Naturalistic Pain Synthesis for Virtual Patients

Maryam Moosaei, Michael J. Gonzales, and Laurel D. Riek

Department of Computer Science and Engineering,
University of Notre Dame, Notre Dame, IN, 46556, USA
{mmoosaei,mgonza14,lriek}@nd.edu

Abstract. Within the clinical education community, there is a desire to improve learners' pain observation skills. Virtual patients can be used as a training tool for this purpose. In this paper, we present a pioneering approach for synthesizing naturalistic pain on virtual patients. Using the UNBC-McMaster pain archive and a CLM-based face tracker, we performed naturalistic pain synthesis. We conducted an experiment to validate our synthesis approach and compared it to manual methods that use FACS-trained animators. Our results suggest that our approach was effective, and yielded higher pain labeling accuracies compared to manually animated painful faces. This research offers a new tool to both the virtual patient and clinical education communities.

Keywords: Virtual patients, pain synthesis, facial expression synthesis, healthcare simulation, patient simulation.

1 Introduction

Many researchers in the fields of affective computing and clinical education are interested in patient simulation (c.f., [1–4]). Simulated patients provide safe experiences for clinical trainees, where they can practice communication, assessment, and intervention skills, without fear of harming a real patient. (See Fig. 1). Although this technology is in widespread use today, commercial patient simulators lack sufficient realism. They have static faces with no capability to convey facial expressions, despite the vital importance of these non-verbal expressivity cues in how clinicians assess and treat patients [5, 6].

This is a critical omission, because almost all areas of health care involve face-to-face interaction [7]. Furthermore, there is overwhelming evidence that providers who are skilled at decoding communication cues are better healthcare providers: they have improved patient outcomes, higher patient compliance and satisfaction, greater patient safety, and experience fewer malpractice lawsuits [6, 8, 9]. In fact, communication errors are the leading cause of avoidable patient harm: they are the root cause of 70% of sentinel events, 75% of which lead to a patient's death [10].

In studying how individuals, teams, and operators interact with inexpressive simulators, our work suggests that commercially available systems are inadequate for the task of training students due to their complete inability to provide human communication cues [11–14]. In particular, these simulators cannot

T. Bickmore et al. (Eds.): IVA 2014, LNAI 8637, pp. 295–309, 2014.

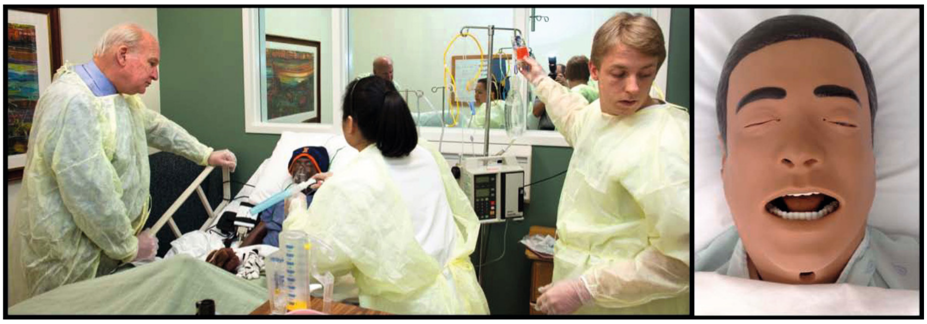

Fig. 1. Left: A team of clinicians treat a simulated patient, who is conscious during the simulation, but has no capability for facial expression. Right: A commonly used inexpressive mannequin head.

convey visual signals of pain to medical trainees even though perceiving a patient's nonverbal pain cues is an exceptionally important factor in how clinicians make decisions. Existing systems may be preventing students from picking up on patients' pain signals, possibly inculcating poor safety habits due to a lack of realism in the simulation [15, 16].

Our work focuses on making patient simulators more realistic by enabling them to convey realistic, patient-driven facial expressions to clinical trainees. We are designing a new type of physical patient simulator with a wider range of expressivity, including the ability to express pain and other pathologies in its face [4]. This paper presents one aspect of this project, which includes research questions surrounding synthesizing naturalistic[1] painful faces on a virtual avatar and evaluating how they are perceived.

This research fills a gap in the virtual patient problem domain, because although there is a growing body of literature on automatic pain recognition [18–20], there is little published work on automatic pain synthesis, particularly using naturalistic data. This work also will enable medical educators improve their face-to-face communication skills and pain recognition skills, which, according to the literature, are both in need of attention [5, 6, 21].

1.1 Our Work and Contribution

In this paper, we describe a technique for naturalistic pain synthesis on virtual patients, and report on several perceptual studies to validate the quality of the synthesis. For synthesis, we used a constrained local model (CLM)-based facial feature tracker applied to examples from the UNBC-McMaster Pain Archive [22].

