Capturing Interaction Quality in Long Duration (Simulated) Space Missions with Wearables

Gedik, Ekin; Olenick, Jeffrey; Chang, Chu-Hsiang; Kozlowski, Steve W.J.; Hung, Hayley

DOI
10.1109/TAFFC.2022.3176967

Publication date
2022

Document Version
Final published version

Published in
IEEE Transactions on Affective Computing

Citation (APA)
Gedik, E., Olenick, J., Chang, C.-H., Kozlowski, S. W. J., & Hung, H. (2022). Capturing Interaction Quality in Long Duration (Simulated) Space Missions with Wearables. *IEEE Transactions on Affective Computing*, *14*(3), 2139-2152. Article 9780004. https://doi.org/10.1109/TAFFC.2022.3176967

Capturing Interaction Quality in Long Duration (Simulated) Space Missions With Wearables

Ekin Gedik, Jeffrey Olenick, Chu-Hsiang Chang, Steve W.J. Kozlowski, and Hayley Hung

Abstract—Space exploration is evolving with the recent increase in interest and investment. For the success of planned long-duration crewed missions, good interpersonal interactions between crew members are crucial. In this study, we evaluate the use of wearables for detection and estimation of the quality of each social interaction participants have throughout a long mission rather than aggregate measures of interactions. Our proposed method utilizes Temporal Convolutional Networks(TCNs) for extracting individual representations from acceleration and audio streams and learnable pooling layers(NetVLAD) to aggregate these representations into fixed-size representations. Use of NetVLAD layers provides an intelligent alternative to simple aggregation for handling variable-sized interactions and interactions with missing data. We evaluate our method on a 4-month simulated space mission where 5 participants wore Sociometric Badges and provided reports on their interactions in terms of effectiveness, frustration, and satisfaction. Our method provides an average ROC-AUC score of 0.64. Since we are not aware of any comparable baselines, we compare our method to hand-crafted features formerly utilized for cohesion estimation in similar scenarios and show it significantly outperforms them. We also present ablation studies where we replace the components in our approach with well-known alternatives and show that they provide better performance than their respective counterparts.

Index Terms—Learnable pooling, long duration space missions-, missing data, social interactions, temporal convolutional networks, wearable sensing

1 Introduction

Although humankind set foot on the moon a half century ago, human space exploration has largely been confined to near-Earth orbit since those pioneering Apollo missions so many years ago. That is beginning to change as major investments are being made to develop capabilities that enable humans to return to the moon and, using it as a base, to embark on long duration (LD) missions ranging from months to years to explore asteroids and eventually Mars. Although there are many physical challenges to surmount, LD space missions are also fraught with threats to the psycho-social health of the crew due to the isolated, confined, and extreme (ICE) nature of the mission. The crew will be isolated from family, friends, and colleagues in a small social world consisting of just four to six astronauts. Mars is some 450 million miles away from Earth, necessitating approximately eight to nine months for one-way transit. As space explorers embark on an interplanetary mission, increasing distance will make synchronous communication with Earth impossible due to transmission time lags. In addition, although space is vast, human habitation for space travel is spartan and confined, entailing very limited personal space; virtually no privacy; and minimal creature comforts for sleeping, personal hygiene, and recreation. Finally, space is an extreme environment, with persistent dangers from low gravity (i.e., loss of muscle mass, lower bone density), radiation (i.e., genetic mutations, increased cancer risk), and potential equipment failure that stress well-being and threaten life itself. In such ICE environments, it is essential that the crew members maintain good interpersonal interactions and team cohesion [1]. How can interaction quality (measured in terms of how efficient, satisfying and/or frustrating the interaction was perceived by the interacting group members) be captured unobtrusively and assessed automatically?

Pervasive sensing technologies have proven themselves to be good candidates for human and group behavior research since they allow unobtrusive monitoring of individual actions and interactions between peers with minimal disruption [2], [3], [4]. In addition to studies that utilize the onboard sensing capabilities of smartphones [3], [5], custom

- *Ekin Gedik and Hayley Hung are with the Socially Perceptive Computing Group, Technical University of Delft, 2628 Delft, XE, Netherlands. E-mail: {egedik, h.hung}@tudelft.nl.*
- *Jeffrey Olenick is with the Department of Psychology, Michigan State University, East Lansing, MI 48824 USA, and also with the Old Dominion University, Norfolk, VA 23529 USA. E-mail: jolenick@odu.edu.*
- *Chu-Hsiang Chang is with the Department of Psychology, Michigan State University, East Lansing, MI 48824 USA. E-mail: cchang@msu.edu.*
- *Steve W.J. Kozlowski is with the Department of Psychology, Michigan State University, East Lansing, MI 48824 USA, and also with the University of South Florida, Tampa, FL 33620 USA. E-mail: skozlowski@usf.edu.*

Manuscript received 28 Mar. 2021; revised 11 May 2022; accepted 12 May 2022. Date of publication 23 May 2022; date of current version 13 Sept. 2023.
This work was supported in part by the Netherlands Organization for Scientific Research (NWO) through the MINGLE (Modelling Social Group Dynamics and Interaction Quality in Complex Scenes using Multi-Sensor Analysis of Non-Verbal Behaviour) Project under Grant 639.022.606. Data used in the preparation of this paper were provided by S. W. J. Kozlowski, S. Biswas, and C.- H. Chang. Their research, which generated the data, was supported by the National Aeronautics and Space Administration (NASA) under Grant NNX13AM77G.
This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by National Aeronautics and Space Administration, John Space Center (NASA JSC IRB) under Application No. Pro0909, and performed in line with the Key contributors to the maintenance and regulation of team function and performance on long duration exploration missions.
(Corresponding author: Ekin Gedik.)
Recommended for acceptance by A. Vinciarelli.
Digital Object Identifier no. 10.1109/TAFFC.2022.3176967

wearable sensing platforms such as Sociometric [6] and Rhythm Badges [7] have been designed and developed specifically for organizational scenarios in mind. Such platforms aim to augment traditional ID badges worn in organizations with sensing capabilities, allowing large-scale, longitudinal collection of individual and group behavior. Even though in theory, such collection platforms Pervasive sensing may also be a good candidate for capturing interaction behavior and quality for ICE teams.

With respect to prior work using pervasive sensing in organizational and other settings, we argue that representations of those social interactions were generally quite simplistic and were based on basic aggregated measures such as the frequency of the interactions, length of the interactions, etc. [4], [5], [6], [8], [9]. However, findings from social psychology showed that the quality of interactions is, at least, as important as their frequency for a healthy social environment [8]. In this paper, we address this shortcoming by explicitly focusing on the automatic estimation of interaction quality in terms of multiple labels (effectiveness, frustration, and satisfaction of the interaction) provided by participants through experience sampling methods. This capability is important if sensors are to be useful for detecting interaction quality problems and, potentially, triggering an intervention to maintain group functioning.

One recurring issue in longitudinal studies is the case of missing data [10]. It is common to have people dropping out of the studies, faulty sensors, and even participants forgetting to switch on and/or wear sensors for some intervals of the data collection. In particular, non-monotone missing data patterns, where data is missing for a subject in some slices and present in the others, were known to present a considerable modeling challenge. Such problems can be mitigated to an extent in scenarios where representations are aggregated either over a group or a long time interval. However, for studies that focus on fine-grained analysis and estimation, the challenge presented by missing data cannot be ignored. In addition, varying lengths of interactions also poses a challenge for the analysis and estimation, especially when individual sensing mediums such as personal wearable sensors are used. A fixed-size representation of an interaction should be extracted from the data of varying number of individuals. Traditionally, basic statistical measures such as the mean, maximum, and minimum values of individual representations were used for this purpose [4]. However, this might result in the loss of crucial information for fine-grained analysis tasks.

Another challenge specific to our research focus is the association of provided labels to the sensing data. Basically, exact timestamps when a reported interaction started and finished is not known and it should be estimated from the sensor data. Due to the large body of work on interruptibilty [11], we know that it is hard to obtain ambulatory assessments that match the period of time for which an event occurred. We are not aware of any works that have tried to balance ambulatory assessments in such settings while keeping ease of use for the subject. In an ideal setting, we would ask participants to exhaustively report all interactions they had and what the quality of that interaction was. However, we have seen in the case of the collected data that even with highly motivated participants who are willing to be stuck in an isolated situation such as a simulated space mission, compliance is still not complete. In this case, we consider an easier approach for the participants where they are asked only to report twice a day and only the last interaction that they had.

In this study, we propose a method to overcome the aforementioned challenges. The primary contributions of this paper are as follows: (i) most importantly, we estimate the quality of social interactions a participant has throughout a long-term mission rather than aggregate measures of interactions across the time frame, (ii) we propose an heuristic-based approach to identify the intervals of interactions from Infrared and Bluetooth data, effectively providing a solution to problem of associating relatively sparse labels to continuous longitudinal sensor data, (iii) we employ learnable pooling layers that are mostly used in computer vision (NetVLAD [12]) in a novel way to pool individual representations of interacting participants into a fixed-size representation of the interaction, providing flexibility of dealing with interactions of varied sizes and making it possible to analyze cases where data from one or more participants are missing, (iv) we evaluate the use of Temporal Convolutional Networks (TCNs) [13], a recent variation of Convolutional Neural Networks (CNNs) which are proposed to model audio data, to automatically extract informative representations from the raw sensor readings rather than using handcrafted features (v) we treat the estimation problem as a multi-task learning one [14] where multiple interaction quality labels provided by the participants (effective, frustrated and satisfied) are jointly estimated.

