This WACV 2020 paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Multiparty Visual Co-Occurrences for Estimating Personality Traits in Group Meetings

Lingyu Zhang
Rensselaer Polytechnic Institute
zhangl34@rpi.edu

Indrani Bhattacharya
Rensselaer Polytechnic Institute
bhatti@rpi.edu

Mallory Morgan
Rensselaer Polytechnic Institute
morgam11@rpi.edu

Michael Foley
Northeastern University
foley.mic@husky.neu.edu

Christoph Riedl
Northeastern University
c.riedl@northeastern.edu

Brooke Foucault Welles
Northeastern University
b.welles@northeastern.edu

Richard J. Radke
Rensselaer Polytechnic Institute
rjradke@ecse.rpi.edu

Abstract

Participants' body language during interactions with others in a group meeting can reveal important information about their individual personalities, as well as their contribution to a team. Here, we focus on the automatic extraction of visual features from each person, including her/his facial activity, body movement, and hand position, and how these features co-occur among team members (e.g., how frequently a person moves her/his arms or makes eye contact when she/he is the focus of attention of the group). We correlate these features with user questionnaires to reveal relationships with the "Big Five" personality traits (Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism), as well as with team judgements about the leader and dominant contributor in a conversation. We demonstrate that our algorithms achieve state-of-the-art accuracy with an average of 80% for Big-Five personality trait prediction, potentially enabling integration into automatic group meeting understanding systems.

1. Introduction

An individual's personality, typically measured using the "Big-Five" model of *Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism* [42], has been shown to be closely related to job performance, motivation, and creativity [32, 50, 31, 36, 55]. Further, rapport among team members and team productivity often depend on the personalities of the individuals in the team. In this paper, we use computer vision feature extractors to identify behavior cues and investigate actions that are closely related to personality traits, emergent leadership, and the perceived contribution of individuals in groups performing a collaborative task.

Psychologists have found that body language such as facial expressions, body postures, arm gestures, or eye gaze are related to individuals' personality and instantaneous mood [46, 20, 7, 45]. Body movement is related to emotional arousal [23]; for example, strong and fast movements are correlated with rejection emotions such as anger or antipathy [18], while raising one's eyebrows slightly conveys a sense of uncertainty [47]. People tend to move their eyes more while thinking [22], and people who have higher scores in Openness tend to increase eye fixation points [41]. Additionally, behavior patterns are often affected by other people during social interactions [27, 39]; this change in behavior can also reflect the personality of the individual [38]. One challenge is that some emotional reactions (mostly extreme emotions) can only be seen under certain uncommon conditions (e.g., anger or fear in a fight, sadness or comfort at a ballgame). Our work focuses on group discussions common in everyday work life, where such extreme emotions as fear, anger, or elation are rare. We propose algorithms to automatically extract non-verbal interaction cues from participants' individual behaviors and their response to the other people involved, and estimate personality traits and group perceptions of leadership and contribution.

As shown in Figure 1, we extract a set of key action events from frontal videos of each meeting participant, including visual focus of attention (VFOA), level of body movement, and relative hand-face position. We then investigate how the *co-occurrence* of these visual events can be used to accurately estimate and predict individual personality traits and the perceived emergent leadership/contribution

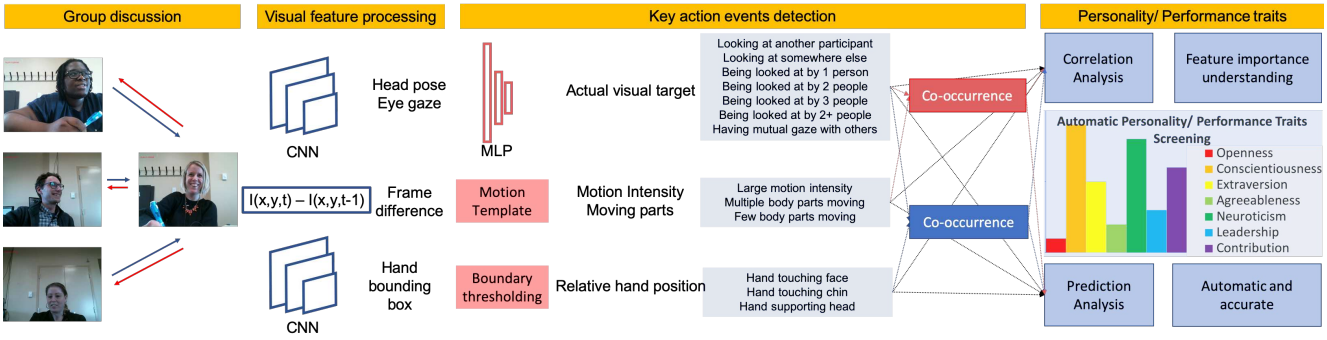


Figure 1: The overall pipeline of our algorithm, indicating the types of visual features we extract and their co-occurrences, for the purposes of estimating and predicting personality and perceived leadership and contribution.

in the group discussion. These algorithms could be incorporated into future virtual meeting agents that can automatically evaluate personality and performance, and ultimately facilitate a discussion to keep it on track.

This paper has three key contributions in the context of the large literature on video-based personality assessment:

- **Multimodality:** We investigate both spatial and temporal information from several distinct streams of data (face/eye, body, and hand activity), demonstrating significant performance boosts over using a single modality alone.
- **Multiparty:** For each person, we construct co-occurrence features based not only on her/his own behavior patterns, but also from the context of her/his interactions with others in the group (e.g., distinguishing how someone acts when she/he is the center of attention vs. being ignored).
- **Interpretability:** We perform correlation analysis between each feature and every output variable, enabling an understanding of which features are more related to which traits, providing a path to build more interpretable prediction models.

