Discerning Affect from Touch and Gaze During Interaction with a Robot Pet

Xi Laura Cang, Paul Bucci, Jussi Rantala, and Karon E. MacLean *IEEE Senior Member*

Abstract—Practical affect recognition needs to be efficient and unobtrusive in interactive contexts. One approach to a robust realtime system is to sense and automatically integrate multiple nonverbal sources. We investigated how users' *touch*, and secondarily *gaze*, perform as affect-encoding modalities during physical interaction with a robot pet, in comparison to more-studied biometric channels.

To elicit authentically experienced emotions, participants recounted two intense memories of opposing polarity in *Stressed-Relaxed* or *Depressed-Excited* conditions. We collected data (N=30) from a touch sensor embedded under robot fur (force magnitude and location), a robot-adjacent gaze tracker (location), and biometric sensors (skin conductance, blood volume pulse, respiration rate).

Cross-validation of Random Forest classifiers achieved best-case accuracy for combined touch-with-gaze approaching that of biometric results: where training and test sets include adjacent temporal windows, subject-dependent prediction was 94% accurate. In contrast, subject-independent Leave-One-participant-Out predictions resulted in 30% accuracy (chance 25%). Performance was best where participant information was available in both training and test sets. Addressing computational robustness for dynamic, adaptive real-time interactions, we analyzed subsets of our multimodal feature set, varying sample rates and window sizes. We summarize design directions based on these parameters for this touch-based, affective, and hard, realtime robot interaction application.

Index Terms—Affective touch, multimodal interaction, human-robot interaction, therapeutic robot, emotion classification.

1 Introduction

Social interfaces such as robots, smart cars or game systems must facilitate complex and believable interactions where programmed machines appear to respond to human social cues [1]. Because people often prefer to interact with machines as they do with other people [1], systems may need to understand nonverbal emotional behaviours mediated through naturally affective modalities like touch or gaze. Affective, interactive therapies for anxiety management may use haptically available emotion indicators: touchable robots (baby harp seal Paro [2], teddy-bear-like Huggable [3]) map simple touch gestures to simple emotions. Studies with the Haptic Creature, a zoomorphic robot with an embedded touch sensor array [4], link a large and varied set of touch gestures to nuanced emotion expression.

Machine recognition of human emotion presents methodological challenges surrounding measurement instruments, study task framing, and computationally modeling emotions [5]. Training data behavior should reflect that of an interaction "in the wild", i.e., spontaneous emotion [6]. The emotion model should accurately describe that person's state. Furthermore, while people can be differentiated by idiosyncrasies in their touch behaviors (a *touch signature* [7], [8]), this also makes it difficult to generalize the connection between emotions and associated touch behaviors: the extent to which individuals exhibit similar touch behaviours during similarly labeled emotional states is unclear.

Here, we wish to enable machine recognition of human emotions for touch-centric social robots, with therapeutic applications in mind. Touch interactions can affect emotional state: the Haptic Creature's motion lowered *anxiety* in users who were stroking it on their laps [9], based on biometric indicators. This suggests physiological benefits analogous to those conferred by animal-assisted therapy [10], [11], [12], [13] – especially valuable where patients are unable to engage with real animals. However, this requires unobtrusive sensing, e.g., through already-occurring touch.

Gaze is another unobtrusive modality that could improve recognition performance. Since the points where a user's gaze focuses on a computer display can indicate feelings of curiosity or boredom [14], we posit that gaze as an indicator of visual attention could help determine when a user is focusing on the robot pet and thereby predict affect. Specifically, we compare the combination of touch and gaze to key biometric channels which have been well-researched in association with various emotions [15], [16].

To investigate these ideas, we set touch as the primary interaction modality in order to leverage the natural human inclination to express emotional closeness with physical contact. Gaze has also been shown to capture emotion data [14], and both (touch and gaze data) can be collected without the disruption of physiological sensors. Previous work has shown that affect-related information can be extracted from emotionally-directed touch gestures such as *Excited*-stroking and *Depressed*-rubbing [17]. However, identifying a gesture as 'stroke' *vs.* 'rub' is insufficient for revealing the user's emotional state while performing that gesture [17]. Furthermore, these studies collected "intent" data, where the emotions were *acted out to* a sensed robot, but not necessarily *experienced by* a participant. We needed a model built from data of participants who are truly experiencing the emotions being studied.

- *X.L. Cang, P. Bucci, and K. E. MacLean are with the University of British Columbia, Vancouver BC, Canada. E-mail: {cang, pbucci, maclean}@cs.ubc.ca*
- *J. Rantala is with the University of Tampere, Finland. E-mail: jussi.e.rantala@tuni.fi*

Manuscript version received Jun 2021.

1.1 Approach and Research Questions

The central purpose of this paper is to narrow the design space of an emotionally interactive robot pet's computational system for predicting an interacting user's emotion: touch-supportive sensing modalities that balance accuracy with ease-of-use; a training procedure that generates truly felt emotional sample data; and an appropriate classification model for touch behaviour in a computationally restricted environment.

To elicit naturally felt, spontaneous human emotion (hard to do in a lab setting [6]), we asked participants to interact with a robot while they relived a significant emotional event, touching it without constraint during the task. This approach departs from previous work [4], [17] that attempts to direct touch behaviours and gestures, i.e., by asking a participant to *pat* the robot *as if* they were *scared*. Relived emotion or emotion recall is regarded as a way to elicit true experiences of emotion [18], [19].

We are interested in touch and gaze as modalities that support low-cost, low-intrusion sensing apparati and explore their viability in comparison to biometric data. To that end, we compared affect measures derived from touch interaction with a robot pet with the more studied but intrusive reference point of biometric indicators, and investigated how recognition performance can be improved with gaze data. Furthermore, analysis methods that originate from social touch gesture classification are well documented [7], [17], [20]. We calculate features from force magnitude and touch location [7], [8], [20] as well as frequency [17] (referred to herein as pressure-location domain and frequency domain respectively) for emotion classification in *touch*. To minimize overlap in label interpretation, we collected and evaluated machine recognition of four emotions (*stressed*, *excited*, *depressed*, and *relaxed*) – quadrant extrema of Russell's dimensional affect model [21].

Choice of the Random Forest algorithm (RF) is motivated by our need for a classification system that performs well with social touch behaviour [7], [8], [17], [22], [23], [24] for our interactive robot pet application. We want to explore the feasibility of realtime emotion prediction from touch interaction with a emotionally interactive robot pet, where we anticipate being compute-restricted. Thus, we chose a computationally simple model favouring flexibility to accommodate quick training and customizable rebuilding.

We specified four main research questions for this study.

RQ1 Modality Effectiveness: How does touch or touch + gaze compare with biometrics in classifying affect? What minimal feature set optimizes performance accuracy?

Touch can be a natural avenue for communicating affect, but to use it computationally, we must access the encoded emotions and consider the relative performance of touch alone and with multimodal support. Gaze, also known to encode affective content [14], could supplement emotional signals from touch. Multimodal datasets are likely to provide a more complete picture than touch alone, due to asynchronous activation, or interaction information.

We expect *classification accuracy to improve with increased modality support*. We thus ask whether the combination of touch and gaze is a viable substitute for the more intrusive sensing apparatus required of tracking biometric signals.

However, multimodality increases compute time and phase delays, potentially undermining real-time feasibility. To optimize tradeoffs, we analyze each feature in terms of repeated occurrence in automatically-selected best-feature subsets. Finally, we suggest an optimal touch-with-gaze feature set, assessing both the *pressure-location domain* and *frequency domain*, hypothesizing that *classification accuracy is best where features are present from both domains.*

RQ2 Individuality: How important is system calibration and knowledge of user in affect classification?

Social touch gesture studies suggest that because individuals have distinctive ways of physical, expressive interaction with objects, recognizing *identity* is realistic [7], [8]. Thus a system that has learned a specific user's behaviour may be better at gesture recognition. Leveraging this result for affect, we assess how well the system can distinguish Participant – high performance suggests high individuality – then perform Emotion classification across three different levels of system knowledge of participant (hereby referred to as *participant knowledge*) and discuss results. We expect that *recognition rates will increase with greater participant knowledge*, i.e., participant-labelled data where instances from the same individual are in both training and test sets will yield the highest classification accuracy (subject labels used as a feature in subject-dependent classification); and lowest accuracy will coincide with testing and training on different individuals (subject-independent classification).

RQ3 Sample Density and Realtime Responsiveness: Is classification during continuous sampling robust to interruptions in signal, and to sample size variation?

Outside of polling rate, we define *sample density* across two window dimensions: (1) size and (2) adjacency. We investigate the accuracy trade-offs of various *window sizes* – which represent the time intervals of continuously sampled data. In the context of an interactive robot, longer windows gives the system time to respond, employs less computation resources and allows for the capture of "slow" behaviours. But where the window is too long, we introduce inappropriate response delays. For example, if our robot body is struck, it needs to present a behaviour demonstrating an immediate reaction. While shorter windows may help with the agility needed for interactive scenarios, the higher throughput requires more computational resources and may not recognize the slower developing interactions. *Window adjacency* refers to continuity of time series classification data. Since adjacent windows share more characteristics than distant samples (temporal dependence), we ask about the effect of non-continuous or 'gapped' data collection under weak or interrupted signal conditions. Removing adjacent instances allows us to quantify any effect from a dropped or intermittent signal as well as the likelihood of overfitting due to recency-based similarities, particularly when using easy-to-build classification models (like Random Forest) without parameter tuning. Here, we leave time-series analysis for future work and focus on the influence of sample density on accuracy. In order to construct early specifications for a touch-cognizant robot, we explore the trade-off between computational load and classification robustness.

We examine the influence of window size and continuity by aggregating data instances in four window sizes and

comparing classification accuracy of the same data set. We downsampled *with "gap"* by dropping 2s of data between windows so adjacent windows are not evaluated) and *without gap* data (adjacent windows are included in the training and test sets). We posit that across both parameters, *reducing sample density reduces classification accuracy,* anticipating the worst performance for small windows with gapped data.