2 Related Work

To our knowledge, there exists no prior work that specifically focuses on the automatic detection of interaction quality from wearable sensors in longitudinal scenarios. Hence, we will be presenting, in no specific order, existing works from the literature that are similar to our task in one or more of the following categories: used sensing modalities (wearables), types of analyzed scenarios (long-term), final goals (estimation of social concepts), type of techniques used for handling missing data, feature extraction (TCNs) and, estimation (multi-task learning).

During the past two decades, wearable and mobile sensors have been used to monitor and analyze various individual and group related phenomena in the long term. One of the first works in this direction was from Choudhory *et al.* where sociometers, a former version of the sociometric badges, were employed for approximately 21 days by a total of 31 participants [15]. Using IR and audio from the badges, authors have mined the interaction networks of users and extracted information related to the group structures. Olguin *et al.* then moved the focus to organizational scenarios with a similar approach where wearable sensors are utilized to measure the frequency of face-to-face interactions. Together with several other information sources (e-mail, surveys, etc.), they analyzed concepts such as the personality of the participants in terms of the "Big Five" model [16] and the cohesion of the teams [6]. Another study that evaluated the use of sociometric badges in organizational

scenarios is [17] where Lepri *et al*. proposed a corpus of digital data (sensor and phone and email logs) collected from 53 participants for six continuous weeks. This corpus also included personal and situational data collected via surveys and experience sampling, and focused on constructs such as personality, affect and productivity. Rhythm, a platform that combines wearable electronic badges (which are more pervasive compared to those mentioned earlier) and online applications for analyzing social interactions between teams, divisions, and locations were introduced by Lederman *et al*. in 2018 [7]. In addition to an in-depth analysis of short-term meetings, they also presented a use-case where the platform is employed for analyzing connectivity patterns during a three-day workshop.

Another promising direction for the long-term use of wearable and mobile sensors is the monitoring of well-being. There exists a large body of work that focuses on the estimation of concepts related to well-being such as stress, mood, and affective states. Various information sources have been used for this purpose, ranging from physiological signals such as skin conductance and heart rate [18] to video [19] and digital traces [20]. The most relevant works in the scope of this paper are the ones that employ some type of information related to the participants' interactions over time, either sensed through mobile sensors [5], [9], [21], [22] and/or mined from digital traces [20], [22], [23]. Perhaps out of these studies, the one closest to our task in terms of methodology and setting is [23], where they predicted health, stress and, happiness of participants from longitudinal sensor data with a multi-task learning formulation [14]. They employed multi-task learning with two different setups, one for predicting these three metrics simultaneously and one for personalization and showed both approaches perform better than using a single-task learning setup. The results of these studies have shown that interactions play a crucial role in a person's well-being. However, to our knowledge, none of these studies have tried to estimate the quality of individual interactions as we do in this paper, but they rather relied on aggregations such as the frequency of interactions. Moreover, interactions were generally simplified into dyadic connections.

In the last five years, multiple researchers have shifted their focus from corporate, hospital, and campus settings to a more in-depth analysis of the group dynamics of small teams in ICE (Isolated, Confined and Extreme) scenarios [1], [2], [4], [24]. Such teams were shown to have inherently different dynamics than larger organizational settings [25]. Since providing outside intervention is challenging in ICE scenarios, maintaining healthy interaction patterns between team members is crucial. Perhaps [24] and [4] are the two studies that are most similar to ours in terms of the setting, used data sources, and the task of automatically estimating interaction patterns in teams. Both studies used Sociometric Badge data collected during a 4-month simulated space exploration mission. Zhang *et al*. employed topic models to mine interaction patterns from IR data [24] and a later paper by the same authors focused on the automatic estimation of individual affect and group cohesion using IR, acceleration, and audio data [4]. Still, both works analyzed dyadic interactions only, used simple aggregation, and did not try to explicitly model the quality of interactions.

The majority of the works mentioned up until this point relied on hand-crafted features for representing the concept they are trying to estimate [4], [5], [9], [16], [20], [21]. With the so-called deep learning revolution first sparked in the computer vision domain, many researchers who utilize wearable sensing data recently shifted their focus to neural network models that can automatically extract representative features from raw sensor readings [22], [26], [27], [28], [29]. The results of these studies have shown that learned features consistently outperform hand-crafted ones for a variety of tasks such as activity and speech recognition. Temporal Convolutional Networks (TCNs) are a relatively new neural network architecture designed for modeling sequential data [13], [30]. TCNs are shown to outperform competing methods such as recurrent neural network architectures like Long-Term Short Memory (LSTM) and Gated Recurrent Units (GRU) [31], [32] in various sequence modeling tasks [33]. Especially, their competitive performance in modeling acceleration [34], [35] and audio data [30] makes them a great candidate for our task.

Missing data is a well-known challenge for field research. The challenge is especially acute for longitudinal research where multiple measurements for the same individuals or phenomena are collected over a long time period [10], [36]. There could be various reasons for the absence of data such as participants dropping out from the study, faults in the measurement tools, and even people forgetting to use the measurement tools. Thankfully, mechanisms of missing data are theoretically well-studied. The types of missing data can be grouped under three categories: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) [36], [37]. Detailed analysis of these concepts is out of the scope of this paper (please refer to [37] for more information) but their consequences are. Basically, statistical analysis in the case of MCAR and MAR is expected to yield unbiased parameters estimates whereas it can't be guaranteed for MNAR cases. So, in case of missing data, the underlying causes of the absence need to be checked if it is possible as best practice to make sure the estimates are unbiased.

Even though missing data is highly probable for large-scale, real-life, longitudinal ambulatory assessment studies, most of the mentioned works do not discuss the effects and outcomes of missing data on the presented results. This is mostly acceptable when the analysis is done on a day-level or on even longer time periods since one can then argue that a couple of hours of missing data is negligible in the greater scheme. However, studies that focus on the fine-grained analysis of events that happen throughout a long-term data collection, like our scenario, do not have this possibility. Conceptually, ambulatory health monitoring has a similar task to ours where short-term analysis of continuous readings is crucial. Scientists working in this domain have discussed the effects of missing data and proposed imputation based solutions [38], [39], [40], [41]. However, imputation is mostly applicable for either non-complex data or short intervals of absence covered by present data. Montori *et al*. discussed various solutions to the sparse data problem in the context of mobile crowd sensing for Internet of Things, such as compressive sensing, piggybacking and edge deduplication [42]. Recently, reinforcement learning

(a) (b)

Fig. 1. (a) The structure participants resided. (b) A Sociometric (SS) badge.

based approaches started to become popular, such as the one presented in [43]. In this study, we propose a novel scheme by utilizing learnable pooling layers, more specifically NetVLAD [12]. Basically, using a learnable pooling layer makes it possible to utilize the data of remaining participants in an interaction even though data from some are missing. Moreover, it also allows mapping individual representations to a fixed-size interaction representation in a smarter way than simple statistical aggregation which was the preferred methodology for former studies [24].

3 Data

We use the dataset HI-SEAS (Mission 2) which was collected during a simulated space mission over four months, early 2014 [44]. Originally, there were 6 volunteers, 2 males and 4 females, identified as white and each having at least a bachelor's degree. One participant, who retired early from the study for personal reasons, is not included in our analysis. During the mission, participants lived in a confined structure that simulates an environment the crew would inhabit during a short-duration space flight. This structure is approximately 6 meters in diameter and has three stories and is shown in Fig. 1a.

Each crew member has formal a role imitating a real-world flight crew: a commander, medical officer, engineer, science officer, architect, and a biologist. During the lifetime of the mission, volunteers were tasked with performing team-oriented objectives with some induced constraints mimicking real ones that would be faced by a Mars flight crew, such as an outside communication delay of 20 minutes. In addition to the assigned tasks, crew members also have unstructured personal time, mostly in the evenings.

Volunteers wore Sociometric Badges (SS Badge) [6], shown in Fig. 1b, while they are awake, excluding personal times such as showering and exercising. They were also expected to provide daily reports twice a day via experience sampling method (ESM). Due to the fact that the team was mostly autonomous after the missions start and not directly supervised by the outside researchers, there were numerous cases where the SS Badge or ESM data is missing.

3.1 Wearable Sensing Data

The SS Badges were worn around the neck and recorded the following data types:

- Movement: The SS Badges have an onboard tri-axial accelerometer with a sampling rate of 20 Hz. The raw data is processed online and only the average values over a pre-configured time resolution, 2 Hz in our case, are logged. The logged data are the acceleration in X, Y, Z dimensions, movement energy, and consistency. The movement energy is the amplitude of the acceleration computed over the three dimensions. The movement consistency is the stability of the movement energy, ranging between 0 (no change) and 1 (maximum variance).
- Audio: The SS Badges have an onboard microphone with a sampling rate of 8 kHz. Similar to the accelerometer, the data is processed and only statistical vocal features computed from 0.5-second windows are logged. These vocal features are the average, standard deviation, variance, minimum and maximum values of the amplitude of the audio signal.
- Infrared: The SS Badges have a forward-facing IR receiver and transmitter. IDs of other badges that are in the transmission range are logged with a time resolution of 1 s.
- Bluetooth: The badges periodically transmit their unique ID via Bluetooth and scan for other badges' broadcasts. Received Signal Strength Indicator (RSSI) values, which act as a rough proxy for distance, received from other badges are logged every 30 seconds.

