Emotion Elicitation and Capture among Real Couples in the Lab

George Boateng
ETH Zürich
Zürich, Switzerland
gboateng@ethz.ch

Laura Sels
Ghent University
Ghent, Belgium
laura.sels@ugent.be

Peter Kuppens
KU Leuven
Leuven, Belgium
peter.kuppens@kuleuven.be

Janina Lüscher
University of Zürich
Zürich, Switzerland
janina.luescher@psychologie.uzh.ch

Urte Scholz
University of Zürich
Zürich, Switzerland
urte.scholz@psychologie.uzh.ch

Tobias Kowatsch
ETH Zürich, University of St. Gallen
Zürich,St. Gallen Switzerland
tobias.kowatsch@unisg.ch

Paper presented at the 1st Momentary Emotion Elicitation & Capture (MEEC) workshop, co-located with the ACM CHI Conference on Human Factors in Computing Systems, Honolulu, Hawaii, USA, April 25th, 2020.

Abstract

Couples' relationships affect partners' mental and physical well-being. Automatic recognition of couples' emotions will not only help to better understand the interplay of emotions, intimate relationships, and health and well-being, but also provide crucial clinical insights into protective and risk factors of relationships, and can ultimately guide interventions. However, several works developing emotion recognition algorithms use data from actors in artificial dyadic interactions and the algorithms are likely not to perform well on real couples. We are developing emotion recognition methods using data from real couples and, in this paper, we describe two studies we ran in which we collected emotion data from real couples — Dutch-speaking couples in Belgium and German-speaking couples in Switzerland. We discuss our approach to eliciting and capturing emotions and make five recommendations based on their relevance for developing well-performing emotion recognition systems for couples.

Author Keywords

Emotion; Couples; Multimodal Sensor Data; Smartphone; Smartwatch

CCS Concepts

•**Applied computing** → **Psychology;** •**Human-centered computing** → *Ubiquitous and mobile computing systems*

and tools;

Introduction

Extensive research shows that intimate relationships have powerful effects on people's mental and physical health (see e.g. [23] for an overview). For instance, conflicts and negative qualities of one's intimate relationship are associated prospectively with morbidity and mortality [16]. Increasingly, researchers are zooming in on the emotional processes that take place in intimate relationships as underlying mechanisms for this relationship-health link (e.g. [9]. However, assessing these dynamic emotional processes is challenging.

In studies of intimate relationships, two methods predominate: self-reports and observer reports. Most often, a standard dyadic interaction paradigm is used, in which couples participate in an emotionally charged discussion that is videotaped [22]. Next, couples can watch these videos and report on the emotions that they have experienced during the interaction (resulting in self-reported emotion); or observers use a coding scheme to rate the interaction on specific emotional behaviors (e.g., the SPAFF [7]). Both methods have their own advantages and limitations and provide unique information on the emotional processes in couples. The power of observational data is that it goes beyond people's own awareness, and is not subjective to reporting biases. However, its greatest limitation is the resource use required in coding. First, a coding scheme has to be developed, which is a whole process in itself [12]. Next, multiple observers have to be trained in a systematic manner to obtain sufficient inter-rater agreement. When the actual coding can start, this process is slow and costly, and multiple coders have to code the same videos to allow obtaining inter-rater reliability.

Automatic emotion recognition holds important promise in meeting these limitations and significantly advancing the field. Hence, it is important to develop a system for automatic recognition of couples' emotions using information such as speech, facial expressions, gestures etc. Works that develop emotion recognition systems using speech data collected from individuals are not adequate for our purpose as such works do not capture the complexity of dyadic conversations such as turn-taking in couples' conversations. As a result, works that focus on couple dyads are most relevant.

Several emotion-recognition works using data from couple dyads involve data collected from actors in artificial dyadic interactions. Examples of these datasets are the IEMOCAP dataset [5], USC CreativeIT dataset [19], and MSP-IMPROV dataset [6]. To elicit emotions, actors are either asked to use a script or they are given hypothetical situations to act out so as to make the acting seem natural and more like a real couple. To capture ground truth, these works tend to be annotated later using either dimensional and or categorical labels and also either moment-by-moment or using global emotion labels of whole recordings.

