Using Facialy Expressive Robots to Calibrate Clinical Pain Perception

Maryam Moosaei¹, Sumit K. Das², Dan O. Popa², Laurel D. Riek³

¹Computer Science and Engineering, University of Notre Dame, USA
²Electrical & Computer Engineering, University of Louisville, USA
³Computer Science and Engineering, University of California, San Diego, USA

ABSTRACT

In this paper, we introduce a novel application of social robotics in healthcare: high fidelity, facially expressive, robotic patient simulators (RPSs), and explore their usage within a clinical experimental context. Current commercially-available RPSs, the most commonly used humanoid robots worldwide, are substantially limited in their usability and fidelity due to the fact that they lack one of the most important clinical interaction and diagnostic tools: an expressive face. Using autonomous facial synthesis techniques, we synthesized pain both on a humanoid robot and comparable virtual avatar. We conducted an experiment with 51 clinicians and 51 laypersons (n = 102), to explore differences in pain perception across the two groups, and also to explore the effects of embodiment (robot or avatar) on pain perception. Our results suggest that clinicians have lower overall accuracy in detecting synthesized pain in comparison to lay participants. We also found that all participants are overall less accurate detecting pain from a humanoid robot in comparison to a comparable virtual avatar, lending support to other recent findings in the HRI community. This research ultimately reveals new insights into the use of RPSs as a training tool for calibrating clinicians’ pain detection skills.

CCS Concepts

•Computer systems organization → Robotics; •Social and professional topics → Medical technologies; •Applied computing → Health informatics;

Keywords

patient simulation; facial expressions; emotion; social robots; humanoid robots; affective communication; non-verbal communication; human robot interaction; medical technologies; health informatics

1. INTRODUCTION

Robots are becoming more common in our daily lives, and soon robots and people will be interacting proximately [50]. There is a desire among robotics researchers and designers to enable robots to communicate quickly and effectively with humans. While there are ongoing questions within the Human Robot Interaction (HRI) community about the most appropriate form of communication between robots and people and the degree of anthropomorphism robots should employ, nonverbal communication is an exceptionally efficient means for communication which people are adept at perceiving [52, 50, 12, 1, 56, 69]. Therefore, it is worth exploring both how to enable robots to express nonverbal behaviors, and explore how successfully they are perceived by people.

People employ multiple kinds of nonverbal signals when communicating with each other, including: changing the prosody of their voice, performing gestures, shifting their posture, signaling attention with their eyes or head direction, and creating facial expressions. Across all of these communication modalities, the face is the most spontaneous channel for the rapid communication of social and emotional displays. It can be used to convey not only basic emotions such as fear, sadness, and disgust, but also complex cognitive mental states, such as confusion, contemplation, and nostalgia [53, 4]. Facial expressions can also be used to enhance conversation, show empathy, and acknowledge the actions of others [6].

Thus, employing human-like facial expressions may allow robots to communicate and interact with humans more effectively in some domains [36]. Several applications for facially expressive robots and virtual avatars bridge the fields of entertainment, education, community healthcare, and neurorehabilitation. Our work focuses on a specific medical application of facially expressive robots: robotic patient simulators (RPS). RPSs are the most frequent application of humanoid robots. They are human-sized robots that can breathe, bleed, react to medication, and convey hundreds of different biosignals (see Fig. 1). RPSs enable clinical learners to safely practice clinical skills without the fear of harming real patients, and can include: patient communication, patient condition assessment, and procedural practice. RPS usage can help reduce preventable medical errors, which kill approximately 400,000 people per year in hospitals alone, and are the third leading cause of death in the United States [28, 15].

Although commonly used in clinical education, one capability of RPSs is in need of attention: enabling face-to-face interactions between RPSs and learners. Despite the importance of patients’ facial expressions in making diagnostic decisions, commercially available robotic patient simulators are not currently equipped with expressive faces. They have static faces that cannot express any visual pathological signs or distress. Therefore, clinicians can fall into the incorrect habit of not paying attention to a patient’s face, which can result in them going down an incorrect diagnostic path.

