Automatic multi-modal detection of team cohesion in meetings

M. M. van der Wel

Automatic multimodal detection of team cohesion in meetings

by

M. M. van der Wel

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Friday December 13, 2019 at 10:00 AM.

Student number:	4323602	
Project duration:	February 1, 2019 – December 13, 2019	
Thesis committee:	Dr. H. S. Hung,	TU Delft
	B. Timmermans,	IBM
	Dr. C. Oertel,	TU Delft
	Prof. Dr. N. Lehmann-Willenbrock,	University of Hamburg
	S. Tan,	TU Delft

Abstract

In this thesis the automatic multimodal detection of social and task cohesion in meetings is studied. The presence of social and task cohesion has positive benefits on employee well-being, creativity and productiveness, and can therefore be used to assess meeting quality.

Conversational partners imitate each other's body language and speech characteristics to have smoother interactions, to increase liking, and it can cause a coordination of expectations. We hypothesize that social and task cohesion are therefore positively related to the imitation of both body language and speech characteristics. As the group-level alignment of non-verbal speech behaviour has been previously linked to social and task cohesion in meetings, this thesis investigates the relationship between cohesion in meetings and both motion and posture mimicry from accelerometer and video data. Motion mimicry is described using accelerometer features previously used for the detection of friendly and romantic attraction in pairs. We propose a method to convert these features to group-level descriptors of mimicry. The same quantifications of mimicry are also applied to video-based motion quantifiers that have been previously used to detect team cohesion. Appearance is described using HOG descriptors from densely-sampled feature points that are tracked over time. To our knowledge this is the first-time appearance similarity is used to estimate cohesion. We also test if a multimodal mimicry model performs better than a unimodal mimicry model to investigate if the different forms of mimicry contain complementary information.

Our group-level movement mimicry features detect social cohesion with average an area under the ROC-curve of 0.64 for social cohesion, and the accelerometer-based movement mimicry features specifically detect task cohesion with an average AUC of 0.63. The multimodal combination of the different features does improve over the unimodal models with an area under the ROC-curve of 0.68 for social cohesion and 0.65 for task cohesion. This shows both that movement mimicry is an indicator of verbal expressions of cohesion, and that measuring mimicry in different modalities can better model cohesion than a unimodal model. Further experiments are recommended for general appearance feature mimicry, specifically in the method used to measure mimicry, to confirm or deny if these are related to cohesion.

Preface

I have always been fascinated with the social signals people show. Communication is not just about the choice of words but also their context, the tone used, and the body language shown. What we can tell about people and their relationships with and impressions of each other by studying these signals is something I was very interested in examining. If these behaviours can be detected automatically, we can train computers to communicate with humans better and help people that are not as adept at these forms of communication. I am therefore very grateful for the opportunity to conduct a thesis that tries to find such an automatic detection method that can help people with this.

Before you lies my thesis 'Automatic multimodal detection of team cohesion in meetings'. This thesis is the final step towards completing my Master's Degree in Computer Science at the Delft University of Technology. It was written in a way that should make it understandable for anyone with a university degree in engineering.

I would like to thank the thesis committee for agreeing to be a part of the final part of my masters' journey, and reading my thesis. A big thank you to both my academic and external supervisor for their assistance during my thesis. Hayley for always pushing me to think critically of my approach, and enabling me to gain new perspectives. Benjamin for his many tips on writing, and ensuring the whole story makes sense, but also for showing me when to slow down.

Next, I would like to thank Stephanie for all her help. You were always willing to discuss my ideas, and were of great help in ensuring these were properly put on paper so that others could see them as well. Without your input many of my insights would have remained in my head.

Nale, thank you for always replying to my emails and being willing to Skype when I had questions about the meaning of the verbal expression-based labels. Your input helped me better understand the psychological dimension of my topic.

I would also like to thank the members of the socially perceptive computing group that have contributed valuable insights and support, and who where always willing to talk about specific aspects of my approach. The same should be said about the wonderful people at IBM Netherlands with whom I have had to pleasure of working, and in-particular the members of CAS. Being in such an environment and observing and discussing their projects has helped me improve my own.

A special thanks goes to Zoltán for entrusting me with this project and being a great supervisor during the first half of my thesis. I was very sorry to see him go because he was of great support and helped me make my own decisions.

The last and certainly not least important people to thank are my family and friends. My complaints where numerous and the doubts ever present, but you always helped me get back on track. Their support was invaluable during this project and motivated me to give my best.

(Marissa) M. M. van der Wel
Delft, December 2019

Contents

1 Introduction 1

1.1 Research Goals . . . 1

1.2 Contributions . . . 2

1.3 Research context . . . 2

1.4 Outline . . . 2

2 Background 3

2.1 Meetings and employee well-being . . . 3

2.2 Team cohesion . . . 3

2.2.1 Definition . . . 3

2.2.2 Value of team cohesion . . . 4

2.3 Mimicry as indicator for cohesion . . . 4

2.3.1 Defining mimicry . . . 4

2.3.2 Relationship cohesion and mimicry . . . 4

3 Related Works 5

3.1 Automatic detection of team cohesion . . . 5

3.2 Automatic detection other social constructs . . . 6

3.3 Automatic mimicry detection . . . 7

3.3.1 (Windowed Cross-Lagged)Correlation. . . . 7

3.3.2 Recurrence analysis . . . 8

3.3.3 Time Warping . . . 8

3.3.4 Mutual Information . . . 8

3.3.5 Probability distribution . . . 9

3.4 Event detection . . . 9

3.4.1 Video features . . . 9

3.4.2 Accelerometer features. . . . 10

3.5 Multi-modal classification . . . 11

3.5.1 Combined-modality features . . . 11

3.5.2 Early Fusion . . . 11

3.5.3 Late fusion . . . 11

4 Sensing Lab Dataset and Annotations 13

4.1 Overview meetings . . . 13

4.2 Sensing Lab set up . . . 13

4.3 Verbal Expression categories . . . 15

4.4 Multimodal feature alignment . . . 16

5 Methodology 19

5.1 Pre-processing . . . 20

5.2 Individual feature extraction . . . 20

5.3 Pair-wise features . . . 22

5.4 Group-level features . . . 24

5.5 Classification . . . 25

5.6 Evaluation . . . 25

6 Results 27

6.1 Design choices . . . 27

6.1.1 Size sliding window . . . 27

6.1.2 Determining number of mixture components for HOG . . . 28

6.2 Unimodal classification . . . 29
6.3 Multimodal classification . . . 31
6.4 Correlation analysis . . . 32
6.5 Assessing the impact of an alternate pair-wise comparison method . . . 34

7 Relating verbal expressions to team cohesion 37
7.1 Correlation questionnaire and coding categories . . . 37
7.2 External observer experiment . . . 37
7.3 Conclusion . . . 39

8 Conclusions and Discussion 41
8.1 Discussion . . . 41
8.2 Limitations . . . 42
8.3 Future work . . . 42
8.4 Main conclusions . . . 43

Bibliography 45

A Available data 53

B Act4teams observation categories 55

C Window distribution 57

D External observer instructions 59

1 Introduction

Meetings are a common occurrence in organisations, but their quality is not always perceived as good. Estimates on the number of hours spent in meetings vary between studies from 6 hours[66] to 13.5 hours per week[89]. This number is higher for managers who spend up to 80% of their time in meetings[66]. Employees and managers rate the quality as poor for almost half of these meetings[82]. When the perceived quality of these meetings is high, there is a strong and positive relationship between the number of meetings in a day and job-related comfort. However, if this perceived quality is low, this leads to a negative relationship[77]. Therefore, if we can improve meeting effectiveness, we can also improve job satisfaction.

Humans are constantly emitting subconscious social signals when communicating with each other. When we are talking not only the words themselves but also the voice pitch, speech rate, and accompanying posture and gestures[45] convey the message. A field of research investigating these phenomena is Social Signal Processing. Herein these subconscious signals are automatically measured and used to predict social constructs like dominance level[47, 49], amount of attraction[51] and depression[36]. Knowing what these signals are telling us can both help computers become more humanlike by imitating these signals[5], as well as help people identify and possibly adjust these signals. Our objective is to use these signals to identify when meetings are effective and when they are not.

The goal of this thesis is to automatically identify social and task cohesion in meetings using these social signals. For this an existing dataset that contains audio, video and accelerometer data is used. These modalities are all considered because different signals can be captured with different measurements. The overall approach will be based on mimicry between participants. The idea is that participants are more in tune with each other when they are aligning their movements and posture. If we can automatically identify social and task cohesion in meetings, we can provide feedback to a team who can then improve their meeting effectiveness and thereby their job satisfaction and performance.

The rest of this chapter will present the research goal and contributions of this thesis. Thereafter an outline of the thesis will be included.

1.1. Research Goals

The main goal of this research is to better understand the multimodal signals regulating small group meetings. Our hypothesis is that participants unconsciously align not only their speech, but also their movements and posture to allow for better communication. We also hypothesize that a better prediction for social and task cohesion can be made with multiple modalities, as one modality could capture information the others do not have. The first goal is to quantify the extent to which general movement patterns and posture are mimicked between people and examine if this can be used to predict social and task cohesion. The second goal of this work is to verify if a multimodal system using movement, posture and speech alignment outperforms a unimodal system.

1.2. Contributions

In this thesis, it is hypothesized that interpersonal coordination of behaviour in small group meetings can serve as an indicator of team cohesion. The contributions of this research involve developing mimicry features inspired from literature to describe general movement and posture similarity using video and accelerometer data, and combining their prediction with paralinguistic mimicry features from audio data to see if they complement each other. The two main contributions are:

1. Prediction of social and task cohesion using accelerometer and video mimicry features

We take accelerometer-based mimicry features previously used for the prediction of different forms of attraction in pairs. These pair-wise features are aggregated to form group-level features which are used to detect cohesion.

We have also created an appearance mimicry measurement. This approach looks at the appearance of the moving points within different videos to create a model describing the appearance of the moving body parts within a given meeting segment. The appearance of the points from another person is then applied to this model to rate their similarity.

Both mimicry features are tested to see if there is a relation with social or task cohesion.

2. Multimodal mimicry as a prediction of social and task cohesion

We create and test a multimodal mimicry approach that combines accelerometer, audio, and video mimicry features to attempt to improve on the performance of a single modality. This is done by fusing the predictions after classification with each modality.

1.3. Research context

This thesis study uses data collected for a shared project between the IBM Center of Advanced Studies, the TU Delft Socially Perceptive Computing Lab and the VU Amsterdam Experimental and Applied Psychology department. The data collection took place at the IBM Sensing Lab. The goal of this project is to improve employee engagement and job satisfaction. A previous study has been done within this project that investigated paralinguistic mimicry as a way of detecting verbal expressions of social and task cohesion [68].

1.4. Outline

1. **Chapter** 2 shows the importance of investigating meeting outcomes and provides motivation for the direction of this research.

2. **Chapter** 3 gives an overview of the related works. In this chapter previously used methods for the detection of team cohesion are listed, followed by a section on the use of social signals for automatic detection of other social constructs. This is followed by an overview of available mimicry detection methods and a section that lists methods that have been used to extract social signals from video and accelerometer data. The last section explains the different ways to work with multimodal data in classification.

3. **Chapter** 4 contains a description of the data collected in the IBM Sensing Lab. It describes the different modalities and their synchronisation, as well as the annotation of the dataset.

4. **Chapter** 5 explains the methodology. The different steps to get from the raw accelerometer and video data to the features used for the prediction of social and task cohesion are explained in this chapter. The motivation for this approach is also explained, as well as the combined modality approach. The final part describes the evaluation method for the classification.

5. **Chapter** 6 lists the results gathered from the experiments. The performance of all the unimodal models and the multimodal models are shown.

6. **Chapter** 7 discusses the experiments conducted to verify the validity of the social and task cohesion labels.

7. **Chapter** 8 discusses what we can conclude from these results. It also describes the limitations of this research and suggests possible experiments to continue this research.

2 Background

This chapter motivates why we are interested in using mimicry as a prediction for social and task cohesion in meetings. The first section explains why it is so important to improve and analyse meetings. The second section explains the concept of team cohesion and the benefits of having high cohesion. The third and last section defines the concept of mimicry and argues why it should be used to detect the presence of cohesion.

2.1. Meetings and employee well-being

Meetings can greatly vary in the number of participants, the purpose and the setting. Rogelberg defines meetings as: "purposeful work-related interactions occurring between at least two individuals that have more structure than a simple chat, but less than a lecture." [76] Meetings can be called to solve problems, give information, develop policies, and can build a sense of community among participants[91]. To potentially attain these benefits a large amount of some employees' time is spend in meetings. Estimates vary from 6[66] to 13.5[89] hours a week for normal employees, with up to 80% of time for managers in large companies[66]. The goal of these meetings must not always be reached as according to a study by Schell [82] 41.9% of meetings have a quality rating of poor.

It has been shown that employee satisfaction with meetings can greatly impact job satisfaction, especially for employees that attend a large number of meetings[78], as well as job-related comfort[77]. If meeting quality can be improved by automatic analysis, this can be very beneficial for both employees and employers.

2.2. Team cohesion

The concept we are measuring in this study is team cohesion. This section first explains what this concept is, followed by why one would like to measure it.

2.2.1. Definition

Team cohesion was originally defined in the 1950s by Festinger as:

"The total field of forces which act on members to remain in the group. These forces may depend on the attractiveness or unattractiveness of either the prestige of the group, members of the group, or the activities in which the group engages." [39]

The attraction to the group has been assessed by asking group members how much they like each other and how long they want to stay in the group[44]. This measures more the attractiveness of the members of the group, and less the attractiveness of the activities in which the group engages. Since then there has been a shift from this unidimensional view of team cohesion to a multidimensional view[21]. Salas et al. [80] have analysed the different dimensions considered under team cohesion that have been used in literature and conclude that team cohesion should be considered multidimensional. Social and task cohesion should be prioritized when measuring team cohesion and as they show a high correlation with performance. Herein task cohesion is defined as:

"An attraction or bonding between group members that is based on a shared commitment to achieving the group's goals and objectives"

and social cohesion is defined as: "A closeness and attraction within the group that is based on social relationships within the group"[80].

In this project social and task cohesion will therefore both be measured. Cohesion was at first mostly assessed in sports teams with a questionnaire developed and tested by Carron et al. [23]. A version specific to work teams was later developed by Carless and De Paola [21].

2.2.2. Value of team cohesion

Team cohesion is an asset of great value to companies. When team cohesion is measured as social and task cohesion it is related with performance according to the survey by Salas et al. [80]. Michalisin et al. [63] consider team cohesion to be a strategical asset and show it provides superior returns. According to Chang et al. [26] it is easier for cohesive teams to be creative. It has been shown that social and task cohesion both need to be present at the same time to successfully perform on tasks requiring group interaction[109].

Team cohesion does not only benefit the companies by increasing employee creativity and performance, but can also benefit employees themselves. Being part of a very cohesive team likely contributes to the well-being of the team members[93]. Finding ways to improve team cohesion can therefore greatly benefit both companies and employees.

2.3. Mimicry as indicator for cohesion

This section provides a definition of the concept "mimicry" as it is used here, and how it is different from synchrony. Thereafter we explain why we think mimicry can be a predictor of social and task cohesion.

2.3.1. Defining mimicry

Across different studies the terms mimicry and synchrony have sometimes been used interchangeably. To distinguish between these two concepts we look at the survey by Delaherche et al. [34]. Herein synchrony is defined as a form of process coordination whereas mimicry is a form of content coordination. Content here refers to what people intend on doing, and process to the systems used to carry out these intentions[31]. Synchrony is about the timing of behaviour and mimicry about the nature of the behaviour. Mimicry can occur in the following types of behaviour: speech, facial expressions, posture, gestures and other types of physical movement and emotions[28]. The amount of time between the original and the mimicked action can vary depending on the type of action[59]. Most facial actions get mimicked within a few seconds while hand gestures can take over thirty seconds to be mimicked.

2.3.2. Relationship cohesion and mimicry

Mimicry is a process that automatically occurs when people interact with each other. It binds and bonds people and can smooth interactions[28]. It also has been shown to stimulate empathy and has been linked with rapport. More specifically it has been shown that when a person mimicked the behaviour of another, the other would report liking that person more and having a smoother interaction[27]. This later was shown to be the case for both the mimicker and the mimickee by Stel and Vonk [88]. It has also been shown that the amount of mimicry increases, when the task difficulty increases[59]. The reaction time of mimicry has been linked to how external observers perceive empathy[54]. Lastly mimicry can achieve coordination of expectancies among participants[34].

Social cohesion relates to members of the group liking each other, which as just stated can be caused by mimicry. We therefore consider mimicry to be positively correlated with social cohesion.

Task cohesion means members share a commitment to achieving the group's goal. This would be enabled by smoother interactions and greatly benefits from the coordination of expectancies among participants. Therefore, we argue mimicry is positively correlated with task cohesion.

3
Related Works

This section provides an overview of literature related to the subject of this paper, and the method used in this paper to measure mimicry in accelerometer and video data, and create a multimodal model. The first section discusses other literature on the automatic detection and prediction of team cohesion, the following section discusses the automatic detection of other social constructs, the next section discusses different methods for comparing time signals and calculating mimicry between signals, and the last section discusses the different types of behaviour that can be detected in video and accelerometer data.

3.1. Automatic detection of team cohesion

Some studies have attempted to automatically detect team cohesion in different settings, and with different ways of measuring this concept. This section lists these different settings and methods, and discusses where they are different from the setting and methods used in the experiments conducted here.

Salas et al. [80] survey some of these methods which include analysing online communication traffic, and badges that track participants' location.

Hung and Gatica-Perez [46] estimate team cohesion for whole meetings as labelled by external observers. To increase the separation between the classes when binarizing into high and low cohesion the middle 50% of the data is discarded. The used dataset contains both real meetings, and meetings conducted according to a scenario by volunteers. Both audio and video turn-taking features, as well features combining these modalities like motion when not speaking are used. Reported accuracies are 90% with audio features, 83% with video features and 82% with multimodal features.

A study by Nanninga et al. [68] predicts verbal expressions of social and task cohesion in meetings. Each verbal expression was coded according to the act4teams coding[52] and these codes were translated into labels for 2- or 5-minute meeting segments. Paralinguistic mimicry features were used to predict these labels. These features lead to an AUC (a classification evaluation metric that considers the trade-off between true positives and false positives) of 0.69 for social cohesion and 0.63 for task cohesion. Combining these features with the audio-based turn-taking features from Hung and Gatica-Perez [46] improved the social cohesion AUC to 0.74. The individual AUC for the audio-based turn-taking features from Hung and Gatica-Perez are also noted at 0.68 for social and 0.52 for task cohesion.

A 2018 study by Zhang et al. [110] investigated social and task cohesion in small and isolated teams. Social and task cohesion are measured by having the participants answer a single question for each concept twice a day. These values are averaged over the participants to acquire group cohesion values. The participants wore a badge that records vocal activity, movement activity and interactions between participants. This experiment resulted in an AUC of 0.64 for social cohesion, and 0.84 for task cohesion. The relative feature importance was also analysed, which showed all the top performing task cohesion features are movement related. This would argue that the inclusion of movement features can improve task cohesion prediction.

Summary

Previous studies on the automatic detection of team cohesion have not tried to predict verbal expressions of social and task cohesion on a small timescale using movement and posture-based mimicry features. Zhang et al. [110] have shown that movement-based features predict long-term task cohesion well, so a logical next step would be to test if the same is true for short-term task cohesion. This thesis will therefore see if movement and posture-based mimicry features can predict verbal expressions of social and task cohesion well, and if combining audio and video/accelerometer-based features improves performance.

3.2. Automatic detection other social constructs

There are other social signals that have been detected automatically. Some of the methods used to detect these signals could be applicable for the automatic detection of task and/or social cohesion. In this section these signals and their detection methods will be discussed, as they might provide inspiration for the detection of cohesion. We will only be discussing methods that involve accelerometer, video and/or audio data as these are the types of data available in our experiment.

Many different types of attraction are possible between a pair of people, and it does not have to be romantic or sexual. Kapcak et al. [51] predict three different types of attraction: romantic, sexual, and social. This prediction was made by looking at mimicry in accelerometer date of the participants during 3-minute speed dating interactions. Social attraction is evaluated by asking the participants if they can rate if they see the other person as a potential friend. Social cohesion is also about being socially close with other people in a group, and therefore this method could be a good predictor of social cohesion.

Rapport can result in smoother social interactions, improved collaboration and improved interpersonal outcomes. Interactions are often labelled for rapport by asking if the participants like each other and enjoy what they are doing[10]. This again is a similar concept to social cohesion. Müller et al. [64] try to detect low rapport from non-verbal behaviour in group interactions. For each participant unimodal features are extracted in the form of speech activity and prosody from audio data, and facial and hand motion features from video data. Multimodal and multi-person features are also extracted and tested. These multi-person features are the synchrony between participants in Action Units and hand velocity, which are measured by calculating Dynamic Time Warping distance with a Sakoe-Chiba band of five seconds(see subsection 3.3.3). It was found that facial features (in the form of Action Units) are the most indicative of failure to establish rapport in group interactions.