3.2 Survey Response Data

All members were requested to fill questionnaires about their individual affective status, perceptions of team cohesiveness, and the quality of interactions they had with other team members, twice a day. In this study, we focus on the labels regarding the interactions. For more information about other survey data that was collected but not used for this study, please refer to [4]. In the questionnaire, participants were asked to give information about the *last interaction* they had with the other team members. They first provided the unique subject IDs of the team members that were a part of this interaction. Then, they rated the following statements using a 6-item scale:

- Please indicate your agreement with the following statements about / your experiences with the interaction:
  - I was satisfied.
  - I was frustrated.
- How effective was the interaction? (That is, did you accomplish what was intended?)

Pearson's correlation coefficient of effective-frustration, effective-satisfaction and frustration-satisfaction are computed to be $-0.44, 0.63$ and -0.55, respectively ($p < 0.01$). These moderate negative and positive correlations show all three constructs are conceptually different while being related and have varying function with respect to our outcome variables.

3.3 Dataset Statistics

As mentioned earlier, there are cases where no data exists for a given ESM entry. Eliminating faulty ESM entries and

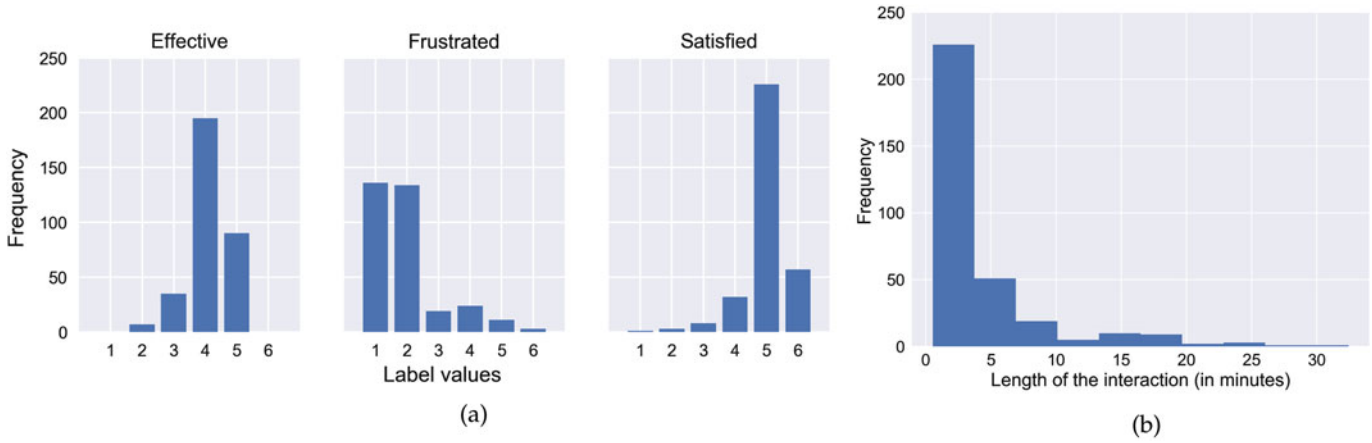

Fig. 2. (a) Label distributions (b) Interaction length distributions.

cases with no wearable sensing data from any of the interacting participants, 327 interactions remain. For more information about how the actual starting and ending timestamps of an interval are determined, please refer to Section 4.1.

Distribution of the Labels. Fig. 2a shows the distribution of effective, frustrated, and satisfied labels. As can be seen from the figure, the distributions are highly imbalanced, favoring higher ratings for effective and satisfied labels and lower ratings for the frustrated labels. In order to reduce this imbalance slightly, we treated the problem as a 3-class classification task where labels 1 and 2 formed the low, 3 and 4 formed the medium, and 5 and 6 formed the high rating classes.

Distribution of the Length of the Interactions. Fig. 2b shows the distribution of the lengths of the interaction in minutes. Since the IR detections can be quite sparse and the Bluetooth is sampled every 30 seconds, we empirically selected one minute as the minimum length for a meaningful interaction. This selection is done by observing interaction lengths in our research group's existing data sets. Thus, every interval that is found to be less than one minute is extended to cover an interval of one minute. The majority of the interactions have a length between one to ten minutes, still, the distribution is quite varied, even showing interactions spanning up to approximately 30 minutes.

Distribution of the Interactions Sizes. Fig. 3a shows the distribution of the number of participants in detected interactions. The blue bars represent the number of participants that are included in the ESM entries. The green bars, on the other hand, shows the actual distribution for which we have wearable sensing data. According to the ESM entries, the majority of the interactions include all the group members. However, for many interactions, data from a varying number of participants are missing, showing a striking difference between subjective reports and those detected by the wearables. This is an interesting result highlighting the challenge of asking people to appraise episodic memories and relating this to sensor data [45].

Missing Data. Fig. 3b shows how many participants' wearable data are missing from the detected interactions. The majority of the cases either have wearable sensing data for all participants or data from one participant is missing. In total, only ~34% of the interactions have data for all the participants that were included in the ESM entry.

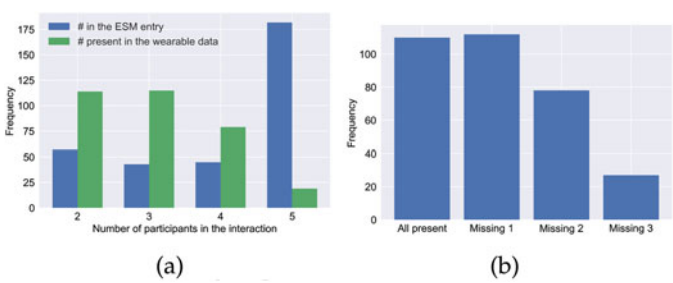

Fig. 3. (a) Distributions of group sizes (b) Missing data statistics.

4 Methodology

Fig. 4 visualizes the overall workflow of the proposed method for interaction quality estimation. Here are the basic steps of the method:

1) Identification of the relevant interaction interval per rating from Bluetooth and IR data (Section 4.1)
2) End-to-end training and estimation (Section 4.2)
   a) Extraction of acceleration and audio data from the detected intervals for participants that are in the interaction,
   b) Extraction of individual representations for each participant's data with TCNs,
   c) Pooling of individual representations with a NetVLAD layer for obtaining a fixed-size representation of the interaction,
   d) Multi-task learning using the fixed-length representation for predicting effectiveness, frustration, and satisfaction labels.

4.1 Identification of the Relevant Interaction Interval Per Rating

In addition to the information presented in Section 3.2, each ESM entry also includes the timestamps for the time the participant started and finished inputting that entry. The task here is to use the information in the ESM entry to detect the actual interval where the interaction depicted in the entry happened. This is a challenging task and most former works circumvented this step by not detecting precise

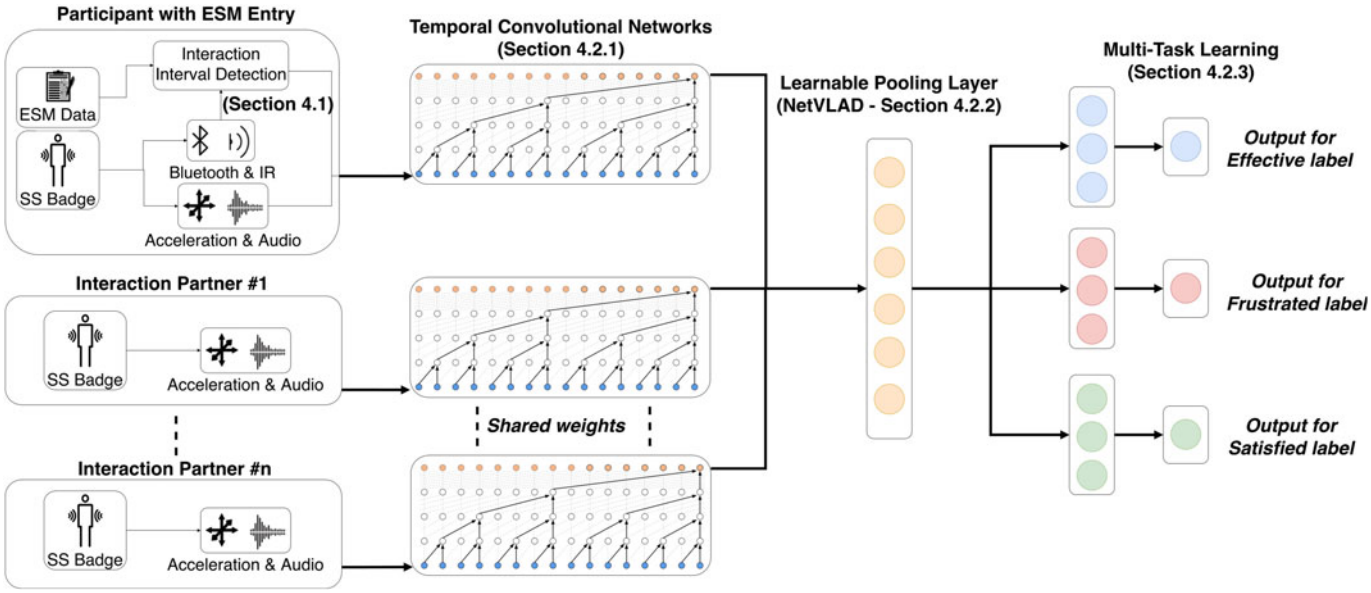

Fig. 4. Simplified workflow of the proposed method (TCN visualizations are taken from [30]).

interactions but rather treating all pings from Bluetooth and IR, which mostly accounts for co-location, as interactions [4]. Since there is no ground truth for the actual starting and ending times of the interactions depicted in the entries, a supervised approach is not possible. Thus, we used a heuristic approach which is built on the following premises:

- Interaction should be on the same day as the ESM entry.
- People in the interaction, detected using the Bluetooth and IR readings from the wearables, should match the people reported in the ESM entry as much as possible. At least one of the participants reported should be in the detected interval. If no interval is found satisfying this requirement on the day of the report, the ESM entry is not used.
- Since the participants are asked to report on their last interaction, intervals that are closer to the ESM entries reporting times are favored.
- Longer intervals are favored.

