DeepFace

Multiple techniques for FR have been developed over the years, though they typically share a similar process pipeline: detect, align, normalise, encode/represent, and verify (Serengil, 2024). In a comprehensive benchmark analysis of FR pipelines, four dimensions were analysed: FR, face detectors, the effects of alignment mode activation or deactivation and distance metrics (Serengil & Özpinar, 2024). DeepFace, developed by Sefik Ilkin Serengil, is a framework for FR tasks such as detection, identification, validation, and facial attribute analysis. The framework combines multiple packages and models into a single Python library for a complete end-to-end FR pipeline.

DeepFace was chosen because of its versatility by virtue of incorporating multiple models into one framework. Thus, implementing DeepFace facilitates initial prototyping of FR models since different models may excel or underperform in different contexts, which has been evaluated previously (Serengil, 2024). Serengil is also invested in the FR community and consistently responds to queries concerning DeepFace, including one from this project group.

Implementing Face Recognition and Gaze Locking

The main problem with this implementation is that the DeepFace face recognition tool does face recognition on images. Meaning that to get our software to work in real time, we would need to extract frames from the video and run them through the face analyser/gaze one by one. To do this we used cv2, a python library developed for handling images. The cv2 library comes with a .read() method which can capture and single out frames from any video source that is set up via cv2. This together with the different face analysis method from deepface, as well as the possibility to retrieve X and Y values made for a rather simple setup. The only thing remaining was sending the movement commands to Epi, which is done via https commands on Epis private network, and then gaze locking was somewhat done. Now Epi moved its head depending on where in the frame epi saw a face. However, this movement did not make any sense. If someone moved right in the camera, epi did not necessarily move right. This was mainly due to Epis warped camera and the way commands for Epis movement are set up. When using direct commands for Epis movement, for example “Move to [X:45, Y:90]”, there are two problems that can occur depending on how exactly you use Epi. One is that, between each movement, Epi always resets at the center position first. This made movement static and unsettling. The other is that Epi calculates where to move depending on its current position. For example, if Epi already is at [X:90, Y:0], if we want Epi to move to [X:120, Y:30], we would have to send a command for [X:30, Y:30]. Meaning we have to account for its current position.

There was code developed to handle these issues. However, at the same time as that code was developed, the project shifted focus from face recognition and gaze locking to instead primarily focus on emotion analysis. This was done to progress our work on the experiment. We never actually tested our developed code, so for now, Epi reacts to movement, but does Epis corresponding movement does not make any sense. Gladly this was not all for nothing. A lot of the code for gaze locking was later used for live emotion analysis.

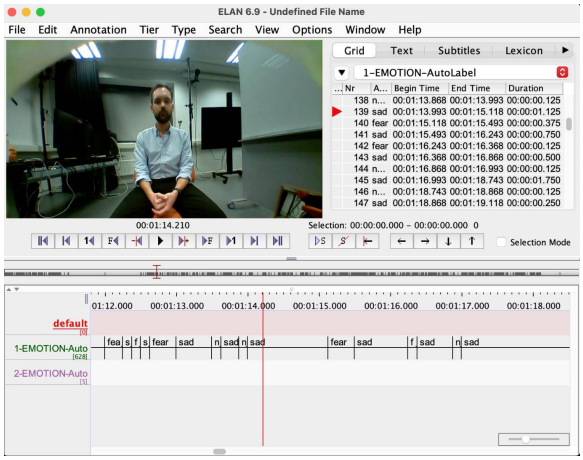

Figure 1. Screenshot of ELAN

Facial expression analysis with DeepFace

As mentioned earlier DeepFace is also capable of facial attribute analysis, meaning DeepFace has the functionality to perform facial expression analysis. Similar to how their face recognition works, the standard detector backend is set to opencv. Given that we used retinaface as the detector backend for our face recognition we decided to also use retinaface as detector backend when using our emotion analyser. To emphasize, the detector backend is referring to which tool is used to detect the face in the picture. To our understanding, this does not affect the emotion analysis. Unlike how you can choose between lots of different detector_backends for the face detection, there seems to be no way to choose emotion analyser through DeepFace. Two emotion analysers were to be implemented. One for real time analysis and another for offline analysis later on. The real time analyser was meant to work as a proof of concept for possible later use. For example, Epi being able to mirror the participants emotions during tests.

