A Gaze-based Method for Relating Group Involvement to Individual Engagement in Multimodal Multiparty Dialogue

Catharine Oertel
KTH Royal Institute of Technology
Linstedtsvägen 44
Stockholm, Sweden
catha@kth.se

Giampiero Salvi
KTH Royal Institute of Technology
Linstedtsvägen 44
Stockholm, Sweden
giampi@kth.se

ABSTRACT

This paper is concerned with modelling individual engagement and group involvement as well as their relationship in an eight-party, mutimodal corpus. We propose a number of features (presence, entropy, symmetry and maxgaze) that summarise different aspects of eye-gaze patterns and allow us to describe individual as well as group behaviour in time. We use these features to define similarities between the subjects and we compare this information with the engagement rankings the subjects expressed at the end of each interactions about themselves and the other participants. We analyse how these features relate to four classes of group involvement and we build a classifier that is able to distinguish between those classes with 71% of accuracy.

Categories and Subject Descriptors

H5.3 [**Information Interfaces and Presentation**]: Group and Organisation Interfaces—*Theory and models*;
J.4 [**Computer Applications**]: Social and Behavioural Sciences—*Psychology, Sociology*;
I.5.4 [**Pattern Recognition Design Methodology**]: Feature evaluation and selection

Keywords

interaction, gaze, engagement, involvement, classification

1. INTRODUCTION

One feature of human-human conversation is that we are able to react to each other. Such a reaction can be based on the interpretation of verbal cues, but is very often also influenced by, or even based on non-verbal cues. Non-verbal cues can be for example smiles, nods or direction of eye-gaze. Their interpretation is particularly crucial when trying to spot irony or interest in our interlocutors, yet reacting onto them also adds to the flow of the conversation in general. Constant monitoring is in fact very important for a successful conversation as for example interest can spark as quickly as it can be lost again, and a prompt reaction on the part of the interlocutor would probably be beneficial for the further course of the conversation. This process of sending, receiving and interpreting cues is quite complex in a dyadic conversation, but gets even more complex in a multiparty conversation. Here, several people are sending and interpreting cues simultaneously, which makes it much harder for the current speaker to keep track of everybody and to react to them. Modelling of the non-verbal exchange between interlocutors in a group and their influence on the general conversational dynamics is, however, crucial in order to make current dialogue systems more adaptive and for the development of software which could aid people diagnosed with autism, for example.

A fair amount of work has been carried out to quantify the individual engagement, interest and/or role of the individual participant in a group as, for example, described in [16, 17, 18, 14] and to a lesser extent also the involvement of a group of participants [5, 13]. First steps have also been undertaken in investigating individual engagement in human-machine dialogue [6] and implementing models of involvement into dialogue systems [15, 2].

To date, very few papers (e.g., [9]) have been concerned with quantifying the relation between group involvement and individual engagement. In order to fully understand group dynamics and make use of this information it is important to relate the engagement measures for the individual person to the general group involvement.

In the current paper we attempt to model this relation in the following way. We define a number of features based on eye-gaze patterns and aimed at describing different aspects of individual engagement and group involvement. In the first part of the analysis, we use these features to find clusters of subjects sharing similar behaviour. We then compare these emerging patterns with those obtained with rank-annotation expressed by the same subjects about themselves and the other participants.

In the second part of the study we analyse the relation between features for individual engagement and those for group involvement and observe the evolution of these in time.

In the last part of the study we test whether it is possible to distinguish between four distinct classes of group involvement using our feature set. We do this both using standard statistical tests as well as carrying out classification experiments.

2. BACKGROUND

Gatica-Perez et al. [5] approach the problem of detecting group involvement or interest levels, as they refers to

ICMI'13, December 9–13, 2013, Sydney, Australia
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-2129-7/13/12 ...$15.00.
http://dx.doi.org/10.1145/2522848.2522865.

them, by asking annotators to annotate 15 second intervals of group involvement in a four party dialogue on a 5 point scale. They define group involvement as "the perceived degree of interest or involvement of the majority of the group". They then extract both speech and visual features of a multimodal corpus and use HMMs in order to detect segments of high and neutral group interest level. McCowan et al. [9] put both the group involvement and individual engagement in relationship to each other by using HMMs. They use both visual and acoustic features, yet no eye-gaze. They train and test their model on a meeting corpus with four participants.

Similarly to Gatica-Perez et al. [5], Oertel et al. [13] carried out a prediction experiment in order to detect involvement in a multimodal corpus recording. Different from [5], however, they used a 10 point annotation scheme for 5 second segments, as described in Oertel et al. [11], and compared the prediction accuracies when trying to predict 2 (high and low) and three (high middle and low) classes of involvement. They also added an analysis of eye gaze and found that the proportionate amount of mutual gaze is a good cues for the prediction of involvement and at least for the three class problem outperformed the audio model.

