Personality Trait Classification via Co-Occurrent Multiparty Multimodal Event Discovery

Shogo Okada [1]
okada@dis.titech.ac.jp

Oya Aran [2]
oaran@idiap.ch

Daniel Gatica-Perez [2,3]
gatica@idiap.ch

[1] Tokyo Institute of Technology, Yokohama, Japan
[2] Idiap Research Institute, Martigny, Switzerland
[3] Ecole Polytechnique Federale de Lausanne (EPFL), Lausanne, Switzerland

ABSTRACT

This paper proposes a novel feature extraction framework from mutli-party multimodal conversation for inference of personality traits and emergent leadership. The proposed framework represents multi modal features as the combination of each participant's nonverbal activity and group activity. This feature representation enables to compare the nonverbal patterns extracted from the participants of different groups in a metric space. It captures how the target member outputs nonverbal behavior observed in a group (e.g. the member speaks while all members move their body), and can be available for any kind of multiparty conversation task. Frequent co-occurrent events are discovered using graph clustering from multimodal sequences. The proposed framework is applied for the ELEA corpus which is an audio visual dataset collected from group meetings. We evaluate the framework for binary classification task of 10 personality traits. Experimental results show that the model trained with co-occurrence features obtained higher accuracy than previously related work in 8 out of 10 traits. In addition, the co-occurrence features improve the accuracy from 2% up to 17%.

Categories and Subject Descriptors

I.5.2 [**Pattern Recognition**]: Design Methodology—*Feature evaluation and selection*; H.2.8 [**Database Applications**]: Data mining

General Terms

Algorithms, Experimentation

Keywords

Personality trait, Inference, Data mining, Multiparty interaction

1. INTRODUCTION

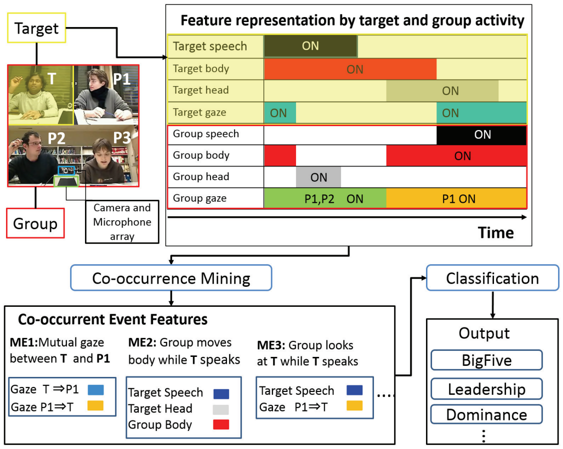


Figure 1: Overview of proposed framework

Automatic nonverbal analysis of small group interaction promises many kinds of applications. In recent years, one challenge in this research is to infer high level characteristics of participants as target variables, such as roles, attitude in conversation, emerging leadership, and personality traits, by combining audio and visual information observed from people. In these works, data-driven approaches are often used to model relationships between nonverbal behavior observed from start to end of an interaction and target variables such as personality traits, leadership, etc. A key factor for success is to extract various types of nonverbal features that have possibility to infer the target variable. To extract effective features for inference of these target variables, previous works have proposed static features from audio and visual data defined based on knowledge of social science.

Common aspects in these features are that these statistics are calculated by accumulating each event observed in the whole meeting. On the other hand, conversational nonverbal patterns exist at multiple time scales [7], ranging from fine-grain features such as presence of speech and head gesture patterns, to conversational context patterns where multiple events co-occur simultaneously, such as events in which listeners nod when a speaker explains something with gestures. Using co-occurrence patterns between modalities yields two main advantages for modeling personality traits. First, it is possible to improve the inference accuracy of the trait value based on rich feature set extracted by capturing the interaction between modalities. Second, discovering key context patterns linking personality traits helps us understand effective conversational contexts to predict the trait variables. For example, when co-occurrence events (gaze and utterance) are observed in a group conversation, emergent leaders likely speak while gazing to people [20].

In this paper, we propose a co-occurrent event mining framework from multiparty and multimodal interaction data to infer personality traits. Figure 1 shows the overview of the proposed framework. The goal is to find patterns between modalities and multiple people: segments of utterance, speech, gaze, head gestures and body gestures. To apply co-occurrence mining for group conversations,

ICMI 2015, November 9–13, 2015, Seattle, WA, USA.
© 2015 ACM. ISBN 978-1-4503-3912-4/15/11 ...$15.00.
DOI: http://dx.doi.org/10.1145/2818346.2820757.

we need to tackle the problem of how to compare the data collected from groups which are composed of different members in one metric space. Here, we separate the multimodal signals from each participant into signals observed from (1) a group member and (2) the other group members. This data representation captures what an individual member does (speak, look, and move) when the other members do certain actions. The data representation enables to compare multimodal activities of individual members and captures the context in which the individual members acts in group interaction each time. After data representation, multiparty multimodal data is converted to 2-D time-series dataset. A pairwise event discovery algorithm based on graph mining is applied to find co-occurrent events and discovered patterns are used as features for the inference.

In this research, we use the ELEA (Emerging LEadership Analysis) corpus including 27 group interactions composed of 3 or 4 members. This dataset includes audio and visual data and personality traits scored from group members and external observers, such as BigFive personality impressions, and perceived leadership [20]. The Big Five model has been proposed in psychology as one capable of capturing the construction of personality [11]. In experiments, we perform binary trait level classification to evaluate our approach with previous work. Experimental results show that co-occurrent event features improve the accuracy in 8 out of 10 traits.

There are two main contributions in our paper. First, we perform co-occurrence mining in multimodal and multiparty interaction for inference of personality trait, which is an unexplored problem setting. Second, we show that using conversational context features, extracted based on co-occurrence of patterns improves the classification accuracy of several personality traits.

We present related work in Section 2. In Section 3, we present data used for inference of personality traits. Section 4 explains the method for feature representation. Section 5 explains the mining framework. Section 6 and section 7 present the experimental setting and the evaluation of our framework, respectively. We discuss the results in Section 8 and conclude this study in Section 9.

2. RELATED WORK

Our research is related to the topics of personality trait modeling, interaction mining, and multimodal recognition using contextual information.

