Modeling Dynamics of Task and Social Cohesion from the Group Perspective Using Nonverbal Motion Capture-based Features

Fabian Walocha
fabian.walocha@telecom-paris.fr
LTCI, Télécom Paris, Institut Polytechnique de Paris
91120 Palaiseau, France

Lucien Maman
lucien.maman@telecom-paris.fr
LTCI, Télécom Paris, Institut Polytechnique de Paris
91120 Palaiseau, France

Mohamed Chetouani
mohamed.chetouani@sorbonne-universite.fr
Institute for Intelligent Systems and Robotics, Sorbonne University, CNRS UMR7222
75252 Paris, France

Giovanna Varni
giovanna.varni@telecom-paris.fr
LTCI, Télécom Paris, Institut Polytechnique de Paris
91120 Palaiseau, France

ABSTRACT

Group cohesion is a multidimensional emergent state that manifests during group interaction. It has been extensively studied in several disciplines such as Social Sciences and Computer Science and it has been investigated through both verbal and nonverbal communication. This work investigates the dynamics of task and social dimensions of cohesion through nonverbal motion-capture-based features. We modeled dynamics either as decreasing or as stable/increasing regarding the previous measurement of cohesion. We design and develop a set of features related to space and body movement from motion capture data as it offers reliable and accurate measurements of body motions. Then, we use a random forest model to binary classify (decrease or no decrease) the dynamics of cohesion, for the task and social dimensions. Our model adopts labels from self-assessments of group cohesion, providing a different perspective of study with respect to the previous work relying on third-party labelling. The analysis reveals that, in a multilabel setting, our model is able to predict changes in task and social cohesion with an average accuracy of 64%(±3%) and 67%(±3%), respectively, outperforming random guessing (50%). In a multiclass setting comprised of four classes (i.e., *decrease/decrease*, *decrease/no decrease*, *no decrease/decrease* and *no decrease/no decrease*), our model also outperforms chance level (25%) for each class (i.e., 54%, 44%, 33%, 50%, respectively). Furthermore, this work provides a method based on notions from cooperative game theory (i.e., SHAP values) to assess features' impact and importance. We identify that the most important features for predicting cohesion dynamics relate to spacial distance, the amount of movement while walking, the overall posture expansion as well as the amount of inter-personal facing in the group.

CCS CONCEPTS

• **Human-centered computing → Collaborative and social computing design and evaluation methods**; • **Computing methodologies** → *Machine learning*.

KEYWORDS

Cohesion; group interaction analysis; machine learning; nonverbal communication; social signal processing

1 INTRODUCTION

Collaboration may be the main cause of our success as a species [45]. This behavior results from the fact that humans naturally feel the need to belong to a group. Groups highly impact individual behavior and they develop their own dynamics. It is at least since the times of the Greek philosopher Aristotle (c. 385 - c. 323 BC) that human behavior and group phenomena are being investigated: "*Man is by nature a social animal*" [3]. Understanding how one specific group phenomenon emerges and evolves over time is a complicated task due to the inter-dependency with other group phenomena and the high variability of the group and context settings. With the advent of new technologies coupled to the emergence of Social Signal Processing (SSP) [36, 47], a lot of efforts have been put into the automatic detection and prediction of these groups phenomena [1]. SSP is a multidisciplinary research domain aimed at automatically detecting and analyzing human social signals and behavior. Automated group interaction analysis is one of the challenging tasks addressed by the SSP community and emphasis is given to the study of emergent states as they play an important role in group dynamics. These are social processes that result from the interactions among group members (e.g., [24]). Cohesion is one of the most studied emergent states [38] due to its influence on desirable group outcomes such as group effectiveness and performance (see [40] for a review). Automatically measuring cohesion is still at its infancy and could potentially help to develop new applications (e.g., software providing feedback on group processes such as meetings). Multiple definitions of cohesion exist, limiting the generalization and the reproducibility over the literature. Recently, Severt and Estrada proposed an integrative framework of cohesion [41]. This framework gathers ideas from the Carron model [9] as well as various influential ideas integrating the multidimensionality of cohesion (e.g., [5, 6, 12, 15]). It acknowledges that cohesion is a dynamic phenomenon that can be expressed in various dimensions and hierarchical levels depending on the context and the relationships among group members. The design of the nonverbal features automatically extracted in this study was informed by its definition of cohesion [41]. This framework refers to social and task dimensions as being part of the instrumental property of cohesion. The former refers to "*social bonds between group members that are bound by the group's working relationship*" while the latter concerns the "*group* *members' shared commitment to the group's tasks*" [41].

The main contribution of this study is to set up a baseline to predict the dynamics of the task and social dimensions of cohesion through nonverbal motion-capture-based features. To the best of our knowledge, this is the first study addressing cohesion from this angle. Previous studies, indeed, focused on predicting the level of cohesion [20, 33]. Here, we focused on the variations of cohesion (i.e., its dynamics) that we modeled as decrease or stable/increase with respect to the previous measurement of cohesion. Furthermore, we chose to study cohesion from a different perspective, using as labels the group members' self-assessments of cohesion instead of relying on third-party labelling. This study also supplies a set of nonverbal group motion capture-based features that is useful to describe the functional property of cohesion at a horizontal level. Finally, this research provides a novel method based on cooperative game theory to assess the impact and the importance of a feature set on our model. This approach helps to understand how are the task and social dimensions related to each other by observing similarities and differences in the way nonverbal behavior manifest and impact each dimension. In order to achieve all of the previously mentioned goals, we first extracted nonverbal features from the MoCap data available in the GAME-ON dataset [29]. We focused our effort on conceiving and extracting features with this technology as body movement and gesture (kinesics) and group members use of space (proxemics) play an important role in nonverbal communication [18] and MoCap data provides reliable and accurate measurements of body movements as opposed to existing video-based features. Then, we ran a supervised machine learning algorithm using as labels the group members' self-assessments to predict the dynamics of both task and social dimensions. Since the GAME-ON dataset also provides six self-assessments of group cohesion for each group member collected all along the data collection, we approximated the dynamics (increase or stable/decrease) of cohesion by taking the mean rank difference between two consecutive measurements. Finally, we assessed the impact of our features on the model prediction and their overall importance. This analysis highlighted what are the common important features to predict the dynamics of both task and social dimensions in our setting. These findings are in line with theoretical models of cohesion (e.g., [41]) that suggest a relationship between social and task dimensions.