2. Related Work

Datasets: Personality trait identification is an active area of research in social signal processing. Relevant works to automatically estimate personality include Al Moubayed et al. [1], where the authors used eigenface features from still face images, together with Support Vector Machines for binary classification of each of the Big-Five personality traits. Dhall and Hoey [19] extracted facial features combined with background environment information to predict user personality from Twitter profile images. Guntuku et al. [29] predicted multiple mid-level cues, such as gender and age, from low-level visual features and combined them to model personality based on "selfies".

One area of research specifically targets estimating personality from first impressions. The ChaLearn First Impression Database [48, 24] is largely used for modelling personality from short 15-second video sequences. Bekhouche et al. [8] proposed a method for automatic job candidate screening for Big-Five personality traits from short videos in [48], by feeding multi-level facial texture features to 5 different Support Vector Regressors. Unlike the video sequences in ChaLearn in which only one person is talking to a camera, online conversational videos have also been used to develop personality models [53, 13, 2, 26]. Biel et al. [13] focused on facial expression analysis from 5-minute Youtube vlog videos [12]. Each video frame was categorized as active or inactive based on the facial expression signal, and metrics such as the proportion of total active time were proposed to correlate with the Big-Five personality traits.

More closely related to our work is personality trait estimation in the context of group discussions. For example, in the job market, personality and potential leadership assessment are used in a common pre-employment test called Leaderless Group Discussion (LGD) [6, 56], in which applicants are formed into groups, given a goal, and observed during their discussion or consensus process. The ELEA corpus [51], originally developed to study emergent leadership in small groups, contains frontal video camera recordings of 3–6 individuals performing a group collaborative task. Aran and Gatica-Perez [4] applied ridge regression on a combination of audio-visual nonverbal features for personality trait classification, with a best result of 75% for predicting extraversion on a subset of the ELEA corpus. Okada et al. [44] extracted multimodal events such as "looking while speaking" for personality trait prediction. Beyan et al. [9] used a deep-neural-network-based approach to estimate leadership and extraversion. The ELEA corpus has also been used for modelling dominance in group discussions. For example, Hung et al. [33] developed a fast speaker diarization algorithm and used speaking length as a feature, while Jayagopi et al. [34] combined audio cues (e.g., interruptions and speaking turns) with visual cues to estimate the most dominant person in the discussion.

In our work, we use entire 15-minute group meetings to analyze participants' body language for estimating personality traits, perceived leadership, and contribution, instead of using still images or short video clips. It is unlikely for a meeting participant or a job candidate to have expressive or exaggerated behavior as in a Youtube vlog, and the background scene in a typical workplace is uninformative compared to a Twitter profile image background. Thus, our dataset represents a more realistic and common environment that should hopefully transfer easily to similar workplace scenarios.

Features and Analysis: A large body of work addresses the automatic extraction of visual behavior cues that relate to an individual's cognitive state. Yan et al. [58] used eye width and eyebrow shape as features for a Support Vector Machine to estimate personality traits. Joo et al. [35] observed that visual features such as hairstyle, body part appearance (e.g., whether the eyeball is fully visible), and relative distances between facial features could be used to infer social judgments about a person. Okada et al. [44] derived measures from visual focus of attention (e.g., the proportion of time that a target person is looking at others) to estimate personality traits. Biel and Gatica-Perez [13] estimated the canonical facial expressions (Anger, Contempt, Disgust, Fear, Neutral, Surprise, Smile) in each frame and applied Support Vector Regression to do personality trait prediction. Bhattacharya et al. [10] studied correlations between visual, non-verbal, and verbal speech features and group leaders/contributors, but did not collect facial video or investigate personality traits. Many researchers have proposed to use the weighted motion energy image (MEI) as a feature for estimating personality or leadership [3, 57, 52, 44, 16]. Celiktukan et al. [16] used a bag-of-words approach to represent a video clip as a histogram of key arm gestures for personality trait classification. Several Convolutional Neural Networks (CNNs) have been proposed for end-to-end inference from image sequences to personality traits [28, 54, 59].

The feature set in our work subsumes many of the above features, and a key novel aspect of our approach is the extraction of *co-occurrences* within this rich feature set based on the dynamics of the group, not just one individual. Additionally, our correlation analysis between input features and output traits is more interpretable than an end-to-end black-box neural network.

3. Dataset Construction

Our dataset contains 15 group meetings, 12 of which had 3 people, and 3 of which had 4 people, resulting in a total of 48 individual recordings. Each individual only participated in one group meeting. Participants were given a well-known group consensus problem called the Lunar Survival Task [30]. In this task, participants imagine they are survivors of a spaceship crash on the moon, and must rank-order 15 items (e.g., tanks of oxygen, a stellar map, a magnetic compass) that they may need to survive a long trek to their base. Participants first rank the items on their own, and are then instructed to reach consensus as a group within 15 minutes. Prior to the task, participants self-assess their own Big-Five personality traits using an instrument called the BFI-10 [49]. After the task, each participant assessed the degree to which the other group members acted as leaders of or major contributors to the discussion.

For personality self-assessment, each participant i has a score Y_i^P ranging from 1 to 10 on each dimension of the Big-Five taxonomy. For leadership and contribution assessments, each participant receives a score from the other participants on a scale of 1 to 5. These scores are averaged to produce a perceived leadership score Y_i^L and perceived contribution score Y_i^C for each participant (participants do not rate themselves on leadership and contribution).

During each 15-minute group discussion, participants sat on either side of a long conference table, and each participant was recorded by an individual front-facing closeup camera that captured their head and upper body. The cameras were rigidly mounted on a custom wooden rig and synchronized by a single mini-PC, resulting in individual videos of 960×720 resolution at 20 fps. Figure 2 illustrates the seating position and room layout.

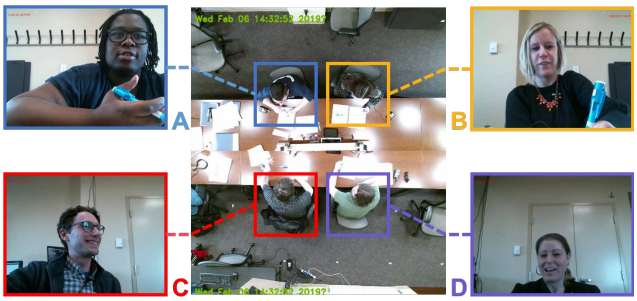


Figure 2: Seating position and room layout. The multi-camera rig is visible in the middle of the table.