2.1 Personality trait inference in group

There is an increasing interest towards inferring higher level concepts such as individual traits, leadership from low level, multimodal signals observed in multiparty interaction. We focus on personality trait modeling in group meetings and in this section we present related research emphasizing group interaction, without including works modeling a single person (e.g.[5]) and works modeling these concepts from dyadic interaction (e.g.[15]).

In multiparty interaction, different works looked at different variables including social roles and, personality traits. [23] proposed an approach to detect functional roles played by participants in group conversation. [22] proposed approaches for speaker role recognition in group conversation. [18] presents an analysis on personality prediction of each participant using self-reported questionnaires. [20] presents an analysis on emergent leadership in meetings. [2] presents an analysis on prediction of participant's personality trait impressions formed by external observers.

In this research, feature extraction from audio and visual data corpus is a key technique. [20] points that when the count of speech turns and the speech length of a member are larger than that of the other members, the member is likely perceived as a leader from the group members. [2] reports that the energy of the speech signal is effective to predict openness to experience, which is a BigFive personality trait. From visual data, body gestures, head gestures, and gaze are also known as important features. In particular, the amount and frequency of body motion are captured as Motion Energy Images (MEI) calculated using difference of images in time. Head gesture and gaze are approximated using head posture and motion. [10] points to visual focus of attention as a key feature to link impressions from group members. [17] also points to gaze states as effective feature to detects important statements.

In a common approach of these works, audio and visual features are calculated thorough mean, medium, min, max and X percentile from various amounts of statistics (count and length) from each pattern observed for the whole meeting or for a part of the meeting [2], [20], [17]. These feature sets are used as input data to train inference models. Though this approach often can fuse total statistics of patterns observed within a duration time, it can not capture co-occurence between multi modal patterns at each time. For example, extracting co-occurrent events between an utterance and a body motion pattern as a feature is useful to model personality traits if the utterance accompanying body gesture will make a stronger impression to the listener than utterance without the gesture. Kendon points out that it is important to analyze co-occurring multi modal patterns to understand multimodal conversation phenomenon [12]. Our objective of this study is to improve the inference accuracy of personality traits by extracting co-occurence features, which are not explored in many previous work. We explicitly discover frequent co-occurrent events from 49-dimensional time-series data observed in multimodal multiparty interaction data by using a graph clustering algorithm.

2.2 Unsupervised interaction analysis

Supervised learning is used in existing works to infer personality traits. On the other hand, unsupervised learning approaches can often find intermediate representations that help understand these higher concepts [10].

The work in [9],[10] use Latent Dirichlet Allocation (LDA) to mine context features in groups. In [10], group features called group looking (or speaking) cues are defined manually for input to LDA. Context features are extracted as topics (clusters) generated by LDA. In contrast, co-occurrence clustering as we propose here is done to discover combinations of each modal pattern, and group features are extracted automatically. The work in [4] models influence from a member to the other members by relating interactions of nonverbal pattern between group members to transition between hidden states (e.g. whose utterance starts after whose utterance) in a Markovian formulation.

In [10],[4], feature extraction is done per group to analyze these group nonverbal patterns and group performance or group composition. On the other hand, we need to compare between individuals belonging to different groups to model personality traits. We thus propose a novel data representation method to apply to a mining framework by separating nonverbal patterns of a member and those of the other members.

In [9], co-occurrence statistics are calculated in each time slice (from 30 seconds to 5 minutes) from visual focus of attention and speech state. However, exact co-occurrence patterns such as "x th utterance is overlapped with y th body movement pattern" are not extracted. In our study we discover such multimodal patterns in finer time scale to apply to the inference of personality traits. Mining a large and more diverse set of co-occurrent events is important to discover effective features however this has been unexplored in previous work.

A data mining framework has also been applied for other kind of multimodal datasets. The work in [13] applied frequent sequence mining as a feature extraction method in predicting user states (anxious, challenging, exciting, and so on) while playing games by using physiological signals, game play information, and user keystrokes. The work in [14] enhanced this framework as an unsupervised feature learning framework using deep learning. In contrast to user's playing games, we tackle in this research multimodal multiparty interaction. In general, the phenomena observed from multiparty interaction is more complex than that observed from a single participant.

3. MULTIMODAL DATA CORPUS

3.1 Dataset and tasks

We used a subset from the Emergent LEAder (ELEA) corpus [20] for this study. The subset consists of audio-visual (AV) recordings of 27 meetings, in which the participants perform a winter survival task with no roles assigned.

The participants in the task, as the survivors of an airplane crash, were asked to rank 12 items to take with them to survive as a group. Participants first ranked the items individually, then as a group. The task itself promotes interactions among the participants.

Participants discuss while being seated around a table. For sensing infrastructure, Dev-Audio Microcone, a commercial portable microphone array (green square in the left picture on Figure 1) is used to collect the audio. Two wide-angle web cameras (right blue square on Figure 1) are used for the video setup.

While the corpus was originally designed to study leadership, the personality of each individual can be made evident through the discussion and negotiation parts of the interaction. There are 102 participants in total (six meetings with three participants and 21 meetings with four participants). Each meeting lasts around 15 minutes. The synchronization of audio and video was done manually by aligning the streams using the clapping activity. More details about the ELEA AV corpus can be found in [19]. Figure 1 shows a snapshot from the data.

3.2 Personality trait impression annotations

The ELEA corpus includes personality annotation data as a result of scoring each participant via questionnaires.

Big-Five Trait Impressions from external observers: Personality impressions of the participants by the external observers were collected in [2]. These annotations include scores for the BigFive: Extraversion (Ext), Agreeableness (Agr), Conscientiousness (Con), Emotional stability (Emo), and Openness to Experience (Ope). More details about modeling can be found in [11].

The Ten Item Personality Inventory (TIPI) was used to measure the BigFive personality traits of the participants [8]. It includes two questions per trait, answered on a 7-point Likert scale. The questionnaire was done for each participant, on a one-minute segment from the meeting, which corresponds to the segment that includes the participant's longest speaking turn, and isolating the video of each participant (Only a single participant is visible). The audio on the other hand is intact and contains speech from all participants in that segment. More details can be found in [2].

