Modeling Dyadic and Group Impressions with Inter-Modal and Inter-Person Features

SHOGO OKADA, Japan Advanced Institute of Science and Technology (JAIST) and RIKEN Center for Advanced Intelligence Project (AIP)
LAURENT SON NGUYEN, Idiap Research Institute
OYA ARAN, *
DANIEL GATICA-PEREZ, Idiap Research Institute and Ecole Polytechnique Federale de Lausanne (EPFL)

This paper proposes a novel feature-extraction framework for inferring impressed personality traits, emergent leadership skills, communicative competence and hiring decisions. The proposed framework extracts multimodal features, describing each participant's nonverbal activities. It captures inter-modal and inter-person relationships in interactions and captures how the target interactor generates nonverbal behavior when other interactors also generate nonverbal behavior. The inter-modal and inter-person patterns are identified as frequent co-occurring events based on clustering from multimodal sequences. The proposed framework is applied to the SONVB corpus, which is an audio-visual dataset collected from dyadic job interviews, and the ELEA audio-visual data corpus, which is a dataset collected from group meetings. We evaluate the framework on a binary classification task involving 15 impression variables from the two data corpora. The experimental results show that the model trained with co-occurrence features is more accurate than previous models for 14 out of 15 traits.

CCS Concepts: • **Human-centered computing** → **Social engineering (social sciences)**; • **Computing methodologies** → *Discourse, dialogue and pragmatics*; • **Information systems** → Clustering;

Additional Key Words and Phrases: Impression, Personality trait, Multimodal interaction, Inference, Data mining

ACM Reference format:
Shogo Okada, Laurent Son Nguyen, Oya Aran, and Daniel Gatica-Perez. 2018. Modeling Dyadic and Group Impressions with Inter-Modal and Inter-Person Features. *ACM Trans. Multimedia Comput. Commun. Appl.* 9, 4, Article 39 (March 2018), 29 pages.
https://doi.org/0000001.0000001

1 INTRODUCTION

The automatic nonverbal analysis of various interaction types is a promising approach for many types of applications. In recent years, one challenge in this research has been to infer the high-level characteristics of participants as target variables, such as their roles, attitudes in conversation, emerging leadership skills, personality traits, and communication skills, by combining audio and visual information obtained from observations of individuals in various social settings, such as monologues, dyadic interactions, and small-group interactions. A key factor to success is the extraction of nonverbal features that can be used to infer the target variable. To extract effective

1551-6857/2018/3-ART39 $15.00
https://doi.org/0000001.0000001

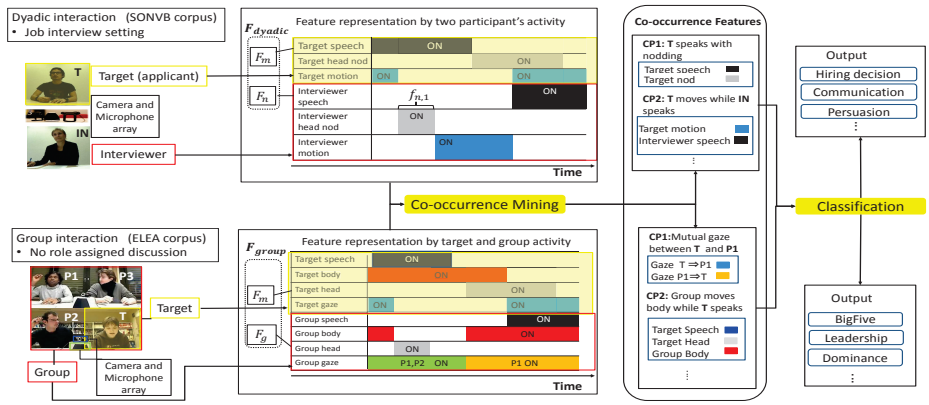


Fig. 1. Overview of proposed framework

features, previous works have defined static features from audio and visual data based on knowledge regarding social science. Audio cues, such as speaking and prosodic features, and visual cues, such as body activity, head activity, hand activity, gaze and facial expression, are used to infer personality traits. Statistics, such as the mean, standard deviation and percentile of these features, are calculated by accumulating each event observed over an entire meeting or conversation. Conversational nonverbal patterns occur on multiple timescales [15] that range from fine-grained features, such as the presence of speech and head-gesture patterns, to contextual conversational patterns, in which multiple events occur simultaneously. Therefore, (1) inter-modal (e.g., speaking with/without gestures) and (2) inter-person (e.g., participant B is nodding while participant A is speaking) co-occurrence features are important for capturing human-human multimodal interactions. From this perspective, static features that are accumulated over an entire meeting do not capture inter-modal and inter-person relationships.

In this paper, we propose a co-occurrence event-mining framework to explicitly extract the inter-modal and inter-person features from multimodal interaction data. The goal of this study is to analyze the effectiveness of the framework in inferring impression scores using multiple datasets representing various social settings. For this purpose, we use the framework to carry out impression inference in both dyadic interactions and group interactions and evaluate its applicability. Via modeling and evaluations, we present guidelines for applying the framework to each type of conversation. The use of co-occurrence patterns between modalities yields two main advantages for modeling the impression scores. First, the inference accuracy of the impression variable can be improved based on a rich feature set extracted by capturing the interactions between modalities (inter-modality) and the interactions between participants (inter-person). Second, discovering key contextual patterns linking personality traits allows understanding the conversational contexts that can be used to predict the trait variables.

In this study, we use the SONVB corpus [31], which includes 62 dyadic interactions that occurred within a real job interview setting. This dataset includes audio and visual data and impression variables for hiring decisions. We also use the ELEA (Emerging LEadership Analysis) corpus, which includes 27 group interactions involving groups of 3 or 4 people. This dataset includes audio and visual data and personality-trait annotations, such as Big Five personality impressions and perceived leadership, scored by group members and external observers [42]. The Big Five model used in psychology is capable of capturing the main personality traits of individuals [22]. In our experiments, we perform binary trait-level classifications to evaluate our approach and compare it

with previous work. The main contributions of this paper are as follows.
(1) To capture inter-modal and inter-person relationships explicitly as features, we propose an efficient co-occurrence mining method that can identify frequent co-occurrences from the combination of $2^{N\,(participants)} \times M\,(modalities)$. Audio-visual features and the corresponding co-occurrence features are extracted automatically by the proposed framework.
(2) We evaluate our approach on two datasets: a dyadic-interaction dataset (SONVB) and a group-interaction dataset (ELEA). We show that the proposed approach is applicable to both dyadic interactions and group interactions.
(3) To demonstrate the effectiveness of co-occurrence features, the well-designed multimodal features proposed in previous studies [5, 31] are used to compare the inference performances. The experimental results show that the use of co-occurring event features improved the accuracy for 9 out of 10 traits in the ELEA corpus and for all five impression indexes in the SONVB corpus.

We present related work in Section 2. Section 3 explains the data-mining framework. Section 4 presents the data corpora used to infer personality traits. Section 5 presents the multimodal features extracted from the corpus. Section 6 and Section 7 present the experimental setting and the results, respectively. Section 8 presents limitations and future works. In Section 9, we conclude the study.

2 RELATED WORK

Our research is related to personality-trait modeling and interaction mining. This study focuses on impression modeling in conversations.

2.1 Impression inferences in dyadic interactions

Several researchers have investigated computational behavior analyses in dyadic interactions for the prediction of outcomes in speed dating [25], job interviews [29, 31], and negotiations[14, 37] and for the identification of personality traits [8] and psychological disorder indicators [43].

To the knowledge, a number of works [29, 31] have used co-occurrence features to infer impression variables. Nguyen et al. [31] extracted not only features from a single modality but also multimodal and interaction (relational) features such as mutual gazing and speaking gestures, which are predefined manually, to infer expert-coded hireability scores. However, the specific contribution of the co-occurrence features to the entire feature set was not reported. Naim et al. [29] proposed a hierarchical coupled hidden Markov model to capture the synchronization of the facial expressions of two participants to infer conversation outcomes. The research shows that synchronized nonverbal templates contribute to the prediction of negotiation outcomes. Although mutual gazing and synchronization of facial-expression patterns are recognized as effective co-occurrence features for impression prediction in [31] and [29], whether other co-occurrence features are effective for such predictions is not clear. The approach proposed in our study focuses on the mining and extraction of various types of co-occurrence features rather than predefined ones (e.g., mutual gaze) to improve prediction accuracy. Co-occurrence features do not have to be manually set, which is required by the methods presented in [31] and [29]. Our approach identifies the co-occurrence patterns that are frequently observed from all possible combinations of modalities.

2.2 Impression inference in groups

For multiparty interactions, different works included different variables: social roles [47, 52], dominance[40], personality traits [5, 38] and leadership [42]. As a common approach of these works, audio and visual features are calculated using the mean, medium, min, max, and X percentile of various statistics (count and length) from each pattern observed throughout an entire meeting or for a part of a meeting [5, 33, 42]. Although this approach can often fuse the total statistics of

patterns observed within a specified duration, it cannot capture co-occurrence between multimodal patterns for each time period. For example, extracting co-occurrence events between an utterance and a body-motion pattern as a feature is useful if the utterance accompanying the body gesture makes a stronger impression on the listener than that utterance without the gesture. Our mining algorithm explicitly extracts such co-occurrence features.

2.3 Inter-modal modeling

Several other studies have focused on extracting the correlations between modalities. Song et al. [44] proposed a multimodal technique that models explicit correlations among modalities via canonical correlation analyses (CCAs) [19]. The algorithm was evaluated using a recognition task for disagreement/agreement with a speaker in political debates [48]. Chatterjee et al. [13] proposed an ensemble approach that combines a classifier based on inter-modality conditional independence with a classifier based on dimension reduction via a multiview CCA. The model explicitly captures the correlations between the modalities but does not focus on extracting co-occurrence patterns that overlap between multiple modalities. These algorithms were applied to a dataset of monologues by speakers presented via social media and spoken during political debates. Our research focuses on extracting features that capture inter-modal and inter-person relationships in interactions but not in monologues.

Feature co-occurrence is often adopted in computer vision [23, 39, 53] and visual search [49, 51, 54]. The idea is also successfully utilized in recommender systems [55]. Going beyond the visual modality, our study shows that feature co-occurrence captures (audio-visual) inter-modal and inter-person relationships to infer impression scores in both dyadic and group face-to-face communication.

2.4 Unsupervised learning and mining for feature extraction

The work in [21] uses latent Dirichlet allocation (LDA)[11] to mine context features in groups. In [21], group features called group looking (or speaking) cues are defined manually and used as input for LDA. Context features are extracted as topics (clusters) generated by LDA. The work in [7] models the influence of one member on other members by relating interactions of nonverbal patterns between group members to transition between hidden states (e.g., one utterance starts after an utterance by another member) in a Markovian formulation.

In [21] and [7], feature extraction is performed for each group to analyze these group nonverbal patterns and group performance or group composition. Moreover, individuals belonging to different groups must be compared within the same metric space. In our study, we propose a novel data-representation method for applying a data-mining framework that separates the nonverbal patterns of one member from those of other members.

A data-mining framework has also been applied for other types of multimodal datasets. The study presented in [26] applied frequent sequence mining as a feature-extraction method for predicting user states while playing games that involved the use of physiological signals, game-play information, and user keystrokes. The study presented in [27] enhanced this framework as an unsupervised feature-learning framework using convolutional neural networks (CNN) [24]. Rather than studying users playing games, the research presented here is focused on multimodal multiparty interactions and dyadic interactions. In general, the phenomena observed in human-to-human conversation involves different issues than those observed for a single participant.

Preliminary works [30, 35, 36] have been performed using co-occurrence pattern mining similar to the proposed approach. Okada et al.[36] used a co-occurrence pattern-mining algorithm, which is a modified version of the algorithm in [46], to extract features to infer the performance level

of storytelling in a group interaction. The main difference with respect to our work is that the research focuses on the modeling of group performance and not individual performance and that nonverbal features are extracted manually. The main limitation of these research works [30, 36] is that only binary event (on/off) features are used for mining. We proposed an approach to convert time-series signals into binary events by using a clustering algorithm in [35]. Using this approach, a time-series event, such as increasing pitch level, could also be extracted. We previously proposed an approach to converting time-series signals into binary events via a clustering algorithm in [35]. In the current study, we reduce the limitations of the framework proposed in [35]. We summarize the main contributions of this paper in the following.

