Feeling (key)Pressed: Comparing The Ways in Which Force and Self-Reports Reveal Emotion

by

Rúbia Reis Guerra

B.Sc., Federal Univeristy of Minas Gerais, 2019

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

Master of Science

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES
(Computer Science)

The University of British Columbia
(Vancouver)

October 2022

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the thesis entitled:

Feeling (key)Pressed: Comparing The Ways in Which Force and Self-Reports Reveal Emotion

submitted by **Rúbia Reis Guerra** in partial fulfillment of the requirements for the degree of **Master of Science** in **Computer Science**.

Examining Committee:

Karon E MacLean, Professor, Department of Computer Science, UBC
Supervisor

Robert Xiao, Assistant Professor, Department of Computer Science, UBC
Supervisory Committee Member

Abstract

Interactive human-computer systems can be enriched to interpret and respond to users' affective states using computational emotion models, which necessitates the collection of authentic and spontaneous emotion data. Popular emotion modelling frameworks rely on convenient, yet static abstractions of emotion (e.g., Ekman's basic emotions and Russell's circumplex). These abstractions often oversimplify complex emotional experiences into single emotion categories. In turn, emotion models guided by such emotion annotations leave out significant aspects of the user's true, spontaneous emotional experience.

Richer representations of emotion, negotiated and understood between participants and researchers, can be created using mixed-methods labelling–assigning an emotion descriptor to a recorded segment of experience–approaches. However, resulting emotion annotations are often not ready-to-use in computational models. In this thesis, we investigate (1) ways to improve meaningfulness of self-reported emotion annotations, and (2) to understand the implicit expression of emotion in touch pressure. For the first, we propose three strategies to interpret multiple versions of self-annotated dynamic emotion through combining (multi-label classification), extracting (of alignment metrics), and resolving (of conflicts between) emotion labels.

We evaluate our label-resolution strategies using the FSR EEG Emotion-Labelled (FEEL) dataset (N=16). The FEEL dataset includes brain activity and keypress force data captured from a 10-minute video of user gameplay experience, annotated with two methods of self-reporting emotion–a continuous annotation and an interview. By featuring multi-pass self-report and user-calibrated scales, the data collection protocol prioritized the capture of genuine emotion evolution. We triangulate multiple self-annotated emotion reports and evaluate classification accuracy of our three proposed label resolution strategies. For our second research question, we compare models built on keypress force and brain activity data in an effort to understand the implicit expression of emotion in touch pressure. Finally, we reflect on the trade-offs of each strategy for developing computational models of emotion. Our findings suggest that touch-based models outperform those built on brain activity, and mixed-methods emotion annotations increase self-report meaningfulness.

Lay Summary

To investigate how touch behaviour reflects our emotion expressions, we used a dataset of 16 people playing a horror video game, capturing their touch pressure and brain activity. We associated (or “labelled”) their touch pressure with their descriptive emotion language from an interview and their joystick drawings of their emotion trajectory while they watched their gameplay recording. We evaluated three strategies for training machine learning models to recognize emotion: (1) predict emotion trajectory after recognizing an emotion word; (2) predict whether or not participants were feeling a given emotion (for all provided emotion words); and (3) predict only after being trained on experiences we were confident represented a specific emotion. Overall, we find that touch pressure is better at predicting emotion trajectory than brain activity and that emotion words are more meaningful than emotion trajectory: knowing the context of an emotion leads to better predictions of emotion trajectory.

Preface

All of the work presented henceforth was conducted in the Sensory Perception and Interaction Laboratory (SPIN) at the University of British Columbia, Vancouver campus. All projects and associated methods were approved by the University of British Columbia's Research Ethics Board [certificate #H15-02611: Interactive Affective Touch].

I used the pronoun "we" in this thesis instead of "I", and rely on this preface to indicate the scope of my own and others contributions. The "we" reflects the collaborative environment in which this research was developed, consisting of myself, my dear graduate and undergraduate colleagues, and my supervisor Dr. Karon MacLean. Other researchers will have different critical reflexive practices than myself. As the writer of this thesis, I will not speak on their behalf, but recognize that the result of much of this work has been a culmination of our collective positionalities, influencing the decisions and reporting style discussed in this thesis.

All the analyses in this thesis are performed with the FSR EEG Emotion-Labelled (FEEL) dataset, originally collected in 2018 by my fellow graduate students Laura Cang and Paul Bucci, and, then undergraduate researchers at SPIN, Laura Rodgers, Hailey Mah, Qianqian Feng, and Anushka Agrawal. Chapter 3 features a detailed description of the study protocol behind FEEL, as well as preliminary results obtained by Laura Cang (PhD candidate) and Bereket Guta (at the time, undergraduate researcher at SPIN).

Chapter 4 (Exploring Multi-Pass Emotion Self-Reports) presents analyses that were done in collaboration with graduate and undergraduate researchers at SPIN. I am responsible for most of the analysis reported in this chapter, with exception of Section 4.2 (Commonality in Interpreting Emotion Words), performed by Chuxuan

Zhang, and 4.4 (Comparing Motion Characteristics of Emotion Dynamics), performed in conjunction by Bereket Guta, visiting graduate researcher Shinmin Hsu, and myself. I carried out results interpretation, reporting, and final editing in conjunction with Laura, Bereket, and Karon.

I led investigation on conceptual development, data preparation, analyses and reporting pertaining to the content discussed in Chapter 5 and 6. Laura, Bereket, and Karon were involved in the early stages of concept formation and contributed to manuscript edits.

A version of Chapters 3, 4 and parts of 7 has been published in:

> Cang, X. L., Guerra, R. R., Bucci, P., Guta, B., Rodgers, L., Mah, H., Hsu, S., Feng, Q., Zhang, C., Agrawal, A., MacLean, K. E. (2022, October 18th). Choose or Fuse: Enriching Data Views with Multi-Label Emotion Dynamics. IEEE 10th Int'l Conf on Affective Computing & Intelligent Interaction (ACII).

Laura Cang was involved in concept formation, results interpretation, and lead manuscript composition. I was the second author, responsible for most of the analysis (presented in Chapter 4 of this thesis), as well as manuscript composition. Paul Bucci contributed significantly in concept formation. Bereket Guta, Shinmin Hsu, and Chuxuan Zhang contributed with analysis and manuscript edits. Dr. Karon MacLean was the supervisory author and was involved throughout the project in concept formation and manuscript composition.

Last, I take responsibility for the concept of this thesis, chapter integration, and all other formal writing.

Contents

Abstract . . . . **iii**

Lay Summary . . . . **v**

Preface . . . . **vi**

Contents . . . . **viii**

List of Tables . . . . **xii**

List of Figures . . . . **xiii**

Glossary . . . . **xvii**

Acknowledgments . . . . **xx**

Dedication . . . . **xxi**

1 Introduction **1**

1.1 Understanding and Interpreting Emotion Language 2

1.2 Is Reading Emotion from Affective Touch Possible? 4

1.3 Research Questions and Approach 5

1.4 Guiding Perspectives 8

1.4.1 Data Analysis Framework 8

1.4.2 Emotion Analysis Guideline 9

1.5 Contributions 10

1.6 Researcher's Position and Ethics Considerations 11
1.7 Organization and Audience 12
1.7.1 Audience 13

2 Related Work 14
2.1 Emotion Data 15
2.1.1 Emotion Elicitation 15
2.1.2 Self-Report Methods 16
2.1.3 Emotion Metaphors 16
2.2 Learning Machines 18
2.2.1 Types of Learning 18
2.2.2 Learning with Multiple Labels 20
2.3 Machine Classification on Implicit Emotion Expression Modalities 21
2.4 Emotion Modelling with Multiple Reporting Passes 23

3 FEEL: Force EEG and Emotion-Labelled Dataset 24
3.1 Collection Protocol 25
3.1.1 Participants 25
3.1.2 Data Capture and Preparation 25
3.2 Preliminary Results Using Brain Activity with Continuous Annotation 29
3.2.1 Time Domain and LSTM Classifier 30
3.2.2 Frequency Domain and Convolutional Neural Network Classifier 31
3.2.3 Classification Performance 32
3.3 Summary 32

4 Exploring Multi-Pass Emotion Self-Reports 34
4.1 Label Metaphors 34
4.2 Commonality in Interpreting **Emotion Words** 35
4.3 Self-Report Modality Consistency via **Time Series** 37
4.3.1 Test Preparation 38
4.3.2 Pearson's Correlation for Signal Similarity 38
4.3.3 Granger Causality Test for Source Plausibility 38

4.4 Comparing Motion Characteristics of Emotion Dynamics 40
4.4.1 Data Preparation 40
4.4.2 Comparing Summary Statistics and Histograms by Parameter 40
4.5 Summary 41

5 Connecting Keypress Force to Multi-Pass Emotion Self-Reports . . 42
5.1 Feature Extraction 42
5.1.1 Data Preparation 43
5.1.2 Windowing 43
5.1.3 Frequency and Statistical Features 44
5.1.4 Keystroke Features 44
5.2 Understanding Affect in FSR Data 46
5.2.1 Labels Sourced from Continuous Annotation 47
5.2.2 Classification Pipeline 47
5.2.3 Classification Performance of Models Built on FSR Data . 49
5.2.4 Feature Evaluation 50
5.2.5 Comparing Models Built on FSR, EEG, and Both 51
5.3 Summary 54

6 Which Self-Report Pass is Best for Computational Emotion Modelling? 56
6.1 Using Only Timeline with Calibrated Words 57
6.2 Combining TwCW and CA 58
6.2.1 Hierarchical Multi-label Classification 58
6.2.2 Extracting Label Alignment 59
6.3 Comparing Performance for Emotion-as-Position Strategies 61
6.4 Comparing Performance for Emotion-as-Direction Strategies . . . 63
6.5 Feature Evaluation 65
6.6 Summary 66

7 Reflections and Recommendations 68
7.1 Re-Establishing Grounding Ideas 68
7.2 Real-Time Predictors of Dynamic Emotion: Keypress Force Outperforms Brain Activity 69

7.3 Personalized Labelling: Trade-Offs 70

7.3.1 How Do Triangulation and Personalization Enrich Emotion Self-Reports? 71

7.3.2 Incorporating Dynamics Into Computational Emotion Models 72

7.4 Considerations in Building Effective Models for Dynamic Emotion Prediction 73

8 Conclusions and Future Work 76

8.1 Conclusions 77

8.2 Future Directions 78

Bibliography 80

A Data Collection Materials 92

A.1 Study Recruitment Poster 93

A.2 Study Consent Form 94

List of Tables

Table 5.1 Aggregated count of data instances by window size. 43

Table 5.2 Percentage of occurring keypress distortion in relation to all Force Sensing Resistor (FSR) data instances, aggregated by window size. 45

Table 5.3 Summary of features extracted from FSR data. 45

Table 5.4 Estimators and parameters used for tuning keypress force models to each participant. 49

Table 6.1 Analysis of Variance on Aligned Rank Data (Type III) on emotion-as-position results. 61

Table 6.2 Analysis of Variance on emotion-as-direction results. 63

List of Figures

Figure 1.1 Overview of the stages of data analysis performed in this thesis, highlighting where each guiding perspective influences our work. 7

Figure 1.2 Cross-Industry Standard Process for Data Mining (CRISP-DM) process cycle applied to the Force EEG and Emotion-Labelled (FEEL) dataset. 9

Figure 2.1 Illustration of single label classification tasks. At the right, we exemplify binary classification–pick one of two categories– and at the bottom, multi-class classification–pick one of many categories). 19

Figure 2.2 Illustration of multi-label learning tasks. At the top right, an example of flat multi-label classification, in which each label is considered an independent category. At the bottom, an example of hierarchical multi-label classification, where decisions on certain label categories can influence each other. . . 20

Figure 3.1 Overview of the Force EEG and Emotion-Labelled dataset collection protocol. 24

Figure 3.2 Illustrative example of contextualized emotion word calibration. 28

Figure 3.3 A sample of the data streams collected during the FEEL study: Electroencephalogram (EEG) from brain activity (top) and the FSR from keypress force collected during gameplay, superimposed with self-reported Continuous Annotation and Timeline of Calibrated words (bottom). 29

Figure 3.4 Overview of classification pipeline for EEG data. 30

Figure 3.5 The AutoEncoder (a), Long Short-Term Memory (LSTM) (b) and Convolutional Neural Networks (CNN) (c) models used for brain activity classification. The AutoEncoder is trained in order to encode the 64 channels at every data point in a given window down to 12 variables. The LSTM uses the encoded features from the AutoEncoder to generate a 3 class output. The CNN uses the 5 bands $\times$ 64 channels $\times$ 64 channels frequency features as an input to generate a 3 class output. 31

Figure 3.6 F1-scores for EEG models by window size. 32

Figure 4.1 Distribution of class labels for each label metaphor and window size. 36

Figure 4.2 Rating variance by calibration word, ordered by number of participants who provided a rating for that word. 37

Figure 4.3 Comparison of summary statistics and histograms by emotion parameter. 39

Figure 5.1 Attack, Decay, Sustain, and Release (ADSR) envelope. 44

Figure 5.2 Example of keystroke force as captured by a key-mounted FSR sensor. W_s and W_e indicate the start and end of a window (dashed boundary). P represents the maximum peak force during a single keystroke (delineated by zero force), and R represents the release point. Distortion occurs when keystrokes span across multiple windows, creating spurious peak and release points (P' and R', P'' and R''), and distorted representations of keypress activity in a window. 46

Figure 5.3 Analysis factors for experiments on continuously annotated emotion experience estimated by dynamic emotion metaphors. We inspect models built on two affective modalities: brain activity (as estimated by EEG) and keypress force (FSR), using two emotion label metaphors of direction (best fit line) and position (window mean), and trained on four window sizes (0.5s, 1s, 2s, 5s). 48

Figure 5.4 Overview of the experiment setup used to evaluate emotion models built on FSR data. 49

Figure 5.5 Relative feature selection count per window size and label type (D: direction, P: position, and A: accumulator). Darker tone indicates most commonly selected features in personalized models that perform above-chance. 51

Figure 5.6 Comparison of FSR, EEG, and multi-modal classification results on Continuous Annotation (CA) by window and label type. 53

Figure 6.1 Overview of the modified experiment setup, first presented in Chapter 5, used in all forecasting schemes using Timeline with Calibrated Words (TWCW) or a combination of TWCW and CA. 57

Figure 6.2 Example of timeline with calibrated words. 57

Figure 6.3 Illustration of flat multi-label classification output with calibrated words (outputting, for instance, the labels “happy” and “stressed”). 58

Figure 6.4 Example of hierarchical multi-label classification with calibrated words and continuous annotation (represented by direction metaphor). 59

Figure 6.5 Illustration of $d(CA,CW)$. 60

Figure 6.6 Comparison of single- and multi-pass classification results on emotion-as-position using CA and TWCW, reported by window and label type. Single-pass CA scores correspond to the first the top two groups of EEG, FSR and fusion results, followed by single-pass TWCW implemented as single- and multi-label models. Finally, the last row indicates models built using information extracted from both self-reports. 62

Figure 6.7 Comparison of single- and multi-pass classification results on emotion-as-direction using CA, TWCW, and a combination of both, reported by window and label type. Single-pass CA scores correspond to the first the group of EEG, FSR and fusion results, followed by combined self-reports models, implemented as hierarchical multi-label tasks. The last row indicates models built using the instance weighting strategy. 64

Figure 6.8 Relative feature selection count by window size in models using TWCW (single-label strategy). Darker tone indicates most commonly selected features in personalized models that perform above-chance. 66

Figure 6.9 Overview of single- and multi-pass classification results on emotion-as-position and emotion-as-direction using CA and TWCW, reported by affective modality, window and label type. 67

Glossary

ADF Augmented Dickey-Fuller Test, used to test whether a given time series is stationary

ADSR Attack, Decay, Sustain, and Release envelope, in music theory, describes how a sound changes over time

ANOVA Analysis of Variance, a set of statistical techniques to identify sources of variability between groups

ART ANOVA Aligned Rank Transform Analysis of Variance, a non-parametric approach to factorial ANOVA that enables analyses of interaction and main effects

CA Continuous Annotation

CNN Convolutional Neural Networks, class of machine learning algorithms, often used in image recognition and classification

CRISP-DM Cross-Industry Standard Process for Data Mining, a process model for data science projects

CV Cross-Validation, re-sampling method that uses different portions of the data to evaluate and build a model on different iterations

DE Differential Entropy, measures the randomness of a random variable and the number of bits required to describe it

ED Emotion Dynamics, describes a set of trajectories, patterns, and regularities with which emotions fluctuate across time

EEG Electroencephalogram, a test that measures electrical activity in the brain

ERP Event-Related Potentials, small voltages generated in the brain structures in response to specific events or stimuli

EXPLICIT EMOTIONAL EXPRESSION Expression of emotion that requires conscious effort to be initiated and completed, some monitoring during implementation, and a certain degree of awareness

EWMA Exponentially Weighted Moving Average, smooths a series of data based on a moving average with weights which decay exponentially

FEEL Force EEG and Emotion-Labelled Dataset

FFT Fast Fourier Transform, an algorithm to convert a signal into individual spectral components and provide frequency information about the signal

FSR Force Sensing Resistor, a material whose resistance changes when a force, pressure or mechanical stress is applied

HCI Human Computer Interaction, a multidisciplinary field of study focusing on the design of computer technology and, in particular, the interaction between humans and computers

ICC Intra-Class Correlation, descriptive statistic to measure the reliability of ratings or measurements among groups

IMPLICIT EMOTIONAL EXPRESSION Expression of emotion that occurs automatically, can happen without insight or awareness, and do not need monitoring in order to be completed

IWCW Interview with Calibrated Words

KFP Keypress Force, amount of pressure applied to press down a key, estimated by force sensing resistors

LABELLING The process of tagging or annotating data with representative labels so that a machine learning model can learn from it

LSE Least Squares Error, a function that describes the smallest sum of distances between data points and a fitted regression line

LSR Least Squares Regression, statistical procedure to find the best fit curve for a set of data points by minimizing the sum of the offsets or residuals of points from the curve

LSTM Long Short-Term Memory Networks, class of machine learning algorithms, often used in tasks with sequential data or temporal dependencies

ML Machine Learning, field of study that focuses on methods that leverage data to improve performance on some set of tasks

PANAS Positive and Negative Affect Schedule, 20-item self-report measure to assess positive and negative affect

RFE Recursive Feature Elimination, feature selection method that fits a model and removes the worst-performing features

SAM Self-Assessment Manikin, non-verbal pictorial self-report technique that measures pleasure, arousal, and dominance associated with a person's affective state

SVM Support Vector Machines, a set of machine learning methods used for classification, regression, and outlier detection

TMMS Trait Meta Mood Scale, self-report assessment of a person's emotional intelligence skill level

TWCW Timeline with Calibrated Words

Acknowledgments

I have been fortunate to be surrounded by many awesome people in the past two years.

I am deeply grateful to Prof. Karon MacLean, for her kind, extensive guidance. Thank you for believing in me and pushing me to grow.

To my second reader, Prof. Robert Xiao, for insightful comments and questions. This work is better after your input.

To Prof. Dongwook Yoon and Prof. Joanna McGrenere, for supporting and guiding me through my first steps at UBC. I would not have been able to start my program without you.

To Laura Cang, for your thoughtful mentoring and kind friendship. Thank you for all the hours spent on video calls with me. I cherish our friendship deeply.

To Hannah Elbaggari, for being such an awesome desk mate and friend. I will miss our workdays together.

To my SPIN and MUX friends, for your insightful feedback and for figuring out grad school with me.

To my CSGSA, SPL, and Systopia friends, for making grad school through a pandemic less lonely.

To my family and friends, for supporting me every step of the way. A special thanks to my friends Paulette, Joseph, Shaurya, Nichole, Unma, Preeti, Amir, Raquel, Traci, Divya, and Matilda, for their companionship, for making my days in Vancouver much more interesting, and for taking me in when I needed a place to stay.

And, finally, to Hương, Sebastian, and Churro, for all the love and for being my closest friends throughout my graduate journey.

Dedication

To my family, chosen or not.

Chapter 1

Introduction

Have you ever felt frustrated that your emotions aren't being understood by others? We all know the feeling of being so gutted you can't speak; so ecstatic that you can only jump for joy. When your emotions get the better of your ability to communicate your feelings, it can take time for other people to interpret and understand what you are feeling and why you are feeling a certain way.

Being able to intentionally communicate emotions as we feel them, verbally or otherwise, is not always feasible. To intentionally communicate feelings, there exists the process of "*I feel $\rightarrow$ I recognize my emotions $\rightarrow$ I communicate them $\rightarrow$ (I am understood)*." This emotion communication is vulnerable to a myriad of different factors: having the language to explain *what* you feel, having the cognitive capacity to *process* your feelings, having enough control over your body to *communicate* your emotional state. This is true whenever there is a communication breakdown, for instance, between adults and young children who are still in the process of acquiring language, or due to difficulty in identifying feelings in people who experience dementia [67]. More broadly, and independently of a person's physical or cognitive capacities, intentional communication becomes challenging when we are at the height of powerful emotions, pushing us to rely on spontaneous emoting.

Emotion-aware technologies are also affected by these challenges. Building such technologies requires an understanding of how people express emotions. In operationalizing this understanding, we rely on self-report of genuine emotions in controlled, although real, situations. However, as noted above, self-reporting itself is a delicate process. Conceptualizing how to build and use trustworthy self-reports that yield adequate models is an open area of research [13].

Conversely, effective emotion-aware technologies should automatically respond to a person's affective state, captured implicitly through different interaction channels. Selecting application-appropriate interaction channels through which emotion expression can reliably be identified is also an open area of research [97], and often depends on self-reports to re-construct a ground-truth of emotion experience.

In this thesis, we explore a novel emotion dataset containing triangulated emotion self-reports. We are interested in exploring (1) ways to improve meaningfulness of emotion self-reports; and (2) the understanding of the physicality of emotion expression in the context of touch pressure. Our overarching goal is to contribute fundamental knowledge which will enable unobtrusive touch-mediated affective technologies that can dynamically and effectively process and respond to ever-changing, human emotions.

1.1 Understanding and Interpreting Emotion Language

To create interactive systems capable of interpreting and responding to users' emotions, we first need to understand how to operationalize subjective emotion experiences as computational models. Whether building robots that detect anxiety through touch interaction or video games that dynamically adjust level difficulty to optimize player engagement, challenges arise in developing these computational models from true and spontaneously evolving emotions.

Emotion theorists have long observed time-varying dynamics of emotion expression, attributing them to complex neurological and physiological regulation mechanisms [33], appraisal effects [72], cognition and contextual factors [68, 76]. To simplify in-lab research, computational emotion modelling often relies on an "emotions-as-point" metaphor [12, 56], represented as a dimensionless point in an emotion plane in which self-reporting static emotion labels for classification involves easy-to-read scales, often along dimensions of arousal, valence, and dominance [15]. While these models are convenient, for real-time use we need to recognize emotion evolution over time, rather than distilling a lengthy event into a single label.

Operationalizing authenticity in emotion data is a significant obstacle. Our memories and emotional assessments are affected by time and reflection [72, 76]; how representative of someone's "reality" can a reporting scheme be? Commonly used labels on the arousal-valence circumplex model [84], Positive and Negative Affect Schedule (PANAS) [105], or Self-Assessment Manikin (SAM) [11] (among others) quickly become intractable for sampling at the rates at which emotion can potentially evolve (ranging from a few seconds to several hours [100]).

Challenges to building emotionally reactive machines exist both in determining the appropriate measurement instrument to detect relevant emotion markers and the right timing to respond to that marker.

Independent of the measurement instrument, self-report of emotion incites questions of generalizability across the population. A researcher's understanding of the instrument scale may be very different from that of a participant [12]; our comprehension of an emotional 'landscape' or internalized emotion frames of reference are highly subjective, influenced by life experiences and personal history [10]. We presume that any set of ground-truth labels for self-reported emotion are similarly personalized: i.e., the experience or scale for *anger* for one person may not be recognizable for another.

For emotionally reactive machines, finding the 'right time' for the machine to act requires that our machines understand the transitional nature or direction of their inherently emotional human patron (users). In particular, machines may need to respond differently to emotions as they increase or decrease in some identifiable parameter, such as intensity or polarity. To forecast the direction of emotion experience, we can predict transition directly or indirectly – by predicting position and then calculating transition. Evaluating emotion based on dynamic qualities will advance the accuracy of machine recognition of human emotion experiences. Better forecasting of a user's near-future emotional expression allows for system responses that are temporally and situationally appropriate.

