

Towards Socially Intelligent HRI Systems: Quantifying Emotional, Social, and Relational Context in Real-World Human Interactions

Jesse Parent, Yelin Kim

Interaction Sensing and Perception in Real Environment (INSPIRE) Lab
College of Engineering and Applied Science
University at Albany, State University of New York
{jtparent, yelinkim}@albany.edu

Abstract

In this paper, we present current issues and findings for developing HRI systems with capabilities to measure and interpret emotional, social, and relational context in real-world human interactions. We discuss our ongoing work to create a dataset that will contribute to achieve this overarching goal. Our source data are documentary films of face-to-face interactions of two-person pairings of real relationships. We select an example video from the source data to explore labeling methods to describe its emotional, social, and relational phenomena. Extending from previous studies, we propose a social label, synthesizing emotion and relationship labels to provide interpretable descriptions of social-relational context in human interaction. We demonstrate how our proposed social label, *engagement label*, is associated with salient emotional and social dynamics during interactions. We further discuss open questions and provide insight into future research directions where a robust and multifaceted approach is necessary. Building from past efforts in classifying and interpreting affective and relational context, this paper opens a new gateway for developing HRI systems that can understand and adapt to real-world interactions.

1 Introduction

Socially intelligent AI systems are capable of measuring, understanding, and adapting to emotional and social context, such as happiness, engagement, relationship, and rapport. These systems will benefit human-robot interaction (HRI) technologies by enabling more natural and human-centered interaction. Understanding nuanced interpersonal relationships is challenging for humans, no less for artificial intelligences. In this work, we present the initial stages of our ongoing AI work that aims to understand social and relational context in naturalistic, multimodal interactions.


Recent research on social signal processing has gained interest among researchers in various fields including affective computing and robotics (Pantic et al. 2011; Taylor and Riek 2016). However, an ongoing question remains the lack of consensus for defining meaningful emotional, relational, and social context for HRI applications. Methods to effectively generate useful annotation, and to identify time units and labels to capture substantial social-relational context are

also underexplored. We consider useful annotation as: intelligible to humans observing the phenomena; founded on established literature and categorization schemes; and being quantifiable for machine comprehension. Moreover, current research faces a central challenge due to limited real-world data, which stems from difficulty in real-world data capturing natural, genuine human expressions during interactions.

To overcome the aforementioned challenges, we propose to develop a new dataset based on documentary films of face-to-face interactions of two-person pairs in *real, established relationships*, ranging from familial, acquaintance, and strangers (Figure 1). This paper presents our preliminary analysis on applying and developing labeling schemes for emotion and relationship, while considering social context labels and proposing the new concept of *engagement label*. By examining audio-video interactions between two human speakers with given relationship metadata — closeness, type, and duration—we explore the efficacy, opportunities, and open questions for using this new labeling approach to interpret social and relational context.

Building on prior work in social signal processing (Pelachaud et al. 2012), the proposed engagement labels provide insight into complex human interactions by examining activation, valence fluctuation, and turn-taking patterns. As shown in Figure 1, we use dimensional analysis with activation and valence for labeling emotion. For relationship labeling, we use a scale of familiarity. The proposed engagement label links affective display and nature of relationship. This will offer insight to social-relational context of human interactions and contribute to develop socially intelligent HRI systems in real-world scenarios.

We further discuss current challenges and findings in developing automatic recognition methods for affective and social cues, using an example filmed session between a mother and daughter. We use given relationship metadata, combined with an established emotion labeling approach, to investigate the social context using the proposed engagement label. We offer research directions that will lead to design and development of socially intelligent AI systems. Through reviewing methodologies for describing affective and social context, we lay groundwork for developing the desired dataset. The extracted audio-visual features and labels will be shared publicly to serve as a benchmark for our research community.



Transcript	Activation	Valence
M1: What scares you the most?	4	3
D1: Oh, I have lots of things that scare me! The first is you being murdered!	5	1
M2: Oh, that is the first one?	5	1
D2: Yes! ...and the boogeyman...	5	3

Figure 1: An example HE region between a mother (“M”) and daughter (“D”) pairing in our dataset, with the transcript and emotion labels below. This HE region shows consistent high activation while varying valence labels. This figure is generated using {THE AND} documentary films created by The Skin Deep Media Corp.