The interaction interval detection procedure has the following steps:

1) Bluetooth and IR data of the participant who input the entry are processed to obtain streams. IR data directly includes the IDs of badges that are in close proximity whereas Bluetooth data has the RSSI values of other badges. Consulting the literature, we found -75 to be a good cutoff value for co-location [46]. This way, we end up with similar streams for both IR and Bluetooth, each having the timestamps and IDs of badges that are found to be in close proximity.
2) Since both IR and Bluetooth entries were found to be quite sparse, we group entries that are temporally close to each other using a predefined threshold t_1 to form an interval.
3) Three fitness values are calculated for each interval as:
   - *How well the participant IDs detected in the interval matches the report:* $f_1 = 2 * precision * recall / (precision + recall)$ where $precision = TP / (TP + FP)$ and $recall = TP/(TP + FN)$. TP, FP and FN corresponds respectively to number of participants that are both in the interval and the ESM entry, the number of participants that are in the interval but not in the ESM entry, and the number of participants that are not in the interval but are in the ESM entry.
   - *The recency of the interval to the ESM entry:* $f_2 = 1 - ((ts_1 - ts_2)/86400)$ where ts_1 and ts_2 are the timestamps for the ESM entry and the ending of the found interval. 86400 is the total number of seconds in a day which used for normalizing f_2 between 0 and 1, in correspondence with $f1$.
   - *The identified interval length:* $f_3 = (ts_{end} - ts_{start})/86400$ where ts_{end} and ts_{start} are the timestamps for the end and the start times of the found interval.
4) These three fitness values are averaged to obtain a final fitness value for the interval: $f_{final} = (f_1 + f_2 + f_3)/3$
5) Intervals with the highest f_{final} value are selected from the Bluetooth and IR data. If the selected intervals from both modalities agree (temporally close to each other than a predefined t_2), intervals are fused to obtain the final interval. If not, the interval with the highest f_{final} value is selected to be the final selected interval.

This procedure is heavily inspired by the one proposed in [45]. Basically, we added a third fitness value based on the length of the discovered interactions since we saw that the majority of interaction intervals detected by the method of [45] was too short and can be easily a false positive. With the addition of the third fitness value we were able to identify intervals where partners are continuously detected for a longer amounts of time, resulting in more robust detections.

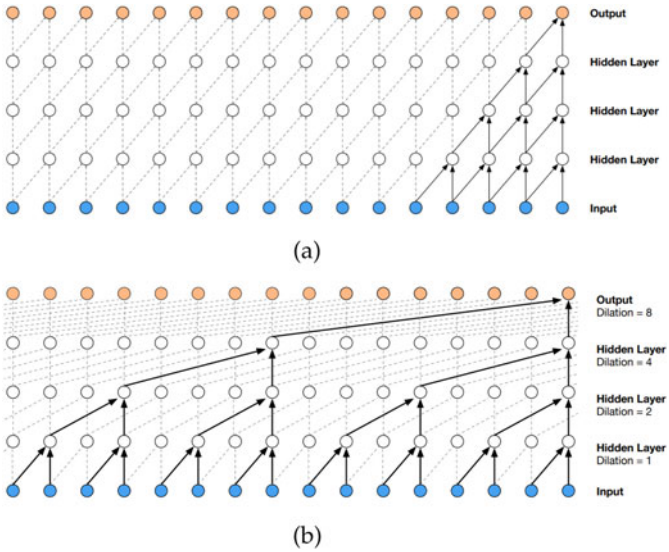

Fig. 5. (a) A stack of causal convolutions, from [30]. (b) A stack of dilated causal convolutions, from [30].

Moreover, we included the Bluetooth readings in the methodology in addition to IR used in [45].

Empirically, we found the best values for t_1 and t_2 to be two and five minutes. Since the entries can be still quite sparse, any detected interval shorter than one minutes is extended on both sides to cover a minute. This procedure results in the identification of 327 interaction intervals with their matching ESM data.

4.2 End-to-End Training and Estimation

After the intervals are found, acceleration and audio data recorded during these intervals, for the participants that were tagged to be in that interaction, are extracted to form the dataset for our automatic estimation experiments. The data are fed into an end-to-end neural network architecture with three distinct components which will be discussed in the following subsections.

4.2.1 Temporal Convolutional Networks

We use Temporal Convolutional Networks to model the accelerometer and audio data of participants in an interaction. In other words, we automatically extract descriptive representations rather than using hand-crafted features. TCNs are selected for this task since multiple previous studies showed that they consistently outperform competing approaches such as recurrent neural network architectures (LSTMs) and hand-crafted features in modeling accelerometer [34], [35] and audio data [30]. TCNs are built on two ideas: causal and dilated convolutions which are visualized in Figs. 5a and 5b. Stacked causal convolutions allow the processing of sequential data while making sure the ordering of the data is not violated during modeling, e.g $p(x_{t+1}|x_1, \ldots, x_t)$, prediction at timestep t, is only dependent on the previous timesteps [30]. However, causal convolutions require many layers for modeling long sequences as it can be seen from Fig. 5a, where the model with three layers has a receptive field of 5 [30]. This shortcoming is mitigated with the use of dilated convolutions where input values to convolve are skipped with a predefined step. The model visualized in Fig. 5b has an effective receptive field of 16 with the same number of layers. With carefully selected filter size, the number of layers, and dilation, it is possible to model sequences of arbitrary sizes.

The interactions in our dataset have varied lengths, as shown in Fig. 2b. Thus, we selected the appropriate number of layers, kernel and dilation size to make sure we have a large enough receptive field to cover even the longest sequences. In order to allow batching while training and testing, each sequence is padded to the longest sequence in a batch. Real lengths of the sequences are stored and the TCNs output at the corresponding timestep is selected as the output. In our setup, we have two important design choices. First, we use the late fusion of the two modalities. Thus, we use separate TCNs for modeling the accelerometer and audio data. This choice was based on a previous study that showed the learnable pooling layer, NetVLAD, performs better when separate layers are used for separate modalities [47]. Second, we distinguish between the data of the specific participant who logged the ESM entry and the other participants reported to be present in the interaction. This choice is based on the fact that multiple participants can report on the same interaction while having different experiences of it. This ego-centric perspective is also in keeping with other prior analyses of this data [45].

In summary, four TCNs are trained: two for the acceleration and the audio data of the participant who logged in the ESM entry; and two for the acceleration and the audio data of the remaining participants in the interaction. Each participant's data is fed into the corresponding TCN and the resulting representations are then fed into the next NetVLAD layers for pooling. So for an N person interaction, a $2 \times N \times D$ representation is obtained where D is the dimension of the representation. The procedure can be also thought of as each participant having their own TCNs but the TCNs for the participants other than the participant who logged the ESM entry share weights.

We compare the performance of TCNs to hand-crafted features [4] and Long Short-Term Memory (LSTM) networks [31] in our experiments. The first comparison serves to show the advantages of automatic extraction of representations over the explicit design of them. Second comparison is to show how TCNs compare to another widely used automatic feature extraction method. Implementation details and results of these comparisons are presented in Sections 5.2.1 and 5.2.4.

4.2.2 Learnable Pooling (NetVLAD)

Traditional pooling that is widely used in Convolutional Neural Networks can be described as a technique for down sampling feature maps. A pooling layer computes one value for describing a $N \times N$ patch of the feature map, either by taking the average or maximum value in the patch. In this paper, we propose to use this process for pooling the representations of different participants wearable sensing data extracted with TCNs. For example, an average pooling procedure will then compute the average value for the each dimension of the extracted representations from individual participants. Computationally, this procedure is same with mean aggregation of feature values.

The advantage of pooling is that it makes it possible to obtain a fixed-size representation of an interaction regardless of the number of the participants inside that interaction. Since any number of representations can be aggregated this way, it does not matter if some of the participants' data are missing. The remaining participant's data can still be used to obtain a representation for the interaction.