RQ4 Experimental Paradigm: How well does our protocol corroborate existing relived emotion techniques to elicit genuine emotion in a controlled laboratory setting?

For affective communicative systems to work under real conditions, they must be trained on data from authentic and spontaneous emotion. Consistently producing *truly experienced* emotions in an artificial setting (and valid training data) is a fundamental challenge in emotion research [25].

We develop a means of implementing a touch variant of relived emotion techniques described in [19], [25], [26] and use self-report measures to explore how our experimental controls influence the *authenticity and intensity of the experienced emotion* generated within a controlled set-up.

1.2 Contributions

Through our research questions, we examine the design space of an affect classification system for an emotionally-interactive touch-centric robot. Specifically, we contribute:

1) *A comparison of affect classification performance* of touch data, with and without gaze support, to biometrics in *experienced-emotion* interactions; and a recommendation of data features from frequency and traditional pressure-location domains in emotion classification.
2) An assessment of subject-independent *vs.* dependent classification; and a *proposal for building a custom personalized system* at various levels of participant knowledge.
3) *An analysis of data factors* to balance classification robustness with computational effort and phase delay, for real-time applications.
4) Through demonstration and evaluation of an ecologically valid elicitation technique (emotional recall) for studies on machine touch recognition, we *assess the methods, models, and task framing required to increase confidence in generating true experienced emotion in a lab setting.*

In the following, we survey previous work, motivating our emotion elicitation method and contextualizing affect classification from each of touch, gaze, and biometrics; then describe our experiment and analysis. We report results that span all our data experiments to target the influence of: multimodal data *vs.* touch alone, participant knowledge, sample density, feature set; and assess emotional experience from participant reports. We discuss our findings and ground them in implications for relevant applications.

2 Related Work

2.1 Targeted Emotion Set

Russell's circumplex model plots affect on arousal (activation) and valence (pleasantness) axes [27]. While valuable in its conciseness, the dimensional model requires we assume (1) emotion labels will be interpreted consistently by every participant at any time; and (2) the axes are truly orthogonal.

Consider the emotional context of approaching the axes or origin when working with such a model: the state of (0,0), presumably a state of full neutrality, may not be meaningful. For example, independent movement, i.e., directly along axes, implies increasing an emotion arousal without changing valence, which belies personal experience. As such, many [28], [29], [30] opt to discretize the 2D space into a grid and rotate it by 45°, such that experimental materials and tasks are aligned with the diagonal axes, namely (high arousal, high valence) ↔ (low arousal, low valence) and (high arousal, low valence) ↔ (low arousal, high valence).

Relevant published studies are not consistent in emotion labels chosen to cover the affective space, making comparison between studies of even common modalities problematic. Understandably, papers utilizing information of gaze use attention-related emotion sets – e.g., *Anxiety, Boredom, Confusion, Curiosity, Excitement, Focus, Frustration* [31]; papers utilizing touch try to span the human experience, namely *Anger, Fear, Happiness, Sadness, Disgust, Surprise, Embarrassment, Envy, Pride* [32]. Yet another method is to partition Russell's affect grid as discrete labels: touch emotion recognition has previously used nine labels[1], while biometric recognition has used four labels corresponding to the quadrants of Russell's grid: *Stressed, Excited, Depressed, Relaxed* [15]. We have elected to use the same four named emotions for consistency with other biometric classification studies, enabling comparison with touch and gaze.

2.2 Elicitation of True Emotion

Our motivating applications center on a social robot that must react to authentic human emotions as they occur in lived experience. In the lab, one unsatisfying approach is to ask participants to imagine and simulate a reaction: (*"Imagine feeling anger, then express it to our robot"*). For example, to collect the data used in [4] and [17], participants were presented with a list of emotions that they acted out by touching a robot, but this does not equate to experiencing it. The difference between expressions of acted and experienced emotions can be significant and counter-intuitive: e.g., truly experienced frustration is often accompanied by a smile, but this is rarely the case for acted frustration [33].

Experienced-emotion studies are difficult to construct. Entertainment media, e.g., emotionally evocative music and/or video, has been employed in emotion elicitation [15]; however, it can be difficult to validate stimulus media.

Following the approach of [18], [19] who found that relived or recalled emotion generated genuine spontaneous reactions, we prompted participants with an emotion word and asked them to recount the story of an intense experience with modifications described in Methods.

2.3 Recognition Modalities

Touch: We can measure touch as force magnitude (pressure) and location – dimensions used for gesture recognition as well as for control directives using trackpads and touch screens. Social touch gesture studies report prediction accuracies ranging from 53% (chance 7%) [20] to 86% (chance

1. Emotions for classification by touch differentiates emotions in the quadrant borders, namely: *Distressed, Aroused, Excited, Miserable, Neutral, Pleased, Depressed, Sleepy, Relaxed* [17].

11%) [7] depending on collection and classification methods (Bayesian classifiers in the former and random forest in the latter case), and like affect studies in general, have no consistent standard. Still, these prediction rates on defined gestural subsets suggest that social touch may be used as directives in systems with embedded recognition systems.

Accurate *emotion* recognition is more difficult. Human recognition of human emotion through touch reaches 59% accuracy (chance 8%) [32]. Machine classification has demonstrated 36∼48% accuracy (chance 11%) [17] depending on inclusion of participant information. Both studies utilized emotion *intent*, not *experience.*

Gaze: Our eyes give affect cues discernible with eye tracking technology, making gaze behaviour an easily accessible emotion-embedding modality to pair with touch without hindering interaction. Like touch, gaze detection technology collects eye behaviour at the focal location and does not require participants to wear sensors on their body. [34] studied the effect of emotional auditory stimulation on pupil size variations, finding that negative and positive stimulation resulted in significantly larger pupil dilation than neutral stimulation but did not differentiate stimulus valence. Other factors, such as changes in luminance [35], can also affect pupil dilation.

An alternative is to analyze where a person is looking. [14] tracked students' gaze when they interacted with a graphical intelligent tutoring system; fixation and saccade features revealed that curious and bored students looked at different interface areas – e.g., engaged students looked more at the table of contents. Overall, boredom and curiosity could be predicted with 69% and 73% accuracy respectively.

We could not find literature on the use of human gaze *point* in classifying emotions using the valence/arousal model. Gaze point is related to boredom and curiosity, and low arousal is correlated with decreased saccadic velocity [36], but can gaze express arousal change too? Does gaze point move more during excitement? Compared to pupil size variation measurements, gaze point can be measured in a less controlled environment (lighting and luminance impact data quality less) with relatively inexpensive tracking technology. Thus, we utilize the Cartesian coordinates of user gaze point in our own classification analyses.

Biometrics: Blood volume pulse (BVP), skin conductivity (SC) and respiratory rate (RR) have been widely used to confirm emotion detection in other modalities – facial expressions [16], affective audio [15], [37], gaze behaviours [38], and touch behaviours [9]. Heart rate variability has been utilized in emotion classification [9], [39], [40].

Like others, we employed three basic signals (BVP, SC, RR) to calculate a set of derived features based on heart rate variability (HRV), breathing rate variability (BRV), or both, such as heart beats per breath. This data is most appropriately compared with studies where emotion elicitation is based on true experience and uses the same emotion sets. For example, [15] uses validated music excerpts to generate authentic responses crossing four musical emotions (positive/high arousal, negative/high arousal, negative/low arousal, positive/low arousal), and reports affect recognition rates between 70% and 95% (chance 25%), with higher rates when participant knowledge is included.

3 METHODS

We asked participants to recall emotionally intense experiences, while interacting with our static (non-mobile, unmoving) robot pet as a tangible focus for emotional interaction. We collected touch, gaze and biometric data; and emotion self-reports before and after each emotion. Of 30 campus-recruited participants (mean age 25.4 years, σ=5.4), 14 identified as female and 18 had corrected vision. Participants were compensated $20 for a ∼60 minute session.

In the following we detail data collection setup and procedure, and describe data pre-processing, feature extraction, and analysis of the study's independent parameters (*window size*, *inter-window gaps* and *participant knowledge*).

3.1 Data Collection

To facilitate emotion elicitation during memory recall, we prioritized participants' comfort. We placed the gaze tracking system coincident with touch site, since the robot body is the focal site for both modalities.

Configuration and Room: We conducted the experiment in a sparsely furnished medium-sized office with a window with a pleasant view. Participants sat, back to the door, comfortably in a half-prone position on a couch, for comfort and to reduce large-scale body movements (Figure 1). An experimenter was in view of the participant except during emotionally intense parts of the session, as described below.

We placed the gaze tracker (designed for mounting beneath a computer monitor) below an angled, monitor-sized board on which we placed the robot, all in comfortable reach of the participant. We fixed the robot position to prevent it from being picked up or substantively moved around to avoid interference with gaze tracking (Figure 1). By coincidence, all participants were right-handed (though the set up was designed to accommodate both right- and left-hand dominance) and we omit a discussion on handedness.

Touch Sensor on a Passive Robot: Figure 1(a) shows the robot's and sensor's construction. We used a custom flexible touch sensing apparatus previously described in [8], which has been validated as for the ability to capture social touch gestures. Similarly to [7], [42], it can detect 5g∼1kg of weight with resolution of 10×10 inches at one taxel per square inch[2]. As with [7], [17], we specified fingerpad-size taxels (touch pixels): emotion tasks in touch generally incite broad rather than precise movements [32]. While higher resolution sensors are needed for precision tasks (e.g., for touch screens, trackpads, or teleoperative mimicry [43]), here we are concerned with cost, sensor flexibility and computational efficiency.

Forming a 10-by-10 grid, this fabric-based device can sense multiple simultaneous touches (multitouch), registering varying pressures on each taxel scaled to 1024 levels and polling at 54Hz. This resulted in 54 frames of 100 cells per second, each reading a touch pressure value in [0-1023].