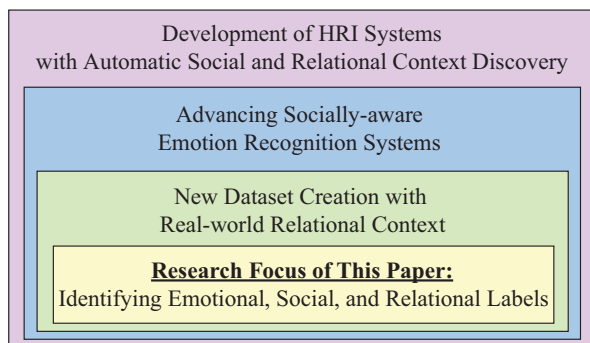


Figure 2: Context for this current research paper with regard to the overall aims of the research.

1.1 Scope of This Paper

It is worth noting the place of this present research in the scope of our overall goal. As shown in Figure 2, the long-term vision is to innovate current HRI systems with comprehension of emotion awareness of social context. To that aim, we seek to create a dataset containing real-world social and relational context. As described in Section 3.2, we start with affective cues (emotion labeling), known relationships (familiarity labeling), and propose a new engagement labeling approach for describing social context.

2 Background and Current Status

2.1 Affective Labeling

Several affective labeling approaches have been proposed for describing emotional context of expressive behaviors. An assumption behind the affective labeling approaches is that perceived emotions labeled by human annotators can describe the emotional context in the dataset. This assumption leads to two widely-used approaches to *affective labeling*:

categorical and dimensional approaches (Gunes et al. 2011; Grandjean, Sander, and Scherer 2008). Categorical emotion labels are discrete, such as *Angry*, *Happy*, *Sad*, etc. In contrast, dimensional emotion labels are represented using continuous values corresponding to different emotional dimensions. The two most common dimensions are *valence* (positive vs. negative) and *activation* (calm vs. excited) (Schlossberg 1954). Valence-activation dimensional space have been used in previous affective computing research (Wöllmer et al. 2008; Ringeval et al. 2015; Chao et al. 2015).

Previous work on the conceptualization of multidimensional representation of emotion has led to exploration of modeling aspects of social and interactive behavior (Vinciarelli, Pantic, and Bourlard 2009). Other metrics include modalities and cues (such as verbal and nonverbal cues, and bio signals), and multimodal dyadic interaction (Metallinou et al. 2011).

With a multitude of methods for interpreting affective data (Anagnostopoulos, Iliou, and Giannoukos 2012; Gunes and Schuller 2013), there is work to be done in creating packages of analytical or interpretive methods that identify specific social context, relational information, and relational significance beyond explicit or instantaneous emotional analysis. This is particularly important for HRI systems that can analyze continuous, natural interactions.

2.2 Related Emotion Databases

Previous work has developed audio-visual emotion recognition models using benchmark datasets, such as IEMOCAP (Busso et al. 2008), SEMAINE (McKeown et al. 2012), and RECOLA (Ringeval et al. 2013). For instance, Parthasarathy and Busso use the SEMAINE database to develop dynamic systems that track emotion continuously, detecting emotionally significant periods of interaction. The difficulty of forming reliable labels has led to much discussion, with one finding that annotators evaluating continuous situations

have stronger concurrence on relative trends in emotional change than on designation of specific values (Yang and Chen 2011).

While current databases are useful in developing emotion-aware AI systems, there is a need for naturalistic data concerning real-world relationship context. The next step in advanced comprehension or functional involvement in human conversations is interpretation of subjects’ relational qualities. Continuous assessment and understanding of affective states, as well as the ability to derive social or relational context in real, ongoing environments, remains an open challenge.

3 Proposed Approach

3.1 Recording of the Corpus

Our proposed dataset includes 100 videos of unedited documentary films of {THE AND}, produced by the Skin Deep Media Corp (Figure 1). Videos are recorded in-studio, with cameras trained on individual speaker’s faces. There is also a wide-angle lens capturing interaction, speaker’s bodies, and other nonverbal factors. To ensure the framing is consistent to the pixel, each video contains triptych with a full 1920 x 1080 (in pixels) shot from the wide angle and two 1080 x 1080 shot of the two speakers, for a total resolution of 4080 x 1080.

Each audio-video file contains a pair of speakers, sitting across from each other in a controlled studio. Participant pairings had diverse relationship characteristics of various ages, genders, and backgrounds. The ground truth relationship data are given by the filmmakers: (i) relationship types that include married; engaged; family; ex-couple; divorced; dating; friends; and first-time encounters, and (ii) relationship durations, where the maximum duration of relationship is 40 years, minimum is 0 days (strangers).