2 RELATED WORK

2.1 Automated approaches to detect cohesion

Over the last decade, an interest for automatically detecting cohesion has emerged. As nonverbal communication has been shown to be a more powerful predictor of group-level cohesion than verbal behavior [25], most of the studies focused on small groups' nonverbal cues. Hung and Gatica-Perez were the first to address this problem, employing both audio and video nonverbal descriptors to study cohesion through multiple dimensions in a meeting context [20]. In their study, they collected external annotations on the established AMI corpus [31]. Thereby, Hung *et al.* provided a first attempt at a general design framework for the automated assessment of group cohesion based on nonverbal behavioral features. They showed that the best performing features were the total pause time between each individual's turns during a meeting segment with audio cues, the total visual activity for each person in the meeting with video cues and the visual activity during periods of overlapped speech with audio-visual cues. They reached more than 80% classification accuracy at estimating high and low levels of cohesion, rated by external observers, using binary classifiers such as SVMs. The task and social dimensions of cohesion were, however, estimated together, making it difficult to assess the impact of each feature on the classification. More recently, Nanninga *et al.* extended this work by integrating pairwise and group descriptors related to the alignment of para-linguistic speech behavior (e.g., convergence and similarity of voice intensity, speech rate) [33]. In a similar meeting context, they showed how combining mimicry and turn-taking based features improved classification when predicting high and low task and social cohesion. They used a Gaussian Mixture Model and Kernel Density Estimation to estimate task and social dimensions separately, achieving a performance of 64% Area under the ROC Curve (AUC) and 71% AUC for task and social dimensions, respectively. Group-level cohesion was estimated by comparing pairwise dyadic mimicry features. It remains an open question whether a group level mimicry feature would further improve performances. For the occasion of the EmotiW 2019 challenge [10], some studies [13, 16, 50] classified high and low levels of cohesion on images from a corpus of images created via web crawling of various keywords related to social events [11]. They showed how facial expressions were impacting external annotations and achieved promising results at predicting cohesion from images.

Other studies explored cohesion at a longitudinal level. They used sociometric badges to collect task and social-relevant features over a long period. These are unobtrusive equipment that can be placed on a person or on its phone and that are able to track the person's movement and activity. A first study using such kind of sensors was conducted by Olguin and Pentland to investigate face-to-face interactions of workers for a period of 20 working days [34]. All the features extracted were based on nonverbal behavior, proximity as well as other sources such as emails and performance data but only concerned individuals. Similarly, Zhang *et al.*, used sociometric badges to explore small group collaborations during long duration missions in confined spaces through a four-month simulation space exploration mission [51]. They defined individual as well as group features to classify cohesion as positive or negative, taking into account both task and social dimensions. They reached promising results (80% AUC) and showed that dyadic interactions and face-to-face communication are important for assessing cohesion. It, however, remains to be seen whether these results are specific to the studied scenario or whether they apply to different groups and contexts. Furthermore, only [33] and [51] integrated task and social dimensions in their models but did not explore their relationships.

At present, automated studies of cohesion only rely on external annotators that evaluate cohesion by answering a questionnaire or using a coding scheme (e.g., ACT4Team [22]) after analyzing a specific group interaction. The group's perspective provided through self-assessments could help to gain additional insights.

2.2 Nonverbal features from the visual channel related to cohesion

In multimodal studies on cohesion analysis, visual features are often found to perform poorly in comparison to other channels [20, 21]. One possible reason for this is that extracting nonverbal behavior from the visual channel might be challenging. Visual data in images or videos contains a lot of information and often requires further processing to extract relevant information on group activity. In a recent study, Kantharaju *et al.*, provided a multimodal analysis of group cohesion using the AMI corpus [4]. They explored cohesion through nonverbal social cues, dialogue acts and interruptions, using features from various modalities (e.g., audio and video). Among the nonverbal social features extracted from the visual channel, they showed that gaze, facial expressions, gestures and body postures were significantly impacting the perceived level of cohesion (e.g., instances of laughter are more frequent in high cohesive groups). In this study, they, however, limited their nonverbal social cues analysis to only four features (e.g., gaze, facial action units, head node, laughter) and assessed their impact with independent t-test on only 16 two-minute segments, which might be insufficient to generalize to different groups and contexts. Furthermore, observing correlations between the features and the cohesion scores is interesting to develop an intuition about how the features might be related to cohesion but it does not prove that they are useful to predict cohesion. These results, however, encourage the development of other nonverbal visual features related to cohesion as they seem to be important for predicting cohesion. Visual features can be very salient in portraying inter-personal relationships, as one can infer a lot of information based on the way people present themselves physically, either by positioning or by movement in space. The *amount of body movement* is partially accessible using low-level, pixel-based information by observing general visual motion found in a video, given a fixed position, fixed lighting camera setup. Overall movement can hereby be approximated for example via the amount of compression in a video and via optical flow [20, 21, 51]. From here, any information related to specific body parts, such as *facial expressions*, *gestures* and *body posture* require an additional model which (1) detects the specific body parts and (2) quantifies the associated motion. The most used strategy to derive information on the facial features of a person is by identifying facial action units (FAUs) when they are active and their respective intensity [4, 32]. Gestures are often expressed by quantifying the total movement of the hands over time. Hand positions are most often detected either by using additional sensors [32, 51], or by first deriving skeleton-data [16, 50], using software solutions such as *OpenPose* [7]. Body language cues can contain additional information on the intensity and synchronicity of the conversation in a group. One technology which is starting to receive growing interest in the context of group interaction research are motion capture (MoCap) systems. The main advantage of MoCap over video data is that 3D data on key body joints of individuals is readily available eliminating the need for an additional model to detect these body parts. Furthermore, as opposed to 3D depth cameras, the final data is not restricted to a single viewpoint, thus positions in 3D space and distances can be calculated with much higher precision. Additionally, due to the high precision of MoCap, body movement and position and distances can be assessed with much higher accuracy. Moreover, MoCap systems are designed to not limit the available range of motions. This makes MoCap an ideal instrument in order to test the impact of proxemics and body language cues on group interactions.