Dominance and Leadership Impressions from group members The ELEA corpus also includes scores for traits relevant to the functioning of individuals with respect to dominance and leadership. After the meeting task, participants filled out a Perceived Interaction Score, that captures perceptions from participants during the interaction, in which they score every participant in the group through four items related to the following concepts: perceived leadership (PLead), perceived dominance (PDom), perceived competence (PCom) and perceived liking (PLike). Afterwards a dominance ranking (RDom) is calculated. PLead captures whether the person directs the group, imposes his or her opinion. PDom captures whether the person dominates, or is in a position of power.

Participants were asked to rank the group, giving 1 to the most dominant participant, and 3 or 4 for the less dominant, such that they have to include themselves in the ranking. More details can be found in [20].

4. MULTIMODAL FEATURE REPRESENTATION

We propose a feature representation to compare nonverbal patterns that are observed from each participant in a group. In the feature representation, we define two kinds of features as

$$\begin{aligned} NF &= \{NF_m, NF_{/m}\}, \\ NF_m &= \{nf_{m,1}, \ldots, nf_{m,t}, \ldots, nf_{m,T}\}, \end{aligned} \tag{1}$$

where NF_m, $NF_{/m}$ is time-series binary data composed of $nf_{m,t}$ and $nf_{/m,t}$, respectively, $nf_{m,t}$ is tth segment observed from a particular member m in a group and $nf_{/m,t}$ is tth segment observed from the group members except the member m. The tth event has a time length and corresponds a segment of "ON" in Figure 1. We defined an event as a segment where the feature is active. T_{ns} is the number of nonverbal patterns NF observed in the whole meeting.

On one hand, $nf_{m,t}$ represents nonverbal features extracted from mutilmodal nonverbal signals observed from an individual participant such as speech utterances, body and head motion and gaze patterns. On the other hand, $nf_{/m,t}$ represents group nonverbal features extracted from signals which are observed from all group members except the participant m. The representation captures how the participant acts when the other members execute any nonverbal activity as a joint probability of each activity by simultaneously observing nonverbal activities both of the individual participant and the other group members.

We define co-occurrence patterns as multimodal events overlapped in time. Our approach requires transforming the continuous signals (pitch, energy, and image template features) into sequences of events. In section 5, we present our data mining algorithm to find frequent co-occurrence events. In this section, we present the input audio and visual signals. The features are summarized in Table 2. We refer the reader to related papers for detailed information on audio-visual feature extraction.

4.1 Audio features

4.1.1 *Speaking turn features*

Binary segmentation is performed to capture the speaking status ST of each participant. This binary segmentation is provided by the microphone array that is used for the audio recordings, which is used for speaker diarization [20]. We define a set of segments where the speech status is on, as speaking turn set ST. The speaking turn set of each participant m in the group is denoted ST_m.

We define three type of features, $SO1$, $SO2$, $Ssil$, as group speaking turn features $ST_{/m}$. $SO1$ is a set of segments where the speech state of one group member (except m) is on. $SO2$ is a set of segments where the speech states of more than two members except m are on. $Ssil$ is a set of segments where the speech states of all members are off.

4.1.2 *Prosodic features*

Prosodic features are extracted for each individual member. Based on the binary speaker segmentation, we obtain the speech signal for

Figure 2: Multimodal feature set

ID	Features	Symbol	Description
F_1	Speaking Status (ST)	ST	Speech segments of the target person
		$SO1$	One person other than target speaks
		$SO2$	More than two people speak.
		$Ssil$	Silent segment
F_2	Pitch (PI)	PUp, PDo	Sign of difference between utterance t and utterance $t-1$
		PCL, PCM, PCH	Cluster index (low medium and high level) after clustering
		$PCNL$, $PCNM$, $PCNH$	Cluster index after clustering of normalized value
F_3	Energy (EN)	EUp, EDo	Sign of difference between utterance t and utterance
		ECL, ECM, ECH	Cluster index after clustering
		$ECNL$, $ECNM$, $ECNH$	Cluster index after clustering of normalized value
F_4	Head Motion (H)	HMT	Motion segments of target person
		$HMO1$	One person other than target moves
		$HMO2$	More than two people move
		$HMsil$	Still motion segment
F_5	Body Motion (B)	BMT	Motion segments of target person
		$BMO1$	One person other than target moves
		$BMO2$	More than two people move
		$BMsil$	Still motion segment
F_6	MEI (MT)	MUp, MDo	Sign of difference between segments
		MCL, MCM, MCH	Cluster index after clustering
		$MCNL$, $MCNM$, $MCNH$	Cluster index after clustering of normalized MEI
F_7	Gaze (G)	GT	Target person looks at person
		$GTSp$	Target person looks at speaker
		$GOT1$	One person looks at the target
		$GOT2$	More than two people look at the target
		MGT	Mutual gaze between target and another person
		MGO	Mutual gaze between two people other than target

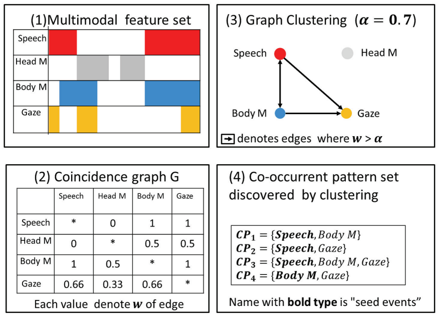

(1)Multimodal feature set
Speech
Head M
Body M
Gaze
(3) Graph Clustering $(\alpha = 0.7)$
Speech
Head M
Body M
Gaze
denotes edges where $w > \alpha$
(2) Coincidence graph G
Speech | Head M | Body M | Gaze
Speech * 0 1 1
Head M 0 * 0.5 0.5
Body M 1 0.5 * 1
Gaze 0.66 0.33 0.66 *
Each value denote w of edge
(4) Co-occurrent pattern set discovered by clustering
$CP_1 = \{\mathbf{Speech}, Body\ M\}$
$CP_2 = \{\mathbf{Speech}, Gaze\}$
$CP_3 = \{\mathbf{Speech}, Body\ M, Gaze\}$
$CP_4 = \{\mathbf{Body\ M}, Gaze\}$
Name with **bold type** is "seed events"

Figure 3: Example of multimodal pattern mining (A case of 4 dimentional patterns)

each participant. Overlapping speech segments are discarded, and only the segments with the participant being the sole speaker are considered for further processing. Two prosodic speech features, energy and pitch, are computed on the signal.