We modeled our perceptual experiment on work by Riva et al. [23], where participants classified videos of three types of synthesized facial expressions: pain, anger, and disgust, across three genders - male, female, and androgynous. Riva et al. considered anger and disgust as reasonable expressions for comparison

[1] Here, *naturalistic* refers to non-acted, real-world data obtained "in the wild". c.f. [17].

because of their “negative valence and threat-relevant nature”. They explored avatar gender variations, because previous work in the field suggested a relationship between actor gender and pain detection accuracy [24]. In their work, the facial expressions were manually created by an animator using FaceGen 3.1, and reviewed by experts trained in the Facial Action Coding System (FACS).

Riva et al. [23] had two findings of note. First, they found participants were less accurate in decoding expressions of pain compared to anger and disgust (similar to other work, c.f. [25–27]). Second, regardless of gender, participants had better pain detection accuracy for male avatars. Given a naturalistic approach to synthesis, we wondered if the findings by Riva et al. [23] would hold, and, thus, replicated their experiment.

We have two main research questions. First, are participants able to distinguish expressions of pain from anger and disgust, and how do their accuracies differ? Based on findings by Riva et al. [23], we predict that overall, participants will be more accurate at detecting disgust compared with pain, and more accurate at detecting anger compared with disgust.

Second, how does an avatar’s gender affect pain detection accuracy? In addition to being curious if we can replicate findings by Riva et al. [23], we also would like evidence-based insights into how to design our physical robotic patient. We eventually will need to make decisions about the apparent gender of the robot, and this will require careful weighing of our findings. Based on findings by Riva et al. [23], we predicted that pain detection accuracy will be lower overall when expressed on a female avatar compared to a male avatar[2].

Our methodology, described in Section 2, addresses these research questions through a 3x3 online study in which subjects labeled videos of male, female, and androgynous avatars displaying pain, anger, and disgust.

Our results, discussed in Section 3, showed that participants were able to distinguish facial expressions of pain from anger and disgust by performing naturalistic synthesis, and were less accurate in decoding disgust compared to pain and anger. Furthermore, we did not find support for the avatar gender finding by Riva et al.; in our data avatar gender did not have significant effect on pain detection accuracy. Finally, our results suggest that naturalistic pain synthesis on virtual avatars is comparable to manual pain synthesis, and arithmetically, may be better. We discuss the implications of our findings for the community in Section 4.

2 Methodology

2.1 Background

Similar to other expressions of emotion, facial expressions of pain are an important non-verbal communication signal, particularly in healthcare [30, 31]. Until

[2] In this work we did not explore the effect of participant gender. The reason is that despite findings about how it affects accuracy in detecting some aspects of expressivity (e.g., arousal and valence) [28, 29], there is no evidence to suggest it affects overall categorization.

recently, self-reporting and clinical observations were the primary ways used to detect pain. However, these methods have several issues. For example, self-report cannot be used for children or patients with communication challenges (e.g. cognitive impairments, unstable states of consciousness or lucidity, etc.). Moreover, there are differences between how clinician and how patients conceptualize pain, which can lead to problems [18, 19, 31, 32].

Psychologists propose that pain is expressed by certain facial movements. While there is some research suggesting pain can be idiosyncratic [33, 34], we approach our research with respect to the Facial Action Coding System (FACS) which states pain can be interpreted universally from face. FACS uses 46 action units (AUs) as its building blocks to code facial expressions. This system was developed initially by Ekman [35] to code basic emotions based on facial muscle activities, and later was applied to pain, notably by Prkachin [19] and Craig [36].

To date, several research groups have worked on techniques for automatic pain detection, a process that involves automatic facial feature extraction and training of classifiers to detect pain. For example, Ashraf et al. [18] classified videos from the UNBC-McMaster Shoulder Pain Expression Archive Database [22] into pain/no pain categories using machine learning approaches. Their feature extraction was based on Active Appearance Models (AAM), which we describe in more detail in Section 2.4. The researchers decoupled shape and appearance parameters from facial images and used Support Vector Machines (SVM) for classification. Others have also explored automatic pain recognition, c.f. Prkachin et al., Monwar and Rezaei, and Hammal et al. [20, 37, 38]

Despite the aforementioned work on automatic pain detection, there is little work on automatic pain synthesis. In our work we studied features that were used in the literature for pain facial expression detection and instead employed them to synthesize facial expression of pain.

