2021

Multimodal Human Group Behavior Analysis

Chongyang Bai
Chongyang.Bai.GR@Dartmouth.edu

MULTIMODAL HUMAN GROUP BEHAVIOR ANALYSIS

A Thesis

Submitted to the Faculty

in partial fulfillment of the requirements for the

degree of

Doctor of Philosophy

in

Computer Science

by

Chongyang Bai

DARTMOUTH COLLEGE

Hanover, New Hampshire

May 2021

Examining Committee:

V.S. Subrahmanian

V.S. Subrahmanian, Ph.D., Chair

Bo Zhu

Bo Zhu, Ph.D.

Soroush Vosoughi, Ph.D.

Srijan Kumar

Srijan Kumar, Ph.D.

F. Jon Kull, Ph.D
Dean of the Guarini School of Graduate and Advanced Studies

Abstract

Human behaviors in a group setting involve a complex mixture of multiple modalities: audio, visual, linguistic, and human interactions. With the rapid progress of AI, automatic prediction and understanding of these behaviors is no longer a dream. In a negotiation, discovering human relationships and identifying the dominant person can be useful for decision making. In security settings, detecting nervous behaviors can help law enforcement agents spot suspicious people. In adversarial settings such as national elections and court defense, identifying persuasive speakers is a critical task. It is beneficial to build accurate machine learning (ML) models to predict such human group behaviors.

There are two elements for successful prediction of group behaviors. The first is to design domain-specific features for each modality. Social and Psychological studies have uncovered various factors including both individual cues and group interactions, which inspire us to extract relevant features computationally. In particular, the group interaction modality plays an important role, since human behaviors influence each other through interactions in a group. Second, effective multimodal ML models are needed to align and integrate the different modalities for accurate predictions. However, most previous work ignored the group interaction modality. Moreover, they only adopt early fusion or late fusion to combine different modalities, which is not optimal.

This thesis presents methods to train models taking multimodal inputs in group

interaction videos, and to predict human group behaviors. First, we develop an ML algorithm to automatically predict human interactions from videos, which is the basis to extract interaction features and model group behaviors. Second, we propose a multimodal method to identify dominant people in videos from multiple modalities. Third, we study the nervousness in human behavior by a developing hybrid method: group interaction feature engineering combined with individual facial embedding learning. Last, we introduce a multimodal fusion framework that enables us to predict how persuasive speakers are.

Overall, we develop one algorithm to extract group interactions and build three multimodal models to identify three kinds of human behavior in videos: dominance, nervousness and persuasion. The experiments demonstrate the efficacy of the methods and analyze the modality-wise contributions.

Preface

The thesis is the original work of the author Chongyang Bai. The work presented was conducted in the Dartmouth Security and AI lab and Stanford Network Analysis group, funded by the US Army Research Office (ARO Grant W911NF1610342). The collection of the major dataset in this thesis, Resistance, was a joint effort by University of Arizona, University of California, Santa Barbara, Rutgers University, Stanford University, the University of Maryland and Dartmouth College.

Acknowledgements

Throughout the journey of my Ph.D and the process of writing the dissertation, I'm truly fortunate to receive plenty of invaluable guidance, advice, and support.

I'd like to express the deepest gratitude to my advisor, Prof. V.S. Subrahmanian, for his dedicated guidance and precious advice over the four years. Thank you for bringing me into the door of the machine learning field, teaching me to discover interesting research problems, and providing me with practical approaches from multiple directions. Your passionate and rigorous research attitude has shaped my working style and will motivate me onwards. What's more, many thanks for offering me the opportunities to collaborate with other brilliant researchers at Stanford University, and supporting me to intern at Google AI for industrial-level multimodal research.

I'm also grateful to Prof. Jure Leskovec and Prof. Srijan Kumar for accepting me and providing their supervision during the two summers at Stanford. There I also got to know smart and diligent researchers who have become good role models

in my research. Jure, you always cut to the quick in our discussions of problems and motivations, and suggest high-impact directions with high standards. Srijan, I could not have progressed so fast without your patient and hands-on help regarding problem formulation, model implementation, paper writing, and many other aspects. The experience at Stanford makes me grow up rapidly and brings me fruitful achievements.

I would also like to thank the rest of my dissertation committee – Prof. Soroush Vosoughi and Prof. Bo Zhu for providing precious feedback on my thesis and serving on my committee. Thanks also to Prof. Lorenzo Torresani for providing valuable suggestions during my thesis proposal, and being my instructor of the deep learning class, which built solid foundation for my research.

I feel very fortunate and excited to have wonderful collaborators along the way of my research adventure: Maksim Bolonkin, Haipeng Chen, Pan Li, Dongkai Chen, Yanbang Wang, Lezi Wang, and Viney Regunath. Maksim, it's been a pleasure working with you, our constructive and critical discussions help me a lot in all the SCAN projects. Thank you Haipeng and Pan for providing the patient guidance for our collaborated papers as well as for showing me how to be diligent and self-motivated as a researcher. You are my close friends and models.