We convert energy and pitch into a sequence of significant increments and decrements of the signal. We calculate the sign of the difference between statistics of utterance j and utterance $j+1$. It is assumed that prosodic features change due to various reasons. For example, when a participant is likely excited, the energy of his/her utterance may increase after hearing an utterance of other participants. Based on this, we extract the sign of the difference for a prosodic signal as a feature.

Pitch samples pi_j are extracted from utterance j. We perform a statistical t-test between pi_j and pi_{j+1}. If significance exists and the mean of pi_{j+1} is larger than the mean of pi_j, we add utterance $j+1$ to a set PUp of utterance segments whose pitch of the current utterance is larger than that of the past utterance. We do something similar for significant decreasing differences generating a set PDo of utterance segments whose pitch of current utterance is small than that of past utterance. A t-test is done between energy samples en_j, en_{j+1} in the same manner. We extract an utterance segment set EUp where energy is increased more than past utterance, and an utterance segment set EDo where energy is decreased more than past utterance.

Next, we perform clustering to convert energy and pitch signals into categorical data. Clustering is done using utterances ST_m of all participants. The procedure of clustering is as follows.

1. Calculate statistics (max, min, average) of all prosodic values in each utterance as input samples for clustering.

2. K-means clustering is done using the data samples calculated from all participants.

3. Each utterance is converted to cluster index.

We set the number of clusters as 3, corresponding to low level (CL), medium level (CM) and high level (CH) cluster. Utterance segments clustered by pitch value are added into feature sets PCL, PCM, PCH. Utterance segments clustered by energy value are added into feature sets ECL, ECM, ECH. We also normalize the prosodic value within segments for each participant considering individual differences, and perform clustering in the same manner.

Utterance segments clustered by normalized pitch value are added into feature sets $PNCL$, $PNCM$, $PNCH$. Utterance segments clustered by normalized energy value are added into feature sets $ENCL$, $ENCM$, $ENCH$. This feature set also captures the variance of each participant. We summarize the feature set for pitch and as
$PI = \{PUp, PDo, PCL, PCM, PCH, PNCL, PNCM, PNCH\}$,
$EN = \{EUp, EDo, ECL, ECM, ECH, ENCL, ENCM, ENCH\}$.
These feature sets, PI and EN, are calculated for the individual participants.

4.2 Visual features

4.2.1 *Visual activity features*

Visual activity features characterize the bodily activity of the participant. We used two different approaches to extract activity features. The first approach is based on head and body tracking and optical flow, which provides the binary head and body activity status and the amount of activity as well. As done for speech states, binary segmentation is done and an activity state set is extracted for head and body motion.

We define a set of segments where the body or head status is on, as the body activity set B and the head activity set H. We also define three types of features: $BO1$, $BO2$, $Bsil$, as group body activity features $B_{/m}$, and three types of features $HO1$, $HO2$, $Hsil$, as group head activity features $H_{/m}$ in the same manner for ST.

4.2.2 *Motion template based features*

As a second approach, we have used Motion Energy Images (MEI) [6] as descriptors of body activity. We used the length of the meeting segment to normalize the images. Motion Energy Images (MEI) are obtained by integrating each difference image from whole video clip. Significant changes of MEI have the possibility to capture behaviors related to personality traits.

Here, we calculate time-series changes of MEI with a sliding window method over the whole meeting. Time-series data of MEI is continuous and we have to segment to categorical patterns. We follow this procedure. 1: Difference images are calculated in between each frame. 2: Calculate motion energy image in windows of 1sec duration. 3: Calculate time-series MEI features with sliding windows. 4: Detect peaks of the time-series after smoothing. 5: Extract segments of MEI as intervals between peaks. 6: Follow the same procedure as that of prosody features to extract MEI-related features.

Motion template feature set MT is defined from segments of MEI. We extract a segment set MUp, where current MEI energy segment has increased compared to the past segment, and a segment set MDo, where MEI energy has decreased compared to the past segment. MEI segments after clustering are added into feature sets MCL, MCM, MCH. Furthermore, after normalization with respect to each participant, MEI energy segments after clustering are added into feature sets $MNCL$, $MNCM$, $MNCH$. We summarize the feature set for motion template as
$MT = \{MUp, MDo, MCL, MCM, MCH, MNCL, MNCM, MNCH\}$,
Feature set MT is calculated for the individual participants.

4.2.3 Visual focus of attention features

Visual focus of attention (VFOA) features were extracted and shared by the authors in [19], where a probabilistic framework was used to estimate the head location and pose jointly based on a state space formulation. We define a set of segments GT where the target participant looks at the other participants through the meeting. We also define a set of segments $GTSp$ where the target participant looks at the speaker. Looking at speaker is an important signal of the listener's interest and politeness [21]. We further define two features $GOT1, GOT2$ as group attention features $G_{/m}$. $GOT1$ is a set of segments where one member looks at the member m. $GOT2$ is a set of segments where more than two members looks at the member m.

Mutual gazing is also an important feature [10]. Therefore, we explicitly define the segment set for mutual gazing (although mutual gazing is defined as co-occurrence pattern with GT and $GOT1,2$). We prepare two group features for mutual gazing. MGT is a set of segments where one member x looks at the member m and vice versa. MGO is a set of segments where two members y,z except the member m look at each other.

5. CO-OCCURRENT MULTIMODAL PATTERN MINING

In this section, we present our mining algorithm to discover co-occurrent patterns between modalities. The goal of this algorithm is to find frequent co-occurring segments from the feature sets in section 4. Co-occurrent pattern discovery is done using a graph clustering algorithm. These multimodal patterns sometimes capture the characteristics of conversation scenes more than single modal patterns. We adopt the star algorithm proposed in [3] to efficiently discover time-series co-occurring patterns from continuous time-series data for our purpose. Figure 3 shows an example of the mining algorithm. The procedure is as follows.