Pentland and Madan [72] focus on the perception of social displays of interest and attraction. Activity, engagement, emphasis and mirroring are quantified from audio and video data. The main indicators of social interest were emphasis and activity. Audio activity is measured by segmenting the speech stream into speaking and non-speaking segments. Conversational activity level is then defined as the percentage of speaking time. Video activity is measured by counting the number of gestures per second. Emphasis is measured by variation in pitch and amplitude for audio, and the standard deviation of gesture velocity and acceleration in video. Although these are called video metrics, in one of the three experiments equivalent measurements from accelerometer data are used, and it is not explained how these gesture values are extracted in the other experiments that use the video data.

Kumano et al. [53] create a graph structure of smiles between participants. This graph is used to find affect and liking between participants. An interpersonal emotion network is defined that shows the amount of smiling one person directs to another during conversation, which indicates both liking between participants and affect. This approach requires both gaze estimation and smile detection. Gaze estimation is difficult in our setting as the angle between participants changes both within and between meetings.

Conversational involvement is used as a measure of interaction quality. Altmann et al. [1] look at windowed cross-lagged regression (see subsection 3.3.1) and a dependence analysis as measures of synchrony between participants. The intensity of facial activity is extracted with the method defined in [83] and [95]. Facial activity synchrony between participants is then compared using the two different synchrony measures. A significant dependence between conversational involvement and frequency of synchrony was found, however the effect size is small. It should be noted that this experiment was only done for two people in two group interactions.

Ramseyer and Tschacher [74] look at coordination of patient and therapist's head and body movement as a predictor of therapy outcomes. Head and body movement was analysed with Motion Energy Analysis which looks at the difference between consecutive grey-scale video frames. The sections capturing the head and the upper body have been pre-defined. Coordination was quantified by looking at windowed-cross correlation with 1-minute window segments and time-lags up to 5 seconds. The method described in subsection 3.3.1 is used to correct for spurious synchrony. The authors asses the influence of coordination on 2 different outcomes: macro and micro. Micro refers here to the result for that specific session, and macro to the result of the whole treatment. An increase in head coordination was found to positively influence macro-outcome of therapy, whereas body coordination positively influenced micro-outcome.

Another construct that has been automatically predicted is dominance. The most dominant person in a conversation influences its direction and can exert power. Hung et al. [47] test if the most dominant person receives more visual attention (is the most watched) both when they are speaking and when others are speaking. Head pose is tracked using a mixed state particle filter[4]. Gaze could not be tracked in this experiment because the image quality is too low. Dominance has been most robustly predicted using speaking activity, but dominant people also have higher movement activity[49]. Jayagopi et al. [49] predict dominance on 5-minute meeting segments with speaking turn and activity features extracted from audio, and equivalent features in video. The audio features by themselves were found to perform better than combining audio and video features; however, the difference in performance between the best audio and best video features was not significant. The video activity features were extracted using compressed domain features and skin-coloured blocks as described by Yeo and Ramchandran [105].

Similar to finding the dominant person is finding the leader within a group. Leaders, like the dominant person, contribute more to the conversation. Therefore, similar features to those used for dominance detection have been used for leadership detection: A combination of visual focus of attention and body/head motion features, as well as audio-based turn taking features. Sanchez-Cortes et al. [81] extract head motion features using optical flow and facial tracking to describe head activity. Body motion was described by frame differencing as the background was static, and only a single person was visible within one angle. Beyan et al. [11] extract visual attention features using optical flow, as well as body and head activity features. The same authors later included audio features to find speaking activity features[12]. Fused audio/video features were found to perform better than audio features by themselves.

Yu et al. [108] propose a method that measures the amount of synchrony between dyads to detect deception. Head pose, head gestures and facial expressions are detected from video. The cross correlation for each feature between both participants is calculated and used to quantify synchrony. An accuracy of 74.2% is achieved after feature selection.

Hammal et al. [43] have found that while the amount of head motion increases during conflict situations, the amount of coordination between partners' decreases. Head motion is measured using a 3D cylinder-based head tracker[32] and the amount of synchrony between partners is compared with Windowed Cross Correlation and peak picking. It should be noted that peoples' movement and vocal qualities change when suffering from depression. Their speech rate slows down and their amount of movement decreases[35, 36]. How this impacts their level of mimicry in normal conversation is unknown, but an increase in coordination has been shown to positively impact therapy outcomes as mentioned before.

Summary

From previous research we can see that social signals occur in both voice and motion. Mimicry in both of these modalities has been investigated as a predictor for other social constructs similar to cohesion. Mimicry in audio has been investigated previously and has shown to be a good predictor of both forms of cohesion, but mimicry in motion in natural meeting settings less so. It seems as if comparing participant's head and body motion is the most common way of detecting mimicry.

3.3. Automatic mimicry detection

There are many different ways to compare two time series. In this section we will provide an overview of these methods. After reviewing the different signals we can extract from the accelerometer and video data, we can conclude which of these methods we can use to define the amount of mimicry between two people for the signals we extract.

3.3.1. (Windowed Cross-Lagged)Correlation

A method that is commonly used to compare the similarity of time series is cross correlation[34]. It calculates the Pearson correlation at different amounts of lag with the following formula:

$$r_{x,y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{3.1}$$

When calculating correlation this formula is applied to the whole signal, but the windowed cross-lagged correlation calculates it for a time window that is shifted over the signal. This results in a correlation value for each shift and each considered lag value. This is often followed by a peak picking algorithm which finds maximum value in a local region[15]. Another follow up method is to take the maximum for each of the different shifts to find a similarity measure for each shift. The resulting values need to be adjusted for similarity that is naturally present between signals. This can be achieved by defining a baseline similarity value, which is calculated by comparing the original signal with a shuffled version of the other signal[3, 75, 90]. A disadvantage of this method is that the size of the window needs to be determined as well as the number of lags considered. Research has been done into the average lag between different types of participant movement which would help determine the considered lags[59].

3.3.2. Recurrence analysis

Like windowed cross-lagged correlation, recurrence analysis compares the similarity of two time series at different time points within both signals. They are unlike in that the similarity metric is often a distance metric in the case of recurrence analysis. The result is a matrix of distances between the two signals starting at different time points within either signal. This matrix is often visualized by applying a threshold on the distance values so that the similar parts of the signals are represented with a 1 and the dissimilar with a 0. An example of such a plot can be seen in Figure 3.1. Diagonal lines in the plot show high correlation between two signals at a specific lag value. Webber and Zbilut [99] have defined five metrics that can be used to describe the contents of the recurrence plot. This metric has been proven successful in predicting attentiveness based on similarity of eye movement[75]. A disadvantage of this metric is that it is difficult to determine which of these metrics define truly define similarity for a given situation, and what the size should be of the signal segments that are compared.

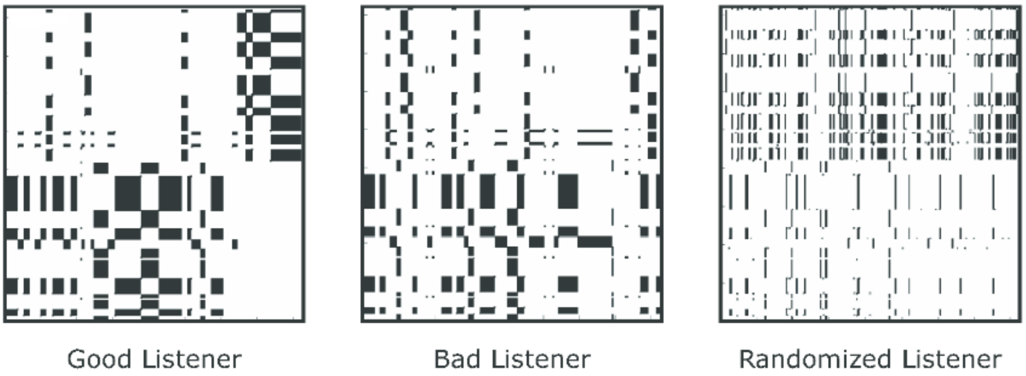


Figure 3.1: Cross-recurrence plots comparing eye gaze of speaker and different types of listeners adapted from Richardson and Dale [75]. The horizontal axis captures the gaze of the speaker and the vertical that of the listener. A good listener here has good alignment with the speaker while a bad listener is not as well aligned. This can be seen by the low amount of diagonal lines that can be drawn in the plot for the bad listener compared to the good listener. There is worse alignment between the speaker and a randomized listener compared to the other two listeners as can be seen in the small amount of black in the plot and no diagonal lines.

3.3.3. Time Warping

Another method used to determine the distance between time series is the time warped distance. This method tries to find the optimal alignment between two time series, with a penalization factor when a shift takes place. It therefore ensures that a specific action does not have to start at the exact same time, but also does not have to last the same amount of time for different people. Dynamic time warping produces a dissimilarity value as a higher value indicates a higher alignment cost and therefore a larger distance. Bilakhia et al. [14] have used Generalized Time Warping to align the motion of humans in both video and accelerometer data. It has also been used to compare the similarity of facial expressions and heart rate[29].

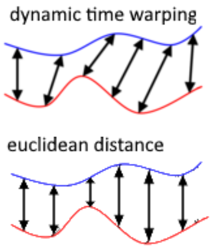


Figure 3.2: Dynamic Time Warping matching in the top image versus Euclidean Distance matching in the bottom image.

3.3.4. Mutual Information

Mutual Information is a metric that indicates how much knowing information about one variable reduces uncertainty about the other. This can be applied to time series, or the time series can be represented in a statistical way. Mutual information is calculated as:

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \tag{3.2}$$

where H(X) and H(Y) are the marginal entropy for both variables and H(X, Y) is the
joint entropy. From this the Normalized Mutual Information can be calculated which gives a score between 0 and 1.

$$NI(X;Y) = \frac{I(X;Y)}{\sqrt{H(X)H(Y)}} \tag{3.3}$$

This metrics was used to quantify the amount of mimicry between the accelerometer data of participants in a speed dating event[51].

3.3.5. Probability distribution

Solanki et al. [87] use the distribution of speech characteristic features to model mimicry within audio. Part of the signal was used to model the individual distribution of each participant with a Gaussian Mixture Model. Features of the other participants were then applied to this model to get the likelihood of the features belonging to the other participant's distribution. This can be used to represent the similarity of speech characteristic between participants, and if this likelihood increases over time they are converging. This approach was also used by Nanninga et al.[68] to model synchrony and convergence between participants and predict cohesion. Instead of applying the samples from one person to that of the other, the trained model can also be used to create samples from the established distribution. These samples can then be fed into the model of the other if the dimensionality of the feature vectors is the same. This is called Monte Carlo sampling and has been used to compare the similarity of MFCC trained GMM's by Pampalk [70] with 2000 samples. Jensen et al. [50] conclude this method performs best for MFCC features when compared with Earth Movers Distance and Normalized L2 Distance.

Summary

As we have shown there are many ways to measure the similarity between sets of signals, and many of these have been previously used to quantify mimicry. Windowed cross-lagged correlation, recurrence analysis and time warping are not directly informing about raw video signals, but can be used for accelerometer data. Recurrence analysis is difficult to use as it requires the definition of a mimicry threshold.

3.4. Event detection

This section describes the different social signals that can be measured in the different modalities. The focus is on the behaviours that can be seen in video, as this has not been summarised before for cohesion.

3.4.1. Video features

Body motion and appearance

Wang et al. [96, 97] have tested a method that has been used to classify actions within a video. Feature points are densely sampled and tracked using optical flow. Four different types of feature descriptors are extracted using the dense trajectories: the trajectory of the point, the Histogram of Oriented Gradients (HOG), the Histogram of Optical Flow (HOF), and Motion Boundary Histograms in the x and y direction (MBHx and MBHy). These describe local appearance, and local motion information.

Cao et al. [20] have developed a neural network that tracks the pose of multiple people within a video. This posture is described by 2D coordinates of body/foot key points. The distance between specific key points and the change in point location can be used to describe posture and change in posture.

Chippendale [30] detects hand and head fidgeting (restlessness shown by nervous movements) with adaptive background modelling and a statistical skin model. Adaptive background modelling is defined based on the temporal difference and temporal stability (remaining 0). The skin model is defined in the YCbCr colour space on the basis of manually segment skin-pixels. Head nodding and shaking is also determined by looking at the change of head position.

Hand gestures

To detect the shape of the hand, the hand is segmented from the rest of the image. This is a relatively easy process if depth is also recorded[58]. If this is not the case a form of background subtraction[42, 56] or skin-filtering[107] is usually applied.

Pisharady and Saerbeck [73] give an overview of hand gesture classification methods. HMMs are the most common method in this survey, which recognises a set of predefined gestures and needs labelled examples[107]. The motion of the hand can be compared directly using synchrony measures such as Dynamic Time Warping[56] and correlation coefficients[103], or the shape is classified using a neural network[58].

Xie et al. [102] outfit subjects with a ring that detects motion in the x, y and z direction. Basic gestures are described by simply noting the direction of movement. These rings provide a very accurate depiction of hand motion, but requires special equipment.

A subset of hand gestures is hand-over face gestures. These gestures used to be treated as noise in facial expression recognition, but can actually convey meaning themselves[71]. Mahmoud and Robinson [60] detect the handshape and location on the face using 3D cameras and assign one of six mental states. It should be noted that hand-over-face gestures occur in 21% of video segments, and are therefore not a continuous expression of mental state. Mahmoud et al. [61] developed a method that can find hand-over-face gestures from non-depth video. To find the facial region that is occluded, a Constrained Local Neural Field (CLNF) facial landmark detector [7] and tracker is used. This calculates a likelihood of a landmark being aligned, and if this likelihood is low it is either not aligned or occluded. The shape of the hand is described using the HOG of Space Time Interest Points (STIP).

Head motion

If head pose is tracked in time it can be used to find head gestures and motion, but head pose is not always used to find head motion. Head pose is often detected using facial landmarks/keypoints which are also used for facial action unit and emotion detection[55, 104]. An example of this is the OpenFace behavioural analysis toolkit by Baltrusaitis et al. [8] and its new version OpenFace 2.0 [9]. This approach results in head pose estimation, eye gaze estimation, and facial action unit recognition. These facial action units are a subset of those described by Ekman and Friesen [37]. OpenFace 2.0 has a mean absolute degree error of 2.6 on the BU dataset[57]. It should be noted that OpenFace 2.0 has not been trained from an elevated angle and its performance in such a situation is therefore unclear.

Ruiz et al. [79] estimate 3D head pose without the use of landmarks using ResNet 50 with a Mean Squared Error loss for each direction. The performance was evaluated on the BIWI dataset[38] which resulted in an absolute degree error of 3.2. Older methods include the use of a 3D Constrained Local Model with a generalised adaptive view-based appearance model[6, 65] which do not perform as well.

Yu et al. [108] compare head pose between two people using cross correlation to detect deception. The presence of head gestures has been measured by Sharma et al. [84]. These gestures are classified by extracting head poses in the FIPCO dataset[104] which are used as input for a Multi-Scale Convolution-LSTM.

Xiao et al. [101] compute optical flow in the facial region to estimate head motion. A Gaussian Mixture Model is used on the extracted features to separate motion from non-motion segments. Symmetry is measured using the Kullback-Leibler (KL) divergence of the two models.

Yeo and Ramchandran [105] look at the residual coding bit-rate and DCT coefficients to find motion within video blocks. The skin-coloured blocks are used as a description of head motion.

Summary

The dense trajectory features by Wang et al. seem like a promising way to capture both body motion (trajectories, HOF) and appearance(HOG). Hand gestures are more difficult to capture as they often require labelled examples of specific gestures. Head motion can be captured using landmarks if the subject is facing the camera frontally, an elevated camera decreases the effectiveness of this method. The residual coding bit-rate method has been tested in the team cohesion setting by Hung and Gatica-Perez [46], and an analogous method is here used as baseline features. The thresholded pixel-wise difference represents a similar concept and is used as code for the residual coding bit-rate method has not been published.

3.4.2. Accelerometer features

The thesis by Öykü Kapcak [106] gives an overview of the automatic quantification of interpersonal interest using accelerometer based synchrony measures. Therein the research by Cabrera-Quiros et al. [18] is listed as the only other known research that uses accelerometer data for the prediction of social constructs. The mean and variance of the magnitude of acceleration is extracted together with the variance over 1-second sliding windows with a shift of 0.5 seconds to emulate video features by Veenstra and Hung [94]. The video features were constructed to measure if people move at the same time, how participants are angled, and if they move closer to each other.

Kapcak et al. [51] create statistical and spectral representation of the movement in each direction and the total magnitude for windows in the accelerometer signal. For these representations windowed cross correlation, mutual information, distance between consecutive windows, and convergence are calculated for pairs of participants.

3.5. Multi-modal classification

This section looks at the different ways that we can combine different modalities. The most common combination is audio and video data. The most commonly used audio features are turn taking or prosodic(paralinguistic) features, which together are called non-verbal. There are three main ways in which we can combine different modalities which are explained in this section.

3.5.1. Combined-modality features

Features can be created from a combination of different modalities. Beyan et al. [12] use features for activity while speaking, and attention received while speaking and not speaking, which requires both audio and video information. Müller et al. [64] also look at the difference in activity while speaking and not speaking, but this time for AU activations and intensities. Hand movement is only measured when the participant is speaking. Creating extra video feature representations for speaking and non-speaking sections seems to be the most common combined feature[2, 46, 47, 69].

Another metric that uses information from the audio and video modality considers the synchrony between both as a feature. Example of this are the prediction of monologues from audio-visual synchrony by Iyengar et al. [48], and the inclusion of audio-visual synchrony measured with mutual information by Hung and Gatica-Perez [46]

3.5.2. Early Fusion

As the term early fusion suggests it fuses different features together early in the classification process. Early means before feeding the features into the classifier, and so the classifier is trained on features from all different modalities. Examples of this approach include combining audio, video and audiovisual features for the detection of dominance[49], interest [72], predefined meeting behaviours[62], mimicry [13], rapport with virtual agents[25] and personality[2].

Dibeklioglu et al. [36] select the top x least redundant features for each modality before combining the modalities. A higher dimensionality increases the required number of samples to reduce the chance of error, and it can therefore be beneficial to reduce the number of features[92]. This approach can lose features that provide complementary information within the different modalities.

3.5.3. Late fusion

The last way to use multi modal features that we describe is late fusion. For late fusion a model is trained for each individual modality or feature, and their predictions are afterwards combined in some way. This combination can be done with a rule, or by training a model on the prediction probabilities for each model[19]. Rule-based approaches include averaging the probabilities and a weighted combination (depending on the performance of each classifier).

A visualization of early and late fusion from a paper by Snoek et al. [86] can be found in Figure 3.3.

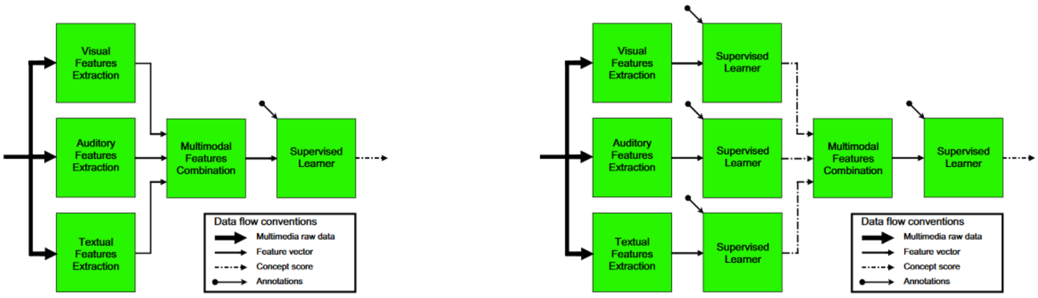


Figure 3.3: Early and late multimodal fusion models. The left image represents a general early fusion model where the features are combined before classification. The right image represents a late fusion model where a model is learned for each modality individually and the predictions of these models are combined after. Taken from [86].

Sanchez-Cortes et al. [81] look for the leader in a group by ranking the participants according to their classification outcome for each feature and selecting the person with the highest rank as most leader. The ideal weighted combination can be determined by testing varying weights, and determining the best combination according to some metric[40, 48]

Summary

As we hope to see if the different modalities compliment each other, each modality is modelled individually and the prediction of these models is combined afterwards using a late fusion approach.

4

Sensing Lab Dataset and Annotations

In this chapter the data collected in the IBM Sensing Lab is described. This is the data that has been used in this research. The goal of the experiment was to understand how well a real-life meeting is progressing by automatically measuring its quality via sensing equipment. The first section provides an overview of the recorded meetings, thereafter the characteristics of the different types of data used in this research are explained in more detail. After that we elaborate on how the ground truth was obtained and how the different modalities were synchronised within and between each other.