Recently, a smarter procedure for pooling is proposed. Rather than taking the average or maximum value, authors of [12] proposed to learn this computation,. Their method named NetVLAD is based on Vector of Locally Aggregated Descriptors (VLAD), a widely used descriptor pooling method in image retrieval and classification which stores the difference vector between a descriptor and its corresponding cluster center [48]. Originally, it has one parameter K which denotes the number of clusters. Given N D-dimensional descriptors, it returns a representation V which is $K \times D$-dimensional. With descriptors as $\boldsymbol{x_i}$ and c_k as cluster centers, the (j, k) element of V is computed as

$$V(j,k) = \sum_{i=1}^{N} a_k(\boldsymbol{x_i})(x_i(j) - c_k(j)) \tag{1}$$

where $x_i(j)$ and $c_k(j)$ are the jth dimension of the ith descriptor and kth cluster center [12]. The matrix V is first L2-normalized column-wise and then transformed into a vector to obtain the final representation. This representation is also L2-normalized. The problem with this formulation is that it is not differentiable with respect to all parameters and the input, hence it can not be used as a neural network layer that is trained with backpropagation. The discontinuity of VLAD is caused by the hard assignments of descriptors to the cluster centers. In order to make VLAD differentiable, authors of [12] proposed to replace hard assignment to a single cluster with soft assignments to multiple clusters as follows:

$$a_k(\boldsymbol{x_i}) = \frac{e^{-a||\boldsymbol{x_i} - c_k||^2}}{\sum_{k'} e^{-a||\boldsymbol{x_i} - c_{k'}||^2}}, \tag{2}$$

which can be rewritten as

$$a_k(\boldsymbol{x_i}) = \frac{e^{\boldsymbol{w}_k^T \boldsymbol{x}_i + b_k}}{\sum_{k'} e^{\boldsymbol{w}_{k'}^T \boldsymbol{x}_i + b_{k'}}} \tag{3}$$

by expanding the squares. The vector $w_k = 2a\boldsymbol{c_k}$ and the scalar $b_k = -a||\boldsymbol{c_k}||^2$. a is a parameter that decays the soft assignment with increasing distance to the cluster center. Replacing this into Equation (1), the final form of NetVLAD is then obtained as

$$V(j,k) = \sum_{i=1}^{N} \frac{e^{\boldsymbol{w}_k^T \boldsymbol{x}_i + b_k}}{\sum_{k'} e^{\boldsymbol{w}_{k'}^T \boldsymbol{x}_i + b_{k'}}} (x_i(j) - c_k(j)), \tag{4}$$

where $\boldsymbol{w_k}$, b_k and $\boldsymbol{c_k}$ are all trainable parameters for each cluster k [12].

Based on NetVLADs prior success on various tasks [47], [48], we decided to use it in our architecture. To see NetVLADs advantages over simple aggregation, we compare its performance to average pooling in Section 5.2.2. Based on the former findings presented in [47], we utilize two NetVLAD layers: one for the TCN representations from the accelerometer data; and one for the TCN representations from the audio data. In this step, the representations of the participant who logged the ESM entry are pooled together with the representations of remaining participants. For an interaction with N participants, the TCNs will return a $2 \times N \times D$ representation which will be transformed into a $2 \times (D \times K)$ representation by the NetVLAD layers. Inspired by the implementation by [47], we also added a fully connected layer to the output of the NetVLAD layer which allows to control the dimension of the output, transforming the $(D \times K)$ dimensional representations into a selected new dimension of D_{new}. Thus, the final output of the pooling step is two vectors, representing the interaction extracted from the accelerometer and the audio data, each with a dimension of D_{new}.

4.2.3 Multi-Task Learning

The proposed architecture for MTL is quite similar to one presented in [23]. Two output vectors from the previous step are concatenated to form a $(2xD_{new})$ vector which is then fed into three different branches of fully-connected layers, each corresponding to the one of the labels the network will predict: effective, frustration, and satisfaction. With this multi-task learning setup, the representations extracted from the TCNs and the the parameters for pooling will be learned jointly for all three labels. Separate branches of fully-connected layers after the pooling will be trained specifically for the corresponding label. After the forward pass, the loss for each label is calculated separately and then summed and backpropogated. To evaluate the contribution of the MTL setup, we compare it to a single-task learning setup in Section 5.2.3.

5 Results and Discussion

5.1 Experiment Setup

We used a slightly modified 5-Fold cross-validation scheme to evaluate the performance of our proposed approach. In this scheme, 60%, 20% and 20% of the intervals are used as the training, validation and test sets. This split is repeated 5 times, ensuring that all of the intervals were part of the test set in one of the splits. Since we are trying to estimate multiple labels, the folds are stratified using iterative stratification [49]. This procedure ensures that the class distributions are similar in the training, validation and test sets. This cross-validation scheme does not guarantee a person-independent approach since data from the same subject (but not the same interaction) can be both in training and test sets. We believe there is an intrinsic person-dependent aspect since different participants can have different labels for the same meeting and all data is intrinsically associated as a pair or group in the team. Moreover, this setup allows us to utilize and evaluate for interacting groups of all sizes in contrast to a pure person-independent setup where groups of four and five would need to be excluded completely. The most apparent real-life implication of the current setup is that it does not guarantee generalization to new subjects out-of-the-box. Since the method is proposed for longitudinal scenarios, an onboarding process at the start where data

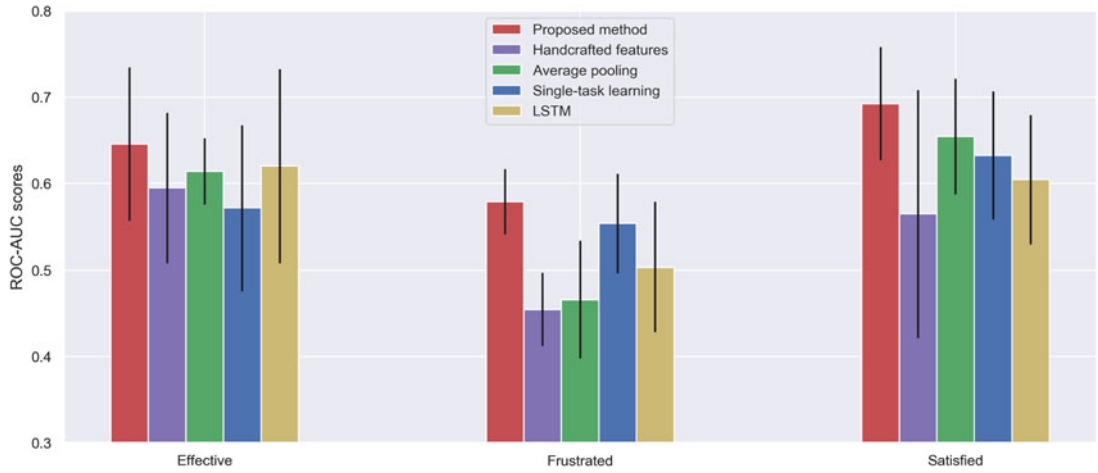

Fig. 6. Visualization of ROC-AUC scores for the proposed method and various competing approaches. Mean score of 5-folds are presented with ± standard deviation).

from a newcomer is gradually added to the training data should help in circumventing this issue. Because of the class imbalance, we used weighted cross entropy loss and selected Receiver Operating Characteristics-Area Under Curve (ROC-AUC) as the evaluation metric. The loss is weighted with respect to the inverse-frequency of the labels in the training set. To extend ROC-AUC to the multi-class case, a One-vs-One scheme is used where every unique pairwise combination of classes are compared and the resulting metrics are averaged.

For selecting the hyperparameters of the model, we used Bayesian optimization [50]. Hyperparameter values that performed best on the validation sets for the majority of the folds selected for training the final models and obtaining the performances on the test sets. Selected values for the hyperparameters are presented throughout the rest of this paragraph. Adam with a weight decay parameter of 0.03 is used as the optimizer [51]. For more regularization, dropout with a probability of 0.3 is also used for the TCNs and fully-connected layers. Best performing number of hidden units are found to be as 64 and 64 for the TCNs and the NetVLAD layers. The kernel size and the number of layers for TCNs were 10 and 8. With this setup, The receptive field of the TCNs were able to approximately cover even the longest sequences. For the NetVLAD layers, best performing value for the number of clusters K is found to be 3. We had one fully-connected layer for each branch with 64 hidden units and the output layers mapped these 64 dimensional vectors to the predicted label. The transformation of data through a forward pass, for one branch of the fully-connected layers, are as follows: $2 \times N \times S \times 5 \rightarrow 2 \times N \times 64 \rightarrow (2 \times 64) = 128 \rightarrow 64 \rightarrow 1$ where N is the number of participants in the interaction and S is the length of the sequence. Each model was trained for 25 epochs and the model performed best on the validation set is used to evaluate the performance on the test set.

5.2 Results

Fig. 6 shows the performances obtained with the proposed method and several competing approaches mentioned in the methodology section. Visualized values are the mean ROC-AUC scores for the 5 folds and the error bars correspond to ± standard deviation. Our proposed architecture managed to outperform all the other approaches and provided an average (of the performances for the three labels) ROC-AUC score of 0.64. Since we are using ROC-AUC scores as the evaluation metric, the random baseline, the performance obtained by a dummy classifier that will classify each sample into the majority class, is 0.5 for each task. In order to check the significance of our results, we applied paired one-tailed t-tests to the distributions formed by the average ROC-AUC scores of each fold, obtained by the proposed method and the other setups. The proposed method is shown to perform significantly better than all other methods ($p < 0.05$ for LSTM, average pooling and single-task learning, and $p < 0.01$ for handcrafted features). We also checked the significance for each label separately and results of this analysis are shared in the following subsections. The p-values for these separate analyses are corrected using Benjamini-Hochberg False Discovery Rate [52] since multiple tests are conducted for each comparison between our method and a competing one. In summary, our method was not significantly better than all the other setups for all labels but the first test conducted on the means proved that it is significantly better than others when the overall performance is considered. Detailed explanations of each competing method and their comparison to our approach are presented in the following subsections.

5.2.1 *Comparison to Feature Engineering*

As mentioned in the methodology section, we used the hand-crafted features presented in [4] for comparing our automatic representation extraction routine to feature engineering. These features are, per person, are as follows:

- Total number of IR pings, grouped under the IDs of participants
- Mean and standard deviations of movement energy and consistency
- Mean and standard deviations of the audio amplitude and standard deviation
- Mirroring and influence features computed per dyad in the interaction:

- Pearson correlations of movement energies and consistencies
- Pearson correlations of audio amplitudes and standard deviations
- Cross-correlations of movement energies and consistencies
- Cross-correlations of audio amplitudes and standard deviations

As proposed in [4], mirroring and influence features are computed using sliding windows of one minute, covering the entire interaction. These individual features are then aggregated to the interaction level by computing the minimum, maximum, mean, median, and standard deviations of the features. After the features per interaction are computed, we used the same branching MTL architecture we explained in Section 4.2.3, for a fair comparison to our approach.

