



EÖTVÖS LORÁND UNIVERSITY

FACULTY OF INFORMATICS

DEPT. OF ARTIFICIAL INTELLIGENCE

Analysis of Dominance, Leadership and Likability in Small Groups from Perceived Personality Synchrony and Body Gesture Energy Synchrony

Supervisor:

Dr. Kristian Arian Fenech

Assistant Professor

Author:

Shenghe Wang, Wei Xiao

Computer Science MSc

Budapest, 2022

Abstract

In this thesis, we propose a novel approach to predict dominance, leadership and likability using features related to the synchrony of perceived personality and the synchrony of body gestures energy. We evaluate these methods on the ELEA (Emerging LEadership Analysis) dataset, which provides audio and video recordings of small groups (3-4 participants) engaged in the winter survival task. To ensure the reliability of results and accommodate the ELEA dataset, we split the train and test set by leave-one-out and leave-one-group-out cross-validation. To estimate the synchrony between participants, we use the cross wavelet transform, which determines both the magnitude and relative phase over time between perceived personality pairwise signals and body gesture energy pairwise signals.

We focus on the following three tasks to predict dominance and leadership traits using perceived personality synchrony. The first task is the prediction of the more dominant and leading person pairwise between group members. The second task is to classify people in all groups if they are above or below the median dominance and leadership of the group. The third task is to estimate the absolute most dominant/leading person in the group. For the likability trait, we use body gesture energy synchrony to classify people in all groups if they are above or below the median likability of the group which can be considered the fourth task.

We compared different machine learning models and found the optimal results were produced by the fully connected neural network and XGBoost classifier with leave-one-out cross-validation estimated protocol. For the dominant and leading person pairwise task, our method achieves 67.47% and 66.25% respectively. As for the second task, our method can achieve 68.75% and 64.58% separately. On the third task, the predicting accuracy of the most dominant person is 75% and the leading person is 81.5%. For the last task, the best accuracy result is 78.6%. Our experiment proves that our proposed method is feasible.

Contents

1	Introduction	3
2	Related Work	5
3	Method	8
3.1	Dataset	8
3.2	Perceived Personality Estimation	9
3.3	Gesture Energy Calculation	10
3.3.1	Key points extraction	10
3.3.2	Energy time series	12
3.4	Interpersonal Synchrony	13
3.5	Cross Wavelet Transform (XWT)	13
3.5.1	Relationship with Synchrony	13
3.5.2	Mathematical Background	14
3.6	Features	17
3.7	Evaluated models	17
3.8	Cross-Validation	18
3.8.1	Leave-One-Out Cross-Validation	19
3.8.2	Leave-One-Group-Out Cross-Validation	19
4	Results	20
5	Discussion	24
6	Conclusion	26
	Bibliography	27
	List of Figures	32

List of Tables	33
----------------	----

Chapter 1

Introduction

Dominance and leadership analysis play an important role in the workplace and education [1]. Dominance works best when the leader’s job is to get everyone to align quickly and move in the same direction and people with high dominance are more likely to lead the team to achieve higher achievements. Likeable people have a good ability to get along with others and will get more opportunities for cooperation. Educational leadership is the process of mobilizing and directing the teachers, students and parents to achieve common educational goals [2]. Excellent educational leadership can improve the quality of teaching and academic management methods to provide students with a good education and academic environment. Likeable teachers also have high teaching ability, can better guide and teach, and impart knowledge in the form that students prefer [3].

Interpersonal synchrony is when two parties in an interaction engage similarly due to the rhythmic coordination of their behavioural patterns [4], and may be an important factor for social impression prediction. In this study, we use interpersonal synchrony (perceived personality and body gesture energy) indicated by related features to predict participants’ social impressions of their group members in a small group containing three to four people.

In this thesis, we focus on the following four tasks. The first is to find the more dominant and leading person pairwise. Secondly, to classify whether the person’s dominance and leadership are above or below the median value in his group. Thirdly, to find the most dominant and leading person in the group. Finally, to judge if the person’s likability is above the median value in the group. Aiming at the first three tasks, we propose using the synchrony of perceived personality to solve them. The

perceived personality traits we focus on are plasticity and stability. People with high plasticity and low stability are more likely to behave creatively [5]. In the fourth task, we propose using the synchrony of body gesture energy to analyze it. During a group meeting, the rhythm of participants' body gesture changing may follow the other members, so body gesture energy can be used to analyze likability. We use the cross wavelet transform to estimate the synchrony of perceived personality and synchrony of body gesture energy because it can measure the degree of synchronization between different components and the evolution of synchronization over time [6].

We test the proposed method on the ELEA dataset that includes 27 groups facing the winter survival task. The ELEA dataset consists of about ten hours of audio, video and transcription of the performance of the people in the group, as well as the annotation such as perceived leadership (PLead), perceived dominance (PDom) and perceived likability (PLike) [7].

The perceived personality was estimated through the transformer-based method proposed by Ádám *et al.* [8], which passes features extracted from the audio, video and transcription into a multimodal transformer [9] to estimate the Big Five personality traits [10], which contain openness to experience, conscientiousness, extraversion, agreeableness and neuroticism (OCEAN) five personalities. These five properties are then summarized into plasticity and stability.

To estimate the body gesture energy, we used software which applies a composite approach combining MaskRCNN [11], MediaPipe [12] and MiVOS [13] to extract key points from the hands, head and body of participants in the meeting video from the ELEA dataset. Then, we calculated the difference of key points between frames which defines the gesture energy to be used as the cross wavelet transform input. After getting the cross wavelet transform spectra of each pair, pair of features were extracted from these spectra. We tried a variety of different combinations of features to find out which combination would be better and used them to train several common machine learning models to get the best results.

During the evaluation, we utilised both the leave-one-out and the leave-one-group-out cross-validation strategies to ensure the accuracy of our experiments. The evaluation results show that our proposed method is feasible to analyse the perceived dominance, leadership and likability.

Chapter 2

Related Work

Personality has been studied for decades. A widely used personality model was proposed by Costa *et al.* in 1992 [14], called the Big Five personality traits, which divides the human personality into five categories that are openness to experience, conscientiousness, extraversion, agreeableness and neuroticism. Since these five personalities were found to interact with each other and they are not orthogonal factors without connection, Digman *et al.* made a factor correlation analysis of the Big Five personality traits based on this and found that there were two obvious high-order factors, one was related to agreeableness, conscientiousness and emotional stability (reverse neuroticism), and the other was related to extraversion and openness. The two higher-order factors were named alpha and beta [15]. After that, DeYoung *et al.* interpreted alpha and beta as stability and plasticity through biological characteristics and social relationships [16].

Plasticity means the ability to adapt to changes in the environment. People with high plasticity can adapt quickly to unfamiliar situations and acquire skills proficiently [17]. Stability means that this kind of person is not often sad, moody or has other emotionally unstable behaviour [18]. Therefore, in this study, it is appropriate to analyze dominance, and leadership using the plasticity and stability that can generalise the Big Five personality traits.