2.2 Overview of Our Work

We employed performance-driven synthesis of pain, anger, and disgust on three virtual avatar faces (female, male, and androgynous) to answer the aforementioned research questions. Performance-driven synthesis is a commonly used animation technique that tracks motions from either a live or recorded actor and maps them to an embodied agent, such as a virtual avatar or physical robot [39, 40]. This technique has been used in the literature to synthesize a wide range of naturalistic facial expressions [41, 42], but not pain.

In our work, the source videos we used for pain synthesis came from the UNBC-McMaster Pain Archive [22] and the source videos for anger and disgust from the MMI database [43]. We used ten source videos for each expression type. Our stimuli creation process included four steps. First, we used a Constrained Local Model (CLM) based tracker to extract 68 feature points frame-by-frame from each source video. Next, we mapped the extracted feature points to the virtual character control points for animation in Steam Source SDK. Then, we animated three different virtual characters (female, male, androgynous) per each expression type, resulting in 90 stimuli videos. Finally, we ran several pilot studies to label both gender and expression, and to establish which videos to include in our main study. This resulted in 27 stimuli videos.

2.3 Source Video Acquisition

For the source videos depicting painful expressions, we used the UNBC-McMaster Shoulder Pain Expression Archive Database [22]. This is a fully labeled, naturalistic data set of 200 video sequences from 25 participants suffering from shoulder pain (52% female). Participants performed range-of-motion tests on both their affected and unaffected limbs under the instruction of a physiotherapist. At the frame level, each frame was coded using the facial action coding scheme (FACS), and contained 66-point AAM landmarks. Each frame also received a pain score ranging from 0 to 12. At the sequence level, each video has both self-report and observer ratings of pain, the latter ranging from zero to five.

In our study we only included videos in which pain was present. Similar to Ashraf et al., [22] we considered pain to be present in a sequence if its observer rating was three or greater, and pain to be absent if its observer rating was zero.

For the source videos depicting anger and disgust, we used the MMI database [43]. This is a database of posed expressions from 19 participants (44% female) who were instructed by a FACS expert to express six basic emotions (surprise, fear, happiness, sadness, anger and disgust). Each video begins with a neutral expression and then transitions into the target expression.

We included source videos from these two databases that were determined by two human judges to be accurately tracked by our face tracker. Judges watched the source videos with the CLM mesh drawn on the face and rated all videos on a scale from one to four, depending on how well the mesh aligned with the face throughout the video. We only included videos in which both judges gave the video a tracking score of one. This resulted in 10 source videos from each expression category (pain, anger, disgust). Figure 3 (top) shows some sample frames from these databases.

2.4 Feature Extraction

For tracking facial features, we used Constrained Local Models (CLMs), which are a shape-based tracking technique similar to Active Appearance models (AAM). AAMs are statistical methods for matching the model of a user's face to an unseen face. A CLM-based approach is similar to an AAM-based approach, except it is person-independent, and does not require any manual labeling of an actor's face [44–46].

To our knowledge, CLM-based models have been mostly used in the literature for face tracking, or expression detection, not for synthesis [44, 45]. In our work we use this technique to synthesize facial expressions of pain, anger, and disgust.

In a CLM-based method, the shape of the face is estimated by labeling some feature points on several facial images in the training set [47]. There are also several extensions of CLM-based tracking approaches [46]. For example, Baltrusitus et al. [44] introduced a 3D Constrained Local Model (CLM-Z) method

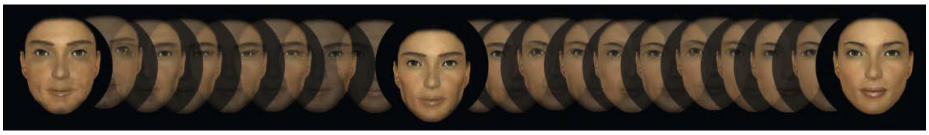

Fig. 2. The spectrum of avatar genders we created, with the final three highlighted. From right to left: male, androgynous, female.

for detecting facial features and tracking them using a depth camera, such as a Kinect. This method is robust to light variations and head-pose rotations, and is thus an improvement over the traditional CLM method.

For our work, we were able to create our stimuli based on a wide range of source videos using the CLM-Z tracker [44]. However, because our source videos were pre-recorded and did not contain depth information, we were not able to take full advantage of the CLM-Z tracker. In the future when we transition to performing real-time facial synthesis on an android robot, we will employ depth information to increase synthesis validity.

2.5 Avatar Model Creation

We used three avatar models in this work: female, male, and androgynous. When creating the avatar models, we aimed to remove any effects of age and ethnicity. In order to generate our avatars, we extracted avatars from the video game Half-life 2 from the Steam Source SDK. We used the program GCFscape to extract our textures from the video game files, and used a program called VTFEdit to convert the textures to modifiable TARGA files (a raster graphics file format).