DeepFace comes with a method named “streaming” which in essence is all of their functionalities cramped into one method that does everything live. At first we thought this would work great for our real time analyser, but we quickly realised that we wanted more flexibility and control over what the method was supposed to do and not do. Streaming was quickly scrapped and we moved on to making our own real time analyser.

Like face detection, face analysis works by analysing an already existing image file. We had already found a workaround for this, using cv2 software, with the live face analysis method we had made previously. Mainly the only thing we had to do was to change that existing method from using face detection, to now having face analysis, and the live face analysis was complete. However, the live analysis requires a lot more processing power, and since we do not actually need live analysis for this experiment we decided to also make an offline analyser. This meant that during the experiment we only had to record a video through Epis cameras and then later run those videos through the offline analyser.

The facial expression offline analyser, *offline-emotion-analyzer.py*, used the RetinaFace backend for face detection. A "frame skip" function was added as each frame takes about 1.5 seconds to analyse on a modern Apple laptop, and the temporal resolution of 25 to 40 readings per second seemed superfluous. The program's output is saved as a CSV file with one row per reading.

An additional program for converting the CSV file to an ELAN-compatible TXT files, *convert_csv_to_elan.py*, was also constructed so that the dominant facial expression could

easily be imported for further processing. ELAN is a popular video annotation software with extensive use in the field of cognitive science. The ability to automatically code ("auto label") facial expressions could be a massive time saver for researchers using ELAN.

3 Method

Experiment

The experiment was an HRI experiment employing a WoZ methodology with the robot Epi. For this experiment, Epi's left eye was used to collect video footage of the participants. A custom experiment paradigm, named Emotion and Stress Evoking Protocol, ESEP, was devised with the goal of evoking diverse emotional reactions, and thus facial expressions, which could be analysed by Epi and human observers. The protocol was designed to be similar to typical experiments within the field of psychology and cognitive science, and sought to test whether the DeepFace facial attribute analysis method could generate reliable data for other experiments.

Query	Script
1	Hello and welcome to our study! My name is Epi, and I will be interviewing you today. Thank you for taking the time to participate. Please have a seat in the chair in front of me and let me know when you're ready.
2	We will be doing a few activities and asking some questions. There are no right or wrong answers, so feel free to share your thoughts and experiences. And please try to keep your answers within two minutes.
3	Please remember that your participation is entirely voluntary. You can withdraw at any time without providing a reason. Are you okay with continuing?
4	Great!
5	I'd like to ask you a few questions about personal experiences. Could you tell me about one of your best memories, something that truly made you happy?
6	That sounds fantastic! What was it about that experience that made it so special for you?
7	Thank you for sharing.
8	Have you ever encountered something that you found really disgusting or repulsive?
9	Can you explain why that experience was disgusting?
10	Thank you!
11	Could you tell me about a time when you felt sad or disappointed?
12	How did that experience affect you, and how did you cope with it?
13	Thank you for sharing.
14	Can you recall a situation where you felt really frustrated or angry?
15	What was the most frustrating aspect about that situation?
16	When thinking back to that experience, do you still get angry?
17	Thank you for sharing.
18	Next, we will move onto working memory tasks. First however, I need to reboot. Feel free to take it easy and mentally prepare in the meantime.
19	Okay I am now rebooted and prepared for our second task. Now we are ready to start.
20	Your next task is a working memory task. I will read you a list of words and I want you to sort them alphabetically. If I say Pear, Apple, Orange, I want you to read back to me Apple, Orange, Pear. Are you ready?
21	Bed, Rest, Awake
22	Tired, Dream, Snooze, Nap
23	Peace. Yawn. Snore. Blanket. Drowzy.
24	Very good. Thank you.
25	Could you count backward from 1022 in steps of 13 as quickly as possible?
26	Try to do it faster, and make sure it's correct. If you make a mistake, please start again from 1022.
27	Wait
28	I am sorry for the interruption. Please start again from 1022.
29	Great. Thank you.
30	Thank you so much for participating in all these tasks and sharing your thoughts with us today. We really appreciate your participation. Have a great rest of your day!