In 2012, Sanchez-Cortes *et al.* analyzed how emergent leaders were identified in small, newly formed groups. Communicative non-verbal cues were extracted from the group's audio-visual recordings to describe the characteristics of each participant, and the correlation between video and audio features was analyzed from the questionnaire. A method was developed to automatically infer emergent leaders using

these nonverbal cues [19].

In 2015, Okada *et al.* proposed a framework for extracting features from multi-party multimodal dialogues, which can be used to infer personality traits and emergent leadership. They calculated the motion energy of the head and body as the visual features and combined the visual features with audio features as the multimodal features. The proposed framework represented multimodal features as a combination of each participant’s audio and video non-verbal activities (*e.g.* intonation, visual focus, etc.) and group activities. The framework effectively improved the accuracy by 2% to 17% over the baseline [20].

In 2016, Fang *et al.* began to study the social dimension (personality traits and social impression) in small group interaction. They extracted three types of audio and visual features: intra-individual features (related to only one participant), dyadic features (associated with a pair of participants), and multivariate features (associated between one participant and the other members of the group). After using these three types of features to predict personality traits and social impressions, the interaction between personality traits and social impressions was analyzed. Finally, they found that the combination of intra-individual characteristics and multivariate characteristics could obtain the prediction accuracy of personality traits and social impressions (81.37%) [21].

In 2019, Okada *et al.* proposed a new feature extraction framework based on the one proposed in 2015. The framework can infer not only emergent leaders as the old framework did, but also personality traits, communication skills, and hiring decisions. The new framework still extracts multimodal features and describes the non-verbal activities of the participants. The difference with the old framework is that the new one can find co-occurrences between two corpora (SONVA and ELEA) in multiple modal sequences and then the features of these events were combined as input for training. Therefore, the new framework could not only extract cross-modal features and non-verbal activities of participants but also capture the relationships and interactions between people in multimodal. The final accuracy would be slightly higher than the model proposed in 2015, which was 2% to 13% corresponding to different personalities [22].

Also in 2019, Bai *et al.* focused on the most dominant person in the group (MDP) and prediction of more dominant person pairwise (PDP) tasks. They came up with three methods to improve the prediction results, which are Dominance

Ranking features (DR) that capture people’s interactions in a video, Dominance Ensemble Late Fusion algorithm (DELF) that combines DR with other short-term or long-term traits as an input for training, and Group Dominance Prediction (GDP) that compares each person with the others in the group. They used two methods of assigning labels, one based on whether a person’s score was above the group’s median, and one based on whether a person had the highest score in the group [23].

In 2020, Dunbar *et al.* proposed a method based on analysing interpersonal synchrony to detect the deceivers [24]. The trained coder watched split-screen videos of each conversation showing both partners side-by-side, recorded synchrony featured 11 categories such as laughing along with the partner, gestural mirroring, postural mirroring, eye synchrony (e.g. rolling eyes), head nods/shakes, verbal repetition, vocal synchrony and made the time series. By analysing the cross wavelet coherence, they found deceivers were more synchronized with their partner than truth-tellers in their faster (rather than slower) movements.

In 2022, Dunbar *et al.* summarized a much more efficient manner to analyse interpersonal synchrony [4]. An automated computer vision technique named OpenPose was used to handle the videos, estimated people’s joints as coordinate points, and people’s movements are represented as differences in coordinate points. The cross wavelet coherence was evaluated on time series bodily movement data, and the synchrony was determined by analysing the features of coherence such as relative phase information and the pattern of synchrony.

Chapter 3

Method

3.1 Dataset

To analyze emergent leaders in newly formed teams, Sanchez-Cortes *et al.* created the ELEA dataset [7]. In the English version, 102 participants were divided into 27 groups including 21 groups of four and 6 groups of three to face the winter survival task, which was the airplane crash. These participants were asked to rank 12 items according to how important they were in the survival task, then performed a five-minute individual ranking task and a group ranking task of up to 15 minutes just like Figure 3.1 shows. The group discussion parts were recorded in audio and video files over a total of about 10 hours. To ensure objectivity, two external observers watched the first five minutes of the videos and filled out the same questionnaires as the participants. In our study, due to missing data, perceived personality could only be assessed on a subset of 17 videos from the original dataset and body gesture energy could be estimated on a subset of 21 videos.

The annotation includes participants' self-reports on the Big Five personality traits, perceived personality, performance in survival tasks and the English version of the video and audio transcripts. Among them, perceived social impression includes perceived leadership (PLead), who was leading the group, perceived dominance (PDom) refers to a person in a position of power, perceived liking (PLike) refers to a friendly and kind person, perceived competence (PCom) shows experienced and competent people and dominance ranking (RDom), where each person in the group was asked to rate the dominance of the entire group on a scale of 1-5 (1 being the highest). In this study, we focus on perceived leadership, perceived

dominance and perceived liking.



Figure 3.1: Video frames from a group discussion in the ELEA dataset

3.2 Perceived Personality Estimation

In this study, the perceived personality time series we use are derived from the personality estimation method proposed by Ádám *et al.* [8]. This method took into account the factors contained in the audio, video and transcription text and combined these three kinds of features. Then, these features were passed in a multimodal transformer [9] to estimate the OCEAN five personalities.

Among the combination feature, the audio features were extracted by OpenSMILE [25] toolkit according to the eGeMAPS, the extended Geneva Minimalistic Acoustic Parameter Set [26] that contains 25 low-level audio parameters. The action unit represents the basic behaviour of a muscle or a group of muscles when facial expression changes, which can decrease appearance bias or discrimination when making decisions. There were 35 face units on the human face extracted by OpenFace [27] toolkit as the video features. Bidirectional Encoder Representation from Transformers (BERT) [28] was used to extract the semantic content in the transcriptions as the text feature.

After the three features corresponding to each frame were extracted, a 30-second sliding window with a 15-second step was made for each group’s audio, video and transcription text. Next, the average value of each sliding window was taken, and the signal time series of OCEAN personalities were estimated. Then, using plasticity and stability to generalise the OCEAN personalities. Finally, the ELEA dataset’s

audio, video and transcription were estimated into plasticity and stability time series. However, because there were some periods in which there was no corresponding change detected in participants, the values for these periods were set to null as Figure 3.2 shows.