Input: Multimodal feature sets: ST, PI, EN, H, B, MT, G ((1) in Figure 3)
Output: Co-occurrent pattern set: CP

Step1: Construct a coincidence graph G. G is a directed graph $G = (V,E)$, Vertex V contains the features, E are the weight values calculated based on coincidence frequency between segments nf. Co-occurent pattern set CP is initialized as an empty set $\emptyset$ and number n of patterns in the set is 0.

Step2: A feature NF_k (Equation 1) in the feature set is represented by a vertex v_k. The weight value $w_{k,l}$ is represented by an edge $e_{k,l}$ (edge connecting vertex k to vertex l) is calculated as $w_{k,l} = overlap(NF_k, NF_l)/N_k$, where $overlap(NF_k, NF_l)$ is the total number of count that there is a temporal overlap between occurrences of NF_k and NF_l ((2) in Figure 3). Here, NF_k is different than NF_l (e.g. $\{NF_k, NF_l\}$ correspond to $\{ST, GT\}$, $\{SO1, HMT\}$, etc.).

For example, $overlap(ST, BT) = 80$, where 80 of total events (nf in Equation 1) in the speaking turns of the target person ST are overlapped temporally with the body motion segments BT. N_k is the number of total segments in NF_k.

Step3: Features NF_k and NF_l is grouped if weight value is more than threshold α ((3) in Figure 3). A set $\{NF_k, NF_l\}$ is extracted as a Multimodal co-occurent pattern CP_{n+1} and n is updated to $n+1$. Here NF_k is defined as a seed modality of the pattern CP_n. This step is done for all pairs of NF_k and NF_l.

Step4: Patterns in CP, which are newly discovered in Step3 are added as new features to Vertex V ((4) in Figure 3).

Step5: Coincidence graph is updated by re-calculating weight between CP_n and all of NF.

Step6: Step3 - Step5 are iterated until all the weights in E is less than α.

α ($0 < \alpha < 1$) is used to determine the minimum between two segments. To find many kind of features for inference, we detect various sub-graphs (e.g. Co-occurrence of 2 pairs $\{ST, GT\}$, 3 pairs $\{ST, GT, BO1\}$, etc) from the coincidence graph in Step4. If we set α as a small value, a large amount of patterns are discovered. In this study, The alpha is set manually as $\alpha = 0.8$ by considering balance of number of mined features and number of training samples.

After the mining process, total count that CP_n is observed in the whole meeting is used as a feature value for inference of personality traits. The value of feature CP_n for a particular member m in a group is calculated as the total count that CP_n is observed in whole meeting in the group (in feature sets $\{NF_m, NF_{/m}\}$).

6. EXPERIMENTAL SETTING

To evaluate the effectiveness of the proposed co-occurrence features, we compare the inference accuracy by a model trained using co-occurrence features with the accuracy by the same model trained using multimodal features used in previous work [2]. The personality traits include 10 variables as described in Section 3: personality trait impressions including 5 variables: Extraversion, Agreeableness, Conscientiousness, Emotional Stableness, Opennes to Experience and perceived impressions including 5 variables: Leadership, two versions of Dominance (ranked and scored), Competence, Likeness.

6.1 Setting of Co-occurrence features

Each feature corresponds to the count (frequency) of each mined event and Counts of CP are concatenated into one vector. The total number of dimensions is equal to the total number of mined co-occurrence events. As our dataset is rather small (102 samples) compared to the total number of mined events, we reduce the abundant mined features by the different heuristic methods.

First, we calculate cosine similarity (maximum 1 and minimum 0) between vectors of the feature value in all data samples. We merged the pairs where the similarity is more than 0.99. Second, we count the number of co-occurring events per modality, finding that this number for pitch, energy and MEI is huge because these features are likely to overlap with other patterns. We perform PCA for co-occurring features with pitch, energy and MEI. We use the cumulative energy ratio (which is the value obtained by dividing the sum of j eigenvalues by the sum of all eigenvalues) to decide the number of features, setting the cumulative energy ratio to 0.90. Totally, the number of features is reduced to 188 from 184143.

6.2 Setting of baseline feature set

A multimodal feature set is proposed in a previous research [2], which has been used for inference of personality traits by accumulating the statistics over whole meetings. This feature set has 37 dimensions, and is composed of speaking turn, energy, pitch, visual activity, MEI, and gaze features. The features are extracted from both target and others for gaze, and from only target for other modalities. These features are well engineered because these feature set includes various kind of statistics. For these features, the min, max, mean, medium, standard deviation, quantile value, count (e.g. of speaking turns) and total time length (e.g. speaking length) were calculated. The original feature set used in [2] was kindly shared for comparison purposes.

6.3 Inference task and Classification model

In [2], regression and classification tasks are tested to infer personality traits from nonverbal features. The authors of [2] reported that it is difficult to regress personality traits except the extraversion trait, because R^2 values from regression results via leave-one-out validation, are less than 0.1 for each trait. Based on these results we decide to perform classification of binary levels of personality traits. In the classification task, trait values are converted to binary values (high or low) by thresholding using median value, e.g. to represent people scoring high/low in extraversion, respectively. The trained model is evaluated by classification accuracy of test data.

To precisely compare the results obtained in [2], we followed the evaluation procedure and used ridge regression model and linear SVM as classification models in same manner with experiment in [2]. For training a ridge regression classifier, the original personality impression scores is used and the median score is used as the threshold for prediction (this method is called R_{SCR} in [2]).

In the experiments below, we use leave-one-out cross validation and report the average accuracy over all folds. We normalize the data such that each feature has zero mean and one standard deviation. The ridge parameter in ridge regression model is optimized using a cross validation scheme, with values in the range of [2, 150]. The parameters of SVM are optimized similarly with a nested cross validation scheme, with C parameter values selected from [0, 0.01, 0.1, 1].

7. EXPERIMENTAL RESULTS

Table 1 shows the classification accuracy, and the bold values indicate the highest accuracy. Only accuracies above 62.7% are considered significantly better (with 99% confidence level) than the 50% random assignment baseline.