The "bot" was assembled in layers. The interior was a compliant structure of flexible binder plastic, roughly the size and weight of a football. The robot's body and passive

2. Built from commercially available piezoresistive and conductive fabric. Fabric is commercially available at www.eeonyx.com.

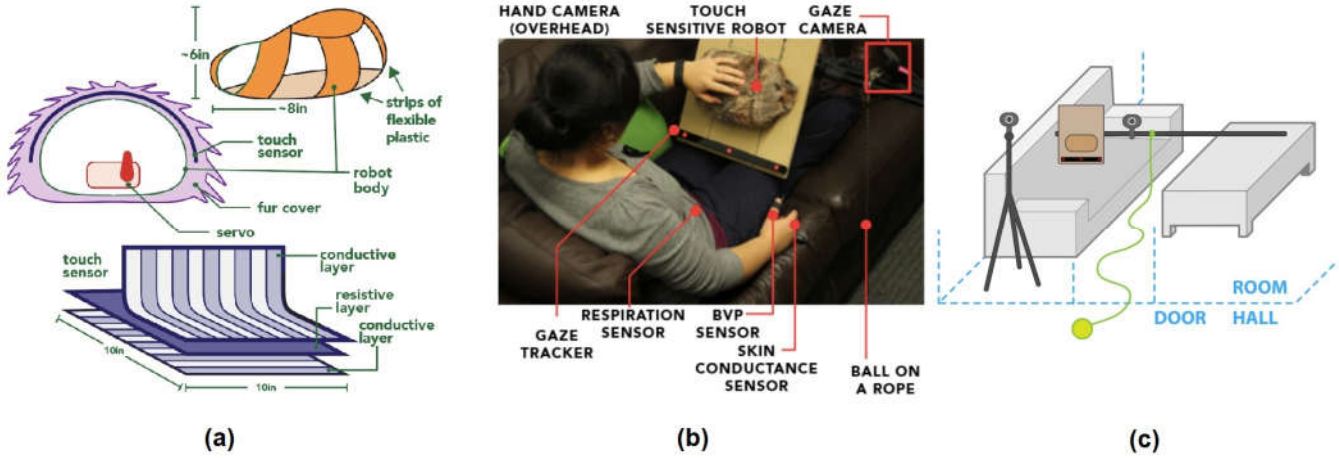

Fig. 1: Study setup overview: robot description and participant experience. **(a)** The robot was constructed from pliant plastic sheets actuated by a pulley, covered with a custom touch sensor, then jacketed in furry fabric to invite touch [41]. It was stationary during the study to eliminate reaction to robot motion. **(b)** A participant sits supported by pillows and facing the gaze tracker, one hand on the sensor-clad, stationary robot, biometric sensors on chest (RR), thumb (BVP), and index / ring fingers (SC) of resting hand. **(c)** A schematic of the study room, depicting camera locations relative to where the participant sits by the robot platform.

feel were designed to invite touch as an ambiguous mammalian form that does not resemble any definitive animal in order to remove behaviour expectation [41], [44]. Movement was disabled here to reduce confounds from novelty effects, sounds, or expectations. The touch sensor was wrapped over the structure, affixed with velcro. Finally, the sensor was covered with a uniformly-textured short, soft brown minky fabric (such as that used in baby blankets; described as *"pleasant to touch...[and] reminded me of my chocolate lab's head"* – P04). To minimize visual clutter, all sensors were wired through the robot platform and gathered in a compact tether for connection to a single laptop.

Gaze and Biometric Sensors: We sampled gaze behaviour via a Tobii EyeX gaze tracker at 60Hz – as with our touch data sampling (Figure 1). We gave no specific instructions regarding gaze direction, but informed participants that gaze data collection worked best when they were facing forward and did not make large body movements.

We collected three biometric signals using the pre-packaged Bio-Graph Infiniti Physiology Suite[3], namely blood volume pulse (BVP), skin conductivity (SC), and respiratory rate (RR), all at 2048Hz. Following established procedures [15], these were expanded to include features on heart rate variability (HRV), breathing rate variability (BRV), and cross-signal indicators such as heart beats per breath.

Participants wore a respiration band around their chest, with the closest fit that did not impede breathing. Once the participant was comfortably seated, we positioned the BVP sensor at the thumbpad, then positioned the SC sensors on the index and ring finger pads. Both BVP and SC sensors were held in place by a small velcro band on the right hand (not used for touching the robot).

3. System manufactured by Thought Technology Ltd. FlexComp ∞ SA7550 Hardware Manual can be found through manufacturer website at http://bit.ly/29A5NIC.

Video Data: We video-recorded participants' hands and face to supplement missing gaze or touch data. For participant privacy, no sound was recorded. The hand camera was placed behind, and the face camera on the right of the participant. Figure 1 shows placement of the gaze tracker.

Emotion Labels: Genuine emotion is taxing. To minimize fatigue, we administered just two emotions per participant, based on discussions with field experts, piloting and literature. The second emotion task was determined by the first; participants experienced either *Stressed - Relaxed* OR *Depressed - Excited*, counterbalanced. The four named emotions [*Stressed, Relaxed, Depressed, Excited*] comprised the emotion label set and validated via self-report on intensity and authenticity and coordinates on Russell's affect grid [21].

Procedure: Table 1 summarizes our study procedure, in which neutral steps delineated experiment steps. Emotion tasks were counterbalanced across participants.

Introduction and Calibration: To reduce novelty effects, we introduced the robot, invited touch exploration, described the robot including its sensing abilities, and explained that its movement was disabled. We then calibrated all sensors.

Neutralization and Self-report: For each stage, we first presented an emotionally neutralizing reading task, wherein the participant read aloud from a short report from a technology magazine for $\sim$5 minutes. We instructed the participant to read each word, told them that no questions would be asked of the readings, and encouraged them to let go of residual emotions from their day.

We then asked the participant to report their current emotional state. Before each emotion self-report, an experimenter explained or reminded the participant of concepts of arousal and valence, answered questions about reporting emotional state, and showed them how to indicate their current emotional state on a form displaying Russell's [27] 2D affect grid varying in arousal and valence [25]. This

TABLE 1: Experimental procedure and data acquisition.

Step	Description (duration)	Data or Output
(1) Intro	Describe study tasks	informed consent
	calibrate sensors	verify data quality
(2) Neutral 1	Read neutral text (5 min)	biometrics
	Self-report	emotional state
	Calibrate gaze/touch sensor	calibration logs
(3) Emotion 1	Recall memory ($\mu = 4.23$ min, $\sigma = 3.09$)	biometrics, gaze, touch
	Self-report	emotional state + authenticity rating
(4) Neutral 2	Read neutral text (5 min)	biometrics
	Self-report	emotional state
	Calibrate gaze/touch sensor	calibration logs
(5) Emotion 2	Recall memory ($\mu = 4.23$ min, $\sigma = 3.09$)	biometrics, gaze, touch
	Self-report	emotional state + authenticity rating
(6) Debrief & Interview	Interview	qualitative data
	Self-report	emotional state

self-report was repeated before and after each neutralizing and emotion task. For emotion tasks, participants were also asked to rate how strongly or authentically they experienced the emotion, compared to the original incident.

Reliving Emotion Task: We next asked the participant to recall an emotionally intense memory pertaining to an assigned emotion word {*Stressed*, *Excited*, *Relaxed*, or *Depressed*} as they interacted with the robot. To elicit strongly emotion-influenced touching, we invited them to relive the emotion as intensely as possible while keeping their non-instrumented hand on the robot. We explained that audio recording was disabled in the video camera and we could not hear them speak from outside the room. They received no other touch instruction or reminder. After we left the room, they described their memory with its associated feelings to the robot in any language, at a volume of their choosing. The participant indicated task completion by pulling a signal rope. Data was collected for a single recalled memory (duration μ=4.23 min, σ=3.09 min).

When the rope was pulled, the experimenter returned and administered the self-report grid, then repeated the steps for the second set of neutralization and emotion tasks.

Debrief and Interview: We conducted a short debriefing interview to learn of any unexpected eventuality during their experience, and ensure that participants were comfortable, emotionally stable, and departing in an emotional state no worse than when they arrived. We provided university counselling contacts after we found in piloting that participants could become distraught during this protocol.

3.2 Features, Pre-Processing, Extraction & Analysis

We recorded touch, gaze, and biometric data for affect classification features (see Table 2 for a full list). Here, we describe the feature extraction process.

Distribution statistics: We included conventional touch statistics [7], [8], [17]: min, max, mean, median, variance, total variance, area under the curve (AUC) for location X- and Y-centroid and touch pressure. Touch pressure is computed by frame: pressure values per capture of the 10x10 sensor. For the centroid, we found the cell containing the coordinates of the touch-pressure centre of mass (X-centroid, Y-centroid); i.e., the weighted average of all taxels in a frame based on their row and column locations, or (X, Y) coordinates respectively. Gaze focal location (x,y) and biometric channels of blood volume pulse (heart rate), skin conductance, and respiration rate were similarly calculated.

TABLE 2: Summary of features extracted from *touch*, *gaze*, and select *biometric* signals.