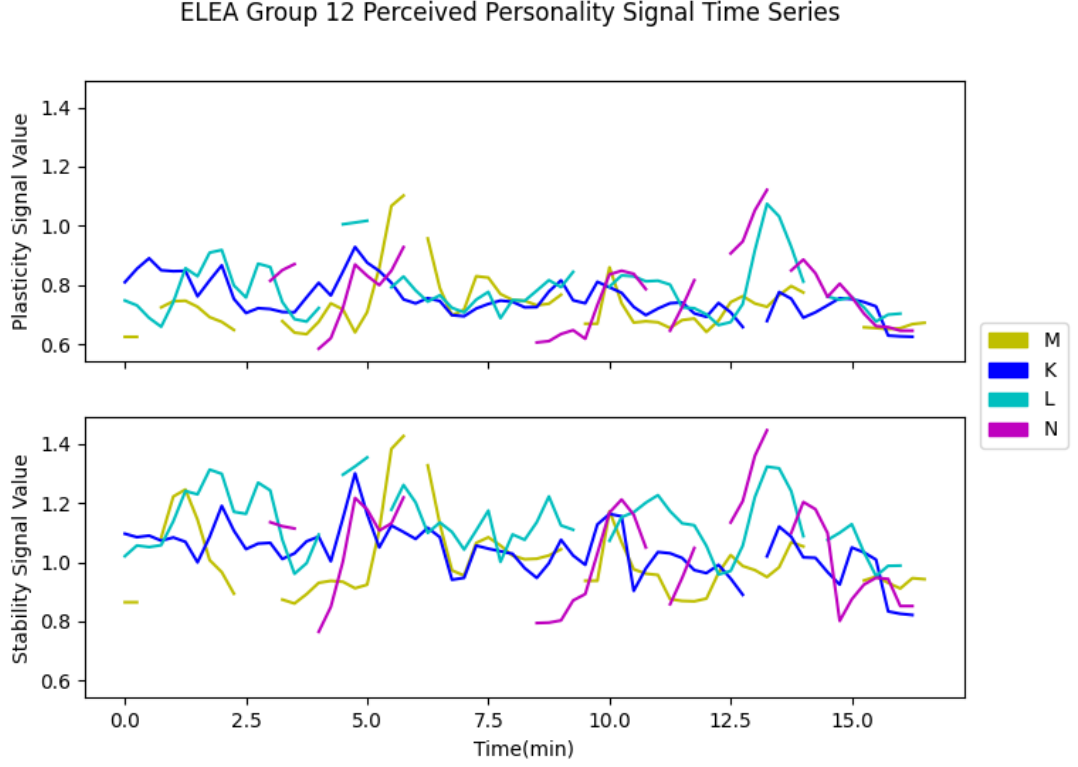


Figure 3.2: Plasticity and stability time series of the 12th group (4 persons) in the ELEA dataset, K, L, M, and N are code names for members within the group and gaps are periods in which changes cannot be detected

3.3 Gesture Energy Calculation

3.3.1 Key points extraction

We use software from the ELTE NIPG group to extract the key points of the hand, body and head. This software is a pipeline utilizing several pre-existing solutions to accurately perform hand segmentation and hand 3D key points prediction on videos, and output the key points of the head and body. It relies on the following three parts, MaskRCNN [11] which was trained on hand segmentation tasks and used the mmdetection implementation, MediaPipe Pose and Hands [12] and MiVOS [13].

The whole process consists of 4 steps:

1. Obtaining hand masks (MaskRCNN).

The current model detects a large number of false positives in addition to hands not belonging to the person of our interest. It was trained on VLM - Hand Over Face, HandOverFace and EgoHands datasets, most of which contain people viewed from the front. This might impact the accuracy in cases where the input video shows people viewed from other angles.

2. Retaining only the relevant masks (MediaPipe Body)

MediaPipe’s body pose model is able to detect a single central figure. The script only keeps the masks whose center is close to the center of the MediaPipe body pose’s hand key points (calculated by taking the center of the pinky, thumb and index key points). This step will output the head and body key points.

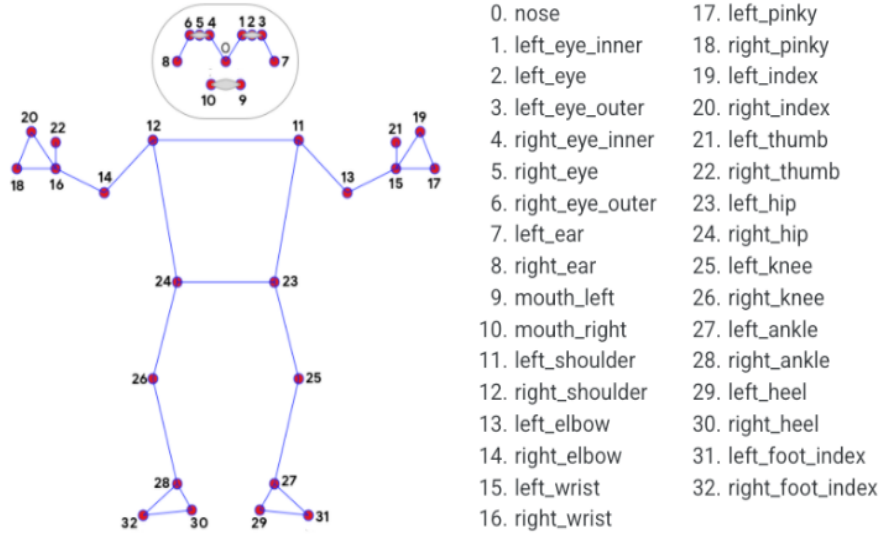


Figure 3.3: Body landmarks of MediaPipe [29]

3. MiVOS tracking.

Step 1 may fail to detect hands sometimes. This step aims to fill in these gaps by tracking undetected hands across frames using a modified MiVOS.

Instead of selecting the objects to be tracked through a GUI, the masks from Step 2 are used. The masks for the left and right hands are passed separately to MiVOS - this was found to improve tracking accuracy at the cost of having to run it twice on a single video.

4. Obtaining hand key points (MediaPipe Hands).

The masks from Step 3 are enlarged and used to crop out the hands from the original frames. MediaPipe’s 3D hand pose estimation is applied to the resulting frames.

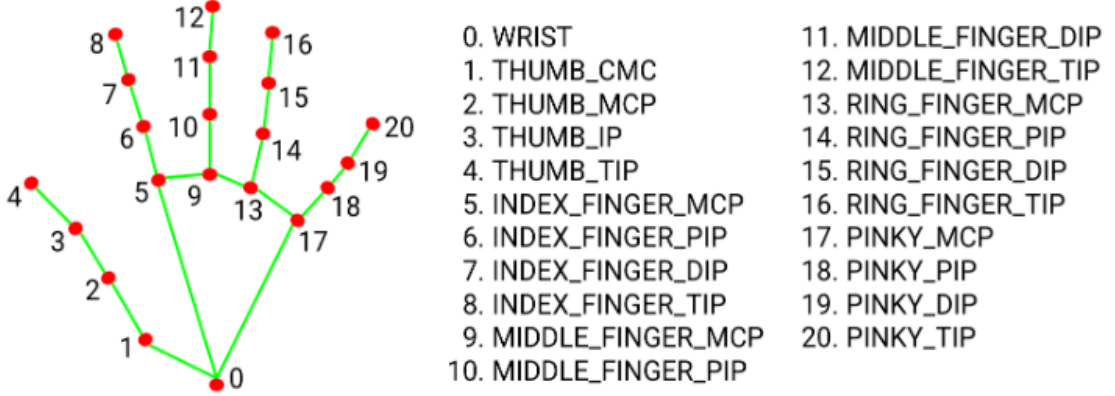


Figure 3.4: Hand landmarks of MediaPipe [29]