We computed the standard deviations for the output variables of Extraversion, Agreeableness, Conscientiousness, Neuroticism, Openness. Leadership and Contribution as 2.1, 1.4, 1.5, 1.7, 1.8, 0.5, and 0.9, respectively.

4. Approach

Our goal is to investigate which kinds of body language are related to meaningful social signals including personality and performance traits. We specifically hypothesize that:

H1: The combination of facial, hand, and body behavior cues can reveal important information regarding a participant's cognitive state.

H2: The combination of co-occurrent events can better describe the complex behavior patterns than individual events alone.

Typical methods (e.g., [10, 44]) compute the frequency F_α, F_β of the binary occurrences I^t_α, I^t_β of individual event α and β over a given time span, i.e.,

$$F_\alpha = \frac{\sum_t^T I^t_\alpha}{T} \qquad F_\beta = \frac{\sum_t^T I^t_\beta}{T} \tag{1}$$

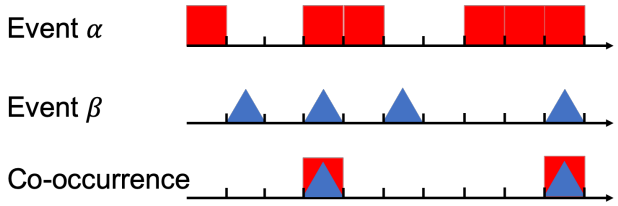


Figure 3: Event co-occurrences over time.

In contrast, we consider the co-occurrence $I^t_{\alpha\wedge\beta}$ at time t of events α and β, as shown in Figure 3, and compute the corresponding frequency

$$F_{\alpha\wedge\beta} = \frac{\sum_t^T I^t_{\alpha\wedge\beta}}{T} \tag{2}$$

We first design the individual key event detection algorithms built upon several visual feature extractors, and then investigate how co-occurrent events contribute to the final social signal estimation based on correlation and prediction experiments.

4.1. Visual Feature Preprocessing

Based on hypothesis **H1**, we first extract behavior cues from the face, hands, and body. We apply several feature descriptors for visual feature preprocessing to estimate relative head pose, eye gaze angle, consecutive frame difference in a short time window, and hand bounding box location. These features will later be used to detect higher-level key visual action events.

Facial features. For the front-facing videos, we apply a feature pre-processing step based on the OpenFace Facial Behavior Analysis Toolkit [5] to extract 68 facial landmarks, head pose, and eye gaze angle relative to the camera center. These are later used for estimating the visual focus of attention and the relative position of the hands and face.

Body features. For body features, we focus on frame-level movements. Since in our individual recordings, the background is constant and the participant takes up most of the frame, we apply a motion estimation algorithm based on frame differencing to calculate the motion observed during the video clip.

Hand features. We apply a pre-trained CNN model [43] consisting of 3 convolutional layers and 1 fully connected layer to extract the hand bounding boxes from each video frame in individual recordings. The coordinates of the box are later used to determine the key visual events relating hand positions to the face.

4.2. Key Action Event Detection

Given the pre-processed visual features, we detect key action events from the synchronized videos of each participant, including the relative frequency of events based on visual focus of attention (VFOA), hand-face relationship, and the aggregated body movement over short time windows. We discuss each key action event in turn, as well as the critical aspect of their co-occurrences.

4.2.1 Visual Focus of Attention

For each participant, we define the visual focus of attention (VFOA) at time t as the location where she/he is looking. Based on the seating positions, the VFOA is quantized into labels for the person seated adjacent to, directly across from, or diagonally across from the participant, as well as a category for "somewhere else", which is generally the piece of paper with the lunar survival task items in front of them. We annotated 52960 frames (from 4 different meetings, 7 to 10 minutes of data from each meeting) of VFOA and split the data into 85% for training and 15% for testing.

Figure 4: Complex head pose and eye gaze angles while looking at the same target.

Our goal is to identify the VFOA for participants in different seating positions based on the raw angles output from OpenFace. Since the calibration of the camera system is unknown, and the spatial locations of participants were not strictly controlled, we cannot immediately infer the visual targets from the raw head pose or eye gaze angles. As demonstrated in Figure 4, the VFOA target for each pair of participants is the same while the apparent head pose and eye gaze angles vary significantly. Thus, we constructed a multilayer perceptron (MLP) model to do VFOA classification for each video frame. The input is the 5-dimensional concatenated vector of head pose and eye gaze angles, and the model contains 7 fully connected layers with ReLu activation to model the nonlinearity. For each seating position, the model maintains the same structure, but is trained sep-

arately using different hyperparameters for optimized performance. Our optimized VFOA model [60] achieves an average prediction accuracy of 90% on the testing set.

We next extract 7 binary per-frame, per-participant visual events based on the frame-by-frame VFOA estimation, which include looking at another participant (I_{AG}), looking somewhere else (I_E), being looked at by another participant (I_{AR}), being looked at by 2 participants (I_{AC2}), by 3 participants (I_{AC3}), or by at least 2 participants (I_{AC2+}), and having mutual gaze with others (I_{MG}). We then calculate the frequency of these events $F_{AG}, F_{AC2}, F_{AC3}, F_{AC2+}, F_{MG}$ for the entire meeting duration as well as two extra metrics F_{ATR} and F_{ATQ}. Denoting I^t_{AB} as the binary frame-level decision that Person A is looking at Person B at frame t:

- F_{ATR}, the amount of attention received by participant A. If there are K participants and N frames of video,

$$F_{ATR} = \frac{1}{KN} \sum_{t=1}^{N} \sum_{k=1}^{K} I^t_{kA} \tag{3}$$

- F_{ATQ}, the attention quotient, computed as F_{ATR}/F_{AG}.