The documentary film directors of {THE AND} created scripts of questions expected to evoke emotional responses of interest to viewers. Questions were written on cue cards placed on a table between two participants. A participant would draw a card, read, then ask a partner the question. After the response, roles would alternate: the person responding would then become the asker, drawing a new question card. The questions were open-ended, such as: “What are your feelings about (a person, topic, or experience)?”

Responses were not scripted in format. Participants were free to clarify, ask follow-ups, and enact other organic elements of conversation. There were no time limits for questions or total session length.

3.2 Labeling and Feature Extraction

To be consistent with previous emotion recognition studies (Metallinou et al. 2011; Mariooryad and Busso 2013), we label emotions at the *utterance* level, which is defined as a turn when a speaker is actively speaking. We use ELAN Linguistic Annotation software to manually segment interaction source films, and apply labels, dimensions, and other annotation (Wittenburg et al. 2006; Max Planck Institute for Psycholinguistics). Utterance-level segmentation offers labeling of specific emotional expressions, whereas more en-

compassing segmentation may allow trend and dynamic labeling.

Emotion Label: From prior work in emotion recognition (Grandjean, Sander, and Scherer 2008; Gunes et al. 2011), we use dimensional analysis from Russel’s Circumplex of Affect model (Grandjean, Sander, and Scherer 2008; Gunes et al. 2011; Mehrabian 1996; Russell 1980). We use valence as unpleasant (1) to pleasant (5), range 1-5. We use activation as calm (1) to excited (5), range 1-5.

Relationship Label: Based on previous work of LaFrance, Hecht, and Paluck, we use discrete levels to describe degrees of closeness:

1. completely new encounter: a stranger;
2. minimal familiarity: limited, occasional prior history;
3. normalized familiarity: common encounters but not extensive relational depth (social acquaintances, coworkers); and
4. advanced familiarity: high proximity and depth (family, best friends).

In our sample video, mother and daughter have advanced familiarity. A supervisor at work is normalized familiarity; a part-time library assistant you’ve spoken to twice, minimal familiarity.

Social Label: To analyze social signals in the proposed dataset, we seek to identify *high engagement (HE)* regions in each video. We consider turn-taking in line with past social signal processing research (Pantic et al. 2011; Sacks, Schegloff, and Jefferson 1974). In this paper, we define a HE region as a region of interaction that includes:

- (i) at least 4 utterances from both speakers, i.e., two turn changes between the speakers, and lasts at least 20 seconds.
- (ii) activation levels higher than 4 from both speakers, and
- (iii) valence level of a speaker deviates more than ± 2 within the segment, e.g., a speaker changes the valence from 2 to 4.

We use engagement labels as categorical labels that designate social context and identify highly engaged moments — other potential labels are discussed in Section 4.1. Functionally, HE regions designate, from the stream of incoming data, regions of interaction that may offer insight for social and relational context when paired with the other labeling approaches.

4 Findings and Discussion

In this paper, we discuss our findings via a session of interaction between a mother (‘M’) and daughter (‘D’), with length of 39 minutes and 40 seconds. The session contains 10 periods that qualify for HE regions, the longest being 80 seconds and the shortest 22 seconds. As shown in Table 1, during the entire session, we find 7:23 minutes of HE regions, or 18.61% of the session’s length. We note that 8 of 10 HE periods occur within the first 24 minutes of the session. We refer to questions in the format “Q#” in the order asked, so the first question is listed as Q1. There are 31 questions total, M asked 14, D asked 17.

Table 1: Statistics of high engagement (HE) regions during a mother (M)-daughter (D) interaction. Top section results are aggregated totals from the session, lasting 39 minutes and 40 seconds. Beneath, each speaker’s emotion labels, for all utterances, and within HE regions.

High Engagement Regions	Duration
Average length of a HE region	44.3 seconds
Total time of HE regions	7:23 minutes (18.61%)
Utterance-Specific Ratings of Mother and Daughter	Activation and Valence Mean (Standard dev.)
M- All utterances	A 3.24 (0.74), V 2.73 (0.80)
M- HE regions	A 4.23 (0.59), V 2.54 (1.05)
D- All utterances	A 3.49 (0.89), V 3.08 (0.84)
D- HE regions	A 4.31 (0.48), V 3.00 (0.91)

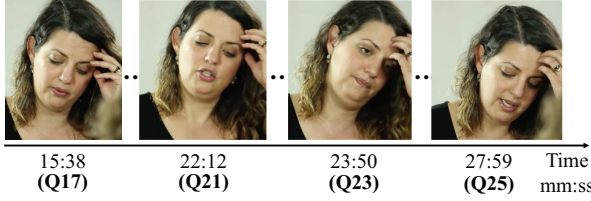


Figure 3: M shown at four different points during the session, having similar head postures and expressions, appearing uneasy when asking certain questions to D. This figure is generated using {THE AND} films by The Skin Deep Media Corp.