3.3.2 Energy time series

The energy was calculated by the frame difference of key points which are determined with the previously described method as equation (3.1), E is the energy, k is the index of key points from 1 to N , and x_i and y_i are the points’ coordinates of frames.

$$E = \frac{\sum_{k=1}^N \sqrt{(x_i - x_{i+1})^2 + (y_i - y_{i+1})^2}}{N} \quad (3.1)$$

For the head, the energy was determined by the positions of the eyes and nose. For the body, the energy was determined by the positions of the shoulders and arms. For the hands, the energy was determined by the positions of a centre point from the hand. Savitzky–Golay filter was performed on energy series to smooth them, which is a digital filter that can be applied to a set of digital data points to smooth the data, to increase the precision of the data without distorting the signal tendency.

Figure 3.5 shows the visualisation for the left-hand energy time series where the left part is a frame from the video and the right part is the plot of energy.

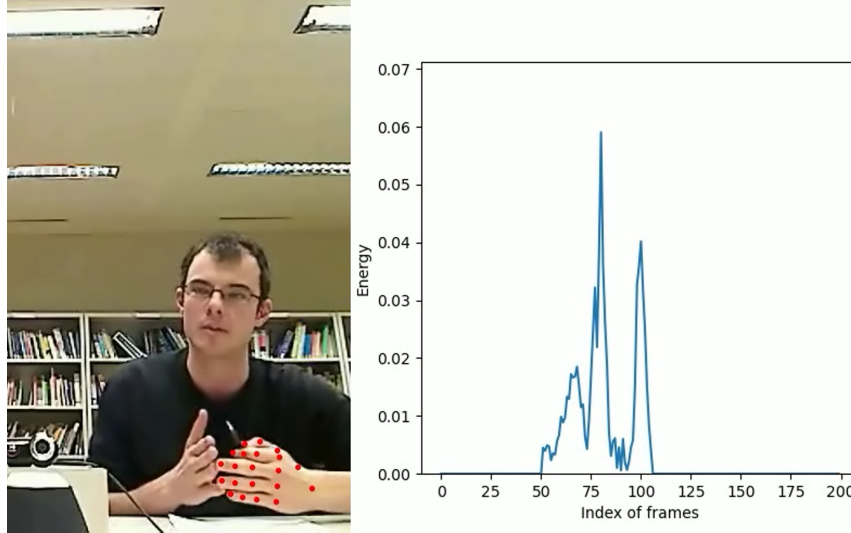


Figure 3.5: Visualisation for gesture energy of left hand. The left part is an example frame from the video, red points are the key points of the left hand. The right part is the plot of energy, the x-axis indicates the index number of frames and the y-axis indicates the energy's magnitude.

3.4 Interpersonal Synchrony

Interpersonal synchrony means that two or more people that change over time maintain a certain relative relationship during the interpersonal communication process. The temporal coordination of behaviour during interactions is a foundation for effective social exchange with synchronized actions enhancing perceptions of rapport and interpersonal connectedness [30].

The speaker's dominance, leadership and likability may also be reflected by the separate personality and body gesture behaviour synchrony. Therefore, in this study, we use the synchrony of perceived personality and the synchrony of body gesture energy to analyze the personalities.

3.5 Cross Wavelet Transform (XWT)

3.5.1 Relationship with Synchrony

In this study, we use the cross wavelet transform [6, 31, 32] to estimate the synchrony of perceived personality and synchrony of body gesture energy. The cross wavelet transform can reflect the interaction of a pair of time series, which is exactly in line with the definition of synchrony. For example, it can show information and

similarities and differences between two time series about how behavioural states change, how long it takes to progress from one state to another, and how many states occur in a given period. The spectrum of cross wavelet transform is shown in Figure 3.6.

The cross wavelet transform can calculate the relative phase of two signals, and the relative phase is an important element that can reflect the synchrony relationship between two or more components [33]. The relative phase is the angle between the closest corresponding points of two signal waves. If the angle equals 0° , it is in-phase, and 180° represents anti-phase. The relative phase is represented by the angle, but it can reflect whether there are differences between the two signals in the temporal or spatial sequence relation. For example, relative phases can show the relationship between two body joints during movement. Galgon *et al.* used relative phases in 2016 [34] to evaluate the interarticular relationship between shoulder-hip (proximal) and hip-ankle (distal) postural coordination in early and late learning.

Frequency is also an important factor related to synchrony. Behaviour can be nested in multiple time scales (multifrequency time series), and each of the scales contributes to the emergence of human communication and interaction [35]. The combination of frequency and relative phase provides a better understanding of synchronous action.

3.5.2 Mathematical Background

Since we use the time series of team members' perceived personality and body gesture energy through video, audio and transcription, there would be some people who did not speak or act in a certain period of time, so a small amount of null value would appear in the time series. As the cross wavelet transform can not be used in the case of missing data in the time series, it is necessary to use interpolation to fill in the missing value. In this study, we chose the mean of the non-null values in the time series to fill in the null values after comparing some interpolation methods (e.g. linear and quadratic interpolation) due to less loss.

The wavelet transform is the basis of the cross wavelet transform, which inherits and develops the idea of short-time Fourier transform localization and overcomes the shortcoming that the size of the time-frequency window does not change with frequency. It can carry out multi-scale refinement through stretching or squeezing

(dilation) analysis function (the mother wave function) and move it (translation) to analyze the frequency locally. The choice of analysis function is neither unique nor arbitrary. It is a function with unit energy, chosen so that it has compact support or sufficiently fast decay which enables one to obtain localization in time. The Morlet wavelet has a complex value that enables one to extract information about the amplitude and phase of the process being analyzed. The use of a complex wavelet with phase computation is fundamental in quantifying the interaction between signals. Therefore, the Morlet wavelet $\psi(t)$ is used as the analysis function in this study

$$\psi(t) = \frac{1}{\pi^{1/4}} e^{i2\pi f_0 t} e^{-t^2/2} \quad (3.2)$$

and we define parameter a which can shrink the analysis function and parameter b which can shift the analysis function. After replace t with $(t-b)/a$, we get

$$\psi\left(\frac{t-b}{a}\right) = \frac{1}{\pi^{1/4}} e^{i2\pi f_0 (t-b)/a} e^{-\frac{1}{2}[(t-b)/a]^2} \quad (3.3)$$