4.2.2 Body Movement Level

Given the difference between consecutive video frames in the preprocessing step, following the method in the Motion Energy Image (MEI) framework [17], the frame difference is thresholded resulting in a binary image $D(x, y, t)$ at time t to indicate that motion occurs at pixel (x, y). In a predefined short time window τ leading up to time t, the motion energy image $E_\tau(x, y, t)$ is the union of the binary images as illustrated in Figure 5.

Figure 5: Original image and MEI sample output.

Movement Intensity. We quantify the body motion of participants by the intensity of the motion detected in the short time window. Previous work [52, 44, 16] used the average or standard deviation value of the MEI over the entire frame/video clip as features for personality or leadership identification. In contrast, we seek to model co-occurrent situations such as a person who tends to move her/his fingers frequently while thinking, or one who moves her/his body more when receiving attention from others. Therefore, we compute MEI on a frame-by-frame basis instead of aggegrating out the temporal resolution. Specifically, for the detected MEI $E_\tau(x, y, t)$ at each time frame t, we compute the motion intensity as the summation of MEI over the whole image and normalize it through the entire video clip.

$$\hat{I}_m(t) = \frac{\sum_x \sum_y E_\tau(x, y, t)}{max_t(\sum_x \sum_y E_\tau(x, y, t))} \tag{4}$$

Number of Moving Parts. To additionally distinguish scenarios in which a person moves several body parts independently vs. moving her/his whole body, we estimate the number of moving body parts to supplement the motion intensity. In particular, we apply a motion segmentation model [14] to roughly segment the moving pixels into independent moving parts. Specifically, through adding a decay operator to the MEI, we compute a grayscale motion history image (MHI) in which the intensity of a pixel is proportional to the recency of motion at that pixel. If there is motion at (x, y) at current time t, the value of the pixel in the MHI image would be τ, the largest value. Otherwise, the MHI will decrease by 1 for every previous frame in which motion was not observed.

$$H_\tau = \begin{cases} \tau & \text{if } D(x, y, t) = 1 \\ max(0, H_\tau(x, y, t-1) - 1) & \text{otherwise} \end{cases}$$

Based on a connected component analysis of the MHI image using currently-moving pixels as seeds, we extract the number of moving parts, marked as a red number in the images in Figure 6. The blue bounding box on the MHI shows the motion boundary of each independent moving part, and the angle of the red line indicates the motion direction of the moving part (calculated from the gradient orientation in the MHI).

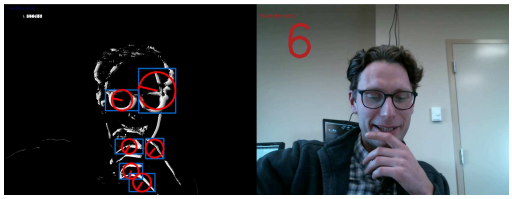
We compute the normalized value for the number of moving parts at each frame as

$$\hat{N}_m(t) = \frac{N_m(t)}{max_t N_m(t)} \tag{5}$$

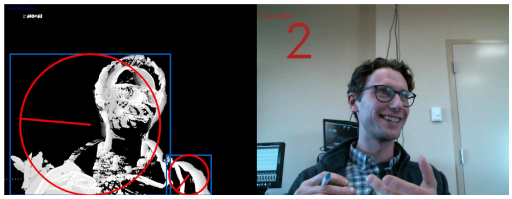
Since the per-frame measurements for motion intensity and number of moving parts are normalized by the maximum value observed for the same person throughout the whole meeting, the measurements we obtained are basically invariant to environmental conditions, the subjects' clothing, or their distance from the camera.

We can now define binary motion events in each frame for each participant by thresholding the quantities above on frames that contain large motion intensity I_{lm} or multiple moving body parts I_{mm}. Specifically, the visual events related to motion dynamics are defined as:

$$I_{lm} = \begin{cases} 1 & I_m(t) > \kappa \\ 0 & else \end{cases} \quad I_{mm} = \begin{cases} 1 & \hat{N}_m(t) > \lambda \\ 0 & else \end{cases} \tag{6}$$



(a) Multiple body parts moving.



(b) Few body parts moving.

Figure 6: Moving part extraction with (a) multiple and (b) few moving parts.

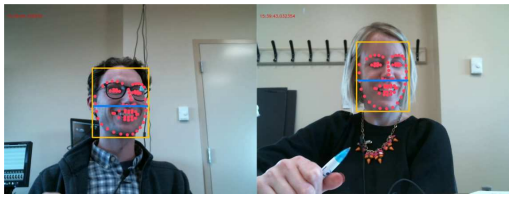
Since I_{lm} and I_{mm} are normalized across the whole video clip, we set κ and λ to be 0.5. We found that the major correlations discussed in Section 5.1 are robust to the values of these thresholds. We then derive motion-based features per participant, including:

- F_{LM}, the fraction of frames in which a participant has large motion intensity
- F_{MM}, the fraction of frames in which a participant has multiple body parts moving

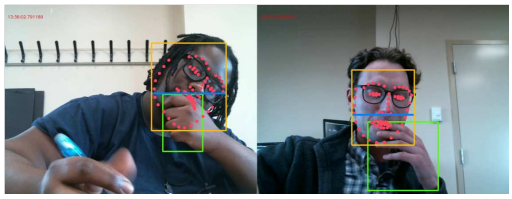
Based on hypothesis **H2**, in addition to the motion dynamics features on their own, we also detect co-occurrence events related to the VFOA features in the previous section (e.g., a participant having large motion intensity while being looked at, or moving multiple body parts while giving visual attention to others). By combining 7 VFOA events and the 3 motion events I_{lm}, I_{mm}, and $I_{fm} = \neg I_{mm}$, we construct 21 co-occurrence events in total, which are normalized to produce the fraction of total frames that each participant is in each co-occurrent event condition. These co-occurrent features play the primary role in the correlation and prediction framework discussed in Section 5.