As shown in Table 1, M’s mean valence for HE periods (2.54) was more negative than the mean for the entire session (2.73). D, on the other hand, had almost the same average for both HE (3.00) and all utterances (3.08). When HE periods of an individual have lower valence than general interaction, but for the other there is no change, this may offer relational-social insight. For M, this may indicate she was more negatively affected by discussion in HE periods, or that she had a greater concern or lack of ease on the specific topics with D.

We observe the longest periods of HE involved direct questions concerning significant relationships. D shows consistent high activation around those topics, and M shows consistent interest in associated information. The two longest HE regions (80 and 70 seconds), which are well above average (44.3 seconds, standard deviation 21.45), predominantly discuss the closest relationships to the speakers.

As shown in Figure 3, we see examples of M shielding her head. M shows this posture on 7 of 14 questions she asked to D, such as “What is the most ridiculous thing I’ve done?” (Q3, 01:15), and “What do you think are my best qualities as a Mom?” (Q7, 05:02). The phenomenon also appears in Figure 1, with M asking “What scares you the most?” (Q11, 07:47). In Figure 3, M asked D: for (Q17), “What do you love and hate about having a younger sister?”; (Q21) “Do we spend enough time together?”; (Q23) “How is our relationship different than with your dad?”; (Q25) “How can I be a better mom?”. When asking certain questions, M may

show distance or uneasiness about D’s forthcoming answer.

D’s valence scores are more varied, ranging from 1 to 5, (standard deviation: 0.84) than M ranging from 1 to 5, (standard deviation: 0.80) during the session. D often only scored in extremes of 1 or 5 when there is specific excitement or reaction. We note D displays less contrived expression. This raises questions of how to account for or label contrived emotional expression.

Our results show a difference in how M and D handle discussion. This may relate to age differences, relative maturity, or a greater sense of responsibility that M has towards D. Regions not meeting criteria for HE lack valence variation, duration, and generally did not have many turn-takings, interest in clarification, or follow-up questions.

4.1 Open Questions

In this section, we discuss open questions for future research and potential research directions to address them. First, given diversity in relationship type, closeness, and duration, many different behaviors may be informative for relational context. There are many combinations of observable phenomena and interaction context mapping to be explored.

For HRI systems with no prior knowledge, procuring substantial relational-social context requires a time dynamic and various reference points to draw from. Being able to measure multiple sets of indicators and their frequency over time may allow deriving of social or relational insight. We speculate that more time spent in HE periods may lead to developing rapport or increasing closeness. To that end, we look to explore trajectory or vector periods of relative increase or decrease in closeness based on social, relational, and emotional dynamics.

While our paper notes the longest HE regions predominantly focused on significant relationships, we look to investigate whether this holds across non-familial interactions. We also see opportunity to enhance diversity of social labeling, such as noting acts of reassurance, emotional repression, or emotional contrivance.

5 Conclusion

In this paper, we explore methods to derive social and relational context by applying emotion, relation, and social labels to high quality film. We aim to build an HRI system that can interpret emotional, social, and relational context. We find a robust and multifaceted approach is necessary, with more experimentation with categories, dimensions, and other labeling critical for future research. We introduce engagement labels to describe social context, which may offer insight to human interaction when paired with emotion and relationship labeling. When valence is uniform throughout a session for a participant yet becomes more negative for a second participant during HE periods, we consider this may offer a relatively greater sense of concern from the second participant about engaging or sharing with the first. Our efforts will contribute to research improving HRI systems, working towards greater comprehension of social and relational context in real-life human interaction.

Acknowledgement

The authors gratefully acknowledge {THE AND} videos shared by the Skin Deep Media Corp.