Almost all areas of healthcare include face-to-face interaction [40] and healthcare providers who are skilled at detecting communication cues have improved outcomes, better patient satisfaction,
expressions of pain were perceived using a virtual patient simulation [42], and tested how well synthesized patient facial expression decoding.

Thus, perhaps physically-embodied, expressions with less accuracy than laypersons [21]. Some researchers have hypothesized this is due to the decline in empathy as clinicians progress through their training [27]. Unfortunately, ignoring patients’ emotional needs can lead to incorrect diagnostic decisions, poorer health outcomes, and reduced patient satisfaction [70, 30].

Fortunately, training can be effective in improving clinicians’ ability to detect patient facial expressions [23, 11]. Thus, RPSs may prove to be an excellent training tool to facilitate this learning process. Certainly this effect has been shown to be true in the virtual space (c.f. [20]); however, it is difficult to use virtual simulations in situ at the bedside, where clinicians must visually assess patient faces while simultaneously conduct physical exams [29, 41].

Furthermore, augmented reality and projected displays are unfortunately ill-suited to most hospital simulation domains due to space and power issues [22], and can cause physical barriers between the patient and clinician. Thus, perhaps physically-embodied, expressive RPS systems may be an effective educational intervention for patient facial expression decoding.

In previous work, Moosaei et al. introduced a new method for patient-driven pain synthesis [42], and tested how well synthesized expressions of pain were perceived using a virtual patient simulator [41]. This study replicated an experiment by Riva et al. [55], who used manual Facial Action Coding System (FACS)-animation to synthesize pain on virtual avatars. Moosaei et al. found that participants are able to accurately distinguish pain from commonly conflated expressions (anger and disgust), and also showed that automatic, patient-driven pain synthesis is a viable alternative to laborious key-frame animation synthesis techniques.

While this was an interesting finding, we were curious to know if the results might change if the expression synthesis was conducted on a humanoid robot (in comparison to a matched virtual avatar). Furthermore, we were curious if there might be differences between clinicians and non-clinicians in their ability to accurately classify and perceive painful facial expression intensity.

Therefore, we have three main research questions. First, are people more accurate in detecting pain expressed on a humanoid robot compared to a comparable virtual avatar? While other researchers in the HRI community have explored the robot vs. avatar question with somewhat conflicting results ([7, 9, 33]), the work has often been conducted on zoomorphic or other non-anthropomorphic platforms, and usually just focused on the six basic emotions. Both our application space (high-fidelity patient simulation) and our intended user population necessitates a higher-degree-of-freedom (DOF), human-like, facially expressive robot, and also necessitates the synthesis of more nuanced, varied-intensity expressions (such as pain, neurological impairment, etc).

Second, does having a clinical background affect synthesized pain detection accuracy? The literature suggests that clinicians might have lower accuracy in overall facial expression detection compared to laypersons [21]. Additionally, other researchers have found that clinicians tend to miss emotional cues and signs of distress in patients [43, 35, 23]. Based on these findings, we predict that clinicians will be overall less accurate in detecting synthesized painful facial expressions compared to lay participants.

Third, do participants with a clinical background perceive pain intensity on robots and avatars differently than lay participants do? Some researchers have shown that clinical experience with patients results in underestimating patients’ pain [47, 17, 48]. Based on these findings, we predict that the clinical trend toward an underestimation bias will also be present when rating intensity of synthesized pain (e.g., they will not be as good as laypersons).

We address these research questions through an experiment in which participants perceived the expression of and the intensity of synthesized facial expressions (pain, anger, disgust) depicted on either a humanoid robot or its virtual model. We describe our methodology in Section 3, results in Section 4, and contextualize our findings for both the HRI and clinical education communities in Section 5.

2. BACKGROUND

Facial expressions of pain are important nonverbal communication signals, particularly in healthcare [25, 47]. Communicating pain via facial expressions is important because it signals to clinicians that a patient might need medical attention. While self-reporting is the main method of assessing pain, there are issues with this method. The most critical issue is that self-report cannot be used for children or patients with communication challenges such as patients with cognitive impairments or unconscious patients. In such cases, an observer must visually assess patients’ pain. However, there are differences between how clinicians and how patients conceptualize pain (clinicians are substantially impaired at pain perception), which leads to problems when making diagnostic decisions for these patients [47, 66, 2, 39].
We performed a pilot study. This resulted in 18 stimuli videos. Finally, to decide which stimuli videos to include in the main study, we ran a pilot study. This resulted in 18 stimuli videos.