4.2.3 Hand-Face Relative Position

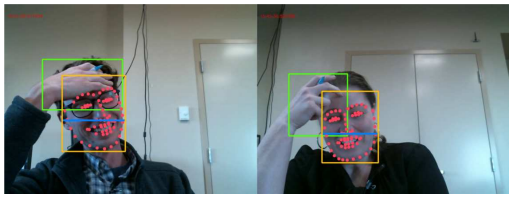
Our last set of features is based on hand position with respect to the face, which has direct relationships to a person's cognitive state (e.g., covering one's mouth with a hand while talking conveys a sense of uncertainty; supporting one's head using a hand near the ear often accompanies thinking and listening [40]). To distinguish these cases, we first detect whether a person's hand touches her/his face in each frame, and if so further detect two binary cases: whether the hand is supporting the head (near the ear) or not, and whether the hand is touching the chin or not.



(a) Face bounding box detection



(b) Hand on chin



(c) Hand supports head

Figure 7: Hand position detection

The detection is based on the relative positions of the face bounding box and hand bounding boxes.

We construct the face bounding box based on the same OpenFace landmark detection framework used for the VFOA features. As shown in Figure 7(a), the red dots are the 68 facial landmarks, the blue box is the estimated face bounding box, and the gray line represents the bottom line of the nose.

Based on the detected hand bounding box, if the overlap between the face bounding box and a hand bounding box is larger than a certain threshold, the event is classified as *hand on face* ($I_{HF} = 1$). In this case, we further classify the binary events *hand on chin* if the hand bounding box is sufficiently below the nose as shown in Figure 7(b), and *hand supporting head* if the hand bounding box is sufficiently above the bottom line of the face bounding box by a threshold as shown in Figure 7(c). To tune the parameters of these algorithms, including the threshold values, we manually annotated 18000 video frames for these classes, with resulting classification accuracy 88.6% for detecting hand on face, 93.3% for detecting hand on chin, and 95.8% for detecting hand supporting head.

Similar to the previous features, we derive three hand-based features per participant:

- F_{HF}, the fraction of frames in which a participant has hand on face.
- F_{HC}, the fraction of frames in which a participant has hand on chin.

- F_{HSH}, the fraction of frames in which a participant has hand supporting head.

As before, we consider the 21 co-occurrences between the VFOA events and the hand position events, enabling us to extract meaningful multiparty features such as the fraction of the time that a participant looks at their paper while supporting their head, or that they are looked at by others while touching their chin.

5. Correlation and Prediction Analysis

We first analyze the correlation between each feature and output variable to understand the feature importance, and then develop non-linear classification models for automatic prediction of personality/performance traits.

5.1. Feature Importance Understanding

We performed a Pearson correlation analysis using the co-occurrence features, to understand which features are most useful for predicting which personality/leadership/contribution dimensions. Figure 8 summarizes the results. Each subgraph shows a different target variable (not including Openness, for which we found no significant correlations). The value of each marked cell shows the correlation coefficient as a ρ value. The cell is blue for positive correlations and red for negative correlations, and the saturation of color represents the significance value ($p < 0.05$ or $p < 0.10$). Cells without sufficient significance are left blank.

We can see that many of our automatically extracted visual metrics have significant correlations with our target variables. Some of the significant results include:

- Participants high in Extraversion tend to have multiple body parts moving or hands supporting their heads while receiving attention from others, which aligns with the findings in [11] showing that extroverts are more likely to be motivated by social rewards.
- Participants low in Extraversion tend to look at the table more with their hands touching the face, which corresponds to the result in [15] that it is less likely for introverts to be sociable.
- Participants low in Agreeableness tend to have large motion intensity while being looked at and while having mutual gaze with others.
- Participants low in Agreeableness tend to use their hands to touch the chin while looking at others, which is consistent with the findings in [46] that people touching their face show a sense of suspicion about what they are hearing and in [15] that people with low Agreeableness scores are less cooperative.
- Participants low in Conscientiousness tend to have more gaze interactions with others, which agrees with the conclusion that Conscientiousness has negative correlation with extrinsic motivations [21, 37].
- Participants low in Neuroticism tend to look at the table more with few moving body parts.
- Participants with low Leadership scores tend to touch the chin while having gaze interaction with others.
- Participants with high Contribution scores tend to have more gaze interactions with others, and support the head more while having gaze interactions with others, corroborating the findings from [40] that supporting the head conveys a sense of thinking and learning.

5.2. Output Variable Prediction

Since we observed several strong correlations in the previous section, we next investigate how co-occurrence events can be used to predict important social signals.

Following similar analysis in [1, 44], we convert the Big-Five personality scores into binary values by thresholding on the median values among the 48 participants in the dataset. That is, for each personality trait, 50% of the participants are marked as high in that trait (above the median) and 50% of them are marked as low in it. We also identify the participant in each group with the highest leader/contributor score, producing a binary target variable. Since our dataset contains 15 meetings with 48 participants, we use k-fold validation to evaluate our model, in which k=3. Specifically, we randomly split the 15 meetings into 3 sets and run training and testing in each set. During each run, 2 sets are used as training data, the remaining set is used for testing, and the accuracy on the testing set is recorded. The final k-fold validation accuracy is the average value of the accuracy over 3 rounds of testing. Therefore, participants belonging to the same meeting group will never be in the training or testing set at the same time.

To investigate how well the visual action events could both individually and jointly predict each target variable, we applied a decision tree classifier with a bootstrap aggregation strategy to further improve the classification performance; multiple decision trees are learned and the final decision is made from the majority vote of the trees. As shown in the top four rows of Table 1, by including the co-occurrent action events, we achieve a much higher accuracy in predicting the binary target variables compared with individual action events only. These results support our hypotheses **H1** and **H2** about the value of combining multiple modalities and investigating feature co-occurrences.