The wavelet transform produces complex numbers for a real signal when a complex mother wavelet function is used. Thus, it can be written as

$$T(a, b) = |T(a, b)| e^{i\phi(a, b)} \quad (3.4)$$

If we consider its real and imaginary components, $Re(T(a, b))$ and $Im(T(a, b))$, the modulus of the wavelet transform can be written as

$$|T(a, b)| = \sqrt{Re(T(a, b))^2 + Im(T(a, b))^2} \quad (3.5)$$

and the phase of wavelet transform can be written as

$$\phi(a, b) = \tan^{-1} \left[\frac{Im(T(a, b))}{Re(T(a, b))} \right] \quad (3.6)$$

The cross wavelet transform is obtained by multiplying the wavelet T_g by the complex conjugate of another wavelet T_h

$$XWT_{g,h}(a, b) = T_g(a, b) \overline{T_h(a, b)} \quad (3.7)$$

and relative phase ϕ_{xwt} is

$$\phi_{xwt}(a, b) = \phi_g(a, b) - \phi_h(a, b) \quad (3.8)$$

The frequency of cross wavelet transform is calculated by the Fourier Transform $\hat{x}$ for wavelets of different scales

$$\hat{x}(f) = \int_{-\infty}^{+\infty} x(t) e^{-i(2\pi f)t} dt \quad (3.9)$$

Cross wavelet transform power E can highlight the location and scale of dominant energetic features within the signal

$$E(a, b) = |T(a, b)|^2 \quad (3.10)$$

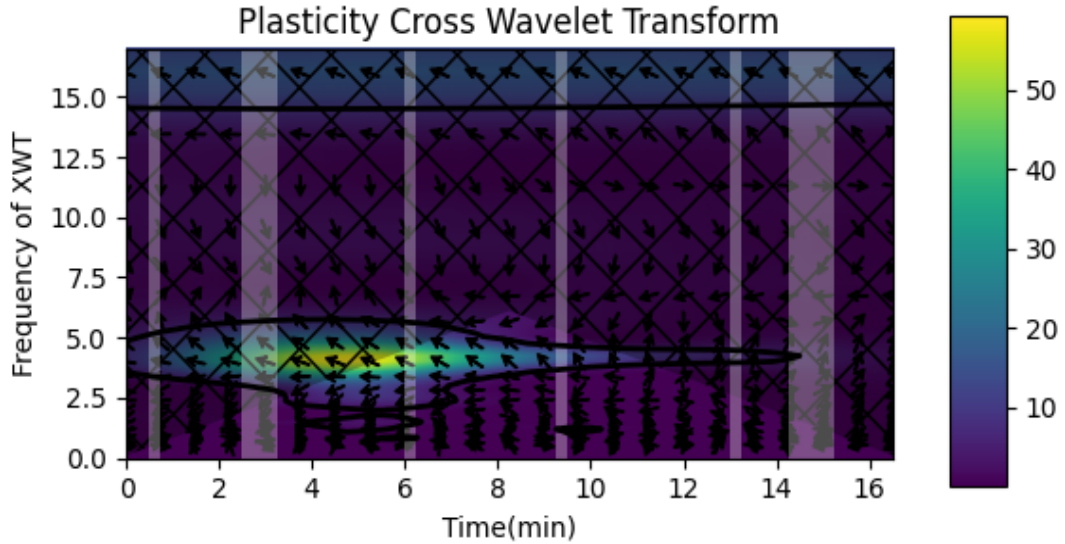


Figure 3.6: The x-axis represents the time, the y-axis represents the frequency of the cross wavelet transform, the background is the cross wavelet transform power, and the arrow represents the relative phase of the two sets of signals. If the arrow points to the right (in-phase), it means signal A leads signal B and left (anti-phase) means A lags behind B. The white shadows are the missing value parts.

As shown in Figure 3.6, there are several light regions surrounded by thick black lines in the background of the spectrum, which is called the significant region. The confidence level is computed as the product of the cross wavelet transform power at each scale by the desired significance level from the chi-square (χ^2) distribution. When the cross wavelet transform is higher than the associated confidence level, it is regarded as significant. In this study, according to the experience of Grinsted *et al.*

[36], the significance level is set to 86.46%. Significant areas can represent short-term changes, while other areas represent long-term changes. We analyze the synchrony of two people by focusing on short-term changes.

3.6 Features

In this study, we focus on short-term changes in synchrony, so the extracted features are also short-term features. Features were extracted from the significant regions of the cross wavelet transform spectra for each of the two individuals.

Maximum Cross Wavelet Transform Power. In the cross wavelet transform spectrum, there may be multiple significant regions, all of which represent different short-term changes, among which the region with the highest power can more prominently show the synchrony relationship between the two persons. Therefore, we identify the highest point in the entire power spectrum and use this to identify the region where short-term changes are most pronounced.

Frequency. After finding the most significant region, we then find the top 20% part of the region to further enhance the features. We take this list of frequencies in this region and average them.

Relative Phase. Similar to frequency, we take the average of the relative phases of this significant region. From this, we can intuitively see the synchronous relationship between two people. If the degree is less than 90° , it means that person A is leading B.

Standard Deviation of Relative Phase. After getting the list of relative phases, the standard deviation is calculated. The difference between the relative phases of the significant region can be seen by the standard deviation.

Duration of Significant Region. That is the duration of the significant region, and we can see how long the synchrony of this short change lasted.

3.7 Evaluated models

We use the grid search method to determine the hyperparameters of the following models and the optimal case of different combinations.

Support vector machine (SVM) is a kind of generalized linear classifier which uses a supervised learning method to classify binary data. It is one of the common

kernel learning methods and can perform nonlinear classification by kernel method. In this study, the *RBF* function is found to be the best kernel function for SVM.

The random forest is a typical ensemble learning classifier, contains multiple decision trees and adopts the method of voting, the classification of the individual trees with the most votes is the final one. In this work, the suitable number for decision trees is *100* and the maximum depth of the trees is *2*.

Gradient boosting is a machine learning technique used in regression and classification tasks, among others. It gives a prediction model in the form of an ensemble of weak prediction models, which are typically decision trees [37]. The best maximum depth of the decision trees is *3* and the learning rate is *0.1* in this study.

XGBoost is an improvement of the gradient boosting algorithm, and it's an optimized distributed gradient boosting library designed to be efficient, flexible, and portable. The best maximum depth of trees is *2* and the learning rate is also *0.1* in this study.

The fully connected neural network draws on both perceptron and bionics to achieve the recognition and classification of information sources. The fully connected neural network in this study is set as two hidden layers and both of which contain *10* neurons, with *tanh()* and *sigmoid()* as the activation function of each layer respectively. *Adam* optimizer is used to iteratively update the weight between each layer and the learning rate is set to *1e-5*. Finally, *cross entropy* is used as the loss function.

3.8 Cross-Validation

The types of data in the training set are not evenly distributed may result in the model underfitting or overfitting. The cross-validation estimation was proposed by Stone in 1974 [38] to solve this problem. Cross-validation divides all the data into *k* equal parts and selects each part as the test set (*i. e.* also called *k-fold* cross-validation), the complement set as the training set, and finally, take the average value of all the test results to evaluate the model.