First to fifth rows in Table 1 shows the classification results for BigFive impressions from external observers, and sixth to tenth rows in Table 1 shows the personality traits perceived from the group members. Each column corresponds to a personality trait and rows correspond to feature sets and classification methods. In each feature set, Co-occurrence feature denotes our proposal, Baseline denotes the feature set defined in [2].

For extraversion and agreebleness, we obtain best accuracies as high as 67.6% using SVM, 68.6% using Ridge Regression, with proposed features, respectively. In contrast, [2] reports that the results for BigFive are significantly different than the random baseline for only extraversion with all features. In particular, the model trained with co-occurrence features improve the Agreeableness accuracy in about 10% compared to the feature set proposed in [2]. In addition, though the results for conscientiousness and emotion stability are not significantly different than the random baseline, the use of both features ((3) Fusing) improves the accuracy about 3% and 2% respectively more than that with the feature set proposed in [2]. For Opennes to Experience, we obtained the best accuracy as high as 61.7% using the model trained with the feature set of [2].

For both perceived competence and ranked dominance, we obtain better than the feature set defined in [2] as high as 66.6% and 64.7% using Ridge regression with proposed features. On the other hand, for perceived dominance and perceived likeness, the result of model with feature set in [2] is better. For perceived leadership, the results with feature set in [2] and our feature set are equal (72.5%).

Furthermore, the better classification model varies for each trait. Ridge regression is effective for agreeableness, emotion stability, perceived dominance, perceived competence. Linear SVM is effective for extraversion, conscientiousness, openness to experience, perceived likeness. Results with both classification models are equal for perceived leadership, ranked dominance. In summary, the proposed co-occurrence features obtained best results for 5 traits, which are significantly better than the random baseline, and obtained equal or the better results than the feature set defined in [2] for 7 traits.

8. DISCUSSION

8.1 Contribution of specific co-occurrence

In this section, we analyze the contribution of specific co-occurrence features to classification performance. Mined co-occurrence can be classified into two types. One is defined as a combination of modalities of the subject (target) participant, such as "speaking with head gesture”. We call this feature set as "Self Context”. The second type is defined as a combination of nonverbal patterns observed from multiple people, such as mutual gaze. We form the "Self Context” event set by removing the second type events from all events.

Next, multimodal features can be also classified into two types. One is the feature set extracted from "On/Off Features” : Speech turn ST, Body motion B, Head motion H, Gaze state G ($F_{1,4,5,7}$ in Table 2) which means whether or not the pattern is observed. The second is the categorical feature set (including three levels: High, Middle, Low, change of time-series) extracted from continuous data: audio pitch P, audio energy E, motion energy image M. The feature set extracted from signal dataset:$\{P, E, M\}$ ($F_{2,3,6}$ in Table 2) can extract detailed information. We obtain the feature event set by removing the events composed from a combination of signal dataset from all events. Table 2 shows the classification results with each co-occurrence feature set: "Self Context”, "On/Off Features”, and All (Both on/off and categorical features). The table also shows the best results obtained with baseline feature set in Table 1.

From Table 2, classification accuracy using all co-occurrence features is the best for agreeableness and perceived competence. On the other hand, the table shows that best classification model is obtained with features extracted from on/off features for conscientiousness, perceived leadership, ranked dominance and with "Self Context” for emotional stability. For extraversion and perceived dominance, the results with features extracted from on/off features and "Self Context” are equal and best with 69.6% and 67.6%, respectively.

The proposed co-occurrence features with on/off data obtained best results for 5 traits, which are significantly better than the ran-

Table 1: Classification accuracy for 10 personality traits (Ridge and L-SVM denote Ridge regression based classifier and Linear SVM respectively)

		Extra- version	Agree- ableness	Conscien- tiousness	Emotional stability	Opennes to Experience	Perceived Leadership	Perceived Dominance	Perceived Competence	Perceived Likeness	Ranked Dominance
(1) Baseline [2]	Ridge	66.67	58.82	51.96	51.96	54.90	**72.55**	**65.69**	52.94	60.78	51.96
	L-SVM	63.44	52.94	52.94	53.92	**61.73**	67.65	60.78	52.94	**64.71**	48.04
(2) Co-occurrence Features	Ridge	64.71	**68.63**	53.92	53.92	57.84	69.61	57.84	**66.67**	53.92	**64.71**
	L-SVM	**67.65**	64.71	50.00	46.08	48.04	**72.55**	61.76	64.71	53.92	**64.71**
(3) Fusing (1) + (2)	Ridge	57.84	60.78	50.98	**55.88**	53.92	67.64	56.86	64.71	55.88	59.80
	L-SVM	66.67	55.88	**55.88**	48.04	43.14	59.80	63.73	59.80	54.90	55.88

Table 2: Classification accuracy for 10 personality traits with various kinds of co-occurence features (Co-occurence features (Self Context, On/Off Features and All) VS Best of baseline [2])

Co-occurrence Features		Extra- version	Agree- ableness	Conscien- tiousness	Emotional stability	Opennes to Experience	Perceived Leadership	Perceived Dominance	Perceived Competence	Perceived Likeness	Ranked Dominance
(1-1) Self context	Ridge	**69.61**	60.78	47.06	47.06	52.94	62.75	60.78	60.78	54.90	62.75
	L-SVM	56.86	65.69	42.16	**56.86**	54.90	64.71	**67.65**	55.88	59.80	58.82
(1-2) On/Off Features ($F_{1,4,5,7}$ in Table 2)	Ridge	**69.61**	63.73	**59.80**	55.88	57.84	**74.51**	52.94	59.80	50.98	**68.63**
	L-SVM	60.78	56.86	50.98	48.04	56.86	69.61	**67.65**	58.82	43.14	63.73
(1-3) All ((2)in Table 1)	Ridge	64.71	**68.63**	53.92	53.92	57.84	69.61	57.84	**66.67**	53.92	64.71
	L-SVM	67.65	64.71	50.00	46.08	48.04	72.55	61.76	64.71	53.92	64.71
Best of Baseline [2]		66.67	58.82	52.94	53.92	**61.73**	72.55	65.69	52.94	**64.71**	51.96

dom baseline, and obtained equal or better results than the feature set defined in [2] for 8 traits. The feature set also improves the accuracy about 7%and 3% for conscientiousness and emotion stability respectively more than best accuracy with the feature set proposed in [2]. This results shows that combinations of only on/off features contribute positively to classification performance in general.