Table 1. All prompts used during the experiment.

The ESEP protocol consisted of two distinct phases: an emotion elicitation phase and a stress induction phase. During the emotion elicitation phase, Epi engaged participants in personal dialogue, posing questions designed to elicit emotional responses grounded in personal narratives. These questions gauged for different emotions, and thus evolved from general prompts (e.g., queries 5, 8, and 14 in Table 1) to more focused emotional probes (e.g., queries 6, 9, and 15 in Table 1).

The stress induction phase incorporated working memory tasks. As was the case in the previous phase, Epi had predetermined movements alongside most responses. However, during this session, there was a simulated malfunction movement intended to evoke stress responses from participants, as opposed to facilitating the interaction.

To control the ESEP a custom Python program was created. The script was loaded by the program from a CSV file containing both the spoken parts and which pre-recorded animation to trigger. Both speech and animation triggers were called using HTTP GET requests (see Johansson, n.d.).

The animations used most of Epi's motors and LEDs. The LED eyes conveyed emotional signals through colour changes (blue for sadness, green for approval, and red for simulated malfunction), head movements to facilitate a natural interaction and its speech function, and pupil dilation signalled interest. The speech was vocalised by Apple's South African English voice (Tessa Enhanced), played through the speaker in Epi using the Ikaros module EpiSpeech.

Participants

For the experiment, six participants were recruited through word of mouth. Due to technical issues with data collection with one of the participants, the data for that participant had to be omitted. This yielded five participants (four female) included in the final data collection, with a mean age of 28.2±3.5 years. All participants signed informed consent prior to starting the experiment.

Procedure

The experiment was conducted at the Robotics Lab at LUCS. Participants were informed that Epi would ask a series of questions of them, and they were seated in front of Epi at a socially appropriate distance while being close enough to allow for the DeepFace algorithm to reliably detect the faces of the participants. While Epi spoke in English, participants were informed they could answer in both English and Swedish, depending on whichever felt more comfortable.

Prior to the experiment starting, the participants got to sign informed consent (Appendix A), and the room was evacuated by everyone except the experiment leader and the participant. The experiment leader remained in the room under the pretense of being a fail-safe should the robot start to malfunction.

Once the experiment finished, the participants were asked to fill out the questionnaires. In addition, the experiment leader conversed with the participants about their impressions of the experiment.

Data Collection

The primary measuring point of this experiment was to assess the viability of using the DeepFace algorithm for emotion detection by quantifying how much Epi would potentially deviate from human observers. As such, the main material was the video recording gathered from Epi's left eye during the experiment. The video was then analysed by the DeepFace algorithm using the previously mentioned offline-emotions-analysis.py program, and two independent, naïve human observers not present during the experiment using the video annotation software ELAN (ELAN, 2024). The assessments were then compared with each other to assess how much they deviate from each other.

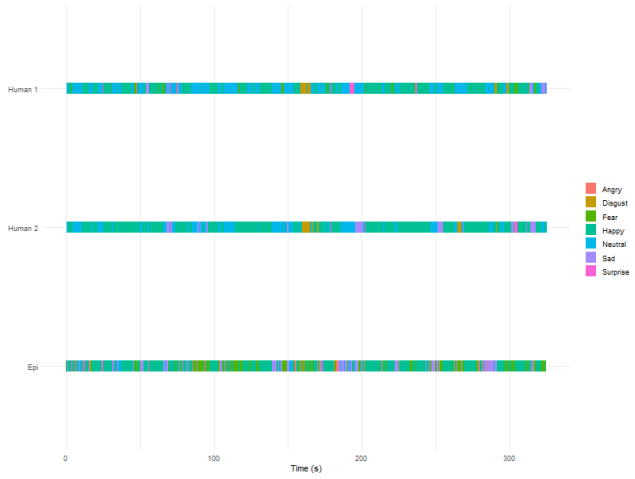

Figure 2. Illustration of how observers differed in their emotion attribution.