Bednarik and Hradis [1], also investigated conversational engagement in multiparty video conversation, however their emphasis lay on the analysis of gaze. They investigate how 6 different levels of engagement (no interest, following, responding, conversing, influencing, managing) annotated by at least two annotators for a 15 second intervals, are related to different gaze patterns. Different from [5], their annotation scheme describes the individual person within a group rather than describing the group as a whole. The authors distinguished between number of switches between faces, number of unique faces attended and number of faces attended. Finally they used a classifier to distinguish between low and high levels of engagement.

Levitski et al.[8] carried out an analysis on a three-party conversation corpus. In their set-up they tracked the eye-gaze of one participant. Two further participants were placed opposite this participant. One of them received to task to remain silent; he was however allowed to produce non-verbal signals and to follow the conversation actively. The remaining participant was allowed to talk normally. Engagement annotations were carried out for the silent participant. Here, a binary distinction is made between "engaged" and "passive". The state "Engaged" was annotated when the person actively gazes at the other participants . "Passive" was annotated when the person's gaze and gesturing indicate that he/she is not involved in the conversation to the same degree. They found that the participants gaze is more focused on the background when the silent interlocutor is perceived as "passive".

Also Bonin et al. [3] propose an annotation scheme which takes into account both the group and the individual. They propose to not predefine chunks in which annotators are supposed to give ratings but rather let the annotators decide where they see changes in involvement. They also rely on the annotators intuition to come up with a definition of involvement and do not define annotation instructions or a definition of involvement.

Our work differs and in part extends the discussed studies in the following way. First of all, our analysis is based on a eight party conversation rather than a 4 or 5 party conversation. Second, we propose a new set of features based on eye-gaze data which are different from the features used in [13, 5, 1] and [8]. Third, differently from [3] and [8], our individual engagement annotations are based on the rankings of the participants themselves and we are using a predefined annotation scheme rather than relying solely on the third-party annotators' intuitions. However, similarly to [3], we are not using fixed window length for the group involvement annotations.

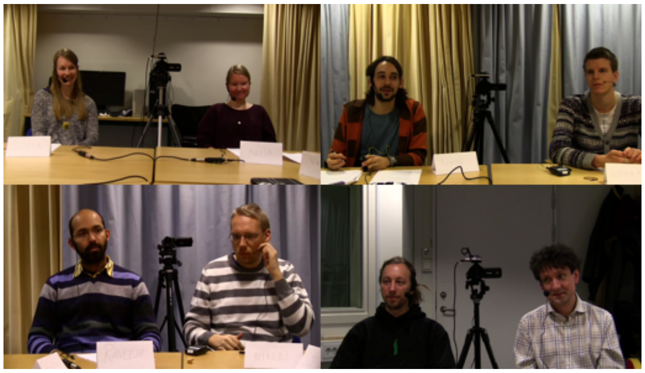


Figure 1: Setting of Werewoolf Corpus

3. THE CORPUS

In the current study we are using the Stockholm Werewolf Corpus (SWC), which is part of a bigger collection of game based interactions called the KTH Games Corpora, and will be released under the Creative Commons Attribution - Noncommercial-Share Alike License. The SWC currently consists of a set of four sessions and three rounds in which participants are engaged in playing the "Werewolf" game (for a more detailed description see Section 3.1). One round ranges between 15-20 minutes depending on the dynamics and involvement of the group of participants. So far a total of one hour of game sessions has been recorded but further recordings are underway. In the current study we are only using sections of round 1 and round 3. Those sections were chosen as, unlike in the other rounds, the role of the game master was played by two different participants.

There were eight participants in total of which two were female and six were male. There were four native speakers of Swedish, one native speaker of Hindu, one German, one Finnish and one Bulgarian. All participants had a high proficiency in English. Some were friends, some colleagues and some had never met each other before. Yet everybody knew at least two more participants before commencing with the recordings. Each participant participated in each session but were assigned new roles in each session of the game. The recordings were carried out in English. Participants were seated around a table and were recorded by 4 cameras. Each camera captured two participants each (see figure 1). The cameras recorded at a frame rate of 25 frames per second.

3.1 The Werewolf Game

The Werewolf Game is a role-playing game for 8 or more participants. One of the participants takes on the role of a game master who leads the participants through the different rounds of the game, each consisting of a "night" phase, in which the participants are asked to close their eyes, and a "day" phase. The role assigned to each of the other participants is only known to that participant. Among the different roles, two participants are "werewolves", all others are simple "villagers". Among the villagers, two participants are "lovers".

Werewolves and villagers try to win the game by eliminating participants from the game. The werewolves by killing villagers during the night phase of the game, the villagers by trying to guess who is a werewolf and killing their choice during the day phase of the game.

The werewolves know each other's identity and collaborate in killing another participant, but for an individual werewolf there is no direct consequence if the second werewolf dies. Among the villagers, the two lovers try to protect one another and both die if one of them gets killed. Both collaborating pairs (werewolves and lovers) are hiding their collaboration from the other participants.

During the day phase of the game, all participants are trying to convince the others of their innocence and make hypotheses about who might be a werewolf. The choice is made by voting, lead by the game master. If there is a tie, one participant, who has been appointed mayor in the beginning, can break the tie. For more information on the corpus please refer to [12].