For the purposes of this experiment, we used the character ("Alyx") [48] as the base for our virtual avatars to ensure consistency of ethinicity and age. Within Adobe Photoshop, certain areas were darkened or lightened to exhibit qualities normally attributed to male or female characteristics. For example, we enlarged the jaw-line, cheekbones, and chin during the creation of male-looking avatars. Similarly, androgynous textures also employed similar changes, but to a smaller degree. We also changed other areas of the face to create variation among these textures.

We created a total of twenty different texture variations employing these changes for our pilot to determine our androgynous, female, and male avatars. Similar to the stimuli created by Riva et al. [23], we cropped the face to remove any neck, clothing, or hair visibility to avoid any unintentional conveyance of gender cues.

We ran a pilot study to establish ground truth gender labels for each avatars' gender, following the methodology of Riva et al. [23]. The goal of this pilot was to select three avatar models as distinctly female, male, and androgynous out of the 20 models we made. We had 16 American participants, 11 female, mean age 44.5 years old. Participants were recruited using Amazon MTurk.

Participants viewed 20 still images of the avatars in a random order and labeled their gender using an 11-point Discrete Visual Analogue Scale (DVAS). A zero on the scale corresponded to "completely masculine" and ten to "completely feminine". We used these results to select our female, male, and androgynous avatar models. We chose the female model with score of 9.13, the male model with a score of 1.81, and the androgynous model with a score of 4.88. See Figure 2 for the final three avatar models.

The average score for each of the three chosen avatars ensured us that we could use these three avatar models as female, male, and androgynous in our main experiment. The next step was to animate each of these three avatars using the 30 source videos.

2.6 Stimuli Creation and Labeling

We initially created 90 stimuli videos. We had three avatar models (female, male, androgynous), three expression types (pain, anger, and disgust), and ten samples of each expression. We tracked 68 facial feature points frame-by-frame from each of our 30 source videos. We removed rotation, translation, and scaling based on eye corner positions. We measured movement of each point in relation to a normalized frame that was calculated during runtime.

We then measured the movement of each point. In order to map our facial points to our avatars, we generated source files that the Source SDK is capable of understanding. To do this, we ran the CLM tracker twice on each video. In the first run, we calculated the maximum movement of each point on the face. This was done to have a maximum scale factor for Source SDK to have as a reference point. In the second run we computed the movement of each feature point in relation to this maximum scale factor. To do this, we divided the movement measured in the second run by the maximum movement for the same point measured in calculation step. This gave us a ratio value between zero and one (zero being neutral, and one being the maximum amount a given point can move) for use with the featuring mapping in Source SDK.

We recorded each of the three avatar types enacting the expressions. The playback duration for each expression was manually adjusted to be 0.3 times slower to compensate for slight variations in how some of the source videos were tracked. After generating the recordings using CamStudio [49], we cropped the stimuli videos to be three to five seconds long to ensure consistency. Figure 3 shows example frames of the created stimuli videos.

We ran a second pilot study to decide which stimuli videos would be included in our main study, and to establish their ground expression truth labels. Each video was modified before being used in the pilot. A black screen with a white crosshair was added to the beginning of each video and appeared for exactly 2.5 seconds to prepare participants for the stimulus video. Then, a facial expression on a virtual character was presented for 2-4 seconds, followed by a black screen. We hosted the videos on Vimeo, and used an HTML 5 video player to remove all logos or player options. The pilot was conducted on SurveyMonkey.

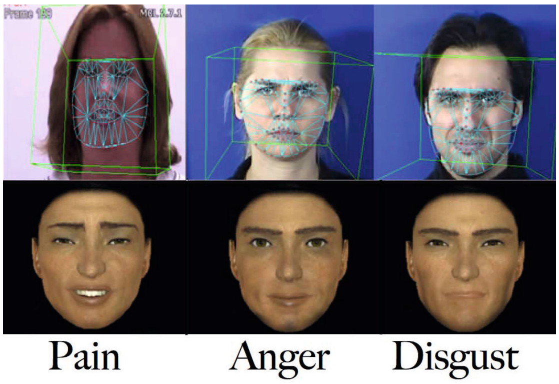

Fig. 3. Sample frames from the stimuli videos and their corresponding source videos, with CLM meshes. The pain source videos are from the UNBC-McMaster pain archive [22]; the others are from the MMI database [43].

The source and stimuli videos had slightly different lengths due to the fact that source videos of pain came from a different database than the videos of disgust and anger. We cut these videos to be nearly equal in length without removing informative frames from each video.