1.2 Is Reading Emotion from Affective Touch Possible?

Studies exploring the bi-directional connection between emotion and touch have been prominent in neuroscience [36, 62, 71] and social sciences [29, 53], suggesting that touch can both communicate and influence a person's emotional state. Kinnunen and Kolehmainen [53] argues that touch is a vital part of affective histories after analyzing 68 touch biographies, in which authors narrate their lives through the ways in which they have touched, been touched, experienced touch and been socialized to touch.

The uptick in research of touch-based affective technologies is far newer, however, having particularly accelerated during the past decade [25]. Aside from limitations in sensing technology, the disparity of progress among fields can be partly attributed to the difficulties in capturing genuine, non-scripted touch in scenarios characterized by authentic emotion expression. Touch pressure (defined broadly) has been shown to encode implicit emotional content that can be interpreted by people [40]. Machines can distinguish identity [14, 30] and many gradations of social touch gestures that we tend to consider emotionally expressive [4, 15, 50].

Although the biological mechanisms through which emotion modulates touch are still unclear [38, 40], touch is a promising modality for inferring–and potentially altering–emotion experiences [59]. Touch is a concrete, perceivable and expressive act [53]. When compared to signals commonly used in emotion research–Electroencephalogram (EEG), brain imaging, heart-rate, facial configurations, body posture, speech [92, 97]–touch is less intrusive, and gives the participant more agency over what data they consent to have recorded–not feasible when recording biological signals, for example.

Studies of the recipient of affective touch demonstrates that 'pleasant' touches activates C-fibre tactile afferents[1]in the skin which trigger pleasure centres in the brain and has evolutionary impact on how we form attachments to one another – pleasant physical touch is crucial in forming prosocial behaviours and relationships for humans throughout our lifetimes [8, 38, 66]. In contrast, investigation into the performer of affective touch has largely been behaviour observation-based where researchers record and evaluate the behaviour while the 'toucher' is expressing a self-reported emotional experience [15] or acting in an emotional context [111].

While it is not entirely clear if emotion "leaks" into non-affective touch, studies into typing behaviour on pressure sensitive keyboards demonstrate that stressed typing is linked to harder key strikes resulting in higher keypress forces and shorter keypress duration (and incidentally, a higher error rate) [37, 64, 103]. Now we ask, could people exhibit similar *implicit* emotional touch where emotion expression is not the primary purpose of the interaction?

1.3 Research Questions and Approach

In order to model spontaneously evolving emotions, we analyze the data from the Force EEG and Emotion-Labelled (FEEL) dataset, previously collected using the emotion annotation (also referred to as "emotion labelling") protocol described in Cang et al. [16]. The data includes 64-channel brain activity and 5-key keypress force data as well as two emotion labelling passes[2], self-reported at high and low densities, collected from 16 participants. Both input modalities–keypress force and brain activity–have been shown to encode emotion [3, 37, 64] and are reasonable to collect during video-game play. All data is time-aligned to an emotion task, playing Playdead's Inside [79] – a horror video game featuring chase and puzzle scenes, navigated serially and thus amenable to temporal alignment of game-play.

From FEEL's multi-pass self-reports, we construct two emotion-labeled time series – Continuous Annotation (CA) and the Timeline with Calibrated Words (TWCW). Diverse self-reports may capture perspectives that are authentic in different ways. There may be conflicts that arise between labelsets generated on the same timeline, the resolution of which could improve classification performance and allow for flexible emotion outputs, which can be personalized based on the application scope.

1. How do we improve meaningfulness of emotion self-reports, capturing subjective and dynamic emotion in computational emotion models?

Nuances in users' emotion language, manifesting as apparent inconsistency in emotion self-reports, can interfere with emotion model performance and validity [49]. We ask: **[RQ1a]** *does user-centring of emotion self-reports add new*

[1] C-fibre tactile afferents are nerve receptors in human skin that generally respond to non-painful stimulation such as light touch [66].

[2] Also referred to as "multi-pass".

information? For example, do people rank common emotion words similarly? In what ways does labelling data differ by pass? What do we gain from quantifying the differences? When designing computational models, **[RQ1b]** *what are effective ways of incorporating triangulated labels for modelling dynamic emotion?*

Our present focus is to demonstrate and evaluate Machine Learning (ML) models of dynamic emotion in terms of the trajectory or direction of emotional movement and compare to those of momentary state. We investigate use of triangulated emotion self-reports collected using mixed methods which play out at different timescales (continuous annotation and participant interview), and consider the benefits and drawbacks of model parameters in terms of label definitions and window size.

2. What does touch pressure tell us about emotions?

User data collected during operation of an emotionally responsive device can be used to track a person's dynamic emotion expression. We are interested in exploring implicit emotion encoded in pressure data from keystrokes in the context of gameplay.

We investigate touch pressure as a non-intrusive modality through which emotion expression can be captured, and, in the future, leveraged to create emotion-responsive technologies. We further refine our initial research question: **[RQ2]** *What does keypress force tell us about emotions in a tense gameplay?* In a video game play scenario, we compare keypress force–a readily available modality captured by game controllers[3]–with brain activity at a much higher temporal and spatial resolution (over the head rather than just the fingers). EEG for emotion state classification is well-studied using Event-Related Potentials (ERP) as snapshots of expression [61, 74, 104]. Keypress force (and touch pressure more generally) has been shown to encode emotion during purposed tasks like typing and other gesture behaviours [15, 37, 64]. We wonder how touch performs, in comparison to more intrusive methods, over an extended time-series for emotion prediction and what key contributions this modality offers.

[3] In this case, arrow keys and the action key "ALT".

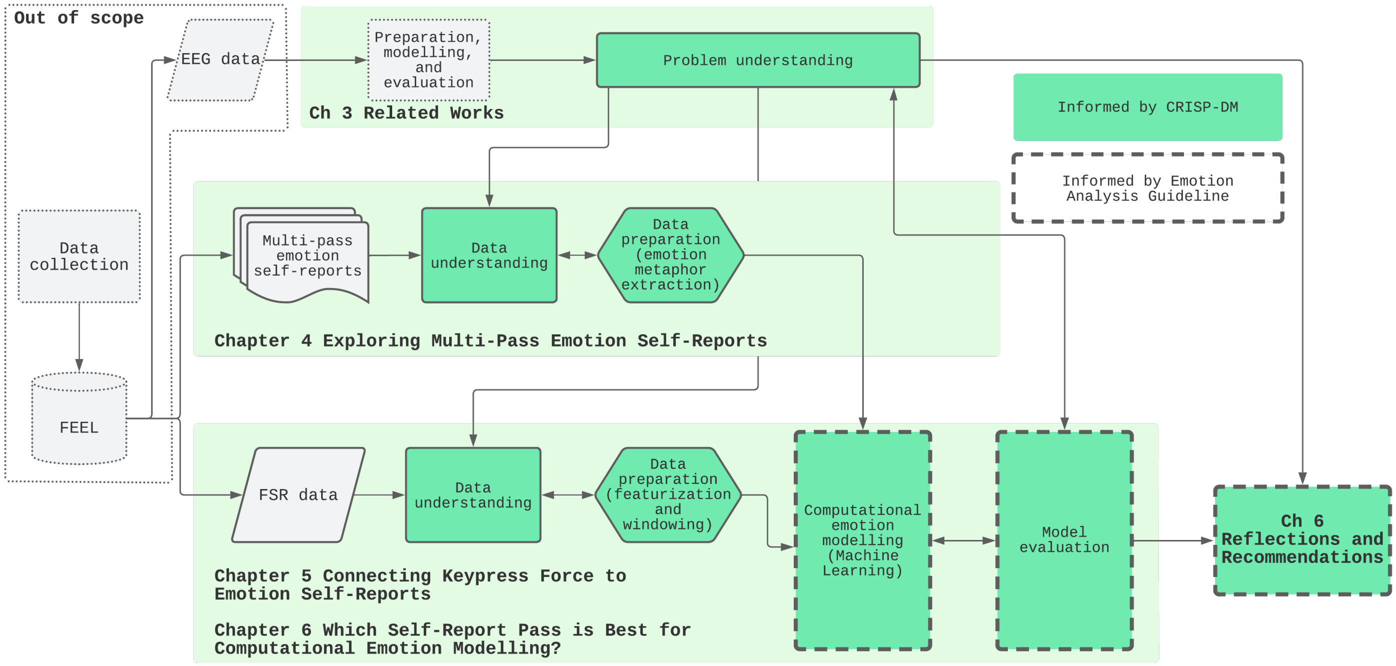

Out of scope
EEG data
Preparation, modelling, and evaluation
Problem understanding
Ch 3 Related Works
Informed by CRISP-DM
Informed by Emotion Analysis Guideline
Data collection
Multi-pass emotion self-reports
Data understanding
Data preparation (emotion metaphor extraction)
Chapter 4 Exploring Multi-Pass Emotion Self-Reports
FEEL
FSR data
Data understanding
Data preparation (featurization and windowing)
Computational emotion modelling (Machine Learning)
Model evaluation
Ch 6 Reflections and Recommendations
Chapter 5 Connecting Keypress Force to Emotion Self-Reports
Chapter 6 Which Self-Report Pass is Best for Computational Emotion Modelling?

Figure 1.1: Overview of the stages of data analysis performed in this thesis, highlighting where each guiding perspective influences our work.

1.4 Guiding Perspectives

Given the exploratory nature of our focus, we guide our analysis efforts with methodologies informed by literature in the field of data mining (Chapter 1.4.1) and social sciences (Chapter 1.4.2). Figure 1.1 presents an overview of the high-level components of this thesis, highlighting how the frameworks presented in this section influence our work in subsequent chapters.

1.4.1 Data Analysis Framework

Data mining involves exploring and analyzing large blocks of information to glean meaningful patterns and relationships. For projects involving exploration and understanding of complex data relationships–such as between Force Sensing Resistor (FSR) and emotion self-reports–the Cross-Industry Standard Process for Data Mining (CRISP-DM) [106] proves to be a useful process model. The CRISP-DM offers an adaptable framework that focuses on problem understanding, while providing a uniform set of processes relevant to a data analysis project. In this thesis, we use the CRISP-DM to guide our analysis efforts of the FEEL dataset. Figure 1.2 presents an overview of the CRISP-DM process applied to the FEEL dataset.

Problem Understanding in this stage, we focus on understanding the problem area, objectives, and requirements of the project. We do so in this thesis by looking at related works that explore emotion labelling procedures and computational affect modelling in varied modalities.

Data Understanding focuses on identifying, collecting, verifying quality, and analyzing the datasets that can help in accomplishing the project goals. We highlight that although collecting data usually takes place at this stage, the data collection efforts precedes the involvement of the author. At this stage, we performed rounds of data exploration, cleaned up the data by removing artifacts of the collection methods, and devised experiments to further understand the information contained in different streams.

Data Preparation comprises of preparing the dataset for modelling. This step can include selecting the appropriate data streams, and cleaning, extracting, integrating, and formatting data as necessary.

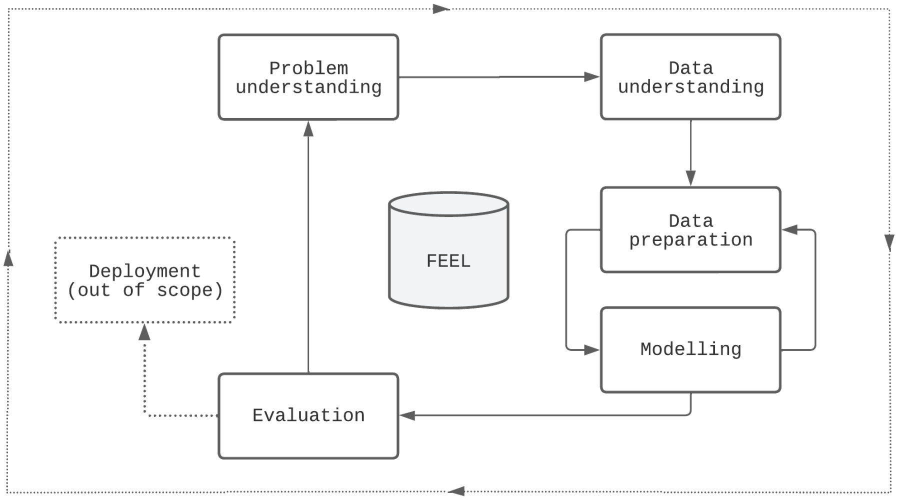


Figure 1.2: Cross-Industry Standard Process for Data Mining (CRISP-DM) process cycle applied to the FEEL dataset.

Data Modelling includes devising the test design, selecting modelling techniques, building, and assessing models. At this stage, we selected relevant strategies from previous works in affect modelling, as well as consulted theory work grounded in Psychology for further insights.

Evaluation while *technical assessment* of models happens during data modelling, the evaluation stage looks more broadly at whether the initial problem (or research question) has been addressed. We also reflected on the process and determined future work during this stage.

Deployment is the final phase for most data mining projects. In this stage, we prepare models to be deployed to the final user and create maintenance plans, as well as finalize any relevant reports that will accompany the work. Due to the exploratory focus of this work, deployment to users is out of scope of this thesis.

1.4.2 Emotion Analysis Guideline

In studies of emotion, researchers often capture instances of an emotion experience, coupled with a set of defined emotion labels, in an attempt to discover biomarkers in the brain or body for the corresponding emotion categories. In regards to analytical methods, these experience-label pairs are commonly evaluated on a single Machine Learning or Deep Learning technique [97]. This practice relies on the assumption that reported emotion labels refer to objective patterns that can be discovered. In this thesis, we adopt the premise that emotion expression and categorization are processes heavily influenced by a person's social and cultural experiences [9, 17].

The study of human emotions spans over multiple research fields. As an effort to ensure that our methodological decisions and reporting are consistent with our assumptions of emotion as psychologically and socially constructed subjective experiences, we performed a brief review of previous works on emotion modelling beyond the domain of Computer Science. As a result, we implement the following recommendations when analyzing emotion data [5, 52, 97, 112]:

1. Create personalized models and analyze emotion expression at an individual level [52, 97, 112].

2. Use more than one computational method to analyze data. This mitigates the extent to which modelling methods are responsible for any observed inconsistencies [5].

3. When using Machine Learning, compare multiple supervised classification algorithms and feature selection methods on the same dataset to explore implications of methodological decisions discovered [5].

1.5 Contributions

- Insights into the descriptive properties of a multi-pass labelling protocol.

- A set of evaluated strategies of how to employ triangulated emotion labels in computational models.

- Evidence of effectively modelling emotion expression in keypress force (KFP), contrasting with better studied comparable EEG-based models.

1.6 Researcher's Position and Ethics Considerations

Models are representational tools that can help us think, reason, or make predictions around a phenomenon of interest. We create models which are adequate and limited to a purpose, and not to provide true, holistic, perfectly accurate renditions of real-life processes and experiences. No model is morally neutral or free of subjective value judgments. When creating models, we have to make choices on what and how to model, what "truths" or assumptions we rely on, and what to conclude when interpreting our outputs [77]. As the primary decision-maker on most of the modelling work done in this thesis, I would like to provide the reader with a bigger picture of my trajectory up to this point.

I have a personal motivation in this topic of research as having grown around family members who struggle to communicate and process their emotions. I believe in the potential of emotion-responsive applications to benefit them and those around them. Academically, I come from a background in Systems Engineering, with an emphasis in Computational Intelligence. I moved to Vancouver in 2020 to pursue my Master's degree in Computer Science, and have since focused my efforts in learning more about how to design *with* and *for* people. Prior to my affiliation to University of British Columbia (UBC), I had the privilege to undertake basic Cognitive Psychology courses at the University of California, Berkeley, which gave me a jump-start in understanding emotion modelling research. Personally, as a South American, gender non-conforming, international student, I have both benefited and been harmed by societal structures around racism, sexism, homophobia, capitalism, and educational inequality. While my experiences made me aware of broader life perspectives, I acknowledge that I am not immune to personal biases.

As an former student at the Federal University of Minas Gerais, I would like to acknowledge that the land on which I studied is the traditional, ancestral, and unceded territory of the Aranãs, Xakriabás, Kaxixós, Pataxós, and Pataxós Hã-hã-hãe, and many others whose history and culture remains undocumented. As a graduate student at UBC, I would like to acknowledge that the land on which I currently live and study is the traditional, ancestral, and unceded territory of the xʷməkʷəýəm (Musqueam People). I recall the unjust, racist, and colonial practices that have had a lasting legacy, and continue to create prejudiced obstacles for Indigenous peoples across Canada and Brazil.

Finally, I used the pronoun “we” in this thesis instead of “I”. The “we” reflects the collaborative environment in which this research was developed, consisting of myself, my dear graduate and undergraduate colleagues, and my supervisor Dr. Karon MacLean. Other researchers will have different critical reflexive practices than myself. As the writer of this thesis, I will not speak on their behalf, but recognize that the result of much of this work has been a culmination of our collective positionalities, influencing the decisions and reporting style discussed in the chapters to follow.

1.7 Organization and Audience

This thesis is organized as follows:

- Chapter 2 presents related works in emotion modelling, followed by a brief background of the Machine Learning terms used in this thesis.
- Chapter 3 expands on the data collection protocol for the FEEL dataset and summarizes preliminary results using brain activity data.
- Chapter 4 dives into multi-pass emotion self-reports, exploring the information contained in each reporting pass.
- Chapter 5 compares emotion models built on FSR data with those built on EEG data.
- Chapter 6 presents different strategies for implementing computational models with multi-pass labels on FSR data, further relating to how emotion expression occurs via touch.
- Chapter 7 highlights reflections and recommendations to computational modelling of emotions based on results of Chapters 4, 5 and 6.
- Chapter 8 concludes this thesis, summarizing key findings and next steps.

1.7.1 Audience

In an effort to make this thesis accessible to the general Human Computer Interaction (HCI) community, we include a summary of key Machine Learning (ML) concepts (see Chapter 2.2). This section is self-contained and may be skipped without loss of significant information for those with a general understanding of ML theory.

Chapter 2

Related Work

What is emotion?

Clore and Ortony [17] describes emotion as the ability to make evaluations about our environments, the appraisal of “something as good or bad in some way”. Our understanding of emotions has changed considerably throughout the years. Much of current emotion modelling work is based in the theory of Core Affect [97, 112], which propose modelling these evaluations on a two or three-dimensional scale of level of arousal, valence, and/or dominance that reflect subjective emotional experiences [84]. Language and culture play major roles in emotion categorization within these different dimensions [9, 10, 17].

We can motivate recognizing and understanding of emotions through emotionally reactive applications, which often begin with developing computational models. We build models as representational tools capable of estimating emotion given a set of inputs.

Most studies in computational emotion modelling (also referred to as “emotion modelling” in this thesis) share a set of building blocks: (1) an emotion elicitation task, where we take care to consider what emotions we are inducing and in what context; (2) a set of inputs through which we infer emotion expression–touch, speech, images of facial configurations; (3) emotion reports, through which either the participant provides an estimate of a "ground-truth", or someone else annotates that data with what they perceive the participant is feeling. The resulting data from these blocks build up to the models themselves, often involving statistical analysis, qualitative methods, or, as in this thesis, machine learning.

In this chapter, we provide an overview of decisions that factor in emotion data collection and summarize previous work in computational emotion models relevant to what we propose in this thesis. In interest of making this thesis accessible to the general public, we add an expanded glossary (Chapter 2.2) covering ML terms used in the coming chapters.

2.1 Emotion Data

Machine interpretation of spontaneous emotion encoded in expressive modalities involves building models on ecologically valid labelled emotion data. We relate to existing literature key considerations around three aspects of building a model: emotion elicitation, self-reports used to label emotion data collected in real-time, and the selection of an emotion metaphor on which to focus the computed model.

Labelled datasets are necessary to train and evaluate models for estimating emotion. Studies that aim to collect emotion data generally involve key decisions in defining the emotion elicitation task, selecting (or creating) emotion measurement instruments, capturing one or more affective expression modalities, and construing emotion metaphors from collected labels.

2.1.1 Emotion Elicitation

Where applications require in-time recognition of emotion, data must represent realistic emotion expression [32, 43]. Relived or recalled emotion is one proxy [15, 27]. Participants are prompted with an emotion word (the single label) and asked to recount the story of a past intense experience.

While successful in eliciting authentic and wide-ranging responses, this oversimplifies an episode to emotive homogeneity [15]. Furthermore, participant stories are hyper-individualistic, not amenable to a search for commonalities. Conversely, entertainment media can root participants in a more uniform elicitation stimulus, with many validated video and music clips used successfully for this purpose [18, 34]. Video games have shown promise in producing physiological responses analogous to that of real life evocations [96].

2.1.2 Self-Report Methods

Classifying emotion requires capturing and labelling emotional experiences. Representation thus impacts how we ask users to report their experience.

Russell's circumplex [84] is a commonly used instrument depicted as a spatially continuous 2D space of arousal and valence (plus dominance in 3D [6]). It underlies popular labelling schemes, most involving a participant locating emotion words on its axes; e.g., words associated with PANAS [105].

Self-Assessment Manikin (SAM [11]) simplifies emotion labelling with a non-verbal pictorial assessment representing a combined Likert scale on pleasure, arousal, and dominance dimensions [93, 108].

Natural language reporting methods are used when experiences (maybe a self-contained memory [14], or a touch [39]) are sufficiently brief, simple to fit a single label, and precede an opportunity for the participant to report without experiential interference. They become intractable for segments that are longer than a few moments, span multiple emotions, and/or require rapid computed response (before the segment ends).

Still with a dimensional representation, others have collected *temporally continuous* emotion ratings using a mouse- [19] or a joystick [88, 110]. For hands-free activities, a joystick allows for high temporal-resolution concurrent reporting, but at the cost of emotional intrusiveness. Post-hoc ratings require review of a recorded experience.

2.1.3 Emotion Metaphors

In order to estimate an amorphous quantity like emotion, for classification purposes or otherwise, one must first explicitly or implicitly choose a representation metaphor which defines how we regard the emotion experience, the language we use to describe it, and the parameters with which we attempt to capture it [12, 56, 76]. Altogether, this descriptive framing and parameterization is sometimes referred to as *emotion modelling* [72]. To avoid overloading the term "model" or confusing the use of emotion models with classification models of emotion, we adopt [12]'s terminology of emotion *metaphor* to refer to how we think about emotion representation. We use *model* to refer to the computational (e.g., machine learning) model implementation. In preparation for building computational models of emotion, we start by addressing commonly used emotion metaphors: emotion *states*, *dynamics*, and *appraisal*.

Emotions-as-State

Classifying emotion as *state* has many practical benefits for machine recognition. There are many validated instruments for identification and measurement, such as Russell's circumplex grid [84] and the Self Assessment Manikin or SAM scale [11]. These are beautifully simple measurement scales which employ forced choice and offer simple and straightforward classes for data labelling. In contrast, emotions rarely fit into convenient boxes. Rather than clean-cut 2D quadrants of arousal and valence or elegant linear scales of dominance, our emotional lives are complexly dynamic in situation-dependent ways: who we are with, how recently our physical and emotional needs have been met, and why we are in the present moment with all the baggage of our cultural and personal history [10].

Emotion Dynamics

Emotions evolve throughout the course of a single event or experience, as well as longer extents of time [56]: consider the emotional journey followed by your favourite engaging movie scene. Psychologists Kuppens and Verduyn [56] propose dynamic emotion metrics to describe changes across an emotional experience, with the most prominent being *emotion inertia* (resistance to variation, quantified as signal autocorrelation); *emotion instability* (mean square of successive differences as the amount of change); and *emotion variability* (within-subject variance respectively to represent the range of change) [44, 94]. Operationalizing concepts rooted in emotion dynamics for computational applications requires labels capturing transitional emotional experiences as they happen. We propose the use of emotion direction as a dynamic emotion metaphor to describe where a present emotional experience may evolve towards.

Before diving into the literature around computational emotion modelling, it is important to define the common terminology underlying Machine Learning (ML) theory used in this thesis. The following section is framed as an expanded glossary of terms mentioned in the next chapters, and may be skipped without loss of information.

2.2 Learning Machines

The concept of "learning" in ML represents the process through which we build self-improving models based on existing data, that is, *computational models* can *learn* to achieve better performance without being explicitly programmed.

2.2.1 Types of Learning

Machine "learning" comprises of three different processes through which a computational model can be estimated [73]:

1. Predictive or supervised learning: in this modality, training data comes in observation-label pairs. The end goal is to derive a model that generalizes well to new data, i.e., , by being capable of mapping new, unlabelled observations in the label space.

2. Unsupervised learning: in cases where only a collection of inputs is available, it is possible instead to derive underlying patterns, i.e., , describe possible correlations between features, cluster observations in a few groups based on similar behaviour, and detect *outliers*. This type of learning is often referred to as knowledge discovery.