While, a fair number of studies are available on automatic pain detection [2], there are only a few on automatic pain synthesis on virtual avatars [55, 41], and, to our knowledge, none on patient-driven facial expression synthesis on physical robots. Moreover, none of these synthesis studies included clinicians as participants, despite the fact that clinicians perceive pain differently (worse) than lay people. Thus, in our work, we decided to explore pain facial expression synthesis on a humanoid robot in addition to its virtual model, and included both lay participants and clinicians in our study to explore if a clinical education affects pain perception.

This work builds on work by Moosaei et al., which extended the state-of-the-art in facial expression synthesis by employing models of patients to automatically synthesizing pain on virtual avatars [41]. The researchers used a CLM-based face tracker [5] to extract 66 feature points frame-by-frame from videos from two facial expression data sets: the McMaster-UNBC Pain Archive [38] and the MMI database [45]. They then mapped the extracted feature points to control points of a virtual avatar in Source SDK, and conducted a perceptual experiment. The researchers had several notable and relevant findings. First, they found that participants had higher accuracy in detecting pain compared with anger and disgust using naturalistic performance-driven synthesis. Second, they found that there is no significant relation between an avatar’s gender and pain detection accuracy. Finally, they showed that realistic pain synthesis on virtual avatars is comparable with manual pain synthesis, and is even arithmetically better [41].

3. METHODOLOGY

We used performance-driven synthesis to map facial expressions of pain, anger, and disgust onto both a humanoid robot and a comparable avatar, employing techniques similar to those described by Moosaei et al. [42, 41]. We applied a Constrained Local Model (CLM)-based face tracker to realistic source videos acquired from the Binghamton Pittsburgh 4D Spontaneous Expression Database (BP4D-Spontaneous) [68].

Our stimuli creation process included three steps. First, using a CLM-based face tracker [5], we extracted 66 feature points frame by frame from each source video. Second, for each source video we mapped the extracted feature points to servo motors of Philip K. Dick (PKD) [24, 26] (a humanoid robot from Hanson Robotics which resembles the late science fiction author) and control points of a modified virtual character from Steam Source SDK. We mapped the five source videos of the three expressions (pain, anger, and disgust) to two embodiments (PKD and avatar) resulting in 30 stimuli videos. Finally, to decide which stimuli videos to include in the main study, we ran a pilot study. This resulted in 18 stimuli videos.

3.1 Source video acquisition

The BP4D-Spontaneous is a fully labeled database of realistic expressions [68]. Figure 2 shows some sample frames from this database. The authors designed an activity for naturally eliciting eight facial expressions (happiness, sadness, surprise, embarrassment, fear, pain, anger, and disgust) from people, across 41 participants (23 women, 18 men). The first task was to talk to the experimenter and listen to a joke to evoke happiness or amusement. Participants watched a documentary video about a real emergency situation which triggered sadness. Participants experienced surprise by suddenly being exposed to a loud siren. Embarrassment was triggered by playing a game in which the subjects improvised a silly song. Participants played a game with danger of physical injury which stimulated fear. Pain was experienced when the participants submerged their hands in ice water. Subjects were insulted by the experimenter to stimulate anger. Last, subjects smelled an unpleasant odor to evoke disgust.

The researchers captured both 2D and 3D videos for each task. At the frame level, each video has metadata including manually annotated facial action units (FACS AUs), automatically tracked head poses, and 2D/3D facial features. In 3D feature tracking, 83 feature points were tracked using a method which involved fitting a multi-frame constrained 3D temporal deformable shape model (3D-TDSM). For 2D feature tracking, researchers used a CLM-based tracker to automatically track 49 feature points frame by frame in each of the videos.

All the 2D tracked videos were reviewed offline and labelled as “Good tracking,” “Multiple errors,” “Jawline off,” “Occlusion,” and “Face out of frame.” The researchers chose a tracking score for each frame and the frame was reported as lost-track if the score was lower than a threshold.