In summary, the proposed co-occurrence features with all features, on/off features and self context features obtained equal or better results than the feature set defined in [2] for 7, 8, 6 traits in 10 traits, respectively. These results show the promise of our approach to improve the inference accuracy of personality traits compared to simple feature sets obtained by accumulated statistics of nonverbal patterns observed from a whole meeting.

8.2 Analysis of Co-occurrence features

We analyze the relationship between these co-occurrence events and each personality trait by correlation analysis. We use the Pearson product-moment correlation for relationship analysis between each trait and the mined patterns with threshold $\alpha = 0.7$.

We discuss for the results of correlation analysis for extraversion, which co-occurrence events were more effective than previous work. Note that there are correlations with other traits as well, but are omitted for space reasons. As the result of correlation analysis, a total of 27586 events (including overlap of events between traits) have significant correlation with $p < 0.1$ for extraversion among 28427 mined events.

To visualize event co-occurrence, we show the co-occurrence matrix in Figure 4, where each value shows the count of each event which has significant correlation with the extraversion trait. Each column corresponds to the type of "seed events" (i.e. the event is used as reference of starting point in mining process) and each row corresponds to the type of event co-occuring with the seed events. In each matrix, red and blue region denotes total frequency of co-occurrence events which has significant positive and negative correlations, respectively. If there are N events such as ith pattern (e.g. ST) co-occurs to jth and kth patterns in co-occurrence events which are positively correlated with extraversion, value (i,i), (i,j), (i,k) in the matrix is N (colored with red) in Figure 4, and these values are $-N$ (colored with blue) where these events have negative correlation. The symbol > 99 denotes when the value is larger than 99. The symbol in each column and row corresponds to that in Table 2. We observe high co-occurrence of nonverbal patterns and the co-occurrence patterns significantly correlated to extraversion from Figure 4. These patterns relate to the traits differently. Some patterns are positively correlated and some patterns are negatively correlated. In the following paragraphs, we observe the relationship of the correlation between co-occurrence patterns and the trait.

Speech (col.1-4): In red regions, speech turn of target persons (ST) is likely to co-occur with that of the others ($SO1$), head motion of target (HMT) and the other. In blue regions, silence ($Ssil$) is likely to co-occur with head motion of others ($HMO1$) and still body motion ($BMsil$) The results show that participants who are perceived as extraverted likely speak with head gestures, start to speak while another member speaks and is given focus of attention by the other members. When there is a participant who is not perceived as extraverted in a group, more silent segments are observed during the interaction.

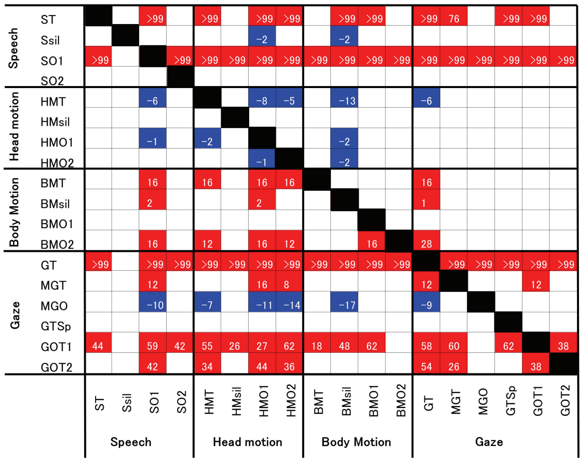
Head and Body motion (col.5-12): Events co-occurred with head motion are negatively correlated to extraversion (blue). On the contrary, events co-occurred with body motion are positively correlated to extraversion (red).

Gaze (col.13-18): Participants who are perceived as extraverted are likely to look at another person (GT) and receive attention from another person ($GOT1,2$). When there is a participant who is not perceived as extraverted, mutual gazing between the others are likely observed.

8.3 From co-occurence to time-series

Through the experiments, we showed that co-occurence features are effective to predict personality traits. The result leads to an open question. The annotators of personality traits do not observe all of visual interaction patterns between the target participant and the group; nevertheless, the patterns improve the classification of the traits. A hypothesis is that personality or nonverbal behavior of the target influences the nonverbal behavior of the other participants.

As an approach to answer the question, we should analyze the time-series structure of co-occurence patterns (e.g. a group head gesture is observed after (or before) the target's utterance), the interval between patterns and their possible causal relations. This structure might reveal which individual patterns influences (or is influenced by) the group activity. Therefore, a future direction in this research is to analyze the structure in multimodal multiparty



		ST	Ssil	SO1	SO2	HMT	HMsil	HMO1	HMO2	BMT	BMsil	BMO1	BMO2	GT	MGT	MGO	GTSp	GOT1	GOT2
Speech	ST	■		>99		>99		>99	>99		>99	>99		>99	76		>99	>99	
	Ssil		■					−2			−2								
	SO1	>99		■	>99	>99	>99	>99	>99	>99	>99	>99	>99	>99	>99	>99	>99	>99	>99
	SO2				■														
Head motion	HMT			−6		■		−8	−5		−13			−6					
	HMsil						■												
	HMO1			−1		−2		■			−2								
	HMO2							−1	■		−2								
Body Motion	BMT			16		16		16	16	■				16					
	BMsil			2				2			■			1					
	BMO1											■							
	BMO2			16		12		16	12			16	■	28					
Gaze	GT	>99		>99	>99	>99	>99	>99	>99	>99	>99	>99	>99	■	>99	>99	>99	>99	>99
	MGT			12				16	8					12	■			12	
	MGO			−10		−7		−11	−14		−17			−9		■			
	GTSp																■		
	GOT1	44		59	42	55	26	27	62	18	48	62		58	60		62	■	38
	GOT2			42		34		44	36					54	26			38	■
		Speech				Head motion				Body Motion				Gaze					

Figure 4: Co-occurrence matrix between multimodal features, where each value is the count of the co-occurrence pattern having significant correlation ($p < 0.1$) with extraversion trait, symbol > 99 denotes when the value is larger than 99 and red (blue) color denotes positive (negative) correlation pair.

conversation with reference of time-series reasoning algorithms [1] and an effective search method for structural temporal multimodal data [16].