FEATURE	SIGNAL	#
TOUCH (54Hz)		
distribution: max, min, mean, var, total var, AUC (Area Under Curve)	Xcentroid, Ycentroid, frame pressure	21
frequency: peak count, fundamental frequency, amplitude max, mean, var & total var	Xcentroid, Ycentroid, frame pressure, pressure of centroid cell + 8 nearest neighbours (9 vals)	72
GAZE (60Hz)		
distribution: max, min, mean, var, total var, AUC	X, Y, saccade length, velocity, fixation duration	25
sample counts	total samples, on/off-robot, off-on robot ratio, rate within platform range, saccade count, saccade rate, fixation count, fixation-saccade ratio	9
frequency: peak count, fundamental frequency, amplitude max, mean, var & total var	X, Y	12
BIOMETRICS (2048Hz)		
summary statistics: mean, median, variance	*Blood Volume Pulse (BVP):* amplitude, high frequency power (FP), low FP, very low FP, heart rate, inter-beat interval, peak amplitude	228
	Skin Conductance (SC): mean, epoch mean	228
	Respiration pattern: abdominal amplitude, respiratory rate, period	228

Thought Technology's commercially available calculations were used for biometric feature extraction: http://www.thoughttechnology.com

Frequency statistics: Based on prior indications of promise [17], we extracted frequency-domain features to assess how well they encode emotion content. We calculated six frequency statistics for 12 touch signals and the same six for two gaze signals. We directly calculated frequency-domain touch and gaze features, and used Thought Technology's pre-packaged signals[4]for biometrics.

3.2.1 Feature Extraction

We calculated distribution and frequency statistics for touch and gaze. For biometric features, we relied on prepackaged calculations but also computed simple statistics (mean, median, variance) for insight into distribution characteristics. Table 2 summarizes the full feature set.

Touch features: We reprised known procedures for social touch recognition by constructing three parameters [7], [8]: *touch pressure* (sum of pressure readings from taxels in frame); and *column* and *row centroids* (weighted measure of row, column centres of mass based on frame taxel pressure, or X-centroid and Y-centroid respectively). We computed 7 statistics per pressure parameter, for 21 features.

For frequency-based features of emotive touch, we performed a Fast Fourier Transform (FFT) of the three frame-level pressure and the centroid coordinates (*x*,*y*) described above; and then calculated 6 frequency statistics for each as well as the pressure readings from the centroid cell and

its eight nearest neighbors [17], comprising 72 more touch features in the frequency-domain.

Gaze features: From the gaze data, we collected raw (X, Y)-coordinates of focal points from the Tobii eye tracker and calculated 34 features: distribution statistics for each of {focal coordinate pair (X-, Y-location), saccade length, velocity, fixation duration} as well as 9 summary features of gaze presence and location including saccade and fixation ratios. We used Salvucci's I-VT algorithm [45] to differentiate between fixations and saccades. Gaze samples with point-to-point velocities $<30°/s$ were classified as fixations and those with velocities $>=30°/s$ as saccades. We calculated 6 frequency statistics for gaze data on the 2D focal location, generating 12 frequency-domain gaze features.

Biometric features: We computed mean, median, and variance across all signals provided from the Thought Technology physiology suite, including both base signals (BVP, SC, RR), and channels dependent on the original signals (HR, HRV, IBI, etc.), for a total of 228 features across 76 channels.

3.2.2 Data Instances / Partitioning on Independent Factors

Each data instance is comprised of a list of touch, gaze, and biometric features computed across a single time window. We omitted windows that provided insufficient samples for FFT (<10) for any modality[5] – generally due to gaze data loss when gaze was outside of the tracked area. We partitioned our data and analyzed how key computational factors influence classification accuracy: window size (data density), inter-window gaps (continuity), and participant knowledge (content) (Table 3).

TABLE 3: A motivating overview of analysis factors.

WINDOW SIZE: [0.2s, 0.5s, 1s, 2s]	
Description	Data was all sampled to 54Hz. Window size is the length of time over which a feature is calculated. e.g., a two-second window has 108 samples.
Implication	With a static sample speed, shorter windows simulate a system with faster update cycles, resulting in less information per window, but faster system response.
Question	How do accuracy rates change with different sample sizes?
INTER-WINDOW GAPS: [Without gaps, With 2s gaps]	
Description	With no gap, all windows are calculated contiguously, i.e., every window is directly adjacent to the one previous. With gap, after every window is calculated, two seconds of data is discarded.
Implication	Social touch gestures take a little under a second to make [7] so a 2s gap increases the likelihood that each window captures different gestures.
Question	How robust is the system to data loss?
PARTICIPANT KNOWLEDGE: [Explicit, Implicit, None]	
Description	The system may select participant labels if included in the training data. We have three levels of participant knowledge: participant labels included, participant labels excluded (both subject dependent), all participant data excluded (subject independent).
Implication	When labelled, the system can tell whose emotions it is attempting to predict. When unlabelled, the system still has knowledge of the participant's behaviour, but cannot determine from whom. The most challenging case: testing on a participant's data without her training samples.
Question	How much does *a priori* identification of an individual influence classification accuracy?

Window Size: Impact of window size on classification is crucial for compute-constrained real-time gesture classification. 2s windows (54Hz, or 108 frames) have been used to capture touch gestures [7]; however, human hands and fingers can move at ∼100-200ms [46], [47].

We therefore partitioned data in 2s non-overlapping windows and extracted features for training and test instances. Each data instance has features extracted from a 2s window to build a classification model. This partitioning and feature calculation were performed on the same data at other window lengths, resulting in four distinct sets of data instances at [0.2s, 0.5s, 1s, and 2s] windows.

Inter-Window Gaps: Even though our Random Forest classification model treats instances without temporal dependence, we consider that temporally-neighbouring instances can be exceedingly similar, particularly in the smallest windows. We investigate whether, and by how much, recency effects influence accuracy rates by adding 2s gaps between instances thereby eliminating adjacent instances. We compare classification performance of the data with and without this artificial gapping (gapped vs un-gapped data.)

Participant Knowledge: We report accuracy for **emotion** classification across three levels of the classifier's knowledge of the participant in increasing information order:

1) **No participant knowledge –** *subject-independent* classification simulates the task where an interactive system's emotion model cannot be trained on all possible users. E.g., a robot in a museum or institutional context must be modelled on a training set that could not include all possible users, who are not known ahead of time.
2) **Implicit participant knowledge –** this *subject dependent* system simulates a classification task where the interactive system's emotion model has been trained on all expected users before classification but not explicitly informed which data is associated with the current user. We imagine a system that lives in a limited private domain, where all users have completed a calibration period, informing the model's training set.
3) **Explicit participant knowledge –** the training set includes participant labels as a feature (subject-dependent where instances are attributable by subject). This system knows whose emotions it is attempting to classify and loads a personalized emotion model for each user.

We also ran **participant** classification to determine not only how well these feature sets can determine *what interaction* was performed, but also *who* performed it.

3.2.3 Classification

Here we summarize the classification tasks: predicting *emotion* and *person* experiencing the emotion while experimenting with data instances comprised of our statistical features and varying window size, inter-window gaps, and participant knowledge. For literature comparison, we report classification accuracy as the ratio of correctly classified instances over all instances as well as multi-class weighted F1-scores based on the instance count of each class.

We used Weka, an open-source machine learning platform [48], for k-fold cross-validation (CV) using a Random Forest (RF) classifier – so chosen for its known efficacy

5. On average, usable data instances dropped by 36% with shorter data windows being more affected.

for touch recognition [7], [22] and low training and computational threshold – to assess classification accuracy on both pressure-location and frequency domain features. We chose a relatively moderate value of $k = 20$ for our CV, to support comparison with other studies which have shown this method to be effective in touch classification [7], [8], [17], [20]. We included subject-dependent tests for models trained on all participants as there is no restriction on whose data instances are included as training or test data, so long as the same data instance is not in both.

Subject-independent Emotion Classification: For subject-independent analysis (no participant knowledge), we use two types of Leave-One-participant-Out (LOpO) classification: (1) one participant's data is left out and the training set includes *all other participants (LOpO-All)* (i.e., training $N = 30 - 1$) and (2) one participant's data is left out and the training set includes all other participants *who performed the same emotion tasks (LOpO-Half)* (i.e., training $N \approx 15 - 1$[6]).

LOpO-ALL simulates a system that has no knowledge of a new user and has been trained on all emotional touch behaviours (chance $\approx 25\%$). **LOpO-HALF** simulates a system that has no knowledge of a new user and has been trained only on the 50% subset of behaviours this user *will* be performing (chance $\approx 50\%$).

Subject-dependent Classification: Given the highly individual nature of the touch behaviours we observed, it is possible to expect LOpO classification to perform at or near chance. We also performed CV for conditions classifying:

1) *Participant:* represents a system trying to identify *who* is performing the interaction.
2) *Emotions given explicit participant knowledge:* participant labels are included as a feature;
3) *Emotion given implicit participant knowledge:* participant labels are omitted.

4 RESULTS

Consistently with past studies on biometric-based emotion classification [15], our biometric data alone gave accuracy rates from ~90% to near 100% (Table 4).

This section describes our results from running classification using our full feature set on emotional touch and gaze behaviour across a number of experimental conditions (compared to that of biometrics alone) (Table 4). We also look at subject-independent tests of emotion classification which also employed the maximal combination of modalities (touch + gaze + biometrics) (Table 5). F1-scores and accuracy differ by less than 0.03 (3%), with most within 0.001 (0.1%) difference. Since they follow the same patterns by condition, we discuss them in terms of accuracy outcomes for comparability to other multiclass affective classification literature [15], [17], [32], [49], [50].

4.1 Subject-Independent Emotion Classification

Subject-independent classification, LOpO-ALL (chance $\approx 25\%$) was run as a single RF trained on all emotions while LOpO-HALF was built on two RFs trained independently for *excited-depressed* and *stressed-relaxed* respectively. For each LOpO level, classification was performed at each window size and gap condition.

Some participants fit the model well, most performed at chance, and, interestingly, a few consistently contradicted the generalized model. For all LOpO levels, window sizes, and gap conditions, **accuracy was very near chance** (Table 5, LOpO-ALL and LOpO-HALF).

6. Test sets are comprised of data instances from participants who performed *Stressed* and *Relaxed* ($N_{SR} = 16$) or ones who performed *Excited* and *Depressed* ($N_{ED} = 14$)

TABLE 4: Weighted F1-scores from 20-fold cross validation varying factors of Gap(+/-), Participant Labels(+/-), and Window Sizes (0.2s, 0.5s, 1s, 2s) on touch T, gaze G, and biometric B features, classifying emotion ($25\% \leq chance < 50\%$). Classification accuracy is within 0.003 from these values. Weighted F1-scores that are from 0.01 to 0.03 below classification accuracy are indicated with *.