- This study significantly expands upon [35] by adding a second case of use (dyadic interactions) and demonstrates that the techniques improve classification performance for all the variables in the dyadic-interaction dataset. The application of inter-modal and inter-person feature extraction in impression recognition tasks involving dyadic interactions has not been explored in previous works. We show the versatility of inter-modal and inter-person feature extraction by studying a dyadic-interaction dataset and a group-interaction dataset.
- Via the analysis of co-occurrence features, we clarify which type of co-occurrence feature set contributes to the classification performance for each variable in SONVB, for the dyadic interaction case.
- In [35], the early fusion of different types of co-occurrence feature sets occasionally fails because of the unbalanced number of binary event features, such as on/off speech, and categorical features To avoid this problem, a late (score) fusion method is used to fuse these feature sets. The co-occurrence feature set obtained from each event sequence is projected into low-dimensional space using principle component analysis (PCA). We show that combining these techniques with the proposed framework can result in the extraction of effective inter-person and inter-modal features for both datasets, even if the hyperparameters of the model have the same values for both datasets.
- In the appendix, we conduct a sensitivity analysis of the main parameters and evaluate the area under the receiver operator characteristic (ROC) curve for binary classification accuracy to clarify the effectiveness of co-occurrence features. The sensitivity analysis presents a guideline for adopting the proposed framework for other tasks related to dyadic and group interactions.

3 CO-OCCURRENCE MULTIMODAL PATTERN MINING

In this section, we present our mining algorithm to identify the co-occurrence patterns between modalities. The goal of this algorithm is to find the frequently co-occurring features in the feature sets presented in Section 3.1.

3.1 Multimodal feature representation

We propose a feature representation method for capturing the co-occurrence of the nonverbal patterns observed for each participant. We define co-occurrence patterns as multimodal events that overlap in time. Each event has a time length and corresponds to a segment denoted by "ON" in Figure 1. We define an event as a segment in which the feature is active. Multimodal features are represented as follows. First, the feature representation for a dyadic interaction is described. Let F_{dyadic} be the feature set for a dyadic interaction:

$$\boldsymbol{F}_{dyadic} = \{F_m, F_n\}, F_m = \{f_{m,1}, \ldots, f_{m,i}, \ldots, f_{1,N_m}\}. \quad (1)$$

F_* denotes one specific feature (e.g., speaking status); F_m is the feature representation for one specific person, who is the subject for which the impression is inferred; and F_n is the feature

representation for the member n who is sitting opposite member m. In SONVB, member m is the applicant, and member n is the interviewer. F_m and F_n represent the time-series binary data composed of $f_{m,i}$ and $f_{n,i}$, respectively, where $f_{m,i}$ is the ith event observed for member m. The ith event is composed of the binary value for a specific nonverbal feature. We defined an event as a segment in which the feature is active. N_m is the number of nonverbal patterns observed throughout an entire meeting. F_n is defined in the same manner as F_m. Examples of $\{F_m,\ F_n\}$ are presented in Figure 1.

Second, the feature representation for a group interaction is described. We propose a feature representation for comparing nonverbal patterns that are observed for each participant in a group. The representation captures how a participant acts when other members execute any nonverbal activity by simultaneously observing the nonverbal activities of both the individual participant and the other group members. Let F_{group} be the feature set for a group interaction:

$$\boldsymbol{F}_{group} = \{F_m, F_g\}. \tag{2}$$

F_m is the feature representation for one specific person in a group, and F_g is the feature representation for a group composed of the other members without m. An example of $\{F_m,\ F_g\}$ is shown in Figure 1. The co-occurrence pattern mining requires conversion of the time-series signal data into a sequence of events ($f_{m,i}$) with a finite time length as a preprocessing step.

Multimodal behavior is inherently observed as time-series signals in a session. The binarization or discretization of continuous time-series data is described in Section 5. The modified audio-visual features f in F_{dyadic} and F_{group} are also described at the bottom of Table 1 and Table 2, respectively.

3.2 Co-occurrence pattern-mining procedure

We adopt the star algorithm proposed in [6] to efficiently identify co-occurring patterns in time series from continuous time-series data. Figure 2 shows an example of the mining algorithm. To input the multimodal feature set $\boldsymbol{F}$ extracted from all participants into the mining algorithm, $\boldsymbol{F}_{dyadic}$ and $\boldsymbol{F}_{group}$, when each participant is the target, member m is concatenated along the time-series dimension. For example, on SONVB, $\boldsymbol{F}_{dyadic}$, which is extracted from nonverbal behavior of the applicant in session 2, is concatenated after the end frame (end of session) of $\boldsymbol{F}_{dyadic}$ in session 1.

3.2.1 Notation. The multidimensional time-series binary data are represented as $F = \{f_1, \ldots, f_m\}$, which are extracted from all modalities and all participants in the interaction. Each f is composed of events in which the feature is active (e.g., speaking status is "on"). An index pair that includes the start frame and end frame of each event is defined as $fi = \{s, e\}$, and the index set of feature f_m is defined as $FI_m = (fi_{m,1} \ldots fi_{m,N_m})$, where N_m is the number of events. The co-occurrence pattern set is defined as $CF_L = \{cf_1, \ldots, cf_{N_L}\}$. cf denotes a co-occurrence pattern, which is represented as a subset of F (e.g., $cf_1 = \{f_1, f_5\}$ when $L = 2$). L denotes the number of features in a subset. Note that CF_1 equals to F because $cf_* = f_*$.

3.2.2 Mining algorithm. The co-occurrence mining is performed via **Algorithm 1** . The data input into the algorithm are multimodal feature sets F and threshold hyperparameter α, while the data output are the co-occurrence feature sets CF. The goal of the algorithm is to find co-occurrence patterns, which are often observed in all interactions of a data corpus. The pattern-mining process is iterated until the set of co-occurrence patterns CF_L is empty set (row 3 in **Algorithm 1**). In each iteration, similarity based clustering (**Algorithm 3**) is conducted to find a co-occurrence pair between the co-occurrence pattern set CF_L and the feature set in row 4 of **Algorithm 1**.

Algorithm 3 is utilized to find pairs of the following: a co-occurrence pattern cf and feature f. The pairs are merged when more than $\alpha * 100[\%]$ of the total events of cf are temporally

ALGORITHM 1: Co-occurrence Pattern Mining

Input : Multimodal feature set: F; Threshold: α
Output: Co-occurrence pattern set: $\boldsymbol{CF}$;
Frame index of $\boldsymbol{CF}$: $\boldsymbol{FI}$

1 Set initial pattern set: $CF_1 \leftarrow F$
2 Initialization: $L \leftarrow 1$
3 **while** $CP_L \neq \emptyset$ **do**
4 $[CF_{L+1}, FI_{L+1}] = FindingClusters(CF_L, F, FI_L, FI_F, \alpha)$
5 Reject equivalent co-occurrence patterns
6 $L \leftarrow L + 1$
7 **end**
8 $\boldsymbol{CF} = \{CF_1, \ldots, CF_L\}$, $\boldsymbol{FI} = \{FI_1, \ldots, FI_L\}$

ALGORITHM 2: CountOverlap

Input : Frame index: FI_a, FI_b
Output: Number of count: N;
Co-occurrence frame index: FI_{ab}

1 $FI_{ab} = \emptyset$, $N \leftarrow 0$
2 **for** $fi_a \in FI_a$ **do**
3 **for** $fi_b \in FI_b$ **do**
4 **if** $(s_a <= e_b) \cap (s_b <= e_a)$ **then**
5 $fi_{ab} = \{s_a, e_a\}$
6 $FI_{ab} \leftarrow \{FI_{ab}, fi_{ab}\}$
7 $N \leftarrow N + 1$
8 **end**
9 **end**
10 **end**

ALGORITHM 3: FindingClusters

Input : Pattern set: CF_a, F_b;
Frame index: FI_a, FI_b;
Threshold:α
Output: Merged pattern set: CF^*;
Frame index of CF^*: FI^*

1 $N_{ab} \leftarrow 0$
2 **for** $cf_a \in CF_a$ **do**
3 **for** $f_b \in F_b \cap f_b \notin cf_a$ **do**
4 $[Count, fi] = CountOverlap(FI_{cf_a}, FI_{f_b})$
5 $w = Count/N_{cf_a}$
6 **if** $w > \alpha$ **then**
7 $N_{ab} \leftarrow N_{ab} + 1$
8 $cf^*_{N_{ab}} \leftarrow \{cf_a, f_b\}$
9 $fi_{N_{ab}} \leftarrow fi$
10 **end**
11 **end**
12 **end**
13 $CF^* = \{cf^*_1 \ldots, cf^*_{N_{ab}}\}$
14 $FI^* = \{fi_1 \ldots, fi_{N_{ab}}\}$

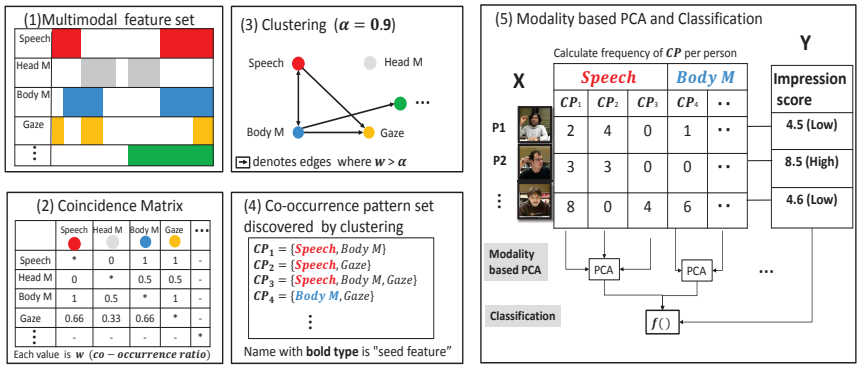


Fig. 2. Example of multimodal pattern mining and representations of features and learning

overlapped with that of f (rows 5, 6 in **Algorithm 3**). The overlap frequency is calculated as $CountOverlap(FI_a, FI_b)$ (**Algorithm 2**). **Algorithm 2** outputs the overlap frequency and the frame index set in which the patterns are temporally overlapped. Here, the ratio of the overlap frequency N to the total number of events N_{cf} is defined as the co-occurrence ratio w in **Algorithm 3**. α $(0 < \alpha < 1)$ in **Algorithm 1, 3** is the threshold of the co-occurrence ratio for the merging of

patterns. α is the hyperparameter in this mining algorithm. If we set α to a small value, the features are merged even if the co-occurrence ratio is low and a large number of co-occurrence features are discovered. If we set α to a large value, the features are merged only when the co-occurrence ratio is high. The dependency of the number of features on the value of α is discussed in Section A.3.1 as an appendix.

Similarity based clustering (**Algorithm 3**) is conducted to find the co-occurrence features in each iteration. After clustering, equivalent co-occurrence patterns are removed (e.g., $cp_1 = \{F_1, F_3, F_5\}$ and $cp_2 = \{F_1, F_5, F_3\}$) from CP_{i+1} to speed up the process. The frame index FI_1 of cf in $CF_1 (= \boldsymbol{F})$ is transferred as that of the co-occurrence feature CF_L. (row 9 of **Algorithm 3**). The feature f in $\boldsymbol{F}$, as a seed of the co-occurrence feature, is defined as the seed feature f_o ((4) in Figure 2). Finally, the co-occurrence feature set is output as $\boldsymbol{CF} = \{CF_1, \ldots, CF_L\}$.

3.2.3 Feature set developed from CF . The co-occurrence features are converted into the feature of each participant as follows. The total number $CFI_{n,m}$ of times that co-occurrence feature $CF_{n,m}$ is observed in the meeting is used as a feature value (table of (5) in Figure 2). We define the co-occurrence feature-vector set as $\boldsymbol{CFI} \in \mathbb{R}^{M \times N}$, where M is the number of participants and N is the number of co-occurrence features. $\boldsymbol{CFI}$ still includes nearly equivalent vectors because the value is composed of similar co-occurrence sets (e.g., $\{F_1, F_3, F_5\}$ and $\{F_1, F_3, F_5, F_7\}$). We reduce the number of dimensions of $\boldsymbol{CFI}$ using PCA for co-occurring features (a group) with each type of seed feature f_o ((4) in Figure 2). If $\boldsymbol{CFI}$ is rewritten as a combination of feature sets with different seed features, then $\boldsymbol{CFI} = \{CFI(1), \ldots, CFI(o), \ldots, CFI(O)\}$. After performing the dimension-reduction process for each group via PCA, the co-occurrence feature set $CFI(o)$ of the oth group is projected onto the low-dimensional vector set $CF^*(o)$. Finally, the feature set is defined as $\boldsymbol{CF}^* = \{CF^*(1), \ldots, CF^*(o), \ldots, CF^*(O)\}$. $\boldsymbol{CF}^*$ is used to develop the classification model.