I gratefully acknowledge the assistance and accompaniment from the rest of my labmates and classmates: Qian Han, Rui Liu, Tommy White, Dongkai Chen, Chao Chen, Yanhai Xiong, and Bowen Dai. It is your encouragement, valuable research ideas and our extracurricular activities (e.g. hotpot and snowboarding!) that offer me the motivation and happiness along the four-year journey.

My completion of Ph.D. would not have been possible without the nurturing of collegues during my internships in industry: Xiaoxue Zang, Ying Xu, Abhinav Rastogi, Srinivas Sunkara, and Jingdong Chen from Google research; Fangyang Shen and Yu Lei from OppenFuture Technologies; Yang Liu, Xiaoming Fu and Xin Tong

from Microsoft Research Asia. Thank you all for providing unwavering support for me, in particular, for offering me the opportunity to approaching industrial-level research and real-life applications of my work!

Last but not the least, I give my sincere thanks to my family for their constant love, encouragement and faith in me. You are my original power and make my dream come true. The thesis is my sincere gift to you!

Thank you all again!

Contents

List of Tables

List of Figures

Chapter 1

Introduction

This chapter describes the motivation for identifying group human behavior followed by the formal statement of the problems we aim to solve. Then it provides the organization of the following chapters and lists the main contributions in the thesis.

Section 1.1 Motivation

Human group behavior is exhibited when people interact with each other, desiring to conform to the group, to be liked by others, and to gain more information about members of the group. For example, a manager often behaves dominantly in meetings, a student can be nervous when others talk about him/her in class, a politician wants to be persuasive in an election. Social scientists conduct studies to discover the factors (e.g. facial and vocal clues) that are highly related to such group behaviors [49, 65, 45]. Moreover, with a large amount of videos such as meetings, discussions, public speech, family activities recorded and available online, companies and governments expect automatic and accurate prediction and analysis of group behavior, with the help of Machine Learning (ML) technologies in audio, video and language understanding.

Automated identification of human group behavior is beneficial in many ways.

Companies would like to have persuasive salespeople and advocates to gain profits on their behalf. Robots are expected to monitor nervous response to adjust the ways to approach humans. In a diplomatic negotiation, analyzing the dominant person on the opposite side is helpful for strategy making. Safety agents want to detect lies to identify suspicious people and prevent harmful events. Eventually, an ideal scheme is to build an artificial intelligence to detect the existence of all kinds of behavior when the people of interest interact with surroundings.

Although important, identifying such group behaviors is a challenging task. First, datasets with rich multimodal communication signals and large-scale task-specific annotations are needed to train the ML models. Most existing datasets are either uni-modal ones (e.g. Image dominance data [79], text persuasion data [157]), or support only a few tasks [84]. Conversely, the ELEA multimodal dataset [124] has a wide range of annotations, but it only contains 102 subjects and simple group interactions. Second, domain-specific knowledge (e.g. facial cues of nervousness) is helpful for the model design. Without such knowledge, it is impossible to train models via features extracted from limited annotated data. Third, computational ways to model group interactions and extract multimodal features are essential for model performance. For example, accurate and compact representation of facial emotions is a key input to the model. Last, due to the heterogeneity of signals in videos, proper multimodal fusion strategies are the key to align these audio, visual, linguistic and communication signals together for better prediction.

Plenty of research has been conducted on human behavior prediction. Although much work has focused on individual features such as head movements [58], speaking turns [75], voice pitch [69], few efforts have paid attention to modeling the (non-)verbal communication among a group or considered the mutual influence of individuals when making predictions. For instance, a dominant person can gain others' attentions when

speaking [50], which can only be captured with group interactions. Besides, human behavior is a complex result of events that happened in a period of time, but most approaches (e.g.[3, 25]) simply computed the average or other simple statistics of each feature over time, ignoring the distribution and the ordering of the features. As an example, a nervous man may change gaze rapidly [89], taking average cannot account for such dynamics. Moreover, these approaches combine different modalities naively by either concatenating the features or averaging the predictions made by single-modal models trained separately. By doing so, the importance and inter-dependency of modalities are not taken into account.

This dissertation is motivated to address these challenges and improve the previous methods. To sum up, we propose an algorithm to extract non-verbal communications from videos, and study three kinds of group behavior from individual cues and *group influence*. We also propose a multimodal fusion framework applicable to human-centered videos and collect a new debate video dataset to demonstrate the framework.

Section 1.2

Problem statement

Given a video where a group of people interact with each other, the ultimate problem is to predict different kinds of behavior of each person. Figure 1.1 shows all the problems we aim to solve along the way.

First, comprehensive signals need to be acquired automatically. The audio and visual signals are usually easier to acquire than the text. In this work, we rather focus on the audiovisual signals and non-verbal interaction, including the text transcripts if they are available. More importantly, we emphasize the problem of extracting non-verbal interaction signals such as looking at and speaking to.