		Participant Labels-				Participant Labels+			
	Win	**T**	**G**	**TG**	**B**	**T**	**G**	**TG**	**B**
Gap +	**0.2s**	.666	.412*	.704	.997	.871	.744	.884	1
	0.5s	.693	.448*	.735	.996	.881	.773	.897	1
	1s	.719	.489*	.759	.996	.886	.781	.909	.999
	2s	.566	.465*	.597	.892	.793	.651	.765	.942
Gap-	**0.2s**	.754	.475*	.822	1	.923	.788	.944	1
	0.5s	.761	.505*	.823	1	.921	.803	.939	1
	1s	.768	.530*	.821	1	.921	.811	.937	1
	2s	.761	.569	.815	.999	.918	.813	.931	1

TABLE 5: Overall classification performance across all test conditions and modality combinations by accuracy and weighted F1-scores.

TEST	DESCRIPTION	CHANCE	ACC	F1
LOpO-ALL	Predict one of four emotions	25.0%	34.5%	0.318
LOpO-HALF	Predict one of two emotions	50.0%	58.0%	0.574

4.2 Participant Classification

Previous results have demonstrated that participants have a *touch signature*: ways or styles of touching which can be sufficiently idiosyncratic to identify the toucher [7], [8]. Individual touch behaviours were both internally consistent and externally unique.

To see if this was true of our data, we performed 20-fold CV on the full set of data instances, to predict subject label (*who* performed the gesture) on touch instances, resulting in a classification accuracy of 78%, where chance is 1/30 or 3.33%. High accuracy rates on participant prediction confirms that individual differences are indeed highly expressed in this type of behavioural data.

4.3 Subject-Dependent Emotion Classification

With participant classification (Section 4.2), we looked for touch behaviour high in both individual differences and consistency. With emotion classification we seek *commonalities* in touch behaviours across individuals, under given emotional conditions. We expect one of the following to be true: (a) participants feeling the same emotions touch the robot similarly, s.t. we can differentiate solely on emotion condition; (b) given knowledge of a participant, we can

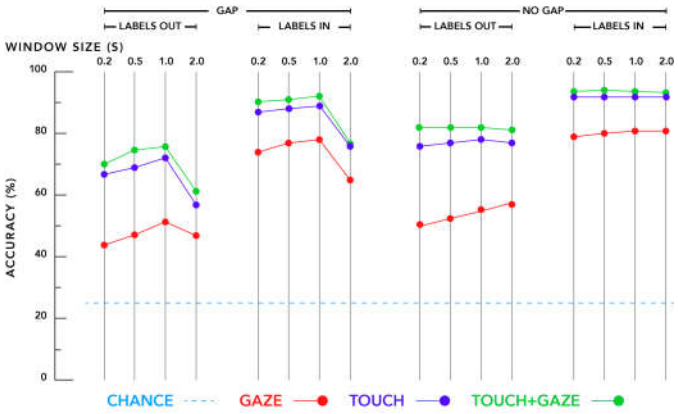

Fig. 2: Emotion classification accuracy rates from 20 fold cross-validation by *modality* (Touch + Gaze, Touch only, and Gaze only), *window size* (0.2s, 0.5s, 1s, 2s), as weighted averages from Table 4. Comparisons are also made between having participant labels included (b) & (d) vs excluded (a) & (c), and where 2s gaps are imposed to simulate data loss (a) & (b) vs no gaps (c) & (d). Including biometric data consistently achieves 90-100% accuracy across windows, labels, and gaps (accuracy dips only under the sparsest data conditions: gapped-2s window cases, regardless of whether subject labels are present).

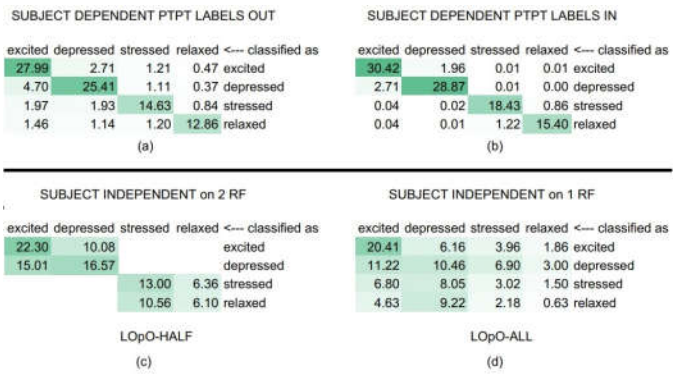

Fig. 3: Comparing how each classification task performed by emotion using touch and gaze features. For subject independent analysis (c) we trained 2 RFs–trained on *Excited-Depressed* and *Stressed-Relaxed* separately (no between-set classification – blank entry for *Depressed-Stressed*). In contrast, a single RF was trained on all 4 emotions in (d).

differentiate between two emotion tasks; or (c) some combination where a system does not explicitly know who a participant is, but can differentiate given a touch signature characteristic of a specific participant.

(A) is unsupported based on our LOpO results where named emotions are recognized at near chance. We focus this section on the feasibility of personalized models of emotional touch: the consequences of (b) and (c); the effect of noisy or inconsistent data to simulate real-world operation; and finally, how the relative contribution of touch and gaze compare with respect to classification accuracy.

We review classification performance with respect to data factors described in Table 3.

4.3.1 Accuracy by Emotion

We break down the average accuracy rates for emotion classification and compare how the classification task affected performance for each emotion (see Fig 3).

Unsurprisingly, subject-dependent CV performs significantly better than subject-independent LOpO; notably, however, *Excited* behaviours can be classified at roughly similar rates. There are a few contributing factors to be considered: (1) *Excited* behaviours were of consistently high arousal with quick motions; while *Stressed* was also high arousal, participants often associated it with fighting *Depressed* feelings. (2) Participants provided longer samples of *Depressed* and *Excited* expressions, which led to more data instances when cut into equal-length windows (see Table 6 in Appendix).

4.3.2 Window size and Gapping

Comparing classification accuracy by window size, we see that overall, increasing window size improves performance.

We imposed data gaps to simulate real-world loss, reducing temporal inter-dependency. Where data was uninterrupted (Figure 2c,d), classification rates are relatively stable regardless of window size.

While introducing gaps (data discontinuity) causes expected dips in performance, larger window sizes suffer disproportionately. Closer inspection reveals that this accuracy drop-off coincides with a decrease of training instances – most severely at 2s, where data instance count drops from 7435 instances down to 676, an over 90% data loss.

4.3.3 Participant knowledge

Where participant labels are known (Figure 2b,d), classification accuracy improves over cases with no participant knowledge (a,c). This effect is seen consistently across modalities with jumps as high as 10-20% for touch- and gaze-only, respectively.

4.3.4 Comparing modalities

We refer to Table 4 to assess how touch (T), gaze (G), touch + gaze (TG), and biometrics (B) compare in subject-dependent emotion classification performance (20-fold CV).

Taking modalities alone, we see that gaze performs comparatively lower than touch. When participant labels are available (Figure 2b,d), classification on both single modalities improve. However, combining touch and gaze further increases accuracy. Particularly under the best condition of maximal information ((d) – with participant labels, no gaps), touch and gaze together can approach that of biometrics performance (97-100%) – in line with previous work showing high classification performance on physiological data [15].

4.4 Feature Set Analysis

To understand feature contribution, we ran Weka's Best First Attribute Evaluator [51] on the Touch and Gaze feature set. This tool iteratively selects the best feature subset for each classification trial in 20-fold CV, producing a list of features and the frequency with which they are selected.

Figure 4 breaks down each parameter by modality and relative selection count as a heat map, where each cell represents the number of features of a statistical type selected at each iteration. Higher saturation indicates a higher number

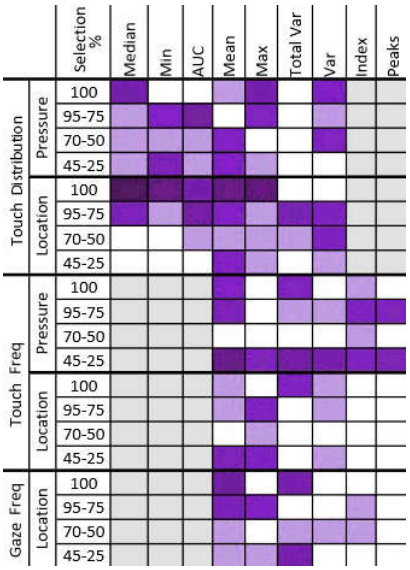

Fig. 4: Feature selection count by statistic as ranked by Weka's Best First Attribute Evaluator. Selection % represents how often the feature is selected for use in 100 iterations of 20-fold CV. The dark box for Touch Distribution-Location x Median indicates that this feature is selected 100% of the time; white boxes indicate features that were never selected.