4 DATA CORPUS

4.1 SONVB: Dyadic-interaction dataset

SONVB is a dyadic-interaction dataset that is used for analyzing impressions related to hiring decisions by the interviewer and the communication skills utilized in a job interview [31]. The corpus includes 62 real employment interviews. More female than male job applicants (45 females, 17 males) were included. The interview structure and hireability annotations followed the same sequence of questions to ensure that comparisons could be made between candidates. In total, the dataset is composed of 670 minutes of recordings (average interview duration: 11 minutes). In this study, five hireability scores were defined: " hiring decision (HirDecision)", " communicative competence (Communication)", " persuasion skill (Persuasion)", " work conscientiousness (Conscience)", and " stress resistance (StressRes)". The hireability score ranged from 1 to 5 except for hiring decisions, which ranged from 1 to 10. The hireability measures were primarily annotated by two professionals in organizational psychology, who are trained in recruiting applicants. Additional details on the SONVB corpus can be found in [31].

4.2 ELEA group-interaction dataset

We used a subset of the ELEA corpus [42] for this study. The subset consists of audio-visual (AV) recordings of 27 meetings in which the participants performed a winter survival task with no roles assigned. The participants in the task played the role of survivors of an airplane crash and were asked to rank 12 items to take with them to survive as a group. Participants first ranked the items individually and then as a group. Participants engaged in a discussion while seated around a table.

To sense the infrastructure, Dev-Audio Microcone [1], a commercial portable microphone array (the green square in the bottom-left picture in Figure 1), was used to collect the audio. Two wide-angle web cameras (the blue squares in the bottom-left picture in Figure 1) were used for the video setup. A total of 102 participants were included (six meetings with three participants and 21 meetings with four participants). Each meeting lasted approximately 15 minutes. The synchronization of audio and video was performed manually by aligning the streams according to the clapping activity. Additional details on the ELEA AV corpus can be found in [41].

Big Five trait impressions from external observers: Personality impressions of the participants according to the external observers were collected in [5]. These annotations include scores for the Big Five traits: "Extraversion", "Agreeableness", "Conscientiousness", "Emotional Stability", and "Openness to Experience". Additional details regarding the Big Five traits can be found in [22]. The Ten-Item Personality Inventory (TIPI) was used to measure the Big Five personality traits of the participants [16]. The TIPI includes two questions per trait, answered on a 7-point Likert scale. Additional details can be found in [5].

Leadership impressions from group members: The ELEA corpus also includes scores for traits of individuals with respect to dominance and leadership. After the meeting task, the participants completed a Perceived Interaction Score, which captures perceptions from participants during the interaction, in which they scored every participant in the group based on four items related to the following concepts: "Perceived Leadership (Leadership)", "Perceived Dominance (Dominance)", "Perceived Competence (Cometence)" and "Perceived Liking (Likeness)". Afterwards, the "Dominance Ranking (Ranked Dominance)". Leadership captures whether a person directs the group and imposes his or her opinion. Dominance captures whether a person dominates or is in a position of power. Participants were asked to rank the group, assigning 1 to the most dominant participant and 3 or 4 to the less dominant participants. Additional details can be found in [42].

5 MULTIMODAL FEATURES

Multimodal features are extracted automatically from audio and visual cues in this study. The feature sets of the SONVB and ELEA used in this study are summarized in Table 1 and Table 2, respectively. We extract co-occurrence features from the feature set in [5, 31] to model the impressions in SONVB/ELEA. First, we explain the audio and visual features in Section 5.1 and Section 5.2, respectively. Then, we present specific features of the SONVB and ELEA in Section 5.3 and Section 5.4, respectively.

Note that different multimodal feature sets for different datasets (Table 1 and Table 2) are used in this study because the original feature sets used for SONVB in [31] and for ELEA in [5] are different. The main difference is that the gaze features are used only in the ELEA corpus and not used in the SONVB corpus. While the original SONVB feature sets [31] include manually coded gaze features, they are coded only as an aggregate for the whole interaction and not as time series data. Because our current study requires input features in the form of time series data, gaze features are not used for SONVB in this study.

Because the ELEA dataset is provided with gaze features and our proposed framework is able to use those features, it is better to use all of these features for comparison with the feature set proposed in[5], which also includes gaze features.

Our specific goal was not to perform a direct comparison between the two datasets, but rather to compare our proposed features with those from previously reported works ([31] and [5]), each of which uses a separate feature set. Thus, our feature set is aligned with those used in the original works [5, 31] to enable a proper comparison of classification accuracy.

[1] Microcone: Intelligent microphone array for groups (now discontinued): http://www.dev-audio.com/

Table 1. Multimodal feature set for the impression index inferred via dyadic interactions: SONVB (Feature set corresponds to $\boldsymbol{F}_{dyadic}$)

ID	Feature	Symbol	Descriptions
Binary features (on/off)			
F_1	speaking status (*ST*)	*LST*	long speech segments of the target person;
		LSA	long speech segments of another person;
		LSsil	silent long speech segment;
		SST	short speech segments of the target;
		SSA	short speech segments of another person;
		SSsil	silent short segments;
		SilT	all silent segments of the target;
		SilA	all silent segments of another person;
		SilAll	silent all segments;
F_2	Head nod (*HN*)	*HNT*	head-nod segments of target person;
		HNA	another person nods;
		HNsil	still nod segment;
Categorical and time-series features			
Prosody features			
F_3	Pitch (*PI*)	*PU*, *PD*, *PN*	$*U$ (up), $*D$ (down), $*N$ (no change): magnitude relationship of the statistics in $(t-1)$ and (t)th utterance or motion segment
		PL, *PM*, *PH*	
F_4	Energy (*EN*)	*EU*, *ED*, *EN*	
		EL, *EM*, *EH*	
F_5	Voice rate (*VR*)	*VU*, *VD*, *VN*	$*L$ (Low), $*M$ (Medium), $*H$ (High): magnitude level of the statistics
		VL, *VM*, *VH*	
Motion of upper body			
F_6	wMEI (*MT*)	*MU*, *MD*, *MN*	Weighted Motion Energy Images (wMEI) used in [4]
		ML, *MM*, *MH*	
Head motion			
F_7	OPT Mag. (*OM*)	*OMU*, *OMD*, *OMN*	OPT: optical flow [9]
		OML, *OMM*, *OMH*	
F_8	OPT Vel. (*OV*)	*OVU*, *OVD*, *OVN*	
		OVL, *OVM*, *OVH*	
F_9	OPT Acc. (*OA*)	*OAU*, *OAD*, *OAN*	
		OAL, *OAM*, *OAH*	
Feature set: $\boldsymbol{F}_{dyadic}$ in Equation 1			
Features from target person: F_m			*LST*, *SST*, *SilT* in F_1, *HMT* in F_2, F_3, F_4, F_5, F_6, F_7, F_8, F_9
Features from another person: F_n			$*A$, $*sil$, *SilAll* in F_1 and F_2

5.1 Audio feature baseline

5.1.1 Speaking status. Binary segmentation is performed to capture the speaking status (*ST*) of each participant. This binary segmentation is provided by the microphone array, and all speaking activity cues are based on the speaker segmentations obtained using the Microcone, which is used for the audio recordings and speaker diarization in [5, 42] and [31]. We define a set of segments in which the speech status is "on" as the speaking-turn set *ST*.

5.1.2 Prosodic features. Prosodic features are extracted for each individual member. Based on the binary speaker segmentation, we obtain the speech signal for each participant. Overlapping speech segments are discarded, only the segments in which the participant is the sole speaker are considered for further processing. Three prosodic speech features (energy, pitch and voice-rate) are determined based on the signal.

We calculate the sign of the difference between the statistics of utterance j and utterance $j+1$. Prosodic features are assumed to change for various reasons. For example, when a participant is likely excited, the energy of his/her utterance may increase after hearing an utterance of another participant. Let pi_j denote the pitch samples extracted from utterance j. We define three types of relationships between feature magnitudes; U, D, and N which indicates up, down, no change, respectively.

Table 2. Multimodal feature set for the impression index used to infer group interactions: ELEA (feature set corresponds to $\boldsymbol{F}_{group}$.)

ID	Feature	Symbol	Descriptions
Binary features (on/off)			
F_1	Speaking status (ST)	ST	speech segments of the target person;
		$SO1$	one person other than the target speaks;
		$SO2$	more than two people speak;
		$Ssil$	silent segment;
F_2	Head Motion (H)	HMT	motion segments of the target person;
		$HMO1$	one person other than the target moves;
		$HMO2$	more than two people move;
		$HMsil$	still motion segment;
F_3	Body Motion (B)	BMT	motion segments of the target person;
		$BMO1$	one person other than the target moves;
		$BMO2$	more than two people move;
		$BMsil$	still motion segment;
F_4	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 the target;
Categorical and time-series features			
Prosody features			
F_5	Pitch (PI)	PU, PD, PN	$*U$ (up), $*D$ (down), $*N$ (no change): magnitude relationships of the statistics in $(t-1)$th and (t)th utterance or motion segment
		PL, PM, PH	
F_6	Energy (EN)	EU, ED, EN	
		EL, EM, EH	
Motion of upper body			$*L$ (Low), $*M$ (Medium), $*H$ (High): magnitude level of the statistics
F_7	wMEI (MT)	MU, MD, MN	
		ML, MM, MH	Weighted Motion Energy Images (wMEI) used in [4]
Feature set: $\boldsymbol{F}_{group}$ in Equation 2			
Features from target person: F_m			ST, HMT in F_2, BMT in F_3, GT, $GTSp$ in F_4, F_5, F_6, F_7
Features from group: F_g			$*O1(O2)$, $*O2$, $*sil$ in F_1, F_2, F_3 $GOT1$, $GOT2$, MGT, MGO in F_4

We perform a statistical t-test between pi_j and pi_{j+1} to determine whether there is the difference between the mean of pi_{j+1} and the mean of pi_j with $p < 0.05$. We categorize utterance $j+1$ into the set PU of utterance segments for which the pitch of the current utterance is larger than that of the previous utterance via the t-test. A similar method is applied to significant decreasing differences, and the set PD of utterance segments for which the pitch of the current utterance is smaller than that of the previous utterance is generated. If the difference is not significant, then the utterance $j+1$ is added to the set PN. For the energy samples, ED, EU, and EN are calculated in the same manner. Voice-rate samples vs per second are calculated as the number of voiced segments per second and we calculate a sample set of vs in utterance j. VD, VU, and VN are calculated in the same manner using the t-test.

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

(1) Calculate the statistics (max, min, average) of prosodic values in each utterance as input samples for clustering.
(2) Perform K-means clustering using the data samples obtained from all participants to assign each utterance into a cluster.

Three clusters corresponding to low-level (L), medium-level (M) and high-level (H) utterances are set. Utterance segments clustered by pitch value are added into the feature sets PL, PM, and PH. The segments clustered by energy value are added into the feature sets EL, EM, and EH. The segments clustered by voice-rate value are added into the feature sets VL, VM, and VH.

5.2 Visual features baseline

Visual activity features characterize the bodily activities of each participant. These features are composed of binary features and categorical features extracted from continuous activity.

5.2.1 Binary activity status. Binary features capture the "on/off" state of the modality.
Body activity: Body activity is measured based on simple motion differences with respect to a stationary background. Hence, all the moving pixels outside the tracked head area are considered to belong to the body area. Each frame is converted to a grayscale image, F_t, and the difference image, $D_t = F_t - F_{t-1}$, is calculated. The difference image is thresholded to identify the moving pixels, and then the total number of moving pixels in each frame is recorded. Binary segmentation is performed using the recorded time-series data, and an activity-state set is extracted for body motion. We define a set of segments in which the body status is "on" as the body-activity set B.
Head activity: As performed for the speech states, binary segmentation based on head tracking and optical flow is performed, and an activity-state set is extracted for the head motion. We define a set of segments as the head-activity set H. The details of the procedure can be found in [42].
Head nod activity: Head nods are defined as vertical movements of the head in which the head is rhythmically raised and lowered. The method proposed in [32] is used to automatically extract head nods. This method calculates the Fourier transform of the optical flow in the head region and inputs it into a support vector machine (SVM) classifier. Framewise classification is performed to detect the nodding state. We define a set of segments as the head-nod set HN.