3 THE GAME-ON DATASET

All of the previous studies focusing on automatically detecting cohesion used datasets that are not specifically designed for its study. These datasets are composed of various groups' interactions in specific contexts and often contain scripted scenarios (e.g., the AMI corpus [31]). Furthermore, the measurements of the level of cohesion only relied on external annotators. In order to explore and advance our understanding of the expression of task and social dimensions of cohesion, we chose to use the GAME-ON dataset. Since the GAME-ON dataset [29] was specifically designed to monitor and elicit changes in cohesion over time, it is suited for the study of cohesion dynamics. Additionally, GAME-ON reports the change of cohesion through group self-assessments all along with the game. These were collected before the data collection and after each task using the GEQ questionnaire [9] with a 9-point Likert-scale. This questionnaire is aimed at measuring social and task dimensions of cohesion from a group member's point of view. GAME-ON consists of more than 11 hours of multimodal data (i.e., audio, video and MoCap data) where a total of 17 groups interact during an escape game. The dataset is composed of five different tasks. For all tasks (except the second), game instructions were given but participants were free to move and interact as they liked. As all of the participants considered themselves as friends and did not have any hierarchical status among them, it provided us with a relevant framework to study these dimensions at a horizontal level.

4 METHODOLOGY

The aim of this analysis is to detect *decreases* in cohesion on a given fixed time segment. For this aim, we present the following modeling pipeline (see Figure 1), used to extract MoCap-based features from the GAME-ON dataset, to train a statistical classifier, and finally, to extract information from the trained model on the most informative features for the task. The modeling pipeline can be separated into three major functional components: the *feature extraction* of both established as well as novel nonverbal features, the *temporal aggregation* by separating each task into fixed-length temporal slices with early feature selection, and finally, the *model training and prediction.* Additionally, we present our labeling strategy to quantify task-to-task changes in group cohesion by calculating the mean rank difference among self-assessments between subsequent tasks. Finally, we present our evaluation strategy by reporting our selected metrics associated with performance and feature importance. As we predict the cohesion dynamics for each slice and each group separately we jointly train and predict the cohesion state overall interactions.

4.1 Feature Extraction

One aim of this work is to explore mechanisms and behavioral correlates of group cohesion using motion capture. The MoCap data available yield translations and rotations of 17 key body joints in 3D space. Thus, the kind of information we can extract from MoCap

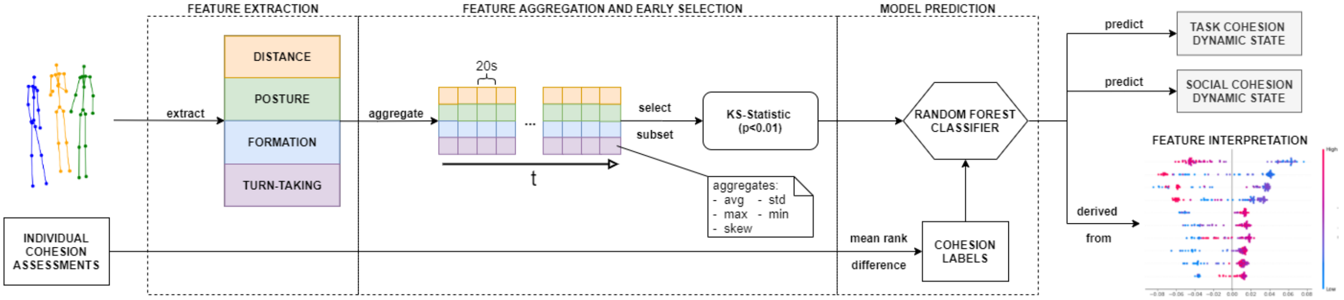

Figure 1: The modeling pipeline. Components: Feature extraction and engineering; Temporal integration and feature aggregation using slices of 20s with an early feature selection using Kolmogorov-Smirnov statistics; Modeling and prediction of the dynamic cohesion state given by the current and previous self-assessments.

data is related to the body language of the participant, their movement in space as well as the distance and orientation towards each other. Additionally, we introduce a new set of features by proposing a framework of *positional turn-taking* (inspired by equal notions of conversational turn-taking [39]), by approximating speaker behavior from facing behavior in the group. Table 1 shows the set of group features used in this work. We differentiate between two types of group features. Aggregated group features refer to features which first are calculated per individual and then aggregated over the group using a set of group descriptors (e.g. average, maximum, variance). Innate group features are features which are directly calculated over the group as a whole.

4.1.1 Aggregated group features. We calculated the *maximum spatial distance* between group members (encoded as *dist_max*) over the distance between positions of the chest joint on the XZ-plane. We expect groups which are standing closer together to not interpret the presence of others as invading, meaning they have a stronger social bond to each other. Conversely, if participants choose to move further away in a task, this can be interpreted as diminishing social cohesion. Additionally, we related the interpersonal distance to notions from proxemics literature by categorizing the inter-personal distance as being inside *public space* (maximum distance above 3.6 meters), *social space* (between 3.6 and 1.2 meters) or *personal space* (below 1.2 meters) as defined by Hall *et al.* [17]. For this aim, we collected the *dist_max* values for each temporal segment in a histogram, where the bins reflect the three proxemic ranges. The histogram related features, encoded as *Public Space*, *Social Space*, *Personal Space* respectively, reflect the normalized size of each bin given by the amount of *dist_max* associated with each interval. We expect groups which share strong social bonds to stand closer to each other (e.g. a lower amount of distance in public space found over social space). Based on spatial distance, we further introduce a binary indicator for *body contact* (*touch*) between group members, which indicates that the hand joints of a participant are less than 15cm away from the upper body joints of another participant. We choose a conservative threshold for touch detection as for one sensor locations are not located at the fingertips but rather close to the palm. Secondly, the threshold is further aimed to enable us to capture touch at areas around the sensor (e.g. touching a participants elbow instead of their forearm). While in theory, touch could also be approximated from image or video data, the sensitivity of MoCap makes it much better suited to approximate haptic communications without the use of tactile sensors. We expect that signalling by touch can work both at communicating task-related information as well as convey social status [19]. We quantify kinesics related features on the *amount of walking* by calculating overall body movement in space and the *amount of hand movement* while standing still. Spatial body movement (encoded as *mov_walk*) is calculated by taking the average change in the position of participants' chest-joint in the XZ-plane. The amount of hand movement (encoded as *mov_hands*) is calculated by taking the average change in position of the hand joints in space for all time points where subjects aren't moving in space (i.e. the change in position of the Hip-joint on the XZ-plane less than 50cm over 1 second). Movement and gesturing may indicate active engagement in the activity and thus are expected to have a positive impact on predicting cohesion [14]. The next set of features relates to postural cues and posture differences among group members. Specifically, we considered the *postural expansion*, given by the bounding box volume of the body joints (inspired by [37]). This feature is given by the volume of the box being spanned by the maximum and minimum body joint coordinates in X-,Y- and Z-direction for each participant. The bounding box volume is then normalized over the arm-span and overall height of each participant in order to account for different body types. We computed both the average expansion (*pos_box*) and the expansion difference ratio (i.e. the difference of highest and lowest expansion at any time point, encoded as *pos_box_ratio*) in the group. We expect posture expansion to be related to notions of dominance and hierarchy, where small differences and big overall expansion to be positively correlated to social cohesion [49]. Finally, we defined a *visual facing* detector that detects whenever the chest-joint of a participant enters the *line of sight* of another participant. The *line of sight* of a participant is modeled as a cone extending orthogonally from the chest point of the participant with a 60° angle and a 3.6 meters depth (see Figure 2). These specifications are derived given our understanding of the biological findings about the extend of our focused field of view [48], and the meaningful distance intervals according to Hall's definition of proxemics [17]. Based on this, we calculated the total facing time (encoded in *fac_total*) as the total time a participant faces another one.