4.1. Overview meetings

A total of 25 meetings were held and recorded in the Sensing Lab. As the lab was set up to measure real-life meetings all participants were asked to hold their meeting as they normally would have. The participants had to read and fill out a consent form before the start of the meeting, as well as a questionnaire afterwards. Questions were asked about the demographics of the participant and about the participant's feelings about the meeting. The number of people in a meeting varied from 3 to 8 with an average of 4.6. Participants are 64% male and have an average age of 37, the race of the participants is not recorded in the questionnaire. Care should be taken with the generalization of conclusions for other demographics.

4.2. Sensing Lab set up

The meeting room in which the data was recorded had a table in the centre, 8 chairs and a beamer on one of the walls. The layout of the room can be seen in Figure 4.1. This room was outfitted with 8 overhead cameras, each one centred on the seat on the opposite side of the room. The cameras were mounted overhead to cause the least amount of influence on the behaviour of the participants. Furthermore, an array of microphones was placed in the centre of the table, and all participants were outfitted with a microphone and a Chalcedony accelerometer to record their speech and movement respectively. This accelerometer was hung around the neck of each participant and recorded the acceleration in the x, y and z direction with a sampling rate of 20 Hz. An overview of all the data available for each meeting can be seen in Table 4.1 and Table A.1. Out of the 25 recorded meetings 14 have labels and all modalities available.

Table 4.1: Available data for each modality. N represents the data is either not available, not synchronised with the other modalities, or participants walk out the camera angle. The accelerometer column represents the number of participants for which the accelerometer data could be recovered. Meeting 8 only include accelerometer data for 2 participants, which does not constitute a group and is therefore also excluded for the accelerometer data.

Meeting nr	Label	Video	Accelerometer	# participants
1			N	5
2			N	4
3	N			5
4			7	8
5			4	6
6			3	4
7			3	3
8			2	3
9			3	5
10			N	5
11			6	6
12			5	5
13			3	4
14			N	4
15			4	5
16			5	5
17	N			3
18			4	6
19		N	3	3
20			4	6
21			6	6
22		N	5	5
23		N	3	4
24		N	3	3
25			3	3

A more in-depth explanation of the audio recorded can be found in the study conducted by Nanninga et al. [68]. For a more detailed description of the chalcedony accelerometer we refer the reader to the description in the MatchNMingle dataset paper[18]. The specifics of the video data will be described in this section.

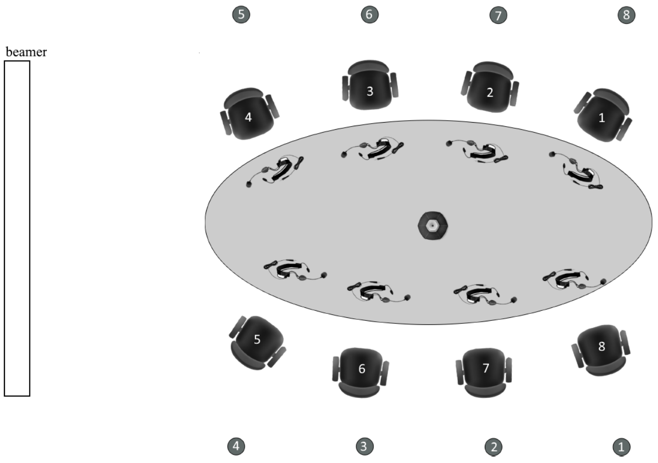


Figure 4.1: Schematic overview of the Sensing Lab layout. The cameras are represented with circles and film up to three people sitting on the opposite side of the table. A beamer is present on the left side of the room. The middle of the table shows an array. On the table in front of each seat lies a headset that is to be worn by each participant.

Video data

The room was outfitted with 8 fixed cameras, with 4 cameras each facing the opposite side of the room from slightly overhead. There is an overlap in the field of view of adjacent cameras. A person visible in the middle for camera 2 can be seen on the left for camera 1, and on the right for camera 3. In the same way up to 3 people can be seen on the same camera. For 3 of the 25 meetings only a video that consists of 4 camera angles combined is still available. These meetings will therefore not be included in the video analysis. The cameras were ABUS HDCC72510's which record at a resolution of 1920 x 1080 and at 25 fps[1]. However, some videos have been saved with a resolution of 704 x 576 instead. Out of the 96 videos that have been used in this experiment 17 are not available in HD. Although the cameras should have a constant 25 fps, all the available videos had a variable frame rate.

4.3. Verbal Expression categories

Twenty-three meetings were coded with the act4teams coding[52]. This coding was created to analyse team interactions, and encodes the verbal statements for each member of a team in one of 44 observation categories. These 44 categories can be grouped into 12 interaction aspects, which can in turn be further grouped into four global facets of communication:

- **Problem-focused statements**, which contains categories like describing the problem, and describing the solution.
- **Procedural statements**, which are statements that steer the meeting and contains categories such as goal orientation, task distribution and summarising.
- **Socio-emotional statements**, which are statements that for example encourage participant or provide support.
- **Action-oriented statements**, which contains categories like interest in change, and taking on responsibility.

The observation categories can be both positive and negative. An overview of all 44 categories can be seen in Appendix B. For each of the 23 coded meetings, any statement made has been assigned to one of the 44 observation categories. For each of these statements the onset and offset time, as well as the person making the statement has been noted. There are observation categories which are not related to a verbal statement, such as shared laughter and pause. Pause being the absence of any verbal statement, which has to last either at least ten seconds or has to be identified as a non-dramatic pause for it to be encoded.

Label creation

The same method as described in the study by Nanninga et al.[67] is used here to convert from act4teams coding categories to social and task cohesion labels. This means that for the meeting segment in which cohesion is being assessed, the length of the utterances that correspond to cohesion is divided by the length of all utterances. This produces social and task cohesion values which range between 0 and 1. Seven of the coded categories are established to positively characterize social cohesion, and six of the coded categories to indicate task cohesion. An overview of these categories can be found in Table 4.2.

Table 4.2: Act4teams codes that indicate team cohesion.

Social cohesion	Task cohesion
Offering praise	Goal orientation
Support	Prioritizing
Encouraging participation	Distribution of tasks
Humour	Taking personal responsibility
Laughter	Interest in change
Expressing feelings	Action planning
Active listening	

The distribution of these values for 2-minute windows within all 20 used meetings can be seen in Figure 4.2. As this research is looking for regions of high and low cohesion, and not necessarily the cohesion

[1] https://www.abus.com/eng/Archive/Video-surveillance/Other/Outdoor-Analogue-HD-Dome-IR-1080p-Vario-2.8-12-mm

value itself, these continuous values are converted to labels. However, there is no clear separation between high and low cohesive regions for both the social and task cohesion. Both social and task cohesion have more windows with a low fraction of cohesive statement duration. Task cohesion especially has a lot of windows with close to no task cohesion related utterances.

To take the high and low cohesive parts of the meeting, windows that are not high or low in value should be discarded. This is done by taking the windows with a cohesive value that is in either the top or bottom 25%.

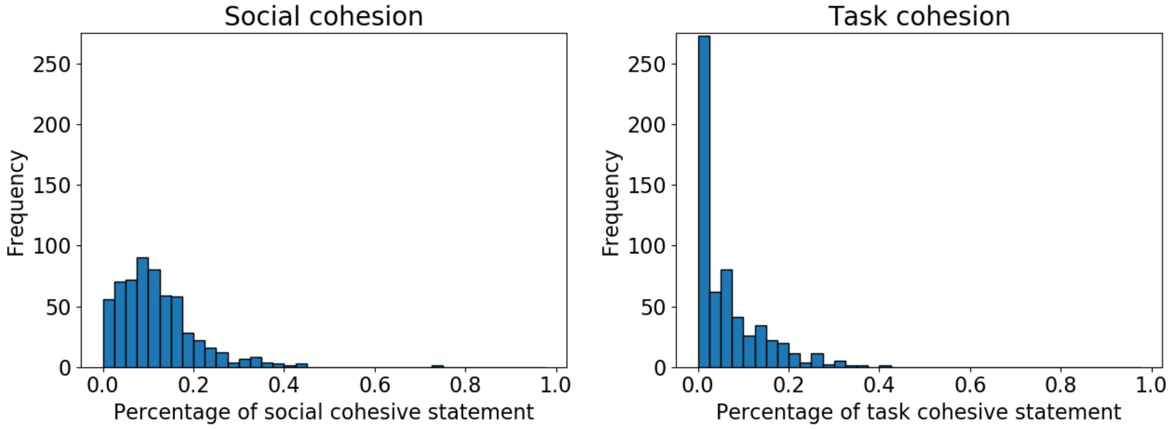


Figure 4.2: Histogram illustrating distribution of cohesion percentage values for 2 minute windows. The left figure shows the distribution for social cohesion, and the right figure for task cohesion.

This method gives us social and task cohesion labels based on verbal expression categories. Social and task cohesion are usually assessed with a questionnaire as shown in section 2.2. **Chapter** 7 discusses how it can be assessed if the labels used in this experiment capture the same concept as questionnaire-based labels.

4.4. Multimodal feature alignment

This section discusses the different steps taken to align the audio and label data with the video and accelerometer data.

The different videos first needed to be aligned among themselves, as the videos were saved with a variable frame rate. A variable frame rate means that the time between consecutive frames is not constant within a video. This means that frame x in one video did not necessarily occur at the same time as frame x in another video. As the actions users take are compared in time, the frames that occurred on approximately the same timestamp within the different videos need to be found. To ensure actions with the same approximate timestamp are compared the videos are converted from variable to constant frame rate. The frame rate is fixed at 25 fps, as this was the camera's original frame rate, and frames with their timestamps closest to the new timestamp are used to construct the new video with ffmpeg. The video recordings were also not started completely synchronously. A person visible on both camera 1 and 2 does not make the same movement at the exact same time. This difference could be more than five seconds in some cases. As we are investigating similarity between participants it is important that the videos recorded with different cameras are in sync. All of the 176 available videos were therefore manually verified and synchronised with each other and the audio by comparing the statements heard in the audio with the mouth movement visible in the video. Manual synchronisation reduced the difference between modalities to less than a second. The video data from one meeting had to be excluded because the participants do not stay within sight of the cameras for enough of the meeting. This leaves us with 19 meetings that have usable video recordings (v-dataset).

To align the accelerometer data with the other modalities, the experimenters were supposed to clap a specific accelerometer on the table after reading out the time the accelerometer showed. For some of the meetings this protocol was either not followed, or this moment is no longer present in the audio/video data. In these cases it was attempted to align the accelerometer data by comparing the video and audio footage to the accelerometer readings. Events like the accelerometer lying unmoving on the table for a few seconds can easily be seen in both accelerometer and audio/video data. In four of the meetings the protocol was not followed and such an event was not found, which means we were unable to align the accelerometer data with the other audio/video data and therefore also the labels.

For 14 out of the 90 accelerometers in the remaining meetings the data could not be retrieved from the device, or the extracted data showed only zeros. The data from these accelerometers was not included in the dataset. This meant for one meeting that the number of participants with valid accelerometer data was only 2. This meeting was also excluded, because we are analysing group dynamics and not pair dynamics. This leaves us with 18 meetings in the accelerometer part of the dataset (acc-dataset).

Table 4.1 shows for each meeting if labels, video data and accelerometer data are available. For 14 meetings we have data available for all modalities, which we will refer to as the shared set. These meetings are highlighted in Figure C.1.

5 Methodology

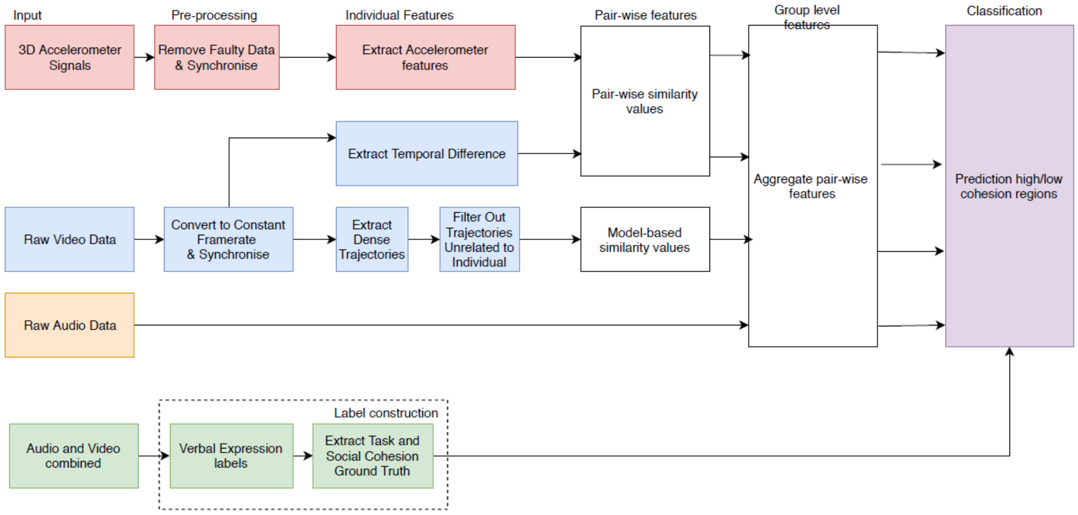

Input
Pre-processing
Individual Features
Pair-wise features
Group level features
Classification
3D Accelerometer Signals
Remove Faulty Data & Synchronise
Extract Accelerometer features
Pair-wise similarity values
Extract Temporal Difference
Aggregate pair-wise features
Prediction high/low cohesion regions
Raw Video Data
Convert to Constant Framerate & Synchronise
Extract Dense Trajectories
Filter Out Trajectories Unrelated to Individual
Model-based similarity values
Raw Audio Data
Label construction
Audio and Video combined
Verbal Expression labels
Extract Task and Social Cohesion Ground Truth

Figure 5.1: Flow chart representing the methodology. The three different modalities were used to find group level similarity values between participants. These values were used to predict if a segment has high/low cohesion according to the verbal expression based labels. The bottom part of the pipeline shows the label extraction pipeline as discussed in section 4.3.

This chapter presents both the unimodal method and the multimodal method developed to predict high/low social and task cohesion in meeting segments. An overview of the multimodal method can be seen in Figure 5.1. The group level audio features are based on paralinguistic mimicry. The extraction has been done by Nanninga et al. [68] and the interested reader is referred to that paper for an in-depth explanation of the method used there. The accelerometer and video data were used separately for the extraction of group level motion similarity features. Each step of this process is explained for both accelerometer and video data within the following sections. Two different video features are extracted, the thresholded pixel-wise difference as a baseline feature similar to that of Hung and Gatica-Perez [46] and the dense trajectories because of their good performance in action detection.

The general steps in the pipeline are the following:

1. Pre-processing.
2. Individual feature extraction.
3. Creation of pair-wise similarity values from individual features.
4. Aggregation of pair-wise similarity values into group-level features.

The pre-processing part of the pipeline contains in both cases a form of time alignment with each other as well as the audio as explained in section 4.4. Comparing the similarity of time series is usually done in pairs (see the examples in section 3.3). Therefore, the comparison is first made for each pair of participants, which are then aggregated into group-level features. The group-level features for both video features and accelerometer are used separately and combined to predict social and task cohesion. How this is done is explained in section 5.5.

5.1. Pre-processing

The temporal difference calculation required no pre-processing except for the time alignment as explained in section 4.4. The pre-processing steps for the accelerometer feature extraction process and the dense trajectory extraction are explained in this section.

Accelerometer features

To eliminate interpersonal differences in the amount of movement in the raw accelerometer signals, each axis is first standardized using the z-score. These standardized values are then used in 3 different ways: the values themselves, the absolute values, and the magnitude of all axes combined. The participants are sitting around a table sometimes opposite each other, which means mirroring a movement can be in a different axial direction. The absolute value is therefore included which removes the within axis direction. The magnitude is a representation of the total amount of movement a participant made and is calculated as:

$$magnitude = \sqrt{x^2 + y^2 + z^2} \tag{5.1}$$

This gives us 7 different signals, the raw and absolute value for each axis and the magnitude for all axes combined.

Video features - Dense Trajectories

For the dense trajectory extraction the only other pre-processing step is that the videos are resized to 960 x 540 pixels to reduce the processing time for each video.

5.2. Individual feature extraction

The individual features are the features extracted to describe an individual participant's behaviour. The features extracted from the accelerometer data represent movement, just like the temporal difference features extracted from video. The dense trajectory feature vector contains several different types of features that are described within the corresponding section.

Accelerometer features

For the accelerometer-based features the output of the pre-processing step could be used as individual level features themselves. Kapcak et al. [51] instead argue for the use of statistical and spectral representations of this data with a sliding window approach as this is common approach for detecting body movement with accelerometer data. This sliding window will have a size of n and will shift by $n/2$ so that consecutive windows are half overlapping. The size n of the sliding window is determined by trying varying sizes and comparing feature performance. The tested sizes are 1, 3, 5 and 10 seconds.

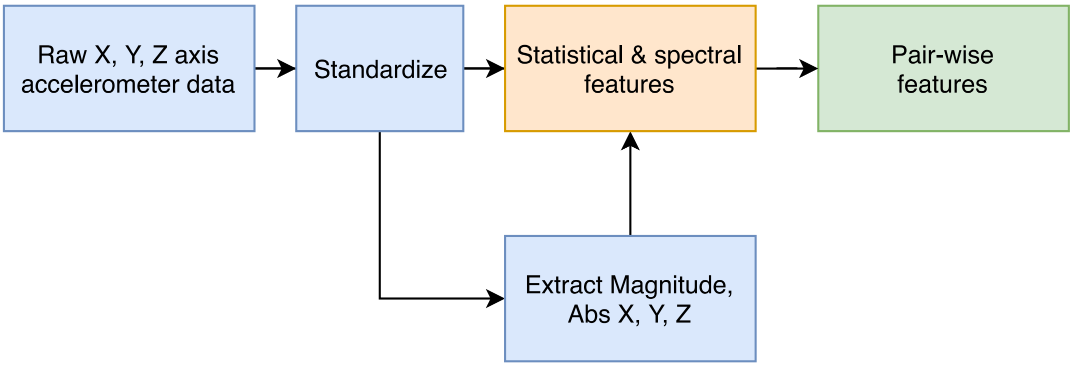

Raw X, Y, Z axis accelerometer data
Standardize
Statistical & spectral features
Pair-wise features
Extract Magnitude, Abs X, Y, Z

Figure 5.2: Flow chart representing the accelerometer pre-processing and individual feature extraction steps. The pre-processing steps are explained in section 5.1 and are shown in blue. This is followed by the individual feature extraction step explained in section 5.2 shown in orange. From these individual features the pair-wise similarity is extracted as explained in section 5.3.

Video features - Temporal Difference

Pixel-wise difference is extracted by taking the absolute difference between pixel values of consecutive video frames. It was explained in Figure 4.2 that up to three different people can be seen on the same camera. It is therefore necessary to separate the extracted features for each individual. A bounding box was therefore defined for each participant that attempts to best separate participants visible in the same video. Only the difference values for the region within the bounding box are used, and a threshold is applied to this difference to get a representation of which pixels have moved place. The number of pixels that have changed over the whole bounding box is used to present the magnitude of motion for a participant between consecutive windows, which we call temporal difference. An example of the (thresholded) pixel-wise difference can be seen in Figure 5.3.

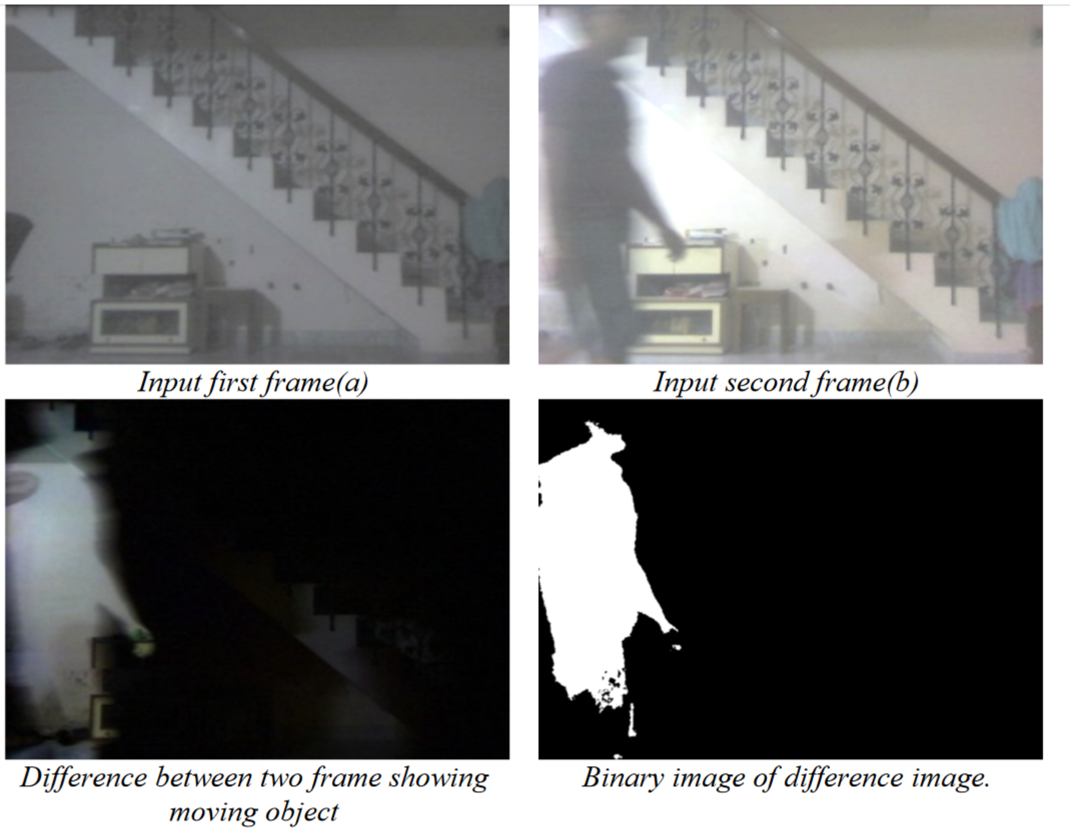

Input first frame(a)
Input second frame(b)
Difference between two frame showing moving object
Binary image of difference image.