At the sequence level, each video has participants’ self report ratings and observer ratings. After each task, participants labeled the intensity of their emotions using a 6-point Likert-type scale (0-5). The participants experienced more than one emotion for each task. However, the results showed that for all the expressions except anger, the intended emotion was the one most frequently labeled by the participants.

In the subjective experiment, five naive observers were asked to watch all the videos and label each video with two expressions for all the tasks. Each observer was also asked to label their confidence level of each choice using a 3-point scale indicating high, moderate, and low confidence. Their results showed that the intended emotion for each task was most frequently labeled by observers.

Since participants’ self reports and observer ratings of each expression’s intensity were not released with the database, we analyzed the labeled action units to decide which source videos we should include in our experiment. Five action units were released with the database: AU 6,10,12,14, and 17. For pain videos, we computed the average intensity of AU6 and AU10 across all the frames of the video and chose the five videos that had the highest average intensity across all their frames. We chose AU6 and AU10.
because they are important in expressing pain [46]. Since AU10 is important in expressing disgust, we computed the average intensity of AU10 across all the frames of each disgust source video and chose the five videos that had highest average intensity across all their frames.

For expressing anger, AU4 would have been ideal to use, but it was not available in the database. Therefore, we asked two naive observers to watch anger videos one by one and rated their perceived intensities using a Discrete Visual Analogue Scale (DVAS), which ranged from “none” to “very intense.” We chose the five anger videos that had the highest average intensities across the two observers.

### 3.2 Feature extraction

For tracking facial features, we used a CLM-based face tracker to track 66 facial feature points frame by frame from each source video. Constrained Local Models are person-independent, shape-based tracking techniques that do not require any manual labeling of a performer’s face [5, 16, 18]. There are also several extensions of CLM-based tracking approaches [18]; we used one developed by Baltrusaitis et al. [5], which is robust to light variations and head-pose rotations.

CLM-based models have been mostly used in the literature for face tracking, or expression detection, not for synthesis [5, 16]. In our work, we used this technique to synthesize facial expressions of pain, anger, and disgust on a humanoid robot and its virtual avatar model.

### 3.3 Humanoid robot

We used Philip K. Dick (PKD), a humanoid robot modeled after the famous science fiction writer Philip K. Dick, who authored many well-known novels such as The man in the high castle and A scanner darkly. The robot was built by Hanson Robotics [26], and is an expressive robot with 24 DOFs on its face, head, neck. It has a Frubber skin, which is a flesh-like rubber that gives PKD a very realistic appearance [59, 19]. This spongy material used to make PKD’s skin can be stretched to mimic human-like facial expressions. PKD has 24 servomotors, which enables it to express a wide range of facial expressions, as well as head and eye movements [24]. These servo motors are connected to the skin through tenders which push and pull the skin to create facial expressions. Figure 3 (right) shows PKD’s face.

### 3.4 Avatar model creation

To ensure a fair comparison of pain detection accuracy between the robot and the avatar, we created a similar virtual model of the robot PKD. This was to mitigate appearance difference effects. To generate the avatar, we first extracted a base avatar model, “Breen”, from Source SDK [62], which enables FACS-based virtual character animation. We then used GCFScape to extract textures from the model, and used VTFEdit to convert the textures to modifiable TARGA files (a raster graphic file format). We then had an artist use Adobe Photoshop to modify Breen’s texture to look like PKD. This included enlarging the jaw-line, cheekbones, and chin. Figure 3 shows the final avatar model side-by-side with PKD.

### 3.5 Stimuli creation and labeling

We initially created 30 stimuli videos. We had two types of embodiments (robot and avatar), three expression types (pain, anger, and disgust), and five videos of each expression. In our work, we used the 66 CLM facial feature points that were tracked frame by frame from each of the 15 videos from the BP4D-Spontaneous database. We used these feature points as the raw features in our work. We first removed rotation, translation, and scaling based on the eye corner positions. Then, we measured the distance of each of the 66 CLM feature points to the tip of the nose as a reference point and saved the 66 distances in a vector.