3.8.1 Leave-One-Out Cross-Validation

Leave-One-Out cross-validation (LOOCV) is a special case of k -fold cross-validation. When k is set to be equal to the length of the dataset, they are the same. LOOCV traverses each sample in the dataset as the test set to evaluate the model, which can further eliminate the influence of data distribution.

3.8.2 Leave-One-Group-Out Cross-Validation

Similar to the Leave-One-Out Cross-Validation, the Leave-One-Group-Out Cross-Validation (LOGOCV) estimation method leaves each group of data as a test set in each iteration and calculates the average of the test results for all groups at the end. The aim is to adapt the model to situations where each set of data has its unique characteristics, such as some widely used fMRI data analysis practices may be exaggerated due to the independence of subjects [39]. By doing this we can interpret the output of cross-validation as the capacity of the models for generalizing to unseen clusters. The ELEA dataset used in this study was divided into groups, and the meetings were also conducted by groups. The scope of our task is to analyze people’s personalities within the group, the extent to which the members of each group influence each other is different and cannot be ignored. Therefore, it is necessary to train and evaluate the model by leave-one-group-out cross-validation.

Chapter 4

Results

In the prediction of the more dominant and leading person pairwise task, we used the features mentioned above as input to the model and assign labels by comparing the PDom and PLead scores of two people. Those with a high score are assigned 1 and those with a low score are assigned 0 (1 is the maximum and 5 is the minimum in the score ranking rule). As far as we know, we are the first to test the task on the ELEA dataset, so there is no baseline. The optimal accuracies of several machine learning models in this task are shown in Table 4.1. The best results are achieved by the fully connected neural network and the estimated protocol is leave-one-out cross-validation.

Model	Dominance Acc.	Leadership Acc.
SVM	0.6024	0.5625
Random Forest Classifier	0.5904	0.5625
Gradient Boosting Classifier	0.6145	0.5625
XGBoost	0.5783	0.5625
Fully Connected Neural Network	0.6747	0.6625

Table 4.1: Comparison of the more dominant and leading person pairwise predictions from different models and the best results are from the fully connected neural network using leave-one-out cross-validation

When classifying whether people’s dominance and leadership are above or below the median value in the group, we have followed the label distribution method of Okada *et al.* in 2015 and 2019 [20, 22]. Firstly, we calculated the median value of PDom and PLead in each group as the threshold, and then judged if people’s scores were above this threshold. People who had higher scores than the threshold were assigned 1 , and in contrast, if someone’s scores were below this threshold, he would

be assigned to 0. In this task, we set the best baseline from the study of Bai *et al.* (dominance) [23] and Okada *et al.* (leadership) [22]. As shown in Table 4.2, the optimal accuracies in this task are from the fully connected neural network and XGBoost classifier with leave-one-out cross-validation estimated protocol, both of which are slightly lower than the accuracy of the baseline.

Model	Dominance Acc.	Leadership Acc.
SVM	0.625	0.625
Random Forest Classifier	0.6042	0.5833
Gradient Boosting Classifier	0.6042	0.625
XGBoost	0.6042	0.6458
Fully Connected Neural Network	0.6875	0.625
Baseline	0.7647	0.7549

Table 4.2: Comparison of dominance and leadership in small group based on median value predictions by different models, the best results kept by the baseline

This task is to identify the most dominant and leading person in each group. We use the label assignment method the same as the study done by Sanchez-Cortes *et al.* in 2012 [19]. We compared the PDom and PLead scores in each group to find the person with the highest score in his group. Only this person was seen as the most dominant and leading, whose label was set to 1, while others' labels were set to 0. The best baseline in this task was set from the study of Bai *et al.* (dominance) [23] and Sanchez-Cortes *et al.*, (leadership) [19]. The prediction accuracies and baseline of this task are shown in Table 4.3. The best results are achieved by the XGBoost classifier and the estimated protocol is leave-one-out cross-validation.

Model	Dominance Acc.	Leadership Acc.
SVM	0.75	0.7083
Random Forest Classifier	0.75	0.7917
Gradient Boosting Classifier	0.75	0.7083
XGBoost	0.75	0.8125
Fully Connected Neural Network	0.75	0.7917
Baseline	0.775	0.857

Table 4.3: Comparison of the most dominant and leading person in small group predictions by different models, the best results were given by the baseline

For the likability prediction, we collect hands, head and body's energy time series, and do an ablation study by using different combinations of these three parts' cross wavelet transform results as the input of classification models to find which

combination is most effective for the prediction. The accuracy of leave-one-out cross-validation to evaluate the model’s performance is shown in Table 4.4. After the comparison, the best performance was achieved by a fully connected neural network, and the best combination is hands, head and body energy.

To our knowledge, the most accurate method of likability prediction on the ELEA dataset is proposed by Fang *et al.* [21]. They extracted features such as Mean Length of Speaking Turns, Head (Body) Motion Length per Turn, and Successful Speech Interrupt from videos and use Ridge Regression to predict a likability score. By comparing with the median score in the group, output a prediction label (0 if smaller, 1 if larger). Their method’s best performance is 77.45%, and the accuracy of the method we proposed is 1.15% more than it.

Combination	Method	Likability Acc.
Hands, head, body	SVM	0.549
	Random Forest	0.618
	Gradient Boosting	0.608
	XGBoost	0.627
	Fully Connected Neural Network	0.786
Hands, head	SVM	0.549
	Random Forest	0.608
	Gradient Boosting	0.510
	XGBoost	0.510
	Fully Connected Neural Network	0.767
Hands, body	SVM	0.549
	Random Forest	0.559
	Gradient Boosting	0.578
	XGBoost	0.539
	Fully Connected Neural Network	0.747
Head, body	SVM	0.598
	Random Forest	0.607
	Gradient Boosting	0.549
	XGBoost	0.598
	Fully Connected Neural Network	0.739
Hands	SVM	0.549
	Random Forest	0.559
	Gradient Boosting	0.510
	XGBoost	0.421
	Fully Connected Neural Network	0.723
Head	SVM	0.559
	Random Forest	0.618
	Gradient Boosting	0.490
	XGBoost	0.480
	Fully Connected Neural Network	0.709
Body	SVM	0.607
	Random Forest	0.589
	Gradient Boosting	0.509
	XGBoost	0.559
	Fully Connected Neural Network	0.769

Table 4.4: Comparison of different classification models in an ablation study, and the best results are from hands, head and body.