The second problem is how to design and extract features from the signals. The

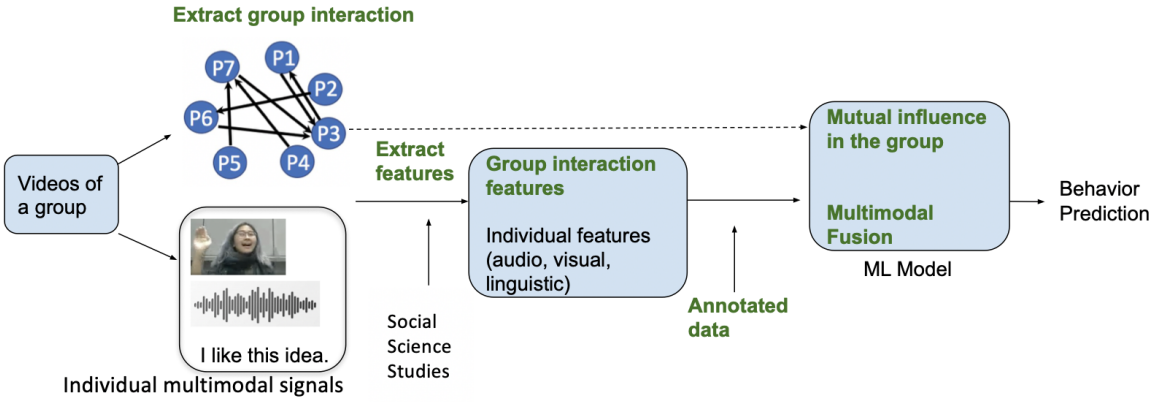


Figure 1.1: Roadmap for group behavior prediction. The green texts highlight the key problems to solve.

factors uncovered by social science studies are considered and converted to mathematical representations for each person with the help of pre-trained neural networks ([112, 93]) and classical signal processing methods ([41, 123]). In particular, group interactions can be modeled as a network, where the nodes and edges are people and their interactions. As such, node features incorporating the network effect (e.g. [34, 88]) can be extracted. To capture the activity dynamics, temporal aggregation of the features is also significant.

The third problem is to collect video datasets which cover dynamic human interaction activities and annotate them with various labels of interest. During collection, the alignments between modalities and different people is advantageous to the training process.

Last but not the least, given the extracted multimodal features and the labels, we build machine learning models to (i) consider the mutual influence among people when making the predictions for each person, and (ii) fuse the modalities with different importance and better representation. These techniques are missing in most of the past efforts.

Table 1.1: Outline and contribution of the thesis.

Chapters	Data	Group interaction extraction	Group mutual influence	Domain feature & representation	Multimodal fusion
4: Visual attention focus pred.	✓	✓	✓		
5: Dominance pred.			✓	✓	
6: Nervousness pred.			✓	✓	
7: Persuasion pred.	✓			✓	✓

Section 1.3

Overview and contributions

Overall contribution To solve the problems stated in the previous section, this thesis comes up with a suite of four methods and algorithms ranging from group interaction extraction to multimodal behavior prediction. We collect a video dataset, Qipashuo, to fill a lack of multimodal dynamic datasets with annotated human behaviors. We propose a scalable, lightly supervised algorithm to extract face-to-face interaction networks from videos as the foundation to model group effects. We then develop models which make accurate predictions for the behaviors of dominance, nervousness and persuasion by considering the multimodal individual cues and group interactions. We also study the significant modality cues towards sdifferent behaviors. Our contribution is outlined according to chapters in Table 1.1.

Chapter 3 introduces all the major datasets used in this thesis, including the Qipashuo dataset [14] we collected, and other datasets (Resistance[16], ELEA [103], IQ2US [14]) to evaluate our models.

In Chapter 4, we propose an iterative collective classifier to predict the visual focus of attention in group interaction videos. We also extract the who-look-at-whom dynamic interaction network in Resistance data using the proposed light-supervised model [16]. The network is the basis of richer interactions in Chapters 5,6, and employed in other authored publications [150, 87].

Chapter 5, 6 explore the identifications of dominance and nervousness in group interaction videos respectively. For dominance prediction, we propose a novel class of dominance rank features based on group interaction and social science studies, together with one multimodal system and one group prediction algorithm, to incorporate the visual and audio modalities of each person. For nervousness prediction, a hybrid algorithm is developed, combing the trained individual facial emotional representation and the proposed class of nervousness score features based on group interaction. Inspired from social science theory , the nervousness scores take the audiovisual emotions and the dominance behavior into account. The methods are evaluated on both Resistance and ELEA data, outperforming six baselines in dominance prediction and seven in nervousness prediction.

Chapter 7 designs an adaptive multimodal fusion framework to learn proper representation for persuasion prediction. The framework aims to align the heterogeneous modality inputs into a common space while learning modality-wise importance. It outperforms three baselines on two persuasion tasks on the Qipashuo and IQ2US datasets.

In addition to making accurate predictions, we further conduct studies of specific modalities, features, and raw inputs that contribute to each kind of group behavior. Our findings are highlighted below:

(a) Dominant people tend to draw more attention when speaking, although they may not speak a lot.