3.2 Annotations

Sections of SWC were annotated for both speech activity as well as eye-gaze. Eye-gaze annotations were carried out on a frame-by-frame basis. For each participant and for each frame in time, the annotations indicate the target of the gaze. Possible targets are any of the other participants as well as other objects in the visual vicinity such as a sheet of paper with the game instructions. Other targets are "eyes closed " and "down". In addition to this distinction further eye-gaze targets such as "looking at the sheet" were annotated.

The SWC has also been annotated for Group involvement. A distinction between four different classes was made. The classes are as follows:

Class "high" is annotated when the group displays a high level of involvement. This could be compared to the verbal involvement levels 7 to 9 in the annotation scheme described in [11]. Class "low" is annotated when the group displays a low level of involvement. This can be mapped to involvement levels 1-3 in [11]. Class "lead" is annotated when the game leader is steering the conversation. This should be seen as a minimal category which is only chosen when participants are not actively taking initiatives themselves. This category could be mapped to involvement level 4. Category "org" is annotated when the group is forming itself. It is the period before the actual start of the game in which participants can get to know each other and questions can be asked. It was included in the corpus and subsequently also into the analysis in order to have an example of more non-task directed conversation. Class "org" is annotated for round 1 but does not exist in round 3.

Group involvement annotations were carried out by three annotators. Each annotator annotated the corpus first separately. Afterwards annotations were compared and a consensus on which label to use was reached. There are only two instances in the corpus were no consensus could be reached (the 4th "lead" in round 1 and the first "low" in round 3). In this case the annotation of the most experienced annotator was used. These annotations are used for the ANOVA analyses as well as the classification experiments.

In addition to the group involvement rating participants in the game also rated themselves and each other in terms of individual engagement. In order to make these ratings as unobtrusive as possible for the game, participants were asked to carry out the rating after each "day" phase and just before the "night" phases. These annotations were used in the rank based clustering experiments.

4. METHOD

For each round of the game, we analyse the gaze data at regular time intervals corresponding to the video frame rate (25 fps). From the annotations, we construct, for each time instant t, a binary $N \times M$ matrix $G(t)$ with elements $g_{ij}(t)$ indicating whether subject i is looking at target j at time t. Among the M targets, the first N correspond to the subjects themselves, and the remaining $M - N$ are respectively "away", "down", "sheet" and "eyes closed"[1]. In this study we will call the latter "other targets". Because the annotations assign the gaze of each subject to only one possible target at any time instant, we have that:

$$\sum_{i=1}^{N}\sum_{j=1}^{M} g_{ij}(t) = N, \quad \forall t$$

Also, because the subjects cannot gaze at themselves,

$$g_{ii} = 0, \quad \forall i \in [1, N]$$

In Figure 2 four different examples of possible values for matrix $G(t)$ are given. Note that the dashed vertical line separates subject targets from other targets. These examples will be used in this section to illustrate the method.

Similarly, a binary vector $S(t)$ of length N was created from the speech annotations, with each element $s_i(t)$ indicating for each time step t whether speaker i was talking or silent. Contrary to the gaze matrix, no constraints can be imposed on the elements of this vector, as any of the N speakers can be talking simultaneously or they may all be silent.

From the above representations, we compute two kinds of measures: Measures of the first kind attempt to describe the general group behaviour at each time instant. We call these measures "group features". Measures of the second kind, instead, try to describe the behaviour of individual subjects and they are computed integrating the gaze information over time.

4.1 Group Features

The aim of these features is to describe the kind of interaction that is taking place at every time instant, disregarding the identity of the subjects involved. We want, e.g., to be able to distinguish between interactions that involve the whole group, from interactions that happen between subgroups of subjects.

The most simple measure, called **presence** in this study, is the fraction of subjects that are looking at other subjects

[1] The original annotations include more detailed target that we decided to collapse into these four categories

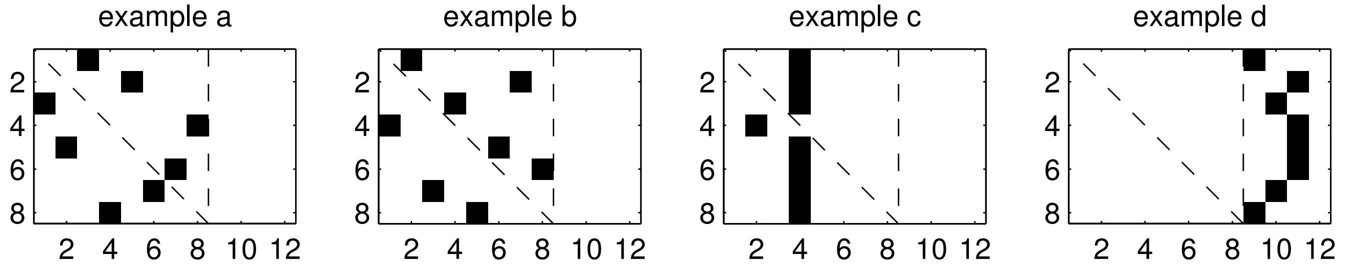


Figure 2: Examples of instantaneous gaze patterns (matrix $G(t)$). Rows indicate subjects and columns targets of the gaze. The vertical dashed line separates subject targets from other targets.

as opposed to other targets. This is simply defined by:

$$f_p(t) = \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} g_{ij}(t)}{N} \quad (1)$$

where we only sum over the first N targets. Examples a-c in Figure 2 correspond to $f_p(t) = 1$, whereas example d corresponds to $f_p(t) = 0$.