5.2.2 Continuous activity features. The amount and the time-series changes of motion activity capture more informative nonverbal characteristics of the participant than binary features. The features are extracted in the same manner as used for the prosodic features (Section 5.1.2).
Motion template [4] of upper body: After the difference images are calculated between consecutive frames, weighted motion energy images (wMEIs) are obtained by integrating each difference image from the whole video clip. In this study, a wMEI is calculated from a window of 1 second, and a time-series wMEI is calculated using a sliding-window method. We sum the values of all pixels in the nth wMEI and define the summed value as the amount of activity mt_n in the nth window. As a result, we obtain time-series activity data $MT = \{mt_1, \ldots, mt_N\}$.
Optical flow in head region: The optical flow (OPT) [9] in the head region quantifies the amount of head motion displayed by a person, and it is based on the parametric optical-flow estimation method described in [34]. The overall optical flow between two consecutive frames is calculated inside the face-bounding box using a parametric affine model. The estimated model is then used to calculate the motion at three predefined points within the bounding box that correspond to the eyes and mouth of the person under analysis. We then determine the average motion of these three points, extract the absolute magnitude of OM of the vertical velocity components, and calculate the velocity magnitude OV and the acceleration magnitude OA as the change in velocity. We calculate these features from one-second windows, and the same sliding-window method used for the wMEI is used to obtain the time-series head-activity features.

We segment continuous time-series data wMEI (MT) and OPT (OM, OV and OA) into finite-length patterns by peak detection. The patterns are clustered in the same manner as that used for the prosodic features (Section 5.1.2). We summarize the feature set for the motion template as MT = MU, MD, ML, MM, MH. The feature set MT is calculated for each individual participant. We calculate the feature set for head activity using optical-flow statistics and summarize the feature set for magnitude values as $OM = \{OMU, OMD, OMN, OML, OMM, OMH\}$;
the feature set for velocity magnitude as $OV = \{OMU, OVD, OVN, OVL, OVM, OVH\}$;
and the feature set for acceleration magnitude as $OA = \{OAU, OAD, OAN, OAL, OAM, OAH\}$.

5.3 Features for SONVB

We extract multimodal features individually from both participants in a dyadic setting (corresponding to the applicant and interviewer in the SONVB setting). This feature description is available for dyadic interactions, including those in the SONVB dataset. In this section, we define the target participant, who is the subject of the inference of the impression, as P_a (the applicant in SONVB). We define the other participant as P_i (the interviewer in SONVB).

5.3.1 Audio features. We define features from the speaking-status set *ST*. We define three types of utterance-segment sets, i.e., *LST*, *LSA*, and *LSsi*l, for both participants. *LST* is a set of utterance segments of P_a. *LSA* is a set of utterance segments of P_i. We explicitly extract non-individual-level features *LSA* from another person P_i to capture what the target person P_a does (e.g., moves his/her body or head) when P_i is speaking to capture the listening behavior of the target participant.

LSsil is a set of segments for which the utterance state of both participants is "off". Short utterances are defined as speaking segments with durations of less than 2 seconds. We define three types of short-utterance-segment sets, i.e., *SST*, *SSA*, and *SSsil*, in the same manner as used for *LST*, *LSA*, and *LSsil*, respectively. The aforementioned non-speaking (silent) segments are defined as having the "pause" status. We define three types of pause statuses, i.e., *SilT*, *SilA*, and *SilAll*, in the same manner as used for *LST*, *LSA*, and *LSsil*, respectively. We define *LST*, *LSA*, *LSsi*l, *SST*, *SSA*, *SSsil*, *SilT*, *SilA*, and *SilAll* as speaking status F_1 in Table 1. The prosodic features defined in Section 5.1.2 are the pitch (*P*) (F_3 in Table 1), energy (*E*) (F_4 in Table 1), and voice rate (*VR*) (F_5 in Table 1), which are used in [31].

5.3.2 Visual features. Head nod (*HN*) was defined as nodding segments in which the nod state is "on". We define three types of head-nod segment sets: *HNT*, *HNA*, and *HNsil* (F_2 in Table 1) , which are defined in the same manner as used for *LST*, *LSA*, and *LSsil*, respectively. For the body motion, feature set *MT* in Section 5.2.2 is used. For the head motion, optical flow features are used. The categorical feature sets *MT* (F_6 in Table 1), *OM*, *OV*, and *OA* (F_7, F_8, and F_9 in Table 1) are extracted from the target participant.

5.3.3 Non-individual-level features. In dyadic interactions, the non-individual-level features of an applicant P_a (*LSA*) are equivalent to the individual-level features of interviewer P_i (*LST*). For SONVB, we extract *LST*, *LST*, *SST*, *SilT* in F_1, and *HMT* in F_2, F_3, F_4, F_5, F_6, F_7, F_8, and F_9 as features corresponding to the target person F_m. We extracted $*A$, $*sil$, and *SilAll* as non-individual features in F_1 and F_2 from another person: F_n in Table 1.

5.4 Features for ELEA

In this section, we define the target participant, who is the subject of the inference task, as P_t.

5.4.1 Audio features. The speaking-turn set (F_1 in Table 2) of P_t in the group is denoted as *ST*. We define three types of features, i.e., *SO*1, *SO*2, and *Ssil*, as group speaking-turn features. *SO*1 is a set of segments in which the speech state of a member who is not P_t is "on". *SO*2 is a set of segments in which the speech states of more than two members, not including P_t, are "on". *Ssil* is a set of segments in which the speech states of all members are "off". Pitch (*P*) and energy (*E*) are used as prosodic features; they were also used in [5] ($F_{5,6}$ in Table 2).

5.4.2 Visual features. Head-activity set *H* (F_2 in Table 2) and body-activity set *B* (F_3 in Table 2) are extracted. We also define the features *HMO*1, *HMO*2, and *HMsil* as group head-activity features *H* and the features *BMO*1, *BMO*2, and *BMsil* as group body-activity features in the same manner as used for *ST*. Feature set *MT* in Section 5.2.2 (F_7 in Table 2) is used as a visual

feature. We define a set of segments GT (F_4 in Table 2) in which the target participant looks at the other participants during the meeting. We also define a set of segments $GTSp$ in which the target participant looks at the speaker. We further define two features, $GOT1$ and $GOT2$, as group attention features. $GOT1$ and $GOT2$ are sets of segments in which one member and more than two members, respectively, look at P_t. We define the segment set for mutual gazing (although mutual gazing is defined as a co-occurrence pattern with GT and $GOT1$, 2). We prepare two group features for mutual gazing. MGT is a set of segments in which one member x looks at P_t and vice versa. MGO is a set of segments in which two members y and z, who are not P_t, look at each other.

5.4.3 Non-individual-level features. Other people's features (except those of the target person) are aggregated as group-level features composed of some people's features. For example, consider the case where the target person G1-A is moving his/her body while person G1-B is speaking within time interval T. In this case, the speech segment from person G1-B is aggregated into the group-level feature of $SO1$ or $SO2$ (if the other person is also speaking). The group-level features are different depending on the behavior of the other person (G1-C, G1, D).

This event is also observed from the point of view of person G1-B within the same time interval T. Consider the case where target person G1-B is speaking while G1-A is moving his/her body. In this case, the body movement segment of person G1-B is aggregated into the group-level feature of BMO1 or BMO2. For ELEA, we extract ST and HMT of F_2, and BMT of F_3, and GT and $GTSp$ of F_4, F_5, F_6, and F_7 as features from target person F_m. We also extract $*O1$ ($O2$), $*O2$, and $*sil$ of F_1, F_2, and F_3 and $GOT1$, $GOT2$, MGT, and MGO of F_4 from group: F_g in Table 2.

6 EXPERIMENTS

To evaluate the effectiveness of the proposed co-occurrence features, we evaluate the inference accuracies of the impression variables in the SONVB corpus and ELEA corpus. The impression variables include the 5 variables described in Section 4.1 that are used for capturing hireability in the SONVB corpus and the 5 variables used for capturing personality traits and the 5 variables used for capturing leadership in the ELEA corpus described in Section 4.2.

6.1 Inference tasks and classification model

The inference tasks on the two interaction settings we study have been either binary classification or regression in previous work. For the SONVB dyadic interaction, the inference task was regression in [31]. For the ELEA group interaction, the inference tasks were classification and regression in [5], and then further studied as classification in [35]. In this paper, we decided to focus only on the binary classification task for the two interaction settings. This is a deliberate choice guided by brevity. To evaluate the effectiveness of co-occurrence features, we classify binary levels of the impression index. In the classification task, impression values are converted to binary values (high or low) by thresholding using the median value. For example, this method is performed to represent people scoring high/low in terms of extraversion. The trained model is evaluated based on the classification accuracy of the test data.

We follow the evaluation procedure in [5] and use the ridge regression model (Ridge), linear SVM (L-SVM) and random forest [12] (RF) as classification models. The ridge regression model and linear SVM are used in [5]. We add RF as a classifier, which has different characteristics from those of ridge regression and the SVM in terms of the machine learning mechanism, because it is an ensemble learning method with tree-based classifiers. To train a ridge regression classifier, the original personality-impression scores are used, while the median score is used as the threshold for predictions (this method is called R_{SCR} in [5]).

In the experiments presented 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 a zero mean and one standard deviation. The ridge parameter in the ridge regression model is optimized using a cross-validation scheme, with values in the range of [2, 150]. The parameters of L-SVM are optimized similarly using a nested cross-validation scheme, with C parameter values selected from [0, 0.01, 0.1, 1]. The parameters of RF are optimized similarly using a nested cross-validation scheme, with the numbers of trees per forest selected from [10, 50,100]. The number of random samples per tree is set as the square root of the training sample set. The maximum tree depth is set to 10.

6.2 Late fusion method for co-occurrence features

We use an ensemble-classification technique to fuse the two types of feature sets. The ensemble classier is a linear weighted combination of two classifiers: $f_b(x)$ is trained using the binary co-occurrence feature set, and $f_c(x)$ is trained using category co-occurrence features. The binary co-occurrence feature set includes co-occurrence features for which the seed event is the binary feature (F_1 , F_2 in Table 1 and F_{1-4} in Table 2). The co-occurrence features are discovered from combinations of only binary features via co-occurrence mining.

The categorical co-occurrence feature set includes co-occurrence features for which the seed event is the categorical feature ((F_{3-9} in Table 1 and F_{5-7} in Table 2)). The co-occurrence features are discovered from combinations of all features. For the task of classification, we calculate the posterior probability for x as follows: $Score(x) = \beta f_b(x) + (1 - \beta) f_c(x)$, where β is a weighting parameter. The parameter β is also optimized using a nested cross-validation scheme from [0, 0.25, 0.5, 0.75, 1].

6.3 Feature sets for SONVB

We identify co-occurrence patterns in multimodal feature sets from both the applicant and interviewer, as [31] reports that multimodal features from both enable the effective inference of traits. The total number of combinations of category co-occurrence features (CP_*), is calculated as ($2^{12} \times 3^{14}$) (category) + 2^{12} (binary)= 19591045120 (almost 19 billion) in the SONVB. In total, 247242 co-occurrence patterns are identified after mining the SONVB corpus. After performing PCA, the number of features is reduced to 297. We prepare five types of feature sets to compare the contributions to the classification performance as follows.

Baseline feature set (1): The original feature set used in [31] was kindly shared for comparison purposes. This feature set has 143 dimensions and is composed of audio and visual features which are extracted automatically.

All feature set (2-1): This feature set is composed of all co-occurrence features (297 dimensions), including the inter-modal features, which are observed from the co-occurrence relationships between modalities, and inter-person features, which are observed from interactions between the applicant and the interviewer. The feature set also includes corresponding co-occurrence features as viewed by both the applicant and the interviewer.

Binary feature set (2-2): The binary co-occurrence feature set (22 dimensions) is composed of only "on/off" features: F1 and F2 in Table 1. This feature set is calculated by performing PCA after subtracting the co-occurrence patterns in which the seed event is a categorical pattern. We set $\beta = 1$ as the weight of $f_c(x)$ in Section 6.2.

Applicant self-feature set (2-3): The applicant self-feature set (142 dimensions) contains co-occurrence patterns that comprise a combination of modalities of the applicant in the job interview. This corresponds to the specific feature set by subtracting the patterns observed from the interactions with the interviewer.

Table 3. Classification accuracy for 5 impression traits in the SONVB for various types of co-occurrence features vs. the baseline features [31] ("Ridge", "L-SVM" and "RF" denote the ridge regression-based classifier, the linear SVM and random forest, respectively). The bold values indicate accuracies that are higher than the baseline accuracy.