The bottom section of Table 1 reports several recent results from different papers that use video of each participant to classify personality traits in a group meeting scenario. It is important to note that the results are not directly comparable since the competing methods use different datasets, but we provide them to give a sense of the state of the art. All the methods use the same evaluation methodology.

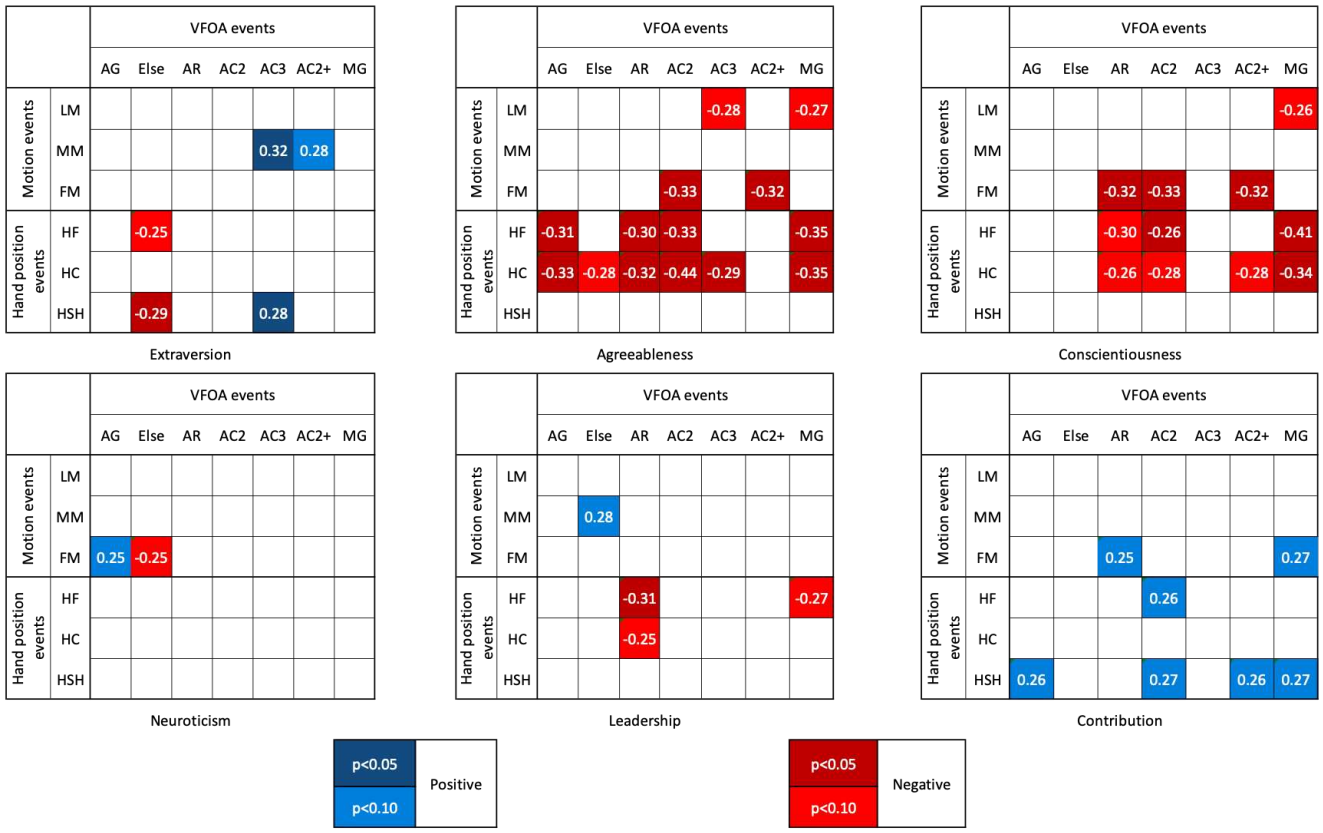


Figure 8: Correlation analysis between visual cues and personality/performance traits.

Table 1: K-fold classification accuracy on our testing set using individual and co-occurrence features. The figures in the bottom rows of the table are not directly comparable with our methods since the underlying datasets are different, but are provided to give a sense of the state of the art.

	Feature set	Features	Extra.	Agree.	Consci.	Neuro.	Open.	Lead.	Contri.
Proposed	Individual	VFOA	64.1%	45.3%	45.8%	64.6%	43.6%	54.2%	56.0%
		Motion	56.0%	68.7%	52.1%	77.1%	43.9%	66.7%	58.6%
		Hand	66.2%	66.7%	54.2%	70.8%	47.5%	52.1%	43.7%
	Co-occurrence	All	**81.0%**	**81.3%**	**81.3%**	**83.3%**	70.8%	**77.1%**	**77.2%**
Others' methods	Okada et al. [44]		69.6%	68.6%	59.8%	56.9%	61.7%	–%	–%
	Fang et al. [25]		77.5%	77.5%	79.4%	79.4%	**73.5%**	–%	–%

6. Conclusions

We proposed a computational framework to effectively predict participants' personality and perceived leadership/contribution traits in a group discussion scenario, using multiparty co-occurrence events. Our group meeting dataset also includes overhead RGB-D videos recorded from two ceiling-mounted Kinect sensors, which were not used in this work. One possible direction for future work is to fit kinematic skeleton models to the 3D data to get accurate estimates of body and arm pose. These could be integrated into our existing framework as another aspect of multimodality. Additionally, meaningful actions including head nodding and eye blinking that are also known to correlate with the target variables could be detected from the multimodal data. Finally, we have not yet integrated the audio from our dataset into our framework, which would enable a new set of non-verbal and verbal metrics, and their co-occurrences with visual metrics.

7. Acknowledgements

This material is based upon work supported by the U.S. National Science Foundation under Grant No. IIP-1631674 and by the U.S. Army Research Lab under Grant No. W911NF-19-2-0135.