3. Reinforcement learning: the basic components of these types of systems involve perceptions, actions, and rewards. A *agent* interacts with a *environment* and is rewarded depending on whether or not its actions are in accordance with the main purpose of the system. This way, the goal of the agent is to learn the behaviour that maximizes its expected cumulative reward over time.

This thesis focuses mostly on supervised learning, with unsupervised learning as an exploratory analysis tool. In particular, we focus on supervised *classification* tasks. A classification task involves mapping observations to discrete categories (or "labels")[1]. The number of outputs that can be attributed to an observation determines the type of classification task at hand. Figure 2.1 illustrates the different types of classification tasks explored in this thesis.

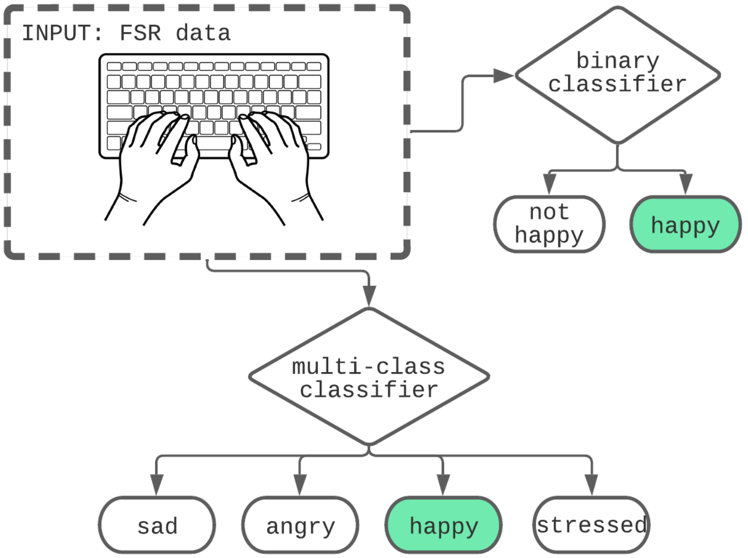


Figure 2.1: Illustration of single label classification tasks. At the right, we exemplify binary classification–pick one of two categories–and at the bottom, multi-class classification–pick one of many categories).

Single-label binary supervised learning: In binary classification, we attribute a single output from a set of two possible labels to an observation. For instance, we could train a model to predict whether someone is "stressed" or "not stressed" based on their heart-rate data.

Single-label multi-class supervised learning: We attribute a single output from a set of many possible labels to an observation. For instance, we could train a model to predict whether someone is emoting "anger", "sadness", "happiness", or "disgust" based on their facial configuration.

Multi-label supervised learning: In multi-label classification, we attribute one or more outputs from a set of many possible labels to an observation. For instance, we

[1] A *regression* task can be described as a generalized classification with *continuous* labels

could train a model to predict whether someone is feeling more than one emotion (e.g., sleepy and angry) at a given time based on their EEG data.

Soft-label supervised learning: In the previous examples, a "hard label" represents a binary membership relationship–someone is either feeling "sad" or "happy". Soft labels allow for flexibility: we output a score (probability or likelihood) of the observation belonging to each one of the possible categories (e.g., predict a 30% chance that someone is feeling "sad", and a 20% chance of feeling "angry").

2.2.2 Learning with Multiple Labels

Single-label learning is the default implementation for many off-the-shelf ML algorithms [78]. Conversely, there are many ways of approaching multi-label learning. In this thesis, we choose two strategies based on simplicity of implementation and interpretation of outputs: flat multi-label learning and hierarchical learning. Figure 2.2 illustrates both approaches.

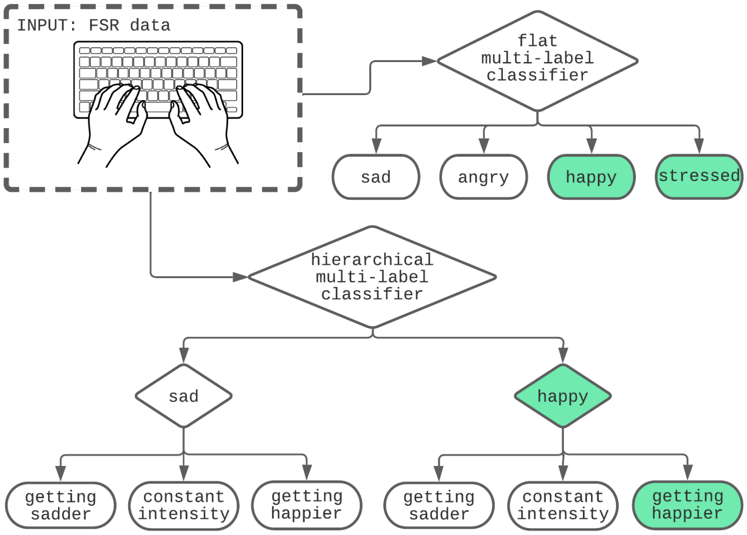


Figure 2.2: Illustration of multi-label learning tasks. At the top right, an example of flat multi-label classification, in which each label is considered an independent category. At the bottom, an example of hierarchical multi-label classification, where decisions on certain label categories can influence each other.

Flat Multi-Label Models In our flat multi-label approach, we adopt a one vs. all strategy, creating several binary classifier models, one per category. The models are trained independently, and the results are outputted as a set.

Hierarchical Multi-Label Models In our hierarchical multi-label approach, we adopt a multi-stage one vs. all approach. In each stage, we train several binary classifier models, one per category. These models are trained independently, and the results of the upper levels inform the classifiers at the lower levels, i.e., , the label outputs of each level become feature inputs to the following levels.

2.3 Machine Classification on Implicit Emotion Expression Modalities

Humans reveal a wealth of emotion non-verbally, by communicating it to others through modalities such as facial expression [7, 70], touch behaviour [15, 49], body language and posture [95]. Our work builds on the body of work on machine classification of emotion; here, using multi-pass emotion self-report as labels over brain activity and keypress force during video gameplay–a dynamic emotion experience.

Affect in Typing

Social touch pressure has been shown to communicate current affect [15, 39]. While not social touch, keyboard typing behaviour or keystroke dynamics in the context of emotionally intense experience offers a view of the connection between manual activity and real experienced emotion [58, 65]. Classification systems using keystroke dynamics have largely employed features related to keypress timing (e.g., typing speed, time between keystrokes, and delete/backspace frequency) to predict the emotion experienced during keyboard use [28, 65, 75], resulting in accuracy rates approaching 88% (chance at 50% for two-level models [28]).

Investigating keypress force or pressure during keyboard use for emotional content, however, is relatively new [37]. Using pressure-sensitive keyboards, emotion has been classified using typing pressure with up to 93% accuracy (chance 17%) [64], with Hernandez et al. [37] finding a positive correlation between stress and keypress force in typing behaviour. In our work, we explore how keypress force might communicate emotional transitions between Stressed and Relaxed on pressure-sensitive control keys.

Affect in Brain Activity Measured by EEG

Although brain activity data is out of scope for this thesis, we include an overview of related works with this affective modality for later comparison with results from FSR data.

EEG measures changes in electrical potential around the scalp through the use of electrodes placed at various locations around the head [55]. This data stream captures aspects of the physiological state of an individual and has been shown to predict symptoms of various brain disorders (epilepsy, seizures, sleep disorders, etc.) [20, 45, 99]. Due to EEG's high (millisecond) temporal resolution, various studies have used it to capture short-lived emotion experiences [23, 60, 74]. Its relatively low signal-to-noise ratio is a recognized challenge, requiring the creation of reliable features from the original signal [55]. Energy spectrum-based features tend to perform well, achieving accuracy rates near 87.53% using an Support Vector Machine (SVM) classifier (chance at 50%) [104].

Building on the success of deep learning models, Xing et al. [109] applied a Long Short-Term Memory (LSTM) model using features generated with an AutoEncoder to classify time-based features directly; and obtained a mean accuracy of 81.10% on valence labels (chance 50%) [109]. Recently, 2D differential entropy-based features, which are capable of capturing spatial relationships, have been combined with Convolutional Neural Networks (CNN) models to classify emotional experiences (positive, negative, neutral) at an accuracy of 97.10% (chance 33%) [1].

Classification of brain activity is dominated by data instances highlighted by Event-Related Potentialss (ERPS), wherein time windows are often constructed around 100ms - 750ms after an event [22, 26, 83]. ERP is not conducive to capturing emotion *evolution,* where brain activity may change over the course of minutes and hours [102].

2.4 Emotion Modelling with Multiple Reporting Passes

Emotion may also evolve under scrutiny and cognitive processing [72, 76]; with some time and reflection, emotional assessment of an experience may be categorically different than the initial evocation [76]. Emotions may be most intense while directly in or just after an experience [89, 102] but a person may only be able to articulate it after an emotional peak has passed, requiring time to assess and consider the appropriate language [98]. This timing tension suggests there may be an ideal time for naming the emotion [24] somewhere shortly after an experience to give enough time for processing [89] but before memory degrades [81].

Although some studies report analysis on multiple labelling schemes (e.g. Ekman's emotion categories and Russel's dimensions) [82, 92, 97, 103, 112], most research on computational emotion models relies on a single pass of either observed[2] or self-reported emotion annotations [82, 92, 97, 112]. To the best of our knowledge, the present work is the first attempt to conduct research on multiple passes of self-reported emotion.

We posit that combining self-reported emotion labels collected at different time frames, but close enough to the primary elicitation task to avoid memory degradation, will lead to better performing models. Moreover, we are interested in exploring touch data as a potential modality for capturing implicit emotion expression in computational models, comparing performance to more intrusive, but more studied, brain activity signals. To this end, we elect the Force EEG and Emotion-Labelled (FEEL) Dataset as the base from which we build our analyses. In the following chapter, we detail the collection protocol for FEEL, discuss preliminary results, and identify next steps.

[2] Annotated without the participant by one or more judges

Chapter 3

FEEL: Force EEG and Emotion-Labelled Dataset

To model spontaneously evolving emotions, we analyze data from participants while they played an emotionally evocative videogame. The data, collected prior to the work reported in this thesis[1], follows a multipass data labelling protocol [16], recording brain activity in the form of Electroencephalogram (EEG) data and key-press force via an Force Sensing Resistor (FSR)-embedded keyboard. In this chapter, we detail the collection protocol for the Force EEG and Emotion-Labelled (FEEL) dataset, illustrated in Figure 3.1. Last, we describe preliminary results obtained from modelling emotions using Electroencephalogram (EEG) data, which we refer to in Chapter 5.

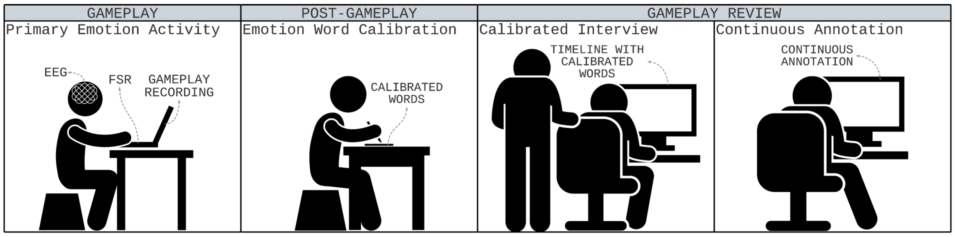


Figure 3.1: Overview of the Force EEG and Emotion-Labelled dataset collection protocol.

[1] Data collected without the author's involvement (see Preface).

3.1 Collection Protocol

In the following, we report the FEEL collection protocol and data schema in detail, to put the data in context for its use in Chapters 4-5. Some is drawn from [16] (for which the thesis author is a co-author) and further describes the data collection.

The Force EEG and Emotion-Labelled (FEEL) dataset contains 64-channel brain activity and 5-key keypress force data as well as two emotion labelling passes, self-reported at high and low densities, collected from 16 participants. All data is time-aligned to an emotion task, playing Playdead's Inside – a horror video game featuring chase and puzzle scenes, navigated serially and thus amenable to temporal alignment of gameplay. The dataset consists of comma separated value (.csv) files which are organized by participant. Video data is excluded for participant privacy. Gameplay averaged 13min 24sec (minimum: 8min 25sec, maximum: 21min 37sec, std 3min 08sec).

3.1.1 Participants

FEEL is collected from N=16 participants, 8 female and 8 male; 8 between 19-24 and the other 8 between 25-34 years of age. All played videogames regularly from a few hours a month up to 4 hours daily, nearly all of whom report 1-6+ hours per week; none had played the videogame featured in the data collection protocol. See Appendices A.1 and A.2 for recruitment poster and consent form, respectively.

As part of recruitment, participants completed a questionnaire adapted from the Trait Meta Mood Scale (TMMS) [85]. Only those who self-reported as having high emotion clarity and low emotion suppression were invited to participate in the data collection. Participants were compensated with an honorarium of $30 for the 2-hr data collection session.

3.1.2 Data Capture and Preparation

Collection Sequence Overview: Data was collected in four steps that produced the FEEL dataset components which are illustrated in Figure 3.1 and described below in detail.

1. *Initial Gameplay* generated streams of participant brain activity (EEG) and

keypress force (from an FSR-embedded keyboard). Scree recordings and player video was excluded from FEEL due to privacy considerations.

2. In *Word Scale Calibration*, participants provided their personal understandings of emotion words relative to one another, for later use.
3. Then in the first self-report cycle (*Interview with Calibrated Words* (IWCW)) participants reviewed the gameplay video and annotated it with their calibrated word sets.
4. Finally, in the second self-report cycle (*Continuous Annotation* (CA)) they used a 1-D joystick to annotate the video.

Task Order: Task order was carefully chosen to minimize influence on emotion elicitation while increasing the likelihood that participants would use a common set of emotion words to describe their experience. During Step 3, the interview allowed players to explicitly process their emotions out loud, guided by researchers looking for notable emotional events – e.g., strong emotions, startling or uncomfortable moments, odd behaviour. Leaving the joystick evaluation as the final step lets participants internalize and contextualize the emotion scale in preparation for the continuous annotation.

Data Alignment: It was crucial that all temporal data (collected from initial gameplay and continuous annotation) was synchronized. For these time-linked steps 1 and 4, the setup was configured such that a single button press sent a synchronization signal from the game-play computer to the modality logging systems handling each data stream, and triggered a screen colour change on the game-play computer's screen. This produced a synchronized timestamp to align the modality streams and, later, the continuous annotation input with each frame of the videogame play.

Brain Activity Data Stream (EEG)

Brain activity signals were captured using EGI's EEG 400 system[2] which features a 64-channel Routine Hydrocel geodesic sensor net and proprietary NetStation data

[2] EGI EEG research system details available at: https://www.egi.com/research-division/eeg-systems/geodesic-eeg-systems

collection and visualization software. Data was sampled at 1kHz, with a band pass filter of 1Hz to 50Hz applied in post-processing. This range was chosen to remove high frequency jitter and 60Hz mains noise [113] yet retain the frequency bands (α, β, and γ) associated with emotion processing [2].

Initially, the data cleaning procedures included: artifact removal with MNE-Python tools[3] and used Independent Component Analysis (ICA) to detect and remove electro-oculogram (EOG) artifacts, such as eye blinks. Parallel to the work in this thesis, we developed scripts to identify channels with unusually high noise levels, verified by visual inspection, and removed affected segments. Since these measures did not produce significant classification improvement, we report classification results with only the band-pass filter as described above.

Keypress Force Data Stream (KFP)

Keypress force was recorded from a standard keyboard with Whadda force-sensitive resistors[4] (FSRS) embedded in game-specific control keys (four direction keys and ALT). The analog signals from these 5 FSRS were processed into digital signals using an Arduino Mega 2560[5] running Standard Firmata and a custom node.js server keylogger using the johnny-five API. Amplitude or force ranged from 0 (no contact) to 1024 units (∼1kg) at 52Hz[6]. Reported results and the published dataset use downsampled FSR data matched to the videogame framerate at 30Hz.

Timeline with Calibrated Words (TwCW)

The Timeline was created from collection sequence Steps 2 and 3, described now in more detail.

Word Calibration: Following the gameplay, players first calibrated a Stressed-Relaxed emotion scale, contextualizing points along the scale with memories of their recent gameplay experience and marked with 13 pre-selected emotion words from the PANAS [105] ("Calibrated Words"): Cautious, Satisfied, Hopeful, Frus-

[3] MNE description and tutorials available at https://mne.tools/stable/index.html

[4] Whadda FSR WPSE334 available at: https://whadda.com/product/force-sensing-resistor-fsr-wpse334/

[5] Arduino Mega 2560 available at: https://store.arduino.cc/products/arduino-mega-2560-rev3

[6] As defined by the FSR specifications available commercially at https://whadda.com/product/force-sensing-resistor-fsr-wpse334/.

trated, Anxious, Nervous, Threatened, Resigned, Alert, Accomplished, Fearful, Dread, and Curious. Write-ins were allowed though not used for classification in present analysis. Figure 3.2 exemplifies the result of this step.

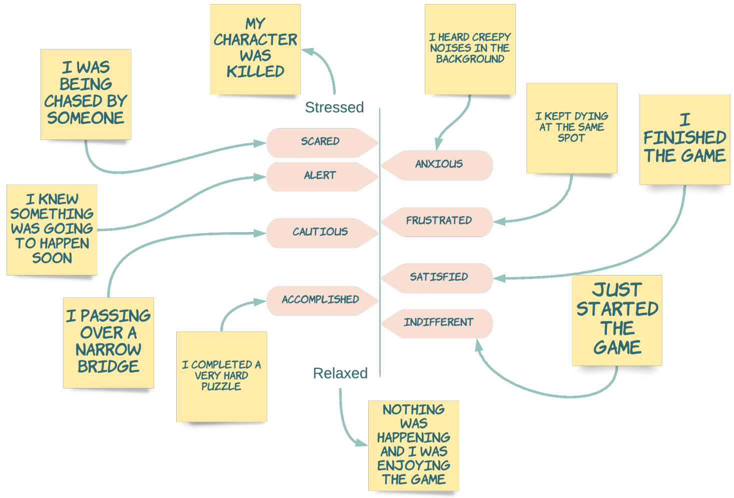

MY CHARACTER WAS KILLED
I HEARD CREEPY NOISES IN THE BACKGROUND
I WAS BEING CHASED BY SOMEONE
Stressed
I KEPT DYING AT THE SAME SPOT
I FINISHED THE GAME
SCARED
ANXIOUS
ALERT
I KNEW SOMETHING WAS GOING TO HAPPEN SOON
FRUSTRATED
CAUTIOUS
SATISFIED
ACCOMPLISHED
JUST STARTED THE GAME
INDIFFERENT
I PASSING OVER A NARROW BRIDGE
I COMPLETED A VERY HARD PUZZLE
Relaxed
NOTHING WAS HAPPENING AND I WAS ENJOYING THE GAME

Figure 3.2: Illustrative example of contextualized emotion word calibration.

Calibrated Interview Self-Report Procedure: Players then carried out their first, low data density annotation pass of gameplay video review by indicating which (calibrated) emotion word applied at gameplay points that they recalled having some emotional intensity.

TwCW Construction: We constructed a TwCW by integrating the Calibrated Words and the Interview results. Each interview annotation consists of a specifically calibrated word plotted as its calibrated value at the annotation timestamp along the videogame play.

Continuous Annotation Stream (CA)

In the second gameplay review, the CA is generated from a non-biased joystick (one that holds the last position rather than returning to centre) tracing an emotion time series, where the height of the curve represents the participant experience between Relaxed and Stressed over the timeline of the gameplay experience. Joystick position readings were matched with video frame rate of 30Hz to ensure alignment with video playback. To reduce some of the analog jitter in the joystick data, we used a simple moving average filter as a smoothing step before analysis and dataset publication. Joystick values are normalized to range from 0 to 1.

In total, the outputs of this protocol are player specific gameplay streams (EEG and FSR), emotion word calibrations, and the TWCW and CA – two time-series of emotion self-report annotated on the same dimensional plane of the Stressed-Relaxed scale over the gameplay timeline. Figure 3.3 illustrates the data collected during the study.

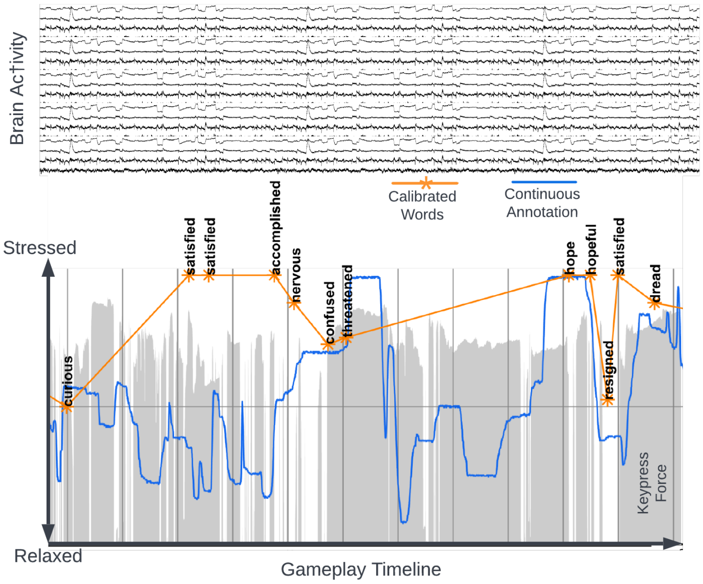


Figure 3.3: A sample of the data streams collected during the FEEL study: EEG from brain activity (top) and the FSR from keypress force collected during gameplay, superimposed with self-reported Continuous Annotation and Timeline of Calibrated words (bottom).

3.2 Preliminary Results Using Brain Activity with Continuous Annotation

For the FEEL dataset, preliminary results of emotion expression in EEG data have been obtained in analysis spearheaded by the authors' collaborators Guta and Cang (see Preface). As the work is not yet published, we overview it here with their permission. It is relevant to Chapters 4-5.

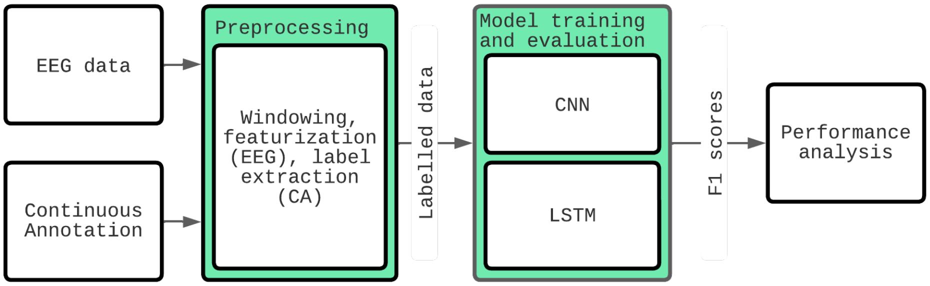


Figure 3.4: Overview of classification pipeline for EEG data.

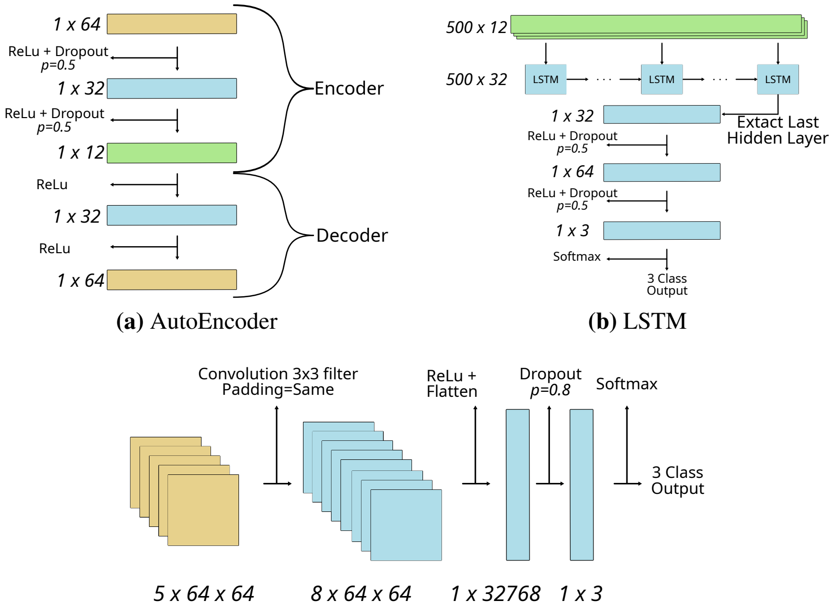
Using the same time window boundaries as the emotion labels (0.5s, 1s, 2s, and 5s), data instances were generated from the brain activity signal. The EEG data is high density, ideal for classification using deep learning models. Cang and Guta trained time- and frequency-domain models–using Long Short-Term Memory (LSTM) and Convolutional Neural Networks (CNN) models respectively–with the Pytorch[7] deep learning framework. Figure 3.4 summarizes the classification pipeline implemented for EEG data.