4.1.2 Innate group features. Innate group features, as opposed to aggregates, describe the group as a whole, rather than its composing members. Using our *facing* detector, we derive two group aspects by assessing the quality and amount of *turn-taking* in the group as well as the current *group formation.* Turn-taking commonly refers to the dynamics of group discourse, such as the amount of speech or the rate of speaker changes. We visually approximate active speaker role by assessing the person who is being *visually faced* using our *facing* detector. We derive the *active speaker* of the group as the person who is being faced the most at a point in time and a *floor exchange* as a change of the active speaker. If no participant is being faced or multiple participants are being faced equally, no active speaker is being detected. This way, we can estimate the *amount of floor exchanges* (fex), as well as the *participation equality* (peq) and *turn taking freedom* (ttf), as described in [26]. We introduce an additional turn-taking feature, called *facing balance* (fac_ratio), which captures balance of facing times towards other participants in the group. More precisely, let facing balance $\mathcal{V}$ be defined as:

$$\mathcal{V} = 1 - \alpha \cdot (max_p\{\mathcal{V}_p\} - min_p\{\mathcal{V}_p\}) \qquad with$$

$$\mathcal{V}_p = \frac{H_{max} - H(p)}{H_{max}}$$

where p is each member of the group, α is the proportion of time where anybody is being faced, $H(x)$ is the entropy over the facing probabilities towards each other member in the group and H_{max} a normalization coefficient as the maximum possible entropy for a given group size. Groups with a clearly defined leader who is often faced would thus have a lower balance than groups without clear leadership structure where facing is equally distributed. We expect that turn-taking features are positively correlated with social cohesion as the equal active engagement was shown to be predictive of social cohesion [20, 33]. Lastly, we quantified the formation of the group by visually detecting *the facing formation* (or F-formation) of the group. Facing formations indicate the presence of a shared interaction space and conversation floor, thus are expected to be positively correlated with both social and task cohesion. In our work, building on the idea of visual facing, we introduced a new strategy of F-formation detection (encoded as f_form), which approximates the interaction space (or *o-space*) of a group using intersecting *line-of-sight.* An F-formation at any point in time is detected if the space in which all three line-of-sight cones intersect is non-empty (see Figure 2). Our method this way is able to detect any F-formation where individuals circumvent a shared convex space. For three people, the detected F-formations are *circular* and *semi-circular* formation, where mutual facing may occur. Notably missing are non-enclosed o-spaces such as a *line* formation or *L-arrangement* (cf. [30]). F-Formation is quantified by the amount of time any type of F-formation is detected. We calculated overlapping line-of-sight by calculating polygon intersections using the python *Shapely* package[1]. Table 1 summarizes all the features developed for this study.

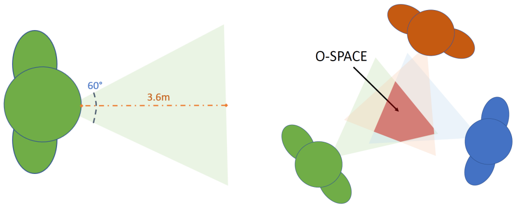

Figure 2: Visual attention and F-formation detection. An F-formation is detected when all visual attention cones intersect in a shared space (o-space).

Table 1: Name, description and type of the extracted features. Aggregate-Type refers to features assessed over group individuals while Group-Type asses the group state as a whole.

Encoding	*Description*	*Type*
max_dist	Max distance between group members	Aggregate
{Public/Social/Private} Space	Spacial association of max distance	Aggregate
mov_hands	Amount of hand movement when not walking	Aggregate
mov_walk	Average amount of walking	Aggregate
mov_walk_ratio	Movement difference in group	Aggregate
pos_box	Average posture expansion	Aggregate
pos_box_ratio	Posture difference in group	Aggregate
touch	Time touch is detected	Aggregate
fac_total	Time someone is being faced	Aggregate
fac_ratio	Facing balance	Group
peq	Participation equality	Group
ttf	Turn-taking freedom	Group
fex	Number of floor exchanges	Group
f_form	Presence, type of F-formation	Group

4.2 Feature aggregation and early selection

Each time-series was split into fixed-length slices of 20-second length over which the features were calculated. The short time window is inspired by the notion of *thin-slicing*, showing that affect and attitude can be accurately assessed when observing people over period of a few seconds [2]. We expect that these findings for general social interactions also translate to the perception of group cohesion. As part of our study, we analyzed 15 groups, resulting in a total of 1881 fixed-length slices of 20-second length (10 hours and 27 minutes of data). We then aggregated each feature over each slice by computing: average (avg), standard deviation (std), maximum value (max), minimum value (min) and distribution skewness ($skew$). Aggregation was done for all features aside from turn-taking features which were directly computed over each slice. This finally resulted in a set of 48 candidate features. We further reduced the feature set by using Kolmogorov-Smirnov statistic [23, 43] to find features whose distribution over the label set is different with a significance level of $p < 0.01$. This early feature selection ensures that we only consider features which are potentially meaningful for the prediction model.

4.3 Labeling strategy

One of the core aims of this work is to study dynamic changes in group cohesion over the set of consecutive tasks. For this aim, we computed task- and participant-specific labels which reflect the change in cohesion between subsequent tasks using the mean rank

[1] https://github.com/Toblerity/Shapely

difference of individual cohesion scores, which are provided in the dataset. We used a rank difference of scores since the scaling of Likert-scale responses as found in the dataset is often highly subject dependent. Equation 1 shows the notation to compute our prediction label given by the mean rank difference, where *rank* refers to the rank-transformation, *GEQ* refers to the sum of individual associated responses in the dataset, the so-called GEQ-score.

$$lab_t = \sum_{i \in \{1,2,3\}} \frac{rank(GEQ(i,t)) - rank(GEQ(i,t-1))}{3} \qquad (1)$$

This work is aimed at predicting whether or not a given short-length time window displays a decrease in group cohesion. Because of this, we binarized the social and task cohesion labels for our prediction setup, where a label $lab < 0$ indicates a *decrease* in cohesion and $lab >= 0$ indicates *no change or increase* in cohesion. Focusing on decreases in affect and behavior is an established method and has previously been done in similar works (e.g., [32, 44]).