Figure 5.3: An illustration of the (thresholded) pixel-wise difference. The difference between the two images on the top row, and a binarization of this difference is shown on the bottom row. Taken from [85].

The same sliding window approach as for the accelerometer features is then used to create statistical and spectral representations of this feature.

Video features - Dense Trajectories

From each of the synchronised videos the camera that the participant faces most straight-on was used to extract Dense Trajectories for each person. This approach samples points densely from each frame and tracks them using a dense optical flow field. The background is static, and the camera unmoving, which means most flow present should be an actual person moving. The features that are extracted from these trajectories have been successfully used for action recognition[96]. The default parameters have been tested on many different datasets with good performance throughout[97, 98] and are also used in this experiment. This means the trajectory length (L) = 15, the sampling stride (W) = 5, the neighbourhood size (N) = 32, the number of spatial cells (nxy) = 2, and the number of temporal cells (nt) = 3.

Four different types of feature descriptors are extracted using the dense trajectories: the trajectory of the point, the Histogram of Oriented Gradients (HOG), the Histogram of Optical Flow (HOF), and Motion Boundary Histograms in the x and y direction (MBHx and MBHy). HOG features describe the local object appearance with the distribution of local intensity gradients, which gives a static description of the region around the point[33]. HOF features on the other hand capture local motion information[96]. Motion Boundary Histograms represents the gradient of optical flow. This is mainly useful for the elimination of noise due to background motion. It should be noted that these videos should not have background motion as the camera position does not change, and the background is static. The trajectories are the x and y coordinate of the tracked point over time. Just the HOG descriptors are used as the video features, as we use these features to describe the posture of the participant and not their movement.

The coordinates of the bounding box are used together with the starting point of the trajectory to filter out trajectories that belong to other participants.

5.3. Pair-wise features

We investigate not only how similar each pair is behaving within a meeting segment, but also if this becomes more similar within this segment. These phenomena are called synchrony and convergence respectively by both Kapcak et al. [51] and Nanninga et al. [68], and are referred to with the same names in this section for ease of reference. Because the accelerometer and temporal difference features are represented in statistical and spectral values these approaches still measure mimicry.

Pair-wise similarity values

The accelerometer features and video-based temporal difference features use the same pair-wise similarity extraction method as shown in Figure 5.1. This method is based on the synchrony and convergence extraction methods defined by Kapcak et al. [51] for the detection of attraction with accelerometer features. The same approach is used for the temporal difference features because of their similarity in meaning (amount of movement), and to compare performance between the 2 feature types.

Synchrony

Three different metrics are used that measure the amount of similarity.

(Normalized) Mutual Information. Mutual information is calculated as:

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \tag{5.2}$$

where H(X) and H(Y) are the marginal entropy for both variables and H(X, Y) is the joint entropy. As can be seen in the formula this metric is a quantification of the amount of information that can be obtained about one variable by observing the other. From this the Normalized Mutual Information can be calculated which gives a score between 0 and 1.

$$NI(X;Y) = \frac{I(X;Y)}{\sqrt{H(X)H(Y)}} \tag{5.3}$$

A higher score will mean the pair's behaviour is more similar.

Correlation. A common comparison method for two signals is correlation as shown in subsection 3.3.1. The Pearson correlation is calculated as:

$$r_{x,y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{5.4}$$

The x and y representing the features for the two different participants. This results in a value between -1 and 1. A value closer to 1 means the features for both participants are more linearly related and are therefore more similar.

Mimicry. When a person is mimicking someone, they are imitating their behaviour with a delay. To measure short term mimicry the values from consecutive windows are compared by calculating the distance between these windows as previously done by Kapcak et al. [51]. This is done for all sliding windows within a 2 minute meeting segment and from all these values the minimum, maximum, mean and variance are taken as a representation.

This measure is asymmetric meaning you get different values for comparing participant A to B and participant B to A. As it should not matter which participant is taken to be A and which one to be B the mimicry comparison with the highest maximum is added to the feature vector first followed by the one with the lower maximum.

Convergence

To measure if participants get more similar within a meeting segment two types of convergence are measured: global and symmetric convergence.

Global convergence For global convergence this difference for each sliding window is calculated and summed for the first half and second half of the segment. The difference between these sums is calculated to evaluate if the difference has increased or decreased. A higher difference means the participants have become more similar in the first half compared to the second half of the meeting segment. An illustration of this method can be seen in Figure 6.9.

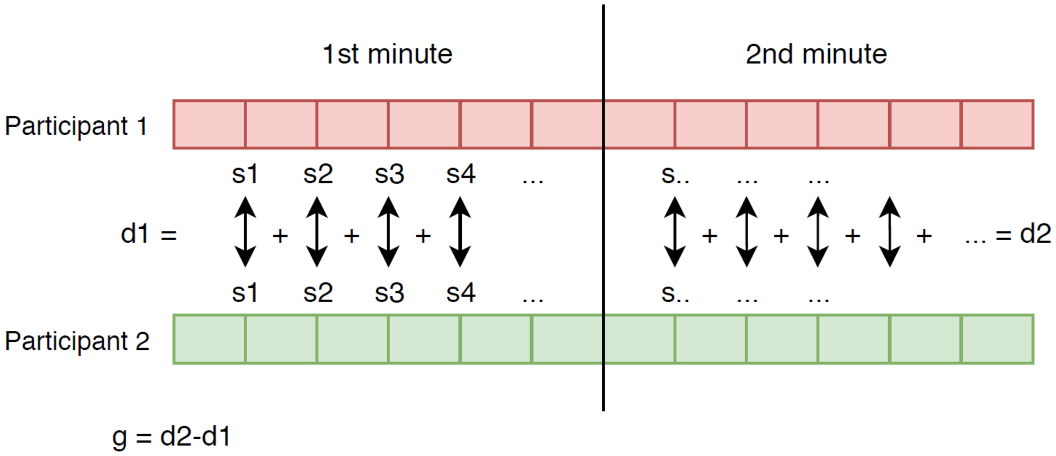


Figure 5.4: An illustration of the calculation of the **global convergence** value. The distance between participants for each of the overlapping sliding windows (s) in the first half of the meeting segment is summed. This is also done for the second half of the meeting segment. These values are then subtracted to get the difference between both halves.

Symmetric convergence For symmetric convergence this difference for each sliding window is used together with the timestamp of the sliding window as input for a Pearson correlation calculation. A negative Pearson correlation coefficient would mean the difference decreases over time and therefore the participants have become more similar. An illustration of the symmetric convergence can be seen in Figure 5.5.

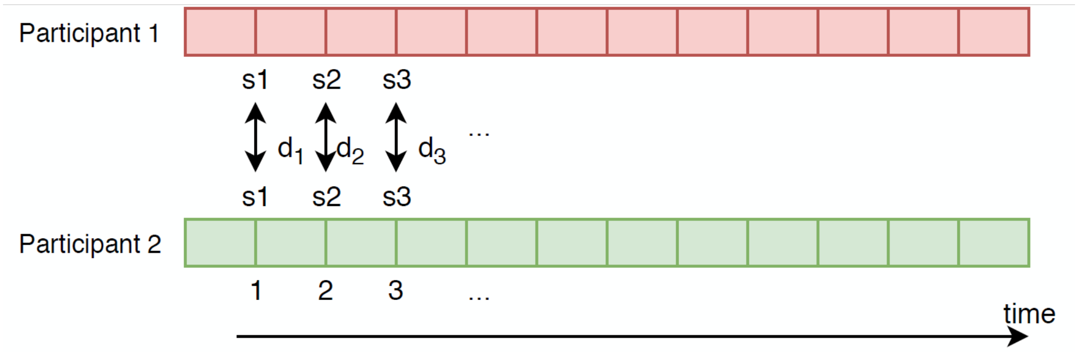


Figure 5.5: An illustration of the calculation of the **symmetric convergence** value. The distance between participants for each of the overlapping sliding windows (s) is calculated. The correlation between these distance values over time is taken to be the symmetric convergence rate.

Model-based similarity values

For dense trajectories a different pair-wise comparison method was used than for the accelerometer features and video temporal difference features as can be seen in Figure 5.1. This approach was chosen because it is more common to create a modelled representation of the appearance features like described in subsection 3.3.5. The extracted dense trajectory features are used to create a model for each meeting section for each participant. If we feed this model a feature vector it can tell us if this movement is likely to occur based on previously shown data. By applying the features from another person on this trained model we can see how well they match with the movements of this other person by obtaining the likelihood. This approach is inspired by the one used by Nanninga et al. [68] to compare paralinguistic mimicry features between participants. A Gaussian Mixture Model (GMM) will be trained to learn a representation for a participant. The number of components is kept constant over all windows and all participants. To select the number of components for the GMM the Bayes Information Criterion (BIC) is utilized. A model will always fit better with a higher number of components, but the BIC penalises for a higher number of parameters used in the model[100]. To determine the number of components 10 participants are randomly selected as a representation of the average participant. Some people participated in several meetings, with 3 participants partaking more than twice. The choice was made to only include each participant once to potentially get more different types of motion. On their samples a model is trained with 1 to 20 components in steps of 2. This random selection is made to decrease the amount of time needed to select the number of components while still representing the average participant. The difference between the average BIC for each number of components is plotted to determine the optimal value.

Synchrony

A model with the selected number of components is then trained for each participant for each meeting section. This model is a representation of the movement of the participant within that meeting section. The average likelihood of the samples from another participant belonging to this model are used to represent the pair-wise similarity between participants within that meeting segment. This comparison method was chosen to measure a form of synchrony between participants, as is also done for the accelerometer and temporal difference features.

Convergence

This pair-wise similarity is also calculated over time by finding the likelihood for each timestamp and calculating the Pearson correlation for the frame number and the likelihood. This is a representation of if their posture becomes more similar within that meeting segment and is therefore converging.

5.4. Group-level features

Each pair of meeting participants now has a set of features describing their similarity in the form of synchrony, mimicry and convergence. As cohesion is a group-level dynamic all of these pair-wise features are aggregated into group-level features. The pair-wise features are aggregated in the following ways:

- Maximum value

- Minimum value
- Median value
- Standard deviation

This leaves us with features that describe group-level motion similarity in the form of synchrony, and convergence for accelerometer, temporal difference and dense trajectory-based features. These features will be used to classify the high and low cohesion meeting sections.

5.5. Classification

The group-level synchrony features are used to train a classifier that classifies windows into low or high cohesion. These low and high cohesion labels are created as described in section 4.3 by taking the top and bottom 25% continuous cohesion values as low and high cohesion segments. This 25% threshold is defined on all the meetings with labels. All windows in which no utterances have been made are afterwards removed from the data. The first meeting segment is skipped as one minute of data is needed to train the audio model, and the other modalities only have features for every 2-minute segment combined. As we are interested in predicting the cohesion in new meetings that do not have labels, we split the data so that no segments from one meeting can be in both the training and test set. Most classifiers learn better when the distribution of classes is equal in the training and test set. Because of the small number of meetings and thereby data points we want to use k-fold cross-validation. For this method you divide the data into k-folds of which one is used as test set and the others are used as training set. This is done k times so that each fold is taken as the test set once. Cross-validation reduces over-fitting and therefore gives a more accurate estimation of performance on previously unseen data[24]. For these reasons we chose to use stratified group k-fold cross validation[1]. This method ensures no data from the same meeting can be both in the training and testing fold, and also tries to make the distribution of classes equal in the training and test set.

As we are not just interested in getting a good prediction model, but also in interpreting which specific behaviours predict cohesion the choice was made to use a simple classifier. This allows us to look at the weights of each feature and identify the features contributing more to successful classifications. The choice was therefore made to train a Logistic Regression model to classify high and low cohesion using the group synchrony features.

For the multimodal version of the pipeline the certainty of the label probability is averaged over the different modalities.

Not all meetings have data for all modalities, therefore the combined approach is tested only on those meetings that have all modalities available. The performance of the individual modalities when training on all available meetings is also included.

5.6. Evaluation

The performance will be evaluated by looking at the Area Under the Curve (AUC) of the Receiver Operating Characteristic curve (ROC-curve) for both the unimodal and multimodal approach. The ROC-curve represents the true positive rate against the false positive rate at different classification thresholds. An AUC of 0.5 for our two class classification problem means we have a performance equal to a random classifier. A higher AUC means when gaining more true positives you gain less false positives. This metric has been previously used by Nanninga et al. [68] to evaluate the performance of the paralinguistic mimicry features on the same data set.

[1] source code: `https://www.kaggle.com/jakubwasikowski/stratified-group-k-fold-cross-validation`

6

Results

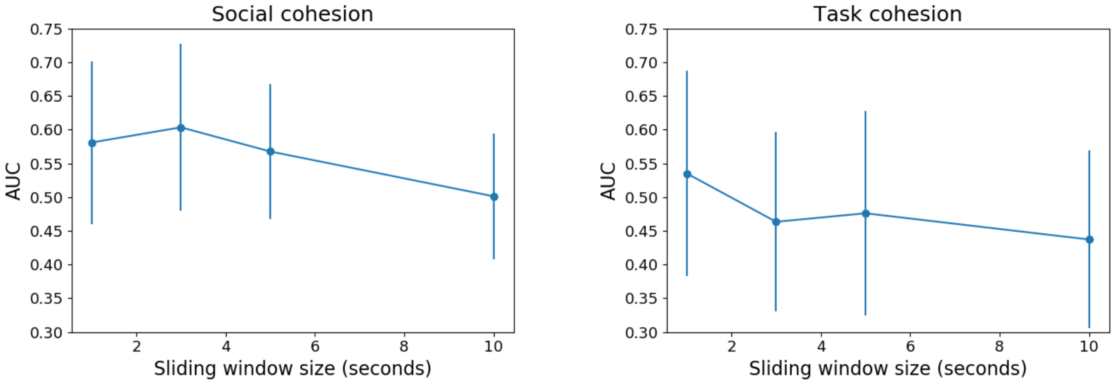
This chapter shows the results for each of the executed experiment. The first experiments are executed to determine some of the settings for the feature extraction process. The second set of experiments evaluates the classification performance for the unimodal models. The third set of experiments compares the performance of a multimodal model with that of unimodal models. The last experiments are done to help analyse the performance of the different feature types.

6.1. Design choices

As explained in section 5.2 the size of the sliding window size for the statistical and spectral representations needs to be determined, which is done in the first experiment. The second experiment determines the number of components in the Gaussian Mixture Model for the HOG features as explained in section 5.3.

6.1.1. Size sliding window

Part of the accelerometer pipeline is the creation of statistical and spectral features for sliding windows. As described in section 5.2 different sliding window sizes are tested to see which one performs best. The tested sizes are 1, 3, 5 and 10 seconds which are the same sizes tested by Öykü Kapcak [106] for the attraction prediction experiment. A toy model was trained for each of the different window sizes to evaluate the change in performance. The average area under the ROC-curve and the range of values within different classification folds can be see in Figure 6.1. We can see for social cohesion that the best classification happens with a window size of 3 seconds and sizes higher than that cause a decrease in performance. The task cohesion classification is not that good in general. The choice was therefore made to select a sliding window size of 3 seconds. The same sliding window size is used for the Temporal Difference features.



(a) Social cohesion

(b) Task cohesion

Figure 6.1: AUC for sliding windows of size 1, 3, 5 and 10 seconds.

6.1.2. Determining number of mixture components for HOG

As explained in section 5.3 on the pair-wise dense trajectory features we are creating a Gaussian Mixture Model(GMM) that describes the motion of each participant within a 2-minute meeting segment. The number of components for these GMMs needs to be selected so that it best describes the data without using too many parameters. Adding more components will always fit better on the input data but might over-fit. We therefore analyse the fit of the model using the Bayes Information Criterion (BIC) which penalizes for the number of parameters. This section presents the normalized BICs of modelling with different numbers of components for 10 randomly selected participants. These models were created with the HOG features of the available dense trajectory features, as these are used to model appearance. In Figure 6.2 we can see the normalized BICs for 1 through 19 components with steps of 2. The increase in BIC is lower for the diagonal covariance matrix than for the full covariance matrix for each increase in number of components. The error bar shows the standard deviation across the different meeting segments. The difference between consecutive values can be seen in Figure 6.3.

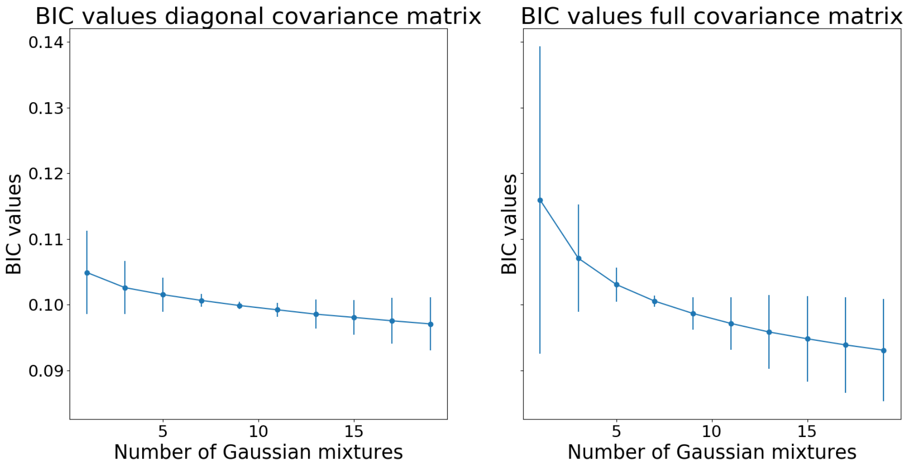


Figure 6.2: BIC values for 1 through 19 components with steps of 2. The left shows the models with a diagonal covariance matrix, and the right with a full covariance matrix.

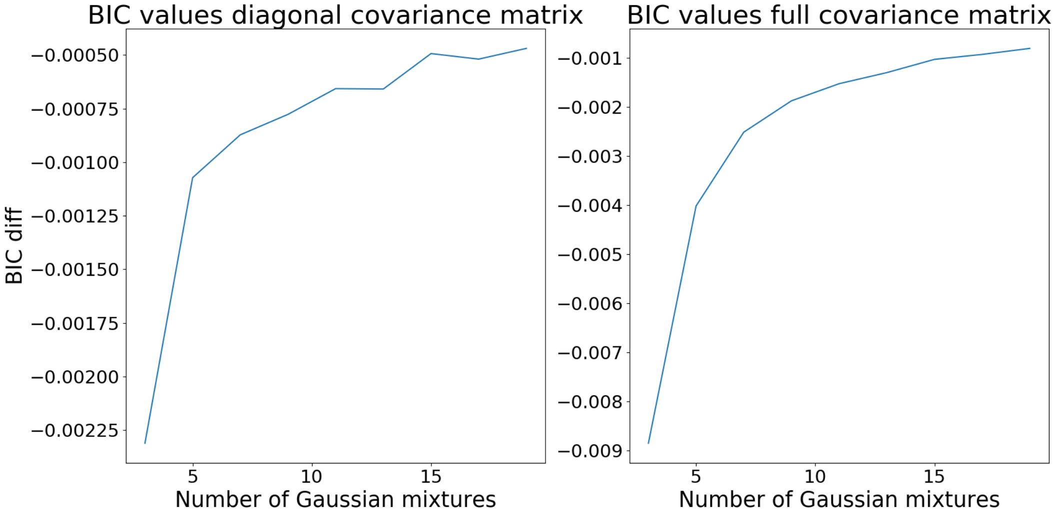


Figure 6.3: Difference between consecutive BIC values for 3 through 19 components with steps of 2. The left shows the difference for the models with a diagonal covariance matrix, and the right with a full covariance matrix. The scale is different for both plots to better show when the difference starts to decrease for both.

These models show a very similar average normalized BIC for both the full and the diagonal covariance matrix models. A good number of components is selected by determining where the benefits of adding more components greatly decreases. The chosen number of components for a model with a full covariance matrix would therefore be 7, and for the diagonal covariance matrix 9. This is where the positive difference of BIC for adding more components has decreased and the standard deviation is the lowest. To make the choice between the full and the diagonal covariance matrix we use the average values of the non-normalized BIC which are $6.959 \cdot 10^6$ and $7.712 \cdot 10^6$ respectively. As a lower BIC means a model has a higher likelihood for the given number of parameters, the 7-component full covariance matrix model is chosen.