3.2.1 Time Domain and LSTM Classifier

LSTM models capture time dependence within a given window and are ideal for tasks where interpreting the current instance of information depends on recent context e.g., natural language speech recognition and translation [41]. For emotion evolution over time, the LSTM as structured in Figure 3.5(b) is used to model EEG emotion expression, classifying time domain features directly.

Time-Domain Feature Extraction by Autoencoder: To combat overfitting to signal noise, Cang and Guta explored dimensionality reduction by applying an autoencoder for feature extraction [109], compressing the 64-channels down to 12 variables. For a given window of length N timepoints, they constructed a time domain feature set of size $N \times 12$ variables as shown in Figure 3.5(a).

[7] https://github.com/pytorch/pytorch



(a) AutoEncoder

(b) LSTM

(c) CNN

Figure 3.5: The AutoEncoder (a), LSTM (b) and CNN (c) models used for brain activity classification. The AutoEncoder is trained in order to encode the 64 channels at every data point in a given window down to 12 variables. The LSTM uses the encoded features from the AutoEncoder to generate a 3 class output. The CNN uses the 5 bands × 64 channels × 64 channels frequency features as an input to generate a 3 class output.

3.2.2 Frequency Domain and Convolutional Neural Network Classifier

The use of Convolutional Neural Networks (CNN) models has increased significantly as they have come to solve a wide range of 2D based classification tasks. Specifically, Cang and Guta build upon the model CNN model structure proposed by Ahmed et al. [1] for affect classification. Based on initial experiments, they found a model consisting of a single convolution layer and a single fully connected layer as depicted in Figure 3.5(c) to be best as more complex architecture tended to overfit to the data.

Frequency-Domain Feature Extraction: For emotion classification using CNNs, differential entropy (DE) features calculated in the frequency domain tend to be more robust to noise [90]. Cang and Guta calculated the DE for each of 5 frequency bands [23]: delta (1-4Hz), theta (4-7Hz), alpha (8-12Hz), beta (12-30Hz) and gamma (30-50Hz), so chosen due to demonstrated activity during emotion expression [1, 47].

3.2.3 Classification Performance

Cang and Guta report F1-scores over chance. Chance is dependent on participants' class distribution by emotion label generated from self-report and window size. Figure 3.6 summarizes scores for CNN and LSTM models.

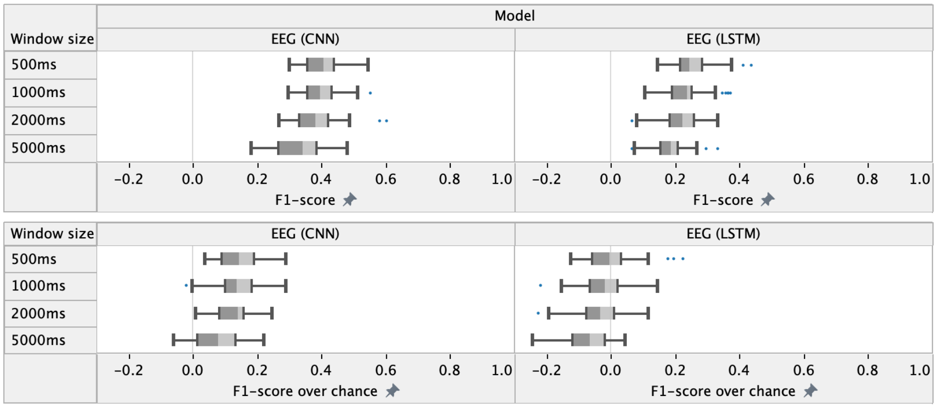


Figure 3.6: F1-scores for EEG models by window size.

Based on their results, Cang and Guta highlight that:

(1) CNN models outperform LSTM in every experimental condition, with LSTM landing at or below chance;

(2) the highest F1-scores are achieved with the smallest windows.

3.3 Summary

In this chapter, we detailed the collection protocol for the Force EEG and Emotion-Labelled dataset, which features brain activity and keypress force data annotated with two distinct time series of self-reported emotion. In addition, we summarized preliminary work performed using EEG data and continuous annotation of emotion. Before further exploring FEEL, we acknowledge some of its limitations:

1. Single-session data: while FEEL includes a collection of curated emotion annotations, we cannot assess how participant behaviour might change from one session to another.

2. One-dimensional emotion scale: participants annotate their experiences from relaxed to stressed, and are primed to use emotion words aligned with this scale, narrowing the range of emotions our models can capture.

3. Single calibration pass: calibration might change from one session to another, or even before and after an emotion experience.

4. Elicitation activity: the tense videogame provides a dynamic emotion experience, but we cannot assess how well the emotion annotations conform to other types of experiences.

5. Limited touch data: while keypress force is unobtrusive to collect in this scenario, we are only able to infer keyboard activity and pressure information. We wonder how implicit expression of emotion happens in other dimensions of touch (e.g., shear, hover).

In the next chapters, we investigate how different reporting passes align and what kinds of information can be extracted from each. Then, we analyze the impact of annotation style on machine classification of dynamic emotion using touch data.

Chapter 4

Exploring Multi-Pass Emotion Self-Reports

Our present analytical goal is to explore the properties of and relationships among the reports obtained in the FEEL dataset, primarily by examining the degree and nature of their [dis]similarity over a range of metrics, and probing for physical intuition among them.

4.1 Label Metaphors

In labelling the emotion experience, participants directly annotated the gameplay timeline with a Continuous Annotation (CA) curve. For label definition purposes, we call the curve $\mathscr{E}$ where $\mathscr{E}(t)$ is the value on the Relaxed-Stressed scale at time t. We use zero-hold interpolation of the CA (downsampled to 30Hz) for temporal alignment with the processed FSR (30Hz) data.

To track dynamic emotion, we refer to emotion direction as the overall rate of change of $\mathscr{E}$ in a given time window, calculated by finding the slope m of the best fit line (minimizing the Least Squares Error). We restrict the domain of m by transforming it into an angle using $arctan(m)$, bounded by $(-\frac{\pi}{2}, \frac{\pi}{2})$, scaled and translated to $(0,1)$.

Emotion position labels are calculated as the mean reported value over the time window[1]. We calculate this metaphor by taking the arithmetic mean of $\mathscr{E}(t)$ in a given window.

To evaluate whether a reduction in noise would improve classification, we create a third label metaphor of an emotion based on the idea of area, here referred to as an emotion **accumulator**, which is the additive overall sum of emotion values or the area under the $\mathscr{E}(t)$ curve of a given time window. Specifically, we find the mean using the Trapezoid rule

$$\frac{1}{N}\sum_{i=1}^{N}\frac{\mathscr{E}(t_{i-1})+\mathscr{E}(t_i)}{2}$$

where N is the number of data points in a given window.

All emotion metaphor calculations are in the range of (0,1); for the purposes of classification comparison, we transform label metaphors into three distinct classes by splitting the scale into thirds, specifically, $[0,\frac{1}{3})$, $[\frac{1}{3}, \frac{2}{3})$, $[\frac{2}{3}, 1]$.

Figure 4.1 depicts the label distribution of these three classes for each metaphor of direction, position, and accumulator. To overcome class imbalance during model building, we over-sample underrepresented classes by bootstrapping and use F1-macro average as our reporting score.

4.2 Commonality in Interpreting Emotion Words

To assess across-participant similarity of calibration ratings (as a proxy for model generalizability), Figure 4.2 plots *rating variance* for each of the calibrated words in order of decreasing agreement (increasing variance).

For a quantitative view of cross-participant consistency, we also conducted an *Intra-Class Correlation* (ICC) test (also known as *inter-rater reliability test* [54]). For the subset of emotion labels rated by all participants (*Anxious, Cautious, Frustrated* and *Satisfied*), we found ICC(2, k=16)=0.99, $p \ll 0.01$ (α=0.05, CI=[0.97, 1.0]), based on mean rating over an absolute-agreement, 2-way random-effects model. ICC values > 0.9 indicate high reliability [54], suggesting these ratings are overall highly similar across-participant for this set of emotions. Indeed, the four rated by all participants had an ICC(2, k=16) of 0.99.

[1] Considerations about data windowing described in Chapter 5.1.2.

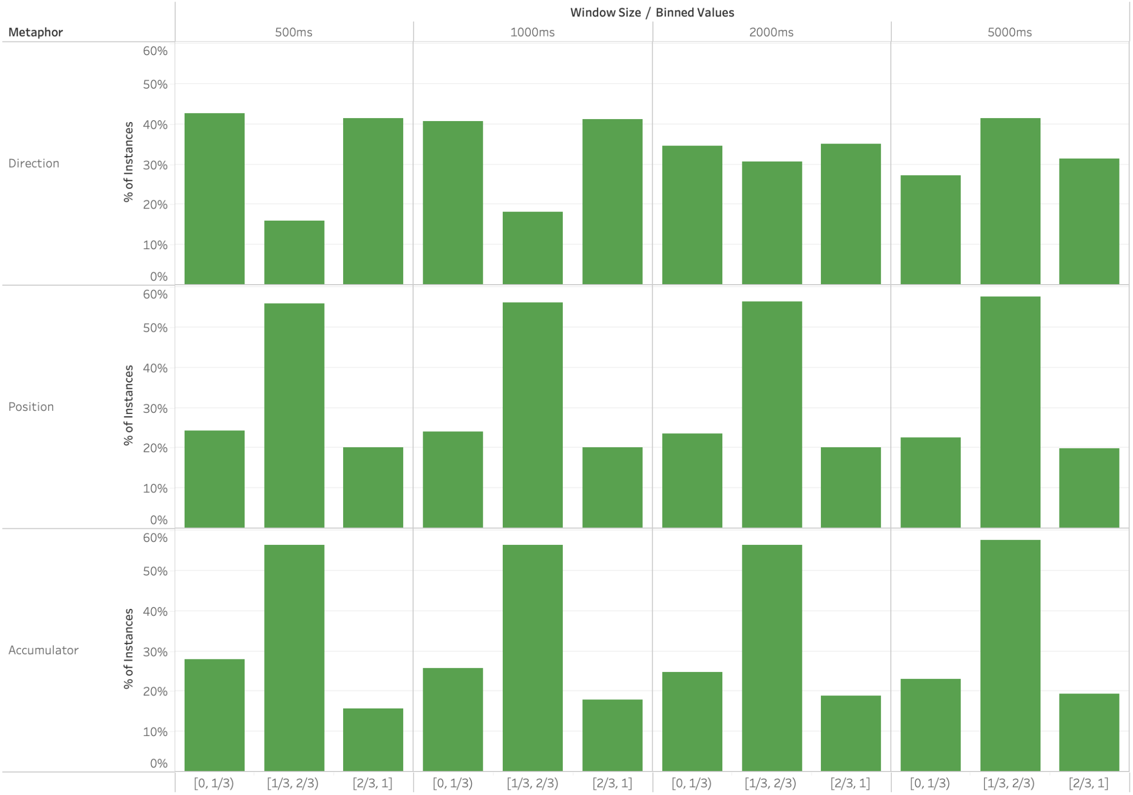


Figure 4.1: Distribution of class labels for each label metaphor and window size.

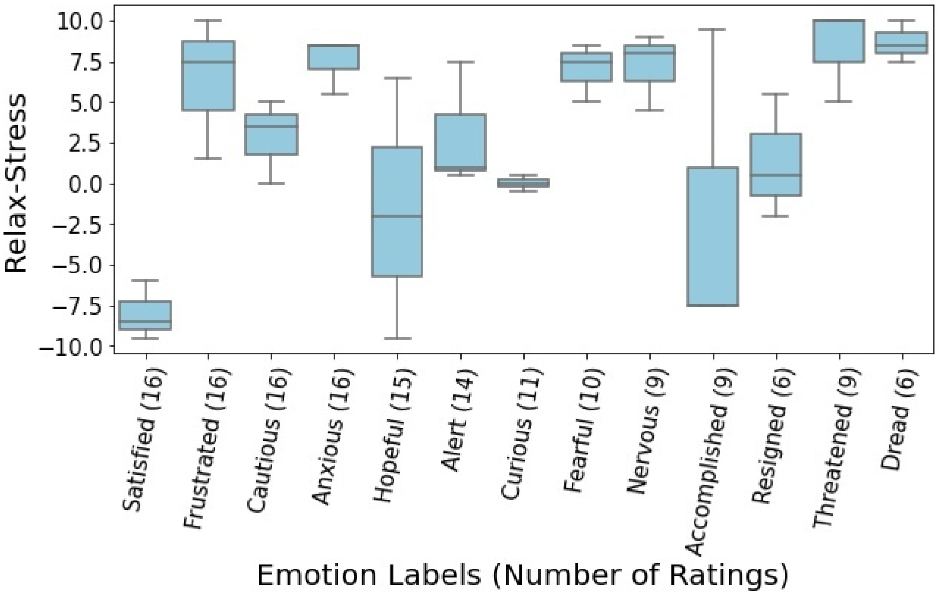


Figure 4.2: Rating variance by calibration word, ordered by number of participants who provided a rating for that word.

However, this agreement varies as set size increases, first decreasing monotonically then dropping sharply at *Satisfied* - *Resigned* to ICC(2, $k = 4$)=0.83. This may be partially due to the relative sparseness of ratings.

Taken together, these results support that there are **substantive differences in how individuals interpret emotion words, highlighting the importance of personalized models.**

4.3 Self-Report Modality Consistency via Time Series

High similarity between self reports indicates consistency and perhaps interchangeability of report modalities; differences might suggest invalidity of one or both, or that they capture different information. Interpreting within-participant TWCW and CA as time-series, we use standard time-series analysis methods [63] (with appropriate condition verification steps) to check for signal similarity – Pearson's correlation – and confirm that both data streams are appropriate responses to a common stimulus – Granger's Causality [87]).

4.3.1 Test Preparation

Using raw report data, we first confirmed that **both time-series were stationary** with the Augmented Dickey-Fuller (ADF) test (Bonferroni-Holm correction $\alpha = 0.05$, $p_{\mathrm{BH}} < 0.02^2$), and that their statistical properties did not change over time [31]. Prior to evaluating cross-correlation between the two reports, we verified that each was not auto-correlated to avoid artificially inflated correlations [21]).

With Python's *statsmodels* [86], all peaks were at lag = 0 for all participants' TWCW and CA auto-correlation plots (i.e., both signals present low correlations at all lagged versions of itself). We conclude that **neither signal is self-similar**.

The TWCW and CA self-reports are sampled at different times and resolutions (0.05Hz and 30Hz respectively). We downsampled the CA series rather than interpolate the sparse TWCW, to minimize bias.

4.3.2 Pearson's Correlation for Signal Similarity

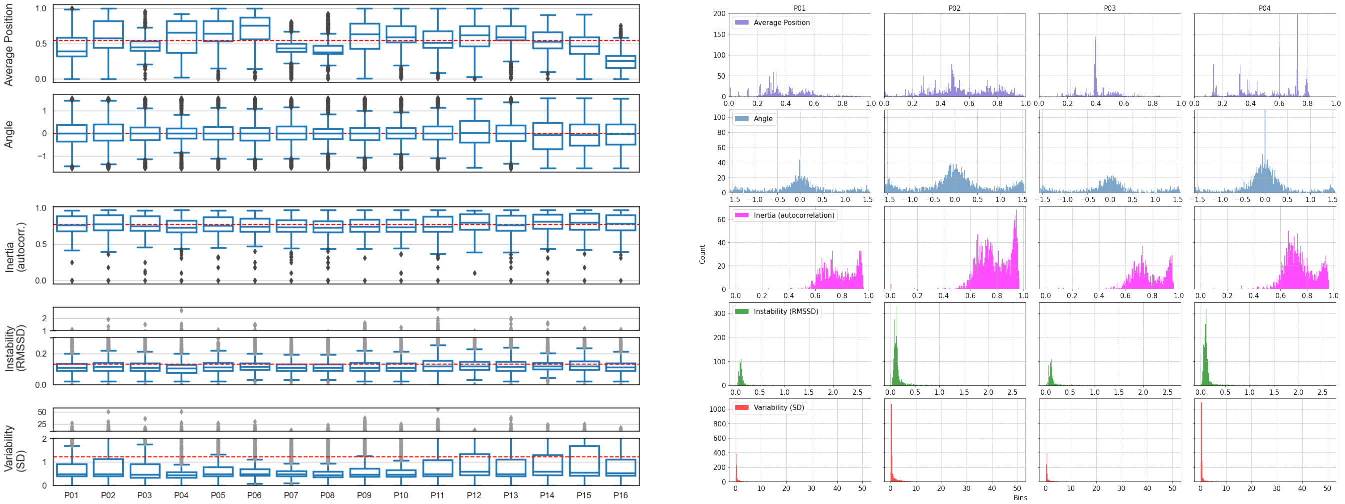
We analyzed signal similarity between participants[3]. P01, P02, P08, and P14 had moderate correlation coefficients for the two emotion self-reports (CA and TWCW) at $\rho > 0.3$ ($p_{\mathrm{BH}} < 0.05$). However, in general there was no significant correlation between the report streams: p-values exceed the threshold after a Bonferroni-Holm's adjustment to $\alpha = 0.003$. We infer that **individuals' self-reports differed** in the metrics we observed.

4.3.3 Granger Causality Test for Source Plausibility

Although Granger cannot confirm direct causality between different variables [91] (i.e., it does not claim TWCW causes the CA values), we employ the test to evaluate whether time-series for CA could *forecast* TWCW and vice versa. We employed a Bonferroni-Holm correction (α_{BH}=$0.05/N$, N = number of participants). We found significance for 15 of 16 participants ($p_{\mathrm{BH}} < 0.048$), suggesting that one label stream could be used to forecast the other for all except P02. This implies **the data streams are appropriate as responses to the same stimulus**.

[2] For all except P01 (TWCW): $p_{\mathrm{BH}} = 0.07$, ADF test statistic $= -2.671$

[3] Pearson's correlation results at $\alpha = 0.05$: P01 ($\rho = 0.38$, $p_{\mathrm{BH}} = 0.142$), P02 ($\rho = 0.38$, $p_{\mathrm{BH}} = 0.235$), P08 ($\rho = 0.43$, $p_{\mathrm{BH}} = 0.235$), P14 ($\rho = 0.37$, $p_{\mathrm{BH}} = 0.245$)



(a) Boxplots of emotion dynamics of Continuous Annotation (Task 4) data, by Participant (*N*=16). Position ($M = 0.5465$, $SD = 0.2221$), Direction (0.0049, 0.7127), Inertia (0.7666, 0.1215), Instability (0.1316, 0.1086) and Variability (1.2165, 2.4079).

(b) Representative subset of label distributions: emotions-as-position (average position; *purple*), emotions-as-direction (direction; *blue*), Inertia (*magenta*), Instability (*green*), Variability (*red*). Note that longer gameplay results in more samples.

Figure 4.3: Comparison of summary statistics and histograms by emotion parameter.

4.4 Comparing Motion Characteristics of Emotion Dynamics

We next examined how various parameters computed on these time series might reveal differing insights. In this scope we included: signal *Position* (the prevalent standard, and following an “emotion-as-state” metaphor); *Direction* (drawing on an alternative metaphor for emotion as directional and changing); and Houben, Van Den Noortgate, and Kuppens’s [2015] three *Emotion Dynamics* (ED) parameters of *Inertia*, *Instability* and *Variability*. Our investigation included comparing these time series (original and computed) through summary statistics and histograms, all by participant.

4.4.1 Data Preparation

We further analyzed each participant’s Continuous Annotation[4] data by first partitioning the continuous self-report data into 500ms windows (window count $\mu = 1587.75$, $\sigma = 462.50$ by participant). Where window boundaries do not coincide with a logged data point, we imputed with the previous data point, turning our time-series into a higher-resolution but stepped signal.

We computed *Position* labels from windows by mean value; and *Direction* labels as the rate-of-change per minute from a least squares linear fit, in the form of an angle $\theta \in [-\pi/2, \pi/2]$. Using R’s *psych* package [80], we calculated *Inertia* (autocorrelation coefficient), *Instability* (Root Mean Square of Successive Differences (RMSSD)) and *Variability* (Standard Deviation (SD)) by window for each participant [44].

4.4.2 Comparing Summary Statistics and Histograms by Parameter

Figure 4.3a shows signal statistics for each participant and parameter. The means for all five measures track closely across participants. However, spread differs: *Inertia* is relatively tight and symmetric, *Variability* is broad and highly asymmetric, *Instability* in between.

[4] Tests for equivalence between the two sets of self-report (CA and TWCW) across each of the three emotion dynamics parameters (two 1-tailed paired samples t-tests [57] per dynamic measure) were inconclusive ($p > 0.5$, $t(15) \ll 0.001$, $d \ll 0.001$). Subsequent emotion dynamics explorations were done on the higher resolution CA data.

In an alternative view, Figure 4.3b shows the same parameters and signals, but now as histogram distributions. Data for these four participants are reasonably representative.

Comparing these two representations of the same underlying data is insightful. For example, while in Figure 4.3a *Position* is clearly less stationary than *Direction*, 4.3b indicates the form that this takes (broader spread, spikiness). And while the dominating feature of the other three ED's boxplots is the uniformity of means across participants, histograms reveal their internal parameters as starkly different: *Inertia* is broad and high-valued, the others low-valued with very long tails.

No insight was gained from visual analysis of spectral qualities (from a Fast Fourier Transform) of all five parameters.

4.5 Summary

The preceding section's results demonstrate that the relatively high resolution of the CA report (30Hz raw, parameters computed at 2Hz) affords computation of a variety of descriptive parameters. We investigated what information can be extracted from different self-report methods, and whether or not self-reports are interchangeable. Our findings suggest that although both continuous annotation and interview methods yield different emotion time series, both seem to be appropriate responses to the same stimulus. Getting to the root of what the differences in label representation mean will require approaches assisted by synchronized physiological data views.

We finish this chapter with the question: **which self-report pass is best for computational emotion modelling?** Before exploring answers, we analyze FSR data as a viable input for emotion models (Chapter 5). Then, in Chapter 6, we return to the question, evaluating computational models using the different self-reports.

Chapter 5

Connecting Keypress Force to Multi-Pass Emotion Self-Reports

To summarize where we are, the outputs of the protocol described in Chapter 3.1 are player specific physiological and activity streams (EEG and FSR), emotion word calibrations, and the TWCW and CA – two time-series of emotion self-report annotated on the same dimensional plane of the Stressed-Relaxed scale over the gameplay timeline.

In this chapter, we model emotion with FSR data, describing our data featurization approach and analysis pipeline. We focus on understanding the viability of computational emotion modelling using manual pressure data, and compare results to a more largely studied affective modality–brain activity.

5.1 Feature Extraction

FSR data from keypress force has an average of < 1 distinct keystrokes in the same window length[1], disqualifying the use of deep learning models. For this dataset, we craft new features (a process known as feature engineering) and perform model selection with classical machine learning models. We extract features using keystroke activity, frequency, and statistical analysis, generating data instances of the same duration employed in brain activity analysis (0.5s, 1s, 2s, and 5s).

[1] According to recordsetter.com, the world record is 886 distinct keystrokes per minute!

5.1.1 Data Preparation

Estimating keypress force data with Force Sensing Resistors introduces signal jitter. To mitigate signal noise while maintaining the overall shape of a keystroke, we apply an Exponentially Weighted Moving Average (EWMA) [46], with smoothing factor $\alpha = 0.5$.

In addition, we aggregate the keypress activity in the game to inspect how pressure alone informs emotion expression. We generate two additional channels as '**composite keys**' by (1) taking the sum of force over all five keys and (2) selecting the max force from all keys, resulting in a total of 7 channels (composite keys "**A5**" and "**A6**" plus the original five).

5.1.2 Windowing

Each time series (FSR data stream and emotion self-reports) is divided into non-overlapping, consecutive, equal duration segments, or windows. Modality-specific features and the associated emotion class calculated from the data points contained within a window's boundaries constitute a data instance to be classified. We experimented with windows of duration w where $w = \{0.5s, 1s, 2s, 5s\}$ in order to assess the effect that window duration has on emotion classification accuracy, recognizing that the choice of window duration limits the lowest possible detectable frequency. Table 5.1 summarizes the count of data instances by window size.

Window size	Mean	Std.
500ms	4834	876
1000ms	2405	526
2000ms	1203	263
5000ms	544	225

Table 5.1: Aggregated count of data instances by window size.