(b) Visual features are more important than audio features for nervousness prediction, and nervousness of a subject is largely influenced by dominance of the people interacting with him or her.

(c) For persuasion prediction, linguistic modality is the most significant, followed by visual then audio modality.

Below we summarize each of the chapters 4-7 briefly.

1.3.1. Visual focus of attention prediction in videos (Chapter 4)

The task is to predict the visual focus of attention of each person (i.e. who the person looks at) at each timestamp in a group-interaction videos. Visual focus of attention in multi-person discussions is a crucial nonverbal indicator in tasks such as inter-personal relation inference, speech transcription, and deception detection. However, predicting the focus of attention remains a challenge because the focus changes rapidly, the discussions are highly dynamic, and the people's behaviors are inter-dependent. Moreover, the tedious training data annotation is not scalable, making the performance drop for unseen videos and people.

To resolve these, we propose ICAF (Iterative Collective Attention Focus), a collective classification model to jointly learn the visual focus of attention of all people. Every person is modeled using a separate classifier. ICAF models the people collectively—the predictions of all other people's classifiers are used as inputs to each person's classifier. This explicitly incorporates inter-dependencies between all people's behaviors. We evaluate ICAF on a subset of 5 videos (35 people, 109 minutes, 7604 labels in all) on the Resistance data and a widely-studied meeting dataset with supervised prediction. ICAF outperforms the strongest baseline by 1%–5% accuracy in two datasets. We further propose a light supervised ICAF to create who-look-at-whom, who-listen-to-whom, and who-speak-to-whom networks from unseen group

interaction videos.

A demo[1] screenshot of our method is shown in Figure 1.2. There are seven subjects in the group, and our methods output the probability of who each person looks at at time t (bottom of the Figure). A network snapshot at t consists of the subjects as nodes and look-at predictions connecting them (upper right of the Figure). Overall, a dynamic who-look-at-who network is constructed from the video.

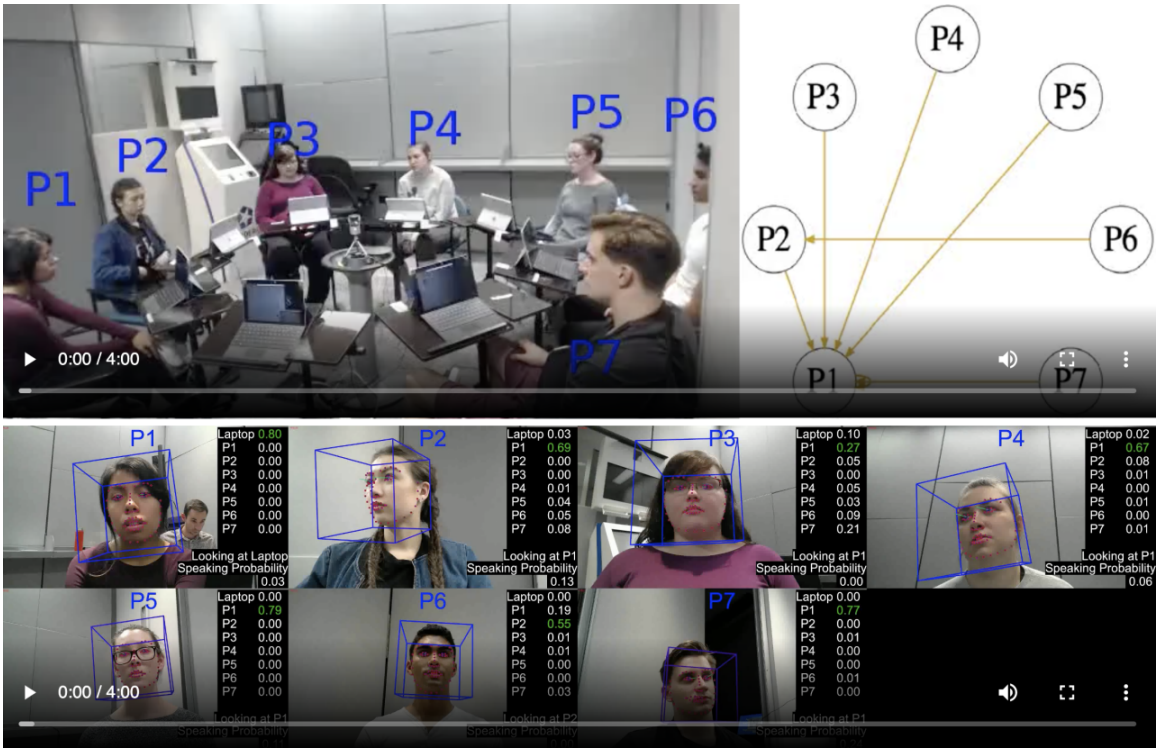


Figure 1.2: Demo videos showing the predicted probabilities of people looking at each other (bottom) and the dynamic social interaction networks built upon the predictions (upper right).