4.4 Model prediction

Since we are trying to predict *decrease* for each slice for the social- and task-related dimension of group cohesion, we used a multilabel prediction approach. We chose a multilabel approach in order to jointly model the influence of the feature space on both dimensions by leveraging commonalities and common differences in the data during training. We further used a tree-based approach to avoid potential scaling issues with our features space which might have been present with distance-based models such as k-nearest neighbors or support vector machines. Because of these considerations we decided to use a random forest classification model to jointly predict social and task dimensions. In this work, we used the scikit-learn [35] implementation of the Random Forest Classifier. We employed a repeated nested 10-fold cross-validation with 5 repetitions across all slices to validate our model. Cross-folds were randomly generated for each repetition. Additionally, the cross-folds were stratified over the set of groups and the set of tasks to ensure that all groups are equally represented in the training and test sets in order not to set preference to tasks and groups with long duration on the experiment. The hyper-parameters of the prediction model were estimated using gridsearch on a 5-fold cross-validation over each train-validation split. The best model was selected using the highest average accuracy over both social and task dimension. Hyper-parameters tuned during model validation are the maximum depth of pruned decision trees, as well as the evaluation criterion. The best scoring model found during evaluation uses pruned decision trees with a maximum depth of 5, a total of 100 estimators, and *Gini imbalance* evaluation criterion. We finally evaluated our model by assessing the average joint accuracy over the test sets in the cross-validation, as well as the separate accuracy and f1-score for social and task cohesion.

4.5 Feature interpretation

Intuitively, we could expect that both social and task dimensions are expressed differently, through various modalities, as they serve different purposes (quantifying social bonds and quantifying task commitment respectively). Psychological models (e.g., [9, 41]), however, assume that theses dimensions are not orthogonal, meaning there may be behavioral correlates which are indicative for both dimensions. Since we trained a model to jointly predict the social and task dimensions of group cohesion, we expect that our predictive model is able to exploit the commonalities in predicting both the task and social dimension to improve generalization on the test data. From the trained model, we can then probe it to reveal (1) which features are informative for the model to predict task and social cohesion and (2) how higher and lower feature values impact the final model prediction. We quantified both the overall feature importance and whether the features' impact is positive or negative by analysing the SHAP values extracted from the fitted model using the overall most common best setup selected during cross-validation. SHAP values [28] are inspired by Shapley values [42], a notion from cooperative game theory where a common payout is distributed according to each participant's individual and shared contribution. In a similar vein, SHAP values assign an additive contribution towards the prediction output assigned by a trained model. This way, each variable of each sample in the dataset can be associated with a score that reflects how important it is for the overall prediction and if it positively or negatively impacts the final prediction. As we used a tree-based model, we employed TreeSHAP [27] to compute the associated SHAP values. The mean absolute SHAP value for each feature provides the overall feature importance over the whole dataset. The features' impact was quantified by calculating the Pearson correlation coefficient (*PCC*) between the feature values and the associated SHAP values. The sign of the correlation thus provides the type of impact (positive/negative) while the value indicates the linearity of this relationship (with $+1/-1$ indicating strong monotonic relationship).

5 RESULTS AND DISCUSSION

The following is a summary of the results of our analysis. Out of the total 1881 slices, we trained during our nested cross-validation on an average of 1354 slices (72%), validated the model on 339 slices (18%) and tested on the remaining 188 slices (10%). Using *mean rank difference*, the resulting labels are balanced in both task (51% *decrease*) and social (54% *decrease*). When viewed as a multiclass problem comprising the four classes *decrease/decrease*, *decrease/no decrease*, *no decrease/decrease* and *no decrease/no decrease*, the label distribution is 30%, 16%, 21%, 33% respectively. Using Kolmogorov-Smirnov statistics, we reduce the total set of considered features from 48 candidate features to a total of 37 features. The 11 removed features include aggregates for the skew of distribution, as well as f_form_min and fac_fex. Skewness, as a higher-order moment, might simply be less informative on the data than lower-order moments, f_form_min is not very informative since it only detects when touch is detected over the whole slice. Finally, fac_fex might not be informative on small time windows. We find that our prediction paradigm is able to correctly identify the cohesion dynamics with an average accuracy of 46% over the test sets for both task and social cohesion. We find that we are able to predict task cohesion with an average test accuracy of 64%(±3%) and social cohesion with an average test accuracy of 67%(±3%). The results show that our model is outperforming a baseline of random guessing (50%). Additionally, we manage to achieve an average f1-score of 64%(±3%) and 67%(±3%) for task and social cohesion, respectively. Figure

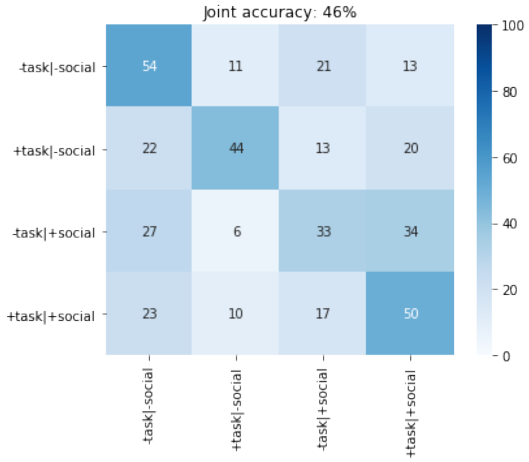

Figure 3: Confusion matrix for predicting dynamics for both task and social cohesion in percentages (baseline=25%). “ − ” and “ + ” refer to decrease or stable/decrease, respectively.