We set the minimum duration window to 0.5s and the maximum at 5s to cover the range of emotion duration of "a few seconds" [101, 102]. One and two second windows match other emotion-related classification studies [15, 60, 104].

5.1.3 Frequency and Statistical Features

Based on previous studies of emotion expression of social touch pressure [15, 50], we calculate a set of descriptive statistics for each window of pressure data–minimum, maximum, variance, mean, area under the curve, and sum of absolute differences. Using Fast Fourier Transform (FFT) and Hamming windows of the same lengths as the time domain windows, we also calculate the most prominent frequency (amplitude and frequency bin), amplitude variance, amplitude mean, and peak count in the frequency spectrum [15].

5.1.4 Keystroke Features

While gameplay is controlled via computer keyboard, participants are not typing but rather activating keys based on gameplay context. Elements of keystroke dynamics–such as travel time between keys–are less relevant as features in this case, so we calculate affective touch features highlighting keystroke fluctuations in force and duration in the time and frequency domains [15, 50]. In addition, we borrow select parameters related to the Attack, Decay, Sustain, and Release (ADSR) envelope (illustrated in Figure 5.1), commonly employed in synthesizers to describe different stages of sound as produced by a piano keyboard.

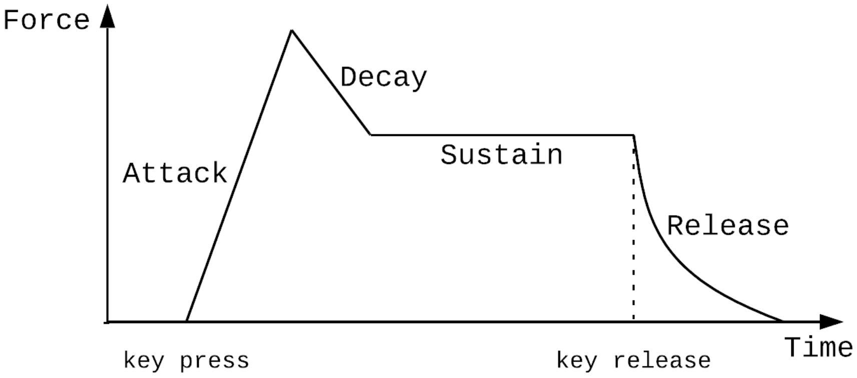


Figure 5.1: Attack, Decay, Sustain, and Release (ADSR) envelope.

After sectioning the data in time windows, we calculate the following parameters for each keystroke in a window: keystroke duration (in ms), peak count, amplitude of maximum peak, time from keystroke start to maximum peak, time from maximum peak to key release, force variance, average force, and area under the keypress curve. We aggregate each of these parameters by taking the mean in a data window. Table 5.3 summarizes all features extracted from FSR data.

Although the windowing strategy is simple to implement, it can cause keypress distortion–i.e., when a keystroke crosses window boundaries, creating spurious peak and release points, and misleading summary statistics. Figure 5.2 shows an excerpt of windowed keypress force activity from a key and illustrates an instance of keypress distortion. For the purposes of multi-modal window alignment and the simulation of real-time application of emotion classification on keypress force, we elect to follow a uniform data windowing pipeline while recognizing this limitation. Table 5.2 summarizes the count of occurrences of keypress distortion by window size.

Window size	Mean	Std.
500ms	0.25%	0.75%
1000ms	0.51%	1.52%
2000ms	1.00%	2.95%
5000ms	0.84%	2.66%

Table 5.2: Percentage of occurring keypress distortion in relation to all FSR data instances, aggregated by window size.

	Features
Statistical features	minimum, maximum, variance, mean, area under the curve, and sum of absolute difference
Frequency features	most prominent frequency (amplitude and frequency bin), amplitude variance, amplitude mean, and peak count in the frequency spectrum.
Keystroke dynamics and ADSR features	keystroke duration (in ms), peak count, amplitude of maximum peak, time from keystroke start to maximum peak, time from maximum peak to key release, force variance, average force, and area under the keypress curve

Table 5.3: Summary of features extracted from FSR data.

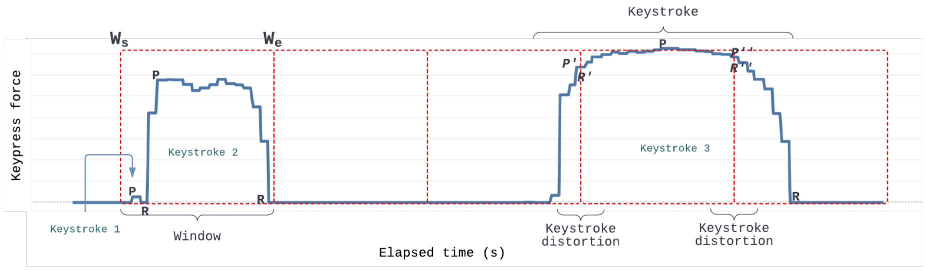


Figure 5.2: Example of keystroke force as captured by a key-mounted FSR sensor. W_s and W_e indicate the start and end of a window (dashed boundary). P represents the maximum peak force during a single keystroke (delineated by zero force), and R represents the release point. Distortion occurs when keystrokes span across multiple windows, creating spurious peak and release points (P' and R', P'' and R''), and distorted representations of keypress activity in a window.

5.2 Understanding Affect in FSR Data

In studies focusing on understanding how emotion can be detected implicitly (i.e., from data channels observed in real-time without voluntary effort from the user, such as through physiological or activity monitoring), brain activity–in particular, as estimated by EEG–is well represented. Keypress force–touch pressure, more generally–forms a significantly smaller subset of computational emotion modelling research [97, 112]. In this section, we compare FSR models on with those built on EEG to investigate the viability of keypress force as an affective modality to capture emotion expression.

An emotion trajectory or path traced through an emotion space [12] suggests a time-series representation of emotion ratings. By windowing the data, we can create data instances on brain activity and/or keypress force modality and generate distinct emotion label sets by metaphor. We vary window size, affect expression modality, label metaphor (position, direction), and report accuracy from classical machine learning and deep learning models. In this section, we describe these experimental factors, and explain their relevance for real-time application. To accommodate high individual variability in brain activity and emotion calibration, all models are built to focus on individualized models where test and training sets are from the same individual (subject-dependent models). Details about preprocessing and model implementation using brain activity data can be found in 3.2 (Figure 3.4) and 5.2.5.

5.2.1 Labels Sourced from Continuous Annotation

Due to the quantitative nature of the continuous annotation (and its representations as label metaphors), we can directly use CA self-report data to build ML models based on real-time physiological data. To allow for comparison between implicit real-time modalities, we follow the same label extraction procedures from CA self-report used for models built on brain activity data, which are described in detail in Chapter 4.1 and illustrated in Figure 5.3.

5.2.2 Classification Pipeline

For the FSR data, we perform grid search cross-validation (k= 5) to select the best fit (per participant) among seven classical machine learning models, comparing performance across linear, ensemble, and boosted models (see Table 5.4 for a list of models and associated parameters tuned through grid search). Models here were developed in Python using SciKit Learn and XGboost[2]. Given the high dimensionality of our feature set ($d = 82$ features), we employ a zero variance threshold – removing all features with zero variance (constant features) – and recursive feature elimination (RFECV) [35] with cross-validation (k= 5) to select features. We run 30 iterations of this pipeline, randomizing our training and validation sets at each iteration. Figure 5.4 illustrates the overall pipeline. Source and preprocessing steps of emotion labels are described in Chapter 4.1 and illustrated in Figure 5.3.

[2] https://github.com/dmlc/xgboost/

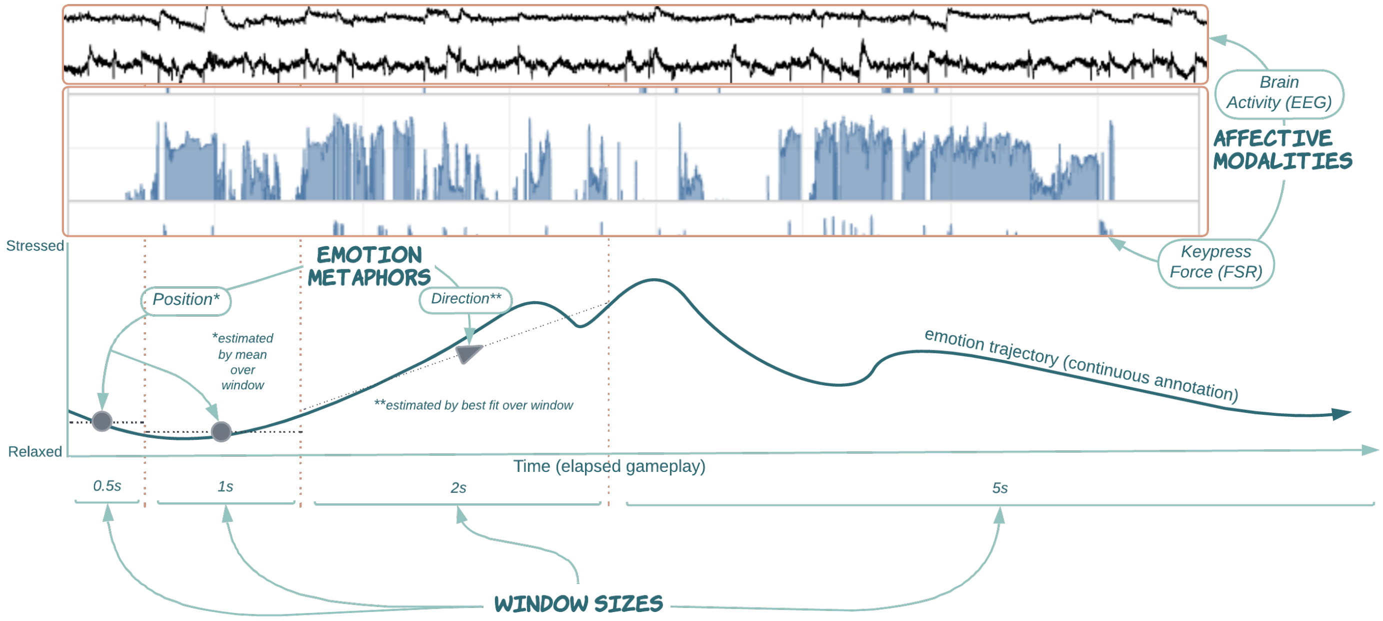


Figure 5.3: Analysis factors for experiments on continuously annotated emotion experience estimated by dynamic emotion metaphors. We inspect models built on two affective modalities: brain activity (as estimated by EEG) and keypress force (FSR), using two emotion label metaphors of direction (best fit line) and position (window mean), and trained on four window sizes (0.5s, 1s, 2s, 5s).

	Parameters
ExtraTreesClassifier	n_estimators: (16, 32),
RandomForestClassifier	n_estimators: (16, 32),
AdaBoostClassifier	n_estimators: (16, 32),
GradientBoostingClassifier	n_estimators: (16, 32), learning_rate: (0.8, 1.0),
XGBClassifier	max_depth: (4, 6, 8), min_child_weight: (1, 5, 10),
LogisticRegression	penalty: (none, l1, l2, elasticnet), C: (0.001, 0.01, 0.1, 1, 10), class_weight: (balanced, None), solver: (newton-cg, sag, saga, lbfgs), l1_ratio: (0.0, 0.3, 0.5, 0.7, 1.0)
SVC	kernel: (linear, rbf), C: (0.001, 0.01, 0.1, 1, 10), gamma: (0.001, 0.0001)

Table 5.4: Estimators and parameters used for tuning keypress force models to each participant.

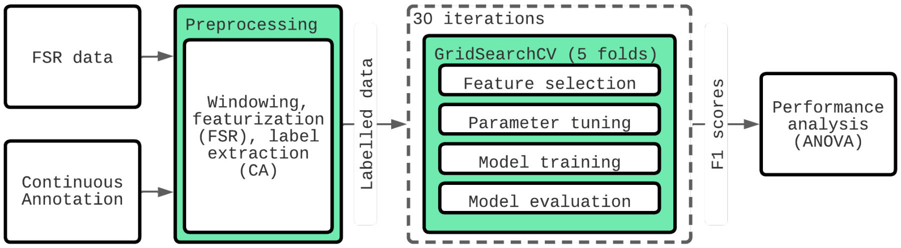


Figure 5.4: Overview of the experiment setup used to evaluate emotion models built on FSR data.

5.2.3 Classification Performance of Models Built on FSR Data

We evaluate classification of FSR models using 5-fold Cross-Validation (CV), where the training set consists of 80% of the data instances with the remainder forming the validation set. Chance is dependent on an individual's class distribution by label metaphor and windowing; in order to compare performance across conditions, we report F1-score over chance (i.e., , diff(F1-score, chance F1-score)) such that performing 'at chance' results in a score of 0.

On FSR data, we report mean cross-validated F1-scores over chance from the best performing algorithm by participant by running 30 iterations of grid search with 5-fold CV with the predefined set of ML models.

We violate Levene's test for equality of variances for the present analysis,

$F(11,180) = 0.293$, $p = 0.987$. Owing to this assumption violation, we chose to subject FSR chance-offset F1-scores to a two-way Aligned Rank Transform Analysis of Variance (ART ANOVA) [107], having three types of labels (direction, position, and accumulator) and four sizes of windows (0.5s, 1s, 2s, 5s). All effects are statistically significant at the 0.05 significance level. The main effects of label type and window sizes, and interaction effect yield F ratios of $F(2,180) = 47.793$, $F(3,180) = 18.003$, $F(6,180) = 4.468$, $p < 0.001$, respectively.

We ran post-hoc tests using Holm correction to further investigate the individual mean differences, significance at $p_{\text{Holm}} < 0.001$ unless otherwise indicated. Results show that the mean F1-score was significantly greater between accumulator and direction scores, as well as position and direction scores, but not at accumulator and position ($p_{\text{Holm}} = 0.671$). On window sizes, post-hoc tests indicate that the mean F1-score was significantly greater between 5s windows and all other window types, and 2s and 0.5s, but not at other sizes (1s-2s and 1s-0.5s, $p_{\text{Holm}} > 0.07$. Finally, on interaction effect, post-hoc results indicate that the combination of accumulator (or position) and 5s windows perform better than direction at both 0.5s and 1s ($p_{\text{Holm}} < 0.001$ and $p_{\text{Holm}} < 0.05$, respectively).

To summarize test results of greatest relevance to our investigation:

(1) FSR data with 5s windows perform better than all other window sizes.

(2) Accumulator and position label types perform better than direction; though at 5s windows all three perform similarly.

5.2.4 Feature Evaluation

To better understand how different FSR features inform classification, we analyze results from Recursive Feature Elimination (RFE) with CV on the feature set of all models, pulling out the upper quartile (top 25%) of features ranked by importance. We report how often these top features are selected, grouped by key and feature type as in Figure 5.5.

We group features by type: "FSR pressure behaviour" features describe the how much pressure is applied to a key when the key is pressed; "aggregate pressure

[2] Implemented using R's ARTool [51]

statistics" describe the pressure force applied to a key, including moments when the key is inactive; "FSR time behaviour" features describe time-related features; finally, "frequency features" (not shown in Figure 5.5) indicates features calculated in the frequency spectrum.

Our analysis indicates that FSR features extracted from the right directional key–the key used to move the character forward through the game–are most important. Particularly, those describing the *area under the curve*, as well as *max, mean and variance* of pressure in the keypress. The only four features that are not directly pressure-related features that make the top ranked important features, are also from the same right directional key. These time-based features are that of *keystroke duration, pressure peak count, time from peak pressure to key release and time from press to peak pressure*.

From the composite keys (defined in Chapter 5.1.1), we find that the pressure statistics (*max, mean, min, total variance, variance, and area under the curve*) are important for A6 – max force over all keys – followed by the same set of pressure statistics for A5 – the sum of the pressure across all keys.

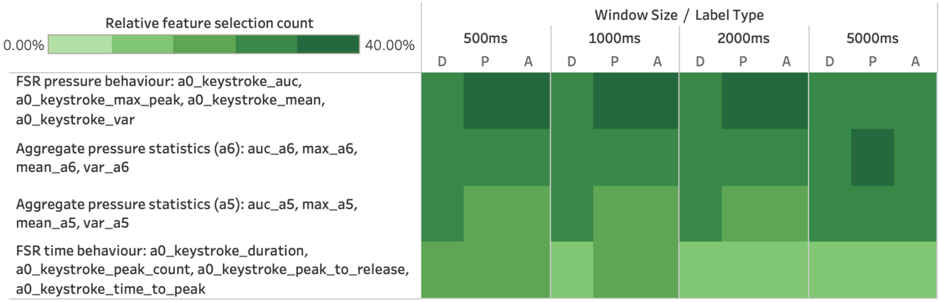


Figure 5.5: Relative feature selection count per window size and label type (D: direction, P: position, and A: accumulator). Darker tone indicates most commonly selected features in personalized models that perform above-chance.

5.2.5 Comparing Models Built on FSR, EEG, and Both

The EEG data is dense but noisy; the FSR data is more stable but sparse–how do these modalities compare to each other in terms of classification performance?

Could integrating modalities offer the best of both to improve performance? The dimension size difference between the data streams prevents a simple feature concatenation of both modalities. We trained two models separately, one for keypress force on FSR data another for brain activity on EEG, and test using soft ensemble voting (sum the probabilistic output and choose the label with the highest sum) to output a single classification label.

We ask if integrating modalities would produce classification accuracy over and above models built on single modality data. We built a Soft Voting Classifier by selecting the highest performing model for each modality. The class with the largest probability after summing is the prediction class. As described in 3.2, we use a CNN for EEG data. For FSR, we build models with the XGBClassifier, which we train on the best performing pressure statistic features calculated over the two composite keys. We report the F1-score over chance in the final row of Figure 5.6.

Classification Performance

We compare chance-offset F1-scores over all four modalities – EEG (LSTM, CNN), keypress force (best fit model), and multimodal (combined CNN and XGBClassifier). The data violates Levene's test for equality of variances so we chose to subject all chance-offset F1-scores to a three-way ART ANOVA, with three types of labels (direction, position, and accumulator), four sizes of windows (0.5s, 1s, 2s, 5s), and four model-modality sets. All effects were statistically significant at the 0.05 significance level. The main effects of label type, window sizes, and modalities yield F ratios of $F(2,720) = 108.920$ ($p < 0.001$), $F(3,720) = 893.344$ ($p < 0.001$), $F(3,720) = 3.026$ ($p < 0.05$), respectively. All interaction effects were also significant, with exception of the three-way interaction among all factors ($F(18,720) = 0.797$, $p = 0.705$).

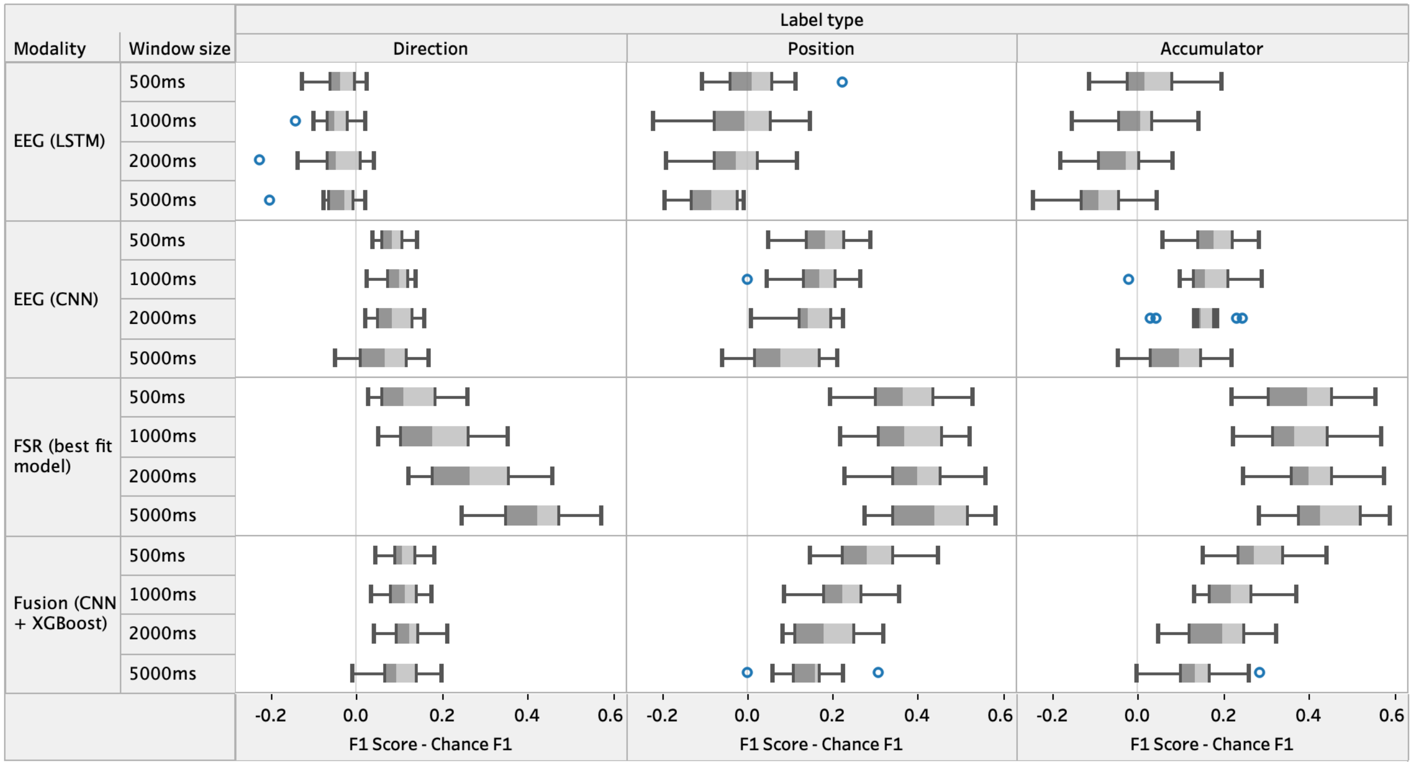


Figure 5.6: Comparison of FSR, EEG, and multi-modal classification results on CA by window and label type.

We ran post-hoc tests using the Holm correction to further investigate the individual mean differences, significance at $p_{\text{Holm}} < 0.001$ unless otherwise indicated. Results show that the mean F1-score was significantly greater between accumulator and direction scores, and position and direction scores, but not at accumulator and position ($p_{\text{Holm}} = 0.583$). On window sizes, post-hoc tests indicated that the mean F1-score was significantly greater between 5s windows and 0.5s ($p_{\text{Holm}} < 0.01$), but identified no significant difference among other sizes ($p_{\text{Holm}} > 0.03$). Among modalities, keypress force performs better than all other strategies, multimodal second, brain activity (EEG CNN) third, and brain activity (EEG LSTM) last. On interaction effects, post-hoc results indicate that the accumulator (or position) perform better than direction at all window sizes ($p_{\text{Holm}} > 0.001$).

In summary, results suggest that **the best performing classifier built on FSR data outperforms the deep learning EEG models**. When looking over the models we personalize, we note that for every participant, there is at least one emotion classifier for FSR that outperforms the EEG-based deep learning models.

5.3 Summary

In this chapter, we analyzed keypress force estimated by FSR as an alternative to EEG data for input to computational models of emotion expression. We varied window size and label metaphor (position, direction), and reported accuracy from classical machine learning and deep learning models. Figure 5.6 summarizes all results discussed in this chapter, presented by affective modality (left column), window sizes (0.5s to 5s), and label type as extracted from CA self-report. On emotion-as-direction, we notice an upward trend when comparing models built on FSR data: increasing window sizes lead to better performance. The opposite is true for CNN models built with EEG data: smaller windows yield better performance. We hypothesize that this difference is tied to how emotion manifests in each modality–fast changes in EEG data follow complex brain activity around emotion processing, while slower changing touch comes through with FSR data as a result of internal emotion regulation. Last, the best performing classifiers for emotion-as-position built on FSR data outperform EEG, FSR, and fusion models built.

In summary, our findings suggest that (1) FSR-based models outperform EEG and multi-modal strategies; (2) emotion expression in FSR data is better captured at larger window sizes; (3) pressure statistics are more informative than time-based features for emotion modelling with keypress force.

Chapter 6

Which Self-Report Pass is Best for Computational Emotion Modelling?

In Chapter 5, we compared models for implicit emotion expression using FSR data with those using a more studied signal–EEG data, finding that keypress force outperforms brain activity in emotion modelling using a single reporting pass (CA). In addition, results from Chapter 4 suggest that while the self-reporting passes are not interchangeable, they do seem to capture responses to the same stimulus.