There are several challenges with these annotations by external raters which are highlighted in this work [20] such as dealing with inter-rater agreement, the subjectivity of each rater, approaches to combine the annotations for moment-by-moment ratings and the laborious nature of these annotations. Additionally, and importantly, the ratings do not reflect the perceived emotions of couples which is necessary to capture rather than the assessment of external raters. Furthermore, it has been shown that algorithms trained on naturalistic data perform worse than those trained on acted data [8] and it is likely that algorithms developed from data collected from actors will not perform well on real people given that actors tend to express emotions with greater intensity as compared to naturalistic contexts and real couples. It is hence important to develop emotion recognition methods using data from real couples along with emotion ratings from them as well.

Towards that end, it is important to adequately collect ground truth information and sensor data to develop a system for emotion recognition among couples. We are developing such a system and, in this paper, we describe our approach to elicit and capture emotions among real couples in two lab studies — one conducted in Belgium with couples speaking Dutch and the other in Switzerland with couples speaking German. We then discuss these studies and make five recommendations for future data collection among couples in the lab to improve automatic emotion recognition. For work focusing on data collected from couples in everyday life, see our paper (under review) [2].

Methods

We used data from two lab studies with real couples, in which the sessions were videotaped and couples provided ratings either of the whole session or retroactively on a moment-by-moment basis while watching the video.

Study 1: Dyadic Interaction Study

A Dyadic Interaction lab study was conducted in Leuven, Belgium with 101 Dutch-speaking couples. These couples were asked to have a 10-minute conversation about a negative topic (a characteristic of their partner that annoys them the most) and a positive topic (a characteristic of their partner that they value the most) [29]. During both conversations, couples were asked to wrap up the conversation after 8 minutes. For the negative topic, they were also asked to end on good terms. After each conversation, each partner completed self-reports on various categorical emotion labels such as anger, sadness, anxiety, relaxation, happiness, etc. on a 7-point Likert scale ranging from strongly disagree (1) to strongly agree (7). Also, they completed the Affect Grid questionnaire [27] which captures the valence and arousal dimensions of Russel's circumplex model of emotions [25]. Each partner also completed their perception of their partner's emotion using the Affect Grid. Additionally, each partner watched the video recording of the conversation separately on a computer and rated his or her emotion on a moment-by-moment basis by continuously adjusting a joystick to the left (very negative) and the right (very positive), so that it closely matched their feelings, resulting in valence scores on a continuous scale from -1 to 1 [11, 24].

Study 2: DyMand Study

We are currently running a Dyadic Management of Diabetes (DyMand) lab study in Zurich, Switzerland with German-speaking couples in which one partner has type 2 diabetes with data from eight (8) couples collected so far [17]. In this lab study, the couple is asked to discuss an illness management–related concern that is causing them considerable distress for a 10-minute period. The session is videotaped and additionally, each partner wears a smartwatch as it collects various sensor data: audio, heart rate, accelerometer, gyroscope, and ambient light. After the session, each partner completes a self-report on a smartphone about their emotions using the Affective Slider [1] which assesses the valence and arousal dimensions of their emotions over the last 10 min of the discussion. Also, the smartphone takes a 3-second video of their facial expression while they complete the self-report.

Discussion and Recommendations

Based on these two studies, we discuss and recommend approaches to collect sensor and ground truth data from couples to aid in developing well-performing systems for emotion recognition among couples.

Elicitation of Emotions

In these studies, we elicited emotions in the couples by asking them to discuss various relationship-relevant topics (Study 1), or a distressing illness management concern (Study 2). In comparison to various elicitation approaches such as watching a video or listening to music, this approach leverages context which mimics a real-world context — partners having a conversation. Hence, the algorithms developed using data from this context like verbal and non-verbal vocalizations could then also be implemented in ubiquitous systems such as smartphones and smartwatch for couple emotion recognition from everyday life. We recommend the use of similar elicitation approaches for couple emotion recognition works.

Self-Report Data Collection

In these studies, we captured emotions using a range of approaches which can generally be grouped into two: global rating (one value or label for the whole conversation) and continuous rating (different values for different parts of the conversation) (only in Study 1).