In addition to the primary measure, additional measures were collected in the form of different questionnaires issued after the experiment. The purpose of these measures were cumulatively to complement the primary analysis and control for other variables, should there have been enough participants to make these data points valuable. The first questionnaire was the Ten-Item Personality Index (Appendix B), assessing personality traits in the Big-Five model of personality (Watson et al., 1988). The second questionnaire assessed the degree of comfortability the participants felt throughout the experiment (Appendix C), the questions and format of which were adopted and revised from an earlier study concerning comfortability in HRI (Redondo et al., 2024). The final questionnaire primarily concerned stress levels (Appendix D), and was adopted from an earlier thesis involving Epi inducing stress to the participants (Sikström, 2021).

Data Analysis

The video collected by Epi during the experiment was analysed by DeepFace and two human observers. The settings used for DeepFace meant it assessed the facial expressions of the participant eight times per second, and with each assessment there was a probability vector indicating how probable different facial expressions were indicative of different emotions. The video did not include audio, and thus the human observers, like DeepFace, only had access to visual information.

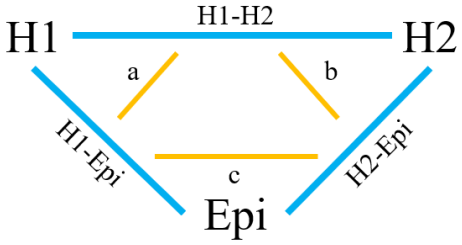

Figure 3. Illustration of the relationships being measured. Blue arrows indicate HDQs. Yellow arrows indicate relationships that can be significantly different.

To quantify observer discrepancies in facial expression assessment, a hamming distance quotient (HDQ) was calculated. Hamming distance is a measure of how similar two vectors are by counting the number of deviations there are between the two. This is illustrated in figure 2, where the vectors are coloured according to the designated emotion for a specific time. Here, it meant that the exported ELAN data could be interpolated and standardised according to the output of the DeepFace algorithm, enabling a count of how much the observers deviated from each other across the videos. This meant that comparisons, on account of DeepFace assessing facial expressions eight times per second, had a 0.125 second granularity. Because the videos were of different length, the hamming distance was converted to a quotient to assess the relative discrepancies across videos, i.e., dividing the total number of discrepancies divided by the total number of facial expression assessments per observer. This quotient is the HDQ used, which in other words measure how much two observers deviate from one another proportional to all observations. Because this measures discrepancies between observers, and the three observers in this setup being Human 1 (H1), Human 2 (H2) and Epi,this yielded three comparisons: H1-Epi, H2-Epi, H1-H2. The analysis then is to assess whether observer discrepancies significantly differ between one another (see Figure 3).

4 Results

The experiment lasted for an average of 11.4±1 minutes, with the emotional phase taking 5.7±0.8 minutes and the stress phase taking 4.2±0.2 minutes. For the H1-H2 comparison, the HDQ was 0.33 ± 0.11, indicating that the human observers agreed with each other 67% of the time. The H1-Epi HDQ was 0.55 ± 0.13, and the H2-Epi HDQ was 0.51 ± 0.13.

A repeated-measures ANOVA was used to evaluate whether the observer couple had any effect on the HDQ, which indicated a significant effect ($F(2, 18) = 16.2$, $p < 0.000$). Post-hoc pairwise comparisons using estimated marginal means revealed significant differences between the H1-H2 HDQ and the H1-Epi HDQ ($M = -0.22$, $SE = 0.04$, $t(18) = -5.35$, $p < 0.001$), and between the H1-H2 HDQ and H2-Epi HDQ ($M = -0.18$, $SE = 0.04$, $t(18) = -4.36$, $p = 0.001$), represented by lines a and b in Figure X. No significant difference was observed between the H1-Epi HDQ and the H2-Epi HDQ ($M = 0.04$, $SE = 0.04$, $t(18) = 0.98$, $p = 0.60$), represented by the line c in Figure X.