To understand how the different labelling passes affect computational modelling of emotions, we compare models built on FSR data using CA, TWCW, or a combination of both. Creating computational models on these disparate self-report modalities requires special considerations. In this section, we add TWCW self-reports to the experimental pipeline introduced in the previous chapter (illustrated in Fig. 6.1), and discuss the strategies for operationalizing multi-pass emotion labels in ML models. Then, we compare results of each strategy evaluated on FSR data.

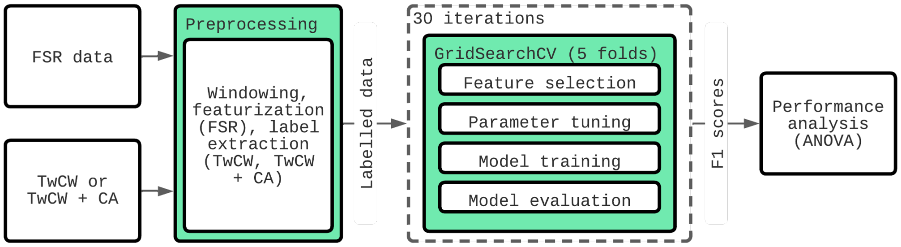


Figure 6.1: Overview of the modified experiment setup, first presented in Chapter 5, used in all forecasting schemes using TWCW or a combination of TWCW and CA.

6.1 Using Only Timeline with Calibrated Words

We assess the use of TWCW as labels in comparison to emotion-as-position from CA. As illustrated in Figure 6.2, multiple calibrated words may appear within one time window. To create a one-to-one label mapping with window instances, we consider (a) the mode of TWCW within a window (i.e., represent window with most used word); (b) a list of words present in a window and use multi-label classification.

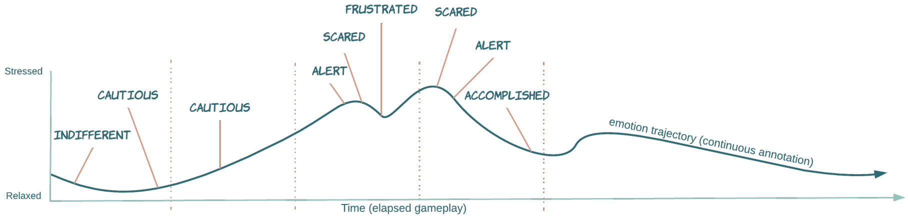


Figure 6.2: Example of timeline with calibrated words.

Single-label TWCW *classification:* Similarly to the experiment pipeline described for CA, we perform grid search cross-validation ($k = 5$) to select the best fit (per participant) among seven classical machine learning models, comparing performance across linear, ensemble, and boosted models (see Table 5.4 for a list of models and associated parameters tuned through grid search). Models here were developed in Python using SciKit Learn and XGboost[1]. Given the high dimensionality of our feature set ($d = 82$ features), we employ a zero variance threshold – removing all features with zero variance – and RFE with cross-validation (k= 5) to select features.

Flat multi-label TWCW *classification:* How can we address mixed-emotions in classification? To answer this question, we propose a flat multi-label model using a list of calibrated emotion words as they occur in a window (illustrated by Figure 6.3). Using Scikit Learn's MultiOutputClassifier[2], we extend the models used in the pipeline described above to allow for multi-label outputs.

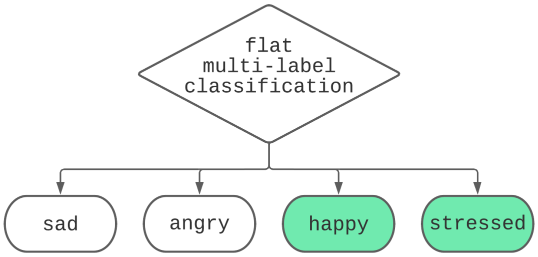


Figure 6.3: Illustration of flat multi-label classification output with calibrated words (outputting, for instance, the labels "happy" and "stressed").

6.2 Combining TwCW and CA

The results described in Chapter 4.3 suggest that each labelling pass provides distinct perspectives on what the participant might be experiencing at each instant of the gameplay. We leverage the information contained in different self-report passes by combining both TWCW and CA into one label pipeline.

6.2.1 Hierarchical Multi-label Classification

Where am I at with my emotions, and where am I heading to? We propose a hierarchical multi-label model using the mode of calibrated emotion words as they

[1] https://github.com/dmlc/xgboost/

[2] https://scikit-learn.org/stable/modules/generated/sklearn.multioutput.MultiOutputClassifier.html

occur in a window (illustrated by Figure 6.3), combined with the direction information extracted from continuous annotation. Using the package HiClass [69], we extend the models used in the pipeline described above to allow for hierarchical multi-label outputs.

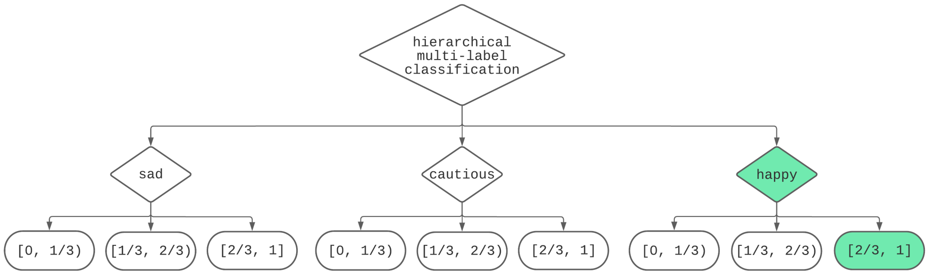


Figure 6.4: Example of hierarchical multi-label classification with calibrated words and continuous annotation (represented by direction metaphor).

We train a classifier for each parent node. For instance, for the structure illustrated in Figure 6.4, we train a multi-class model that outputs the emotion words *sad*, *happy*, and *cautious*. Subsequently, we train a classifier trained for each emotion word, outputting binned direction values (as described in Chapter 4.1).

6.2.2 Extracting Label Alignment

Multiple labelling passes also create inconsistencies. As illustrated in Figure 6.2, emotion words might be associated to incongruous values of continuous annotation, when taking into account the word's calibrated value. We propose two strategies based on an alignment metric (a_m) of a calibrated emotion word: label resolving and instance weighting. Each instance in the dataset receives a score as calculated by:

$$a_m(\mathrm{CW}|\mathrm{CA}) = 1 - d(\mathrm{CW}, \mathrm{CA})$$

We introduce $d(\mathrm{CA}, \mathrm{CW})$ as a penalty that considers both temporal resolution of emotions and disagreement between emotion word calibration values and the current CA score. We bound the distribution of a_m to [0, 1] range by taking calculating $d(\mathrm{CA}, \mathrm{CW})$ as the normalized Euclidean distance between a calibrated word from TWCW and the CA curve:

$$d(\mathrm{CA},\mathrm{CW}) = \sqrt{\left(\frac{t_{CA}}{T} - \frac{t_{CW}}{T}\right)^2 + \frac{(CA - CA_{CW})^2}{CA_{max}^2}} \cdot \frac{1}{\sqrt{2}}$$

where t_{CA}, t_{CW}, CA, CA_{CW} are the timestamps and observed values of CA and TWCW, respectively, T represents the length of a window (in seconds), and CA_{max} represents the maximum value that both CA and CA_{CW} can assume (in this case, 1). To ensure $d(\mathrm{CA},\mathrm{CW}) \in [0,1]$, we multiply the Euclidean distance, illustrated in Figure 6.5, by a normalizing factor of $\frac{1}{\sqrt{2}}$.

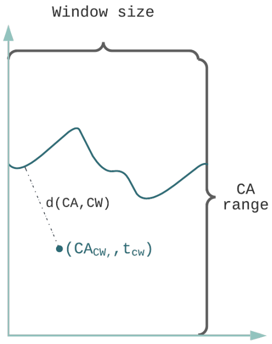


Figure 6.5: Illustration of $d(CA,CW)$.

Resolving Label Inconsistencies: The a_m scores calculated in the previous step can be interpreted as the likelihood of an observed data instance being correctly labelled. We model this information by using this set of a_m scores as a guide for when to keep or discard labels in a window. When a high disagreement occurs between CA and TWCW ($a_m(\mathrm{CW}|\mathrm{CA}) <$ chance), we eliminate the calibrated word from that window. To perform classification, we repeat the multi-label framework described in Section 6.1.

Instance Weighting: The same scores calculated in the previous step can also be interpreted as the likelihood of an observed data instance being correctly labelled. We model this information by using this set of scores as weights for each instance during training. When a high disagreement occurs between CA and TWCW, $a_m(\mathrm{CW}|\mathrm{CA}) \to 0$, eliminating the observed window from training. Conversely, a high agreement score between CA and TWCW amplifies the influence of a given data window during model training. We repeat the hierarchical multi-label framework providing $a_m(\text{CW}|\text{CA})$ as sample weights during training.

We compare performance on keypress force (best fit model) over classification schemes described in Chapter 6.1 using TWCW or a combination of TWCW and CA, separating by emotion-as-position and emotion-as-direction, respectively.

6.3 Comparing Performance for Emotion-as-Position Strategies

We compare chance-offset F1-scores from emotion-as-position models trained using FSR data. Given that the score distribution violates Levene's test for equality of variances, we chose to subject all chance-offset F1-scores to a two-way Aligned Rank Transform Analysis of Variance (ART ANOVA)[107], comparing five types of labels (from CA only: position, accumulator; from TWCW only: mode of TWCW in window, multi-label; combining CA and TWCW: multi- soft-labels) and four sizes of windows (0.5s, 1s, 2s, 5s). Both main effects were statistically significant at the 0.001 significance level. Table 6.1 summarizes the results. We ran post-hoc tests using the Holm correction to further investigate the individual mean differences, significance at $p_{\text{Holm}} < 0.001$ unless otherwise indicated. Results show that the mean F1-score was significantly greater at classification schemes using TWCW than using CA, with no significant distinction between TWCW models. On window sizes, post-hoc tests indicated that the mean F1-score was significantly greater between 5s windows and 0.5s, and 1s ($p_{\text{Holm}} < 0.01$), but no significant difference among other sizes ($p_{\text{Holm}} > 0.02$).

	df	*df_{res}*	*SS*	*SS_{res}*	*F*	*p*	η_p^2
Label type	3	240	267,037.375	1,130,707.250	**18.893**	< **.001**	.191
Window size	3	240	97,687.281	1,300,028.875	**6.011**	< **.001**	.070
Label type × Window size	9	240	5,869.156	1,391,930.000	0.112	.999	.004

Table 6.1: Analysis of Variance on Aligned Rank Data (Type III) on emotion-as-position results.

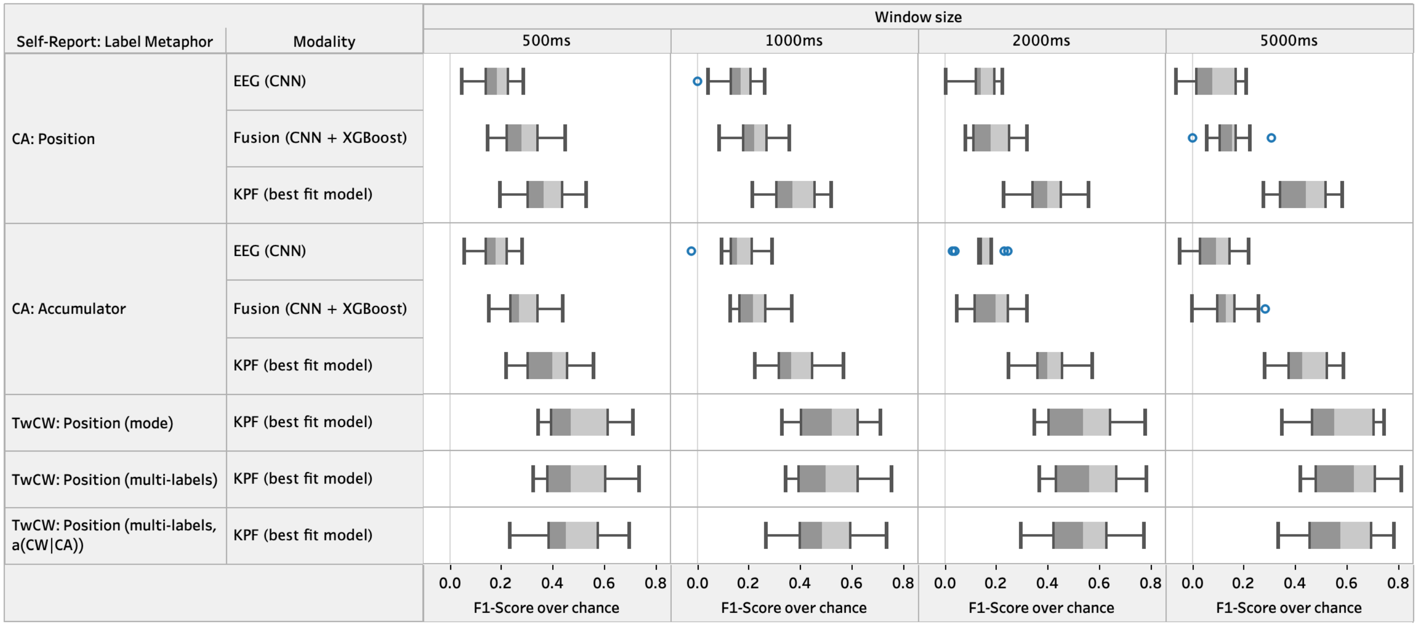


Figure 6.6: Comparison of single- and multi-pass classification results on emotion-as-position using CA and TWCW, reported by window and label type. Single-pass CA scores correspond to the first the top two groups of EEG, FSR and fusion results, followed by single-pass TWCW implemented as single- and multi-label models. Finally, the last row indicates models built using information extracted from both self-reports.

We summarize results in Figure 6.6, comparing label metaphors computed using the different self-report passes (left column), affective modality, and window sizes (0.5s to 5s). We notice no significant difference among the models that use TWCW (last three rows). The best performing classifiers for emotion-as-position built on FSR data with TWCW outperform EEG, FSR, and fusion models built on CA only.

6.4 Comparing Performance for Emotion-as-Direction Strategies

We analyze emotion-as-direction chance-offset F1-scores using a two-way ANOVA, with three types of labels – direction using CA, emotion word and direction labels obtained with hierarchical classification, and emotion and CA direction labels with hierarchical classification and instance weighting – and four sizes of windows (0.5s, 1s, 2s, 5s). As summarized in Table 6.2, both main effects were statistically significant at the 0.001 significance level. The interaction effect between label type and window size was also significant at $p < 0.05$. We ran post-hoc tests using the Holm correction to further investigate the individual mean differences, significance at $p_{\text{Holm}} < 0.001$ unless otherwise indicated. Results show that the mean F1-score was significantly greater when combining TWCW and CA self-reports than when using CA alone, but not between the different approaches combining TWCW and CA. On window sizes, post-hoc tests were consistent with previous results: F1-scores were significantly greater at larger windows, but no significant difference between 1s and 0.5s ($p_{\text{Holm}} > 0.16$).

	df	*SS*	*MSE*	*F*	*p*	η_p^2
Label type	2	3.011	1.506	**188.369**	< **.001**	.702
Window size	3	0.949	0.316	**39.594**	< **.001**	.426
Label type × Window size	6	0.109	0.018	2.275	.039	.079
Residuals	160	1.279	0.008			

Table 6.2: Analysis of Variance on emotion-as-direction results.

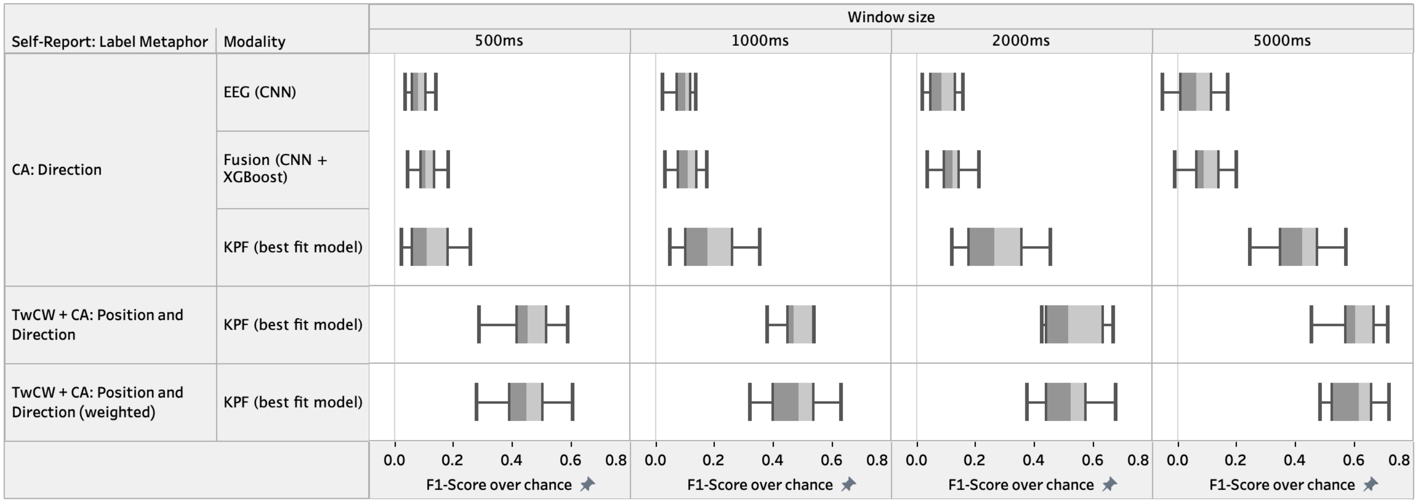


Figure 6.7: Comparison of single- and multi-pass classification results on emotion-as-direction using CA, TWCW, and a combination of both, reported by window and label type. Single-pass CA scores correspond to the first the group of EEG, FSR and fusion results, followed by combined self-reports models, implemented as hierarchical multi-label tasks. The last row indicates models built using the instance weighting strategy.

We summarize results in Figure 6.7, comparing label metaphors derived from the different self-report passes (left column), affective modality, and window sizes (0.5s to 5s). Similar to emotion-as-position results, we notice no significant difference among the models that use TWCW (last two rows). We hypothesize that window sizes are too short to capture significant differences in TWCW (sampled at 0.05Hz). Last, the best performing hierarchical classifiers for emotion-as-direction built on FSR data with TWCW and CA outperform EEG, FSR, and fusion models built on CA only.

6.5 Feature Evaluation

As a comparison with the feature evaluation procedure described for CA, we analyze results from RFE with CV on the feature set of the single-label model trained using TWCW, pulling out the upper quartile (top 25%) of features ranked by importance. We report how often these top features are selected, grouped by key and feature type as in Figure 6.8.

Consistently with previous results, keypress force features on the right directional key (used to move the character forward in the game) are most important, particularly those describing the *area under the curve*, as well as *max, mean and variance* of pressure in the keypress. The only four features that are not directly pressure-related features that make the top ranked important features, are also from the same right directional key. These time-based features are that of *keystroke duration, pressure peak count, time from peak pressure to key release and time from press to peak pressure*.

Similarly to models built on CA, we find that the pressure statistics (*max, mean, min, total variance, variance, and area under the curve*) from the composite keys are important for A6 – max force over all keys – followed by the same set of pressure statistics for A5 – the sum of the pressure across all keys.

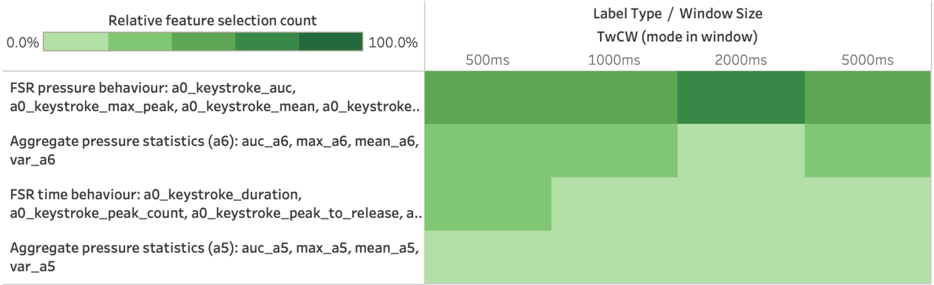


Figure 6.8: Relative feature selection count by window size in models using TWCW (single-label strategy). Darker tone indicates most commonly selected features in personalized models that perform above-chance.

6.6 Summary

In this chapter, we analyzed different schemes for including different self-report passes in FSR-based computational emotion models. We introduced five variations of models derived from FSR data that include TWCW self-reports as labels. Figure 6.9 summarizes all results presented in this chapter. We varied self-report pass (CA, TWCW), label metaphor (position, direction), window size, and reported accuracy from classical machine learning, comparing these FSR-based models to EEG-based deep learning models. We do not observe significant difference among the models that use TWCW (last five rows). As discussed in Chapter 6.4, we hypothesize that window sizes are too short to capture significant differences in TWCW. Overall, we observe increased performance when contrasting FSR-based models built on TWCW to those built using only CA. More investigation is required to understand how models built on EEG data perform when using TWCW as labels.

In summary, findings suggest that (1) emotion words are sparse, but offer better performance than continuous annotation when using FSR data; (2) multiple self-report passes used in combination lead to better performing FSR models.

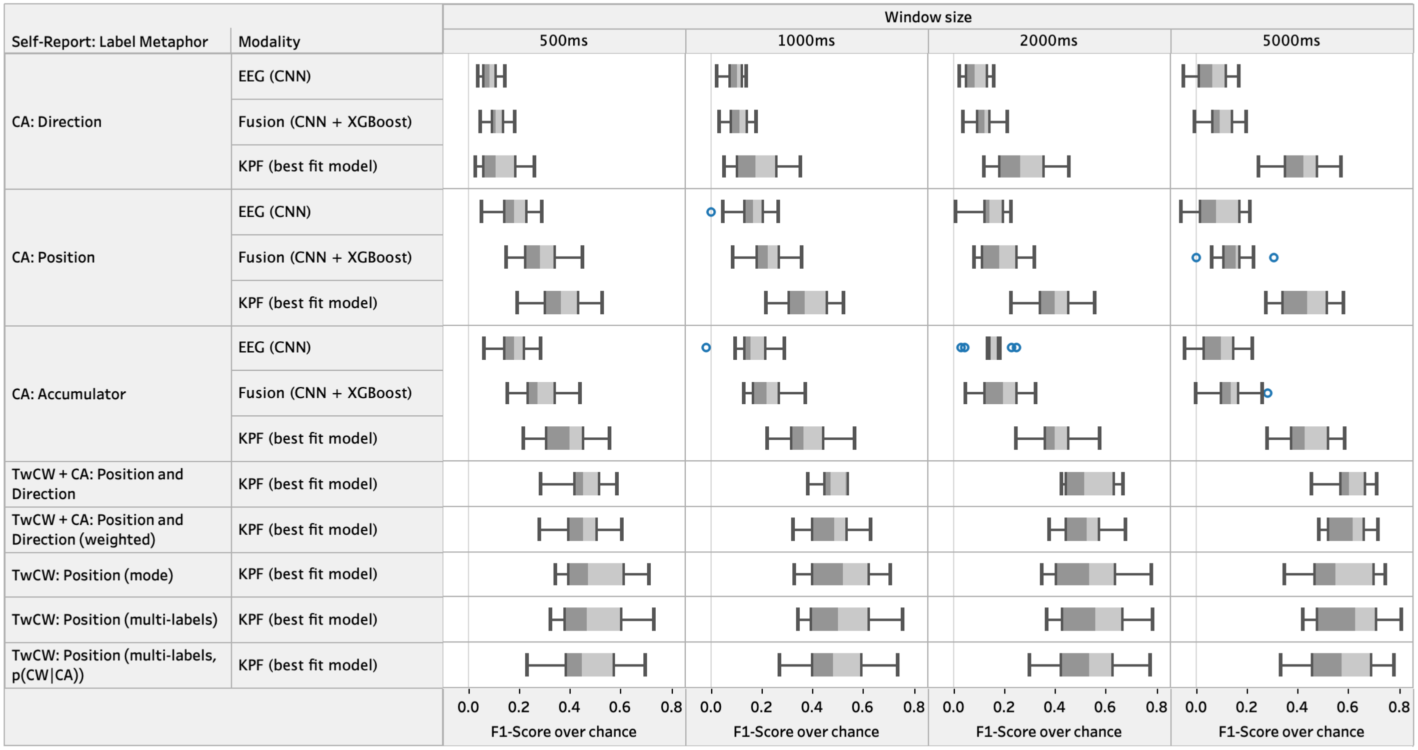


Figure 6.9: Overview of single- and multi-pass classification results on emotion-as-position and emotion-as-direction using CA and TWCW, reported by affective modality, window and label type.