The tip of the nose stays static regardless of the expression one is expressing. In-plane rotations and translations do not affect distances of facial points to the tip of the nose. Therefore, any changes in the distance of a facial point to the tip of the nose means a change in facial expression is occurring. In each frame, we kept track of the changes each facial point had to the tip of the nose. We then mapped these changes to servo motors of the PKD robot and control points of its virtual model. We performed this in three steps.

First, for each of the PKD’s servo motors and each of the avatar’s control points we found a group of one or multiple CLM facial points whose movement would affect that motor or control point. For example, as PKD has two motors to control each of its eyebrows, one to control its inner eyebrow and another to control its outer eyebrow. However, as figure 4 shows, our face tracker tracks five points on each of the eyebrows (e.g. 46, 47, 48, 49, 50 for the left eyebrow). Therefore, we considered a group of points (48, 49, 50) for the left inner eyebrow and a group of points (46, 47, 48) for the left outer eyebrow. If a performer moves her left inner eyebrow (e.g. in anger expression), the distance of these three facial points (48, 49, 50) to the tip of the nose (32) decreases.

Second, for each servo motor or control point, we averaged the movements of all the points in that group to calculate only one number as the command for that motor or control point.
Servo motors of PKD and control points of the avatar have a different actuation range of values compared with feature points. Therefore, in the third step we created a conversion between these values. We found the minimum, maximum, and default values of each group of CLM facial points, each servo motor, and each control point. We then used linear interpolation to map movements of each group of CLM facial points to each servo motor of PKD or control point of the avatar.

We recorded PKD and its virtual model expressing each of the 15 source videos. After generating the 30 raw stimuli videos, we processed them before running our study. Stimuli videos had different lengths because their source videos from BP4D-Spontaneous database had different lengths. We cropped the videos to be three seconds long to ensure consistency. We added two seconds of a black screen with a white crosshair to the beginning of each video to prepare participants for the stimulus video. Then, a facial expression on PKD or the virtual character was presented for three seconds, followed by a black screen. Figure 5 shows example frames of the created stimuli videos.

### 3.6 Pilot study

We ran an online pilot study to determine which stimuli videos should be included in our main study and to establish their ground truth labels, following established methods in the literature [52, 41]. 20 participants were recruited via Amazon Mturk (12 female and 8 male). Participants were all native English speakers, and aged between 23 to 46 years old (mean age 35.6 years old).

Participants were asked to watch each video once, and label the expression they perceived and its intensity. The expression labels were semi-forced choice: pain, anger, disgust, and none of the above. We employed this paradigm to align with the methodological labeling approach proposed by Tottenham et al. [61]. They suggested that a semi-forced choice method is less strict than a forced choice method [57], and that it is easier to interpret findings in the semi-forced choice paradigm compared with a free-choice paradigm.

For labeling the intensity of perceived expression in each video, we used a 6-point DVAS. We decided to have an even number of choices (0 to 5) for intensity, because otherwise participants tend to always pick the middle if there are an odd number of choices. Zero was associated with none and five was associated with very intense.

We calculated the accuracy of each of our videos across our participants, and chose the three best videos of each expression that had the highest average accuracy across the two embodiments (robot and avatar).

We had average accuracies of 82%, 65%, and 67% for pain, 62%, 35%, and 32% for anger, and 27%, 25%, and 17% for disgust. We were not surprised by the low detection accuracies for disgust, since it is known to be a poorly distinguishable facial expression in the literature [8, 7, 10]. One reason for low detection accuracies for anger was that the source videos came from a naturalistic dataset (BP4D) and therefore had very low intensities.

Zhang et al. designed an activity for simulating different natural expressions including anger and disgust in their participants [68]. Therefore, the naturally stimulated anger videos had very low intensity compared to an acted dataset such as the one we used in our previous work [41]. Naturalistic anger source videos were hard to recognize due to their very low intensity. However, since our research questions were focused on pain, we decided to continue with the study and include anger and disgust in addition to pain to make sure we do not ask participants to only watch and label only one type of expression (pain).

### 3.7 Main experiment

Following the pilot study, we selected three videos of each expression to use in our main experiment, resulting in a total of 18 stimuli videos. The videos were prepared and presented in the same format as our pilot and randomized accordingly. Participants were asked to select a semi-forced choice expression label (pain, anger, disgust, none of the above) and expression intensity (on a 6-point DVAS scale).