Although the above feature grasps a general involvement in the interaction, it does not model the kind of interaction that is taking place. The following features try to be more specific.

First of all, we want to distinguish between the cases where the gazes are mutual or not. Mutual gazes may indicate that the subjects are involved in interactions in subgroups (pairs). The way we measure this is by estimating the degree of **symmetry** of the square part of matrix $G(t)$:

$$f_s(t) = \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} g_{ij}(t) g_{ji}(t)}{N} \quad (2)$$

If the matrix is symmetric (example a, Figure 2), the double sum in Eqn. 2 is equal to N and $f_s(t) = 1$. If, on the contrary, there is no symmetry in the matrix (example b in the figure), the double sum and $f_s(t)$ are equal to 0. This is also the case if no subject is looking at other subjects as in example d. In example c, $f_s(t) = \frac{2}{N}$ because only one pair of subjects have mutual gaze.

Another interesting aspect to model is if several subjects are looking at the same target at a certain time instant. This may indicate that a global interaction is taking place. We define two features that attempt to grasp this, the first, called **maxgaze**, is simply the maximum fraction of subjects looking at the same target and is defined as

$$f_m(t) = \frac{\max_{j \in [1,M]} \sum_{i=1}^{N} g_{ij}(t)}{N} \quad (3)$$

Note that $f_m(t)$ can only reach the value of 1 if the target is among the "other targets". If the target is a subject, only the other $N-1$ subjects can gaze at it at the same time, and $f_m(t) = \frac{N-1}{N}$ (see example c in Figure 2).

The last gaze-based feature, called **entropy**, interprets the fraction of subjects looking at each targets as probabilities that any subject would look at that particular target:

$$P(\text{target} = j|t) = \frac{\sum_{i=1}^{N} g_{ij}(t)}{N}$$

and then computes the normalised entropy of this probability distribution:

$$f_e(t) = -\frac{\sum_{j=1}^{M} P(\text{target} = j|t) \log(P(\text{target} = j|t))}{\log(M)}$$

If each subject is looking at a different target, the distribution is uniform and the entropy is maximum ($f_e(t) = 1$, examples a and b in Figure 2), if all the subjects are looking at the same target, the entropy is minimum ($f_e(t) = 0$). Example c in Figure 2 is an example of very low entropy.

The above four features may also be computed over a sequence of subsequent observations by first averaging the $G(t)$ matrices in a window in time. This results in smoother feature values.

From the speech data, two kinds of features were calculated. Considering the vectors $S(t)$ over a time window of length W, we first computed the ratio of silent time steps, i.e., time steps when no one was speaking over the length W of the time window (we call this feature **silence**). For the remaining time steps, we calculate the average number of speakers talking simultaneously. We expected these two measures to give us an indication of the global involvement in the conversation.

4.2 Subject-specific Features

The features defined in the previous section disregard the identity of the subject and target of the gaze. In this section we define measures that aim at discriminating between the behaviour of different subjects. In order to do this, we average the gaze matrices $G(t)$ over a time window of W observations and obtain an $N \times M$ matrix $\bar{G}(t)$ as:

$$\bar{G}(t) = \frac{\sum_{\theta=t}^{t+W} G(\theta)}{W}$$

In this matrix the elements $\bar{g}_{ij}$ indicate the fraction of observations in which subject i was looking at target j in that particular time window, and can also be interpreted as probabilities:

$$\bar{g}_{ij} = P(\text{subject} = i, \text{target} = j)$$

where we have omitted the time dependency for simplicity.

This allows us to define subject dependent features similar to the ones introduced in the previous section. In particular we can define the subject dependent **entropy** and **presence** features by considering single rows of the $\bar{G}(t)$ matrices.

Furthermore, we can compare the behaviour of different subjects, e.g., by computing the pair-wise distance between the subject-specific gaze distributions. A good metric between two probability distributions P and Q is the Jensen-Shannon divergence $JSD(P||Q)$ [4]. This is a symmetric version of the more popular Kullback–Leibler divergence $D(P||Q)$ [7] and is defined as:

$$JSD(P||Q) = \frac{1}{2}D(P||M) + \frac{1}{2}D(Q||M)$$

where $M = \frac{1}{2}(P + Q)$.

5. EXPERIMENTS

We considered five minutes from Round 1 and 3 of the SWC described in Section 3. The group features and subject-specific features described in Section 4 were evaluated at a rate of one second by averaging the $G(t)$ matrices over windows of 25 observations. The resulting features are displayed in Figure 3. These features can be used to relate individual subject engagement to group involvement on a time dependent fashion.