Chapter 7

Reflections and Recommendations

In preceding chapters, we described analyses of data collected using the mixed-method labelling procedure described by Cang et al. [16], and extracted directional changes in emotion expression in real-time on keypress force as estimated by and modelled on FSR. Here, we summarize our high-level findings, root them in our research questions, and discuss the impact each has on the design of emotionally responsive devices. Finally, we reflect on the challenges faced during the modelling process, suggesting areas for improvement.

7.1 Re-Establishing Grounding Ideas

Model building requires making choices about assumptions we rely on, what, and how to include model, and what to conclude from model evaluation. We restate those assumptions for transparency.

What is a model? Models are representational tools, built to address a specific purpose. When combining both computational approaches and emotion theories, the term "model" becomes overloaded–it can mean either a *computational* model, algorithmically created to simulate and study complex systems or a *theoretical* model, or a framework created to structure, define concepts and explain phenomena. We disambiguate the term, using "models" solely to refer to computational emotion models.

What is emotion? There is no consensus of what defines emotions. We interpret emotions as a spectrum of possible reactions to internal and environmental stimuli, and operate under the assumption that the language used to describe and process emotions is highly subjective [9, 17, 97]. Consequently, we focus on building individualized models for each participant.

How do we evaluate emotion models? Our motivation behind creating computational emotion models is purely exploratory. We aim to understand whether computational models are able to capture implicit emotion in touch data. Since, at this stage, deployment is not a concern, we evaluate our models using statistical accuracy metrics.

What is our gold-standard for computational emotion models? Emotion expression in EEG data is well-studied, while implicit emotion conveyed through touch pressure takes a significantly smaller share of the literature [97, 112]. We use EEG-based models as a benchmark against which we evaluate performance of our touch-based models.

What is our gold-standard for annotating emotion data? Emotion labels can be generated automatically–by relying on predefined computational models built on a "gold standard" signal for emotion detection–observed by third parties, or self-reported [97, 112]. Given that we are focused on understanding the benefits of curated subjective emotion language in modelling implicit emotion expression, we rely on self-reporting of emotions. Specifically, we looked at a continuous report of emotion trajectory, drawn in a slice of Russell's circumplex [84] (stressed-relaxed). We further enriched these annotations by extracting notable emotion words from a participant-led recount of the emotion activity.

7.2 Real-Time Predictors of Dynamic Emotion: Keypress Force Outperforms Brain Activity

Personalized classification models built on FSR data perform better than the CNN models built on brain activity at all window sizes and label metaphors, as shown in Figure 5.6. Classification of emotion activity in brain activity performs better at higher frequency bands at the α, β, and γ bands [47, 104, 113], combined in the 8-50Hz range. Larger windows are more likely to capture lower frequency information, which is ideal for manual keyboard interactions for models of keypress force but of little benefit for higher frequency dependent brain activity models.

Exploit different time resolutions when creating multi-modal models.

Our simple multi-modal model, fusing estimates of brain activity and keypress force, is a demonstration of multiple classifiers running simultaneously. Here, we try a fusion model built from brain activity and keypress force modalities, which we capture in the same time window, but results in Figure 5.6 suggest that there is an argument to be made for a different approach. We propose that future work exploit the best window resolution for each modality in an interactive system. Employing clever voting strategies could be valuable for alerting when behaviours are widely different between distinct time scales – a fast and sharp hit interrupting slower stroking behaviour, for instance.

Manual touch pressure encodes valuable emotion content.

We used feature extraction techniques from a variety of affective touch interactions: keystroke dynamics in typing behaviour [64], pressure and location features from social touch [15], and ADSR features from sounds produced from a music keyboard [48]. Feature evaluation reveals that of the top 20 most important features from all three domains, 16 are pressure- or force-related. We know that increases in typing force is correlated with higher stress experiences [37], and that machine-mediated social touch [50] can be differentiated by variations in pressure. Now, we have evidence that emotion expression can also be captured through keypress force using a keyboard for videogame controls. More investigation is needed to assess other contexts where we might be subconsciously expressing emotion via touch pressure. In the meantime, we posit that the benefits to tracking pressure in devices where interactions feature manual affective touch far outweigh the cost.

7.3 Personalized Labelling: Trade-Offs

Compared to past studies of dynamic changes in behaviour or mood [44], the video game task which we used to elicit emotion is short and densely reported. With its data we reflect on our questions and protocol, highlighting implications for high-resolution real-time models.

7.3.1 How Do Triangulation and Personalization Enrich Emotion Self-Reports?

To estimate emotion evolution by-the-second, we can select a single dimensional emotion scale and collect self-reports (as in our CA data). How does adding scale calibration and a review/interview phase enrich this report stream?

Personalized scales clarify what may be generalizable, as well as improving personal models' accuracy.

Asking participants to project a set of emotions onto a specified emotion axis grounds the ratings in an individualized experience between the Stressed-Relaxed extremes. Plotting the ratings across commonly used words (as in Figure 4.2), we see that words with low rating variation – *Satisfied* and *Anxious* – may be useful as emotion reference frames. In contrast, high variance words like *Hopeful* or *Accomplished* may be less useful for labelling without additional interpretation.

Multi-pass reporting increases label versatility.

A continuous annotation of emotion communicates a highly personal experience at a resolution that is otherwise difficult to solicit. As a continuous quantitative signal, we can model emotion as a regression for high-resolution forecasting or elect to discretize (or bin values) for categorical classification. Additionally, we can compose an entirely new time-series by incorporating our personalized scale into an interview as a lower resolution signal where continuous annotation is impractical or unnecessary.

Disagreement may indicate synergy, not conflict.

Data from our two self-report passes (annotation and interview) are not sufficiently correlated to be interchangeable, yet causality results indicate they are highly related. Models combining both self-report passes outperform single-pass strategies, indicating that each has its own authenticity and value. Choices over which self-reporting pass is best could be optimized in protocol refinement to fit the application needs.

7.3.2 Incorporating Dynamics Into Computational Emotion Models

Reading signal characteristics (like auto-correlation, mean successive differences, variance) as measures of emotion inertia, instability and variability connects them to lived experience. What can they mean for intuitive predictive models?

Momentary emotion dynamics as characteristic, not label.

Inertia, instability, and variability can help uncover "slow emotion" in mood disorders [94], but they lose meaning in rapid-response timescales, and thus as emotion labels. Re-framed as signal statistics, they yield hints such as emotion variability's larger spread suggesting extra *sensitivity* (Figure. 4.3a) which could inform model development, e.g., by identifying archetypal behaviours for improved model selection.

An abundance of metaphors to fit the need.

The metaphor of "emotion-as-position" does not capture "fast" emotion dynamics. For example, *angle*, which captures relative differences in emotional intensity, has a natural physical meaning of directionality – *where I'm going*, not *where I am*. We have seen that *Inertia* and *Instability* respectively lend insight into responsiveness of emotion to stimuli, and emotive range.

Context may dictate choice of label metaphor. To identify if someone is *Excited*, we may choose a **position** representation; to *getting Sadder*, **direction** may work best. A **position** metaphor is more versatile; **direction** can be estimated from a set of points but the reverse requires additional information. Depending on the context of the application, a combination of both yields promising results.

The case for dynamic emotions: timing is everything.

If you had to choose between starting to regulate your emotions before you became angry and starting to do so after you rage has taken place, which would be tougher to manage?

When modelling human emotions, we may consider how the emotion space itself changes over time: when you feel sadder, it may be easier to get angry than calm, despite the fact that they are separated by equivalent Euclidean distances on the Affect Grid. In this situation, an emotion experience is a trajectory over a constantly changing landscape rather than a point [12].

Predicting directionality of an emotion trajectory allows for the creation of interventions that respond timely and appropriately. This is especially important, for example, when we are interested in delivering interventions to help the application user in regulating their emotions [89]. Different regulation strategies target emotional response at different intensities and cognitive processing stages.

7.4 Considerations in Building Effective Models for Dynamic Emotion Prediction

Choose or Fuse: is report divergence an opportunity?

Diverse self-reports capture perspectives that are authentic in different ways. We have inspected characteristics of emotion self-report in the time- and frequency-domains. Based on analysis insights, we might *choose* one approach, for its sensitivity or practicality. Or, we might *fuse* them, e.g., using discrepant moments as a spotlight on emotional conflict or low-confidence labels. We note that although our fusing strategies for resolving conflicts in CA and TWCW did not add benefits, classification incorporating both label streams performed overall better than other schemes. Emotion words are sparse when compared to feeltrace, but carry more consistent meaning.

Fusing yields better performance, but *choosing* might be more adequate to reduce label collection load. Next, we plan to evaluate choose-fuse strategies from the standpoint of applications. In which scenarios might we need to capture both label streams? How do we evaluate utility, interpretability, and usability of these unified labels?

Emotion metaphor and windowing influences classification performance.

From Figure 6.9 we see that the effect of window size is dependent on modality and label metaphor. For both position and accumulator metaphors, performance decreases as window size increases – except in high performing FSR models of keypress force where the opposite is true. Notably, models of keypress force using the CA and direction metaphor improve the most when window size increases.

As results and statistical analyses show, emotion label metaphors of position and accumulator behave very similarly (both distinct from direction); we continue discussion focusing on position and direction.

Over larger window sizes, a more stable trendline can be seen in the joystick labelled Continuous Annotation data. The direction metaphor, which we compose as the slope of the best fit line over the window, benefits most here since shorter window labels are less robust to outliers. For position labels, the larger windows can act as a 'smoothing' process, dampening valuable variation in emotion.

As expected, models using TWCW are more robust to increases in window span, although windows explored in this thesis are likely small to detect significant differences for an interview-style labelling pass with low sampling frequency (∼0.05Hz).

Short (0.5s) windows increase classification performance of dense-input models, such as the ones using EEG data, with emotion-as-position or accumulator labels calculated on CA. We note that label precision at this scale is unreasonable to obtain by any other means except labelling procedures that feature simple continuous annotation.

Modality capture:

The collection of this dataset was a very time-intensive and effortful procedure (Chapter 3.1). Setup and calibration of the EEG data collection system comprised a significant portion of the effort. Given how involved and invasive EEG data collection is, how noisy the data can be, and the classification performance relative to FSR in the same conditions, we are inclined to rely on keypress force or other manual touch data for emotion interaction data in future.

Labelling effort:

Collecting multipass emotion self-report also incurs a high time cost – the combination of personalized calibration, emotion elicitation task, interview, and continuous annotation would run a minimum of 3 to 4 times that of the emotion elicitation task. Where tasks run long, multipass reviewing procedures would be impractical unless annotation could happen contemporaneously and would not interfere with the natural evolution of the emotional experience.

Emotion elicitation and affect scale:

Calibrating an emotion scale finds shared meaning in common emotion words along the Stressed-Relaxed scale. While participants had personalized understandings of the measurement scale, they all engaged in the same emotion elicitation experience (here, a horror video game). Since we are building personalized models on individually defined emotion measurements, we look forward to fully committing to personalizing the entire experience. Where we want to understand the building of true personalized models that work “in the wild”, we will need models that are built on participant-defined emotion experiences that are able to evolve longitudinally. If the human emotional experience is ever-evolving, so should the measurement scales, training data and accompanying models. We plan future work to address how a data collection protocol can address evolution of emotion models over multiple data collection sessions.

Chapter 8

Conclusions and Future Work

We present an analysis of touch pressure data using self-reported emotion extracted from the FEEL dataset. We compare results from multi-modal data streams (brain activity and keypress force estimated by EEG and FSR data), collected during an emotional videogame play experience and labelled using multi-timescale and personally calibrated emotion labels [16].

Using 16 participants' data, we determine that this multi-pass labelling implementation adds **versatility** to collection options–one might choose to collect a single pass, depending on application requirements, different affect encoding modalities, or even propose different elicitation strategies; provides personalized and triangulated **insight into nuanced meanings of emotion**, and offers new options for **signal selection or integration**–models can be created from one or multiple affective modalities, and combined with respective emotion reporting techniques and metaphors to fit the application's needs. We **show how emotion dynamics measures and metaphors can add value**, in particular **emotions-as-position**–a mean value of the emotion reported on a personalized calibrated scale over a particular time window–and the **emotion-as-direction**–the slope of the best fit line over the data in the same time window.

In this chapter, we summarize concluding remarks and highlight future directions for this work.

8.1 Conclusions

This thesis demonstrates machine learning classification schemes on both sparse and dense self-reports. Comparing performance over factors of window size, feature set (features extracted from FSR or EEG data), and label set (Continuous Annotation (CA), Timeline with Calibrated Words (TWCW)), we find that, overall

1. Ensemble classifiers built on keypress force outperform deep learning models built on EEG features.

2. Feature evaluation of the FSR data reveals that pressure features used in machine-mediated social touch rank highest in terms of selection frequency, outperforming temporally-based features.

3. With dense, continuous self-reports, the best ensemble classifiers using keypress force perform comparably well on window sizes between 0.5s - 5s using a **emotion-as-position** label metaphor. Classifying **emotion-as-direction** label metaphor is best under the largest windows.

4. Models built with interview-based self-reports perform well across multiple window sizes, but are limited to emotion-as-position interpretations given the sparsity of available emotion labels.

5. A combination of both dense and sparse reporting methods yields customizable label sets that are more robust to temporal resolution–emotion words from interviews are few and far in between, but meaningful; while continuously annotated labels provide a sense of directionality.

From these findings, we propose that designers of real-time emotionally interactive devices should ensure that (1) the selected label metaphor should be planned with window size in mind, particularly when explicitly describing dynamic emotion labels like that of direction; (2) multi-modal systems should conform to the best conditions for each modality (i.e., multi-scale windows or multi-metaphor labelling) in order to optimize model performance; and (3) if the interaction involves manual touch, incorporating a pressure sensor in the areas where human touch is enacted may offer improved emotion recognition opportunities.

8.2 Future Directions

We presented an initial exploration into the combining labelling procedures for dynamic emotion modelling. We highlight where future directions are highly promising.

Parameters computed on high-resolution data are different. What does this mean?

To get behind different characteristics in computable descriptive parameters, one approach is to compare with other high-resolution data streams such as EEG and facial encoding. We plan to do this by focusing analysis on particular events (e.g., timeline regions stimuli known to trigger reactions in all – a scary spot in the game), and see how these parameters look across multiple participants when calibrated in a variety of ways.

At what time scale does calibration change?

Calibration happens immediately after the elicitation task, and all emotion words are contextualized around the task. Could engaging in a highly emotionally charged activity influence the rating scale upon reflection? In future iterations of the multi-pass protocol, we envision performing calibration tasks both at the beginning and end of the self-report labelling allowing us to investigate how calibration may drift within and between sessions.

How must models of dynamic emotion evolve?

Longitudinal studies will reveal how to create personalized models that evolve with the individual. Mood, life and situational context influence perception of emotional events [42] but also change dramatically over time: we wonder how repeat data collection over the course of months impacts emotion models.

How to capture a range of emotion experiences?

We report analysis on a single-dimensional scale, designed to simplify annotation; real-life events may trigger far more complex emotion landscapes where emotions are in conflict simultaneously (e.g., feeling excited and sad about graduation). How can we make it more intuitive to document multiple simultaneous scales?

How to account for subjectivity when evaluating personalized emotion models?

We create models as representational tools, adequate to a purpose. Emotion models cannot provide a true, holistic and perfectly accurate rendition of real-life processes and experiences. In this work, we attempt to incorporate contextualized subjective reporting to computational models, quantifying *performance* based on the machine learning understanding of *evaluation*. As we make strides to *humanize* emotion reports beyond a quantitative scale, how can we impart subjectivity into evaluation procedures to best reflect the user's values and needs? A model that is able to accurately detect emotions but misses the right time for delivering interventions might not satisfy its purposes. Or conversely, perhaps users are less interested in receiving correct outputs of 'satisfied' or 'content', and rather focus on specific emotions, such as 'attentive' or 'determined'. Machine Learning (ML) metrics are interesting for investigating and comparing performance, but they do not capture all the nuances of what a model might need to be considered effective.

Bibliography

[1] M. Z. I. Ahmed, N. Sinha, S. Phadikar, and E. Ghaderpour. Automated Feature Extraction on AsMap for Emotion Classification Using EEG. *Sensors*, 22(6):2346, Mar. 2022. ISSN 1424-8220. doi:10.3390/s22062346. URL https://www.mdpi.com/1424-8220/22/6/2346. → pages 22, 31, 32

[2] K. Alarabi Aljribi. A comparative analysis of frequency bands in eeg based emotion recognition system. In *The 7th International Conference on Engineering & MIS 2021*, pages 1–7, 2021. → page 27

[3] S. M. Alarcao and M. J. Fonseca. Emotions recognition using eeg signals: A survey. *IEEE Trans on Affective Computing*, 10(3):374–393, 2017. → page 5

[4] K. Altun and K. E. MacLean. Recognizing affect in human touch of a robot. *Pattern Recognition Letters*, 66(November):31–40, 2014. → page 4

[5] B. Azari, C. Westlin, A. B. Satpute, J. B. Hutchinson, P. A. Kragel, K. Hoemann, Z. Khan, J. B. Wormwood, K. S. Quigley, D. Erdogmus, J. Dy, D. H. Brooks, and L. F. Barrett. Comparing supervised and unsupervised approaches to emotion categorization in the human brain, body, and subjective experience. *Scientific Reports*, 10(1), Nov. 2020. doi:10.1038/s41598-020-77117-8. URL https://doi.org/10.1038/s41598-020-77117-8. → page 10

[6] I. Bakker, T. Van der Voordt, P. Vink, and J. De Boon. Pleasure, arousal, dominance: Mehrabian and russell revisited. *Current Psychology*, 33(3): 405–421, 2014. → page 16

[7] M. Balconi and G. Fronda. How to induce and recognize facial expression of emotions by using past emotional memories: A multimodal neuroscientific algorithm. *Frontiers in Psychology*, page 1658, 2021. → page 21

[8] L. Barnett. Keep in touch: The importance of touch in infant development. *Infant Observation*, 8(2):115–123, 2005. → page 4

[9] L. F. Barrett. The theory of constructed emotion: an active inference account of interoception and categorization. *Social cognitive and affective neuroscience*, 12(1):1–23, 2017. → pages 10, 14, 69

[10] L. F. Barrett, B. Mesquita, K. N. Ochsner, and J. J. Gross. The experience of emotion. *Annu. Rev. Psychol.*, 58:373–403, 2007. → pages 3, 14, 17

[11] M. M. Bradley and P. J. Lang. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry*, 25(1):49–59, 1994. → pages 3, 16, 17

[12] 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)*, pages 1–7, 2019. → pages 2, 3, 16, 46, 72

[13] P. Bucci, D. Marino, and I. Beschastnikh. Affective robots need therapy. *ACM Transactions on Human-Robot Interaction*, 2022. → page 2

[14] X. L. Cang, P. Bucci, A. Strang, J. Allen, K. MacLean, and H. S. Liu. Different strokes and different folks: Economical dynamic surface sensing and affect-related touch recognition. In *Proceedings of the 2015 ACM on Int'l Conf on Multimodal Interaction*, pages 147–154, 2015. → pages 4, 16

[15] X. L. Cang, P. Bucci, J. Rantala, and K. Maclean. Discerning affect from touch and gaze during interaction with a robot pet. *IEEE Trans on Affective Computing*, Early Access(01):1–1, 2021. → pages 2, 4, 6, 15, 21, 43, 44, 70

[16] X. L. Cang, R. R. Guerra, P. Bucci, B. Guta, L. Rodgers, H. Mah, S. Hsu, Q. Feng, C. S. Zhang, A. Agrawal, and K. E. MacLean. Choose or fuse: Enriching data views with multi-label emotion dynamics. In *IEEE 10th Int'l Conf on Affective Computing & Intelligent Interaction (ACII)*, 2022 (to appear). → pages 5, 24, 25, 68, 76

[17] G. L. Clore and A. Ortony. Psychological construction in the occ model of emotion. *Emotion Review*, 5(4):335–343, 2013. → pages 10, 14, 69

[18] J. A. Coan and J. J. Allen. *Handbook of emotion elicitation and assessment.* Oxford University Press, 2007. → page 15

[19] R. Cowie, E. Douglas-Cowie, S. Savvidou*, E. McMahon, M. Sawey, and M. Schröder. ’feeltrace’: An instrument for recording perceived emotion in real time. In *ISCA tutorial and research workshop (ITRW) on speech and emotion*, 2000. → page 16

[20] K. Crowley. Sleep and sleep disorders in older adults. *Neuropsychology review*, 21(1):41–53, 2011. → page 22

[21] R. T. Dean and W. Dunsmuir. Dangers and uses of cross-correlation in analyzing time series in perception, performance, movement, and neuroscience: The importance of constructing transfer function autoregressive models. *Behavior research methods*, 48(2):783–802, 2016. → page 38

[22] C. M. Deveney and D. A. Pizzagalli. The cognitive consequences of emotion regulation: an erp investigation. *Psychophysiology*, 45(3): 435–444, 2008. → page 22

[23] R.-N. Duan, J.-Y. Zhu, and B.-L. Lu. Differential entropy feature for eeg-based emotion classification. In *2013 6th International IEEE/EMBS Conference on Neural Engineering (NER)*, pages 81–84. IEEE, 2013. → pages 22, 32

[24] B. Dudzik and J. Broekens. A valid self-report is never late, nor is it early: On considering the “right” temporal distance for assessing emotional experience. In *2nd Momentary Emotion Elicitation & Capture Workshop at CHI*, 2021. → page 23

[25] M. A. Eid and H. A. Osman. Affective Haptics: Current Research and Future Directions. *IEEE Access*, 4:26–40, 2016. ISSN 2169-3536. doi:10.1109/ACCESS.2015.2497316. Conference Name: IEEE Access. → page 4

[26] M. Eimer and A. Holmes. An erp study on the time course of emotional face processing. *Neuroreport*, 13(4):427–431, 2002. → page 22

[27] P. Ekman, R. W. Levenson, and W. V. Friesen. Autonomic nervous system activity distinguishes among emotions. *Science*, 221(4616):1208–1210, 1983. → page 15

[28] C. Epp, M. Lippold, and R. L. Mandryk. Identifying emotional states using keystroke dynamics. In *Proceedings of the sigchi conference on human factors in computing systems*, pages 715–724, 2011. → page 21

[29] T. Field. Touch for socioemotional and physical well-being: A review. *Developmental review*, 30(4):367–383, 2010. → page 4

[30] A. Flagg and K. E. MacLean. Affective touch recognition for a furry therapeutic machine. In *Affective Touch Recognition for a Furry Therapeutic Machine*, pages 25–32, Barcelona, 2013. → page 4

[31] W. A. Fuller. *Intro to statistical time series*. John Wiley & Sons, 2009. → page 38

[32] Y. Gaffary, J.-C. Martin, and M. Ammi. Haptic expression and perception of spontaneous stress. *IEEE Trans on Affective Computing*, 11(1):138–150, 2020. → page 15

[33] J. J. Gross. Antecedent-and response-focused emotion regulation: divergent consequences for experience, expression, and physiology. *Journal of personality and social psychology*, 74(1):224, 1998. → page 2

[34] J. J. Gross and R. W. Levenson. Emotion elicitation using films. *Cognition & emotion*, 9(1):87–108, 1995. → page 15

[35] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1): 389–422, 2002. → page 47

[36] H. F. Harlow and R. R. Zimmermann. The development of affectional responses in infant monkeys. *Proc. Am. Philos. Soc.*, 102(5):501–509, 1958. → page 4

[37] J. Hernandez, P. Paredes, A. Roseway, and M. Czerwinski. Under pressure: sensing stress of computer users. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 51–60, 2014. → pages 5, 6, 21, 70

[38] M. J. Hertenstein. *The communicative functions of touch in adulthood*, page 299–327. Springer Publishing Company, New York, NY, US, 2011. ISBN 978-0-8261-2191-2. → page 4

[39] M. J. Hertenstein, D. Keltner, B. App, B. A. Bulleit, and A. R. Jaskolka. Touch communicates distinct emotions. *Emotion*, 6(3):528, 2006. → pages 16, 21

[40] M. J. Hertenstein, J. M. Verkamp, A. M. Kerestes, and R. M. Holmes. The communicative functions of touch in humans, nonhuman primates, and

rats: a review and synthesis of the empirical research. *Genetic, social, and general psychology monographs*, 132(1):5–94, 2006. → page 4

[41] J. Hirschberg and C. D. Manning. Advances in natural language processing. *Science*, 349(6245):261–266, 2015. → page 30