[%]		HirDecision	Communication	Persuasion	Conscience	StressRes
(1) Baseline [31]	Ridge	70.97	62.90	56.45	75.81	66.13
	L-SVM	59.68	53.23	54.84	79.03	50.00
	RF	66.13	61.29	54.84	69.35	72.58
(2-1) Co-occurrence features all	Ridge	**77.42**	**66.13**	46.77	62.90	61.29
	L-SVM	56.45	**66.13**	51.61	58.06	61.29
	RF	59.68	62.90	54.84	72.58	72.58
(2-2) Co-occurrence features binary	Ridge	62.90	**64.52**	**61.29**	72.58	64.52
	L-SVM	48.39	**64.52**	48.39	**85.48**	69.35
	RF	56.45	56.45	59.68	77.42	61.29
(2-3) Co-occurrence features applicant self	Ridge	66.13	50.00	48.39	43.55	58.06
	L-SVM	64.52	56.45	**62.90**	46.77	56.45
	RF	67.74	62.90	59.68	53.23	69.35
(2-4) Co-occurrence features interviewer self	Ridge	58.06	**69.35**	**66.13**	67.74	58.06
	L-SVM	61.29	**72.58**	51.61	67.74	48.39
	RF	51.61	64.52	58.06	70.97	**74.19**

Interviewer self-feature set (2-4): The interviewer self-feature set (173 dimensions) contains co-occurrence patterns that comprise a combination of modalities of the interviewer in the job interview. It is calculated in the same manner as set (2-3).

6.4 Feature sets for ELEA

We conduct the co-occurrence pattern mining for the multimodal feature set, whose members are constructed by concatenating target and group features in the ELEA corpus. The total number of combinations of category co-occurrence features, CF_c, is calculated as $3^6 \times 2^{18} + 2^{18} = 191365120$ (almost 0.19 billion). In total, 124639 co-occurrence patterns are identified after mining the ELEA corpus. After performing PCA, the number of features are reduced from 124639 to 80. We prepare four types of feature sets to compare the contributions to the classification performance as follows.
Baseline feature set (1): The original feature set used in [5] was shared for comparison purposes. This feature set has 37 dimensions and is composed of audio and visual features.
All feature set (2-1): This feature set is composed of all co-occurrence features (80 dimensions), including the inter-modal and inter-person features.
Binary feature set (2-2): The binary co-occurrence feature set (28 dimensions) is composed of only "on/off" features: F_{1-4} in Table 2.
Target self-feature set (2-3): The target co-occurrence self-feature set (19 dimensions) is composed of co-occurrence patterns that include a combination of modalities of the subject (target) participant.

7 RESULTS

7.1 Classification accuracy for SONVB

Table 3 shows the classification accuracies. The bold values indicate the accuracies that are higher than the baseline accuracy (1). The underlined bold values indicate the best accuracies of all models. Only accuracies above 65.8% are considered significantly better (with a 99% confidence level) than the 50% random-assignment baseline.

7.1.1 Baseline features vs. co-occurrence features. In this section, we compare the accuracy of the model with all co-occurrence features (2-1) with that of the model with baseline features (1). Table

3 shows that the best classification accuracy is obtained by the proposed model with co-occurrence features in the classification task for 2 impression values: hiring decision and communication. For the hiring decision, we obtain accuracies as high as 77.4% using the ridge regression classifier with all co-occurrence features (2-1), whereas the accuracy with the baseline features is 70.9%. The model trained with co-occurrence features achieves an increase in accuracy of approximately 7% relative to the accuracy of the feature set proposed in [31]. The hiring decision is the target variable in [31], and the total score of the impression is based on the performance during a job interview. For the communication-competence trait, the model with co-occurrence features achieves an accuracy of 66.1% for both classifiers; this result is significantly different and better than the random baseline.

7.1.2 Contribution of specific feature set. In this section, we analyze the contributions of specific co-occurrence features to the classification of the impression trait. Rows 2-5 in Table 3 show the classification results of specific co-occurrence features for hireability impressions. Each column corresponds to an impression variable, and the rows correspond to feature sets and classification methods. In each feature set, "Co-occurrence features" denotes our proposal, and "Baseline" denotes the feature set defined in [31].

For hiring decisions, the other models (2-2, 2-3, 2-4) do not improve the accuracy of the baseline, which means that use of the complete co-occurrence feature set results in effective prediction of the hiring decision. For the communication-competence trait, we obtain accuracies as high as 72.5% using a fusion of the SVM classifier (L-SVM) with the interviewer co-occurrence features (2-4). The model (2-4) improves the best accuracy of the baseline (65.8%) by approximately 7%. For the persuasion impression, the best accuracy (66.1%) is achieved by the ridge regression classifier with the interviewer co-occurrence features (2-4). For resistance to stress, the best accuracy (74.1%) is achieved by the RF with (2-4). For conscientiousness, the best accuracy (85.4%) is achieved using the SVM model with binary features.

In summary, the proposed co-occurrence features yield the best results for all 5 traits, and the results for all impression variables are significantly different from the random baseline feature accuracy of 65.8%. In particular, both the model with binary co-occurrence features (2-2) and the interviewer's co-occurrence features (2-4) improve the accuracy of the baseline [31] for 3 traits. The best accuracy for hiring decisions is achieved by the ridge regression model with all co-occurrence features (2-1), that for communication is achieved by the SVM model with the interviewer's co-occurrence features (2-4), that for persuasion is achieved by the ridge regression model (2-4), that for conscientiousness is achieved by the SVM model with binary co-occurrence features (2-2), and that for resistance to stress is achieved by the RF model (2-4).

7.2 Classification accuracy for ELEA

Table 4 shows the classification accuracies for the ELEA dataset. The bold values indicate the accuracies that are higher than the baseline values in [5]. The underlined bold values indicate the best accuracies of all models. Only accuracies above 62.7% are considered significantly better (with a 99% confidence level) than the 50% random baseline.

Columns 1-5 in Table 4 show the classification results for the Big Five impressions, and columns 6-10 in Table 4 show the results for the leadership and the dominance. Each row corresponds to a type of feature set, and each column corresponds to a type of personality trait. In each feature set, "Co-occurrence features" (2-1,2,3) denotes our proposed feature set and "Baseline" (1) denotes the feature set defined in [5][2].

[2] The accuracy of "Baseline" (1) in Table 4 is different from those in [5] ("WM/WM"of Figure 4 (a) and (b)). [5] reported only the best classification accuracy for "extraversion" and "openness to experience" using most effective feature group and the classification accuracy using all multimodal features (setting in this study) was not reported in [5].

Table 4. Classification accuracy for 10 personality traits in ELEA for various types of co-occurrence features vs. the baseline features [5]. The bold values indicate accuracies that are higher than the baseline accuracy. The underlined bold values indicate the accuracies that are highest among the models considered for comparison.

[%]		Extra-version	Agree-ableness	Conscien-tiousness	Emotional stability	Openness to Experience	Leader-ship	Domi-nance	Compet-ence	Like-ness	Ranked Dominance
(1) [5] Baseline	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	61.76	67.65	60.78	52.94	64.71	48.04
	RF	61.76	49.02	46.08	50.98	49.02	72.55	58.82	50.98	59.80	59.80
(2-1) Co-occurrence features all	Ridge	**70.59**	**66.67**	**53.92**	52.94	60.78	68.63	55.88	**56.86**	53.92	**57.84**
	L-SVM	58.82	**64.71**	52.94	52.94	**65.69**	**73.53**	50.00	49.02	**65.69**	**61.76**
	RF	62.75	**63.73**	53.92	**56.86**	55.88	62.75	57.84	55.88	54.90	60.78
(2-2) Co-occurrence features binary	Ridge	**67.65**	**64.71**	**57.84**	**56.86**	54.90	**75.49**	58.82	**60.78**	56.86	**57.84**
	L-SVM	61.76	**64.71**	**54.90**	52.94	51.96	**73.53**	54.90	51.96	62.75	**64.71**
	RF	60.78	**61.76**	51.96	58.82	49.02	59.80	51.96	50.98	51.96	55.88
(2-3) Co-occurrence features self	Ridge	**72.55**	58.82	**55.88**	**58.82**	58.82	59.80	52.94	**59.80**	46.08	**53.92**
	L-SVM	**71.57**	57.84	44.12	**57.84**	55.88	64.71	58.82	**60.78**	54.90	**56.86**
	RF	62.75	58.82	50.98	47.06	54.90	52.94	51.96	55.88	58.82	50.00

Table 5. Comparison between accuracies of the model in [35] and the proposed model (late fusion of co-occurrence features (proposed) vs. early fusion of co-occurrence features ([35]) VS. best of baseline [5]). The bold values indicate the accuracies that are highest among the models considered for comparison.

[%]	Extra-version	Agree-ableness	Conscien-tiousness	Emotional stability	Openness to Experience	Leader-ship	Domi-nance	Compet-ence	Like-ness	Ranked Dominance
Best of late fusion ((2-1) in Table 4)	**70.59**	66.67	**53.92**	**56.86**	**65.69**	**73.53**	55.88	56.86	**65.69**	61.76
Best of early fusion in [35]	67.65	**68.63**	**53.92**	53.92	57.84	72.55	61.76	**64.71**	53.92	**64.71**
Best of baseline [5]	66.67	58.82	52.94	53.92	61.76	72.55	**65.69**	52.94	64.71	51.96

7.2.1 Baseline features vs. proposed co-occurrence features. Table 4 indicates that the best classification accuracies for 9 traits are achieved by the model (2-1) using all co-occurrence features. Our proposed features do not improve the accuracies for perceived dominance. For extraversion, agreeableness and openness to experience of the Big Five, the best results are obtained with the co-occurence features. For extraversion, the best accuracy is as high as 70.5% obtained with ridge regression. For agreeableness, an accuracy of 66.7% is achieved using the ridge regression model, and 65.6% is achieved for openness to experience using the SVM with co-occurrence features. The baseline feature set [5] generates significantly better accuracy for only extraversion compared with the random baseline accuracy. In particular, the model trained with co-occurrence features improves the accuracy for agreeableness by approximately 8% compared with the accuracy of the feature set proposed in [5]. However, the results for conscientiousness and emotional stability are not significantly different than the random baseline for both methods.

For perceived leadership, perceived competence, perceived likability and ranked dominance, better results are obtained with the proposed model than with the feature set [5], with the accuracies being as high as 73.5%, 56.8%, 65.6% and 61.7%, respectively. The proposed model with all co-occurrence features (2-1) yields, for 5 traits, results that are significantly better than the random baseline, and the results for 9 traits are better than those of the feature set in [5].

7.2.2 Contribution of specific co-occurrence features in group. In this section, we analyze the contributions of specific co-occurrence features to the classification performance of three models (2- 1, 2-2, 2-3). Rows 2-4 in Table 4 show the comparison between specific features. The underlined bold value indicates the best accuracy; the results reveal that the proposed model with all features

(2-1) achieves the best accuracy for agreeableness, openness to experience, and perceived likeness. The proposed model with on/off features (2-2) achieves the best accuracies for conscientiousness, perceived leadership, perceived competence and ranked dominance. In particular, the accuracy for leadership is 75.4%, which is the best accuracy among all classification tasks for the ELEA corpus. The accuracy for ranked dominance (64.7%) is significantly better than the random baseline. Although the on/off features are simple, they capture both inter-modal and inter-person properties. For the co-occurrence self-feature set (2-3), the table shows that the best classification model is that with features extracted from the "Self-Context" features, with an accuracy of 72.5% achieved for extraversion and 58.8% achieved for emotional stability.

In summary, the proposed co-occurrence features with all features, on/off features and self-context features yielded equally good or better results than the baseline feature set for 9, 7, and 5 traits, respectively, out of 10 traits. These results show the potential of our approach to improve the inference accuracy of personality traits compared with well-designed feature sets developed from the accumulated statistics of nonverbal patterns observed over an entire meeting. For nine traits, the models (2-1, 2-2) yielded better results than the feature set in [5] (in 6 of these traits, the results are significantly better than the random baseline). These results also show that the inter-modal and inter-person features are common features for many traits.

7.3 Contribution of late fusion

This section compares the model presented in our previous work [35] and the model proposed here. A unimodal event with a long duration frequently co-occurs with many types of events. The inherent characteristics result in an imbalance in the number of features between modalities. In the ELEA dataset, the number of co-occurrence features based on categorical events tends to be larger than the number of that based on binary events.

To avoid an imbalance of features, PCA is performed for only co-occurring features with categorical features, such as pitch, energy and wMEI in our previous work [35]. The early fusion method is adapted to fuse co-occurrence features with binary events and features with categorical events, which are projected to low-dimensional space via PCA. However, the ad hoc method does not always work well because the number of co-occurrence features varies for different modalities. In the proposed method, PCA is adapted for each co-occurrence feature set, using each seed feature to balance the number of features between the modalities and dimensions. In addition, we use the late-fusion method to fuse features co-occurring with the events and binary events. Table 5 shows a comparison of the best accuracies among those of the model in [35], the proposed model and the baseline feature set in[5]. According to [35], the threshold parameter for the mining was set as $\alpha = 0.8$. Normalized features for individual participants are included in the feature set. In this study, the threshold parameter in the mining is set as $\alpha = 0.9$, and normalized features are not used. We report only results that are significantly better than the random baseline results. The table shows that the best classification model is obtained using the proposed late-fusion method, which presents accuracies of 70.5% for extraversion, 65.6% for openness to experience, 73.5% for perceived leadership, and 65.6% for perceived likability.