As it can be seen from Fig. 6, hand-crafted features performed the worst and the proposed method significantly outperformed them for each label ($p < 0.05$ for effective and satisfied and $p < 0.01$ for frustrated). For frustrated labels, it even fails to perform better than a random baseline. On one hand, such results are expected since the analyses of [4] and this paper differ significantly in terms of the time resolution. In [4], the estimation of team cohesion and individual affect were done on a daily level, while our experiments included intervals as short as 1 minutes. Statistical features proposed in [4], such as the mean and the standard deviations of sensor readings might manage to capture a general trend happening throughout a day but might fail to be informative for shorter time intervals. However, we would have expected mirroring and influence features to be relatively representative of interaction quality, considering the existing literature on this topic. It could be possible that the proposed methodology for extracting these features in [4] is not optimal for representing aspects of a more fine-grained phenomenon such as interaction quality.

5.2.2 Comparison to Average Pooling

To analyze the effects of using a learnable pooling layer, we compared our results to an architecture where the NetVLAD layers are replaced by average pooling layers. All the other settings were kept exactly the same. NetVLAD's contribution to the performance can be easily seen. For all labels, the architecture with NetVLAD layers performed better on average than the one with average pooling layers (significant for the frustrated label with $p < 0.05$). Even though the performance obtained with average pooling for the effective and satisfied labels are comparable to remaining methods, it failed to provide a performance better than a random baseline for the frustrated labels. However, relatively lower performance scores for frustrated labels are observed for all approaches, suggesting that the estimation of this label is harder than the others. Frustration, which has a negative valence, might be harder to observe from manifested behavior compared to the other two labels that have a more positive valence.

5.2.3 Comparison to Single-Task Learning

In order to understand the contribution of using an MTL setup, we compared our method to single-task learning (STL). Instead of training one model that outputs three labels, we trained three models separately, each with one of the labels. All the other settings were kept exactly the same.

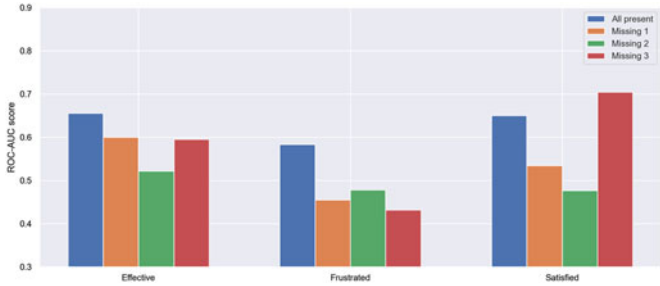

Fig. 7. ROC-AUC scores with respect to absence of data.

Even though the difference is quite marginal, STL still obtained the best overall performance after our proposed approach. Interestingly, it is the only other approach that managed to obtain a performance better than the random baseline for the frustrated labels and the proposed method was significantly better than STL for the effective ($p < 0.05$) and frustrated ($p < 0.1$) labels. This further shows the contributions of the TCN and NetVLAD layers. However, we can see that the differences between the proposed method and STL are quite marginal (0.03) for the frustrated labels, compared to other ones. This might show that the representations jointly learned with other labels might not be as discriminative for estimating frustration. The differences in label distributions and the opposite valence associated with frustration also support this claim.

5.2.4 Comparison to RNNs

As formerly mentioned, recurrent neural networks were the norm for modeling sequential data for quite a long time [53], [54]. Some of the most successful RNN variants are the Long Short-Term Memory (LSTM) networks. For a better analysis of utilizing TCNs for extracting meaningful representations from raw data, we replaced the TCNs in our architecture with one-layer LSTMs. The experiment setup was kept exactly the same other than these replacements.

The model with TCNs outperformed the one with LSTMs on average for all the three labels (significantly for frustrated and satisfied labels with $p < 0.1$). LSTMs are known to be notoriously hard to train and data-hungry [55] which might be an explanation for these results. Since our dataset was relatively on the smaller side, LSTMs might have failed to learn meaningful representation. However, this already shows an advantage of TCNs where less data is required for convergence. TCNs better performance compared to LSTMs was also compatible with the previous findings reported in [33]. On average, LSTMs still outperformed handcrafted features, further supporting the decision to automatically learn representations rather than explicitly designing them.

5.2.5 Missing Data Versus Performance

Fig. 7 shows how the performance varies with respect to missing data for all the three labels. To compute these statistics, we have kept track of predictions in each fold and if any participants data is missing for that interaction. Then, ROC-AUC scores were computed for four different subsets:

interactions with data from all participants present and interactions missing data from one, two and three participants. As it can be seen, for most cases the performance drops with the increasing missing data. Interactions with missing data from three participants (and also two participants for the satisfied labels) have samples from only two classes in the subsets (medium and high for the effective and satisfied and low and medium for the frustrated), rather than three. Moreover, the subset of missing data from three participants has only 13 samples in the test set. These characteristics make the performances for this subset to be more volatile and can explain the unexpected increase in the performance for effective (compared to missing data from two participants) and satisfied (compared to all). As expected, the estimation of quality of interactions is more challenging if the data for some participants are missing. Considering the overall performances presented in Fig. 6, we can argue that our proposed architecture is more robust to absence of data. More importantly, pooling approach makes it possible to analyze cases where some data is missing, rather than completely omitting that sample from the analysis.

While calculating these statistics, models trained on the whole training set with all cases of interactions with respect to missing data are used. Further analysis can be done by training models on subsets with specific missing data patterns only. However, since some cases only have limited number of samples, models trained on them failed to converge. More data with diverse missing data patterns are required for such further analysis.

5.3 Discussion

5.3.1 Ambulatory Assessment of Interactions in Realistic and Longitudinal Scenarios

As we highlighted at the beginning of this paper, unobtrusive sensors appear to be good potential candidates for real or near real-time time automated assessment of group interaction quality. Such a capability is particularly important for future long duration space missions, which are the ultimate ICE experience. We know that group social cohesion begins to destabilize and breakdown over durations as short as 5 to 7 months [56]. For long duration missions in the neighborhood of 3 years, active crew support will be needed. An ability to detect that destabilization early and then to trigger interventions to provide psycho-social support would be critical for maintaining crew effectiveness for these challenging space exploration missions of the future. Moreover, such a capability would also have utility for supporting ICE missions on Earth such as polar science and deep-sea exploration.

Our results in this paper represent a proof of concept of fine-grained analysis of the quality of interactions happening in a real-life ICE scenario. Even though there exist prior works that focused on the detection of interactions in-the-wild with relatively uncontrolled data collection procedures [57], [58], no fine-grained analysis related to the quality of detected interactions were formerly presented. Interactions inherently have rich and valuable information about many different concepts relating to social behavior. With this study, we showed that a more in-depth analysis of interactions is possible, even with sensor limitations, imperfect data, and label information collected in uncontrolled and realistic settings. We also explicitly presented various challenges such settings bring and discussed various solutions for circumventing them. We believe that these challenges, such as the missing data and the requirement for the precise identification of short-term events in longitudinal data, are common problems experienced by all researchers working in this domain. This can be also seen from the recent studies which explicitly focused on tackling challenges introduced by missing data in mobile crowdsensing [42]. With the development of approaches that can handle these challenges better, even a finer-grained analysis will be possible, providing a much deeper understanding of the concepts that are being analyzed. We see finer-grained analysis of longitudinal data is to be the frontier for this domain and hope this paper can inspire researchers to move into this direction more.

5.3.2 Performance of Interaction Interval Detection

One of the biggest limitations of our experiments is the process of finding the interaction intervals. As mentioned, there were no ground truth for the actual starting and ending times for the interactions logged in the ESM data. This forced us to employ a heuristic approach with aggressive parameters, which is explained in the Section 4.1. Since there was no ground truth, it was not possible to evaluate the performance of this step and how it affects the performance of the interaction quality estimation. However, it is heartening to see that there were learnable patterns of behavior for the effective and satisfaction task which suggest that fairly reasonable intervals were chosen. We acknowledge that there are many settings where one cannot properly evaluate the intermediate measures being used. This also echoes the age old problem of objectively evaluating the quality of synchrony measures when trying to estimate aspects of conversation quality. Our results suggest here that the hand crafted features were perhaps only a limited set of possible coordination features that could be better learned by the TCNs.

If there was ground truth, even for a subset of the dataset, a more robust learning-based approach can be employed for the detection of the interactions. Recent studies have shown that dynamics of interactions, which accelerometer readings can act as a proxy for, have valuable information for detecting conversing partners, in addition to the proxemics (which IR and Bluetooth readings acts as a proxy for in the scope of this work) [59], [60]. Having ground truth would allow to utilize all the important information recorded by the wearable sensors.

Of course, obtaining exact starting and ending points of interaction intervals is a quite challenging task. Asking participants to report this information would most probably not result in perfect ground truth considering the missing data problem discussed throughout this paper. More importantly, such a procedure most probably will interfere with the ecological validity of the dataset and cause more burden on the participants since they will need to report many times during the day. However, the procedure for collecting information about interactions can be still enhanced. A possible option is to utilize techniques from the latest interruptibility literature together with the rough detections from

sensors, asking participants to confirm existence of an interaction when they are found to be available for reporting. Another option is to utilize different sensors such as cameras to obtain more information about interactions until sufficient samples are collected for training a supervised interaction detector.