[42] K. Hoemann, Z. Khan, M. J. Feldman, C. Nielson, M. Devlin, J. Dy, L. F. Barrett, J. B. Wormwood, and K. S. Quigley. Context-aware experience sampling reveals the scale of variation in affective experience. *Scientific reports*, 10(1):1–16, 2020. → page 78

[43] M. Hoque and R. W. Picard. Acted vs. natural frustration and delight: Many people smile in natural frustration. In *IEEE Face & Gesture*, pages 354–359, 2011. → page 15

[44] M. Houben, W. Van Den Noortgate, and P. Kuppens. The relation between short-term emotion dynamics and psychological well-being: A meta-analysis. *Psychological bulletin*, 141(4):901, 2015. → pages 17, 40, 70

[45] J. R. Hughes and M. Melyn. Eeg and seizures in autistic children and adolescents: further findings with therapeutic implications. *Clinical EEG and Neuroscience*, 36(1):15–20, 2005. → page 22

[46] J. S. Hunter. The exponentially weighted moving average. *Journal of quality technology*, 18(4):203–210, 1986. → page 43

[47] N. Jatupaiboon, S. Pan-ngum, and P. Israsena. Emotion classification using minimal eeg channels and frequency bands. In *The 2013 10th international joint conference on Computer Science and Software Engineering (JCSSE)*, pages 21–24. IEEE, 2013. → pages 32, 70

[48] K. Jensen. Envelope model of isolated musical sounds. In *Proceedings of the 2nd COST G-6 Workshop on Digital Audio Effects (DAFx99)*, volume 12. Citeseer, 1999. → page 70

[49] M. F. Jung. Affective grounding in human-robot interaction. In *2017 12th ACM/IEEE Int'l Conf on Human-Robot Interaction (HRI*, pages 263–273. IEEE, 2017. → pages 5, 21

[50] M. M. Jung, R. Poppe, M. Poel, and D. K. Heylen. Touching the void–introducing cost: corpus of social touch. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 120–127, 2014. → pages 4, 44, 70

[51] M. Kay and J. O. Wobbrock. Package 'artool'. *CRAN Repository (2016)*, pages 1–13, 2016. → page 50

[52] D. Keltner, D. Sauter, J. L. Tracy, and A. S. Cowen. Emotional expression: Advances in basic emotion theory. *Journal of Nonverbal Behavior*, 2019. doi:10.1007/s10919-019-00293-3. → page 10

[53] T. Kinnunen and M. Kolehmainen. Touch and affect: Analysing the archive of touch biographies. *Body & Society*, 25(1):29–56, 2019. → page 4

[54] T. K. Koo and M. Y. Li. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, 15(2):155–163, 2016. → page 35

[55] J. S. Kumar and P. Bhuvaneswari. Analysis of electroencephalography (eeg) signals and its categorization–a study. *Procedia engineering*, 38: 2525–2536, 2012. → page 22

[56] P. Kuppens and P. Verduyn. Emotion dynamics. *Current Opinion in Psychology*, 17:22–26, 2017. → pages 2, 16, 17

[57] D. Lakens, A. M. Scheel, and P. M. Isager. Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, 1(2):259–269, 2018. → page 40

[58] P.-M. Lee, W.-H. Tsui, and T.-C. Hsiao. The influence of emotion on keyboard typing: an experimental study using visual stimuli. *Biomedical engineering online*, 13(1):1–12, 2014. → page 21

[59] M. Leng, Y. Zhao, and Z. Wang. Comparative efficacy of non-pharmacological interventions on agitation in people with dementia: A systematic review and bayesian network meta-analysis. *International Journal of Nursing Studies*, 102:103489, 2020. → page 4

[60] M. Li and B.-L. Lu. Emotion classification based on gamma-band EEG. In *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 1223–1226, Sept. 2009. doi:10.1109/IEMBS.2009.5334139. ISSN: 1558-4615. → pages 22, 43

[61] Y.-P. Lin, C.-H. Wang, T.-P. Jung, T.-L. Wu, S.-K. Jeng, J.-R. Duann, and J.-H. Chen. EEG-Based Emotion Recognition in Music Listening. *IEEE Transactions on Biomedical Engineering*, 57(7):1798–1806, July 2010. ISSN 1558-2531. doi:10.1109/TBME.2010.2048568. → page 6

[62] L. S. Löken and H. Olausson. The skin as a social organ. *Experimental brain research*, 204(3):305–314, 2010. → page 4

[63] H. Lütkepohl. *New introduction to multiple time series analysis*. Springer Science & Business Media, 2005. → page 37

[64] H.-R. Lv, Z.-L. Lin, W.-J. Yin, and J. Dong. Emotion recognition based on pressure sensor keyboards. In *2008 IEEE international conference on multimedia and expo*, pages 1089–1092. IEEE, 2008. → pages 5, 6, 21, 70

[65] A. Maalej and I. Kallel. Does keystroke dynamics tell us about emotions? a systematic literature review and dataset construction. In *2020 16th International Conference on Intelligent Environments (IE)*, pages 60–67. IEEE, 2020. → page 21

[66] F. McGlone, H. Olausson, J. A. Boyle, M. Jones-Gotman, C. Dancer, S. Guest, and G. Essick. Touching and feeling: differences in pleasant touch processing between glabrous and hairy skin in humans. *European Journal of Neuroscience*, 35(11):1782–1788, 2012. → pages 4, 5

[67] M. F. Mendez. Frontotemporal dementia: A window to alexithymia. *The Journal of neuropsychiatry and clinical neurosciences*, 33(2):157–160, 2021. → page 1

[68] B. Mesquita, L. F. Barrett, and E. R. Smith. *The mind in context*. Guilford Press, 2010. → page 2

[69] F. M. Miranda, N. Köehnecke, and B. Y. Renard. Hiclass: a python library for local hierarchical classification compatible with scikit-learn. *arXiv preprint arXiv:2112.06560*, 2021. → page 59

[70] A. Mollahosseini, B. Hasani, and M. H. Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Trans on Affective Computing*, 10(1):18–31, 2017. → page 21

[71] A. Montagu. Touching, the human significance of the skin. *Perennial Library*, pages 98–99, 1972. → page 4

[72] A. Moors, P. C. Ellsworth, K. R. Scherer, and N. H. Frijda. Appraisal theories of emotion: State of the art and future development. *Emotion Review*, 5(2):119–124, 2013. → pages 2, 3, 16, 23

[73] K. P. Murphy. *Machine Learning*. Adaptive Computation and Machine Learning series. MIT Press, London, England, Aug. 2012. → page 18

[74] M. Murugappan, N. Ramachandran, and Y. Sazali. Classification of human emotion from EEG using discrete wavelet transform. *Journal of Biomedical Science and Engineering*, 03(04):390, Apr. 2010. doi:10.4236/jbise.2010.34054. URL http://www.scirp.org/journal/PaperInformation.aspx?PaperID=1607. → pages 6, 22

[75] A. N. H. Nahin, J. M. Alam, H. Mahmud, and K. Hasan. Identifying emotion by keystroke dynamics and text pattern analysis. *Behaviour & Information Technology*, 33(9):987–996, 2014. → page 21

[76] A. Ortony, G. L. Clore, and A. Collins. *The cognitive structure of emotions*. Cambridge university press, 1990. → pages 2, 3, 16, 23

[77] W. S. Parker and E. Winsberg. Values and evidence: how models make a difference. *European Journal for Philosophy of Science*, 8(1):125–142, 2018. → page 11

[78] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. → page 20

[79] D. Playdead. Playdead's inside. https://playdead.com/games/inside/, 2022. Accessed: 2022-08-21. → page 5

[80] W. Revelle. *psych: Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University, Evanston, Illinois, 2022. URL https://CRAN.R-project.org/package=psych. R package version 2.2.3. → page 40

[81] T. Ritchie, J. J. Skowronski, J. Hartnett, B. Wells, and W. R. Walker. The fading affect bias in the context of emotion activation level, mood, and personal theories of emotion change. *Memory*, 17(4):428–444, 2009. → page 23

[82] P. V. Rouast, M. T. Adam, and R. Chiong. Deep learning for human affect recognition: Insights and new developments. *IEEE Transactions on Affective Computing*, 12(2):524–543, 2019. → page 23

[83] G. A. Rousselet. Does filtering preclude us from studying erp time-courses? *Frontiers in psychology*, 3:131, 2012. → page 22

[84] J. A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161, 1980. → pages 3, 14, 16, 17, 69

[85] P. Salovey, J. D. Mayer, S. L. Goldman, C. Turvey, and T. P. Palfai. Emotional attention, clarity, and repair: exploring emotional intelligence using the trait meta-mood scale. *Emotion, disclosure, & health*, pages 125–154, 1995. → page 25

[86] S. Seabold and J. Perktold. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*, 2010. → page 38

[87] A. K. Seth, A. B. Barrett, and L. Barnett. Granger causality analysis in neuroscience and neuroimaging. *Journal of Neuroscience*, 35(8): 3293–3297, 2015. → page 37

[88] K. Sharma, C. Castellini, F. Stulp, and E. L. Van den Broek. Continuous, real-time emotion annotation: A novel joystick-based analysis framework. *IEEE Trans on Affective Computing*, 11(1):78–84, 2020. → page 16

[89] G. Sheppes and J. J. Gross. Is timing everything? temporal considerations in emotion regulation. *Personality and Social Psychology Review*, 15(4): 319–331, 2011. → pages 23, 73

[90] L.-C. Shi, Y.-Y. Jiao, and B.-L. Lu. Differential entropy feature for eeg-based vigilance estimation. In *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 6627–6630. IEEE, 2013. → page 32

[91] A. Shojaie and E. B. Fox. Granger causality: A review and recent advances. *Annual Review of Statistics and Its Application*, 9:289–319, 2022. → page 38

[92] N. J. Shoumy, L.-M. Ang, K. P. Seng, D. M. Rahaman, and T. Zia. Multimodal big data affective analytics: A comprehensive survey using text, audio, visual and physiological signals. *Journal of Network and Computer Applications*, 149:102447, 2020. → pages 4, 23

[93] A. Simoës-Perlant, C. Lemercier, C. Pêcher, and S. Benintendi-Medjaoued. Mood self-assessment in children from the age of 7. *Europe's Journal of Psychology*, 14(3):599, 2018. → page 16

[94] S. H. Sperry, M. A. Walsh, and T. R. Kwapil. Emotion dynamics concurrently and prospectively predict mood psychopathology. *Journal of affective disorders*, 261:67–75, 2020. → pages 17, 72

[95] D. Stoeva and M. Gelautz. Body language in affective human-robot interaction. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, pages 606–608, 2020. → page 21

[96] T. Terkildsen and G. Makransky. Measuring presence in video games: An investigation of the potential use of physiological measures as indicators of presence. *International journal of human-computer studies*, 126:64–80, 2019. → page 15

[97] T. Thanapattheerakul, K. Mao, J. Amoranto, and J. H. Chan. Emotion in a century: A review of emotion recognition. In *proceedings of the 10th international conference on advances in information technology*, pages 1–8, 2018. → pages 2, 4, 10, 14, 23, 46, 69

[98] J. B. Torre and M. D. Lieberman. Putting feelings into words: Affect labeling as implicit emotion regulation. *Emotion Review*, 10(2):116–124, 2018. → page 23

[99] A. T. Tzallas, M. G. Tsipouras, and D. I. Fotiadis. Epileptic seizure detection in eegs using time–frequency analysis. *IEEE transactions on information technology in biomedicine*, 13(5):703–710, 2009. → page 22

[100] P. Verduyn. Emotion duration. In *Affect Dynamics*, pages 3–18. Springer, 2021. → page 3

[101] P. Verduyn, I. Van Mechelen, and F. Tuerlinckx. The relation between event processing and the duration of emotional experience. *Emotion*, 11(1): 20, 2011. → page 43

[102] P. Verduyn, P. Delaveau, J.-Y. Rotgé, P. Fossati, and I. Van Mechelen. Determinants of emotion duration and underlying psychological and neural mechanisms. *Emotion Review*, 7(4):330–335, 2015. → pages 22, 23, 43

[103] R. Wampfler, S. Klingler, B. Solenthaler, V. R. Schinazi, M. Gross, and C. Holz. Affective state prediction from smartphone touch and sensor data in the wild. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI ’22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391573.

doi:10.1145/3491102.3501835. URL https://doi.org/10.1145/3491102.3501835. → pages 5, 23

[104] X.-W. Wang, D. Nie, and B.-L. Lu. Emotional state classification from EEG data using machine learning approach. *Neurocomputing*, 129:94–106, Apr. 2014. ISSN 09252312. doi:10.1016/j.neucom.2013.06.046. URL https://linkinghub.elsevier.com/retrieve/pii/S0925231213009867. → pages 6, 22, 43, 70

[105] 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*, 54(6):1063, 1988. → pages 3, 16, 27

[106] R. Wirth and J. Hipp. Crisp-dm: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, volume 1, pages 29–39. Manchester, 2000. → page 8

[107] J. O. Wobbrock, L. Findlater, D. Gergle, and J. J. Higgins. The aligned rank transform for nonparametric factorial analyses using only anova procedures. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 143–146, 2011. → pages 50, 61

[108] T. Xie, M. Cao, and Z. Pan. Applying self-assessment manikin (sam) to evaluate the affective arousal effects of vr games. In *Proceedings of the 2020 3rd Int'l Conf on Image and Graphics Processing*, pages 134–138, 2020. → page 16

[109] X. Xing, Z. Li, T. Xu, L. Shu, B. Hu, and X. Xu. Sae+lstm: A new framework for emotion recognition from multi-channel eeg. *Frontiers in neurorobotics*, 13:37, 2019. → pages 22, 30

[110] T. Xue, S. Ghosh, G. Ding, A. El Ali, and P. Cesar. Designing real-time, continuous emotion annotation techniques for 360 vr videos. In *Extended Abstracts of the 2020 CHI Int'l Conf on Human Factors in Computing Systems*, pages 1–9, 2020. → page 16

[111] S. Yohanan and K. E. MacLean. The role of affective touch in human-robot interaction: Human intent and expectations in touching the haptic creature. *International Journal of Social Robotics*, 4(2):163–180, 2012. → page 4

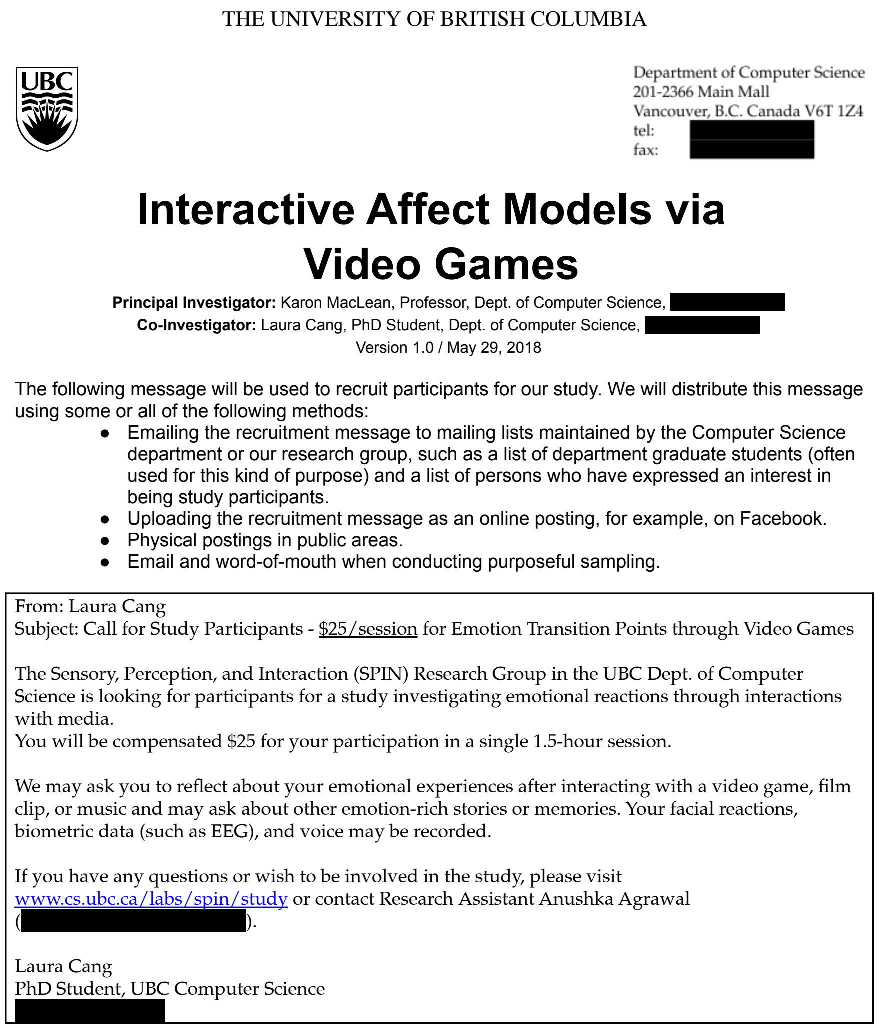
[112] J. Zhang, Z. Yin, P. Chen, and S. Nichele. Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review. *Information Fusion*, 59:103–126, 2020. → pages 10, 14, 23, 46, 69

[113] W.-L. Zheng and B.-L. Lu. Investigating critical frequency bands and channels for eeg-based emotion recognition with deep neural networks. *IEEE Transactions on autonomous mental development*, 7(3):162–175, 2015. → pages 27, 70

Appendix A

Data Collection Materials

A.1 Study Recruitment Poster



THE UNIVERSITY OF BRITISH COLUMBIA

Department of Computer Science
201-2366 Main Mall
Vancouver, B.C. Canada V6T 1Z4
tel: [redacted]
fax: [redacted]

Interactive Affect Models via Video Games

Principal Investigator: Karon MacLean, Professor, Dept. of Computer Science, [redacted]
Co-Investigator: Laura Cang, PhD Student, Dept. of Computer Science, [redacted]
Version 1.0 / May 29, 2018

The following message will be used to recruit participants for our study. We will distribute this message using some or all of the following methods:

- Emailing the recruitment message to mailing lists maintained by the Computer Science department or our research group, such as a list of department graduate students (often used for this kind of purpose) and a list of persons who have expressed an interest in being study participants.
- Uploading the recruitment message as an online posting, for example, on Facebook.
- Physical postings in public areas.
- Email and word-of-mouth when conducting purposeful sampling.

From: Laura Cang
Subject: Call for Study Participants - $25/session for Emotion Transition Points through Video Games

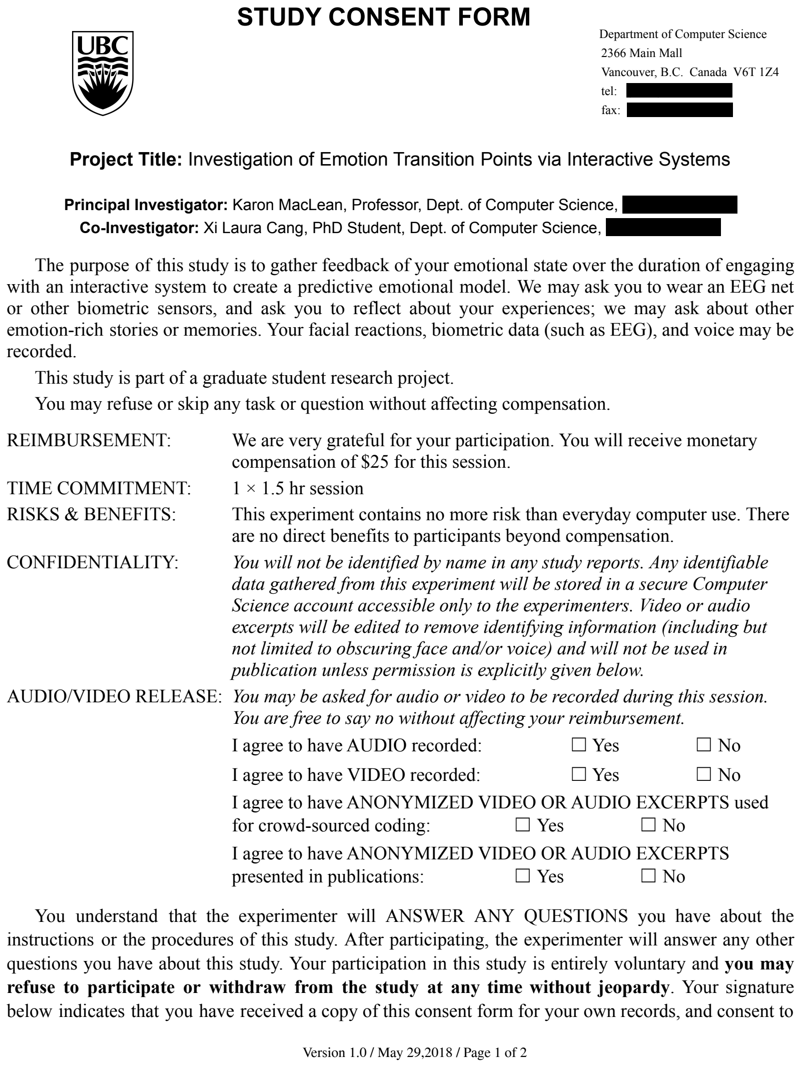
The Sensory, Perception, and Interaction (SPIN) Research Group in the UBC Dept. of Computer Science is looking for participants for a study investigating emotional reactions through interactions with media.
You will be compensated $25 for your participation in a single 1.5-hour session.

We may ask you to reflect about your emotional experiences after interacting with a video game, film clip, or music and may ask about other emotion-rich stories or memories. Your facial reactions, biometric data (such as EEG), and voice may be recorded.

If you have any questions or wish to be involved in the study, please visit www.cs.ubc.ca/labs/spin/study or contact Research Assistant Anushka Agrawal ([redacted]).

Laura Cang
PhD Student, UBC Computer Science
[redacted]

A.2 Study Consent Form



STUDY CONSENT FORM

Department of Computer Science
2366 Main Mall
Vancouver, B.C. Canada V6T 1Z4
tel: [redacted]
fax: [redacted]

Project Title: Investigation of Emotion Transition Points via Interactive Systems

Principal Investigator: Karon MacLean, Professor, Dept. of Computer Science, [redacted]
Co-Investigator: Xi Laura Cang, PhD Student, Dept. of Computer Science, [redacted]

The purpose of this study is to gather feedback of your emotional state over the duration of engaging with an interactive system to create a predictive emotional model. We may ask you to wear an EEG net or other biometric sensors, and ask you to reflect about your experiences; we may ask about other emotion-rich stories or memories. Your facial reactions, biometric data (such as EEG), and voice may be recorded.

This study is part of a graduate student research project.

You may refuse or skip any task or question without affecting compensation.

REIMBURSEMENT: We are very grateful for your participation. You will receive monetary compensation of $25 for this session.

TIME COMMITMENT: 1 × 1.5 hr session

RISKS & BENEFITS: This experiment contains no more risk than everyday computer use. There are no direct benefits to participants beyond compensation.

CONFIDENTIALITY: *You will not be identified by name in any study reports. Any identifiable data gathered from this experiment will be stored in a secure Computer Science account accessible only to the experimenters. Video or audio excerpts will be edited to remove identifying information (including but not limited to obscuring face and/or voice) and will not be used in publication unless permission is explicitly given below.*

AUDIO/VIDEO RELEASE: *You may be asked for audio or video to be recorded during this session. You are free to say no without affecting your reimbursement.*

I agree to have AUDIO recorded: ☐ Yes ☐ No

I agree to have VIDEO recorded: ☐ Yes ☐ No

I agree to have ANONYMIZED VIDEO OR AUDIO EXCERPTS used for crowd-sourced coding: ☐ Yes ☐ No

I agree to have ANONYMIZED VIDEO OR AUDIO EXCERPTS presented in publications: ☐ Yes ☐ No

You understand that the experimenter will ANSWER ANY QUESTIONS you have about the instructions or the procedures of this study. After participating, the experimenter will answer any other questions you have about this study. Your participation in this study is entirely voluntary and **you may refuse to participate or withdraw from the study at any time without jeopardy**. Your signature below indicates that you have received a copy of this consent form for your own records, and consent to participate in this study. Any questions about the study can be directed to Laura Cang, [redacted].

If you have any concerns or complaints about your rights as a research participant and/or your experiences while participating in this study, contact the Research Participant Complaint Line in the UBC Office of Research Ethics at [redacted] or if long distance e-mail [redacted] or call toll free [redacted].

You hereby CONSENT to participate and acknowledge RECEIPT of a copy of the consent form:

PRINTED NAME ________________________________ DATE ____________________________

SIGNATURE ____________________________________