The table also shows that the early fusion method ([35]) achieves the best accuracies for agreeableness (68.6%), perceived competence (66.6%) and ranked dominance (64.7%). The baseline features in [5] result in the best accuracy for perceived dominance (65.6%). The results show that the proposed late-fusion method exhibits the best performance for most of the 4-trait prediction tasks. Moreover, an equivalent approach to fusion and feature extraction is used for SONVB and ELEA, with the accuracy being almost better than the baseline. We conclude that reducing the features mined via PCA for each modality combined with late fusion is a stable method for impression modeling.

7.4 Sensitivity analysis of classification results

The robustness with respect to binary classification is determined by the discriminativeness of the trained models. We discuss the robustness of models with the proposed features for binary classification by calculating the area under the ROC curve in Section A.1. The binary classification accuracy is influenced by the hyperparameters of the model. To analyze the relationship between the classification accuracy and the hyperparameters, we conduct a sensitivity analysis of the hyperparameters: (1) weighting-parameter β on late fusion in Section A.2 and (2) mining-parameter α in Section A.3. The contributions of each modality are discussed in Section A.4.

8 LIMITATIONS AND FUTURE WORK

We now discuss three limitations of our work and research directions to address them.

8.1 Improved extraction of behavioral features

As discussed earlier, gaze features for the dyadic setting were neither extracted nor used. Future work should investigate them, as well as other features that are amenable for similar treatment as we propose here, like facial expressions, which have been shown to be informative of some traits like extraversion [10][45]. A second related issue is the analysis of the time-series structure of the co-occurrence patterns (e.g., a group head gesture is observed after (or before) the target's utterance), the intervals between patterns, and their possible causal relations. This structure might reveal the individual patterns that influence (or are influenced by) the group activity or other participants. Therefore, a future direction is to adopt time-series reasoning algorithms [3] and a method of searching for structural temporal multimodal data [28] to the proposed mining framework.

8.2 Feature extraction with deep learning

The results of consequent experiments show that the proposed inter-modal and inter-person feature representation and extraction is effective for improving accuracy. Based on this finding, more effective feature-extraction methods for inter-modal and inter-person representation (column: N (people) $\times$ M (modalities), row: T (time length of interaction)) should be explored in the future. Along this line, the deep learning family will be a good candidate for replacing the proposed method. Future work could investigate the most useful ways of using deep learning to make a comparison with the proposed approach. Although deep learning techniques promise discriminating performance when a large amount of data is available for training models, the dataset used in this study includes a maximum of 102 samples, and the data size is too small to directly adopt the deep learning approach for this task. For future work, a guideline for applying the deep learning algorithm to the impression recognition task is as follows.

The autoencoder and stacked autoencoder [17], as unsupervised deep learning methods, are effective for reducing the feature space dimensionality. First, a sparse autoencoder can be used instead of PCA in the proposed framework. Second, an approach exists that can extract co-occurrence features directly from multimodal time-series data (Section 3.1) using an autoencoder. In this method, an input vector is concatenated from $(N \times M) \times Xframes$, and the sampling is conducted in the time-series dimension by adopting a sliding-window method to input the network of the autoencoder.

Third, a convolution neural network (CNN) [24] can be used for supervised feature extraction from time-series data in [50]. In our study, prosody time-series data and motion time-series data were discretized into three levels via clustering or a statistical t-test. The process can be replaced with a CNN. A CNN is used to extract discriminative features from time-series data obtained from participants with a binary label (higher trait or lower trait score). In this case, we need to define

the convolution and pooling operators in the time-series dimension by adopting a sliding-window method to segment the time-series signal into a collection of short portions of the signal. The CNN can be replaced with a recurrent neural network (RNN) or long short-term memory (LSTM) networks [18] as the feature extractor for time-series data.

8.3 Toward impression recognition in other types of conversations

In recent years, some studies [2, 20] have focused on the analysis of participants interacting naturally in organized free-standing conversations because of the availability of several wearable and ubiquitous sensing devices to the general public. The feature extraction becomes more complex for free-standing conversations in comparison with face-to-face interactions. More preprocessing of the audio and video (e.g., visual person tracking and group identification, audio separation and speaker identification) or the use of ubiquitous or wearable sensors is required to extract the audio and visual features of each individual participant. After extracting individual audio/visual features, the inter-modal and inter-person feature representation can be developed, and co-occurrence events can be detected in the same manner. Based on the above, we believe that the proposed approach can be used for free-standing interactions. A multimodal dataset [1] collected from free-standing conversational groups in unstructured social settings is publicly available. Adapting the proposed framework to this dataset is another direction for future work.

9 CONCLUSIONS

We presented a novel feature-extraction framework for multimodal conversations to infer personality traits. Our framework represents multimodal features as combinations of each participant's nonverbal activities and the activities of others. Frequently co-occurring events are identified via co-occurrence clustering. We applied the framework to infer 5 hireability impressions for the SONVB corpus and 10 personality-trait impressions for the ELEA corpus. The experimental results showed that classifiers trained with co-occurring features were more accurate than those trained with other features proposed in recent works [5, 31] for all the impressions in the SONVB data and 9 of the impressions in the ELEA data. Moreover, these classifiers were shown to be statistically better than random classifiers for all the impressions in SONVB and 6 personality traits in ELEA. In addition, the co-occurrence features were shown to improve the classification accuracy from 3% to 13%. Our feature representation captures the interplay between the nonverbal behavior of an individual and her/his interactions, and it can be used for feature extraction for other types of conversations (e.g., negotiations, counseling, and group-learning settings).

To validate the versatility of the proposed framework, we plan to extend this framework by using time-series structure mining, and to apply the proposed framework for impression recognition in other types of conversations. To improve the classification accuracy, the discriminative-feature extraction with deep learning will be implemented in future work.

A ADDITIONAL EXPERIMENTAL RESULTS

In this section, we focus on the analysis of the classification results of the main impression traits in the SONVB and ELEA corpora. The target traits of the two corpora are hiring decision and leadership, respectively. In addition to these two traits, extraversion and agreeableness from the Big Five, which are annotated in [5], are the subject of the analysis, as annotations of the two traits have higher agreement between annotators than the other three traits. Thus, a total of four traits serve as the subject of this analysis. From a comparison of the experimental results for each classifier, RF achieved the best accuracy in terms of only "resistance to stress" in SONVB. RF is thus

Table 6. Area under the curve (AUC) calculated from ROC curve based on output scores from models for HirDecision in SONVB

	[AUC]	HirDecision
(1) Baseline [31]	Ridge	0.719
	L-SVM	0.637
(2-1) Co-occurrence features all	Ridge	**0.815**
	L-SVM	0.662
(2-2) Co-occurrence features binary	Ridge	0.701
	L-SVM	0.533
(2-3) Co-occurrence features applicant self	Ridge	0.696
	L-SVM	**0.744**
(2-4) Co-occurrence features interviewer self	Ridge	0.576
	L-SVM	0.637

Table 7. Area under the curve calculated from output scores of models for extraversion, agreeableness, and leadership in ELEA

	[AUC]	Extra-version	Agree-ableness	Leader-ship
(1) Baseline [5]	Ridge	0.745	0.596	0.759
	L-SVM	0.660	0.562	0.732
(2-1) Co-occurrence features all	Ridge	**0.754**	**0.732**	0.705
	L-SVM	0.620	**0.719**	**0.807**
(2-2) Co-occurrence features binary	Ridge	0.722	**0.708**	**0.789**
	L-SVM	0.644	**0.719**	**0.807**
(2-3) Co-occurrence features self	Ridge	**0.787**	**0.629**	0.627
	L-SVM	0.741	**0.618**	0.631

neglected in the following experiments, as the best accuracies are achieved by the ridge regression classifier or linear SVM regarding the four main traits.

A.1 Analysis of robustness against binary classification

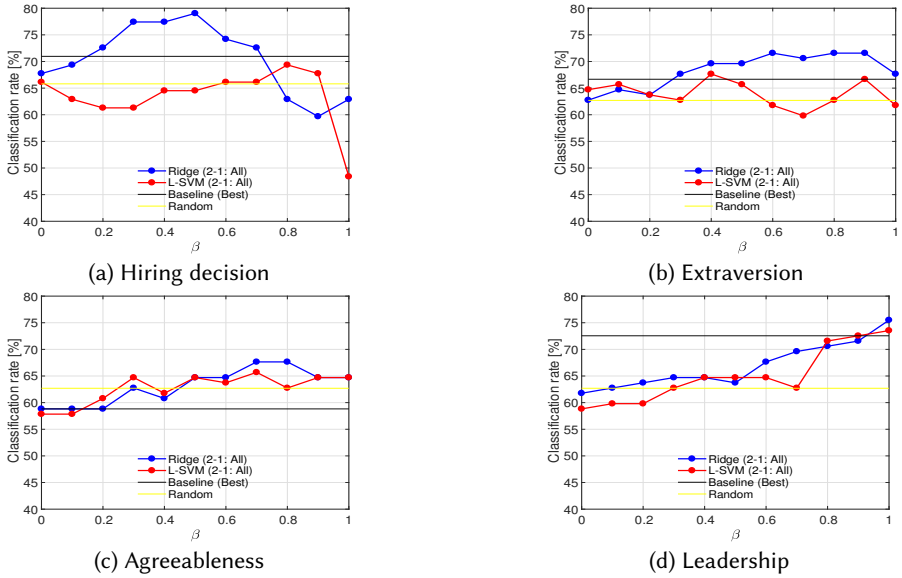
An ROC curve is used to display the performance of a binary classification algorithm. To validate the classification accuracy of models with co-occurrence features, the ROC curve is plotted, and the area under the curve (AUC) is calculated. The AUC is commonly used to compare different classification algorithms for a given dataset. It is well known that the value of the AUC denotes the discriminativeness of a model.

Table 6 shows the AUCs calculated from the ROC curve based on the output scores of the models for HirDecision in SONVB. As in Section 7, the AUC of the models with the proposed co-occurrence features and that of the model with baseline features are compared. The proposed model (ridge regression) with all co-occurrence features (2-1) yielded the best AUC for hiring decision. The AUC was 0.81, which is better than the AUC (0.71) of the best result of the baseline model [31] (ridge regression) by 0.1 point. This result agrees with the result of the classification rate presented in Table 3. In addition to the result, the two models (2-3) with features extracted from the "Self-Context" features of applicants also yielded better AUCs than that of the baseline. The AUC of the SVM were 0.74, respectively, though the classification accuracies of these models were worse than the best accuracy of the baseline. The results show that the co-occurrence features contribute to the development of a robust binary classification model.

Table 7 shows the AUCs for the ROC curves for extraversion, agreeableness, and leadership in ELEA. The proposed model (2-1) with all co-occurrence features yielded a better AUC for extraversion, agreeableness, and leadership than that of the baseline [5]. The best AUC for extraversion was yielded by the ridge regression model (2-3) with the "Self-Context" feature, while that for agreeableness was obtained by the ridge regression model (2-1) with all co-occurrence features and SVM models (2-1 and 2-2) with either binary features or all features. The AUCs were 0.78, 0.73, and 0.80, respectively. Co-occurrence features improved the best AUC of the baseline (0.74, 0.59, and 0.75) by approximately 0.04, 0.14, and 0.05 points, respectively. In particular, six models with co-occurrence features resulted in better AUCs for agreeableness than that of the baseline. Based on these experimental evaluations involving both classification accuracy and the AUC, co-occurrence features contribute not only to improve the binary classification accuracy but also to develop a robust model for the binary classification of impressions.

A.2 Weighting parameter β when fusing binary and category features

In the late fusion, the weighting parameter is used to control the trade-off between the output scores of both models and the binary and category features. In this section, we analyze the dependency of



(a) Hiring decision (b) Extraversion

(c) Agreeableness (d) Leadership

Fig. 3. Dependency of classification accuracy on the weighting parameter (β)

the classification accuracy on the weighting parameter (β) to clarify the contribution of binary and category features. From equation 6.2, when $\beta = 0$, the model is trained with only category features. When $\beta = 1$, the model is trained with only binary features, and the model is equal to the model with the binary co-occurrence feature set (2-2).