Our main experiment was a video-based study. In HRI, these studies are commonly employed in order to include a larger number of participants [65], and to reach a more diverse audience. In video-based HRI studies, participants typically view videos of human-robot interactions and answer questions about them. In our work, instead of watching a human and robot interact, participants watched a robot or its avatar model expressing pain, anger, and disgust, and were asked to label the perceived expression and its intensity.

While some researchers argue that video-based studies are not suitable for answering some research questions considering that participants do not actually interact with the robot, other HRI research findings suggest that video-based studies have comparable results to live studies [63, 32, 67, 31, 44]. For example, Kidd did not find any significant differences between subjects’ ratings of personality traits for live and video-based studies of an interaction with a robotic head [31]. Furthermore, Kiesler et al. studied the subjective perception of live and recorded agents (robots and avatars) and found the results were similar [32].

As more interaction is involved between robot and participant in a study, the less suitable a video-based study would be because of...
the increased importance of aspects of embodiment [63]. Because our work is focused on how people perceive robot expressions, with a particular focus on validating our synthesis algorithms and also on their effect on clinical trainees, we hoped to still gain valuable findings about visual pain perception on an RPS at this stage of our research program. Furthermore, we aimed to recruit a large number of interprofessional clinicians from across the United States which would have been very challenging if we wanted to bring all of them to our lab. In the future, we do plan to replicate with a live study.

We recruited 102 participants for our experiment, 51 clinicians and 51 non-clinicians. We recruited clinicians by sending an advertisement email to over a dozen different medical and nursing schools across the United States. The clinicians were from multi-disciplinary backgrounds, 37% were registered nurses (RNs), 41% were physicians, and the rest included pharmacists, medical technicians, clinical researchers, and therapists (approximately 20%).

Lay participants were recruited via Amazon Mturk. Participants were eligible only if they did not participate in our previous studies, were over 18, and were native English speakers. Additionally, each participant completed a demographic questionnaire at the beginning of the study. In this questionnaire, participants were asked if they ever received a clinical education, or if they ever held a clinical related job. If they answered affirmatively, they were unable to continue the study. Overall, participant ages ranged from 19-64 (mean age = 33.6 years) and they were of mixed heritage, and were native English speakers. In terms of gender, the clinicians were 39% male, and the non-clinicians were 54% male.

Each participant watched the 18 videos (6 pain, 6 anger, and 6 disgust) in a random order. Nine videos were from the robot and nine were from the avatar. Participants were asked to watch each video once and label the expression that they perceived and its intensity. Expressions were semi-forced choice: pain, anger, disgust, and none of the above. For labeling the intensity of perceived expression in each video, we used a 6-point DVAS (0 (none) to 5 (very intense)).

The results from the main experiment are described in the subsequent sections. We measured accuracy (correct or incorrect) across our three independent variables (embodiment type, educational background, and expression type). We describe the statistical details of our analysis below.

## 4. RESULTS

### 4.1 Regression method

To answer the first and second research questions, we used binary logistic regression. We had accuracy as the one dependent variable derived from the expression classification task (i.e. classification of the expressions as pain, anger, or disgust). Accuracy is based on the ground truth that we gained from our pilot studies. We had three independent categorical variables: expression with three levels (pain, anger, and disgust), embodiment type with two levels (robot and avatar), and clinical education background with two levels (clinician and non-clinician).

The dependent variable was analyzed using a within-subjects binary logistic regression since the only dependent variable was binary (1: accurate, 0: inaccurate). In the following analyses, significant effects are those with $p$-values <.05. Table 1 shows the details regarding the exact number of errors that participants made in the classification of 18 videos. In this classification, we considered an answer correct if the participant’s label matched with the source video label. The percentage of correct classifications was computed across each of the three expressions types within each embodiment and each education background:

\[
3 \times (\text{Expression: pain, anger, or disgust}) \times 2 \times (\text{Embodiment: robot or avatar}) \times 2 = 12
\]