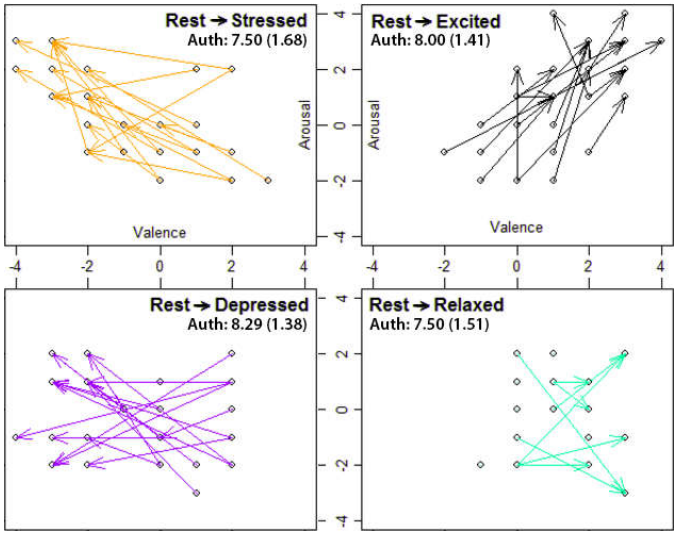

Fig. 5: Changes in individual's self-report of emotion after Neutralization (start) and Emotion tasks (finish); N=14 for *Stressed* & *Relaxed* and N=16 for *Depressed* & *Excited*. Overall, we see a move from the origin to the representative quadrant. *Stressed* and *Excited* show the strongest overall change along both **Arousal** and **Valence** axes. *Relaxed* shows the least change with disconnected points referring to "no change" from neutral state.

of times selected at this percentage. For example, *Median-Touch Location* was selected in every CV trial.

The most selected features were the 11 calculated medians of touch location, chosen 100% of the time during 20-fold CV. Overall, when using *Classic Touch Location* data, we recommend calculating *Median, Min, AUC, Mean, Max* features; in contrast, when using *Classic Touch Pressure* data, *Total Variance* is not chosen at all and may be left out.

4.5 Reports of Experienced Emotion

Participants reported their current emotional state with Russell's 2D affect grid [21] during two neutralization tasks and following two emotion tasks. After completion of all emotion tasks, we interviewed our participants on their experience; highlights are covered in this section.

Self-reported emotion movement: In Figure 5, there is variation where we expected participants to report emotion movement towards the quadrant extremes. In decreasing order: *Excited* (all 14 participants reported moving towards the quadrant extrema); *Stressed* (13/15); *Depressed* (6/15); and *Relaxed* (2/16). In paired t-tests, we found significant differences in self-reports between neutral and emotion tasks for each of *Stressed*, *Depressed* and *Excited* in both arousal and valence ($p < 0.05$).

Paired t-tests showed no significant difference ($p > 0.05$) in neutralization tasks, nor order effect in emotion tasks.

Figure 5 plots each participant's emotion trajectory across the 2D affect grid for each relived emotion instance, from starting state to recall conclusion. Both high arousal emotions (*Excited*, *Stressed*) were consistent with expectations where participants reported a shift in emotion toward the grid corner of the target emotion word.

Authenticity: Each participant self-reported how authentically they experienced the target affect in each emotion task. On a scale of 1–10 with 1 being *completely contrived or artificial* and 10 being *completely authentic as in the original experience*, participants rated authenticity highly (between $7.5 \leq \mu \leq 8.29$) with *Relaxed* and *Stressed* tied, and then *Excited* and *Depressed* in increasing order.

Added insight from interviews: For some, immediacy or recency of recalled events helped to highlight emotions. This experiment was run around final-exam and holiday reunion time. Both are cited as reasons for ease of recall.

> "I'm leaving to see my family for the first time in three years, I can't stop being *Excited*." – P09
> "*Excited* was easy – the situation was more recent and was more important [than my *Depressed* memory]." – P22
> "I have a lot of school assignments right now and I kind of toggled between many memories [*Stressed*]. It was hard to pick one to feel but I think that might have added to the feeling." – P21
> "[W]hen I was doing *Stressed*, I felt like I wanted to punch something it was so gut-wrenching." – P29

The low arousal emotions, *Relaxed* and *Depressed*, moved as expected in valence but not arousal, which remained overall at its neutral "resting" position. In the case of *Relaxed*, this might be explained by perceived similarity between this emotion task and the 'resting' start condition.

> "*Relaxed* was easy to express because it's pleasant and I want to feel it and also, I'm sitting on a couch which helps." – P27; similar reports by P02, P18

For these two emotions, some participants reported that the emotion *Depressed* was linked to *Stressed* in their memories (e.g., feeling stress about exams was also depressing), which may explain some of the unexpected movement in arousal

for *Depressed*. Four participants also reported feelings so strong that their *Depressed* memory evoked active tears, while others indicated that these feelings were somewhat mitigated by the experience of stroking a soft body.

> "My [*Depressed*] memory was very clear and I was able to recall a lot of details. It really helped to be touching a soft thing and felt like it was taking some of my sadness." – P26, also P15, P24

Another possibility for both of these emotion targets is that participants were simply unable to turn down their arousal state to this degree during the short time of the session.

5 Discussion

We summarize result highlights before contextualizing them in our research questions:

- Using both touch and gaze **improved** accuracy rates over touch alone.
- Increasing window size **had little effect** on accuracy.
- Adding data blackouts or gaps **did not noticeably decrease classification accuracy** except for 2s windows.
- Due to individual differences in touch behaviour, it is necessary to include participants in the training set for potentially usable recognition:
  1) Classification accuracy for *whom* (participant performed a data instance was comparable to that of WHAT (emotion), implying that individual differences can be captured;
  2) Both LOpO-ALL and LOpO-HALF analyses performed at or near chance;
  3) Including participant information in the training set **improved** accuracy rates, but **participant labeling is not necessary** for recognition.

5.1 RQ1: Ability of Touch and Gaze to Predict Emotion

As anticipated, accuracy of distinguishing between emotions based on a full suite of biometric signals approached 100% in the best-case model (Figure 2a) trained on participant-labeled data (Table 4, column B). When full-suite biometric signals can be effectively employed, they will give the best result. Even partial biometric sources – e.g., heart rate variability (BVP) alone – do well relative to each of the less intrusive modalities. We can expect improvement in the wearability or embeddability of some biometric channels, so this result is important to note.

Of modalities not requiring sensors to be worn (touch, gaze), touch reaches 92% accuracy[7], improved with gaze to 94%; however, performance worsens in more adverse conditions. This level of classification accuracy may be adequate for many applications, e.g., when the goal is simply to establish large-scale movement between quadrants.

Classification accuracy favours pressure-location distribution features: At 54Hz, touch distribution features of pressure and location were most frequently selected for emotion-classification performance (Figure 4).

Touch frequency-domain features have been used successfully [17]; the contrast may be our relatively low sample rate coupled with short windows (0.2-2s vs 8s in [17]). Further, emotion classification using gaze data appears to consistently benefit from inclusion of features calculated on Fourier transforms of gaze position. Since frequency-domain features are relatively compute-intensive (realtime FFT *vs.* pre-processable pressure-location set), it may be reasonable to reduce the feature set to touch distribution features where efficiency is a priority.

5.2 RQ2: Individuality

Recognition rates increase with greater participant knowledge: LOpO results near chance (for both iterations–ALL and HALF) imply low generalizability of a model to other individuals' emotional behaviour.

Participant knowledge matters, but not labels – We propose a touch-centric robot that exploits individual differences and, instead of an out-of-the-box general training model, builds personalized models of a short list of users. Having participant knowledge is important for classification; all expected users of a single robot should be included in a model's training pool. However, including participant labels adds only minor benefit (Table 4 with labels *vs.* without) when training data already includes the test participant. This may be due to the relatively high participant classification rate (Table 4; chance 3.3%) wherein participant-specific behaviours may influence classification such that even though participants are unlabelled, the system is able to guess. When high accuracy is needed, *a priori* user identification (participant-labelled data) may be a helpful refinement.

Excited is most recognizable emotion – Based on confusion matrices describing per emotion performance (Figure 3), *Excited* may be most generally recognizable. The emotion self-report (Figure 5) shows that *Excited* was experienced consistently (all participants reported the expected emotion direction). Similar emotional experiences may translate to common touch and gaze expressions in these high-arousal, high-valence emotion spaces.

5.3 RQ3: Sample Density for Realtime Responsiveness

Larger windows and including gapped data reduces classification accuracy: With post-hoc classifications, increasing window sizes and eliminating data segments (discontinuities with gapping) reduces data instance count. We discuss the effects from conditions where greatest data instance count are in no gap-0.2s window conditions and least with 2s gap-2s windows, with respect to real-time classification.

Size – From Figure 2, increasing window size from 0.2s to 2s results in marginal improvement of classification under no-gap conditions. In this case, increasing system response rate (by using 0.2s windows rather than 2s of data) may be favourable as little accuracy loss is experienced.

Continuity – Gapping data does indeed drop accuracy by 10% in T, G, and TG (Table 4). We considered the possibility that the performance decrease is related to low data instance count, but even when removing that confound and comparing equal instance intervals of gapped *vs.* non-gapped signals [8] we found that each single modality's performance

7. While classifiers differentiated four emotions, each participant performed only two. Chance is thus more like 50% when participant labels are known.

8. Addition of gaps between 2s windows reduces the data set instance count by over 90% (7435 to 676 instances).

on adjacent data streams (non-gapped) resulted in higher accuracy rates than that of gapped data[9].

Interestingly, for most window sizes (0.2s, 0.5s, and 1s — where gapped and ungapped instance counts are on the same order of magnitude) results suggest data loss should not be devastating to real-time emotion classification of touch, even when the gap (2s) is 10x that of the collected instance (0.2s). Given a relatively predictable signal interruption pattern, we can select a window size range knowing that even if a signal is lost for up to 10x that of the collected window, classification accuracy may still be tolerable.

This performance differential exposes a role of signal continuity in these channels' expression of human behavior and emotion reaction: a possible explanation is that emotion expression evolves in even short timeframes. While larger, adjacent windows may marginally improve classification accuracies for short (single-window) snapshots, they may introduce error for longer interactions. Periodic system re-training may help to build a more robust user model. Since this may interfere with actual system use, re-training could be suggested as participant behaviour changes and participant classification accuracy drops – an indication of significant behavioural departure from the current model.

5.4 RQ4: Experimental Methodology

We chose an experimental approach based on the use case of a robot pet. Several elements were nonstandard: emotion elicitation method, choice of emotions investigated, study framing (including how existing emotion models may influence the emotion task: a participant interacting with an unresponsive furry object), and analysis aspects. With results in hand, we critique these innovations.