Figure 3a, 3b, 3c, and 3d show the dependency of classification accuracy on the weighting parameter (β) for hiring decision, extraversion, agreeableness and leadership. In these figures, the yellow line denotes the best classification accuracy of the baseline, and the black line denotes the significance level of a random baseline. The blue and red lines denote the classification accuracies of the ridge regression models with all co-occurrence features (2-1) and the SVM models with all co-occurrence features. In Figure 3a, the best accuracy (79.0 %) for hiring decision is achieved when $\beta = 0.5$. This means that both feature sets make equal contributions to the classification task. In this case, the fusing of features is an effective way to improve the accuracy for hiring decision.

In Figure 3b, the best accuracy (71.5 %) for extraversion is achieved when $\beta = 0.6$. The accuracies of the fused models are better than that of the baseline when $\beta = [0.3 - 1.0]$. These results indicate that binary features contribute to the improvement of the classification accuracy more than the category features, as the accuracy degrades as the weight value of the category feature increases ($\beta = [0.0 - 0.2]$). From 3c, the dependency of the accuracy for agreeableness is similar to that for extraversion. The best accuracy is 67.6% when $\beta = [0.7, 0.8]$.

In Figure 3d, the dependency of leadership is different from that of the other three traits. The best accuracy (75.4 %) is obtained when $\beta = 1$ (only binary features). This means that category features are unnecessary to improve the classification accuracy for this trait, The effective weighting rate is different among the four traits. In our method, the grid search for the weighting parameter is conducted by applying cross-validation to the training dataset. The grid search is effective when determining the optimal weighting ratio for the classification of each trait.

A.3 Threshold parameter α on co-occurrence mining

In this section, we analyze the dependency of the classification accuracy on the threshold parameter α when finding co-occurrence features via mining.

Table 8. Number of co-occurrence features and dependency of α on the number of co-occurrence features in the ELEA corpus (the mining process (iteration) was stopped if the number of features increases to more than 10^5. The case is marked with $*$.)

α	0.9		0.8		0.7	
	binary	category	binary	category	binary	category
After Mining	59	124580	407	*302523	*146601	*263552
After PCA (99%)	28	52	57	127	155	184
Total dimensions	80		184		339	

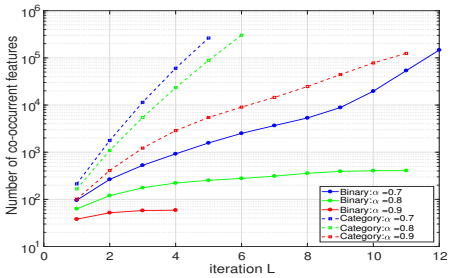


Fig. 4. Number of co-occurrence features per α ("Category" denotes category co-occurrence features and "Binary" denotes the binary features.)

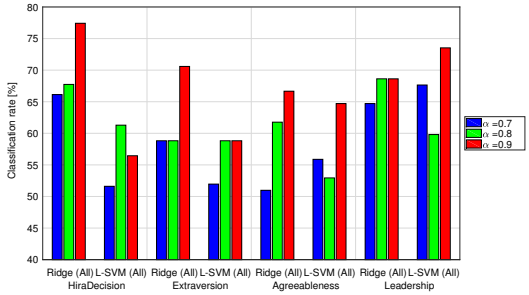


Fig. 5. Classification accuracy based on the change in the threshold parameter to merge features as a co-occurrence feature

A.3.1 Number of co-occurrence features per hyperparameter α. In this section, by analyzing the number of co-occurrence features by varying α from among $[0.7, 0.8, 0.9]$ for the ELEA corpus, we show that $\alpha = 0.9$ is reasonable. The total number of combinations of binary co-occurrence features, CF_b, is calculated as $2^{18} = 262144$, as 18 kinds of binary features are extracted from the ELEA corpus, as shown in Table 2. The total number of combinations of category co-occurrence features, CF_c, is calculated as 191102976.

The co-occurrence mining in Section 3.2 was conducted to extract a feature set from the ELEA corpus by changing α. During the process of pattern mining, the change in the number of features in each step is as shown in Figure 4. Figure 4 shows the dependency of the number of binary co-occurrence features CF_b and category co-occurrence features CF_c. The plot with the dotted line and square shapes denotes the number of category co-occurrence features, while the plot with the circle shapes denotes the number of binary features.

For binary co-occurrence features CF_b, when we set α to 0.9 and 0.8, the iteration of the mining algorithm is terminated after 4 iterations and 11 iterations, respectively. When we set α to 0.7, the number of binary co-occurrence features increases exponentially, and the number exceeds 0.1 million (10^5) after 4 iterations, at which point we stop the mining process. For category co-occurrence features CF_c, when we set α to 0.9, the number of binary co-occurrence features becomes 124580 after 11 iterations. When we set α to 0.8 and 0.7, the number of binary co-occurrence features increases exponentially and exceeds that when $\alpha = 0.9$ after 6 and 5 iterations, respectively.

From the results, when we set α to a low value, too many features are extracted, and it becomes difficult to handle the feature set. We summarize the number of features before/after conducting co-occurrence mining and PCA in Table 8. When we set α to 0.8 or 0.7, we stop the mining process if the number of features exceeds the number of features when $\alpha = 0.9$ (more than 10^5). The number at which we stop the mining process is denoted by $*$.

Co-occurrence features discovered after mining still include irrelevant features. PCA is conducted to remove these irrelevant features. The final total numbers of dimensions of features are 80, 184, and 339 for $\alpha = \{0.9, 0.8, 0.7\}$, respectively. Based on the preliminary experiments, we set α to 0.9 in the following experiments.

Table 9. Contribution of each modal feature group for hiring decisions, extraversion, and leadership (tables show the classification accuracies of the models trained using co-occurrence feature sets excluding a specific modal feature (e.g. F_* ()). Acc. denotes the accuracy of the test data. Diff. denotes the difference of accuracy for cases in which a specific modality is removed. Bold values indicate that the difference is more than 2.0, and underlined bold values indicate that the difference is less than −2.0.)

[%]	HirDecision			
(2-1) in Table 3	Ridge		L-SVM	
	77.42		56.45	
Removed modality	Acc.	Diff.	Acc.	Diff.
F_1 (ST)	59.68	**+17.74**	38.71	**+17.74**
F_2 (PI)	69.35	**+8.07**	59.68	**-3.23**
F_3 (EN)	70.97	**+6.45**	58.06	-1.61
F_4 (VR)	75.81	+1.61	59.68	**-3.23**
F_5 (H)	70.97	**+6.45**	62.90	**-6.45**
F_6 (MT)	72.58	**+4.84**	64.52	**-8.07**
F_7 (OPM)	67.74	**+9.68**	50.00	**+6.45**
F_8 (OPV)	66.13	**+11.29**	62.90	**-6.45**
F_9 (OPA)	67.74	**+9.68**	70.97	**-14.52**

[%]	Extraversion				Perceived Leadership			
(2-1) in Table 4	Ridge		L-SVM		Ridge		L-SVM	
	70.59		58.82		68.63		73.53	
Removed modality	Acc.	Diff.	Acc.	Diff.	Acc.	Diff.	Acc.	Diff.
F_1 (ST)	65.69	**+4.90**	58.82	0.00	53.92	**+14.71**	65.69	**+7.84**
F_2 (PI)	72.55	-1.96	62.75	**-3.93**	67.65	+0.98	73.53	0.00
F_3 (EN)	71.57	-0.98	62.75	**-3.93**	69.61	-0.98	73.53	0.00
F_4 (H)	65.69	**+4.90**	64.71	**-5.88**	72.55	**-3.92**	80.39	**-6.86**
F_5 (B)	66.67	**+3.92**	58.82	0.00	65.69	**+2.94**	69.61	**+3.92**
F_6 (MT)	67.65	**+2.94**	57.84	+0.98	67.65	+0.98	73.53	0.00
F_7 (G)	76.47	**-5.88**	61.76	**-2.94**	67.65	+0.98	63.73	**+9.80**

A.3.2 Classification performance for each hyperparameter α. The classification accuracy of the model with all co-occurrence features ((2-1) in Table 4) is considered in this section. Table 8 shows the dependency of the classification accuracy of models with co-occurrence features extracted via mining when $\alpha = \{0.7, 0.8, 0.9\}$. The dependency is analyzed based on Table 5. The best accuracies for all traits are achieved by the model with co-occurrence features extracted when $\alpha = 0.9$. The second best accuracies for hiring decision, agreeableness, and leadership are achieved by these features extracted when $\alpha = 0.8$. The accuracies are 58.8%, 61.7%, 68.6% and 67.7% for hiring decision, extraversion, agreeableness and leadership.

The accuracies achieved by models with $\alpha = 0.9$ are better than those of models with $\alpha = 0.8$, with an improvement in accuracy from 5% to 11%. These results show that the extraction of co-occurrence features when the value of the threshold parameter α is lower (less than 0.9) is not effective for improving the classification accuracy. Setting α to lower values did not improve the accuracy for the other traits, only for the four traits. The experimental results show that the parameter $\alpha = 0.9$ is optimal for all traits and that we do not need to optimize the parameter for each trait in the ELEA and SONVB datasets.

A.4 Analysis of the contributions of each modality

In this section, we analyze the contribution of each modality in the audio-visual features to classify personality impressions. The classification model is trained using co-occurrence patterns by removing features of specific modalities, and it is evaluated in the same manner as Section 6.1. The contributions of specific modalities are identified by comparing the classification accuracy of the model with all co-occurrence features ((2-1) in Table 4) with the accuracy of feature sets that exclude specific modalities.

If the accuracy is degraded, then the removed feature set is effective, whereas if the accuracy is improved, then the removed feature set is unnecessary. This analysis is performed for hiring decision in SONVB and extraversion and leadership in ELEA as the representative impression variables. Table 9 shows the classification accuracies of the model using co-occurrence feature sets that exclude specific features (e.g., $F_*()$) for hiring decisions, extraversion, and leadership. In these tables, Acc. denotes the accuracy of the test data, and Diff. denotes the difference of accuracy for cases in which the modal feature set is removed. Bolded values indicate that the difference is more than 2.0, and underlined values indicate that the difference is less than −2.0.

A.4.1 Hiring Decision. From Table 9, all modalities contribute to improving the classification accuracy of the ridge regression model. In particular, the three most effective features are speech status (Diff. is +17.7%), magnitude of the vertical activity of optical flow (Diff. is +9.6%) and acceleration of the vertical activity of optical flow (Diff. is +9.6%). Many non-effective feature sets are included in the SVM because the baseline accuracy is low (56.4%). We observe that all co-occurrence features observed from both the applicant and interviewer are effective for inferring the hiring-decision score.

A.4.2 Extraversion. From Table 9, speech status (F_1 (ST)), body activity (F_5 (B)), and the wMEI feature set (F_5 (MT)) are effective for classifying the extraversion level because the difference of accuracy (Diff.) is positive or zero (i.e., not negative) in both the ridge regression model and SVM. This finding is consistent with the results in [5]. In the ridge regression model, the most effective features are speech status (F_1 (ST)) and head activity (F_4 (H)), and the difference of accuracy is +4.90%. However, pitch (F_2 (PI)), energy (F_3 (EN)), and gaze activity (VFOA) (F_7 (G)) are not effective features for classifying extraversion because the difference of accuracy (Diff.) is negative. In particular, the model with the co-occurrence feature set that excluded gaze modality obtained an accuracy of 76.4% for extraversion, which is the best result for this trait.

A.4.3 Leadership. From Table 9, speech status, body activity, pitch, gaze activity, and wMEI are effective in classifying leadership level because the difference of accuracy (Diff.) are positive or zero for both the ridge regression model and SVM. The result for the gaze feature is inconsistent with the result obtained for extraversion. The most effective features are speech status in the ridge regression model (Diff. is +14.7%) and gaze activity in the SVM (Diff. is +9.8%). The results indicate that the best performance for perceived leadership was obtained using "co-occurrence features binary," which includes the gaze features in Table 4. Head activity is not effective for the classification of leadership (Diff. is −6.8% in SVM). The model with the feature set excluding head activity obtained an accuracy of 80.39%, which is the best result for leadership.

ACKNOWLEDGMENTS

We appreciate the support of the Swiss National Science Foundation (SNSF), through the UBImpressed Sinergia project (CRSII2 147611) and the SOBE Ambizione Fellowship (PZ00P2 136811), and the Japan Society for the Promotion of Science (JSPS) KAK-ENHI (25730132, 15K00300, 25280076).