20 participants were recruited using Amazon MTurk, 11 female and 9 Male. All participants were American, and their ages ranged from 22-58 years old with mean age of 38.05 years old. Participants who participated in our previous pilot were excluded from this pilot. Participants were only allowed to view each video once and were allowed two 30 second breaks where a nature video was shown. Participants watched the 90 stimuli videos in a random order and labeled the avatar's gender and expression.

For gender labeling, we used the same scale as in Pilot 1 (an 11-point DVAS). We aimed to ensure that gender classification was the same as the first pilot when expressions were actually animated on the avatars. We found this was the case: Cronbach's $\alpha = 0.961$, indicating high inter-rater reliability on gender labeling.

Expression labels were fixed choice - anger, disgust, pain, and none of the above. This labeling approach was based on work by Tottenham et al. [50], who found a semi-forced choice method was less strict than a forced choice method (to which Russell [51] objects), while being more easy to interpret findings from than a free-choice method.

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 our three genders. We had average accuracies of 80%, 75%, and 63.33% for pain, 80%, 71.67%, and 63.33% for anger, and 33.33%, 31.67%, and 28.33% for disgust[3].

[3] 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 [26, 27].

Table 1. Frequencies and percentages of hits and errors in the main study

	Overall	**Female**	**Male**	**Androgynous**
Pain				
Total number of responses	450	150	150	150
Correct answers	303(67.33%)	100(66.67%)	98(65.33%)	105 (70%)
Judged as anger	2 (0.44%)	0 (0%)	1 (0.67%)	1(0.67%)
Judged as disgust	31(6.89%)	10 (6.67%)	7 (4.67%)	14 (9.33%)
Judged as none of the above	114 (25.33%)	40 (26.67%)	44 (29.33%)	30 (20%)
Anger				
Total number of responses	450	150	150	150
Correct answers	292(64.89%)	101(67.33%)	95(63.33%)	96(64%)
Judged as disgust	84(18.67%)	30(20%)	25(16.67%)	29(19.33%)
Judged as pain	44(9.78%)	10(6.67%)	20(13.33%)	14(9.33%)
Judged as none of the above	30(6.67%)	9(6%)	10(6.67%)	11(7.33%)
Disgust				
Total number of responses	450	150	150	150
Correct answers	133(29.56%)	47(31.33%)	43(28.67%)	43(28.67%)
Judged as anger	120(26.67%)	40 (26.67%)	42(28%)	38(25.33%)
Judged as pain	91(20.22%)	26(17.33%)	34(22.67%)	31(20.67%)
Judged as none of the above	106(23.56%)	37(24.67%)	31(20.67%)	38(25.33%)

Main experiment: Following the pilot, we selected three videos of each expression with the highest accuracy across our three avatar genders to use in our main experiment, resulting in 27 videos. Videos were prepared and presented in the same format as our previous pilot and randomized accordingly.

We recruited 50 participants using Amazon MTurk. Again, participants were eligible only if they did not participate in our previous studies. Participant ages ranged from 20-57 (mean age = 38.6 years). Participants were of mixed heritage, and had each lived in the United States at least 17 years.

Participants were asked to label the avatar's expression in each of the 27 videos. The results from the main experiment are described in the subsequent sections. We measured accuracy (correct or incorrect) across our two independent variables (gender and expression type). We describe the statistical details of our analysis below, but first present a brief summary.

3 Results

3.1 Summary of Key Findings

Our first research question was to explore if participants are able to distinguish pain from expressions of anger and disgust. Table 3 shows that expression is a significant predictor for accuracy. Therefore, we found participants were able to distinguish these three expressions. The results further showed that participants were more accurate in detecting pain than two other expressions. Thus, we did not find the same accuracy ordering as Riva et al. [23]; i.e., disgust > pain > anger.

Table 2. Omnibus Tests of Model Coefficients. $\chi^2(4) = 165.646$, $p <.001$.

		Chi-square	df	Sig.
	Step	165.646	4	0.000
Step 1	Block	165.646	4	0.000
	Model	165.646	4	0.000

Our second research question concerned the effect of the avatar's gender on pain detection accuracy. As seen in Table 3, gender is not a significant predictor for accuracy as the p-values for all the three genders are greater than .05. This suggests that there is no significant relation between an avatar's gender and pain detection accuracy. Thus, we were not able to replicate findings by Riva et al. [23] suggesting that people are more accurate at detecting pain when it is expressed on a male face.

3.2 Regression Method

We had one dependent variable and two independent variables. The dependent variable derived from the expression classification task was accuracy (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 two categorical independent variables. The independent variables were expression with three levels (pain, anger, and disgust) and gender with three levels (androgynous, male, and female).