The global ratings consisted of 7-point Likert scale for categorical emotions such as angry, relaxed, happy, sad, and the Affect Grid in Study 1 which were completed using electronic questionnaires. We collected valence and arousal values using the Affective Slider on a smartphone for Study 2. Global ratings are important to capture (1) a partner's perception of his/her emotion (self-perceived) and (2) his/her perception of his/her partner's emotion (partner-perceived) as was done in Study 1. The assessment of a partner's perception of his/her partner's emotion is useful and could be used to compute the baseline measures for metrics like accuracy (for classification task) and correlation coefficient (for regression tasks) of machine learning experiments.

The continuous emotion rating was done only in Study 1 by each partner separately by continuously adjusting a joystick to the left while watching a video of their conversation on a computer-based software (the rated valence values were displayed in real-time on). This continuous emotion rating is important as it gives a granular assessment of emotions which is important for developing an emotion recognition system that shows how the emotion of each partner is changing on a second-by-second or minute-by-minute basis. Also, the mean value could be used to get an estimate of the global emotion rating. Additionally, it could be useful for the accurate recognition of the global rating. Based on the peak-end rule, which says that the extremes and end of emotional experience influence a person's overall judgment of that emotional experience [10] and prior work exploring this rule using Study 1's data [29], using data from the extremes and or end of the 10-minute conversation might produce better emotion recognition performance of the global emotion rating of the whole conversation.

We did not collect self-reports about the personality of each partner though it might be useful. There are individual differences in the experience and expression of emotions with a concrete example shown in how the relation between arousal and valence varies across individuals [13]. Preliminary evidence suggests the valence and arousal emotional expressions of individuals relates to the five-factor model of personality [14]. Hence, individuals' personality may affect how they express their emotions. Hence, collecting information such as the Big Five Inventory [30] and using as input to an emotion recognition algorithm could potentially improve its performance.

Based on the discussion, we recommend collecting self-perceived and partner-perceived (1) global emotion ratings with smartphone-based valence and arousal instruments such as the Affective Slider and (2) continuous emotion ratings for valence and arousal using for example, a smartphone-based app. Categorical labels could also be collected if they are not additionally burdensome or redundant. We also recommend that personality self-reports also be collected. These will help in developing and evaluating robust emotion recognition systems.

Sensor Data Collection

In Study 1, we collected only audio and video data whereas in Study 2, we additionally use a smartwatch-based system we developed — DyMand — [3] to collect multimodal sensor data: audio, heart rate, accelerometer, gyroscope, and ambient light. The additional data collected from the smartwatch could provide more context for better recognition such as the heart rate providing physiological measures and the accelerometer and gyroscope providing information about hand gestures. Previous works have shown that multimodal approaches to emotion recognition perform better than unimodal approaches [21]. Given that an additional device like a commercial smartwatch is not burdensome to wear, we hence recommend the collection of such multimodal data.

Cross-Cultural Studies

The universality of emotions has been interrogated and questioned [26, 15]. There is evidence that suggest that culture affects how people experience and express emotions, for example, with facial expressions, gestures, physiological reaction, verbal and nonverbal vocalizations [18, 28]. Hence, algorithms developed using data from one cultural context might not work well in others, or worse, contain various biases. Collecting cross-cultural data will be useful in developing algorithms that work across various cultures and reduce bias in the algorithms. We collected data from different cultures albeit, only within Europe as of yet: Dutch-speaking couples in Belgium and German-speaking couples in Switzerland. We are developing and evaluating our emotion recognition systems using cross-cultural data. We hence recommend collecting data from couples in different cultures to develop robust algorithms.

Development of Software Tools

Data collected from real couples can be annotated by them as described previously and as a result, there is no need for manual annotation by external raters which is time-consuming and laborious. However, the data needs to be processed before they are useful for developing emotion recognition systems. There are some challenges involved in this process, some of which are unique to the context of dyadic interactions like couples' dyadic conversations such as turn-taking. Audio is an important data source for emotion recognition because various key information can be extracted such as vocal expression (how things are said), nonverbal vocalizations (eg. sigh, laughs) and verbal vocalizations (what is said which might give more context for recognition). Tools that perform automatic processing of audio data would improve the development of emotion recognition system for couples. Hence, it is important for various software tools to be developed that can easily be used by other researchers.