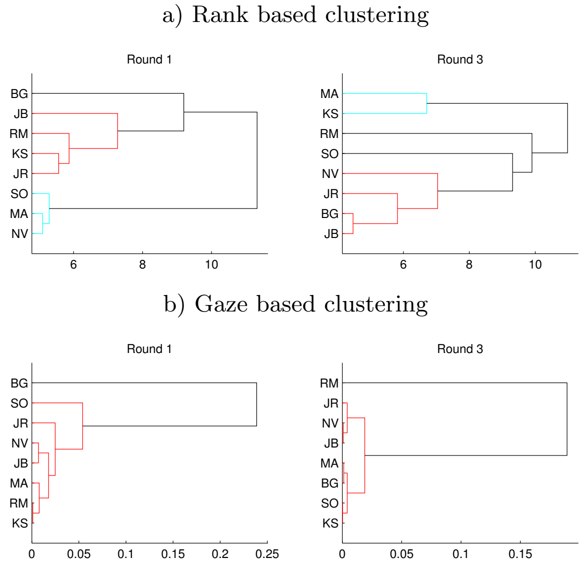
We also computed subject specific features integrating over all the five minutes of each round and used them to describe global subject engagement in the round. This was done by displaying each subject on a **entropy/presence** plane as in Figure 5, or by computing a pairwise Jensen-Shannon divergence between gaze distributions and performing hierarchical clustering with average linkage (see Figure 4b). Note that including the gaze patterns towards the individual subjects, will give results that are strongly dependent on the place each subject occupies around the table. This is the case because each subject has a higher probability to look at the subjects across the table than the ones on the sides. In order to avoid this bias, to compute the dendrograms in Figure 4 we merged all gazes to other subjects into one bin of the distribution.

The second set of experiments were performed to predict the group involvement category. Firstly, we ran a one way ANOVA with subsequent Tukey-Kramer multiple comparison analysis to relate each group feature to each category. Results are in Figure 6 and are discussed in Section 6.

We also run a Gaussian Mixture classifier trained on the group features in Round 1 of the game and tested on both Round 1 and Round 3. Each category was modelled by a 2 component GMM with full covariance matrix. The confusion matrix is shown in Table 1 and discussed in Section 6.

6. RESULTS

As evident from Figure 3, different subjects behave differently in the interactions, due to their idiosyncrasies as well as their role in the game. The first set of results aim at describing these differences.


a) Rank based clustering

b) Gaze based clustering

Figure 4: Rank and Gaze Based Subject Clustering

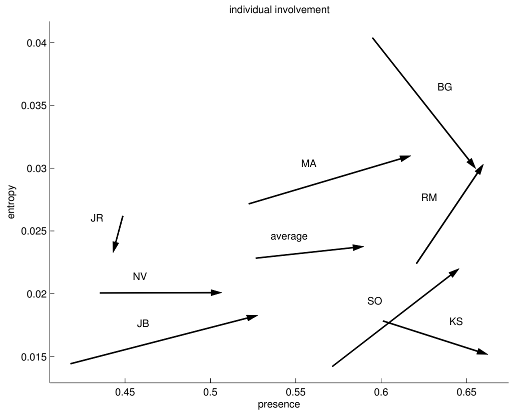


Figure 5: Entropy/presence plot for each subject from Round 1 to Round 3

6.1 Comparing Individual Subjects

Figure 4 compares the difference between the subjects' behaviour as expressed by the subjects' rankings (a) and the gaze patterns (b). In Figure 4 (a) the rankings each subject expressed about themselves and the other participants at the end of each round, are used to define a pairwise distance measure that in turns is used to perform the clustering. In Round 1 three clusters emerge, whereas four clusters emerge in Round 3. The compositions of these clusters is different in the two rounds which can be expected given the different roles the subjects play in the game. In both cases, the game master (BG and RM, respectively) stand alone when it comes to involvement rankings. It has to be noted here, however, that the assignment of a cluster to the game master by himself is not due to high engagement ratings. In both rounds other participants received higher ranks than the respective game master. It is also interesting to note that, in accordance with [13], the emergence of more than two clusters suggests that involvement and engagement should not be treated as binary, but, rather, as gradually varying phenomena.

Figure 4 (b) displays the clustering obtained from the gaze patterns as explained in Section 4 and 5. Also in this case, the game master stands alone. A second, perhaps more interesting similarity is that there seem to be more distinct clusters in Round 1 than in Round 3. This has already been the case for the clustering based on the involvement rankings and could be interpreted in the way that participants differed to greater degree in terms of involvement in Round 1 than in Round 3.