3 further shows the corresponding confusion matrix of the four possible joint classes when analyzing the results as a multiclass problem as described before. Again, we find that we are able to outperform above baseline (25% with random guessing). Furthermore, the results show that our model is confident in predicting the *non-conflicting* classes (decrease/decrease, no decrease/no decrease) while the performance is lower for the *conflicting* classes (i.e., decrease/no decrease, no decrease/decrease). This is likely because (1) these classes are slightly under-represented in the dataset using our labelling strategy and (2) we do not explicitly try to model these classes with our multilabel approach but rather infer each state based on the model's knowledge of social and task dynamics, respectively. The fact that we are still able to perform above chance level in all four classes reveals that general dynamics model for social and task cohesion is sufficient to also model the agreement between social and task cohesion dynamics. Finally, the model seems marginally better at differentiating the classes based on the social dimension than on the task dimension, also reflected in the fact that the overall average accuracy is higher for predicting social cohesion than task. The fact that we achieve higher performance for predicting social cohesion is in line with previous findings from the literature [33]. Figure 4 summarizes our analyses regarding the features' impact. We find that overall, the **maximum distance**, the **amount of walking**, the overall **posture expansion** and the **amount of inter-personal facing** in the group are found to be most meaningful in predicting both task and social cohesion.

When considering the 10 most important features for predicting task and social cohesion, we can extract a common subset of important features to analyze the commonalities and differences for predicting both dimensions. The top-10 features for each dimension comprise 64% and 56% of the total model impact respectively. In our feature set, we find that “*max_dist_min*”, “*pos_box_avg*”, “*mov_walk_avg*” and “*fac_avg*” are the most important features, associated with distance, body movement, posture and finally facing behavior. Among this set, we find that “*max_dist_min*” is negatively correlated with both task cohesion dynamics as well as social cohesion dynamics ($PCC = -0.86/-0.83$, respectively). This means that both task and social cohesion in a group drops when the distance between group members is high. This finding is in line with our previous assumption that groups standing closer together hold higher social bonds and don't feel the presence of others as invading. In a similar vein, we find that “*fac_avg*” is positively correlated with both dimensions ($PCC = 0.71/0.88$, respectively). This indicates that groups, which do not face each other during interactions, are more likely to experience a decrease of the task and social cohesion. Both “*pos_box_avg*” and “*mov_walk_avg*” show opposite expression for the two dimensions. Posture expansion is positively correlated in task cohesion ($PCC = 0.8$) while it is negatively correlated in social cohesion ($PCC = -0.51$). This is contrary to our prior assumption that overall high expansion correlates positively with social cohesion, as an erect posture was previously been found to be an indicator for social success [49]. The amount of movement of the participants in space on the other hand is negatively correlated in task cohesion ($PCC = -0.6$) while it is positively correlated in social cohesion ($PCC = 0.6$). Again, this is contrary to our prior assumptions as movement was assumed to be a sign of active task engagement. To further analyze these weak contingent linearities, Figure 5 shows the dependency plots between shap and feature values for “*pos_box_avg*” and “*mov_walk_avg*”. We find that the model assigns the most impact to low feature values, while assigning lower, consistent impact to higher feature values. For “*pos_box_avg*” we additionally find that there is an increase in feature impact for feature values higher than 0.6. In summary, when trying to monitor potential changes in the dynamics of cohesion in a group, our analysis reveals that one needs to look at the overall spacial distance of the group and the facing behavior, since they might indicate a lack of collaboration in the group. Secondly, the body posture and movement in the group should also be monitored since extreme posture changes or low movement might be indicative of resignation or conflict [8].

6 CONCLUSION AND FUTURE WORK

In this paper, we achieved promising results on automatically predicting variations of the task and social dimensions of cohesion from the group perspective (i.e., using self-assessments of cohesion) by analyzing MoCap-based behavioral features. Results suggest that our set of MoCap-based features, derived from existing video-based features, is successful at capturing relevant nonverbal cues related to cohesion. Our proposed approach uses tools from cooperative game theory to generate knowledge about the features' impact on the model predictions. It highlights that the *maximum distance* between group members, the overall *posture expansion* and the *amount of facing* between each person have the most significant (positive or negative) impact on the model. Moreover, it shows that there is a sizeable intersection of features found to be important for predicting both task and social cohesion. This fact implies that there exists a common set of behavioral correlates for both dimensions, meaning that they are indeed correlated with each other. Furthermore, because some of our common features

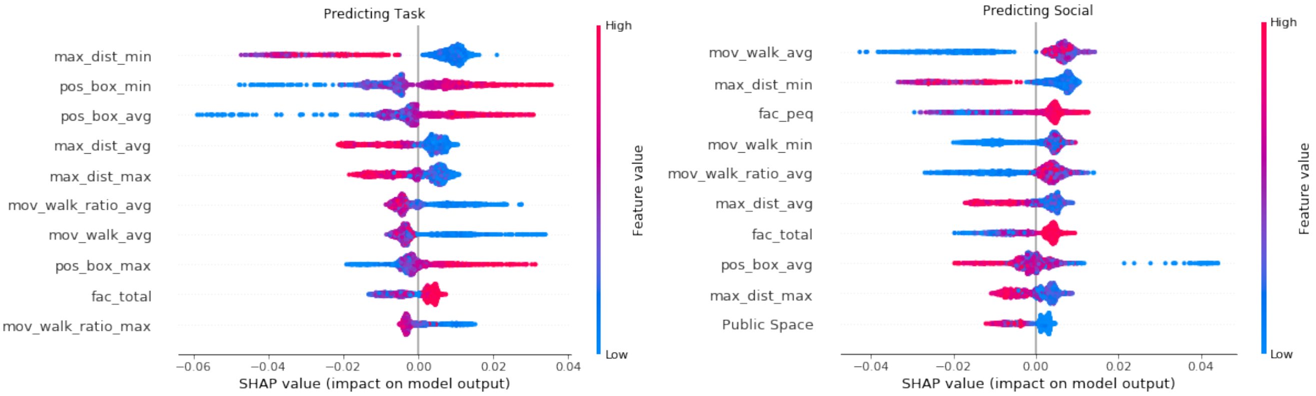

Figure 4: Impact of the 10 most informative features on the model for predicting task (left) and social (right) cohesion dynamics. A high feature value associated with a positive SHAP value indicates that the feature is positively impacting the model output.