There is a need for open source tools for voice activity detection [4] and diarization [31] that are robust — perform well when used with all kinds of audio. The voice activity detection tool is needed to automatically annotate parts of the audio that contain vocalizations so either silent or noisy segments can be discarded. Additionally, the tool could be further refined to annotate specific nonverbal vocalizations like sighs, laughter, chuckles, etc. which might be indicative of specific emotions in various parts of the audio, thereby improve recognition performance. The diarization tool is needed to automatically annotate which parts of the audio correspond to each speaker. It is important to segment audio recordings into parts that correspond to each speaker to aid in developing a well-performing emotion recognition system.

Also, there is a need for open-source tools for automatic transcription of non-English languages (which are lacking) because using the transcriptions could provide more context and improve recognition performance. Doing the annotation and transcription manually for a few hours of audio might not be a problem. However, doing so for data in the tens of thousands of hours is not scalable. Approaches such as using Amazon Mechanical Turk may work for acted data but they cannot work for real couples' data because of their confidentiality. We recommend that efforts be put into developing these tools within the affective computing community to avoid individual duplicate efforts and also because inaccurate annotations would result in poor data input for the emotion recognition algorithms.

Conclusion

We are developing emotion recognition methods using data from real couples and in this work, we describe two studies we ran with real couples — Dutch-speaking couples in Belgium and German-speaking couples in Switzerland. We discuss our approach to eliciting and capturing emotions and make the following five recommendation based on their relevance for developing well-performing emotion recognition systems for couples: 1) Elicit emotions by asking couples to discuss a topic from their relationship, 2) Collect global and continuous emotion self-report and personality data using mobile systems like smartphones, 3) Collect multimodal sensor data using devices like smartwatches, 4) Collect data from different cultures and 5) Develop open-source voice activity detection, diarization, and transcription software tools within the affective computing community.

Acknowledgements

We are grateful to Prabakaran Santhanam and Dominik Rügger for helping with the development of mobile software tools in running the second study. Study 2 is co-funded by the Swiss National Science Foundation (CR12I1_166348).

REFERENCES

[1] Alberto Betella and Paul FMJ Verschure. 2016. The affective slider: A digital self-assessment scale for the measurement of human emotions. *PloS one* 11, 2 (2016), e0148037.

[2] George Boateng, Janina Lüscher, Urte Scholz, and Tobias Kowatsch. 2020. Emotion Capture among Real Couples in Everyday Life. Momentary Emotion Elicitation. In *Momentary Emotion Elicitation and Capture workshop. CHI 2020 (Under Review)*.

[3] George Boateng, Prabhakaran Santhanam, Janina Lüscher, Urte Scholz, and Tobias Kowatsch. 2019a. Poster: DyMand–An Open-Source Mobile and Wearable System for Assessing Couples' Dyadic Management of Chronic Diseases. In *The 25th Annual International Conference on Mobile Computing and Networking*. 1–3.

[4] George Boateng, Prabhakaran Santhanam, Janina Lüscher, Urte Scholz, and Tobias Kowatsch. 2019b. VADLite: an open-source lightweight system for real-time voice activity detection on smartwatches. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*. 902–906.

[5] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation* 42, 4 (2008), 335.

[6] Carlos Busso, Srinivas Parthasarathy, Alec Burmania, Mohammed AbdelWahab, Najmeh Sadoughi, and Emily Mower Provost. 2016. MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception. *IEEE Transactions on Affective Computing* 8, 1 (2016), 67–80.

[7] James A Coan and John M Gottman. 2007. The specific affect coding system (SPAFF). *Handbook of emotion elicitation and assessment* (2007), 267–285.

[8] Sidney K D'mello and Jacqueline Kory. 2015. A review and meta-analysis of multimodal affect detection systems. *ACM Computing Surveys (CSUR)* 47, 3 (2015), 1–36.

[9] Allison K Farrell, Ledina Imami, Sarah CE Stanton, and Richard B Slatcher. 2018. Affective processes as mediators of links between close relationships and physical health. *Social and Personality Psychology Compass* 12, 7 (2018), e12408.

[10] Barbara L Fredrickson. 2000. Extracting meaning from past affective experiences: The importance of peaks, ends, and specific emotions. *Cognition & Emotion* 14, 4 (2000), 577–606.