The above analysis emphasises differences between subjects, but does not indicate which subject is more or less involved in the interaction. In order to display this, in Figure 5 we plotted each subject on a **entropy/presence** plane. The

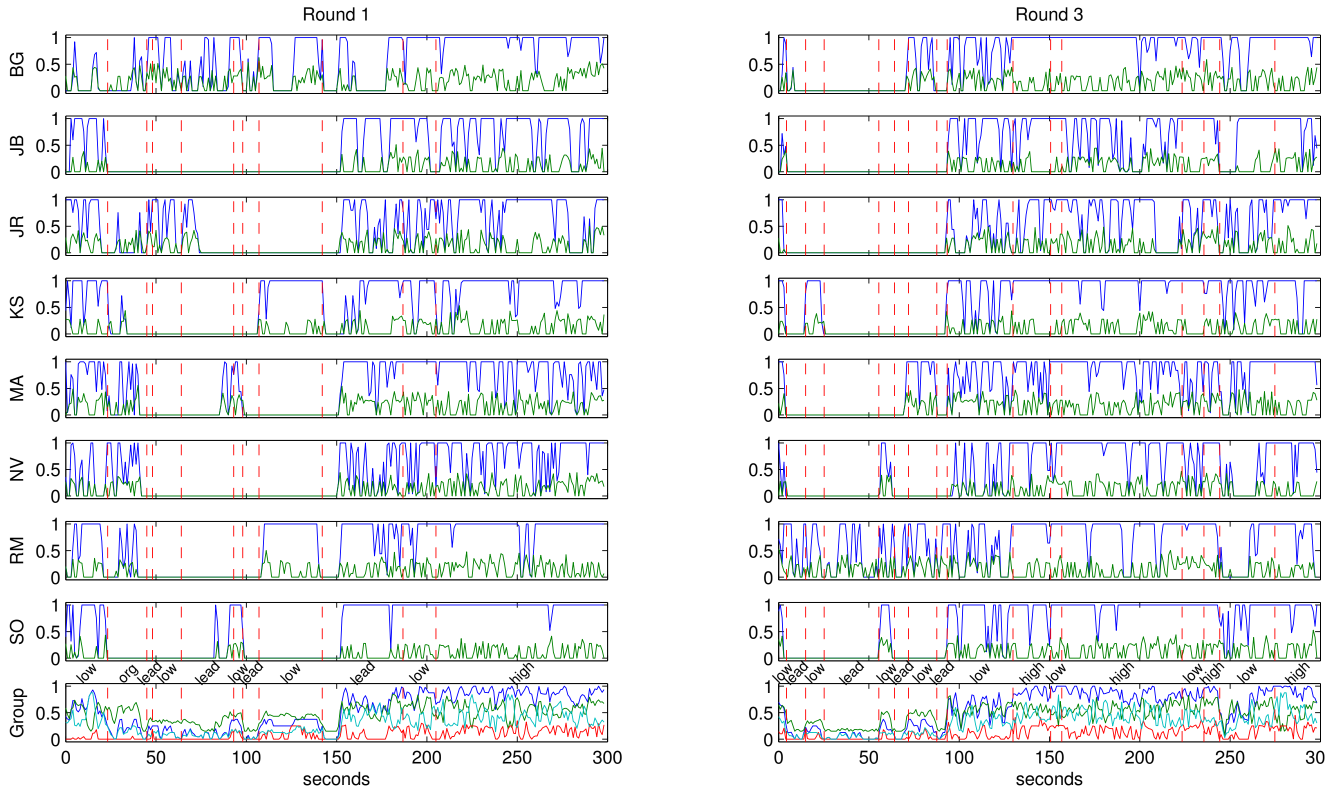


Figure 3: Subject-specific and Group Features computed on the first and third round of the game on 25 sample (1 second) windows. Blue: presence, green: entropy, red: symmetry, cyan: maxgaze.

arrows in the figure show how the involvement of each subject changes from Round 1 to Round 3. The average involvement is also plotted in the figure.

It can be noted from the average that the general trend is an increase in presence from Round 1 to Round 3. Only one subject (JR) is an exception to this trend. Of particular interest is the entropy measurements for participants BG, RM as well as KS, MA and JR. Participant BG was game master in Round 1 and RM took on this role in Round 3. Between the rounds, the entropy of BG's gaze patterns drops whereas RM's increases, indicating that the game master tends to spread gazes across different targets more than other participants.

Similarly results can be reported for subjects KS and MA. KS was in the role of the werewolf in Round 1 and in the role of normal villagers in Round 3 and he decrease in terms of entropy. On the contrary, subject MA, who was normal villager in Round 1 and werewolf in Round 3, has an increasing entropy level.

It can therefore be concluded that participants which fulfill special roles in the game, and therefore should be more involved than other participants, also show a higher level of entropy. This indicates that while presence seems to be the best feature to estimate group involvement, entropy seems to be the best feature to display individual engagement.

6.2 Group Involvement

Figure 6 displays the boxplots of the involvement features presence, entropy, symmetry and maxgaze for different classes of involvement for Round 1 and 3 of the corpus. In Round 1, one additional class ("org") was annotated. This class encompasses segments in the corpus where additional people were in the room, and organisational issues were discussed.