The dependent variable was analyzed using an appropriate within-subjects binary logistic regression since the only dependent variable is 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 each participant made in the classification of 27 videos. In this classification, we considered an answer correct if 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 of the three genders. 3 *(Expression: pain, anger, or disgust)* × 3 *(Gender: androgynous, male, and female).*

Table 1 indicates the details of errors for each expression and each gender. Participants' answers were classified as either accurate or inaccurate. Overall, participants labeled 622 videos incorrectly, representing 46.07% of our responses.

The independent variables (gender and expression) were significant predictors for the dependent variable (see Table 2). We compared the full model with two predictors (gender and expression) with the restricted model with just a constant factor. The results of the analysis of the full model with two predictors (independent variables) suggest a significant effect of the set of predictors on the correct identification rate as the dependent variable.

The Wald test in Table 3 shows the degree to which each expression affected accuracy. While the chi-square value in Table 2 shows that predictors together

Table 3. Variables in the regression equation

	B	S.E	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
							Lower	Upper
Pain			151.609	2	0.000			
Disgust	-0.109	0.141	0.600	1	0.438	0.897	0.680	1.182
Anger	-1.593	0.144	122.017	1	0.000	0.203	0.153	0.270
Step 1 Androgynous	0.759	2	0.684					
Male	0.041	0.143	0.082	1	0.775	1.042	0.787	1.378
Female	-0.081	0.142	0.324	1	0.569	0.922	0.697	1.219
Constant	0.737	0.130	32.118	1	0.000	2.090		

have significant effect on the model, the Wald test is the significant test for each individual predictor separated. 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[4].

The results of the Wald test suggest that disgust and three genders can be dropped from the model for accuracy prediction. The Wald test suggests that pain by itself is a significant predictor for accuracy, W = 151.609, p <.001. Disgust by itself is not a significant predictor for the accuracy, W = .600, p >.05. Anger by itself is a significant predictor for accuracy, W = 122.017, p <.001. None of the three genders are significant predictors for accuracy. Pain has the largest effect on accuracy prediction followed by anger.

4 Discussion

Participants were able to distinguish pain from anger and disgust in virtual patients created using automatic naturalistic synthesis. Thus, we found support for our first research question (RQ1). Our results support Riva et al.'s findings that participants are able to detect pain from anger and disgust when being expressed by a virtual avatar face.

Our results do not support the claim by Riva et al. [23] that participants are less accurate in decoding the facial expression of pain compared to anger and disgust. Our results instead reflect the opposite - participants are more accurate in decoding facial expressions of pain compared with anger and disgust. Also, participants are more accurate at detecting anger compared to disgust.

We have also found support that naturally driven pain synthesis is comparable to FACS-animated pain synthesis. To the best of our knowledge, this is the first work on naturally performance-driven pain synthesis. Riva et al. [23] manually synthesized facial expressions on a virtual avatar, and found 60.4% as the overall pain labeling accuracy rate. Our pain labeling accuracy rate was

[4] SPSS considers pain as the base expression and androgynous as the base gender. Thus, these columns are empty for these two predictors.

67.33%. While we cannot statistically compare these results due to variability in the two experiments, arithmetically they are encouraging.

These findings suggest that our method may be used for automatic pain expression synthesis without requiring a FACS-trained animator to manually synthesize painful expressions. For practitioners and researchers in the clinical education community without such resources, this may prove beneficial.

Our results do not support the previous findings by Riva et al. [23] that pain expression recognition is a function of the gender of the avatar displaying it. We did not find any significant relation between the avatar's gender and the participants' accuracy in detecting pain, anger, or disgust. This further lends support to the idea that automatic pain expression synthesis from naturalistic sources may, in some cases, be preferable to FACS/animator-generated synthesis.

One limitation of our work was that the avatar model from the Source SDK had neither wrinkles nor control points around the nose area. Therefore, we could not map AU9, which is important in expressing disgust and pain [37]. Adding this action unit could help improve the detection accuracy for both disgust and pain. Another limitation was that the source videos for pain came from a naturalistic dataset, whereas the anger and disgust videos were acted and exaggerated. At the time this paper was published, we were not aware of any naturalistic databases for these expressions, but in the future this would be good to explore.

Similar to Riva et al. [23], we ran our experiments with lay participants. The literature suggests that the expressions of pain can be clearly recognized and discriminated by lay participants [36]. However, in the future, we are also interested to see how different populations perceive painful facial expressions, for example, if clinicians at different stages of training perceive an avatar's pain differently. Prkachin et al. [31] showed that clinical experience with patients results in underestimating patients' pain. It would be exciting to test this hypothesis more thoroughly with our avatars and explore if interventions can be designed.