[11] John M Gottman and Robert W Levenson. 1985. A valid procedure for obtaining self-report of affect in marital interaction. *Journal of consulting and clinical psychology* 53, 2 (1985), 151.

[12] Patricia K Kerig and Donald H Baucom. 2004. *Couple observational coding systems*. Taylor & Francis.

[13] Peter Kuppens, Francis Tuerlinckx, James A Russell, and Lisa Feldman Barrett. 2013. The relation between valence and arousal in subjective experience. *Psychological bulletin* 139, 4 (2013), 917.

[14] Peter Kuppens, Francis Tuerlinckx, Michelle Yik, Peter Koval, Joachim Coosemans, Kevin J Zeng, and James A Russell. 2017. The relation between valence and arousal in subjective experience varies with personality and culture. *Journal of personality* 85, 4 (2017), 530–542.

[15] Nangyeon Lim. 2016. Cultural differences in emotion: differences in emotional arousal level between the East and the West. *Integrative medicine research* 5, 2 (2016), 105–109.

[16] Timothy J Loving and Richard B Slatcher. 2013. Romantic relationships and health. *The Oxford handbook of close relationships* (2013), 617–637.

[17] Janina Lüscher, Tobias Kowatsch, George Boateng, Prabhakaran Santhanam, Guy Bodenmann, and Urte Scholz. 2019. Social Support and Common Dyadic Coping in Couples' Dyadic Management of Type II Diabetes: Protocol for an Ambulatory Assessment Application. *JMIR research protocols* 8, 10 (2019), e13685.

[18] David Matsumoto and Paul Ekman. 1989. American-Japanese cultural differences in intensity ratings of facial expressions of emotion. *Motivation and Emotion* 13, 2 (1989), 143–157.

[19] Angeliki Metallinou, Chi-Chun Lee, Carlos Busso, Sharon Carnicke, Shrikanth Narayanan, and others. 2010. The USC CreativeIT database: A multimodal database of theatrical improvisation. *Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality* (2010), 55.

[20] Angeliki Metallinou and Shrikanth Narayanan. 2013. Annotation and processing of continuous emotional attributes: Challenges and opportunities. In *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*. IEEE, 1–8.

[21] Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. 2017. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion* 37 (2017), 98–125.

[22] Nicole A Roberts, Jeanne L Tsai, and James A Coan. 2007. Emotion elicitation using dyadic interaction tasks. *Handbook of emotion elicitation and assessment* (2007), 106–123.

[23] Theodore F Robles, Richard B Slatcher, Joseph M Trombello, and Meghan M McGinn. 2014. Marital quality and health: A meta-analytic review. *Psychological bulletin* 140, 1 (2014), 140.

[24] Anna Marie Ruef and Robert W Levenson. 2007. Continuous measurement of emotion. *Handbook of emotion elicitation and assessment* (2007), 286–297.

[25] James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology* 39, 6 (1980), 1161.

[26] James A Russell. 1994. Is there universal recognition of emotion from facial expression? A review of the cross-cultural studies. *Psychological bulletin* 115, 1 (1994), 102.

[27] James A Russell, Anna Weiss, and Gerald A Mendelsohn. 1989. Affect grid: a single-item scale of pleasure and arousal. *Journal of personality and social psychology* 57, 3 (1989), 493.

[28] K. R Scherer, H Wallbott, D Matsumoto, and K Tsutomu. 1988. Emotional experience in cultural context: A comparison between Europe, Japan and the United States. *Faces of emotion: recent research* (1988), 98–115.

[29] Laura Sels, Eva Ceulemans, and Peter Kuppens. 2019. All's well that ends well? A test of the peak-end rule in couples' conflict discussions. *European Journal of Social Psychology* 49, 4 (2019), 794–806.

[30] Christopher J Soto and Oliver P John. 2017. The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of personality and social psychology* 113, 1 (2017), 117.

[31] Eva Vozáriková and Jozef Juhár. 2015. Comparison of Diarization Tools for Building Speaker Database. *Advances in Electrical and Electronic Engineering* 13 (11 2015), 314–319. DOI: http://dx.doi.org/10.15598/aeee.v13i4.1468