Emotion elicitation: While the technique of memory re-telling was validated by literature [19], [25], we elicited stronger emotional reactions than we expected. In some cases, this could be due to participants playing a 'good-subject role', trying to please experimenters [52] and artificially inflating the perceived efficacy of this protocol. However, we anticipate some degree of this characteristic in any laboratory study. Furthermore, we noted some strong physical and embodied emotional reactions (such as genuine tears) that suggests this method could still be a valuable tool, particularly in a laboratory setting where people may otherwise find it hard to act naturally. We plan to employ variations in our own future studies.

Emotion set: We reported both high and low classification accuracy rates, but nevertheless question whether accuracy is an indicator of a successful emotion model, even when corroborated by F1-scores. There is certainly value in accuracy metrics, but underlying assumptions of both dimensional and discrete emotion models present known problems for classification. Specifically, discrete systems based on dimensional models suffer from a problem of distinguishability in which semantically dissimilar emotional labels are placed in the same bins [53].

Study and Emotion Task Framing: We assumed that participants express a roughly *steady state* emotion, felt across the entire memory recall. However, it is possible that strong emotions may be felt only for an instant before autonomic emotion regulation or coping mechanisms take over [54]. The horizon over which we sample a participant's emotional state, and the assumption of immediacy impact decisions an interactive system should implement. Our discrete classification system can identify differences in minute-long interactions, but cannot estimate an emotional inflection point (i.e., transition from one emotion to another). A truly interactive system would need to react to the *change* in an emotional state and adapt over many samples.

Furthermore, in natural emotional exchanges, interactions with pets or friends allow for error correction: an initial misjudgement can be corrected with further context. An adaptive rather than prescriptive model might go further towards develop a meaningful relationship over a direct and immediate call-and-response instructing interaction [55]. Using touch data in context with gaze and biometric analysis lays the groundwork for extending haptic human-robot interactions from instructional directives to meaningful conversational relationships.

5.5 Implications for Social Robot Applications

From our findings, we consider next steps in designing the classification system for our social touch-centric robot.

Out of the three nonverbal modalities we studied, touch may be most relevant for applications such as social robot therapy. Our findings indicate that for a previously known user, *distinguishing between a few emotional states is feasible for touch-alone.* This provides intriguing opportunities for development of therapeutic robots that could run human-affect recognition and respond by adjusting their behavior.

While gaze and biometrics improved classification, their use in practical scenarios remains challenging. For robust detection of gaze, the user must always face the robot at a certain angle or wear a calibrated head-mounted gaze tracker. Including biometrics is even more restrictive as participants must don a series of body-hugging sensors, then remain emotionless during periods of neutral user calibration before departures from neutrality (emotion) can be detected in signals such as heart rate and skin conductance. Embedding biometric sensors into the robot system may be possible but still poses some difficulty: touch interaction with the robot typically consists of momentary touch contact that may be too short and infrequent for measuring biometric signals. However, these sensory systems can be integrated in situations with careful sensor placement for gaze attention and training data collection sessions.

To be used effectively in therapy, an expert such as a therapist would need to introduce the robot and guide potential users in providing training data for recognition of emotions via touch. As participant-knowledge appears to be a key component to increasing emotion classification performance, we can conceive of a system training procedure that extends beyond simply including participant info. The robot could be personalized to first recognize and then work from a custom user profile where accuracy is crucial. Although this implies a setup cost for use, potential benefits in environments where real animals cannot be used (such as some hospital environments) may compensate.

9. 2s windows / unlabelled participants generated for T: 90.4% (adjacent) *vs.* 56.7% (gapped); TG: improved to 78.8% *vs.* 47.5%.

6 Conclusions

We presented affect classification results from emotionally influenced touch and gaze behaviours, verified against better-understood biometric data. Participants recalled intense emotional memories spanning Russell's 2D arousal-valence affect space, namely *Depressed*, *Excited*, *Stressed*, and *Relaxed*. We collected data across the three modalities via a custom fabric touch sensor embedded in a small furry stationary robot; a gaze tracker; and a biometric suite including skin conductance, respiratory rate and heartrate variability. Our data is both quantitative (sensor capture during interaction, and self-ratings of emotion genuineness and intensity) and qualitative (post-experience interviews).

For models trained with test participant data using pressure-location features, the overall emotion recognition rate was roughly 83% for touch, 87% for touch + gaze, and 99% for touch + gaze + biometrics. Performance drops steeply when test participants were left out of the training model, resulting in 31%, 31%, and 29%, approaching chance (25%). We tried increasing the feature set by incorporating frequency features for touch and gaze modalities. This resulted in emotion recognition rates of 79% for touch frequency features, 85% for frequency and pressure-location touch features, and 85% for touch frequency, touch pressure-location, and gaze frequency features combined. LOpO performed similarly poorly at 30%, 32%, and 35% respectively.

We summarize findings that will inform our next stage of design for robots capable of real-time emotion classification:

1. Emotional behaviour encoded in touch and gaze interaction may be sufficient. While including biometric data greatly improves accuracy, current technology requires they be worn, resulting in a more restrained experience. Setup interferes with natural emotional expression and sensors affixed to the hand and body can feel restrictive.

2. An individualized training or calibration phase is crucial for a personalized prediction system. Increasing participant information greatly improves the classification model's prediction accuracy. While this stage likely requires guidance from an expert or therapist, the training investment facilitates the learning of user-specific characteristics and develops a more robust user behaviour model, thereby allowing for a personalized and productive experience.

3. Sampling density and feature count may be reduced to improve computation load. During real-use, the speed of classification and reaction is a serious concern. Lossless continuous capture is ideal, however, in real-time we may find that packets must be dropped from slow or problematic data captures. We experimented with introducing gaps in data for this reason, and our findings indicate that interruptions in data collection at up to 2s intervals may be tolerable.

4. Limitations of commonly used emotion models should inform future research in this field. Although we achieved possibly usable classification rates, reflections from the field suggest that existing affect models have clear limitations that must be addressed [5]. People do not experience emotions in isolation nor discretely; emotional experiences follow a trajectory with distinctive peaks and valleys. Future detection systems must model the rise and resolution of an experience. While this study used a stationary robot, a deployed interactive system must acknowledge that its response has influence over user emotional reaction, necessitating dynamic adjustments to behaviour modelling.

Acknowledgments

We thank Dr. Jessica Tracy for directing us to relived memories as an emotion elicitation strategy, and Merel Jung for her valuable input in developing the methodology. This work was funded in part by Natural Sciences and Engineering Research Council of Canada (NSERC) and the Academy of Finland project Haptic Gaze Interaction (decision # 260026). The study was conducted under UBC Ethics #H15-02611.

References

[1] T. Fong, I. Nourbakhsh, and K. Dautenhahn, "A survey of socially interactive robots," *Robotics & Autonomous Systems*, vol. 42, no. 3, pp. 143–166, 2003.

[2] K. Wada and T. Shibata, "Living with seal robotsits sociopsychological and physiological influences on the elderly at a care house," *IEEE Trans on Robotics*, vol. 23, no. 5, pp. 972–980, 2007.

[3] W. D. Stiehl, C. Breazeal, K.-H. Han, J. Lieberman, L. Lalla, A. Maymin, J. Salinas, D. Fuentes, R. Toscano, C. H. Tong *et al.*, "The Huggable: a therapeutic robotic companion for relational, affective touch," in *ACM SIGGRAPH Emerging Tech*, 2006, p. 15.

[4] S. Yohanan and K. E. MacLean, "The role of affective touch in human-robot interaction: Human intent and expectations in touching the Haptic Creature," *Int'l J of Social Robotics*, vol. 4, no. 2, pp. 163–180, 2012.

[5] P. Bucci, X. Cang, H. Mah, L. Rodgers, and K. E. MacLean, "Real emotions don't stand still: Toward ecologically viable representation of affective interaction," in *IEEE Int'l Conf on Affective Computing & Intelligent Interaction (ACII)*, 2019, pp. 1–7.

[6] Y. Gaffary, J.-C. Martin, and M. Ammi, "Haptic expression and perception of spontaneous stress," *IEEE Trans Affective Computing*, pp. 138–150, 2018.

[7] A. Flagg and K. MacLean, "Affective touch gesture recognition for a furry zoomorphic machine," in *ACM Intl Conf on Tangible, Embedded & Embodied Interaction (TEI)*, 2013, pp. 25–32.

[8] X. L. Cang, P. Bucci, A. Strang, J. Allen, K. MacLean, and H. Liu, "Different strokes and different folks: Economical dynamic surface sensing and affect-related touch recognition," in *ACM Int'l Conf on Multimodal Interaction (ICMI)*, 2015, pp. 147–154.

[9] Y. Sefidgar, K. E. MacLean, S. Yohanan, M. Van der Loos, E. A. Croft, and J. Garland, "Design and evaluation of a touch-centered calming interaction with a social robot," *Trans Affective Computing*, vol. PP, no. 99, pp. 108–121, 2015.

[10] J. S. Odendaal, "Animal-assisted therapymagic or medicine?" *Journal of psychosomatic research*, vol. 49, no. 4, pp. 275–280, 2000.

[11] S. B. Barker and K. S. Dawson, "The effects of animal-assisted therapy on anxiety ratings of hospitalized psychiatric patients," *Psychiatric Services*, 1998.

[12] M. R. Banks and W. A. Banks, "The effects of animal-assisted therapy on loneliness in an elderly population in long-term care facilities," *Journals of Gerontology: Biological & Medical Sciences*, vol. 57, no. 7, pp. M428–M432, 2002.

[13] N. E. Richeson, "Effects of animal-assisted therapy on agitated behaviors and social interactions of older adults with dementia," *American Journal of Alzheimer's Disease and Other Dementias*, vol. 18, no. 6, pp. 353–358, 2003.