6.2. Unimodal classification

For the first classification experiment Stratified 5-fold group cross-validation is executed on all data available for each modality (see Table 4.1). The classification performance is measured with the average area under the ROC-curve as explained in section 5.6. The subset of the dataset that has accelerometer data available will be referred to as *acc-dataset*, the subset that has video data available *v-dataset*. Audio data is present for all meetings that are labelled. The meetings that have data on all modalities will be called the *shared set* as previously stated in section 4.4.

Accelerometer Figure 6.4 shows the results for both social and task cohesion for all modalities. The accelerometer features detect social cohesion with an average AUC of 0.64 and task cohesion with 0.52 on the acc-dataset. This suggests social cohesion can be detected decently with accelerometer-based motion similarity feature. An AUC of 0.52 is close to random guessing and therefore task cohesion does not seem to be detectable with the accelerometer features on the acc-dataset.

Video The performance of the temporal difference features and HOG features is evaluated on the v-dataset. For both of these features types we see that task cohesion cannot be detected on this dataset as their average AUC is 0.52 for the temporal difference features and 0.44 for the HOG features which is close to random guessing or worse. Social cohesion is detected with an average AUC of 0.57 for the temporal difference and 0.55 for HOG.

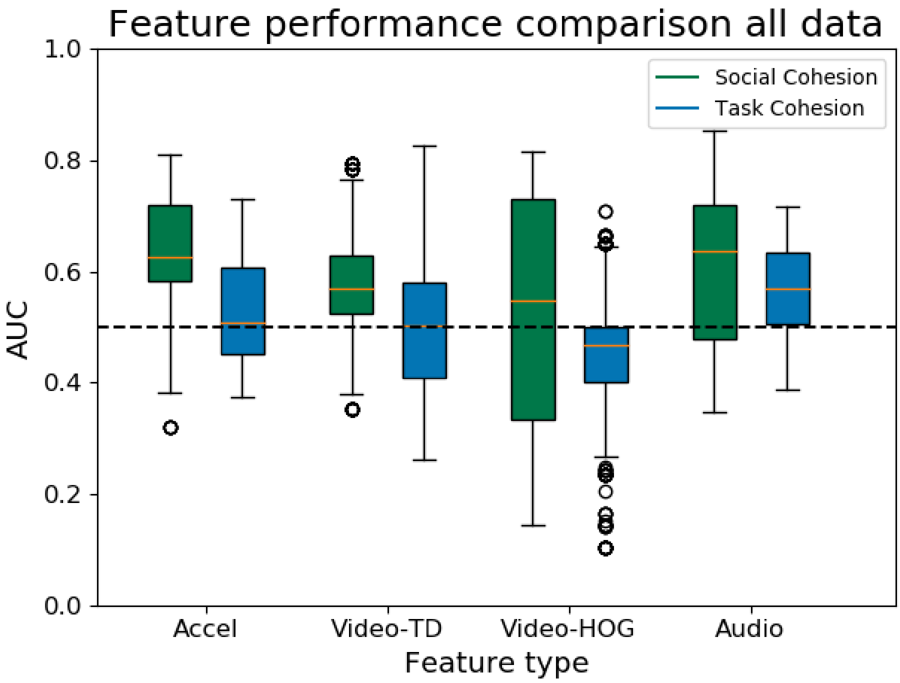


Figure 6.4: AUC Boxplots of social and task cohesion for all different features. Social cohesion is the left(blue) for each feature and task cohesion is right(green).

Audio The AUCs of the audio features have also been calculated on all the available data to verify the original results of the study by Nanninga et al. [68]. The results are different than those of Nanninga et al. [68] on the same part of the dataset with the same features where an average AUC of 0.69 was reported for social cohesion and 0.63 for task cohesion. Our experiment shows an AUC of 0.63 for social and 0.60 for task cohesion. The only difference in the set-up is that the first meeting segment starts at 2-minutes into the meeting instead of 1-minute into the meeting. This can change the labels for all the following segments, and these segments are apparently more difficult to classify for the paralinguistic mimicry features.

Each feature predicts social cohesion better than task cohesion on average. All features but the paralinguistic mimicry features from audio perform close to or worse than random for detecting task cohesion. The audio features might perform best for detecting task cohesion because it has the largest amount of training data available, since all the coded meetings have audio data. Social cohesion is detected with an average AUC of 0.64 for the accelerometer, 0.63 for the audio features and 0.57 for the Temporal Difference features which is all better than random guessing. Because a different set of meetings is available for each modality, it is not possible to definitively conclude which modality detects social and task cohesion the best. Some of the more difficult to classify meetings might be included for some modalities and not for others. We can however conclude that paralinguistic mimicry is a decent detector of both social and task cohesion, whereas motion similarity from accelerometer and temporal difference detect social cohesion. The appearance-based HOG features seem to detect neither.

To enable a fairer comparison between the different modalities, we also test the classification performance for the shared set. The average ROC-curve for each feature type for the shared set is shown in Figure 6.5.

Accelerometer The detection performance of accelerometer features for social cohesion is the same for the shared set as on the acc-dataset with an average AUC of 0.64. The average AUC of the accelerometer features for task cohesion has however increased to 0.63, which would indicate motion mimicry features from accelerometer data can also be used to detect task cohesion in the shared set.

Video The performance of both video feature types increases for the detection of social cohesion in the shared set. The temporal difference shows a 0.05 increase in the average AUC with 0.64 on the shared set compared to 0.57 on the v-dataset, and the HOG shows an average AUC of 0.56 on the shared set which is 1 point higher than the 0.55 on the v-dataset. This suggests something in the meetings that are in the v-dataset but not in the shared set makes them difficult to classify. There are 5 meetings that are in the v-dataset but not in the shared set. Visual analysis of these meetings shows one of these meetings contains behaviour that can cause this meeting to be an outlier when compared to the average meeting. One of the participants in this meeting is present in the meeting room but does not partake in the meeting at all. This would impact the minimum group-level value as this person would be very dissimilar to the other participants, but likely in a different way than the normal minimum pair would be.

Audio The audio features seem to suffer from the smaller amount of data available in the shared set. The average AUC drops from 0.63 and 0.57 in the full dataset, to 0.59 and 0.48 in the shared set for social and task cohesion respectively. By using only the meetings with all features available instead of all that have audio 8 meetings were removed from the 23 total, and a large amount of high task cohesion windows were removed as can be seen in Figure C.1.

The detection of task cohesion has increased for both the accelerometer and temporal difference features to now be better than random at 0.63 and 0.58. Accelerometer features now have about the same average AUC for social and task cohesion. A possible explanation for this increase in task cohesion detection performance is the distribution of the fraction frequencies on which the labels are based. Task cohesion has a more left-skewed fraction frequency than social cohesion (see Figure 4.2), which makes high and low cohesion segments less separable for this form of cohesion. The increase in performance could be caused by having dropped the less separable meeting segments for these features, because the meetings do not contain all modalities.

Conclusion: The accelerometer features are the best detector of both social and task cohesion on the shared set. These features have been previously shown a high performance for the detection of attraction, which is a similar concept to social cohesion, and it is therefore logical that they are the best performing features for the detection of social cohesion. The video-TD features measure a similar form of mimicry to the accelerometer features and their equal performance for social cohesion was therefore also within expectations. The audio features are a decent predictor of both social and task cohesion on the full dataset, but their performance has decreased on the shared set possibly because a lack of training data, or an increase in difficulty of separability between the two classes. The ROC-curve for the HOG features never lies above the curves of all other modalities, and does not show a great detection performance.

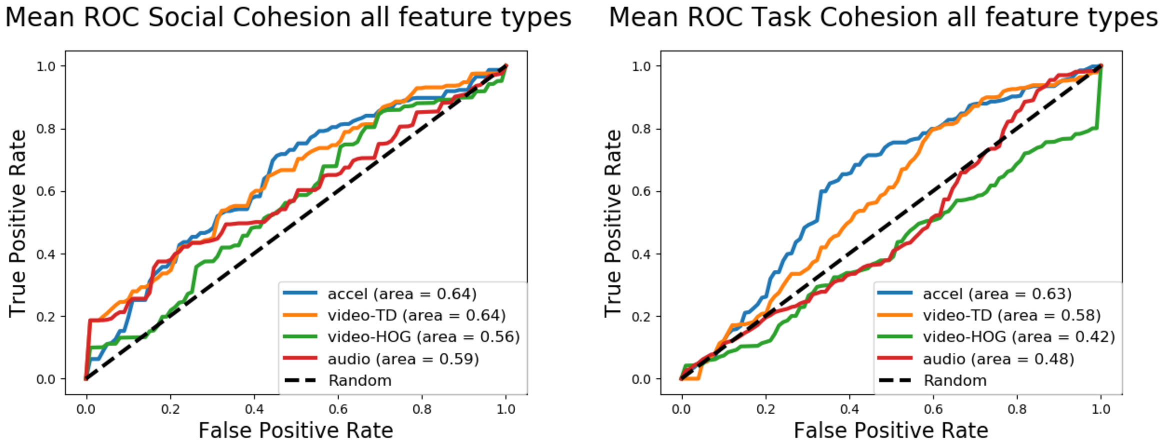


Figure 6.5: Average ROC-curve for each feature type. Social Cohesion is show on the left and Task Cohesion on the right. The legend shows the average AUC for each feature type.

6.3. Multimodal classification

In this section the late fusion of classification models for the different modalities is tested to see if a multimodal detection method outperforms a unimodal detection method. To assess the effect of combining all different modalities we do not need to use both video features. Because the temporal difference and HOG features are both extracted from video, and the HOG features show worse performance than the temporal difference features we do not include the HOG features in our exhaustive analysis. For completeness we do include the detection performance of the combination of accelerometer, temporal difference, audio and HOG. The ROC-curves of all combinations are shown in Figure 6.6.

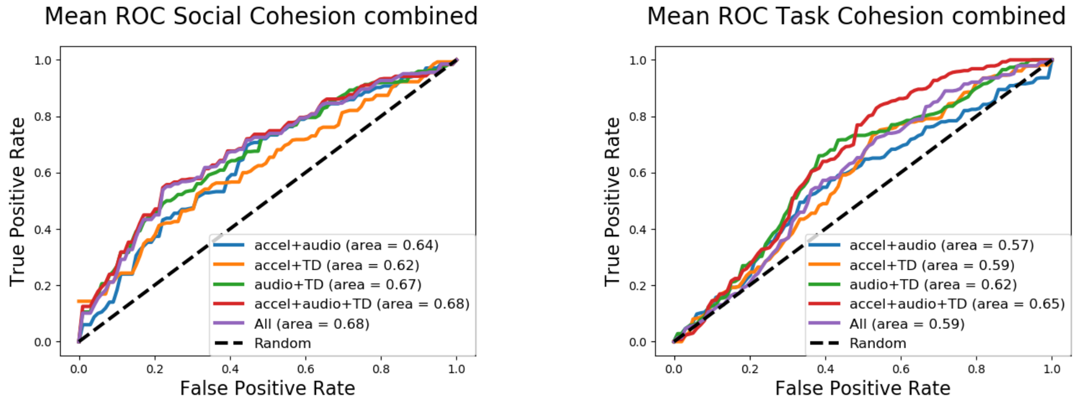


Figure 6.6: Average ROC-curve for all combinations except those with HOG. Social Cohesion is show on the left and Task Cohesion on the right. The legend shows the average AUC for each combination.

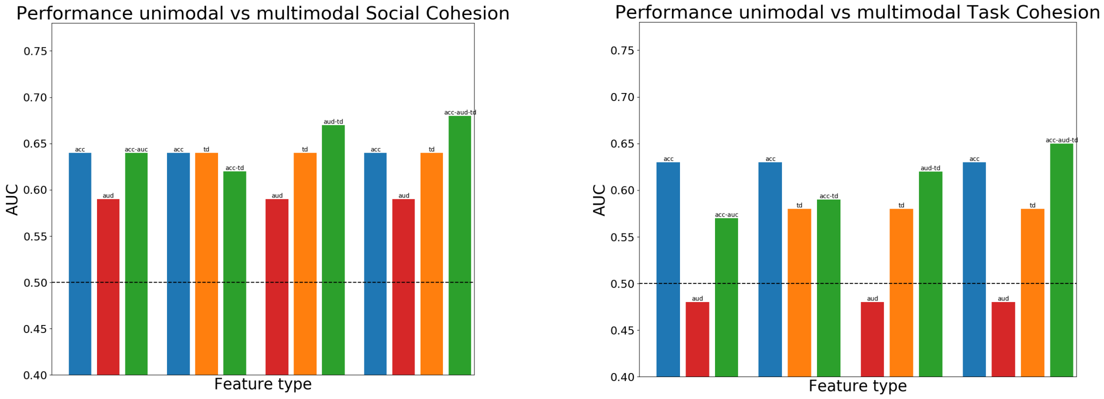


Figure 6.7: Average AUC of each modality combination preceded by the performance of the individual modality. The left figure shows social cohesion, and the right figure shows task cohesion.

Figure 6.7 shows the average AUC of the individual modalities followed by the their combination. We can see that the combination of audio and temporal difference features gives the best average performance for two modalities, and the combination of all 3 modalities always gives the highest AUC. The individual detection performance for task cohesion of the audio and temporal difference video features was not that high at 0.48 and 0.58 respectively, but their combination surprisingly proves beneficial with an average AUC of 0.62. As their combination is beneficial for the detection of both social and task cohesion this suggests there is indeed complimentary information in the audio and video modality. The accelerometer and temporal difference feature both measure movement similarity, and it is therefore less surprising that their combination does not improve over the performance of the accelerometer data alone.

The combination of motion-based mimicry features from accelerometer and video data, and the audio features give the best average classification performance, which suggests the combination of all different modalities does contain complimentary information. However, the increase in performance is just 0.01 for social cohesion when compared to the combination of audio and temporal difference, and most likely not significant. We can therefore not conclude that the combination of all modalities provides a significant increase in average AUC for social cohesion, but the combination of features representing the two different mimicry types does provide beneficial for the detection of social cohesion. The performance for task cohesion when using all modalities is higher than that of the accelerometer features by themselves, which suggests a multimodal model detects task cohesion better than a unimodal model.

Adding the video HOG features to the final combination does not further increase on the combination of accelerometer, video temporal difference and audio features, which suggests appearance mimicry does not complement the other forms of mimicry for the detection of cohesion.

Conclusion: The late fusion of all unimodal models (except HOG) predicts social and task cohesion the best out of all models. This suggests different modalities compliment each other when detecting cohesion.

6.4. Correlation analysis

To gain insight into features that are most related to either social or task cohesion we execute a correlation analysis. This is done for both the accelerometer features, and the video Temporal Difference features as they show the best average AUC on the shared part of the dataset. An extensive analysis of the performance and correlation of the audio features can be found in the thesis of Nanninga [67].

In Table 6.1 we see the 10 accelerometer and video-TD features with the highest absolute correlation and $p < 0.05$ with the social cohesion labels, and Table 6.2 shows the same for task cohesion. The group minimum value is the most common in all tables and is always negatively correlated with the label. One might assume that a group would work together better when they are more similar. However, the speaker is not separated from the rest of the group, and a low minimum might be caused by the rest of the group being very dissimilar to the speaker by silently listening while the speaker shows more movement. This would also be reflected in a higher standard deviation which is indeed positively correlated with social and task cohesion.

For social cohesion we can see a positive correlation with the median and maximum group values for the video-TD features. This suggests that if the non-speakers of the group are more similar their social cohesion is also similar, because this would be reflected in a higher median and maximum. Mutual information is the pair-wise correlation value that is most commonly found in the top 10 correlation accelerometer features. Mutual information is a metric that quantifies how much about one signal can be known by knowing the other, and is therefore more a similarity than a mimicry metric.

Table 6.1: The 10 accelerometer and video-TD features with the highest correlation coefficient with the ground truth labels for **social cohesion**. All features are significantly correlated with $p < 0.01$. NMI: Normalized mutual information, MI: Mutual information, g_conv: Global convergence, s_conv: Symmetric convergence, corr: Correlation, mim: Mimicry, Mag: Magnitude, Var: Variance Y: Y-axis, Z: Z-axis, X: X-axis, Abs: Absolute value, Var:Variance, PSD: Power spectral density and the bin number. In each name the first word is the group-level aggregation method, the second is the mimicry measurement and the last word is the individual feature type.

(a) Accelerometer features Social cohesion

	Feature name	corr	p_value
1	min MI psd_Z_6	-0.2932	0
2	min MI psd_X_3	-0.2916	0
3	std NMI var_XAbs	0.2914	0
4	min NMI psd_X_3	-0.2802	0
5	std NMI mean_XAbs	0.2667	0.0001
6	std NMI psd_Z_4	0.2624	0.0001
7	std NMI psd_X_2	0.2596	0.0001
8	min NMI psd_Z_6	-0.2509	0.0002
9	std NMI psd_X_1	0.2469	0.0002
10	min MI var_X	-0.2462	0.0002

(b) Temporal difference features Social cohesion

	Feature name	corr	p_value
1	median MI var_Mag	0.2202	0.0005
2	median MI mean_Mag	0.1857	0.0037
3	max MI var_Mag	0.1831	0.0042
4	min g_conv mean_Mag	-0.1808	0.0047
5	min mim_1_mean mean_Mag	-0.1638	0.0106
6	std g_conv mean_Mag	0.1623	0.0113
7	min mim_2_mean mean_Mag	-0.1612	0.0119
8	max MI mean_Mag	0.1611	0.0119
9	min mim_2_std mean_Mag	-0.1586	0.0133
10	min g_conv var_Mag	-0.1578	0.0138

Table 6.2 shows mimicry is a common entry in the top 10 features with the highest correlation with task cohesion for both accelerometer and video-TD features. This might be because imitation of behaviour is indicative of aligning oneself with the information. This contrasts with the most common pair-wise comparison method for social cohesion which is more of a similarity metric as explained previously.

Table 6.2: The 10 accelerometer and video-TD features with the highest correlation coefficient with the ground truth labels for **task cohesion**. All features are significantly correlated with $p < 0.01$. NMI: Normalized mutual information, MI: Mutual information, g_conv: Global convergence, s_conv: Symmetric convergence, corr: Correlation, mim: Mimicry, Mag: Magnitude, Var: Variance Y: Y-axis, Z: Z-axis, X: X-axis, Abs: Absolute value, Var:Variance, PSD: Power spectral density and the bin number. In each name the first word is the group-level aggregation method, the second is the mimicry measurement and the last word is the individual feature type.

(a) Temporal difference Task cohesion

(b) Accelerometer features Task cohesion

	Feature name	corr	p_value
1	min mim_1_mean psd_ZAbs_4	-0.3308	0
2	min mim_2_mean psd_ZAbs_4	-0.3278	0
3	min mim_2_mean psd_YAbs_5	-0.314	0
4	min mim_1_std psd_YAbs_5	-0.3126	0
5	min mim_2_std psd_YAbs_5	-0.3106	0
6	min mim_1_std psd_ZAbs_4	-0.3097	0
7	min mim_1_mean psd_YAbs_5	-0.3082	0
8	min mim_2_std psd_ZAbs_4	-0.3062	0
9	min mim_1_max psd_YAbs_5	-0.2982	0
10	min mim_2_max psd_YAbs_5	-0.2961	0

	Feature name	corr	p_value
1	std g_conv var_Mag	0.151	0.0085
2	min g_conv var_Mag	-0.1338	0.0198
3	min mim_1_mean mean_Mag	-0.1329	0.0207
4	min mim_1_std mean_Mag	-0.1271	0.0269
5	min mim_2_std mean_Mag	-0.1255	0.0289
6	min mim_2_mean mean_Mag	-0.1253	0.0292
7	min corr var_Mag	-0.1251	0.0295
8	min corr mean_Mag	-0.1208	0.0355
9	max s_conv var_Mag	0.1201	0.0367
10	min mim_2_max mean_Mag	-0.1187	0.0389

Group minimum values are a large part of the features with the top absolute correlation with social cohesion for temporal difference as can be seen in Table 6.1b. This seems to confirm our suspicions that the increase in performance for the shared set is partially caused by excluding the meeting where one participant is not interacting with the group at all as suggested in section 6.2.

6.5. Assessing the impact of an alternate pair-wise comparison method

The results in section 6.2 suggests appearance measured with video HOG-type features from dense trajectories cannot be used to detect either social or task cohesion. This feature type uses a different pair-wise comparison approach than the temporal difference and accelerometer type features as can be seen in Figure 5.1, but we do not know what the impact of this difference in approach is. To evaluate the impact of the pair-wise comparison method we extract a value more similar to the global convergence described in section 5.3 for the appearance features, which were the worst performing features. As HOG are not the only appearance-based features we also include MBH and y-coordinate of the point. Global convergence for the appearance features is calculated by dividing each meeting segment into two halves. A 7-component GMM is trained for a random sampling of 5000 samples for each half of the segment for each participant. For all pairs of participants, the average log-likelihood of applying a random sampling of one to the model of the other is calculated. This is used as a measure of distance. Then the difference of distance between the second and first half is taken as a measure of global convergence within that meeting segment. A schematic representation of this comparison can be seen in Figure 6.8. These features are aggregated into group-level features the same as before. Performance is tested on the shared set for the global convergence of the appearance features, and the accelerometer features as these performed best in the previous experiments.