Acknowledgement. This material is based upon work supported by the National Science Foundation under Grant No. IIS-1253935.

References

1. Kenny, P., Parsons, T.D., Gratch, J., Leuski, A., Rizzo, A.A.: Virtual patients for clinical therapist skills training. In: Pelachaud, C., Martin, J.-C., André, E., Chollet, G., Karpouzis, K., Pelé, D. (eds.) IVA 2007. LNCS (LNAI), vol. 4722, pp. 197–210. Springer, Heidelberg (2007)
2. Benjamin, L.: Shader lamps virtual patients: the physical manifestation of virtual patients. In: Medicine Meets Virtual Reality 19: NextMed, vol. 173 (2012)
3. Mitchell, S.E., et al.: Developing virtual patient advocate technology for shared decision making. In: 34th Annual Meeting of the Society for Medical Decision Making (2012)
4. Gonzales, M.J., Moosaei, M., Riek, L.D.: A novel method for synthesizing naturalistic pain on virtual patients. In: Simulation in Healthcare (2013)

5. Ryan, K.F.: Human simulation for medicine. In: Human Simulation for Nursing and Health Professions (2011)
6. Henry, S.G., Fuhrel-Forbis, A., Rogers, M.A., Eggly, S.: Association between nonverbal communication during clinical interactions and outcomes: A systematic review and meta-analysis. Patient Education and Counseling 86(3) (2012)
7. Martin, L.R., Friedman, H.S.: Nonverbal communication and health care. In: Applications of Nonverbal Communication (2005)
8. Back, A.L., et al.: Efficacy of communication skills training for giving bad news and discussing transitions to palliative care. Arch. Intern. Med. 167(5) (2007)
9. Brown, J.: How clinical communication has become a core part of medical education in the UK. Medical Education 42(3) (2008)
10. Leonard, M.: The human factor: the critical importance of effective teamwork and communication in providing safe care. Qual. Saf. Health Care 13 (2004)
11. Huus, A., Riek, L.D.: An Expressive Robotic Patient to Improve Clinical Communication. In: 7th ACM International Conference on Human-Robot Interaction (HRI), Pioneers Workshop (2012)
12. Martin, T.J., Rzepczynski, A.P., Riek, L.D.: Ask, inform, or act: communication with a robotic patient before haptic action. In: Proceedings of the International Conference on Human-Robot Interaction, HRI (2012)
13. Rzepcynski, A., Martin, T., Riek, L.: Communication and awareness: the building blocks of a successful clinical environment. In: Proceedings of the International Conference on Clinical Communication (2012)
14. Janiw, A., Woodrick, L., Riek, L.D.: Patient situational awareness support appears to fall with advancing levels of nursing student education. In: Simulation in Healthcare (2013)
15. Rzepcynski, A., Martin, T., Riek, L.: Informed consent and haptic actions in interdisciplinary simulation training. In: Proceedings of the American Public Health Association, APHA (2012)
16. Henneman, E.A., Roche, J.P., Fisher, D.L., Cunningham, H., Reilly, C.A., Nathanson, B.H., Henneman, P.L.: Error identification and recovery by student nurses using human patient simulation: Opportunity to improve patient safety. Appl. Nurs. Res. 23(1) (2010)
17. Douglas-Cowie, E., Cowie, R., Sneddon, I., Cox, C., Lowry, O., Mcrorie, M., Martin, J.-C., Devillers, L., Abrilian, S., Batliner, A., et al.: The humaine database: addressing the collection and annotation of naturalistic and induced emotional data. In: Affective Computing and Intelligent Interaction, pp. 488–500. Springer, Heidelberg (2007)
18. Ashraf, A.B., Lucey, S., Cohn, J.F., et al.: The painful face: pain expression recognition using active appearance models. ACM ICMI (2007)
19. Lucey, P., et al.: Automatically detecting pain using facial actions. In: 3rd Int'l Conference on Affective Computing and Intelligent Interaction, ACII (2009)
20. Hammal, Z., Cohn, J.F.: Automatic detection of pain intensity. In: ICMI (2012)
21. Coll, M.-P., Grégoire, M., Latimer, M., Eugène, F., Jackson, P.L.: Perception of pain in others: implication for caregivers. Pain Management 1(3), 257–265 (2011)
22. Lucey, P., Cohn, J.F., Prkachin, K.M., Solomon, P.E., Matthews, I.: Painful data: The unbc-mcmaster shoulder pain expression archive database. In: IEEE International Conference on Automatic Face & Gesture Recognition (2011)
23. Riva, P., Sacchi, S., Montali, L., Frigerio, A.: Gender effects in pain detection: Speed and accuracy in decoding female and male pain expressions. Eur. J. Pain (2011)