Chapter 5

Discussion

For the prediction of dominance and leadership, we focused on three tasks and all of which are binary classification tasks, so the probability of random guess is 50%. The highest accuracies of the three tasks are all higher than the random guess probability. Except for the first task which has no baseline to compare, the optimal results of the latter two tasks are slightly lower than the best baseline results of the same type of task [19, 22, 23], but the differences are not much. The gap may be caused by the following reasons. First, as we mentioned in section 3.1, we only use the subset of 17 meetings from the original 27 in the ELEA dataset due to missing data, and the baseline results were estimated on the entire dataset, so it is not possible to compare which method is better by the size of the number directly. Second, these baseline tasks directly take the features extracted from audio and video as the input of the model. However, we use the features of audio and video text to estimate the perceived personality and then extracted features from the cross wavelet transform spectra as the input of the model, so there would be more loss than their method. There is inevitable loss here, but it supplies the possibility that we can use more than audio and video features to analyze personality. Based on the accuracy of these three tasks, it can be concluded that our proposed method of predicting dominance and leadership with the synchrony of perceived personality is feasible.

For the prediction of likability, it can be found that the best result which used the hands, head and body combination is higher than the baseline. From the ablation study, the result using the head synchrony has the lower accuracy, probably because head movements are more simple than hand and body movements and cannot detect large energy fluctuations. The combination of hand, body and head neutralizes the

loss of energy from individual parts, improving the prediction accuracy. Compared with the state of the art (Fang *et al.*), the accuracy of the best combination is 1.15% higher than it. Similar to the dominance and leadership tasks, we can only evaluate our method on a subset of the ELEA dataset so we cannot conclude that our method is better through the numbers, but the result can show that using body gesture energy synchrony to predict likability is feasible.

Chapter 6

Conclusion

In this thesis, we show the proposed novel method which is using synchrony to predict dominance, leadership, and likability. Dominance and leadership were predicted by the perceived personality synchrony, while likability was predicted by body gesture energy synchrony. To quantify this synchronization relationship, the cross wavelet transform is applied. The ELEA dataset was used to estimate our proposed method. First, we extracted the body gesture energy time series from the video and audio, then pass it and the perceived personality time series into the cross wavelet transform to generate the spectra and finally extracted the short-term features from the cross wavelet transform spectra into the classification model for prediction.

In our experiment, the accuracy of our predictions for the four tasks reached 64.47% to 78.6%. As discussed in section 5, these results show that our proposed method is feasible and meaningful.

There are also some limitations of our proposed method, and the main one is the features extracted from the cross wavelet transform spectra. In the present experiment, we extracted features from the more significant region in the most significant regions and set the significance level to 86.46%. All these are determined according to the experience, none of these parameters is specified clearly, so more parameter settings are worth trying. Due to the novelty of our proposed method, we do not have a very suitable dataset, so our method can only be evaluated on the subsets of the ELEA data set and the results can not be directly compared with the best baseline. Next, we'll try a different set of features than we did for this project and combine perceived personality and body gesture energy for further analysis.

Bibliography

- [1] Hemant Kakkar and Niro Sivanathan. “The impact of leader dominance on employees’ zero-sum mindset and helping behavior”. en. In: *J. Appl. Psychol.* 107.10 (Oct. 2022), pp. 1706–1724.
- [2] Wikipedia contributors. *Educational leadership*. https://en.wikipedia.org/w/index.php?title=Educational_leadership&oldid=1097606928. July 2022.
- [3] Michael Delucchi. “Don’t worry, be happy: Instructor likability, student perceptions of learning, and teacher ratings in upper-level sociology courses”. In: *Teach. Sociol.* 28.3 (July 2000), p. 220.
- [4] Norah E. Dunbar, Judee K. Burgoon, and Ken Fujiwara. “Automated Methods to Examine Nonverbal Synchrony in Dyads”. In: *Understanding Social Behavior in Dyadic and Small Group Interactions*. Ed. by Cristina Palmero et al. Vol. 173. Proceedings of Machine Learning Research. PMLR, 2022, pp. 204–217.
- [5] Paul J. Silvia et al. “Openness to experience, plasticity, and creativity: Exploring lower-order, high-order, and interactive effects”. In: *Journal of Research in Personality* 43.6 (2009), pp. 1087–1090. ISSN: 0092-6566. DOI: <https://doi.org/10.1016/j.jrp.2009.04.015>.
- [6] Johann Issartel et al. “The relevance of the cross-wavelet transform in the analysis of human interaction – a tutorial”. In: *Frontiers in Psychology* 5 (2015). ISSN: 1664-1078. DOI: [10.3389/fpsyg.2014.01566](https://doi.org/10.3389/fpsyg.2014.01566).
- [7] Dairazalia Sanchez-Cortes, Oya Aran, and Daniel Gática-Pérez. “An Audio Visual Corpus for Emergent Leader Analysis”. In: 2011.

- [8] Ádám Fodor et al. “Multimodal Sentiment and Personality Perception Under Speech: A Comparison of Transformer-based Architectures”. In: *Understanding Social Behavior in Dyadic and Small Group Interactions*. Ed. by Cristina Palmero et al. Vol. 173. Proceedings of Machine Learning Research. PMLR, 2022, pp. 218–241.
- [9] Yao-Hung Hubert Tsai et al. “Multimodal Transformer for Unaligned Multimodal Language Sequences”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 6558–6569. DOI: 10.18653/v1/P19-1656.
- [10] O. P. John. “The "Big Five" factor taxonomy: Dimensions of personality in the natural language and in questionnaires.” In: *In L. A. Pervin (Ed.), Handbook of personality: Theory and research (pp.).* (1990), 66–100.
- [11] Kaiming He et al. “Mask R-CNN”. In: *CoRR* abs/1703.06870 (2017). arXiv: 1703.06870.
- [12] Fan Zhang et al. “MediaPipe Hands: On-device Real-time Hand Tracking”. In: *CoRR* abs/2006.10214 (2020). arXiv: 2006.10214.
- [13] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. “Modular Interactive Video Object Segmentation: Interaction-to-Mask, Propagation and Difference-Aware Fusion”. In: *CoRR* abs/2103.07941 (2021). arXiv: 2103.07941.
- [14] Paul T Costa Jr and Robert R McCrae. “The five-factor model of personality and its relevance to personality disorders”. In: *J. Pers. Disord.* 6.4 (Dec. 1992), pp. 343–359.
- [15] John M Digman. “Higher-order factors of the big five”. In: *J. Pers. Soc. Psychol.* 73.6 (1997), pp. 1246–1256.
- [16] Colin G DeYoung, Jordan B Peterson, and Daniel M Higgins. “Higher-order factors of the Big Five predict conformity: Are there neuroses of health?” In: *Personality and Individual Differences* 33.4 (2002), pp. 533–552. ISSN: 0191-8869. DOI: [https://doi.org/10.1016/S0191-8869\(01\)00171-4](https://doi.org/10.1016/S0191-8869(01)00171-4).
- [17] Adema Ribic. “Stability in the face of change: Lifelong experience-dependent plasticity in the sensory cortex”. In: *Front. Cell. Neurosci.* 14 (Apr. 2020), p. 76.