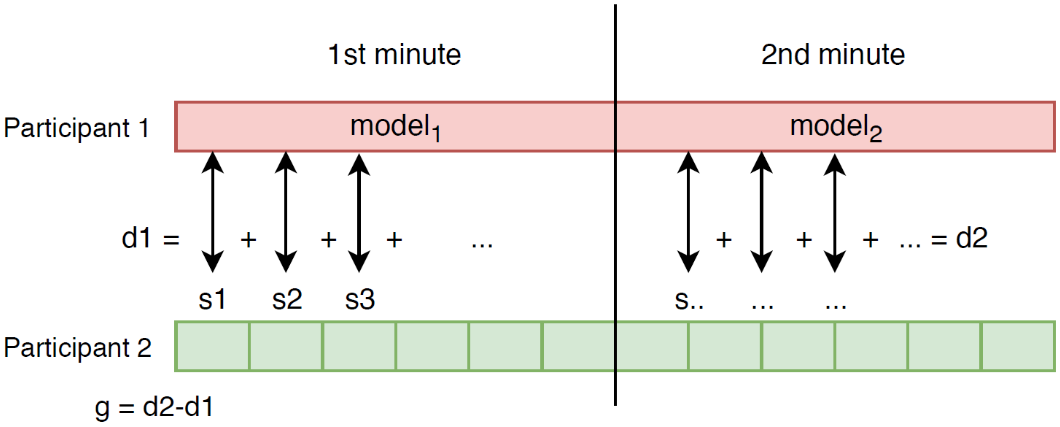


Figure 6.8: An illustration of the calculation of the **global convergence** value for the appearance features. Random samples from the second participant are compared with the model for the first participant and combined to get the distance for each meeting segment half.

The results of the comparison can be seen in Figure 6.9. The appearance features now shown an average AUC of 0.61 for social cohesion compared to the 0.56 when using the other pair-wise similarity methods. They also show a better performance for detecting social cohesion than the accelerometer type features for this specific similarity detection method which shows an average AUC of 0.50. This shows that not only is the choice of feature very important, but also the way similarity between features is measured. Appearance-based similarity seemed a poor predictor of social cohesion, but now shows a better performance than the best performing feature type of the unimodal experiment.

Conclusion: The choice of pair-wise comparison method is very impactful for the performance of the chosen feature. Appearance similarity might be able to detect social cohesion, but different similarity methods need to be tested to draw a more definitive conclusion on how well they perform. Other experiments also need to be conducted to see if they could provide information the other forms of mimicry do not provide, by multimodal testing.

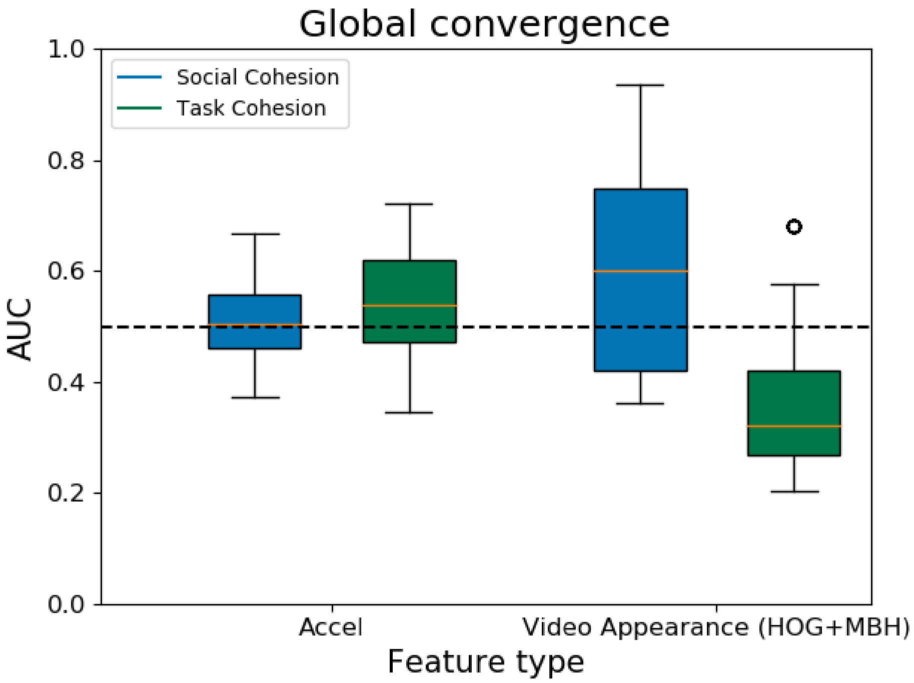


Figure 6.9: AUC Boxplots of global convergence-based features for accelerometer and video appearance features. Social cohesion is the left(blue) for each feature and task cohesion is right(green).

7 Relating verbal expressions to team cohesion

The social and task cohesion labels used in the experiment are constructed based on verbal expression categories present within a specific segment, as explained in section 4.3. The categories that represent social or task cohesion were based on literature, but it has not been verified if the labels that are created with this method agree with standard methods of evaluating social and task cohesion. The standard evaluation method is a questionnaire like the Group Environment Questionnaire[23] developed for sports teams, and a work team version developed by Carless and De Paola[21]. The participants in the meetings conducted for the sense lab experiment were asked to fill in the work team questionnaire version for measuring cohesion. An experiment was conducted that compares the answers to these questions with the frequency of the verbal expression categories across the whole meeting to verify the validity of the labels used in this experiment.

7.1. Correlation questionnaire and coding categories

Social and task cohesion are scored for each participant by averaging over the questions related to either concept, and then averaged across the participants within a meeting to get a general social and task cohesion score for that meeting. This resulted in 23 scores for each type of cohesion. These scores are correlated to the frequency of each coding category within the meeting corrected for meeting duration.

The results of this experiment can be seen in Table 7.1. All coding categories with a p-value of < 0.05 are shown in red. Asking questions (Q) and providing (partial) solutions (S) are the only two categories correlated with task cohesion for $\alpha = 0.05$ (see Table 7.3). Asking questions would be negatively correlated and providing solutions positively correlation with task cohesion. Both Active Listening (AL) and Separating Opinions From Facts (OO) are negatively correlated with social cohesion for $\alpha = 0.05$ (see Table 7.2). This is surprising as (Active) Listening is usually considered as positively correlated with social cohesion[16, 17, 46]. The validity of this comparison should be questioned as the number of data points is very limited, and there is little variability in the cohesion values from the questionnaires (standard deviation social cohesion = 0.4203, task cohesion = 0.3255). From these results we cannot conclude that the categories used to create the cohesion labels in this experiment are correlated with cohesion.

7.2. External observer experiment

Because categorizing verbal statements is an external observation, the constructed labels could be more closely related to external observations of cohesion. Asking external observers to rate social and task cohesion would also allow for the creation of more labels if segments of the meeting were annotated instead of the whole meeting. An experiment was therefore constructed that would have external annotators answer questions that would determine their perception of the teams' cohesion within a meeting segment. To limit the amount of time required for the annotators but still get a good impression of the whole meeting, three sections were labelled for each meeting. These sections are two minutes long to match the length of the sections used in the automatic annotation experiment. The first section starts at four minutes into the meeting to ensure that the set-up that is sometimes still ongoing during the first part of the video is not in any of these sections. The other two sections start at 10 and 15 minutes, because the shortest meeting lasts 22 minutes.

Table 7.1: Correlation analysis questionnaire results and category frequency.
Red shows p-values under 0.05.

Table 7.2: Social cohesion

Code	p_value	Correlation coefficient
A	0.065	-0.391
AL	0.021	-0.478
B	0.836	0.046
C	0.575	-0.123
CB	0.835	0.046
CL	0.225	-0.263
CP	0.800	0.056
CR	0.555	-0.130
CS	0.572	0.124
D	0.995	-0.001
DA	0.988	-0.003
DE	0.376	-0.194
DP	0.896	-0.029
DS	0.198	0.279
E	0.887	-0.031
END	0.619	0.110
ET	0.673	-0.093
F	0.124	-0.330
FB	0.140	-0.317
G	0.069	-0.386
H	0.393	-0.187
HI	0.516	-0.143
I	0.755	-0.069
IN	0.914	-0.024
IS	0.578	-0.122
LA	0.129	-0.326
NC	0.337	-0.210
NF	0.520	-0.141
NI	0.948	0.014
OK	0.979	-0.006
OO	0.043	-0.425
OP	0.308	-0.222
P	0.969	-0.009
PA	0.392	0.187
PRIO	0.688	-0.089
PRQ	0.254	-0.248
PRS	0.885	0.032
PS	0.864	-0.038
Q	0.732	-0.075
RES	0.511	-0.144
S	0.551	0.131
SELF	0.928	-0.020
SIDE	0.559	0.129
SUM	0.712	-0.081
T	0.333	-0.211
TD	0.185	-0.286
TM	0.447	-0.167
VIS	0.561	-0.128
W	0.302	-0.225

Table 7.3: Task cohesion

Code	p_value	Correlation coefficient
A	0.608	-0.113
AL	0.773	0.064
B	0.983	-0.005
C	0.757	-0.068
CB	0.385	0.190
CL	0.561	-0.128
CP	0.283	0.234
CR	0.523	-0.140
CS	0.495	0.150
D	0.267	0.242
DA	0.444	-0.168
DE	0.323	-0.216
DP	0.161	0.302
DS	0.071	0.384
E	0.586	-0.120
END	0.075	-0.378
ET	0.130	-0.325
F	0.961	0.011
FB	0.100	-0.351
G	0.396	0.186
H	0.054	-0.406
HI	0.910	0.025
I	0.787	0.060
IN	0.331	0.212
IS	0.697	-0.086
LA	0.077	-0.377
NC	0.174	-0.294
NF	0.259	-0.246
NI	0.075	0.378
OK	0.588	0.119
OO	0.176	-0.292
OP	0.951	0.014
P	0.273	0.238
PA	0.270	-0.240
PRIO	0.184	0.287
PRQ	0.949	-0.014
PRS	0.869	-0.036
PS	0.589	0.119
Q	0.042	-0.428
RES	0.486	0.153
S	0.008	0.540
SELF	0.629	0.106
SIDE	0.260	0.245
SUM	0.676	0.092
T	0.353	0.203
TD	0.717	0.080
TM	0.384	-0.191
VIS	0.512	-0.144
W	0.096	-0.355

The observers were asked to rate four statements for each type of cohesion on a 5-point Likert scale. These statements were either based on the questions in the work team cohesion questionnaire by Carless and De Paola [21] and adjusted to be suitable for an external observer, or based on findings by Braaten [17], Griffith [41], and Carron and Brawley [22]. A 5-point Likert scale was chosen to stay consistent with the original scale used in the after-meeting questionnaire. The statements that are rated for each cohesion type can be found in Appendix D.

Two annotators rated the three segments for all meetings in the shared set, except for one that was in Dutch, which gives us 42 annotated segments. The scores for each question that related to either social or task cohesion was averaged to get an overall cohesion score for the given segment for that annotator. The second task cohesion related question was first inverted as a positive answer to this question would have a negative effect on task cohesion. The social and task cohesion scores for each segment of both annotators was also averaged to get the average social and cohesion scores for that segment.

Correlation analysis was done for the coding category frequency and the external annotator social and task cohesion scores. The frequency counts were not normalized because each segment lasts the same amount of time. The results can be seen in Table 7.5 with all correlations with a p-value of < 0.05 shown in red. Code C has no value as this coding did not occur in the externally annotated meetings. All the significant correlations ($\alpha < 0.05$) have a positive correlation between the coding category and the cohesion score. For social cohesion this is only Identifying a (partial) solution (S) with a correlation of 0.396. For task cohesion four coding categories have a significant correlation which are: Defining the Objective (D), Interest in Change (IN), Identifying a (partial) solution (S), and Reference to Specialists (W).

Table 7.4: Inter-rater reliability between the two annotators measured with Cohen's κ for each question. Also shows 95% Confidence Interval.

Question	Cohen's κ	95% CI
tc1	0.114671	(-0.080223, 0.3096)
tc2	0.2202970	(0.0077167, 0.4483)
tc3	0.030769	(-0.095394, 0.1569)
tc4	0.230769	(0.042009, 0.4195)
sc1	0.187500	(-0.041342, 0.4163)
sc2	-0.027972	(-0.197244, 0.1413)
sc3	0.212500	(-0.017774, 0.4428)
sc4	-0.031250	(-0.222002, 0.1595)

The inter-rater reliability is also calculated to determine de degree of agreement among raters for the different questions. Cohen's kappa is calculated for each question, the Kappa and its 95% confidence interval can be seen in Table 7.4. The inter-rater reliability is not very high which suggests the posed questions might be difficult to answer consistently for external observers. This could be further confirmed by having more annotators answer the same questions.

7.3. Conclusion

We were unable to definitively conclude that the labels created from the act4teams coding are correlated with the more commonly used ways of measuring social and task cohesion. A better link might be established by looking not just at the duration of the statements, but also the order of statements. The external observer experiment suggests there is a link between certain categories and social or task cohesion, but the inter-rater reliability was low. The questionnaires filled-in by the participants would not have this problem, but a whole meeting cohesion label is too large a scale to compare to coding categories. We therefore suggest refining the external annotator questions to increase inter-rater reliability on observed cohesion, and compare the results of such an experiment with not only the frequency count of these categories, but also consider the interaction pattern by using the order of statements in some way.

Table 7.5: Correlation analysis external annotator labels and category frequency.
Red shows p-values under 0.05.

Table 7.6: Social cohesion

Code	p_value	Correlation coefficient
A	0.212	0.197
AL	0.190	-0.206
C		
CB	0.317	-0.158
CL	0.453	0.119
CP	0.607	-0.082
CR	0.911	-0.018
CS	0.664	0.069
D	0.439	0.123
DA	0.152	0.225
DE	0.509	-0.105
DP	0.606	0.082
DS	0.116	0.246
E	0.141	0.231
ET	0.102	0.256
F	0.393	0.135
FB	0.075	0.278
G	0.926	0.015
H	0.485	0.111
I	0.472	0.114
IN	0.639	-0.074
IS	0.878	-0.024
LA	0.111	0.250
NC	0.366	0.143
NF	0.916	0.017
NI	0.551	0.095
OK	1.000	0.000
OO	0.825	-0.035
OP	0.964	-0.007
P	0.813	-0.038
PA	0.357	-0.146
PRIO	0.964	-0.007
PRQ	0.690	0.063
PRS	0.905	0.019
Q	0.877	0.025
RES	0.392	-0.135
S	0.009	0.396
SELF	0.840	-0.032
SIDE	0.719	-0.057
SUM	0.753	0.050
T	0.168	0.217
TD	0.393	0.135
TM	0.742	0.052
VIS	0.147	0.228
W	0.661	-0.070

Table 7.7: Task cohesion

Code	p_value	Correlation coefficient
A	0.861	0.028
AL	0.731	-0.055
C		
CB	0.419	-0.128
CL	0.786	-0.043
CP	0.583	-0.087
CR	0.437	0.123
CS	0.673	-0.067
D	0.002	0.455
DA	0.695	0.062
DE	0.615	0.080
DP	0.297	-0.165
DS	0.654	0.071
E	0.144	-0.229
ET	0.335	0.153
F	0.121	-0.243
FB	0.260	-0.178
G	0.770	-0.046
H	0.170	0.216
I	0.699	-0.061
IN	0.001	0.507
IS	0.382	0.138
LA	0.204	0.200
NC	0.407	0.131
NF	0.719	-0.057
NI	0.725	0.056
OK	0.953	-0.009
OO	0.524	0.101
OP	0.773	0.046
P	0.169	-0.216
PA	0.286	-0.168
PRIO	0.532	0.099
PRQ	0.186	-0.208
PRS	0.768	-0.047
Q	0.459	0.118
RES	0.809	-0.038
S	0.009	0.400
SELF	0.962	-0.008
SIDE	0.532	0.099
SUM	0.417	-0.129
T	0.053	0.300
TD	0.923	0.015
TM	0.725	0.056
VIS	0.265	0.176
W	0.000	0.526

8 Conclusions and Discussion

This chapter summarizes and discusses the results shown in **Chapter** 6 and **Chapter** 7. First we conclude if our original hypotheses can be confirmed or disproven, or if further work is necessary. Then we discuss the limitations of the conducted experiments, and possible directions to continue this research. We conclude this chapter with the main conclusions that can be drawn from this thesis.

8.1. Discussion

In this section we discuss our conclusions for the different hypotheses. First we discuss if group-level movement mimicry can be used to detect cohesion, next we discuss if group-level posture mimicry can be used to detect cohesion, and the last hypothesis we discuss is if multimodal cohesion detection is better than unimodal.

Group-level movement mimicry as an indicator for cohesion

One of our original hypotheses is that group-level movement mimicry can be an indicator of both social and task cohesion. These features had an average AUC of 0.64 for social cohesion for both the accelerometer and video features, and 0.63 and 0.58 for task cohesion. Our results on the shared part of the dataset for both the temporal difference features and the accelerometer features performed better than the state-of-the-art paralinguistic mimicry features for both social and task cohesion, which shows movement-based mimicry features are an indicator of cohesion.

Analysis We have seen that the performance of the temporal difference features on the shared part of the dataset is not as good for both social and task cohesion, and the same can be seen for accelerometer data and task cohesion. Correlation analysis on the shared part of the dataset has shown that the minimum group-level metric is consistently one of the most highly correlated. We believe that this might be because the listeners are attentively listening and therefore showing very different behaviours from the speaker if cohesion is high. In one of the meetings that gets excluded in the shared set the participant that is most different from the others might not be the speaker. One participant is not partaking in the discussion and they would probably be the least similar with the rest of the group, which can cause an outlying value in the minimum group-level metric. This could be confirmed by evaluating if excluding this participant would improve the results from the first experiment. Meetings that are part of the test data for the folds with a low AUC can be visually inspected to gain further insight into what causes the poor classification for these meetings.

Group-level posture mimicry as an indicator for cohesion

Another hypothesis was that group-level posture mimicry could also be an indicator for cohesion. All experiments seem to indicate that this is not the case for *task cohesion.* With an average AUC of 0.42 these features performed worse than random guessing. The experiments on the shared set for social cohesion show there might be some indication of social cohesion in appearance-based HOG features with an average AUC of 0.56. This pattern is confirmed when testing appearance-based features with a different form of measuring pairwise similarity in the final experiment, and with more appearance descriptors as the average AUC increases to

0.61. We conclude that it cannot be currently determined if appearance-based features are a good indicator of social cohesion, but we can firmly reject the hypothesis that they are a good indicator of task cohesion.

Analysis We originally chose to use a different way of measuring pair-wise similarity for the appearance-based features as we felt it most closely matched the definition of the concepts used by other researchers. It was seen that choice of pair-wise similarity metric was very impactful for the performance of these features. To examine if mimicry appearance-based features truly capture social cohesion it is important to establish how mimicry should be measured, and then verify if mimicry in these features is related to social cohesion.

To gain insight into what the Gaussian Mixture Model actually models, the individual models from participants can be investigated. This can be done by finding the sample that best matches a given mixture and inspecting this sample to see what region of the video it came from to get an impression of what each components represents.

Unimodal vs multimodal detection of cohesion

Our last hypothesis was that cohesion can be better detected by using a multimodal approach, as one modality might contain information another does not. Our experiments showed that the late fusion of paralinguistic mimicry from audio, temporal difference from video and accelerometer features showed better performance than that of individual modalities for both social and task cohesion with average AUCs of 0.68 and 0.65 respectively. This confirms that a multimodal system with motion and paralinguistic mimicry outperforms a unimodal system.

Analysis It should be noted that the audio features performed much better when using the full dataset for evaluation, and this experiment could only be conducted on the shared part of the dataset. Future experiments should test if this relationship still holds if more data is available, as the performance of the paralinguistic mimicry features seem to increase more than that of the other modalities. Another way to gather more audio data would be to include the non-shared part of the dataset as training samples for the audio model. This is possible because a different model is trained for each modality.

8.2. Limitations

Although the creation of the labels was based on connections between the coding categories and either form of correlation as found in literature, the relationship between the act4teams-based cohesion labels and more established ways of measuring social and task cohesion has not been statistically established. We conducted experiments that aimed to establish the link between more common methods of measuring cohesion and the labels used in this experiment, but failed to prove this correlation statistically. It is therefore unsure if the results from this experiment truly measure social and task cohesion in the traditional sense. We also saw that the performance of the paralinguistic mimicry features dropped when shifting the start of the meeting segments. This suggests that the current labels are unstable because the labelling and label distribution is impacted by shifting the segments.