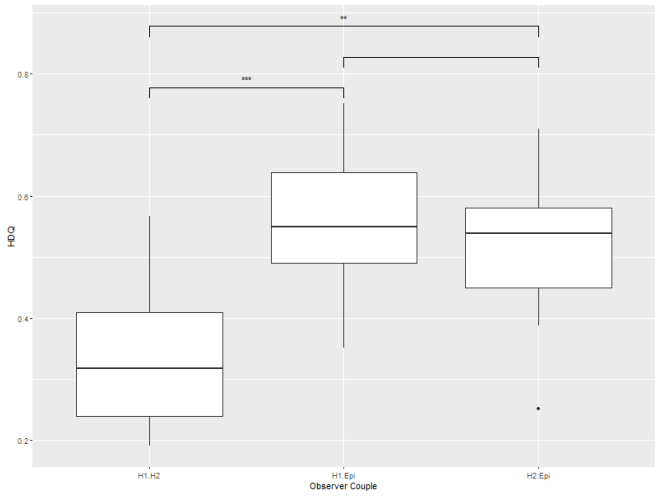

Figure 4. Boxplot of the HDQs across observer couples.

Due to the low participant count, the secondary measures have not been properly analysed or factored into the primary analysis. Plots showing the distribution of the data gathered from the secondary measurements can be found in the Appendix. The Ten-Item Personality Inventory showed a tendency towards high values across personality traits, particularly the trait "Openness to Experiences" (Appendix E). The

Comfortability Self-Report indicated that participants felt rather comfortable initially, with comfortability being reduced as the experiment went on (Appendix F). The Stress Self-Report suggested that Epi can induce stress, as indicated by the increased self-reported levels of stress during the interaction as compared to before the interaction (Appendix G). The qualitative answers provided in the Stress Self-Report showed a borderline universal inclination towards experiencing the mental arithmetic task to be the most stress-inducing.

5 Discussion

Implications for facial and emotion detection algorithms

Current algorithms, while sophisticated in controlled environments, show significant deviation from human perception when deployed in real-world settings. The discrepancy between human observer agreement (67%) and algorithmic detection suggests that existing models may need substantial refinement to better align with human emotional understanding (Serengil, 2024). This challenge is compounded by environmental factors inherent to robotic deployment - varying lighting conditions, inconsistent viewing angles, and dynamic user movements - which stand in stark contrast to the controlled conditions under which these algorithms are typically trained.

The current approach of classifying emotions into discrete categories appears overly simplistic when compared to the nuanced way humans interpret emotional states. Our human observers frequently noted emotional expressions that fell outside DeepFace's seven categorical options, suggesting that future algorithms may need to adopt more sophisticated classification schemes (Turk and Pentland, 1991). The frame-by-frame analysis approach employed by current algorithms fails to capture the multimodal, temporal context that humans naturally integrate into their emotional assessments. This limitation becomes particularly apparent in real-time applications, where the sequential nature of emotional expressions may carry significant meaning beyond individual frame analysis (Roesler et al., 2021).

Implications for social robotics

The implementation of facial expression recognition in Epi revealed several challenges for social robotics. The technical limitations of Epi's camera system - including limited resolution, narrow dynamic range, and variable field of view - demonstrate how hardware constraints significantly impact emotion detection reliability (Johansson et al., 2020). These limitations suggest that social robots require either more sophisticated sensory hardware or compensatory algorithmic approaches. It is also the case that the need for multimodal approaches becomes apparent, as human emotion recognition integrates multiple channels including vocal intonation and body posture (Roesler et al., 2021). Our HDQ results indicate that relying solely on facial analysis may be insufficient for meaningful human-robot interaction.

Our findings revealed a significant gap between human and algorithmic emotion detection, with human observers achieving 67% agreement while Epi's assessments showed notable deviations. This discrepancy has particular implications for applications in healthcare and education (Vishwakarma et al., 2024). The challenge is compounded by real-time processing requirements, where our implementation of DeepFace required approximately 1.5 seconds per frame (Serengil, 2024), creating a critical tradeoff between analysis depth and response speed. However computational advances and more effective algorithms should solve problems of this nature fairly seamlessly.