5.3.3 Learnable Pooling Layers

In this study, we proposed to use a learnable pooling layer, NetVLAD specifically, to aggregate the individual representations of participants into a fixed-size representation of an interaction. The focus was on empirically showing the value of the idea of using learnable pooling for aggregation. The choice of the specific learnable pooling layer as NetVLAD was based on former empirical results where it was shown to outperform other learnable pooling methods [47]. Such other methods include slight modifications to the NetVLAD (NetRVLAD) and implementations of other well-known descriptors such as the bag-of-words (Soft-DBoW) and the Fisher Vectors (NetFV) [47]. Conceptually, NetVLAD layers in our proposed architecture can be replaced by any of these methods. However, more experimentation with more data is needed for a conclusive statement on which implementation should be preferred.

6 Conclusion and Future Work

6.1 Conclusion

In this paper, we presented our approach for estimating the quality of interactions between group members in a longitudinal simulated space mission from wearable sensing data. During the four month mission, participants wore Sociometric Badges which recorded movement (accelerometer), sound (microphone) and proximity (Infrared and Bluetooth) information. They also filled in daily reports about the interactions they had with other group members. These reports included the following information: who the interaction was with, and whether the level of effectiveness, frustration, and satisfaction with the interaction. The task of this paper was to automatically estimate these three labels (effective, frustrated, and satisfied) from the wearable sensing data.

We employed an heuristic approach to detect the interaction intervals; the timestamps for the starting and ending times of the interaction; from the IR and Bluetooth data. We proposed a neural network architecture composed by Temporal Convolutional Networks (TCNs) and learnable pooling layers. TCNs were used to process sequential data for automatically extracting discriminative representations from the individual accelerometer and audio streams. Learnable pooling layers, NetVLAD, were then used to aggregate these individual representations into a fixed-size representation of the interaction. This pooling step provides the flexibility of working with interactions of varied sizes. Also, it allows to tackle the well-known problem of missing data in field studies, by allowing the utilization of remaining data from interactions where some participants data are missing. Finally, we proposed to estimate the three labels jointly with a multi-task learning setup. Representations extracted by TCNs and the parameters of NetVLAD layers were learned jointly, allowing to exploit the similarities between the tasks.

Our proposed approach significantly outperformed various competitors and provided an average ROC-AUC score of 0.64. To better understand each components contribution to the performance, we presented ablation studies where parts of the network are replaced with a competitive approach, such as replacing NetVLAD layers with average pooling layers and replacing TCNs with LSTMs. We also analyzed how missing data affects the performance and shown that there is a negative correlation between the number of participants with missing data and performance.

6.2 Future Work

One of the biggest limitations of our experiments is the interaction interval detection. Currently, it is based on heuristics and was not evaluated due to the lack of ground truth. With ground truth, a learning based detector can be trained on multiple modalities, allowing more precise detections of when interactions occur. Having more precise starting and ending times for interactions is expected to result in superior quality estimation. Learning and evaluating on a low frequency model based on the prior work [59], [60], extended to both audio and accelerometer data could be a promising avenue to partially evaluate the efficacy of such an approach.

Another interesting direction to investigate is evaluating the performance of learnable pooling methods other than NetVLAD. Even though NetVLAD was empirically shown to be superior to others in former work [47], these evaluations were made for a significantly different task. Further investigation in this direction might even result in a new pooling approach, specifically designed for the task of pooling individual representations into a fixed-length one representing the interaction itself.

We believe that more data and further experimentation is needed to fully analyze the effects of missing data on performance. Unfortunately, for this dataset, the absence of data was such that more elaborate analysis was not possible. With more data with varying absence patterns, it would be possible to train estimation models on specific subsets with specific missing data statistics, allowing further analysis.

In this study, we have focused on estimating the quality of interactions. To keep the scope of this paper focused, we have not included any experiments that present the use of interaction quality information. Being able to automatically quantify the quality of interactions a person having throughout their life creates many new possibilities. As mentioned earlier, many existing works that aim to estimate concepts such as affect, stress, productivity, team and social cohesion used some simplistic notion of interactions in their methodology, but to our knowledge, none had access to information regarding the quality of interactions. Considering the diverse literature in social sciences that show the connection between interaction quality and various social concepts [8], [61], [62], we are excited about future work that will build on the findings of this paper.

Acknowledgments

The author S.W.J. Kozlowski is a Principal Investigator. Any opinions, findings, conclusions and recommendations expressed are those of the authors and do not necessarily reflect the views of NASA.

REFERENCES

[1] E. Salas, S. I. Tannenbaum, S. W. Kozlowski, C. A. Miller, J. E. Mathieu, and W. B. Vessey, "Teams in space exploration: A new frontier for the science of team effectiveness," *Curr. Directions Psychol. Sci.*, vol. 24, no. 3, pp. 200–207, 2015.

[2] S. W. Kozlowski and G. T. Chao, "Unpacking team process dynamics and emergent phenomena: Challenges, conceptual advances, and innovative methods," *Amer. Psychol.*, vol. 73, no. 4, 2018, Art. no. 576.

[3] R. Wang *et al.*, "Studentlife: Assessing mental health, academic performance and behavioral trends of college students using smartphones," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, 2014, pp. 3–14.

[4] Y. Zhang, J. Olenick, C.-H. Chang, S. W. J. Kozlowski, and H. Hung, "TeamSense: Assessing personal affect and group cohesion in small teams through dyadic interaction and behavior analysis with wearable sensors," *Proc. ACM Interact. Mobile Wearable Ubiquitous Technol.*, vol. 2, no. 3, pp. 1–12, Sep. 2018.

[5] A. Bogomolov, B. Lepri, M. Ferron, F. Pianesi, and A. Pentland, "Daily stress recognition from mobile phone data, weather conditions and individual traits," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 477–486.

[6] D. Olguin Olguin, "Sensor-based organizational design and engineering," Ph.D. dissertation, Program in Media Arts and Sciences, Massachusetts Inst. of Technol., Cambridge, MA, USA, 2011.

[7] O. Lederman, A. Mohan, D. Calacci, and A. S. Pentland, "Rhythm: A unified measurement platform for human organizations," *IEEE MultiMedia*, vol. 25, no. 1, pp. 26–38, Jan.–Mar. 2018.

[8] D. Fiorillo and F. Sabatini, "Quality and quantity: The role of social interactions in self-reported individual health," *Soc. Sci. Med.*, vol. 73, no. 11, pp. 1644–1652, 2011.

[9] N. D. Lane *et al.*, "BeWell: Sensing sleep, physical activities and social interactions to promote wellbeing," *Mobile Netw. Appl.*, vol. 19, no. 3, pp. 345–359, 2014.

[10] J. G. Ibrahim and G. Molenberghs, "Missing data methods in longitudinal studies: A review," *Test*, vol. 18, no. 1, pp. 1–43, 2009.

[11] A. Mehrotra and M. Musolesi, "Intelligent notification systems: A survey of the state of the art and research challenges," 2017, *arXiv: 1711.10171*.

[12] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN architecture for weakly supervised place recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 5297–5307.

[13] L. Sun, K. Jia, D.-Y. Yeung, and B. E. Shi, "Human action recognition using factorized spatio-temporal convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 4597–4605.

[14] R. Caruana, "Multitask learning," *Mach. Learn.*, vol. 28, no. 1, pp. 41–75, 1997.

[15] T. Choudhury and A. Pentland, "Sensing and modeling human networks using the sociometer," in *7th IEEE Int. Symp. Wearable Comput.*, 2003, pp. 216–222.

[16] D. O. Olguín, B. N. Waber, T. Kim, A. Mohan, K. Ara, and A. Pentland, "Sensible organizations: Technology and methodology for automatically measuring organizational behavior," *IEEE Trans. Syst., Man, Cybern., Part B. (Cybern.)*, vol. 39, no. 1, pp. 43–55, Feb. 2009.

[17] B. Lepri *et al.*, "The sociometric badges corpus: A multilevel behavioral dataset for social behavior in complex organizations," in *Proc. Int. Conf. Privacy, Securi., Risk Trust Int. Conf. Soc. Comput.*, 2012, pp. 623–628.

[18] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 42–55, Jan.–Mar. 2012.

[19] H. Dibeklioğlu, Z. Hammal, and J. F. Cohn, "Dynamic multimodal measurement of depression severity using deep autoencoding," *IEEE J. Biomed. Health Inform.*, vol. 22, no. 2, pp. 525–536, Mar. 2018.

[20] R. LiKamWa, Y. Liu, N. D. Lane, and L. Zhong, "MoodScope: Building a mood sensor from smartphone usage patterns," in *Proc. 11th Annu. Int. Conf. Mobile Syst., Appl., Serv.*, 2013, pp. 389–402.

[21] S. T. Moturu, I. Khayal, N. Aharony, W. Pan, and A. Pentland, "Using social sensing to understand the links between sleep, mood, and sociability," in *Proc. IEEE 3rd Int. Conf. Privacy, Secur., Risk Trust IEEE 3rd Int. Conf. Soc. Comput.*, 2011, pp. 208–214.

[22] S. Servia-Rodríguez, K. K. Rachuri, C. Mascolo, P. J. Rentfrow, N. Lathia, and G. M. Sandstrom, "Mobile sensing at the service of mental well-being: A large-scale longitudinal study," in *Proc. 26th Int. Conf. World Wide Web*, 2017, pp. 103–112.