The three independent variables (expression, embodiment type, and clinical education background) were significant predictors for the dependent variable (see Table 3). We compared the full model with three predictors (expression, embodiment, and background) with the restricted model with just a constant factor. The results of the analysis of the full model with three predictors (independent variables) suggest a significant effect of the set of predictors on the accuracy affected by the expression. While the chi-square value in Table 2 shows that the predictors together have a significant effect on the model, the Wald test is the significant test for each individual predictor. Table 3 shows the effect of each individual independent variable on the classification rate. The standardized Beta value represents the weight that each predictor has in the final model. Negative weight shows a negative relation. Since the regression was run with pain
as the reference value, it does not have a Beta value. The Wald test suggests that all the three expressions are significant predictors for accuracy, \(p < .001\). Medical background is also a significant predictor for accuracy, \(W = 8.92, p < .05\). Embodiment type is a significant predictor for accuracy as well, \(W = 8.33, p < .05\).

### 4.2 Multi-Factor ANOVA

Our third research question explored the effect of a clinical education background on the perceived intensity of pain. To answer this question, we ran a multi-factor ANOVA. We had one dependent variable in this test, intensity, which was a scalar variable between 0 and 5. We had two categorical independent variables, clinical education background with two levels (clinician and non-clinician), and embodiment type with two levels (robot and avatar).

Table 4 shows that a clinical background is not a significant predictor for the perceived intensity. Therefore, we found that there is no significant difference between clinicians and non-clinicians in overall pain intensity perception.

However, Table 4 shows that embodiment type is a significant predictor for perceived pain intensity. Therefore, we found that participants tend to perceive a lower intensity of pain expressed by a robot compared to its similar avatar.

### 5. DISCUSSION

We synthesized naturalistic painful facial expressions on a humanoid robot and its virtual model, and answered different research questions regarding their perception. This work explores new applications of social robots in patient simulation, and gives some insights on how simulator type (avatar or robot) or educational background (clinician or non-clinician) may affect perception of pain.

We found support for our first research question, which supports earlier findings in the literature, that clinicians are less accurate than laypersons in decoding pain [21, 23, 43, 35, 27, 17]. Our work showed that this effect is found regardless of simulation embodiment (robot or virtual character), which suggests even the novelty of the embodiment is not sufficient to incur improved decoding skills. However, since training has been shown to help mitigate these clinical habituation effects, expressive, interactive RPSs may be useful in the medical education curriculum to be used alongside procedural skills training. For example, rather than having infrequent and contextless patient empathy training sessions [58], clinical learners could be continually learning pain decoding skills as they practice other critical care skills on patient simulators (e.g., performing physical exams, assessing patient state, etc.).

Some of these applications have recently been explored in the HRI community. For example, using robots for a range of health communication interventions to empower people with facial masking or facial movement disorders (e.g., Parkinson’s disease, cerebral palsy, dystonia) when interacting with their clinicians [13, 54, 60, 51]. We are in fact currently exploring masked synthesis on RPSs in our research, and believe these kinds of interventions could also be useful in a clinical education context to help with destigmatization and to improve provider empathy.

We did not find support for our prediction that participants with a clinical education background tend to miss lower levels of pain intensity compared to laypersons. This is in contrast to findings by Prkachin et al. [47, 17, 48] who showed that clinicians tend to vastly underestimate patient pain intensity compared to laypersons. One reason for this finding could involve the type of pain studied. Prior work has focused on chronic pain perception by clinicians; however, we employed acute pain models in our work. Further work is needed to explore this finding; preferably longitudinally in situ in clinic.

It also may be interesting to explore if clinicians underestimate pain intensity on expressive RPSs when patient history is provided to the clinical learner. It is well documented that clinicians underestimate expressed pain in patients who exhibit risky lifestyle behaviors (e.g., drug use), that come from low socio-economic strata, culturally-underrepresented minorities, older adults, and women. Lok et al. [64] recently showed that clinicians can achieve self-awareness of their pain perception biases within this context, which suggests that there may be a place for RPS training programs to help mitigate these effects.