[14] N. Jaques, C. Conati, J. M. Harley, and R. Azevedo, "Predicting affect from gaze data during interaction with an intelligent tutoring system," in *Intelligent Tutoring Systems*. Springer, 2014, pp. 29–38.

[15] J. Kim and E. André, "Emotion recognition based on physiological changes in music listening," *IEEE Trans Pattern Analysis & Machine Intelligence*, vol. 30, no. 12, pp. 2067–2083, 2008.

[16] J. Kortelainen, S. Tiinanen, X. Huang, X. Li, S. Laukka, M. Pietikainen, and T. Seppanen, "Multimodal emotion recognition by combining physiological signals and facial expressions: a preliminary study," in *EMBC Annual Conf*, 2012, pp. 5238–5241.

[17] K. Altun and K. E. MacLean, "Recognizing affect in human touch of a robot," *Pattern Recognition Letters*, vol. 66, pp. 31–40, 2015.

[18] P. Ekman, R. W. Levenson, and W. V. Friesen, "Autonomic nervous system activity distinguishes among emotions," *Science*, vol. 221, no. 4616, pp. 1208–1210, 1983.

[19] R. Levenson, "Emotion elicitation with neurological patients," *Handbook of emotion elicitation and assessment*, pp. 158–168, 2007.

[20] M. M. Jung, "Towards social touch intelligence: developing a robust system for automatic touch recognition," in *ACM Int'l Conf on Multimodal Interaction (ICMI)*, 2014, pp. 344–348.

[21] J. A. Russell, "A circumplex model of affect." *Journal of Personality and Social Psychology*, vol. 39, no. 6, p. 1161, 1980.

[22] M. M. Jung, X. L. Cang, M. Poel, and K. E. MacLean, "Touch challenge'15: Recognizing social touch gestures," in *Proc of the 2015 ACM on Int'l Conf on Multimodal Interaction*, 2015, pp. 387–390.

[23] V.-C. Ta, W. Johal, M. Portaz, E. Castelli, and D. Vaufreydaz, "The grenoble system for the social touch challenge at icmi '15," in *Proc 2015 ACM on Int'l Conf on Multimodal Interaction*, 2015, pp. 391–398.

[24] Y. F. Gaus, T. Olugbade, A. Jan, R. Qin, J. Liu, F. Zhang, H. Meng, and N. Bianchi-Berthouze, "Social touch gesture recognition using random forest and boosting on distinct feature sets," in *Proc 2015 ACM on Int'l Conf on Multimodal Interaction*, 2015, pp. 399–406.

[25] J. A. Coan and J. J. Allen, *Handbook of emotion elicitation and assessment*. Oxford University Press, 2007.

[26] R. W. Levenson, "Autonomic nervous system differences among emotions," *Psychological Science*, vol. 3, no. 1, pp. 23–27, 1992.

[27] J. A. Russell, A. Weiss, and G. A. Mendelsohn, "Affect grid: a single-item scale of pleasure and arousal." *J Personality & Social Psychology*, vol. 57, no. 3, 1989.

[28] D. Watson, L. A. Clark, and A. Tellegen, "Development and validation of brief measures of positive and negative affect: the panas scales." *J Personality & Social Psychology*, vol. 54, no. 6, p. 1063, 1988.

[29] J. T. Hancock, K. Gee, K. Ciaccio, and J. M.-H. Lin, "I'm sad you're sad: emotional contagion in cmc," in *ACM Conf on Computer Supported Cooperative Work (CSCW)*, 2008, pp. 295–298.

[30] J. R. Crawford and J. D. Henry, "The positive and negative affect schedule (panas): Construct validity, measurement properties and normative data in a large non-clinical sample," *British Journal of Clinical Psychology*, vol. 43, no. 3, pp. 245–265, 2004.

[31] J. Sabourin, B. Mott, and J. C. Lester, "Modeling learner affect with theoretically grounded dynamic bayesian networks," in *Affective Computing & Intelligent Interaction*. Springer, 2011, pp. 286–295.

[32] M. J. Hertenstein, D. Keltner, B. App, B. A. Bulleit, and A. R. Jaskolka, "Touch communicates distinct emotions." *Emotion*, vol. 6, no. 3, p. 528, 2006.

[33] M. Hoque and R. W. Picard, "Acted vs. natural frustration and delight: Many people smile in natural frustration." in *IEEE Face & Gesture*, 2011, pp. 354–359.

[34] T. Partala and V. Surakka, "Pupil size variation as an indication of affective processing," *Int'l J Human-Computer Studies*, vol. 59, no. 1-2, pp. 185–198, Jul. 2003.

[35] E. H. Hess and S. B. Petrovich, "Pupillary behavior in communication," in *Nonverbal Behavior and Communication*, A. W. Siegman and S. Feldstein, Eds. Erlbaum, Hillsdale, NJ, 1972, pp. 327–348.

[36] L. L. Di Stasi, A. Catena, J. J. Canas, S. L. Macknik, and S. Martinez-Conde, "Saccadic velocity as an arousal index in naturalistic tasks," *Neuroscience & Biobehavioral Reviews*, vol. 37, no. 5, pp. 968–975, 2013.

[37] M. Nardelli, G. Valenza, A. Greco, A. Lanata, and E. P. Scilingo, "Recognizing emotions induced by affective sounds through heart rate variability," *IEEE Trans Affective Computing*, vol. 6, no. 4, pp. 385–394, 2015.

[38] R. Hill, "Perceptual attention in virtual humans: Toward realistic and believable gaze behaviors," in *AAAI Fall Symposium on Simulating Human Agents*, 2000, pp. 46–52.

[39] B. M. Appelhans and L. J. Luecken, "Heart rate variability and pain: associations of two interrelated homeostatic processes." *Biological Psychology*, vol. 77, no. 2, pp. 174–182, 2008.

[40] C. M. Jones and T. Troen, "Biometric valence and arousal recognition," in *Australasian Conf on Computer-Human Interaction: Entertaining User Interfaces*, 2007, pp. 191–194.

[41] P. Bucci, L. Zhang, X. L. Cang, and K. E. MacLean, "Is it happy? behavioural and narrative frame complexity impact perceptions of a simple furry robot's emotions," in *ACM CHI Conf on Human Factors in Computing Systems*, 2018, pp. 1–11.

[42] M. M. Jung, R. Poppe, M. Poel, and D. K. Heylen, "Touching the void–introducing cost: corpus of social touch," in *ACM Int'l Conf on Multimodal Interaction (ICMI)*, 2014, pp. 120–127.

[43] D. Silvera-Tawil, D. Rye, and M. Velonaki, "Interpretation of social touch on an artificial arm covered with an eit-based sensitive skin," *Int'l J Social Robotics*, vol. 6, no. 4, pp. 489–505, 2014.

[44] P. Bucci, X. L. Cang, A. Valair, D. Marino, L. Tseng, M. Jung, J. Rantala, O. S. Schneider, and K. E. MacLean, "Sketching cuddlebits: coupled prototyping of body and behaviour for an affective robot pet," in *CHI Conf on Human Factors in Computing Systems*, 2017, pp. 3681–3692.

[45] D. D. Salvucci and J. H. Goldberg, "Identifying fixations and saccades in eye-tracking protocols," in *Symp on Eye Tracking Research & Applications*. ACM Press, 2000, pp. 71–78.

[46] K. B. Shimoga, "Finger force and touch feedback issues in dexterous telemanipulation," in *IEEE Intelligent Robotic Systems for Space Exploration*, 1992, pp. 159–178.

[47] M. A. Otaduy and M. C. Lin, "High fidelity haptic rendering," *Synthesis Lectures on Computer Graphics and Animation*, vol. 1, no. 1, pp. 1–112, 2006.

[48] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[49] M. J. Hertenstein, R. Holmes, M. McCullough, and D. Keltner, "The communication of emotion via touch." *Emotion*, vol. 9, no. 4, p. 566, 2009.

[50] M. M. Jung, M. Poel, R. Poppe, and D. K. Heylen, "Automatic recognition of touch gestures in the corpus of social touch," *J on multimodal user interfaces*, vol. 11, no. 1, pp. 81–96, 2017.

[51] I. H. Witten, "Data mining with WEKA," *UWaikato, New Zealand*, 2013.

[52] A. L. Nichols and J. K. Maner, "The good-subject effect: Investigating participant demand characteristics," *The Journal of general psychology*, vol. 135, no. 2, pp. 151–166, 2008.

[53] R. A. Calvo, S. D'Mello, J. Gratch, and A. Kappas, *The Oxford Handbook of Affective Computing*. Oxford University Press, 2014.

[54] J. J. Gross, "The emerging field of emotion regulation: an integrative review." *Review of General Psychology*, vol. 2, no. 3, p. 271, 1998.

[55] H. Sharp, *Interaction Design*. John Wiley & Sons, 2003.

Xi Laura Cang is a PhD candidate in Computer Science at UBC with a background in Mathematics Education (BSc / BEd 2010) and Computer Science (MSc 2016). She is interested in affective computing for therapeutic HRI and embedded machine learning systems for affect classification in multimodal interactions.

Paul Bucci is a PhD student in Computer Science at UBC (M.Sc in Computer Science (2017), B.Sc. in Computer Science (2015), B.A. in Visual Art (2012)). His research interests concern interactive affective systems and human-centred design.

Dr. Jussi Rantala is a Postdoctoral Researcher who received his PhD in Interactive Technology (2014) and MSc in Computer Science (2007) from the University of Tampere. His research interests include haptics, mobile and wearable devices, and gaze tracking.

Dr. Karon E. MacLean is a Professor in Computer Science at UBC (B.Sc. in Biology and Mechanical Engineering from Stanford (1986); Ph.D. in Mechanical Engineering from MIT (1996), with industry experience in robotics and interaction design. Her research interests are in situated haptic and multimodal interfaces, and affective, therapeutic human-robotic interaction.