References

[1] N. Al Moubayed, Y. Vazquez-Alvarez, A. McKay, and A. Vinciarelli. Face-based automatic personality perception. In *Proceedings of the 22nd ACM International Conference on Multimedia*, pages 1153–1156. ACM, 2014.

[2] F. Alam and G. Riccardi. Predicting personality traits using multimodal information. In *Proceedings of the 2014 ACM Multimedia Workshop on Computational Personality Recognition*, pages 15–18. ACM, 2014.

[3] O. Aran and D. Gatica-Perez. Cross-domain personality prediction: from video blogs to small group meetings. In *Proceedings of the 15th ACM International Conference on Multimodal Interaction*, pages 127–130. ACM, 2013.

[4] O. Aran and D. Gatica-Perez. One of a kind: Inferring personality impressions in meetings. In *Proceedings of the 15th ACM International Conference on Multimodal Interaction*, pages 11–18. ACM, 2013.

[5] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 59–66. IEEE, 2018.

[6] B. M. Bass, C. R. McGehee, W. C. Hawkins, P. C. Young, and A. S. Gebel. Personality variables related to leaderless group discussion behavior. *The Journal of Abnormal and Social Psychology*, 48(1):120, 1953.

[7] G. Beattie. *Rethinking body language: How hand movements reveal hidden thoughts*. Routledge, 2016.

[8] S. E. Bekhouche, F. Dornaika, A. Ouafi, and A. Taleb-Ahmed. Personality traits and job candidate screening via analyzing facial videos. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1660–1663. IEEE, 2017.

[9] C. Beyan, M. Shahid, and V. Murino. Investigation of small group social interactions using deep visual activity-based nonverbal features. In *Proceedings of the 26th ACM International Conference on Multimedia*, pages 311–319. ACM, 2018.

[10] I. Bhattacharya, M. Foley, N. Zhang, T. Zhang, C. Ku, C. Mine, H. Ji, C. Riedl, B. F. Welles, and R. J. Radke. A multimodal-sensor-enabled room for unobtrusive group meeting analysis. In *Proceedings of the 2018 on International Conference on Multimodal Interaction*, pages 347–355. ACM, 2018.

[11] E. Biddle, E. Lameier, L. Reinerman-Jones, G. Matthews, and M. Boyce. Personality: A key to motivating our learners.

[12] J.-I. Biel and D. Gatica-Perez. The youtube lens: Crowdsourced personality impressions and audiovisual analysis of vlogs. *IEEE Transactions on Multimedia*, 15(1):41–55, 2012.

[13] J.-I. Biel, L. Teijeiro-Mosquera, and D. Gatica-Perez. Facetube: predicting personality from facial expressions of emotion in online conversational video. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction*, pages 53–56. ACM, 2012.

[14] G. R. Bradski and J. W. Davis. Motion segmentation and pose recognition with motion history gradients. *Machine Vision and Applications*, 13(3):174–184, 2002.

[15] C. Brinkman. *The Big Five personality model and motivation in sport*. PhD thesis, Miami University, 2013.

[16] O. Celiktutan, P. Bremner, and H. Gunes. Personality classification from robot-mediated communication cues. In *25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2016.

[17] J. W. Davis and A. F. Bobick. The representation and recognition of action using temporal templates. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 928–934, 1997.

[18] M. De Meijer. The contribution of general features of body movement to the attribution of emotions. *Journal of Nonverbal Behavior*, 13(4):247–268, 1989.

[19] A. Dhall and J. Hoey. First impressions-predicting user personality from twitter profile images. In *International Workshop on Human Behavior Understanding*, pages 148–158. Springer, 2016.

[20] S. K. D'mello and A. Graesser. Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features. *User Modeling and User-Adapted Interaction*, 20(2):147–187, 2010.

[21] A. L. Duckworth, C. Peterson, M. D. Matthews, and D. R. Kelly. Grit: perseverance and passion for long-term goals. *Journal of Personality and Social Psychology*, 92(6):1087, 2007.

[22] H. Ehrlichman and D. Micic. Why do people move their eyes when they think? *Current Directions in Psychological Science*, 21(2):96–100, 2012.

[23] P. Ekman and W. V. Friesen. Head and body cues in the judgment of emotion: A reformulation. *Perceptual and Motor Skills*, 24(3 PT 1):711–724, 1967.

[24] H. J. Escalante, I. Guyon, S. Escalera, J. Jacques, M. Madadi, X. Baró, S. Ayache, E. Viegas, Y. Güçlütürk, U. Güçlü, et al. Design of an explainable machine learning challenge for video interviews. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 3688–3695. IEEE, 2017.

[25] S. Fang, C. Achard, and S. Dubuisson. Personality classification and behaviour interpretation: An approach based on feature categories. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 225–232. ACM, 2016.

[26] G. Farnadi, S. Sushmita, G. Sitaraman, N. Ton, M. De Cock, and S. Davalos. A multivariate regression approach to personality impression recognition of vloggers. In *Proceedings of the 2014 ACM Multimedia Workshop on Computational Personality Recognition*, pages 1–6. ACM, 2014.

[27] J. M. George. Personality, affect, and behavior in groups. *Journal of Applied Psychology*, 75(2):107, 1990.

[28] Y. Güçlütürk, U. Güçlü, M. A. van Gerven, and R. van Lier. Deep impression: Audiovisual deep residual networks for multimodal apparent personality trait recognition. In *European Conference on Computer Vision*, pages 349–358. Springer, 2016.

[29] S. C. Guntuku, L. Qiu, S. Roy, W. Lin, and V. Jakhetiya. Do others perceive you as you want them to?: Modeling personality based on selfies. In *Proceedings of the 1st International*

Workshop on Affect & Sentiment in Multimedia, pages 21–26. ACM, 2015.

[30] J. Hall. *NASA Moon Survival Task-: The Original Consensus Exercise*. Teleometrics International, 1994.

[31] J. Hogan and B. Holland. Using theory to evaluate personality and job-performance relations: A socioanalytic perspective. *Journal of Applied Psychology*, 88(1):100, 2003.