- [18] Kendra Cherry. *What Are the Big 5 Personality Traits?* <https://www.verywellmind.com/the-big-five-personality-dimensions-2795422>. 2022.
- [19] Dairazalia Sanchez-Cortes et al. “A Nonverbal Behavior Approach to Identify Emergent Leaders in Small Groups”. In: *IEEE Transactions on Multimedia* 14 (June 2012), pp. 816–832. DOI: 10.1109/TMM.2011.2181941.
- [20] Shogo Okada, Oya Aran, and Daniel Gatica-Perez. “Personality Trait Classification via Co-Occurrent Multiparty Multimodal Event Discovery”. In: *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ICMI ’15. Seattle, Washington, USA: Association for Computing Machinery, 2015, 15–22. ISBN: 9781450339124. DOI: 10.1145/2818346.2820757.
- [21] Sheng Fang, Catherine Achard, and Séverine Dubuisson. “Personality classification and behaviour interpretation: an approach based on feature categories”. In: Oct. 2016, pp. 225–232. DOI: 10.1145/2993148.2993201.
- [22] Shogo Okada et al. “Modeling Dyadic and Group Impressions with Intermodal and Interperson Features”. In: *ACM Trans. Multimedia Comput. Commun. Appl.* 15.1s (2019). ISSN: 1551-6857. DOI: 10.1145/3265754.
- [23] Chongyang Bai et al. “Predicting dominance in multi-person videos”. English (US). In: *Proceedings of the 28th International Joint Conference on Artificial Intelligence, IJCAI 2019*. Ed. by Sarit Kraus. IJCAI International Joint Conference on Artificial Intelligence. International Joint Conferences on Artificial Intelligence, 2019, pp. 4643–4650. DOI: 10.24963/ijcai.2019/645.
- [24] Dunbar et al. “Strategic Synchrony and Rhythmic Similarity in Lies About Ingroup Affiliation”. In: *Journal of Nonverbal Behavior* 44 (2020), 153–172. DOI: <https://doi.org/10.1007/s10919-019-00321-2>.
- [25] Florian Eyben, Martin Wöllmer, and Björn Schuller. “openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor”. In: *Proc. ACM Multimedia (MM)*, ACM. Florence, Italy, 2010, pp. 1459–1462.
- [26] Florian Eyben et al. “The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing”. In: *IEEE*

- Transactions on Affective Computing* 7.2 (2016), pp. 190–202. DOI: 10.1109/TAFFC.2015.2457417.
- [27] Tadas Baltrušaitis et al. “OpenFace 2.0: Facial Behavior Analysis Toolkit”. In: *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. 2018, pp. 59–66. DOI: 10.1109/FG.2018.00019.
- [28] Jacob Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2018. DOI: 10.48550/ARXIV.1810.04805.
- [29] *Home*. <https://google.github.io/mediapipe/>. Accessed: December 1, 2022.
- [30] Lynden K. Miles, Louise K. Nind, and C. Neil Macrae. “The rhythm of rapport: Interpersonal synchrony and social perception”. In: *Journal of Experimental Social Psychology* 45.3 (2009), pp. 585–589. ISSN: 0022-1031. DOI: <https://doi.org/10.1016/j.jesp.2009.02.002>. URL: <https://www.sciencedirect.com/science/article/pii/S0022103109000328>.
- [31] Paul S Addison. *The illustrated wavelet transform handbook*. Second edition. | Boca Raton, FL : CRC Press, Taylor & Francis: CRC Press, 2016. DOI: <https://doi.org/10.1201/9781315372556>.
- [32] Johann Issartel et al. “A Practical Guide to Time—Frequency Analysis in the Study of Human Motor Behavior: The Contribution of Wavelet Transform”. In: *Journal of Motor Behavior* 38.2 (2006), pp. 139–159. DOI: 10.3200/JMBR.38.2.139-159.
- [33] Scott Kelso. “Dynamic Patterns: The Self-Organization of Brain and Behavior”. In: Jan. 1995.
- [34] Anne K Galgon and Patricia A Shewokis. “Using mean absolute relative phase, deviation phase and point-estimation relative phase to measure postural coordination in a serial reaching task”. en. In: *J. Sports Sci. Med.* 15.1 (Mar. 2016), pp. 131–141.
- [35] R C Schmidt et al. “Measuring the dynamics of interactional synchrony”. en. In: *J. Nonverbal Behav.* 36.4 (Dec. 2012), pp. 263–279.

- [36] A. Grinsted, J. C. Moore, and S. Jevrejeva. “Application of the cross wavelet transform and wavelet coherence to geophysical time series”. In: *Nonlinear Processes in Geophysics* 11.5/6 (2004), pp. 561–566. DOI: 10.5194/npg-11-561-2004.
- [37] S. Madeh Piryonesi and Tamer E. El-Diraby. “Data Analytics in Asset Management: Cost-Effective Prediction of the Pavement Condition Index”. In: *Journal of Infrastructure Systems* 26.1 (2020), p. 04019036. DOI: 10.1061/(ASCE)IS.1943-555X.0000512.
- [38] M. Stone. “Cross-Validatory Choice and Assessment of Statistical Predictions”. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 36.2 (1974), pp. 111–147. ISSN: 00359246. (Visited on 11/23/2022).
- [39] Michael Esterman et al. “Avoiding non-independence in fMRI data analysis: Leave one subject out”. In: *NeuroImage* 50.2 (2010), pp. 572–576. ISSN: 1053-8119. DOI: <https://doi.org/10.1016/j.neuroimage.2009.10.092>.

List of Figures

3.1	Video frames from a group discussion in the ELEA dataset	9
3.2	Plasticity and stability time series of the 12th group (4 persons) in the ELEA dataset, K, L, M, and N are code names for members within the group and gaps are periods in which changes cannot be detected .	10
3.3	Body landmarks of MediaPipe [29]	11
3.4	Hand landmarks of MediaPipe [29]	12
3.5	Visualisation for gesture energy of left hand. The left part is an example frame from the video, red points are the key points of the left hand. The right part is the plot of energy, the x-axis indicates the index number of frames and the y-axis indicates the energy's magnitude.	13
3.6	The x-axis represents the time, the y-axis represents the frequency of the cross wavelet transform, the background is the cross wavelet transform power, and the arrow represents the relative phase of the two sets of signals. If the arrow points to the right (in-phase), it means signal A leads signal B and left (anti-phase) means A lags behind B. The white shadows are the missing value parts.	16

List of Tables

4.1	Comparison of the more dominant and leading person pairwise predictions from different models and the best results are from the fully connected neural network using leave-one-out cross-validation	20
4.2	Comparison of dominance and leadership in small group based on median value predictions by different models, the best results kept by the baseline	21
4.3	Comparison of the most dominant and leading person in small group predictions by different models, the best results were given by the baseline	21
4.4	Comparison of different classification models in an ablation study, and the best results are from hands, head and body.	23