REFERENCES

[1] X. Alameda-Pineda, J. Staiano, R. Subramanian, L. Batrinca, E. Ricci, B. Lepri, O. Lanz, and N. Sebe. 2016. SALSA: A Novel Dataset for Multimodal Group Behavior Analysis. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 38, 8 (2016), 1707–1720.

[2] Xavier Alameda-Pineda, Yan Yan, Elisa Ricci, Oswald Lanz, and Nicu Sebe. 2015. Analyzing Free-standing Conversational Groups: A Multimodal Approach. In *Proceedings of the 23rd ACM International Conference on Multimedia (MM '15)*. 5–14.

[3] James F. Allen. 1983. Maintaining Knowledge About Temporal Intervals. *Commun. ACM* 26, 11 (Nov. 1983), 832–843.

[4] Oya Aran, Joan-Isaac Biel, and Daniel Gatica-Perez. 2014. Broadcasting Oneself: Visual Discovery of Vlogging Styles. *IEEE Trans. Multimedia* 16, 1 (2014), 201–215.

[5] Oya Aran and Daniel Gatica-Perez. 2013. One of a Kind: Inferring Personality Impressions in Meetings. In *Proceedings of the 15th ACM on International Conference on Multimodal Interaction (ICMI '13)*. 11–18.

[6] Javed Aslam, Katya Pelekhov, and Daniela Rus. 1999. A Practical Clustering Algorithm for Static and Dynamic Information Organization. In *Proceedings of the Tenth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA '99)*. 51–60.

[7] Umut Avci and Oya Aran. 2016. Predicting the Performance in Decision-Making Tasks: From Individual Cues to Group Interaction. *IEEE Trans. Multimedia* 18, 4 (2016), 643–658.

[8] L. Batrinca, N. Mana, B. Lepri, N. Sebe, and F. Pianesi. 2016. Multimodal Personality Recognition in Collaborative Goal-Oriented Tasks. *IEEE Trans. on Multimedia* 18, 4 (2016), 659–673.

[9] S. S. Beauchemin and J. L. Barron. 1995. The Computation of Optical Flow. *ACM Comput. Surv.* 27, 3 (Sept. 1995), 433–466.
[10] Joan-Isaac Biel, Lucía Teijeiro-Mosquera, and Daniel Gatica-Perez. 2012. FaceTube: predicting personality from facial expressions of emotion in online conversational video. In *Proceedings of the 14th ACM international conference on Multimodal interaction*. ACM, 53–56.
[11] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research* 3, Jan (2003), 993–1022.
[12] Leo Breiman. 2001. Random Forests. *Machine Learning* 45, 1 (2001), 5–32.
[13] Moitreya Chatterjee, Sunghyun Park, Louis-Philippe Morency, and Stefan Scherer. 2015. Combining Two Perspectives on Classifying Multimodal Data for Recognizing Speaker Traits. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction (ICMI '15)*. 7–14.
[14] Jared R Curhan and Alex Pentland. 2007. Thin slices of negotiation: predicting outcomes from conversational dynamics within the first 5 minutes. *Journal of Applied Psychology* 92, 3 (2007), 802.
[15] Daniel Gatica-Perez. 2009. Automatic nonverbal analysis of social interaction in small groups: A review. *Image Vision Computing* 27, 12 (nov 2009), 1775–1787.
[16] Samuel D. Gosling, Peter J. Rentfrow, and William B. Swann. 2003. A very brief measure of the Big-Five personality domains. *Journal of Research in Personality* 37 (2003), 504–528.
[17] Geoffrey E Hinton and Ruslan R Salakhutdinov. 2006. Reducing the dimensionality of data with neural networks. *Science* 313, 5786 (2006), 504–507.
[18] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.
[19] Harold Hotelling. 1936. Relations between two sets of variates. *Biometrika* 28, 3/4 (1936), 321–377.
[20] Hayley Hung and Ben Kröse. 2011. Detecting F-formations As Dominant Sets. In *Proceedings of the 13th International Conference on Multimodal Interfaces (ICMI '11)*. 231–238.
[21] Dineshbabu Jayagopi, Dairazalia Sanchez-Cortes, Kazuhiro Otsuka, Junji Yamato, and Daniel Gatica-Perez. 2012. Linking Speaking and Looking Behavior Patterns with Group Composition, Perception, and Performance. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction (ICMI '12)*. 433–440.
[22] Oliver P John and Sanjay Srivastava. 1999. The Big Five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of personality: Theory and research* 2, 1999 (1999), 102–138.
[23] V. Kumar, A. Namboodiri, and C. V. Jawahar. 2015. Visual Phrases for Exemplar Face Detection. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*. 1994–2002.
[24] Yann LeCun and Yoshua Bengio. 1998. The Handbook of Brain Theory and Neural Networks. MIT Press, Chapter Convolutional Networks for Images, Speech, and Time Series, 255–258.
[25] Anmol Madan, Ron Caneel, and Alex Pentland. 2004. Voices of attraction. In *Proceedings of International Conference on Augmented Cognition*.
[26] Héctor P. Martínez and Georgios N. Yannakakis. 2011. Mining Multimodal Sequential Patterns: A Case Study on Affect Detection. In *Proc. of ACM ICMI*. 3–10.
[27] Héctor P. Martínez and Georgios N. Yannakakis. 2014. Deep Multimodal Fusion: Combining Discrete Events and Continuous Signals. In *Proceedings of the 16th International Conference on Multimodal Interaction (ICMI '14)*. 34–41.
[28] Chreston Miller, Louis-Philippe Morency, and Francis Quek. 2012. Structural and Temporal Inference Search (STIS): Pattern Identification in Multimodal Data. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction (ICMI '12)*. 101–108.
[29] I. Naim, M. I. Tanveer, D. Gildea, and M. E. Hoque. 2015. Automated prediction and analysis of job interview performance: The role of what you say and how you say it. In *Proceedings of 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, Vol. 1. 1–6.
[30] Yukiko I. Nakano, Sakiko Nihonyanagi, Yutaka Takase, Yuki Hayashi, and Shogo Okada. 2015. Predicting Participation Styles Using Co-occurrence Patterns of Nonverbal Behaviors in Collaborative Learning. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction (ICMI '15)*. 91–98.
[31] L.S. Nguyen, D. Frauendorfer, M.S. Mast, and D. Gatica-Perez. 2014. Hire me: Computational Inference of Hirability in Employment Interviews Based on Nonverbal Behavior. *IEEE Trans. on Multimedia* 16, 4 (2014), 1018–1031.
[32] Laurent Nguyen, Jean-Marc Odobez, and Daniel Gatica-Perez. 2012. Using Self-context for Multimodal Detection of Head Nods in Face-to-face Interactions. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction (ICMI '12)*. 289–292.
[33] Fumio Nihei, Yukiko I. Nakano, Yuki Hayashi, Hung-Hsuan Hung, and Shogo Okada. 2014. Predicting Influential Statements in Group Discussions Using Speech and Head Motion Information. In *Proceedings of the 16th International Conference on Multimodal Interaction (ICMI '14)*. 136–143.

[34] Jean-Marc Odobez and Patrick Bouthemy. 1995. Robust multi resolution estimation of parametric motion models. *Journal of visual communication and image representation* 6, 4 (1995), 348–365.
[35] Shogo Okada, Oya Aran, and Daniel Gatica-Perez. 2015. Personality Trait Classification via Co-Occurrent Multiparty Multimodal Event Discovery. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction (ICMI '15)*. 15–22.
[36] Shogo Okada, Mi Hang, and Katsumi Nitta. 2016. Predicting Performance of Collaborative Storytelling Using Multimodal Analysis. *IEICE Transactions* 99-D, 6 (2016), 1462–1473.
[37] S. Park, S. Scherer, J. Gratch, P. J. Carnevale, and L. P. Morency. 2015. I Can Already Guess Your Answer: Predicting Respondent Reactions during Dyadic Negotiation. *IEEE Trans. on Affective Computing* 6, 2 (2015), 86–96.
[38] Fabio Pianesi, Nadia Mana, Alessandro Cappelletti, Bruno Lepri, and Massimo Zancanaro. 2008. Multimodal Recognition of Personality Traits in Social Interactions. In *Proceedings of the 10th International Conference on Multimodal Interfaces (ICMI '08)*. 53–60.
[39] X. Qian, H. Wang, Y. Zhao, X. Hou, R. Hong, M. Wang, and Y. Y. Tang. 2017. Image Location Inference by Multisaliency Enhancement. *IEEE Trans. on Multimedia* 19, 4 (2017), 813–821.
[40] Rutger Rienks and Dirk Heylen. 2006. Dominance Detection in Meetings Using Easily Obtainable Features. In *Proceedings of the Second International Conference on Machine Learning for Multimodal Interaction (MLMI'05)*. 76–86.
[41] Dairazalia Sanchez-Cortes, Oya Aran, Dinesh Babu Jayagopi, Marianne Schmid Mast, and Daniel Gatica-Perez. 2013. Emergent leaders through looking and speaking: from audio-visual data to multimodal recognition. *Journal on Multimodal User Interfaces* 7, 1-2 (2013), 39–53.
[42] Dairazalia Sanchez-Cortes, Oya Aran, Marianne Schmid Mast, and Daniel Gatica-Perez. 2012. A nonverbal behavior approach to identify emergent leaders in small groups. *IEEE Trans. on Multimedia* 14, 3 (2012), 816–832.
[43] S. Scherer, G. Stratou, M. Mahmoud, J. Boberg, J. Gratch, A. Rizzo, and L. P. Morency. 2013. Automatic behavior descriptors for psychological disorder analysis. In *Proceedings of 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. 1–8.
[44] Yale Song, Louis-Philippe Morency, and Randall Davis. 2012. Multimodal Human Behavior Analysis: Learning Correlation and Interaction Across Modalities. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction (ICMI '12)*. 27–30.
[45] L. Teijeiro-Mosquera, J. I. Biel, J. L. Alba-Castro, and D. Gatica-Perez. 2015. What Your Face Vlogs About: Expressions of Emotion and Big-Five Traits Impressions in YouTube. *IEEE Trans. on Affective Computing* 6, 2 (2015), 193–205.
[46] Alireza Vahdatpour, Navid Amini, and Majid Sarrafzadeh. 2009. Toward Unsupervised Activity Discovery Using Multi-dimensional Motif Detection in Time Series. In *Proceedings of the 21st International Jont Conference on Artifical Intelligence (IJCAI'09)*. 1261–1266.
[47] Alessandro Vinciarelli. 2007. Speakers role recognition in multiparty audio recordings using social network analysis and duration distribution modeling. *IEEE Trans. on Multimedia* 9, 6 (2007), 1215–1226.
[48] A. Vinciarelli, A. Dielmann, S. Favre, and H. Salamin. 2009. Canal9: A database of political debates for analysis of social interactions. In *Proceedings of the 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*. 1–4.
[49] Zhong Wu, Qifa Ke, M. Isard, and Jian Sun. 2009. Bundling features for large scale partial-duplicate web image search. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 25–32.
[50] Jian Bo Yang, Minh Nhut Nguyen, Phyo Phyo San, Xiao Li Li, and Shonali Krishnaswamy. 2015. Deep Convolutional Neural Networks on Multichannel Time Series for Human Activity Recognition. In *Proceedings of the 24th International Conference on Artificial Intelligence (IJCAI'15)*.
[51] X. Yang, X. Qian, and Y. Xue. 2015. Scalable Mobile Image Retrieval by Exploring Contextual Saliency. *IEEE Trans. on Image Processing* 24, 6 (2015), 1709–1721.
[52] Massimo Zancanaro, Bruno Lepri, and Fabio Pianesi. 2006. Automatic Detection of Group Functional Roles in Face to Face Interactions. In *Proceedings of the 8th International Conference on Multimodal Interfaces (ICMI '06)*. 28–34.
[53] S. Zhang, Q. Tian, G. Hua, Q. Huang, and W. Gao. 2011. Generating Descriptive Visual Words and Visual Phrases for Large-Scale Image Applications. *IEEE Trans. on Image Processing* 20, 9 (2011), 2664–2677.
[54] Shiliang Zhang, Ming Yang, Xiaoyu Wang, Yuanqing Lin, and Qi Tian. 2015. Semantic-Aware Co-Indexing for Image Retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.* 37, 12 (2015), 2573–2587.
[55] Guoshuai Zhao, Xueming Qian, Xiaojiang Lei, and Tao Mei. 2016. Service Quality Evaluation by Exploring Social Users' Contextual Information. *IEEE Trans. on Knowl. and Data Eng.* 28, 12 (2016), 3382–3394.

Received February 2007; revised March 2009; accepted June 2009