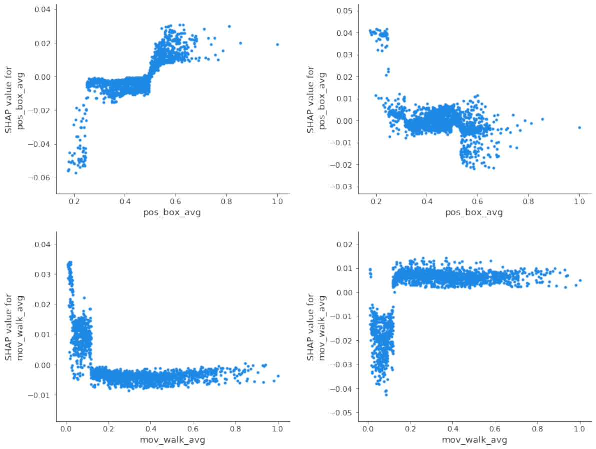

Figure 5: Dependency plots between SHAP and feature values of task (left) and social (right) cohesion for contingent variables "*pos_box_avg*" (top) and "*mov_walk_avg*" (bottom).

hold opposing model impact between both dimensions (e.g., posture expansion), we can further deduce that they do not manifest similarly, confirming the assumptions on the interplay of these dimensions from the original model [41]. Our analysis reveals that our set of MoCap based-features are informative for monitoring group cohesion when labels are based on self-assessments. These features were explicitly modeled to reflect nonverbal behavioral cues which are traditionally computed on the visual channel. This supports our assumption, that a reason for the poor performance of visual features might be the way they are extracted (low-level and model-based extraction on images and videos). Features found in this study to be informative encompass both low-level extracted information (overall body motion) and high-level visual information (inter-personal distance, posture and facing behavior). This study lays the foundations for the development of computational models to study the dynamics of cohesion. Some points, however, remain to be addressed. Indeed, this work focuses on MoCap-based features to detect variations in cohesion. It remains to be seen how much additional information this modality provides when integrating it into a multimodal (audio-visual) framework. One way the presented modeling pipeline could include more fine-grained dynamic changes is by changing the way cohesion dynamics are reflected in each temporal segment. This can be done at the labeling stage or at the model level. Concerning the labels, we used a mean rank difference approach between two tasks in order to approximate the variations of cohesion. A more complex labeling strategy could be developed to integrate information such as the inter-rater agreement. It should also minimize biases introduced by either self or external assessments [46] by potentially combining both ratings. From the model perspective, we used a multilabel approach to jointly predict a decrease in both task and social dimensions by considering each segment independently. Future work could also include the temporal dependencies between slices to directly model temporal developments on a large time scale. As a first exploration of the dynamics of cohesion based on self-assessments, this work provides useful guidelines for the design, use and interpretation of MoCap-based features related to cohesion and how to assess their impact on the model performances.

ACKNOWLEDGMENTS

This work has been partially supported by the French National Agency (ANR) in the frame of its Technological Research JCJC program (GRACE, project ANR-18-CE33-0003-01, funded under the Artificial Intelligence Plan).

REFERENCES

[1] J. A. Allen, C. Fisher, M. Chetouani, M. M. Chiu, H. Gunes, M. Mehu, and H. Hung. 2017. Comparing Social Science and Computer Science Workflow Processes for Studying Group Interactions. *Small Group Research* 48, 5 (2017), 568–590.

[2] Nalini Ambady and Robert Rosenthal. 1993. Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness. *Journal of personality and social psychology* 64, 3 (1993), 431.

[3] Aristotle. 4th Century BC. *Politics.*

[4] Reshmashree Bangalore Kantharaju, Caroline Langlet, Mukesh Barange, Chloé Clavel, and Catherine Pelachaud. 2020. Multimodal Analysis of Cohesion in Multiparty Interactions. In *Proceedings of The 12th Language Resources and Evaluation Conference*. European Language Resources Association, 498–507.

[5] Daniel J Beal, Robin R Cohen, Michael J Burke, and Christy L McLendon. 2003. Cohesion and Performance in Groups: A Meta-Analytic Clarification of Construct Relations. *Journal of Applied Psychology* 88, 6 (2003), 989–1004.

[6] Kenneth A Bollen and Rick H Hoyle. 1990. Perceived cohesion: A conceptual and empirical examination. *Social forces* 69, 2 (1990), 479–504.

[7] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2018. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. *arXiv preprint arXiv:1812.08008* (2018).

[8] Dana R. Carney, Judith A. Hall, and Lavonia Smith LeBeau. 2005. Beliefs about the nonverbal expression of social power. *Journal of Nonverbal Behavior* 29, 2 (2005), 105–123.

[9] A. V. Carron, W. N. Widmeyer, and L. R. Brawley. 1985. The Development of an Instrument to Assess Cohesion in Sport Teams: The Group Environment Questionnaire. *Journal of Sport Psychology* 7, 3 (1985), 244–266.

[10] Abhinav Dhall. 2019. EmotiW 2019: Automatic Emotion, Engagement and Cohesion Prediction Tasks. In *2019 International Conference on Multimodal Interaction.* Association for Computing Machinery, 546—-550.

[11] Abhinav Dhall, Roland Goecke, Shreya Ghosh, Jyoti Joshi, Jesse Hoey, and Tom Gedeon. 2017. From individual to group-level emotion recognition: Emotiw 5.0. In *Proceedings of the 19th ACM international conference on multimodal interaction.* 524–528.

[12] Kenneth L Dion. 2000. Group Cohesion: From Field of Forces to Multidimensional Construct. *Group Dynamics: Theory, Research, and Practice* 4, 1 (2000), 7–26.

[13] S. Ghosh, A. Dhall, N. Sebe, and T. Gedeon. 2019. Predicting Group Cohesiveness in Images. In *2019 International Joint Conference on Neural Networks (IJCNN).* 1–8.

[14] Susan Goldin-Meadow and Martha Wagner Alibali. 2013. Gesture's role in speaking, learning, and creating language. *Annual review of psychology* 64 (2013), 257–283.

[15] James Griffith. 1988. Measurement of group cohesion in US Army units. *Basic and applied social psychology* 9, 2 (1988), 149–171.

[16] Da Guo, Kai Wang, Jianfei Yang, Kaipeng Zhang, Xiaojiang Peng, and Yu Qiao. 2019. Exploring Regularizations with Face, Body and Image Cues for Group Cohesion Prediction. In *2019 International Conference on Multimodal Interaction.* 557–561.

[17] Edward Twitchell Hall. 1966. *The hidden dimension.* Vol. 609. Garden City, NY: Doubleday.

[18] Anjali Hans and Emmanuel Hans. 2015. Kinesics, Haptics and Proxemics: Aspects of Non -Verbal Communication. *IOSR Journal Of Humanities And Social Science (IOSR-JHSS)*, 47–48.

[19] Richard Heslin. 1974. Steps toward a taxonomy of touching. *Paper presented to the annual meeting of the Midwestern Psychological Association. Chicago* (1974).

[20] Hayley Hung and Daniel Gatica-Perez. 2010. Estimating Cohesion in Small Groups Using Audio-Visual Nonverbal Behavior. *IEEE Transactions on Multimedia* 12, 6 (2010), 563–575.