[32] R. Hogan, J. Hogan, and B. W. Roberts. Personality measurement and employment decisions: Questions and answers. *American Psychologist*, 51(5):469, 1996.

[33] H. Hung, Y. Huang, G. Friedland, and D. Gatica-Perez. Estimating dominance in multi-party meetings using speaker diarization. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):847–860, 2010.

[34] D. B. Jayagopi, H. Hung, C. Yeo, and D. Gatica-Perez. Modeling dominance in group conversations using nonverbal activity cues. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(3):501–513, 2009.

[35] J. Joo, F. F. Steen, and S.-C. Zhu. Automated facial trait judgment and election outcome prediction: Social dimensions of face. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3712–3720, 2015.

[36] T. A. Judge, D. Heller, and M. K. Mount. Five-factor model of personality and job satisfaction: A meta-analysis. *Journal of Applied Psychology*, 87(3):530, 2002.

[37] M. Komarraju, S. J. Karau, and R. R. Schmeck. Role of the big five personality traits in predicting college students' academic motivation and achievement. *Learning and Individual Differences*, 19(1):47–52, 2009.

[38] R. M. Krauss and C.-Y. Chiu. Language and social behavior. *Handbook of social psychology*, 1998.

[39] J. S. Lerner, Y. Li, P. Valdesolo, and K. S. Kassam. Emotion and decision making. *Annual Review of Psychology*, 66:799–823, 2015.

[40] M. Mahmoud and P. Robinson. Interpreting hand-over-face gestures. In *International Conference on Affective Computing and Intelligent Interaction*, pages 248–255. Springer, 2011.

[41] K. Matsumoto, S. Shibata, S. Seiji, C. Mori, and K. Shioe. Factors influencing the processing of visual information from non-verbal communications. *Psychiatry and Clinical Neurosciences*, 64(3):299–308, 2010.

[42] R. R. McCrae and O. P. John. An introduction to the five-factor model and its applications. *Journal of Personality*, 60(2):175–215, 1992.

[43] V. Meunier. A python program to detect and classify hand pose using deep learning techniques. https://github.com/MrEliptik/HandPose, 2019.

[44] S. Okada, O. Aran, and D. Gatica-Perez. Personality trait classification via co-occurrent multiparty multimodal event discovery. In *Proceedings of the 2015 ACM International Conference on Multimodal Interaction*, pages 15–22. ACM, 2015.

[45] A. Pease. *Body language: How to read others' thoughts by their gestures*. Sheldon Press, 1984.

[46] B. Pease and A. Pease. *The definitive book of body language: The hidden meaning behind people's gestures and expressions*. Bantam, 2008.

[47] I. Poggi and C. Pelachaud. Emotional meaning and expression in animated faces. In *International Workshop on Affective Interactions*, pages 182–195. Springer, 1999.

[48] V. Ponce-López, B. Chen, M. Oliu, C. Corneanu, A. Clapés, I. Guyon, X. Baró, H. J. Escalante, and S. Escalera. Chalearn lap 2016: First round challenge on first impressions-dataset and results. In *European Conference on Computer Vision*, pages 400–418. Springer, 2016.

[49] B. Rammstedt and O. P. John. Measuring personality in one minute or less: A 10-item short version of the big five inventory in english and german. *Journal of Research in Personality*, 41(1):203–212, 2007.

[50] S. Rothmann and E. P. Coetzer. The big five personality dimensions and job performance. *SA Journal of Industrial Psychology*, 29(1):68–74, 2003.

[51] D. Sanchez-Cortes, O. Aran, and D. Gatica-Perez. An audio visual corpus for emergent leader analysis. In *Workshop on Multimodal Corpora for Machine Learning: Taking Stock and Road mapping the Future, ICMI-MLMI*, 2011.

[52] D. Sanchez-Cortes, O. Aran, M. S. Mast, and D. Gatica-Perez. A nonverbal behavior approach to identify emergent leaders in small groups. *IEEE Transactions on Multimedia*, 14(3):816–832, 2011.

[53] C. Sarkar, S. Bhatia, A. Agarwal, and J. Li. Feature analysis for computational personality recognition using youtube personality data set. In *Proceedings of the 2014 ACM Multimedia workshop on Computational Personality Recognition*, pages 11–14. ACM, 2014.

[54] A. Subramaniam, V. Patel, A. Mishra, P. Balasubramanian, and A. Mittal. Bi-modal first impressions recognition using temporally ordered deep audio and stochastic visual features. In *European Conference on Computer Vision*, pages 337–348. Springer, 2016.

[55] R. P. Tett and D. D. Burnett. A personality trait-based interactionist model of job performance. *Journal of Applied Psychology*, 88(3):500, 2003.

[56] D. A. Waldman, L. E. Atwater, and R. A. Davidson. The role of individualism and the five-factor model in the prediction of performance in a leaderless group discussion. *Journal of Personality*, 72(1):1–28, 2004.

[57] A. S. Wicaksana and C. C. Liem. Human-explainable features for job candidate screening prediction. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1664–1669. IEEE, 2017.

[58] Y. Yan, J. Nie, L. Huang, Z. Li, Q. Cao, and Z. Wei. Exploring relationship between face and trustworthy impression using mid-level facial features. In *International Conference on Multimedia Modeling*, pages 540–549. Springer, 2016.

[59] C.-L. Zhang, H. Zhang, X.-S. Wei, and J. Wu. Deep bimodal regression for apparent personality analysis. In *European Conference on Computer Vision*, pages 311–324. Springer, 2016.

[60] L. Zhang, M. Morgan, I. Bhattacharya, M. Foley, J. Braasch, C. Riedl, B. Foucault Welles, and R. J. Radke. Improved visual focus of attention estimation and prosodic features for analyzing group interactions. In *2019 International Conference on Multimodal Interaction*, ICMI 19, page 385394, New York, NY, USA, 2019.