24. Hirsh, A.T., Alqudah, A.F., Stutts, L.A., Robinson, M.E.: Virtual human technology: Capturing sex, race, and age influences in individual pain decision policies. Pain 140(1) (2008)
25. Kappesser, J.,, A.C., de C. Williams, A.C.: Pain and negative emotions in the face: judgements by health care professionals. Pain 99(1) (2002)
26. Bazo, D., Vaidyanathan, R., Lentz, A., Melhuish, C.: Design and testing of a hybrid expressive face for a humanoid robot. IEEE (IROS) (2010)
27. Berns, K., Hirth, J.: Control of facial expressions of the humanoid robot head roman. In: IEEE/RSJ IROS (2006)
28. Bernardes, S.F., Lima, M.L.: On the contextual nature of sex-related biases in pain judgments: The effects of pain duration, patient's distress and judge's sex. Eur. J. Pain 15(9) (2011)
29. Simon, D., Craig, K.D., Miltner, W.H., Rainville, P.: Brain responses to dynamic facial expressions of pain. Pain 126(1) (2006)
30. Hadjistavropoulos, T., Craig, K.D., Fuchs-Lacelle, S.: Social influences and the communication of pain. Pain: Psychological Perspectives (2004)
31. Prkachin, K.M., Craig, K.D.: Expressing pain: The communication and interpretation of facial pain signals. J. Nonverbal Behav. 19(4) (1995)
32. de C. Williams, A.C., Davies, H.T.O., Chadury, Y.: Simple pain rating scales hide complex idiosyncratic meanings. Pain 85(3) (2000)
33. Aung, M., Romera-Paredes, B., Singh, A., Lim, S., Kanakam, N., de C. Williams, A., Bianchi-Berthouze, N.: Getting rid of pain-related behaviour to improve social and self perception: a technology-based perspective. In: 14th International Workshop on Image Analysis for Multimedia Interactive Services, WIAMIS (2013)
34. Romera-Paredes, B., et al.: Transfer learning to account for idiosyncrasy in face and body expressions. IEEE Face and Gesture (2013)
35. Ekman, P., Rosenberg, E.L.: What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System, Oxford (1997)
36. Simon, D., Craig, K.D., et al.: Recognition and discrimination of prototypical dynamic expressions of pain and emotions. Pain 135 (2008)
37. Prkachin, K.M., Berzins, S., Mercer, S.R.: Encoding and decoding of pain expressions: a judgement study. Pain 58(2) (1994)
38. Monwar, M.M., Rezaei, S.: Pain recognition using artificial neural network. In: IEEE Symposium on Signal Processing and Information Technology (2006)
39. Williams, L.: Performance-driven facial animation. ACM SIGGRAPH Computer Graphics 24(4) (1990)
40. Wan, X., Jin, X.: Data-driven facial expression synthesis via laplacian deformation. Multimedia Tools and Applications 58(1) (2012)
41. Beeler, T., et al.: High-quality passive facial performance capture using anchor frames. ACM T. Graphic 30 (2011)
42. Bickel, B., et al.: Physical face cloning. ACM T. Graphic. 31 (2012)
43. Pantic, M., Valstar, M., Rademaker, R., Maat, L.: Web-based database for facial expression analysis. In: IEEE Int'l Conf. on Multimedia and Expo, ICME (2005)
44. Baltrusaitis, T., Robinson, P., Morency, L.: 3d constrained local model for rigid and non-rigid facial tracking. In: CVPR (2012)
45. Chew, S.W., Lucey, P., Lucey, S., Saragih, J., Cohn, J.F., Sridharan, S.: Person-independent facial expression detection using constrained local models. In: IEEE Int'l Conf. on Automatic Face and Gesture Recognition, FG (2011)

46. Cristinacce, D., Cootes, T.: Feature detection and tracking with constrained local models. Proceedings of British Machine Vision Conference 3 (2006)
47. Abboud, B., Davoine, F., Dang, M.: Facial expression recognition and synthesis based on an appearance model. Signal Process-Image 19(8) (2004)
48. Valve Software: Source SDK, `http://source.valvesoftware.com/sourcesdk.php`
49. Camstudio: Open source streaming video software, `http://camstudio.org`
50. Tottenham, N., et al.: The nimstim set of facial expressions: judgments from untrained research participants. Psychiatry Research 168(3) (2009)
51. Russell, J.A.: Is there universal recognition of emotion from facial expressions? a review of the cross-cultural studies. Psychological Bulletin 115(1) (1994)