All the meetings used in this experiment were recorded in the IBM SenseLab. This can influence the type of participant that would be present in these meetings. It is possible that these results only apply within this company.

The different modalities had to be manually synchronised and might therefore not be perfectly lined up with each other. As the accelerometer features are a statistical description of a 3-second window this should not have a large impact, and the same should apply for the appearance features as these are modelled without the time component, but the impact of possibly imperfect alignment has not been established.

8.3. Future work

Currently the bounding boxes used for the extraction of the video features are annotated by hand. For a fully automated process different automatic bounding box detection methods would need to be tested, or a new one would have to be developed.

The current unimodal vs multimodal experiment uses late fusion to combine the different modalities, as it is easier to analyse the impact of the individual modality on the combined performance. For those interested in a high classification performance early fusion could provide better results, but this would need to be tested.

An experiment was set up to measure the correlation between externally annotated cohesion on 2-minute meeting segments and the labels used in our experiments. Inter-rater reliability on these questions was low, but some logical correlations between the act4teams categories and cohesion annotations were found. If the questions used in this experiment are refined, and more annotators are asked to annotate it might be possible to establish if a relationship between these verbal expression codings and external observation of cohesion exists. It is however possible that the frequency of coding categories is not as important for cohesion, and the order of categories is also an important indication. It is therefore recommended to verify if refining the questions and using more annotators shows that the frequency of these categories is truly linked to cohesion, and otherwise test if the order of specific categories can be related to the presence of cohesion.

We have shown that the pair-wise comparison metric greatly impacts the classification performance. As already suggested in the discussion of the posture mimicry experiments it is vital to investigate if mimicry between participants is truly measured, and establish what mimicry exactly is. This can be done by either annotating for behaviours in which mimicry can occur like head gestures, hand gestures, or posture, or by annotating for mimicry itself. If data is annotated for behaviours this can then be converted to a measure of mimicry by formalizing what mimicry is and applying this formalization to the annotated behaviours. This approach would improve inter-rater reliability as there would be no mismatch in the definition of mimicry, and annotating a behaviour should be less complex.

The experiments conducted in this thesis have investigated general movement mimicry, and posture mimicry. Further experiments could look at the automatic detection of more specific behaviours like head gestures and hand gestures, and see if conversational partners mimic these when cohesion is present.

8.4. Main conclusions

Movement mimicry features extracted from accelerometer and video data can detect social cohesion and task cohesion. Audio based mimicry features can as well when adding more training samples. The right way of measuring pair-wise mimicry still needs to be established to definitively conclude if appearance mimicry as measured in HOG features is or is not able to detect social cohesion. The combination of general movement mimicry from accelerometer data and paralinguistic mimicry improve prediction of both individual modalities for verbal expressions of social cohesion. The performance of the movement mimicry features however does depend on people exhibiting 'normal' behaviour.

Bibliography

[1] Uwe Altmann, Catharine Oertel, and Nick Campbell. Conversational involvement and synchronous nonverbal behaviour. In *Cognitive Behavioural Systems*, pages 343–352. Springer, 2012.

[2] Oya Aran and Daniel Gatica-Perez. One of a kind. In *Proceedings of the 15th ACM on International conference on multimodal interaction - ICMI '13*. ACM Press, 2013. doi: 10.1145/2522848.2522859.

[3] Kathleen T. Ashenfelter, Steven M. Boker, Jennifer R. Waddell, and Nikolay Vitanov. Spatiotemporal symmetry and multifractal structure of head movements during dyadic conversation. *Journal of Experimental Psychology: Human Perception and Performance*, 35(4):1072–1091, 2009. doi: 10.1037/a0015017.

[4] Sileye O Ba and Jean-Marc Odobez. A rao-blackwellized mixed state particle filter for head pose tracking. Technical report, IDIAP, 2005.

[5] J. N. Bailenson and N. Yee. Digital chameleons: Automatic assimilation of nonverbal gestures in immersive virtual environments. *Psychological Science*, 16(10):814–819, oct 2005. doi: 10.1111/j.1467-9280.2005.01619.x.

[6] T. Baltrusaitis, P. Robinson, and L. Morency. 3d constrained local model for rigid and non-rigid facial tracking. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, jun 2012. doi: 10.1109/cvpr.2012.6247980.

[7] Tadas Baltrusaitis, Peter Robinson, and Louis-Philippe Morency. Constrained local neural fields for robust facial landmark detection in the wild. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 354–361, 2013.

[8] Tadas Baltrusaitis, Peter Robinson, and Louis-Philippe Morency. OpenFace: An open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, mar 2016. doi: 10.1109/wacv.2016.7477553.

[9] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. OpenFace 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, may 2018. doi: 10.1109/fg.2018.00019.

[10] Frank J Bernieri, John S Gillis, Janet M Davis, and Jon E Grahe. Dyad rapport and the accuracy of its judgment across situations: A lens model analysis. *Journal of Personality and Social Psychology*, 71(1): 110, 1996.

[11] Cigdem Beyan, Francesca Capozzi, Cristina Becchio, and Vittorio Murino. Identification of emergent leaders in a meeting scenario using multiple kernel learning. In *Proceedings of the 2nd Workshop on Advancements in Social Signal Processing for Multimodal Interaction - ASSP4MI '16*. ACM Press, 2016. doi: 10.1145/3005467.3005469.

[12] Cigdem Beyan, Francesca Capozzi, Cristina Becchio, and Vittorio Murino. Prediction of the leadership style of an emergent leader using audio and visual nonverbal features. *IEEE Transactions on Multimedia*, 20(2):441–456, 2018. doi: 10.1109/TMM.2017.2740062.

[13] Sanjay Bilakhia, Stavros Petridis, and Maja Pantic. Audiovisual detection of behavioural mimicry. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*. IEEE, sep 2013. doi: 10.1109/acii.2013.27.

[14] Sanjay Bilakhia, Stavros Petridis, Anton Nijholt, and Maja Pantic. The MAHNOB mimicry database: A database of naturalistic human interactions. *Pattern Recognition Letters*, 66:52–61, nov 2015. doi: 10.1016/j.patrec.2015.03.005.

[15] Steven M Boker, Jennifer L Rotondo, Minquan Xu, and Kadijah King. Windowed cross-correlation and peak picking for the analysis of variability in the association between behavioral time series. *Psychological methods*, 7(3):338, 2002.

[16] Leif J Braaten. The different patterns of group climate critical incidents in high and low cohesion sessions of group psychotherapy. *International Journal of Group Psychotherapy*, 40(4):477–493, 1990.

[17] Leif J. Braaten. Group cohesion: A new multidimensional model. *Group*, 15(1):39–55, mar 1991. doi: 10.1007/bf01419845.

[18] Laura Cabrera-Quiros, Andrew Demetriou, Ekin Gedik, Leander van der Meij, and Hayley Hung. The matchnmingle dataset: a novel multi-sensor resource for the analysis of social interactions and group dynamics in-the-wild during free-standing conversations and speed dates. *IEEE Transactions on Affective Computing*, 2018.

[19] Cabrera Quiros, L. C. *Automatic analysis of human social behavior in the wild using multimodal streams*. PhD thesis, TU Delft, 2018.

[20] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *arXiv preprint arXiv:1812.08008*, 2018.

[21] Sally A Carless and Caroline De Paola. The measurement of cohesion in work teams. *Small group research*, 31(1):71–88, 2000.

[22] Albert V. Carron and Lawrence R. Brawley. Cohesion. *Small Group Research*, 31(1):89–106, feb 2000. doi: 10.1177/104649640003100105.

[23] A.V. Carron, W.N. Widmeyer, and L.R. Brawley. The development of an instrument to assess cohesion in sport teams: The group environment questionnaire. *Journal of Sport Psychology*, 7(3):244–266, sep 1985. doi: 10.1123/jsp.7.3.244.

[24] Gavin C Cawley and Nicola LC Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11(Jul):2079–2107, 2010.

[25] Aleksandra Cerekovic, Oya Aran, and Daniel Gatica-Perez. Rapport with virtual agents: What do human social cues and personality explain? *IEEE Transactions on Affective Computing*, 8(3):382–395, jul 2017. doi: 10.1109/taffc.2016.2545650.

[26] Song Chang, Liangding Jia, Riki Takeuchi, and Yahua Cai. Do high-commitment work systems affect creativity? a multilevel combinational approach to employee creativity. *Journal of Applied Psychology*, 99(4):665–680, jul 2014. doi: 10.1037/a0035679.

[27] Tanya L Chartrand and John A Bargh. The chameleon effect: the perception–behavior link and social interaction. *Journal of personality and social psychology*, 76(6):893, 1999.

[28] Tanya L Chartrand, William W Maddux, and Jessica L Lakin. Beyond the perception-behavior link: The ubiquitous utility and motivational moderators of nonconscious mimicry. *The new unconscious*, pages 334–361, 2005.

[29] Prerna Chikersal, Maria Tomprou, Young Ji Kim, Anita Williams Woolley, and Laura Dabbish. Deep structures of collaboration. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing - CSCW '17*. ACM Press, 2017. doi: 10.1145/2998181.2998250.

[30] P. Chippendale. Towards automatic body language annotation. In *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*. IEEE, 2006. doi: 10.1109/fgr.2006.105.

[31] Herbert H Clark. *Using language*. Cambridge university press, 1996.

[32] M Cox, J Nuevo-Chiquero, Jason M Saragih, and Simon Lucey. Csiro face analysis sdk. *Brisbane, Australia*, 3, 2013.

[33] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. IEEE. doi: 10.1109/cvpr.2005.177.

[34] Emilie Delaherche, Mohamed Chetouani, Ammar Mahdhaoui, Catherine Saint-Georges, Sylvie Viaux, and David Cohen. Interpersonal synchrony: A survey of evaluation methods across disciplines. *IEEE Transactions on Affective Computing*, 3(3):349–365, jul 2012. doi: 10.1109/t-affc.2012.12.

[35] Hamdi Dibeklioğlu, Zakia Hammal, Ying Yang, and Jeffrey F Cohn. Multimodal detection of depression in clinical interviews. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 307–310. ACM, 2015.

[36] Hamdi Dibeklioglu, Zakia Hammal, and Jeffrey F. Cohn. Dynamic multimodal measurement of depression severity using deep autoencoding. *IEEE Journal of Biomedical and Health Informatics*, 22(2): 525–536, mar 2018. doi: 10.1109/jbhi.2017.2676878.

[37] Paul Ekman and Wallace V Friesen. Measuring facial movement. *Environmental psychology and nonverbal behavior*, 1(1):56–75, 1976.

[38] Gabriele Fanelli, Matthias Dantone, Juergen Gall, Andrea Fossati, and Luc Van Gool. Random forests for real time 3d face analysis. *International Journal of Computer Vision*, 101(3):437–458, aug 2012. doi: 10.1007/s11263-012-0549-0.

[39] Leon Festinger. Informal social communication. *Psychological review*, 57(5):271, 1950.

[40] Daniel Gatica-Perez, L. McCowan, Dong Zhang, and Samy Bengio. Detecting group interest-level in meetings. In *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, volume 1, pages I–489. IEEE, 2005.

[41] James Griffith. Further considerations concerning the cohesion-performance relation in military settings. *Armed Forces & Society*, 34(1):138–147, sep 2007. doi: 10.1177/0095327x06294620.

[42] L. Gupta and Suwei Ma. Gesture-based interaction and communication: automated classification of hand gesture contours. *IEEE Transactions on Systems, Man and Cybernetics, Part C (Applications and Reviews)*, 31(1):114–120, 2001. doi: 10.1109/5326.923274.

[43] Zakia Hammal, Jeffrey F. Cohn, and David Ted George. Interpersonal coordination of HeadMotion in distressed couples. *IEEE Transactions on Affective Computing*, 5(2):155–167, apr 2014. doi: 10.1109/taffc.2014.2326408.

[44] Michael A Hogg. The social psychology of group cohesiveness: From attraction to social identity. *New York*, 1992.

[45] Judith Holler and Katie Wilkin. Co-speech gesture mimicry in the process of collaborative referring during face-to-face dialogue. *Journal of Nonverbal Behavior*, 35(2):133–153, jan 2011. doi: 10.1007/s10919-011-0105-6.

[46] Hayley Hung and Daniel Gatica-Perez. Estimating cohesion in small groups using audio-visual nonverbal behavior. *IEEE Transactions on Multimedia*, 12(6):563–575, oct 2010. doi: 10.1109/tmm.2010.2055233.

[47] Hayley Hung, Dinesh Babu Jayagopi, Sileye Ba, Jean-Marc Odobez, and Daniel Gatica-Perez. Investigating automatic dominance estimation in groups from visual attention and speaking activity. In *Proceedings of the 10th international conference on Multimodal interfaces - IMCI '08*. ACM Press, 2008. doi: 10.1145/1452392.1452441.

[48] G. Iyengar, H.J. Nock, and C. Neti. Audio-visual synchrony for detection of monologues in video archives. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03)*. IEEE, 2003. doi: 10.1109/icassp.2003.1200085.

[49] Dinesh Babu Jayagopi, Hayley Hung, Chuohao Yeo, and Daniel Gatica-Perez. Modeling dominance in group conversations using nonverbal activity cues. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(3):501–513, mar 2009. doi: 10.1109/tasl.2008.2008238.

[50] Jesper Hojvang Jensen, Daniel PW Ellis, Mads G Christensen, and Soren Holdt Jensen. Evaluation distance measures between gaussian mixture models of mfccs. In *Proceedings of the 8th International Conference on Music Information Retrieval, ISMIR 2007, Vienna, Austria, September 23-27, 2007*, 2007.

[51] Öykü Kapcak, José Vargas-Quiros, and Hayley Hung. Estimating romantic, social, and sexual attraction by quantifying bodily coordination using wearable sensors. In *2019 Eight International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2019.

[52] S Kauffeld, G Lorenzo, K Montasem, NK Lehmann-Willenbrock, S Grote, and E Frieling. Die nächste generation der teamentwicklung: Neue wege mit act4teams®[the next generation of team development: New paths with act4teams®]. *Handbuch kompetenzentzwicklung*, pages 191–215, 2009.

[53] Shiro Kumano, Kazuhiro Otsuka, Dan Mikami, and Junji Yamato. Recognizing communicative facial expressions for discovering interpersonal emotions in group meetings. In *Proceedings of the 2009 international conference on Multimodal interfaces - ICMI-MLMI '09*. ACM Press, 2009. doi: 10.1145/1647314.1647333.

[54] Shiro KUMANO, Kazuhiro OTSUKA, Masafumi MATSUDA, and Junji YAMATO. Analyzing perceived empathy based on reaction time in behavioral mimicry. *IEICE Transactions on Information and Systems*, E97.D(8):2008–2020, 2014. doi: 10.1587/transinf.e97.d.2008.

[55] Amit Kumar, Azadeh Alavi, and Rama Chellappa. KEPLER: Simultaneous estimation of keypoints and 3d pose of unconstrained faces in a unified framework by learning efficient h-CNN regressors. *Image and Vision Computing*, 79:49–62, nov 2018. doi: 10.1016/j.imavis.2018.09.009.

[56] Ana Kuzmanic and Vlasta Zanchi. Hand shape classification using DTW and LCSS as similarity measures for vision-based gesture recognition system. In *EUROCON 2007 - The International Conference on "Computer as a Tool"*. IEEE, 2007. doi: 10.1109/eurcon.2007.4400350.

[57] Marco La Cascia, Stan Sclaroff, and Vassilis Athitsos. Fast, reliable head tracking under varying illumination: An approach based on registration of texture-mapped 3d models. *IEEE Transactions on pattern analysis and machine intelligence*, 22(4):322–336, 2000.

[58] Gongfa Li, Heng Tang, Ying Sun, Jianyi Kong, Guozhang Jiang, Du Jiang, Bo Tao, Shuang Xu, and Honghai Liu. Hand gesture recognition based on convolution neural network. *Cluster Computing*, dec 2017. doi: 10.1007/s10586-017-1435-x.

[59] Max M. Louwerse, Rick Dale, Ellen G. Bard, and Patrick Jeuniaux. Behavior matching in multimodal communication is synchronized. *Cognitive Science*, 36(8):1404–1426, sep 2012. doi: 10.1111/j.1551-6709.2012.01269.x.

[60] Marwa Mahmoud and Peter Robinson. Interpreting hand-over-face gestures. In *Affective Computing and Intelligent Interaction*, pages 248–255. Springer Berlin Heidelberg, 2011. doi: 10.1007/978-3-642-24571-8_27.

[61] Marwa M Mahmoud, Tadas Baltrušaitis, and Peter Robinson. Automatic detection of naturalistic hand-over-face gesture descriptors. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 319–326. ACM, 2014. doi: 10.1145/2663204.2663258.

[62] I. McCowan, S. Bengio, D. Gatica-Perez, G. Lathoud, F. Monay, D. Moore, P. Wellner, and H. Bourlard. Modeling human interaction in meetings. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03)*. IEEE, 2003. doi: 10.1109/icassp.2003.1202751.

[63] Michael D Michalisin, Steven J Karau, and Charnchai Tangpong. Top management team cohesion and superior industry returns: An empirical study of the resource-based view. *Group & Organization Management*, 29(1):125–140, 2004.

[64] Philipp Müller, Michael Xuelin Huang, and Andreas Bulling. Detecting low rapport during natural interactions in small groups from non-verbal behaviour. In *Proceedings of the 2018 Conference on Human Information Interaction&Retrieval - IUI '18*. ACM Press, 2018. doi: 10.1145/3172944.3172969.

[65] Louis-Philippe Morency, Jacob Whitehill, and Javier Movellan. Generalized adaptive view-based appearance model: Integrated framework for monocular head pose estimation. In *2008 8th IEEE International Conference on Automatic Face & Gesture Recognition*. IEEE, sep 2008. doi: 10.1109/afgr.2008.4813429.

[66] Joseph E. Mroz, Joseph A. Allen, Dana C. Verhoeven, and Marissa L. Shuffler. Do we really need another meeting? the science of workplace meetings. *Current Directions in Psychological Science*, 27(6):484–491, oct 2018. doi: 10.1177/0963721418776307.

[67] Marjolein C Nanninga. Prediction of team cohesion, by examining prosodic mimicry in small group meetings. Master's thesis, TU Delft, 2017. URL `http://resolver.tudelft.nl/uuid:bbed8ce5-36a8-4773-a461-33f19c5f774d`.

[68] Marjolein C Nanninga, Yanxia Zhang, Nale Lehmann-Willenbrock, Zoltán Szlávik, and Hayley Hung. Estimating verbal expressions of task and social cohesion in meetings by quantifying paralinguistic mimicry. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 206–215. ACM, 2017.

[69] Laurent Nguyen, Jean-Marc Odobez, and Daniel Gatica-Perez. 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*. ACM Press, 2012. doi: 10.1145/2388676.2388734.

[70] Elias Pampalk. Speeding up music similarity. *2nd Annual Music Information Retrieval eXchange, London*, 2005.

[71] Allan Pease and Barbara Pease. The definitive book of body language (bantam hardcover ed.), 2006.

[72] Alex Pentland and Anmol Madan. Perception of social interest. In *Proc. IEEE Int. Conf. on Computer Vision, Workshop on Modeling People and Human Interaction (ICCV-PHI)*. Citeseer, 2005. doi: 10.1.1/467.6810.

[73] Pramod Kumar Pisharady and Martin Saerbeck. Recent methods and databases in vision-based hand gesture recognition: A review. *Computer Vision and Image Understanding*, 141:152–165, dec 2015. doi: 10.1016/j.cviu.2015.08.004.

[74] Fabian Ramseyer and Wolfgang Tschacher. Nonverbal synchrony of head- and body-movement in psychotherapy: different signals have different associations with outcome. *Frontiers in Psychology*, 5, sep 2014. doi: 10.3389/fpsyg.2014.00979.

[75] Daniel C. Richardson and Rick Dale. Looking to understand: The coupling between speakers' and listeners' eye movements and its relationship to discourse comprehension. *Cognitive Science*, 29(6):1045–1060, nov 2005. doi: 10.1207/s15516709cog0000_29.

[76] S. G. Rogelberg. *Encyclopedia of Industrial and Organizational Psychology*, chapter Meetings at work., pages 474,475. Thousand Oaks, 2006.

[77] Steven G. Rogelberg, Desmond J. Leach, Peter B. Warr, and Jennifer L. Burnfield. "Not Another Meeting!" Are Meeting Time Demands Related to Employee Well-Being? *Journal of Applied Psychology*, 91(1):83–96, 2006. doi: 10.1037/0021-9010.91.1.83.

[78] Steven G. Rogelberg, Joseph A. Allen, Linda Shanock, Cliff Scott, and Marissa Shuffler. Employee satisfaction with meetings: A contemporary facet of job satisfaction. *Human Resource Management*, 49(2):149–172, mar 2010. doi: 10.1002/hrm.20339.

[79] Nataniel Ruiz, Eunji Chong, and James M. Rehg. Fine-grained head pose estimation without keypoints. 2017.