[21] Dinesh Babu Jayagopi, Hayley Hung, Chuohao Yeo, and Daniel Gatica-Perez. 2009. Modeling dominance in group conversations using nonverbal activity cues. *IEEE Transactions on Audio, Speech, and Language Processing* 17, 3 (2009), 501–513.

[22] Simone Kauffeld, Nale Lehmann-Willenbrock, and Annika L. Meinecke. 2018. *The Advanced Interaction Analysis for Teams (act4teams) Coding Scheme.* Cambridge University Press, 422—-431.

[23] Andrey Kolmogorov. 1933. Sulla determinazione empirica di una lgge di distribuzione. *Inst. Ital. Attuari, Giorn.* 4 (1933), 83–91.

[24] Steve WJ Kozlowski. 2015. Advancing research on team process dynamics: Theoretical, methodological, and measurement considerations. *Organizational Psychology Review* 5, 4 (2015), 270–299.

[25] Uliyana Kubasova, Gabriel Murray, and McKenzie Braley. 2019. Analyzing Verbal and Nonverbal Features for Predicting Group Performance. In *Proc. Interspeech 2019.* ISCA, 1896–1900.

[26] Catherine Lai and Gabriel Murray. 2018. Predicting group satisfaction in meeting discussions. In *Proceedings of the Workshop on Modeling Cognitive Processes from Multimodal Data.* 1–8.

[27] Scott M Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. 2020. From local explanations to global understanding with explainable AI for trees. *Nature machine intelligence* 2, 1 (2020), 2522–5839.

[28] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in neural information processing systems.* 4765–4774.

[29] Lucien Maman, Eleonora Ceccaldi, Nale Lehmann-Willenbrock, Laurence Likforman-Sulem, Mohamed Chetouani, Gualtiero Volpe, and Giovanna Varni. 2020. GAME-ON: A Multimodal Dataset for Cohesion and Group Analysis. *IEEE Access* 8 (2020), 124185–124203.

[30] Paul Marshall, Yvonne Rogers, and Nadia Pantidi. 2011. Using F-formations to analyse spatial patterns of interaction in physical environments. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work.* 445–454.

[31] I. Mccowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska Masson, W. Post, D. Reidsma, and P. Wellner. 2005. The AMI meeting corpus. *Int'l. Conf. on Methods and Techniques in Behavioral Research* (2005).

[32] Philipp Müller, Michael Xuelin Huang, and Andreas Bulling. 2018. Detecting low rapport during natural interactions in small groups from non-Verbal behaviour. In *23rd International Conference on Intelligent User Interfaces.* 153–164.

[33] Marjolein C Nanninga, Yanxia Zhang, Nale Lehmann-Willenbrock, Zoltán Szlávik, and Hayley Hung. 2017. Estimating Verbal Expressions of Task and Social Cohesion in Meetings by Quantifying Paralinguistic Mimicry. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction.* Association for Computing Machinery, 206—-215.

[34] D. Olguin and A. Pentland. 2010. Sensor-based organisational design and engineering. *International Journal of Organisational Design and Engineering* 1, 1 (2010), 69–97.

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

[36] Alex Pentland. 2007. Social signal processing [exploratory DSP]. *IEEE Signal Processing Magazine* 24, 4 (2007), 108–111.

[37] Stefano Piana, Maurizio Mancini, Antonio Camurri, Giovanna Varni, and Gualtiero Volpe. 2013. Automated analysis of non-verbal expressive gesture. In *Human Aspects in Ambient Intelligence.* Springer, 41–54.

[38] Lisa Rosh, Lynn R Offermann, and Rhonda Van Diest. 2012. Too close for comfort? Distinguishing between team intimacy and team cohesion. *Human Resource Management Review* 22, 2 (2012), 116–127.

[39] Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. 1978. A simplest systematics for the organization of turn taking for conversation. In *Studies in the organization of conversational interaction.* Elsevier, 7–55.

[40] Eduardo Salas, Rebecca Grossman, Ashley M Hughes, and Chris W Coultas. 2015. Measuring team cohesion: Observations from the science. *Human factors* 57, 3 (2015), 365–374.

[41] Jamie B Severt and Armando X Estrada. 2015. On the function and structure of group cohesion. In *Team cohesion: Advances in psychological theory, methods and practice.* Vol. 17. Emerald Group publishing limited, 3–24.

[42] Lloyd S Shapley. 1951. Notes on the n-Person Game—II: The Value of an n-Person Game. (1951).

[43] Nickolay Smirnov. 1948. Table for estimating the goodness of fit of empirical distributions. *The annals of mathematical statistics* 19, 2 (1948), 279–281.

[44] Daniel Szafir and Bilge Mutlu. 2012. Pay Attention! Designing Adaptive Agents That Monitor and Improve User Engagement. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Austin, Texas, USA) *(CHI '12).* Association for Computing Machinery, New York, NY, USA, 11—-20.

[45] Michael Tomasello, Alicia P. Melis, Claudio Tennie, Emily Wyman, and Esther Herrmann. 2012. Two Key Steps in the Evolution of Human Cooperation. *Current Anthropology* 53, 6 (2012), 673–692.

[46] Alessandro Vinciarelli and Gelareh Mohammadi. 2014. A survey of personality computing. *IEEE Transactions on Affective Computing* 5, 3 (2014), 273–291.

[47] Alessandro Vinciarelli, Maja Pantic, Hervé Bourlard, and Alex Pentland. 2008. Social signal processing: state-of-the-art and future perspectives of an emerging domain. In *Proceedings of the 16th ACM international conference on Multimedia.* 1061–1070.

[48] Henry Kenneth Walker, Wilbur Dallas Hall, and John Willis Hurst. 1990. *Clinical Methods: The History, Physical, and Laboratory Examinations.* Butterworths.

[49] Glenn E Weisfeld and Jody M Beresford. 1982. Erectness of posture as an indicator of dominance or success in humans. *Motivation and Emotion* 6, 2 (1982), 113–131.

[50] Tien Xuan Dang, Soo-Hyung Kim, Hyung-Jeong Yang, Guee-Sang Lee, and Thanh-Hung Vo. 2019. Group-level Cohesion Prediction using Deep Learning Models with A Multi-stream Hybrid Network. In *2019 International Conference on Multimodal Interaction.* 572–576.

[51] Yanxia Zhang, Jeffrey Olenick, Chu-Hsiang Chang, Steve WJ Kozlowski, and Hayley Hung. 2018. TeamSense: assessing personal affect and group cohesion in small teams through dyadic interaction and behavior analysis with wearable sensors. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (2018), 1–22.