Overall, this work makes the following contributions:

- Accurate and scalable algorithm for visual focus of attention prediction. The supervised and light supervised algorithms achieve 0.62 and 0.55 accuracy respectively on a highly dynamic testing scenario, when predicting one out of 5-8 attention targets.

[1] `https://www.cs.dartmouth.edu/~cy/icaf/`

- Accurate visual speaking prediction model, providing sub-second speaking probabilities from facial movements of people.

- Face-to-face interaction network dataset. We extract the who-look-at who dynamic interaction networks on the Resistance dataset, leading to 62 time series networks, consisting of ∼3M edges spanning ∼0.1M seconds of videos. The dataset is made public[2].

Impact Based on the who-look-at-who network and speaking prediction, more complicated non-verbal communication are modeled. Specifically, chapter 5 and [87] create the speak-to and listen-to netowrks to study dominance and deception resp. Chapter 6 further annotate the interaction with the emotions one wants to convey, which contributes to nervousness prediction. The authored publication [150] builds a general neural model upon such dynamic networks to study more group human behaviors.

1.3.2. Dominance prediction in multi-person videos (Chapter 5)

Identifying dominant people in a group setting is desired for lots of applications. For example, businessmen in meetings may want to find the decision maker among the customer team to strive for a deal. In a negotiation, delegations may be interested in identifying the most dominant person from the other side. Despite being an important task, dominance behavior may be shown from multimodal personal cues (audio, visual) as well as the interactions among the group, e.g., talking to each other. It is challenging to extract relevant features and design an appropriate model considering the multimodal and group effects.

We consider the problems of predicting (i) the most dominant person in a group of people, and (ii) the more dominant of a pair of people, from videos depicting group

[2] `https://snap.stanford.edu/data/comm-f2f-Resistance.html`

interactions. Inspired by the dominance indicators discovered in social science studies such as looking-while-speaking [50] and visual dominance ratio [55], we introduce a novel family of features called Dominance Rank. Dominance ranks are the relative dominance of people in a group induced by various measurements of group interaction. For instance, how much more probable does person A look while listening to person B than the opposite? Figure 1.3 visualizes this measure during 5 seconds, where an edge exists when the amount of interaction from one person to another is larger than a threshold. Intuitively, since people tend to look and listen to P2 and P3 more, P2 and P3 might be more dominance than others.

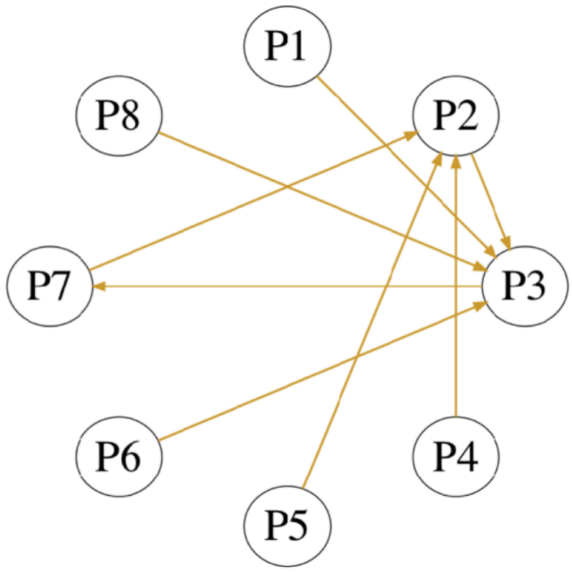


Figure 1.3: Visualization of the interaction network weighted by the look-while-listening ratio difference.

We also introduce features (e.g., facial action units, emotions) that are discovered influential factors [45] but not previously used for dominance prediction. We develop Dominance Ensemble Late Fusion (DELF) to combine multimodal features. For the MDP problem, we further propose the Group Dominance Prediction (GDP) algorithm, which augment our training data by over 120 times and make the prediction from the group instead of a single person. We test our two models against four competing

algorithms in the literature on the Resistance and ELEA datasets. We show 2.4% to 16.7% improvement in AUC compared to baselines on one dataset, and a gain of 0.6% to 8.8% in accuracy on the other.

To summarize, the following contributions are made:

- The Dominance Rank features capturing the mutual interaction of people
- The DELF model to make the prediction from multimodal modalities
- The GDP algorithm to augment the training data and boost the MDP problem performance.

1.3.3. Nervousness prediction in multi-person videos (Chapter 6)

Detecting nervous people in a group has many applications. Understanding that one person is nervous in a social activity may enable others to put that person at greater ease. The security department might identify suspicious subjects through the nervousness clue.

On the one hand, as social science theory suggests that A might be more nervous than B if B is more dominant or B conveys passive evaluation to A ([48, 104]), we define a new class of 54 features called nervousness scores (NSs) from the audio-visual channel to capture the external influence from the group. NSs use dominance relationships between people, as well as gaze (who is looking at who), and speaker (who is speaking) information. In total, 3 interaction types and 9 forms of dominance influence are defined, and 6 (resp. 4) kinds of visual (resp. audio) emotions are considered. Intuitively, the nervousness score of a person is a summation of the evaluations (measured by emotions and dominance) he received from the people interacting with him,

On the other hand, as facial behavior is vital information for nervousness, we develop a Facial Emotion Graph Convolution Network (FE-GCN) to learn facial em-

beddings from images, which uses GCN to capture the facial landmark dynamics together with CNN to capture the appearance. The temporal sequence of embeddings are then aggregated by a Temporal Convolution Network (TCN).