These challenges suggest that future social robot design needs to reconsider its approach to emotion detection uncertainty. Rather than pursuing perfect classification, robots might benefit from probabilistic approaches that maintain multiple hypotheses about emotional states (Roesler et al., 2021). Beyond that, our observation of consistent stress increases during cognitive tasks, particularly during mental arithmetic, raises some ethical considerations. The combination of anthropomorphic features with simulated malfunctions appeared to heighten stress responses (Watson et al., 1988; Sikström, 2021), suggesting the need for "stress-aware interaction protocols" that can monitor and adapt to user stress levels, particularly crucial in sensitive applications like healthcare and education (Ragno et al., 2023; Woo et al., 2021).

Limitations

There are several limitations with the current work. First, the small sample size meant that the data from the questionnaires were not usable for secondary analyses. A larger sample size would have enabled further analyses, such as controlling for personality traits and examining whether there were facial expressions that correlated with the self-reported comfortability and stress levels.

Second, the experiment did not collect data on what emotions the participant themselves experienced during the experiment. This meant that there is technically no way to determine who was more right between H1, H2, and Epi when it came to their emotion attribution. The HDQ as applied here could be used to measure discrepancies between observers and the subjects themselves, rather than, or in addition to, measuring inter-observer reliability. This would have provided a grounding where one could assess in objective metrics the degree to which the facial expression analysis deviates from the subjects as compared to human observers. Such a comparison could shed light on when and how automatic facial expression analysis can be applied to accrue reliable data on par with or improved beyond human observers.

Third, the work was limited by DeepFace categorising facial expressions as belonging to one of only seven emotional categories. Notably, the human observers independently of one another vocalised the issues with such a narrow restrictions, and claimed to see a lot of other emotions not compatible with DeepFace categories, such as disappointed, tired and bored. Further, it remains controversy in emotion research whether emotions ought to be conceptualised as categories at all, or if instead they are better conceptualised as points on multidimensional continuums (e.g., valence-arousal spectrums). Regardless, given that this is a limitation of DeepFace in its current state, little can be done at present beyond waiting for the technology to become better in the near future or choosing another facial expression analysis tool.

A final limitation is that the HDQ does not account for degrees of discrepancies. Hamming distance is a binary measure, merely counting how many values differ at all between vectors. This means that observers disagreeing about whether a facial expression denotes happy or neutral weighs as much as disagreeing whether the facial expression denotes happiness or anger. This point is especially pertinent if emotions are conceptualised as multidimensional continuums, where the distances can be quantified. If a facial

expression analysis could employ such an analysis, the HDQ could measure the distance in the conceptual space for each point, yielding a better measurement, given the validity of the face recognition model.

Future Research

There are multiple ways in which the current work can be expanded. One way is to perform further analyses to see how HDQ potentially varies as dependent on other variables. For instance, one could assess whether certain emotion classifications are more likely than others to yield disagreements between observers, or if different stimuli periods could affect the HDQ.

As discussed in the limitations, another route for future work is to use a facial expression analysis that measures emotion along a set of dimensions. The analysis could be run post-hoc on the video material, and then compared to human observers and facial recognition models using categorical classification to assess the viability of the dimensional model.

A third route is to consider multi-modal integration of multiple sensory cues. Emotion expresses itself in more ways than facial expressions, and as such there is more information to garner by collecting data from other cues as well. This includes other visual cues (e.g., posture), as well as cues from other modalities, notably the auditory to discern cues such as tone of voice, sentence structure and semantics.