According to the ANOVA analysis, all plots include significant differences with very low p values (p varies between 0 and 4.3e-51, F between 85.6 and 1.6e+03). A subsequent Tukey-Kramer multiple comparison analysis shows that all but one group differences are significant. The only difference that is not significant is in Round 1, **entropy** between group "low" and "org"

Both in Round 1 and Round 3 the feature **presence** has lower levels for class "low" (low involvement) than class "high" (high involvement). Class "lead" and "org", which is typically annotated when the game master steers the conversation, have the lowest values of **presence**. This means that participants are looking more at other people than at non-target objects, such as for example the table, or a distant point at the wall. In comparison to any other feature, presence is the feature which is able to distinguish best between low and high involvement.

The feature **entropy**, on the other hand, can discriminate well between class "lead" and the other classes. The fact that the **entropy** is lower for this class can be easily explained by the fact that the participants tend to follow the game master and do not spread their gazes across all possible targets.

The feature **symmetry** indicates that the amount of mutual gaze between participants increases with group involve-

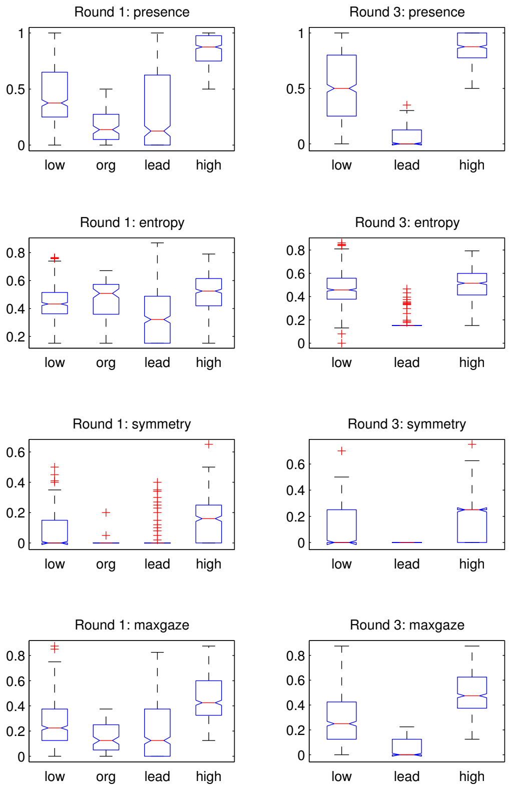


Figure 6: Boxplot of Gaze Data for Involvement Classes presence, entropy, symmetry and maxgaze

ment. This finding has been made for dyadic conversations. It is however, interesting to note that this still holds for an 8 party conversation.

The feature maxgaze similarly to the feature presence is distinguishing best between classes “high” and “low”. This means that people direct their gaze more towards the same participant in the high involvement class than in the low involvement class.

Table 1 shows the results of the GMM classification experiment. The models were trained on Round 1 and tested on both Round 1 and 3. The accuracy for the four categories is high and does not change significantly when testing on the training set (71.9%) or test set (71.3%). This indicates that the models and features generalise well over subjects changing roles.

The confusion matrix on Table 1 indicates that the “high” category is the easiest to correctly classify. “low” is often confused for “high” (but not the contrary). Surprisingly, “lead” is better recognised in the test set than in the training set. Although no example of “org” were present in the test set, the models do not often mis-recognise other categories for “org”.

We also carried out further classification experiments using the speech features described in Section 4 and a GMM classifier as in the gaze-base experiment. Preliminary results, however, indicate that the classifier based on the speech data did not perform as well as the classifier based on the gaze data. It was not able to satisfactorily distinguish between involvement class “high” and “low” as well as “org” and “lead”.

Training set	Round 1 (%)			
Confusion	high	low	org	lead
high	**94.6**	0.0	5.37	0.0
low	21.4	**58.2**	12.3	8.16
org	17.2	13.8	**60.9**	8.05
lead	0.0	4.76	14.3	**80.9**
Accuracy	71.9%			
Test set	**Round 3 (%)**			
Confusion	high	low	org	lead
high	**92.6**	1.64	5.74	0.0
low	37.6	**40.0**	16.0	6.40
org	0.0	3.77	**96.2**	0.0
lead	0.0	0.0	0.0	**0.0**
Accuracy	71.3%			

Table 1: Confusion matrix and accuracy for models trained on Round 1 and tested on Round 1 and Round 3

6.3 Relating Group involvement to individual engagement

So far we focused either on the individual subjects or on the group behaviour. The temporal features in Figure 3 allow us to relate the two aspects: we can monitor each subject’s engagement and the group involvement simultaneously.

For example, we can see how the different gaze patterns elicited by the role of subjects in the game contributes to the group behaviour. Some examples are the “lovers” MA and SO in Round 1, 70–100 seconds and SO and NV in Round 3, 50–60 seconds, the werewolves RM and KS in Round 1, 100–150 seconds and BG and MA in Round 3, 50–100 seconds. Although these patterns are specific to the current game settings, the method is general and can be applied to other settings as well.