We solve two kinds of tasks to predict relative nervousness: Who is more nervous given a pair of people? Is a person more nervous compared to before? Our results show that: (i) either NSs or FE-GCN generate the best performance in head to head comparisons with seven baselines based on past work, (ii) an ensemble that merges NSs and FE-GCN provides high quality results in terms of both F1-score and AUC compared to the baselines, and (iii) the learned FE-GCN identifies landmarks that are highly relevant for nervousness prediction.

Below summarizes our contributions:

- 54 interpretable audio-visual nervousness score features that consider the human interactions annotated by emotions and dominance.
- The FE-GCN model to learn facial embeddings for nervousness prediction.
- An ensemble model combing above outperforms seven baselines on four tasks on the Resistance and ELEA data.
- Comprehensive experiment analysis of the important signals for nervousness prediction, such as specific landmarks, positive and negative audio-visual emotions.

1.3.4. Adaptive multimodal fusion for persuasion prediction (Chapter 7)

Identifying persuasive speakers in an adversarial environment is a critical task. In a national election, politicians would like to have persuasive speakers campaign on their behalf. When a company faces adverse publicity, they would like to engage persuasive advocates for their position in the presence of adversaries who are critical of them.

The persuasiveness depends on the combination of how a person expresses themselves visually and vocally, as well as what the person says. Besides, the temporal dynamics also plays a key role, e.g., change of speech speed and vocal pitch.

Debates represent a common platform for these forms of adversarial persuasion. This paper solves two problems: the Debate Outcome Prediction (DOP) problem predicts who wins a debate while the Intensity of Persuasion Prediction (IPP) problem predicts the change in the number of votes before and after a speaker speaks. Though DOP has been previously studied, we are the first to study IPP. Past studies on DOP fail to leverage two important aspects of multimodal data: 1) multiple modalities are often semantically aligned, and 2) different modalities may provide diverse information for prediction.

Our M2P2 (Multimodal Persuasion Prediction) framework is the first to use multimodal (acoustic, visual, language) data to solve the IPP problem. To leverage the alignment of different modalities while maintaining the diversity of the cues they provide, M2P2 devises a novel adaptive fusion learning framework which fuses embeddings obtained from two modules – an *alignment* module that extracts shared information between modalities and a *heterogeneity* module that learns the weights of different modalities with guidance from three separately trained unimodal reference models.

We test M2P2 on the popular IQ2US dataset designed for DOP. M2P2 significantly outperforms three recent baselines by at least 25% Mean Squared Error (MSE) in IPP and 3% accuracy in DOP on two datasets. The model is able to weight the three modalities and pays attention to the relevant inputs over time.

Figure 1.4 shows a sample of how our M2P2 framework predicts speaker persuasiveness at interim points during a debate from the QPS dataset — the reader can readily see that the M2P2 prediction of number of votes (orange line) closely matches

the ground truth (green line).

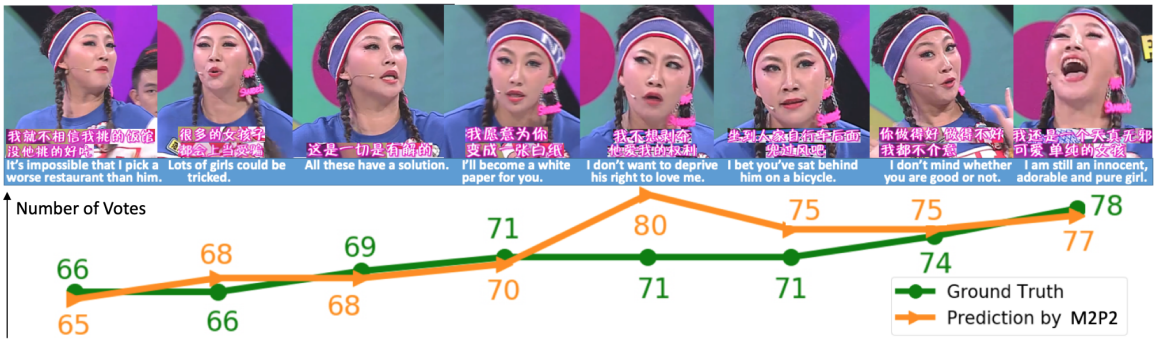


Figure 1.4: Real-time prediction of debate persuasiveness (number of votes) using M2P2. M2P2 closely predicts the ground truth number of votes.

To summarize, our key contributions are:

- Proposal of the fine-grained IPP problem.
- A novel adaptive fusion learning framework to solve the IPP and DOP problems which is applicable to other multimodal learning tasks. It outperforms three recent baselines on both tasks.
- A new persuasion dataset Qipashuo from the well-known Chinese debate TV show Qipashuo.
- Model explanability: weight the modality importance and identify relevant raw inputs (faces, words).