9. CONCLUSIONS

We presented a novel feature extraction framework from mutliparty multimodal conversation for inference of personality traits. Our framework represents multimodal features as the combination of each participants' nonverbal activity and also the group activity. Frequently co-occurring events are discovered using graph clustering. We applied the framework for inference of 10 personality trait impressions on the ELEA corpus. Experimental results show that classifiers trained with co-occurring features produced higher accuracy than other features proposed in recent works for 8 of the traits and are statistically better than random for a total of 6 traits of personality traits. In addition, co-occurrence features provide improvement on the classification accuracy ranging from 2% to 17%.

Our feature representation captures the interplay between the nonverbal behavior of an individual and her/his interactions, and can be used for feature extraction for other types of conversation (e.g. dyadic interaction). To validate the versatility of proposed framework, we plan to apply this framework for multimodal datasets observed from other kind of conversations to infer various types of higher attributes such as conversational roles and persuasiveness in the same manner. We also plan to enhance this framework to time-series structure mining. To improve the classification accuracy, using label information for mining is also part of future work (e.g. discriminative feature extraction).

10. ACKNOWLEDGMENTS

We thank the support of the SNSF through the SOBE (PZ00P2 136811), UBImpressed projects and JSPS KAKENHI (25730132, 15K00300, 25280076).

11. REFERENCES

[1] J. F. Allen. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11):832–843, 1983.

[2] O. Aran and D. Gatica-Perez. One of a kind: Inferring personality impressions in meetings. In *Proc. of ACM ICMI*, ICMI '13, pages 11–18. ACM, 2013.

[3] J. Aslam, K. Pelekhov, and D. Rus. A practical clustering algorithm for static and dynamic information organization. In *Proc. of ACM-SIAM Symposium on Discrete Algorithms*, SODA '99, pages 51–60, 1999.

[4] U. Avci and O. Aran. Effect of nonverbal behavioral patterns on the performance of small groups. In *Proc. of workshop on Understanding and Modeling Multiparty, Multimodal Interactions*, pages 9–14. ACM, 2014.

[5] O. Celiktutan, F. Eyben, E. Sariyanidi, H. Gunes, and B. Schuller. Maptraits 2014 - the first audio/visual mapping personality traits challenge - an introduction: Perceived personality and social dimensions. In *Proc. of ACM ICMI*, ICMI '14, pages 529–530, New York, NY, USA, 2014. ACM.

[6] J. W. Davis and A. F. Bobick. The representation and recognition of human movement using temporal templates. In *Proc. of IEEE CVPR*, pages 928–934. IEEE, 1997.

[7] D. Gatica-Perez. Automatic nonverbal analysis of social interaction in small groups: A review. *Image Vision Computing*, 27(12):1775–1787, nov 2009.

[8] S. D. Gosling, P. J. Rentfrow, and W. B. Swann. A very brief measure of the big-five personality domains. *Journal of Research in Personality*, 37:504–528, 2003.

[9] D. Jayagopi and D. Gatica-Perez. Mining group nonverbal conversational patterns using probabilistic topic models. *IEEE Trans. on Multimedia*, 12(8):790–802, 2010.

[10] D. B. Jayagopi, D. Sanchez-Cortes, K. Otsuka, J. Yamato, and D. Gatica-Perez. Linking speaking and looking behavior patterns with group composition, perception, and performance. In *Proc. of ACM ICMI*, pages 433–440, 2012.

[11] O. P. John and S. Srivastava. The big five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of personality: Theory and research*, 2(1999):102–138, 1999.

[12] A. Kendon. *Gesture: Visible Action as Utterance*. Cambridge University Press, 2004.

[13] H. P. Martínez and G. N. Yannakakis. Mining multimodal sequential patterns: A case study on affect detection. In *Proc. of ACM ICMI*, pages 3–10. ACM, 2011.

[14] H. P. Martínez and G. N. Yannakakis. Deep multimodal fusion: Combining discrete events and continuous signals. In *Proc. of ACM ICMI*, pages 34–41, 2014.

[15] A. Metallinou, A. Katsamanis, and S. Narayanan. Tracking continuous emotional trends of participants during affective dyadic interactions using body language and speech information. *Image and Vision Computing*, 31(2):137–152, feb 2013.

[16] C. Miller, L.-P. Morency, and F. Quek. Structural and temporal inference search (stis): Pattern identification in multimodal data. In *Proc. of ACM ICMI*, pages 101–108, 2012.

[17] F. Nihei, Y. I. Nakano, Y. Hayashi, H.-H. Hung, and S. Okada. Predicting influential statements in group discussions using speech and head motion information. In *Proc. of ACM ICMI*, ICMI '14, pages 136–143, 2014.

[18] F. Pianesi, N. Mana, A. Cappelletti, B. Lepri, and M. Zancanaro. Multimodal recognition of personality traits in social interactions. In *Proc. of ACM ICMI*, ICMI '08, pages 53–60. ACM, 2008.

[19] D. Sanchez-Cortes, O. Aran, D. B. Jayagopi, M. S. Mast, and D. Gatica-Perez. Emergent leaders through looking and speaking: from audio-visual data to multimodal recognition. *Journal on Multimodal User Interfaces*, 7(1-2):39–53, 2013.

[20] D. Sanchez-Cortes, O. Aran, M. S. Mast, and D. Gatica-Perez. A nonverbal behavior approach to identify emergent leaders in small groups. *IEEE Trans. on Multimedia*, 14(3):816–832, 2012.

[21] L. A. Turner, L. H. Perry, and H. M. Sterk. *Constructing and reconstructing gender: The links among communication, language, and gender*. SUNY Press, 1992.

[22] A. Vinciarelli. Speakers role recognition in multiparty audio recordings using social network analysis and duration distribution modeling. *IEEE Trans. on Multimedia*, 9(6):1215–1226, 2007.

[23] M. Zancanaro, B. Lepri, and F. Pianesi. Automatic detection of group functional roles in face to face interactions. In *Proc. of ACM ICMI*, pages 28–34. ACM, 2006.