7. DISCUSSION AND CONCLUSIONS

Our results show that it is possible to estimate individual engagement and group involvement in a multiparty corpus, by analysing the participants’ eye-gaze patterns. We were able to define a set of features that express the time evolution of different aspects of individual and group behaviour.

The individual features were successfully used to find similarities between subjects. The clusters that emerge correspond to the roles the individuals subjects played. They also resemble the clusters obtained by comparing the rankings that the subjects assigned to themselves, and to the other participants, after each round of the game. The fact that the resemblance is not perfect might be due to the fact that we based our analysis on gaze alone. We also only considered 5 minutes for each round of the game. The subjects however, assigned their rankings after each full round (approx. 10 minutes).

The subjects’ change of roles was also evident when we plotted each individual subject in feature space (using the **entropy** and **presence** features for simplicity). Here we were not interested in comparing the subjects with one another, but rather in observing the evolution of each individual subject when their role changes. The feature that best expresses these changes is the entropy of gaze patterns. Although the roles are determined by the specific rules of the game, we argue that similar roles are common in unconstrained conversations, such as those recorded in the D64 corpus [10], as well. To prove this, however, it will be necessary to test the method on unconstrained data, something that we leave for future studies.

All in all we could that we can statistically differentiate between the four classes of involvement, for each individual gaze feature, and that, with an average accuracy of 71%, our classifier is able to generalise well over the different rounds of the game.

8. ACKNOWLEDGMENTS

Catharine Oertel is supported by GetHomeSafe (EU 7th Framework STREP 288667).

9. REFERENCES

[1] R. Bednarik and M. Hradis. Gaze and conversational engagement in multimparty video conversation: An annotation scheme and classification of high and low levels of engagement. In *4th Workshop on Eye Gaze in Intelligent Human Machine Interaction*, 2005.

[2] D. Bohus and E. Horvitz. Models for multiparty engagement in open-world dialog. In *Sigdial 2009*, 2009.

[3] F. Bonin, R. Bock, and N. Campbell. How do we react to context? annotation of individual and group engagement in a video corpus. In *International Conference on Social Computing*, pages 899–903, 2012.

[4] B. Fuglede and F. Topsoe. Jensen-shannon divergence and hilbert space embedding. In *Proc. Int. Symp. on Information Theory*, 2004.

[5] D. Gatica-Perez, I. McCowan, D. Zhang, and S. Bengio. Detecting group interest-level in meetings. In *ICASSP 2005*, pages 489–492, 2005.

[6] R. Ishii and Y. I. Nakano. An empirical study of eye-gaze behaviors: Towards the estimation of conversational engagement in human-agent communication. In *2010 workshop on Eye gaze in intelligent human machine interaction EGIHMI*, pages 33–40, 2010.

[7] S. Kullback and R. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[8] A. Levitski, J. Radun, and K. Jokinen. Visual interaction and conversational activity. In *Workshop on Eye Gaze in Intelligent Human Machine Interaction: Eye Gaze and Multimodality*, 2012.

[9] I. McCowan, D. Garcia-Perez, S. Bengio, G. Lathoud, M. Barnard, and D. Zhang. Automatic analysis of multimodal group actions in meetings. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 27(3):305–317, 2005.

[10] C. Oertel, F. Cummins, J. Edlund, P. Wagner, and N. Campbell. D64: a corpus of richly recorded conversational interaction. *Journal on Multimodal User Interfaces*, 7:19–28, 2013.

[11] C. Oertel, C. de Looze, A. Windmann, P. Wagner, and N. Campbell. Towards the automatic detection of involvement in conversation. In *Analysis of Verbal and Nonverbal Communication and Enactment: The Processing Issues, Lecture Notes in Computer Science*, 2011.

[12] C. Oertel, G. Salvi, J. Götze, J. Edlund, J. Gustafson, and M. Heldner. The kth games corpora: How to catch a werewolf. In *Multimodal-Corpora:Beyond Audio and Gaze*, 2013.

[13] C. Oertel, S. Scherer, and N. Campbell. On the use of multimodal cues for the prediction of degrees of involvement in spontaneous conversation. In *Interspeech 2011*, pages 1541–1544, 2011.

[14] H. Salamin, S. Favre, and A. Vinciarelli. Automatic role recognition in multiparty recordings: Using social affiliation networks for feature extraction. *IEEE Transactions on Multimedia*, 11(7):1373–1380, 2009.

[15] G. Skantze and J. Gustafson. Attention and interaction control in a human- human-computer dialogue setting. In *Sigdial 2009*, pages 310–313, 2009.

[16] W. Y. Wang and J. Hirschberg. Detecting levels of interest from spoken dialog with multistream prediction feedback and similarity based hierarchical fusion learning. In *Sigdial 2011*, pages 152–161, 2011.

[17] B. Wrede and E. Shriberg. The relationship between dialogue acts and hot spots in meetingss. In *ASRU*, pages 180–185, December 2003.

[18] B. Wrede and E. Shriberg. Spotting hotspots in meetings: Human judgments and prosodic cues. In *Eurospeech 2003*, pages 2805–2808, 2003.