Chapter 2

Background

Analysis and prediction of human group behavior is a significant field involving a variety of topics. This chapter introduces the social science studies characterizing different human group behaviors (Section 2.1), features calculated from multiple signals (Section 2.2), efforts to model mutual influence among a group (Section 2.3), models to aggregate short-time features over time (Section 2.4), and methods to make predictions from multimodal signals (Section 2.5).

Section 2.1
Social science studies of human group behaviors

Social scientists take data-driven methods to study human group behaviors. Normally, they design the experiments in which volunteers complete specific tasks and report self or cross evaluations, monitor the hypothesized factors during the tasks, and finally link the factors with the behaviors reported in evaluations.

Dominance Dominance is a strategy of social influence and power exhibited as verbal or non-verbal communication among groups [57]. Dominance is motivated by the intention to control or change the behavior of other group members [40].

Dillard et al. [49] found that dominance is correlated with speaking rates and voice characteristics (frequency, amplitude, pauses, pitch, etc.). Visual cues like people looking at each other, body movements, gestures and facial expressions are also the indicators of dominance in social interactions ([65, 55]). Additionally, dominance can also be expressed in the use of personal space and the artifacts within the space [134]. Moreover, Dovidio et al. [50] studied the relationship between dominance and the combination of looking-while-speaking and looking-while-listening behaviors. For example, the Visual Dominance Ratio was defined as the ratio between the total looking-while-speaking periods to the total looking-while-listening periods for dyadic interactions.

Nervousness Nervousness and social anxiety are closely related to each other. According to the Mayo Clinic [38], "In contrast to everyday nervousness, social anxiety disorder includes fear, anxiety and avoidance that interfere with daily routine, work, school or other activities." In general, nervousness is viewed as a short term form of anxiety. In fact, the terms nervousness, anxiety (and stress to a certain degree) are often used interchangeably in social science studies ([67, 42, 39]). Messenger et al. [104] notes that nervousness is correlated with expectations of negative social evaluations. [137]'s analysis of social interactions in a community suggests that nervousness is linked to public speaking, talking to strangers or being the center of attention – they argue that "social fears" are linked to nervousness. This phenomenon is not limited to Western cultures — Caballo et al. [31] suggest that communities in Latin America exhibit similar behaviors. In group settings, nervousness is heightened by group dynamics. Morrison et al. [105] suggest that negativity by others plays an important role in anxiety. This observation is in line with [31] which notes nervousness is most easily identified when dealing with strangers and criticism. Dijk et al. [48] suggest that anxious individuals are less dominant in group interactions. Morrison et

al. [105] suggest that individuals, once anxious, will not respond to positive stimuli and maintain their state until the end of the interactions. Maner et al. [99] also notes this phenomenon. Moreover, nervousness is linked with facial cues, e.g. nervous individuals often avert their gaze and change their posture more frequently [45]. Using the Hamilton Anxiety Scale, Yazici et al. [153] state that nervous individuals exhibit actions known as "adaptors", e.g. rubbing one's forehead or tapping a pencil. Signs of nervousness can also be identified by facial cues [114]. Cues such as head movements, blink rates and gaze directions are all important considerations taken by previous computer science papers interested in anxiety detection. DePaulo et al. [45] state that nervous individuals raise the pitch of their voice and speak with more hesitation (e.g. "ums" and "ahs") and speech errors.

Persuasion Persuasion is a process to change the attitude or behavior of a person or a group toward some ideas or objects, by using written, spoken words or visual effects to convey information [117]. In the studies of Johnson and Blair [78], the order of messages, the comprehensibility of the content, and the validity as well as the number of arguments presented are all related to persuasion. When it comes to voice influence, paralinguistic features (e.g. pitch, volume) are important because they are predictive social markers [90] and could influence the persuasiveness [27]. Towards the visual persuasion, the related factors are facial emotions, postures, attractiveness, etc. ([122, 79]). Clearly, persuasion is not barely in texts, but a mixture of texts, audio and visual effects.

Section 2.2

Feature extraction from multiple modalities

Based on the findings in Section 2.1, both high-level features and low-level representation have been extracted from audio, visual and linguistic signals.

Dominance Some earlier predictive models use discrete features based on binary speaking variables (during a given time segment, does the person speak or not). These features include statistics such as total speaking length, total speaking turns, and total successful interruptions ([75, 3]). In addition, Sanchez-Cortes et al. [125] use prosodic features such as energy and pitch variation. Other works extensively use visual features in the form of discrete variables. Aran et al. [3] and Jayagopi et al. [75] use statistics on the overall visual activity (binary variable - person either moves or not). Sanchez-Cortes et al. [125] and Beyan et al. [22] analyze more fine-grained activity such as head and body movements, and gestures. In addition to these, a set of proposed methods uses gaze-related features such as looking at the target or at a speaker ([4, 108, 109]).