References

Akalin, N., & Loutfi, A. (2021). Reinforcement Learning Approaches in Social Robotics. *Sensors*, *21*(4), Article 4. https://doi.org/10.3390/s21041292

Balkenius, C., Johansson, B., & Tjøstheim, T. A. (2020). Ikaros: A framework for controlling robots with system-level brain models. *International Journal of Advanced Robotic Systems*, *17*(3), 1729881420925002. https://doi.org/10.1177/1729881420925002

Balkenius, C., Morén, J., Johansson, B., & Johnsson, M. (2010). Ikaros: Building cognitive models for robots. *Advanced Engineering Informatics*, *24*(1), 40–48. https://doi.org/10.1016/j.aei.2009.08.003

Dwijayanti, S., Iqbal, M., & 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. IEEE Access. https://doi.org/10.1109/ACCESS.2022.3200762

ELAN (Version 6.9) [Computer software]. (2024). Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive. Retrieved from https://archive.mpi.nl/tla/elan

Johansson, B. (n.d.). _Additional instructions_. GitHub. https://github.com/birgerjohansson/Epi/wiki/Additional-instructions

Johansson, B., Tjøstheim, T. A., & Balkenius, C. (2020). Epi: An open humanoid platform for developmental robotics. *International Journal of Advanced Robotic Systems*, *17*(2), 1729881420911498. https://doi.org/10.1177/1729881420911498

Keizer, S., Ellen Foster, M., Wang, Z., & Lemon, O. (2014). Machine Learning for Social Multiparty Human—Robot Interaction. *ACM Trans. Interact. Intell. Syst.*, *4*(3), 14:1-14:32. https://doi.org/10.1145/2600021

Ragno, L., Borboni, A., Vannetti, F., Amici, C., & Cusano, N. (2023). Application of Social Robots in Healthcare: Review on Characteristics, Requirements, Technical Solutions. *Sensors*, *23*(15), Article 15. https://doi.org/10.3390/s23156820

Redondo, M. E. L., Niewiadomski, R., Rea, F., Incao, S., Sandini, G., & Sciutti, A. (2024). Comfortability Analysis Under a Human–Robot Interaction Perspective. *International Journal of Social Robotics*, *16*(1), 77–103. https://doi.org/10.1007/s12369-023-01026-9

Roesler, E., Manzey, D., & Onnasch, L. (2021). A meta-analysis on the effectiveness of anthropomorphism in human-robot interaction. *Science Robotics*, *6*(58), eabj5425. https://doi.org/10.1126/scirobotics.abj5425

Serengil, S. (n.d.). *GitHub - serengil/deepface: A Lightweight Face Recognition and Facial Attribute Analysis (Age, Gender, Emotion and Race) Library for Python*. GitHub. https://github.com/serengil/deepface?tab=readme-ov-file

Serengil, S. (2024, July 23). *A gentle introduction to face recognition in deep learning*. Sefik Ilkin Serengil. https://sefiks.com/2020/05/01/a-gentle-introduction-to-face-recognition-in-deep-learning/

Serengil, S., & Özpınar, A. (2024). A Benchmark of Facial Recognition Pipelines and Co-Usability Performances of Modules. *Bilişim Teknolojileri Dergisi*, *17*(2), 95–107. https://doi.org/10.17671/gazibtd.1399077

Sheridan, T. B. (2016). Human–Robot Interaction: Status and Challenges. *Human Factors*, *58*(4), 525–532. https://doi.org/10.1177/0018720816644364

Sikström, K. (2021) *The Trier Social Stress Test with an Epigenetic Humanoid Robot*. [Master's Thesis, Lund University]. Lund University Publications Student Papers.

Turk, M., & Pentland, A. (1991). Eigenfaces for Recognition. *Journal of Cognitive Neuroscience*, *3*(1), 71–86. https://doi.org/10.1162/jocn.1991.3.1.71

Vishwakarma, L. P., Singh, R. K., Mishra, R., Demirkol, D., & Daim, T. (2024). The adoption of social robots in service operations: A comprehensive review. *Technology in Society*, *76*, 102441. https://doi.org/10.1016/j.techsoc.2023.102441

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, *54*(6), 1063–1070. https://doi.org/10.1037/0022-3514.54.6.1063

Woo, H., LeTendre, G. K., Pham-Shouse, T., & Xiong, Y. (2021). The use of social robots in classrooms: A review of field-based studies. *Educational Research Review*, *33*, 100388. https://doi.org/10.1016/j.edurev.2021.100388

World Economic Forum. (2019). *Top 10 Emerging Technologies 2019*. https://www3.weforum.org/docs/WEF_Top_10_Emerging_Technologies_2019_Report.pdf