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

[81] Dairazalia Sanchez-Cortes, Oya Aran, Marianne Schmid Mast, and Daniel Gatica-Perez. A nonverbal behavior approach to identify emergent leaders in small groups. *IEEE Transactions on Multimedia*, 14(3):816–832, jun 2012. doi: 10.1109/tmm.2011.2181941.

[82] A Schell. Meeting-kultur in europäischen unternehmen: Ad-hoc-umfrage unter mitarbeitern und führungskräften, die regelmäßig an business-meetings teilnehmen [european business meeting culture: An ad-hoc survey of employees and managers who regularly participate in business meetings]. *Munich: Schell Marketing Consulting*, 2010.

[83] Stefan Scherer, Friedhelm Schwenker, Nick Campbell, and Günther Palm. Multimodal laughter detection in natural discourses. In *Human Centered Robot Systems*, pages 111–120. Springer, 2009.

[84] Mohit Sharma, Dragan Ahmetovic, Laszlo A. Jeni, and Kris M. Kitani. Recognizing visual signatures of spontaneous head gestures. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, mar 2018. doi: 10.1109/wacv.2018.00050.

[85] Nishu Singla. Motion detection based on frame difference method. *International Journal of Information & Computation Technology*, 4(15):1559–1565, 2014.

[86] Cees G. M. Snoek, Marcel Worring, and Arnold W. M. Smeulders. Early versus late fusion in semantic video analysis. In *Proceedings of the 13th annual ACM international conference on Multimedia - MULTIMEDIA '05*. ACM Press, 2005. doi: 10.1145/1101149.1101236.

[87] Vijay Solanki, Alessandro Vinciarelli, Jane Stuart-Smith, and Rachel Smith. When the game gets difficult, then it is time for mimicry. In *Recent advances in nonlinear speech processing*, pages 247–254. Springer, 2016.

[88] Mariëlle Stel and Roos Vonk. Mimicry in social interaction: Benefits for mimickers, mimickees, and their interaction. *British Journal of Psychology*, 101(2):311–323, may 2010. doi: 10.1348/000712609x465424.

[89] Viktoria Stray. Planned and unplanned meetings in large-scale projects. In *Proceedings of the 19th International Conference on Agile Software Development Companion - XP '18*. ACM Press, 2018. doi: 10.1145/3234152.3234178.

[90] Xiaofan Sun, Anton Nijholt, Khiet P. Truong, and Maja Pantic. Automatic visual mimicry expression analysis in interpersonal interaction. In *CVPR 2011 WORKSHOPS*. IEEE, jun 2011. doi: 10.1109/cvprw.2011.5981812.

[91] Karen Tracy and Aaron Dimock. Meetings: Discursive sites for building and fragmenting community. *Annals of the International Communication Association*, 28(1):127–165, 2004.

[92] G. V. Trunk. A problem of dimensionality: A simple example. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(3):306–307, jul 1979. doi: 10.1109/tpami.1979.4766926.

[93] Adam J. Vanhove and Mitchel N. Herian. Team cohesion and individual well-being: A conceptual analysis and relational framework. In *Team Cohesion: Advances in Psychological Theory, Methods and Practice*, pages 53–82. Emerald Group Publishing Limited, nov 2015. doi: 10.1108/s1534-085620150000017004.

[94] Arno Veenstra and Hayley Hung. Do they like me? using video cues to predict desires during speed-dates. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 838–845. IEEE, 2011.

[95] Paul Viola and Michael J Jones. Robust real-time face detection. *International journal of computer vision*, 57(2):137–154, 2004.

[96] Heng Wang, Alexander Klaser, Cordelia Schmid, and Cheng-Lin Liu. Action recognition by dense trajectories. In *CVPR 2011*. IEEE, jun 2011. doi: 10.1109/cvpr.2011.5995407.

[97] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu. Dense trajectories and motion boundary descriptors for action recognition. *International Journal of Computer Vision*, 103(1):60–79, mar 2013. doi: 10.1007/s11263-012-0594-8.

[98] Heng Wang, Dan Oneata, Jakob Verbeek, and Cordelia Schmid. A robust and efficient video representation for action recognition. *International Journal of Computer Vision*, 119(3):219–238, jul 2015. doi: 10.1007/s11263-015-0846-5.

[99] C. L. Webber and J. P. Zbilut. Dynamical assessment of physiological systems and states using recurrence plot strategies. *Journal of Applied Physiology*, 76(2):965–973, feb 1994. doi: 10.1152/jappl.1994.76.2.965.

[100] Ernst Wit, Edwin van den Heuvel, and Jan-Willem Romeijn. 'all models are wrong...': an introduction to model uncertainty. *Statistica Neerlandica*, 66(3):217–236, jul 2012. doi: 10.1111/j.1467-9574.2012.00530.x.

[101] Bo Xiao, Panayiotis Georgiou, Brian Baucom, and Shrikanth S. Narayanan. Head motion modeling for human behavior analysis in dyadic interaction. *IEEE Transactions on Multimedia*, 17(7):1107–1119, jul 2015. doi: 10.1109/tmm.2015.2432671.

[102] Renqiang Xie, Xia Sun, Xiang Xia, and Juncheng Cao. Similarity matching-based extensible hand gesture recognition. *IEEE Sensors Journal*, 15(6):3475–3483, jun 2015. doi: 10.1109/jsen.2015.2392091.

[103] Yingen Xiong, F. Quek, and D. McNeill. Hand gesture symmetric behavior detection and analysis in natural conversation. In *Proceedings. Fourth IEEE International Conference on Multimodal Interfaces*. IEEE Comput. Soc, 2002. doi: 10.1109/icmi.2002.1166989.

[104] W. Yang, K. Akiyama, K. Kitani, L. Jeni, and Y. Mukaigawa. Head gesture recognition in spontaneous human conversations: A benchmark. In *Workshop on Egocentric (First-Person) Vision (CVPR), 201*, 2016.

[105] Chuohao Yeo and Kannan Ramchandran. Compressed domain video processing of meetings for activity estimation in dominance classification and slide transition detection. *University of California, Berkeley, Tech. Rep. UCB/EECS-2008-79*, 2008.

[106] Öykü Kapcak. Estimating attraction by quantifying bodily coordination using wearable sensors. Master's thesis, TU Delft, 2019. URL `http://resolver.tudelft.nl/uuid:9223bb0e-9497-4969-8319-66699118abf3`.

[107] Ho-Sub Yoon, Jung Soh, Younglae J Bae, and Hyun Seung Yang. Hand gesture recognition using combined features of location, angle and velocity. *Pattern recognition*, 34(7):1491–1501, 2001.

[108] Xiang Yu, Shaoting Zhang, Zhennan Yan, Fei Yang, Junzhou Huang, Norah E. Dunbar, Matthew L. Jensen, Judee K. Burgoon, and Dimitris N. Metaxas. Is interactional dissynchrony a clue to deception? insights from automated analysis of nonverbal visual cues. *IEEE Transactions on Cybernetics*, 45 (3):492–506, mar 2015. doi: 10.1109/tcyb.2014.2329673.

[109] Stephen J Zaccaro and M Catherine McCoy. The effects of task and interpersonal cohesiveness on performance of a disjunctive group task 1. *Journal of applied social psychology*, 18(10):837–851, 1988.

[110] Yanxia Zhang, Jeffrey Olenick, Chu-Hsiang Chang, Steve WJ Kozlowski, and Hayley Hung. Teamsense: Assessing personal affect and group cohesion in small teams through dyadic interaction and behavior analysis with wearable sensors. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(3):150, 2018.

A

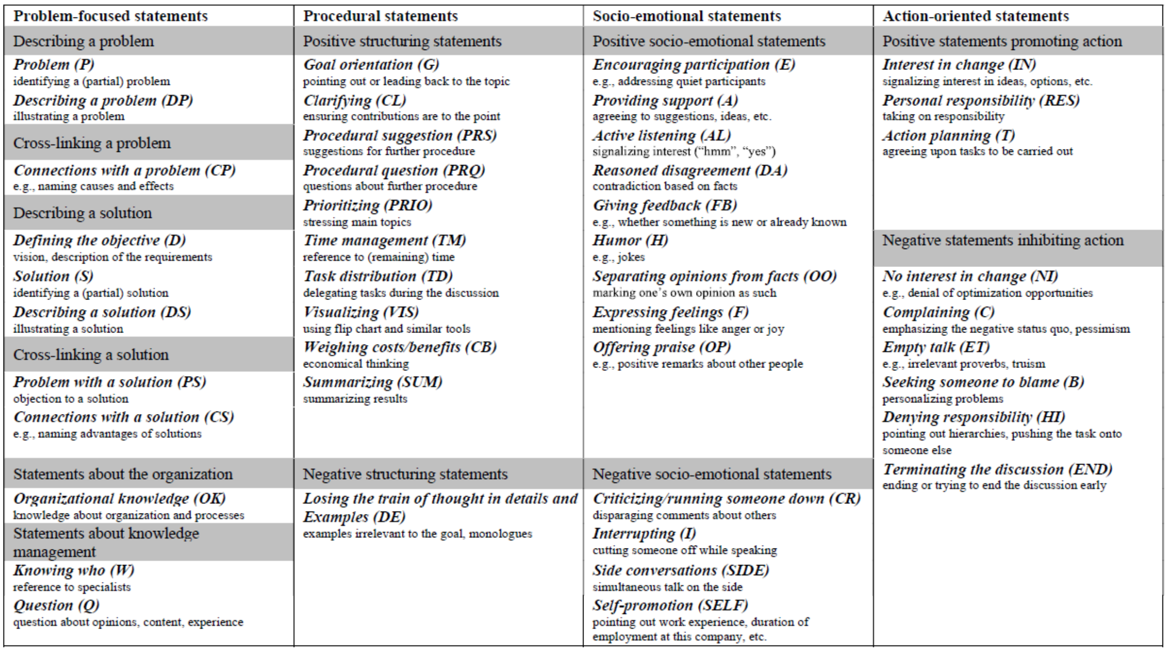
Available data

Table A.1: Overview of available data for each meeting. The greyed out rows are the meetings for which were not coded. NA means this data was not available for this modality, and U means accelerometer data can be used for this meeting.

Meeting nr	Duration	Nr. participants	Average age	Annotation	Video	Accel.	Note
1	01:14:00	5	36.4 (11.1)				
2	01:27:45	4	46.3 (9.3)				
3	02:46:00	5	36.4 (11.1)	NA	-	-	
4	01:04:05	8	35.5 (14.8)			U	
5	00:52:45	6	37.7 (16.0)			U	
6	01:29:41	4	23.7 (1.5)			U	
7	00:41:23	3	24.7 (2.9)			U	
8	00:22:29	3	37.3 (15.3)				Accel data 2 people
9	00:24:30	5	36.3 (21.3)			U	
10	00:55:40	5	46.3 (1.2)				
11	00:47:13	6	31.3 (4.8)			U	
12	00:43:00	5	43.7 (18.3)			U	
13	00:58:38	4	24.0 (1.4)			U	
14	00:36:45	4	29.0 (6.2)				
15	00:54:25	5	28.4 (6.5)			U	
16	00:28:38	5	48.0 (3.4)			U	
17	?	3	38.7 (16.3)	NA	-	-	
18	00:41:29	6	43.5 (5.2)			U	
19	01:13:20	3	43.5 (2.1)			U	Impossible angle video
20	00:42:55	6	26.0 (1.6)			U	
21	00:49:16	6	40.0 (5.6)			U	
22	00:49:13	5	51.3 (5.1)		NA	U	
23	01:01:10	4	57.0 (2.2)		NA	U	
24	01:11:24	3	44.0 (1.4)		NA	U	
25	01:02:28	3	22.3 (0.6)			U	

B

Act4teams observation categories

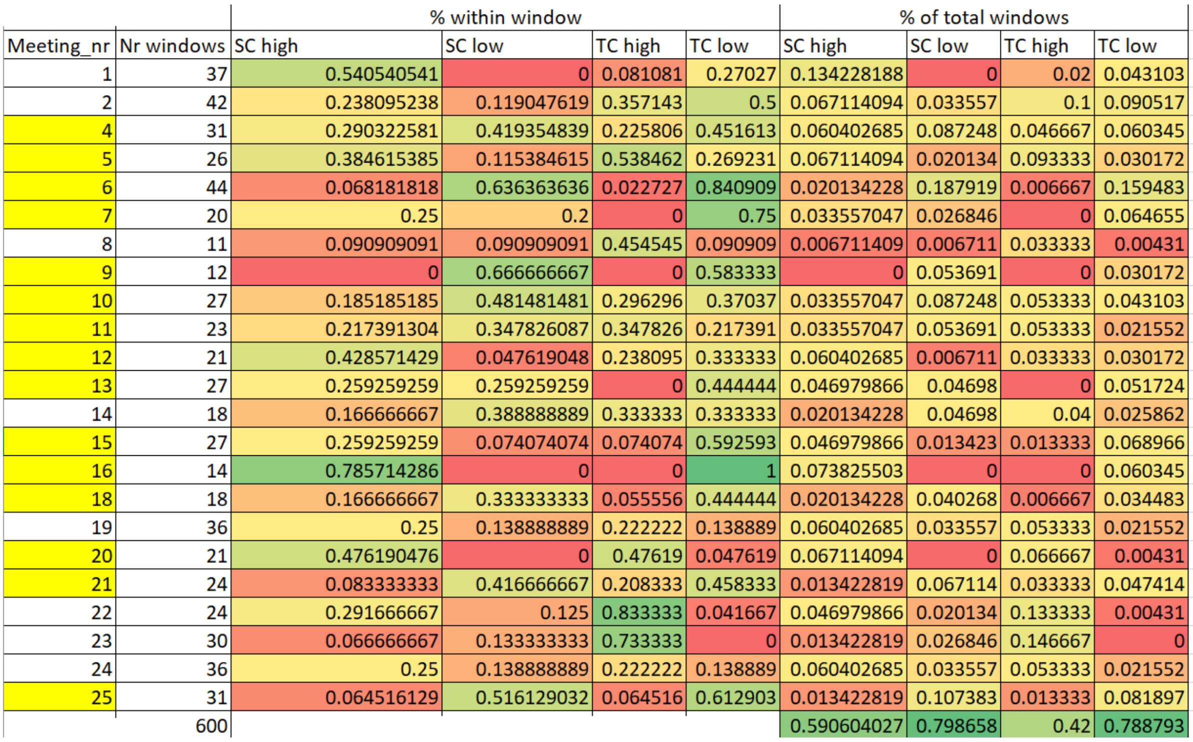


Problem-focused statements	Procedural statements	Socio-emotional statements	Action-oriented statements
Describing a problem	Positive structuring statements	Positive socio-emotional statements	Positive statements promoting action
Problem (P) identifying a (partial) problem	***Goal orientation (G)*** pointing out or leading back to the topic	***Encouraging participation (E)*** e.g., addressing quiet participants	***Interest in change (IN)*** signalizing interest in ideas, options, etc.
Describing a problem (DP) illustrating a problem	***Clarifying (CL)*** ensuring contributions are to the point	***Providing support (A)*** agreeing to suggestions, ideas, etc.	***Personal responsibility (RES)*** taking on responsibility
Cross-linking a problem	***Procedural suggestion (PRS)*** suggestions for further procedure	***Active listening (AL)*** signalizing interest ("hmm", "yes")	***Action planning (T)*** agreeing upon tasks to be carried out
Connections with a problem (CP) e.g., naming causes and effects	***Procedural question (PRQ)*** questions about further procedure	***Reasoned disagreement (DA)*** contradiction based on facts	
Describing a solution	***Prioritizing (PRIO)*** stressing main topics	***Giving feedback (FB)*** e.g., whether something is new or already known	
Defining the objective (D) vision, description of the requirements	***Time management (TM)*** reference to (remaining) time	***Humor (H)*** e.g., jokes	Negative statements inhibiting action
Solution (S) identifying a (partial) solution	***Task distribution (TD)*** delegating tasks during the discussion	***Separating opinions from facts (OO)*** marking one's own opinion as such	***No interest in change (NI)*** e.g., denial of optimization opportunities
Describing a solution (DS) illustrating a solution	***Visualizing (VIS)*** using flip chart and similar tools	***Expressing feelings (F)*** mentioning feelings like anger or joy	***Complaining (C)*** emphasizing the negative status quo, pessimism
Cross-linking a solution	***Weighing costs/benefits (CB)*** economical thinking	***Offering praise (OP)*** e.g., positive remarks about other people	***Empty talk (ET)*** e.g., irrelevant proverbs, truism
Problem with a solution (PS) objection to a solution	***Summarizing (SUM)*** summarizing results		***Seeking someone to blame (B)*** personalizing problems
Connections with a solution (CS) e.g., naming advantages of solutions			***Denying responsibility (HI)*** pointing out hierarchies, pushing the task onto someone else
Statements about the organization	Negative structuring statements	Negative socio-emotional statements	***Terminating the discussion (END)*** ending or trying to end the discussion early
Organizational knowledge (OK) knowledge about organization and processes	***Losing the train of thought in details and Examples (DE)*** examples irrelevant to the goal, monologues	***Criticizing/running someone down (CR)*** disparaging comments about others	
Statements about knowledge management		***Interrupting (I)*** cutting someone off while speaking	
Knowing who (W) reference to specialists		***Side conversations (SIDE)*** simultaneous talk on the side	
Question (Q) question about opinions, content, experience		***Self-promotion (SELF)*** pointing out work experience, duration of employment at this company, etc.	

Figure B.1: Overview of the different act4teams coding categories.

C

Window distribution



		% within window				% of total windows			
Meeting_nr	Nr windows	SC high	SC low	TC high	TC low	SC high	SC low	TC high	TC low
1	37	0.540540541	0	0.081081	0.27027	0.134228188	0	0.02	0.043103
2	42	0.238095238	0.119047619	0.357143	0.5	0.067114094	0.033557	0.1	0.090517
4	31	0.290322581	0.419354839	0.225806	0.451613	0.060402685	0.087248	0.046667	0.060345
5	26	0.384615385	0.115384615	0.538462	0.269231	0.067114094	0.020134	0.093333	0.030172
6	44	0.068181818	0.636363636	0.022727	0.840909	0.020134228	0.187919	0.006667	0.159483
7	20	0.25	0.2	0	0.75	0.033557047	0.026846	0	0.064655
8	11	0.090909091	0.090909091	0.454545	0.090909	0.006711409	0.006711	0.033333	0.00431
9	12	0	0.666666667	0	0.583333	0	0.053691	0	0.030172
10	27	0.185185185	0.481481481	0.296296	0.37037	0.033557047	0.087248	0.053333	0.043103
11	23	0.217391304	0.347826087	0.347826	0.217391	0.033557047	0.053691	0.053333	0.021552
12	21	0.428571429	0.047619048	0.238095	0.333333	0.060402685	0.006711	0.033333	0.030172
13	27	0.259259259	0.259259259	0	0.444444	0.046979866	0.04698	0	0.051724
14	18	0.166666667	0.388888889	0.333333	0.333333	0.020134228	0.04698	0.04	0.025862
15	27	0.259259259	0.074074074	0.074074	0.592593	0.046979866	0.013423	0.013333	0.068966
16	14	0.785714286	0	0	1	0.073825503	0	0	0.060345
18	18	0.166666667	0.333333333	0.055556	0.444444	0.020134228	0.040268	0.006667	0.034483
19	36	0.25	0.138888889	0.222222	0.138889	0.060402685	0.033557	0.053333	0.021552
20	21	0.476190476	0	0.47619	0.047619	0.067114094	0	0.066667	0.00431
21	24	0.083333333	0.416666667	0.208333	0.458333	0.013422819	0.067114	0.033333	0.047414
22	24	0.291666667	0.125	0.833333	0.041667	0.046979866	0.020134	0.133333	0.00431
23	30	0.066666667	0.133333333	0.733333	0	0.013422819	0.026846	0.146667	0
24	36	0.25	0.138888889	0.222222	0.138889	0.060402685	0.033557	0.053333	0.021552
25	31	0.064516129	0.516129032	0.064516	0.612903	0.013422819	0.107383	0.013333	0.081897
	600					0.590604027	0.798658	0.42	0.788793

Figure C.1: Shows percentage what percentage of windows within a meeting is of the given label type, as well as the what percentage of this label is within this meeting. The yellow meeting numbers are all that contain all modalities.

D

External observer instructions

Instructions

Read the questions before you start the video.
The questions will be answered for each 2-minute fragment.
Treat each fragment as independent from the other ones.
Try to answer with your first instinct.
To start a video, enter the link next to it into the command line.
Don't forget to wear headphones!

For each video put in the annotation form how much you agree on a 5-point scale.

1. Strongly disagree
2. Disagree
3. Neither agree nor disagree
4. Agree
5. Strongly agree

tc

1. The team seems united in trying to reach its goals[21].
2. The team seems to have conflicting aspirations[21].
3. The team allows everyone to input into the conversation[17, 22].
4. The team members share the same goal[41].

sc

1. The team members enjoy each other's company[21].
2. Overall, the team members are is involved in the discussion [17].
3. Overall, the team members are attentively listening to each other[17].
4. The team members appear comfortable with each other[17].