References

- Anagnostopoulos, C. N.; Iliou, T.; and Giannoukos, I. 2012. Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011. *Artificial Intelligence Review* 43(2):155–177.
- Busso, C.; Bulut, M.; Lee, C.-C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J. N.; Lee, S.; and Narayanan, S. S. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation* 42(4):335.
- Chao, L.; Tao, J.; Yang, M.; Li, Y.; and Wen, Z. 2015. Long short term memory recurrent neural network based multimodal dimensional emotion recognition. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, 65–72. ACM.
- Grandjean, D.; Sander, D.; and Scherer, K. 2008. Conscious emotional experience emerges as a function of multilevel, appraisal-driven response synchronization. *Consciousness and cognition* 17(2):484.
- Gunes, H., and Schuller, B. 2013. Categorical and dimensional affect analysis in continuous input: Current trends and future directions. *Image and Vision Computing* 31(2):120–136.
- Gunes, H.; Schuller, B.; Pantic, M.; and Cowie, R. 2011. Emotion representation, analysis and synthesis in continuous space: A survey. In *Automatic Face & Gesture Recognition and Workshops (FG 2011)*, 2011 *IEEE International Conference on*, 827–834. IEEE.
- LaFrance, M.; Hecht, M. A.; and Paluck, E. L. 2003. The contingent smile: a meta-analysis of sex differences in smiling. *Psychological bulletin* 129(2):305.
- Mariooryad, S., and Busso, C. 2013. Exploring cross-modality affective reactions for audiovisual emotion recognition. *IEEE Transactions on Affective Computing* 4(2):183–196.
- Max Planck Institute for Psycholinguistics, The Language Archive, N. T. N. ELAN. <http://tla.mpi.nl/tools/tla-tools/elan/>.
- McKeown, G.; Valstar, M.; Cowie, R.; Pantic, M.; and Schröder, M. 2012. The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing* 3(1):5–17.
- Mehrabian, A. 1996. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in Temperament. *Current Psychology* 14(4):261–292.
- Metallinou, A.; Katsamanis, A.; Wang, Y.; and Narayanan, S. 2011. Tracking changes in continuous emotion states using body language and prosodic cues. *IEEE International Conference on Acoustics, Speech and Signal Processing* (June):2288–2291.
- Pantic, M.; Cowie, R.; D’Errico, F.; Heylen, D.; Mehu, M.; Pelachaud, C.; Poggi, I.; Schroeder, M.; and Vinciarelli, A. 2011. Social Signal Processing: The Research Agenda. *Visual Analysis of Humans* 511–538.
- Parthasarathy, S., and Busso, C. 2016. Defining emotionally salient regions using qualitative agreement method. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* 08-12-Sept:3598–3602.
- Pelachaud, C.; Heylen, D.; Pantic, M.; Vinciarelli, A.; Poggi, I.; D’Errico, F.; and Schroeder, M. 2012. Bridging the gap between social animal and unsocial machine: A survey of social signal processing. *IEEE Transactions on Affective Computing* 3:69–87.
- Ringeval, F.; Sonderegger, A.; Sauer, J.; and Lalanne, D. 2013. Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*.
- Ringeval, F.; Eyben, F.; Kroupi, E.; Yuce, A.; Thiran, J.-P.; Ebrahimi, T.; Lalanne, D.; and Schuller, B. 2015. Prediction of asynchronous dimensional emotion ratings from audio-visual and physiological data. *Pattern Recognition Letters* 66:22–30.
- Russell, J. A. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology* 39(6):1161–1178.
- Sacks, H.; Schegloff, E. a.; and Jefferson, G. 1974. A simplest systematics for the organization of turn taking for conversation. *Language* 50:696–735.
- Schlosberg, H. 1954. Three dimensions of emotion. *Psychological review* 61(2):81.
- Taylor, A., and Riek, L. D. 2016. Robot perception of human groups in the real world: State of the art. In *AAAI Fall Symposium Series: Artificial Intelligence for Human-Robot Interaction Technical Report FS-16-01*. Retrieved January, volume 4, 2017.
- Vinciarelli, A.; Pantic, M.; and Bourlard, H. 2009. Social signal processing: Survey of an emerging domain. *Image and Vision Computing* 27(12):1743–1759.
- Wittenburg, P.; Brugman, H.; Russel, A.; Klassmann, A.; and Sloetjes, H. 2006. Elan: a professional framework for multimodality research. In *Proceedings of LREC*, volume 2006, 5th.
- Wöllmer, M.; Eyben, F.; Reiter, S.; Schuller, B.; Cox, C.; Douglas-Cowie, E.; and Cowie, R. 2008. Abandoning emotion classes-towards continuous emotion recognition with modelling of long-range dependencies. In *INTERSPEECH*, volume 2008, 597–600.
- Yang, Y. H., and Chen, H. H. 2011. Ranking-based emotion recognition for music organization and retrieval. *IEEE Transactions on Audio, Speech and Language Processing* 19(4):762–774.