We also found that all participants (clinicians or laypersons) are more accurate in detecting pain from a virtual avatar compared to a similar robot, and able to perceive higher pain intensities on the avatar. These results lend support to some recent findings in the HRI community. For example, Bennett and Sabanovic [9] found that detection accuracy is higher for facial expressions on an avatar compared to a physical robot. Also, Li et al. [37] found that when a human lecturer giving a presentation was replaced with an expressive virtual avatar, knowledge recall was higher than when the lecturer was replaced by an expressive social robot. Bennett and Sabanovic argued that the higher detection accuracy finding for virtual avatars might be because it is easier to maintain FACS fidelity on an avatar compared to a physical robot [9]. This finding relates to a limitation of our work, which was that PKD did not have any range of motion nor wrinkles around the nose area or cheeks. Therefore we could not map AU9 or AU10, which are both important for expressing pain and disgust [46]. (We also disabled these degrees of freedom in the virtual avatar to ensure a fair comparison). Adding these action units to both the robot and avatar could help improve overall detection accuracy and intensity perception.

In the development of our own robot, we have used these findings to improve its design. For example, it is now clear to us that AU9, AU10, and AU4 are critical action units to have on our robot. Upon its completion, we plan to experimentally explore the implications of these AUs, and also replicate this experiment to see if our results differ. Another reason for low detection accuracies of some facial expressions may be because our videos came from a naturalistic dataset (BP4D). It is well understood in the affective computing community that naturally evoked emotions overall have far lower intensities compared to acted datasets (such as the MMI corpus [45]). It is worth noting that the recognition rate itself is not the key finding with regards to the path to applicability - it is the significant difference in accuracy between clinicians (51%) and non-clinicians (67%). This result provides direct support to other findings in the clinical literature that as clinicians progress through their training

<table>
<thead>
<tr>
<th>Model</th>
<th>Chi-square</th>
<th>df</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Step</td>
<td>198.035</td>
<td>4</td>
<td>0.000</td>
</tr>
<tr>
<td>Step 1</td>
<td>198.035</td>
<td>4</td>
<td>0.000</td>
</tr>
<tr>
<td>Model</td>
<td>198.035</td>
<td>4</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Table 2: Omnibus Tests of Model Coefficients. \(\chi^2(4) = 198.035, p < .001\). Combined predictors have a significant effect on the model.
they lose their expression decoding abilities, which affects how well they can provide care. This research offers a new paradigm for training that may enable clinicians to overcome these decoding deficits through repeated interactions and exposure. Expressive RPSs are perfect for providing such training.

Despite their high clinical relevance, commercially available RPSs do not currently support human-like face-to-face interactions with learners. However, this is problematic as clinicians need to learn how to start reading patient facial expressions from their very first day of training, and continue to hone their facial expression decoding skills throughout their training. These face-to-face communication skills should not be kept separate from other beside training skills (e.g., taking vital signs, inserting IV lines), but rather should be well-ingrained into the simulation curriculum in a manner that maintains simulation fidelity.

Our work has implications for the HRI community because it explores a novel application of facially expressive robots in healthcare, which is using these robots as expressive patient mannequins. We also explored how synthesizing facial expressions on robotic patient simulators can make their interactions with clinical learners more realistic. Our work can help design more effective social robots for healthcare applications by studying how these robots are perceived by different groups, and how different factors, such as physical embodiment or educational background, can affect perception of these robots.

The clinical community can benefit from our work because we are developing new training tools for clinical learners to help them improve their skills in reading patients’ faces. Our results support findings by other researchers that compared to lay participants, clinicians have a lower accuracy in detecting pain in patients. However, to the best of our knowledge our work is the first to replicate these findings through simulation with an expressive robot instead of a real patient. Considering that both robotic patient mannequins and virtual avatars can be used as part of a training tool for improving learners’ skill in decoding visual signals of illness and pain in patients, our results can help clinical professionals understand how choosing different simulation systems (virtual or robotic) can affect perception of expressions.

6. ACKNOWLEDGEMENT

This material is based upon work supported by the National Science Foundation under Grant No. IIS-1253935 and Grant No. IIP-1534124. The authors also would like to thank Dr. Roya Ghassemi for providing input for the statistical methods used in Section 4 of this paper.

7. REFERENCES