[23] N. Jaques, S. Taylor, E. Nosakhare, A. Sano, and R. Picard, "Multi-task learning for predicting health, stress, and happiness," in *Proc. NIPS Workshop Mach. Learn. Healthcare*, 2016.

[24] Y. Zhang, J. Olenick, C.-H. Chang, S. W. Kozlowski, and H. Hung, "The I in team: Mining personal social interaction routine with topic models from long-term team data," in *Proc. 23rd Int. Conf. Intell. User Interfaces*, 2018, pp. 421–426.

[25] D. Gatica-Perez, "Automatic nonverbal analysis of social interaction in small groups: A review," *Image Vis. Comput.*, vol. 27, no. 12, pp. 1775–1787, 2009.

[26] J. Wang, Y. Chen, S. Hao, X. Peng, and L. Hu, "Deep learning for sensor-based activity recognition: A survey," *Pattern Recognit. Lett.*, vol. 119, pp. 3–11, 2019.

[27] H. F. Nweke, Y. W. Teh, M. A. Al-Garadi, and U. R. Alo, "Deep learning algorithms for human activity recognition using mobile and wearable sensor networks: State of the art and research challenges," *Expert Syst. Appl.*, vol. 105, pp. 233–261, 2018.

[28] N. D. Lane and P. Georgiev, "Can deep learning revolutionize mobile sensing?," in *Proc. 16th Int. Workshop Mobile Comput. Syst. Appl.*, 2015, pp. 117–122.

[29] P. Georgiev, S. Bhattacharya, N. D. Lane, and C. Mascolo, "Low-resource multi-task audio sensing for mobile and embedded devices via shared deep neural network representations," *Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol.*, vol. 1, no. 3, pp. 1–19, 2017.

[30] A. V. d. Oord *et al.*, "Wavenet: A generative model for raw audio," 2016, *arXiv:1609.03499*.

[31] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[32] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," 2014, *arXiv:1412.3555*.

[33] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," 2018, *arXiv: 1803.01271*.

[34] N. Y. Hammerla, S. Halloran, and T. Plötz, "Deep, convolutional, and recurrent models for human activity recognition using wearables," 2016, *arXiv:1604.08880*.

[35] N. Nair, C. Thomas, and D. B. Jayagopi, "Human activity recognition using temporal convolutional network," in *Proc. 5th Int. Workshop Sensor-Based Activity Recognit. Interact.*, 2018, pp. 1–8.

[36] J. W. Graham, "Missing data analysis: Making it work in the real world," *Annu. Rev. Psychol.*, vol. 60, pp. 549–576, 2009.

[37] J. L. Schafer and J. W. Graham, "Missing data: Our view of the state of the art," *Psychol. Methods*, vol. 7, no. 2, 2002, Art. no. 147.

[38] I. Azimi, T. Pahikkala, A. M. Rahmani, H. Niela-Vilén, A. Axelin, and P. Liljeberg, "Missing data resilient decision-making for healthcare iot through personalization: A case study on maternal health," *Future Gener. Comput. Syst.*, vol. 96, pp. 297–308, 2019.

[39] T. Feng and S. Narayanan, "Imputing missing data in large-scale multivariate biomedical wearable recordings using bidirectional recurrent neural networks with temporal activation regularization," in *Proc. 41st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2019, pp. 2529–2534.

[40] J. L. Hicks *et al.*, "Best practices for analyzing large-scale health data from wearables and smartphone apps," *NPJ Digit. Med.*, vol. 2, no. 1, pp. 1–12, 2019.

[41] J. Ae Lee and J. Gill, "Missing value imputation for physical activity data measured by accelerometer," *Statist. Methods Med. Res.*, vol. 27, no. 2, pp. 490–506, 2018.

[42] F. Montori, P. P. Jayaraman, A. Yavari, A. Hassani, and D. Georgakopoulos, "The curse of sensing: Survey of techniques and challenges to cope with sparse and dense data in mobile crowd sensing for Internet of Things," *Pervasive Mobile Comput.*, vol. 49, pp. 111–125, 2018.

[43] W. Liu, L. Wang, E. Wang, Y. Yang, D. Zeghlache, and D. Zhang, "Reinforcement learning-based cell selection in sparse mobile crowdsensing," *Comput. Netw.*, vol. 161, pp. 102–114, 2019.

[44] B. J. Caldwell, P. G. Roma, and K. Binsted, "Team cohesion, performance, and biopsychosocial adaptation research at the hawai'i space exploration analog and simulation (HI-SEAS)," in *Proc. 31st Annu. Conf. Soc. Ind. Organizational Psychol.*, 2016.

[45] B. Dudzik *et al.*, "Discovering digital representations for remembered episodes from lifelog data," in *Proc. Workshop Model. Cogn. Processes Multimodal Data*, 2018, pp. 1–9.

[46] D. Chaffin *et al.*, "The promise and perils of wearable sensors in organizational research," *Organizational Res. Methods*, vol. 20, no. 1, pp. 3–31, 2017.

[47] A. Miech, I. Laptev, and J. Sivic, "Learnable pooling with context gating for video classification," 2017, *arXiv: 1706.06905*.
[48] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 3304–3311.
[49] K. Sechidis, G. Tsoumakas, and I. Vlahavas, "On the stratification of multi-label data," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discov. Databases*, 2011, pp. 145–158.
[50] M. Balandat *et al.*, "BoTorch: Programmable Bayesian optimization in pytorch," 2019, *arXiv: 1910.06403*.
[51] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2017, *arXiv: 1711.05101*.
[52] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *J. Roy. Statist. Soc.: Ser. B. (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995.
[53] Y. Guan and T. Plötz, "Ensembles of deep lstm learners for activity recognition using wearables," *Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol.*, vol. 1, no. 2, pp. 1–28, 2017.
[54] A. Graves, A.-R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2013, pp. 6645–6649.
[55] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 1310–1318.
[56] S. Kozlowski, "Capturing the dynamics of team processes," in *AAAS Annu. Meeting*, 2019.
[57] G. Vanderhulst, A. Mashhadi, M. Dashti, and F. Kawsar, "Detecting human encounters from WiFi radio signals," in *Proc. 14th Int. Conf. Mobile Ubiquitous Multimedia*, 2015, pp. 97–108.
[58] A. Montanari, S. Nawaz, C. Mascolo, and K. Sailer, "A study of bluetooth low energy performance for human proximity detection in the workplace," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun.*, 2017, pp. 90–99.
[59] E. Gedik and H. Hung, "Detecting conversing groups using social dynamics from wearable acceleration: Group size awareness," *Pro. ACM Interactive, Mobile, Wearable Ubiquitous Technol.*, vol. 2, no. 4, pp. 1–24, 2018.
[60] A. Rosatelli, E. Gedik, and H. Hung, "Detecting f-formations & roles in crowded social scenes with wearables: Combining proxemics & dynamics using LSTMs," in *Proc. 8th Int. Conf. Affect. Comput. Intell. Interact. Workshops Demos*, 2019, pp. 147–153.
[61] T. L. Schuster, R. C. Kessler, and R. H. Aseltine, "Supportive interactions, negative interactions, and depressed mood," *Amer. J. Community Psychol.*, vol. 18, no. 3, pp. 423–438, 1990.
[62] E. D. Heaphy and J. E. Dutton, "Positive social interactions and the human body at work: Linking organizations and physiology," *Acad. Manage. Rev.*, vol. 33, no. 1, pp. 137–162, 2008.

Ekin Gedik received the bachelor's and master's degrees from the Middle East Technical University, Turkey, in 2010 and 2013, respectively, and the PhD degree from the Delft University of Technology, in 2018. He is currently a guest researcher with Socially Perceptive Computing Lab of Delft University of Technology. His research interests are social behavior analysis, wearable sensing, affective computing, computer vision and applied pattern recognition.

Jeffrey Olenick received the bachelor's degree in psychology and history, the master's degree in social sciences from University of Chicago, and the PhD degree in organizational psychology from Michigan State University, in May 2020. He is an assistant professor of industrial/organizational psychology with Old Dominion University. He studies the dynamics of individual and team learning and development, including training and transfer, self-regulation, and team dynamics in extreme environments, as well as the methods used to study such topics.

Chu-Hsiang Chang received the MA and PhD degrees from the University of Akron, in 2002 and 2005, respectively. She is an associate professor of psychology with Michigan State University.

Steve W.J. Kozlowski received the BA degree in psychology from the University of Rhode Island, in 1976, and the MS and PhD degrees in I/O psychology from The Pennsylvania State University, in 1979 and 1982, respectively. He is a world class scholar and professor with the University of South Florida (previously Michigan State University). He is a recognized authority in the areas of multilevel organizational systems theory; team leadership and team effectiveness; and learning, training, and adaptation. The goal of his programmatic research is to generate actionable theory, research-based principles, and deployable tools to develop adaptive individuals, teams, and organizations.

Hayley Hung received the first degree in electrical and electronic engineering from Imperial College, and the PhD degree in computer vision from Queen Mary University of London, in 2007. She is an associate professor with Delft University of Technology and leads the Socially Perceptive Computing Lab. Between 2010 and 2013, she held a Marie Curie Intra-European Fellowship with the Intelligent Systems Lab, the University of Amsterdam. Between 2007 and 2010, she was a post-doctoral researcher with Idiap Research Institute. Her research interests are in social computing, social signal processing, computer vision, and machine learning.