Nervousness We summarize the computational efforts for nervousness as well as anxiety and depression which are highly related to nervousness. Pediaditis et al. [114] and Caballo et al. [63] look at facial cues of anxiety, stress, and nervousness of individuals — not group interaction videos as we do. They extract the movements of head, gaze, mouth, and pupils through estimated facial landmarks. They also estimate heart rate by assessing the frequency of the facial colors' signal. Florea et al. [61] leveraged findings learned from large facial expression datasets to a small annotated anxiety dataset, showing a a significant boost in anxiety detection in images. To predict depression in interview videos, Ray et al. [120] extracted the facial action

units, head pose and eye gaze angles from video frames. Additionally, mel-frequency cepstral coefficients (MFCCs) [43] features and the pre-trained universal sentence encoder [33] embeddings are also computed from audio and transcripts respectively.

Persuasion Extensive amounts of work have studied how to predict persuasion from the text modality. [157, 119, 148, 64] explore the linguistic modality by studying style, context, semantic features and argument-level dynamics in English transcripts. The LIWC features [115] are used by [25] to count psychological and structural word categories. For the visual modality, Joo et al. [79] define nine visual intents related to persuasion (e.g. dominance, trustworthiness) and train SVMs to predict them and persuasion using hand-crafted features. Huang et al. [71] improve these results by fine-tuning pre-trained CNNs to learn suitable face & body representation. Brilman et al. [25] extract facial emotions to predict the debate outcomes. In the case of audio, many ([126, 107, 127]) use MFCC features, and Nojavanasghari et al. [107] also employ voice quality (e.g. Normalized amplitude quotient (NAQ)) and pitch features.

Section 2.3

Dyadic and group influence modeling

While most work extracts comprehensive features from individuals, the dyadic interactions and group-level influence are essential and can be incorporated by both feature engineering and machine learning algorithms.

Feature engineering Aran et al. [4] and Okada et al. [109] mine co-occurrent events in the sequence of visual and audio features of individual players to predict impression (e.g. dominance, likeness) and personality traits (e.g. openness) in a group setting. For example, more than two people move bodies or look at a speaker, two

people look at each other. Bai et al. [12] show that the ranks of each feature among a group is effective for deception prediction. Kumar et al. [87] define the reciprocity of looking and speaking to measure the mutual engagements.

Learning Algorithms Sanchez-Cortes et al. [125] build a graph with the people in a group as nodes and the speaking turns as weighted edges, and employ collective classification, which is helpful to predict the emergent leadership. Kumar et al. [87] build the network weighted by avoidance of gazing, and employ the belief propagation to predict deception. To get rid of depending on domain-specific knowledge to extract features, Wang et al. [150] build a general neural model on dynamic face-to-face interaction networks, and demonstrate its efficacy on multiple tasks such as prediction of dominance, deception, nervousness.

Section 2.4

Temporal aggregation of short-time features

Since the input audiovisual sequences are usually 5-20 minutes long, temporal aggregation is essential to make sequence-level predictions. This section reviews the methods taken by the previous behavior prediction papers, and then introduces three effective techniques employed in our work.

Past work The first category [75, 3, 4, 109, 25, 157] accumulates over time the binary indicators such as speak turns, interruptions, body movements, which are then normalized with the sequence length. However, counting can be noisy since the binary indicators are usually obtained from probability estimations (e.g. speaking probability). The second category computes extensive statistics such as mean, variance, min, max, and percentiles ([125, 120, 25, 107]) of continuous valued features (e.g. emotion prediction probability, prosody energy). To obtain fine-grained descriptions, the

third category ([151, 12]) employs Fisher vector encoding [118] which computes the bag-of-words representation of the sequence of features.

Specifically, given a sequence of feature vectors $\{\boldsymbol{f}_1, \ldots, \boldsymbol{f}_T\}$, $(\boldsymbol{f}_i \in \mathbb{R}^D, \forall i)$, the Fisher vector encoder [118] first builds a Gaussian Mixture Model (GMM) with mean $\boldsymbol{\mu}_i$, diagonal covariance $\boldsymbol{\sigma}_i$ and mixture weights w_i for the i^{th} component $(1 \leq i \leq N)$. Then, the Fisher vector component i is computed as

$$\mathcal{G}_{\boldsymbol{\mu}_i} = \frac{1}{T\sqrt{w_i}} \sum_{t=1}^{T} \gamma_t(i) \left(\frac{\boldsymbol{f}_t - \boldsymbol{\mu}_i}{\boldsymbol{\sigma}_i} \right) \tag{2.1}$$

$$\mathcal{G}_{\boldsymbol{\sigma}_i} = \frac{1}{T\sqrt{2w_i}} \sum_{t=1}^{T} \gamma_t(i) \left(\frac{(\boldsymbol{f}_t - \boldsymbol{\mu}_i)^2}{\boldsymbol{\sigma}_i^2} - 1 \right), \tag{2.2}$$

where $\gamma_t(i)$ is the posterior probability. Finally, the $2DK$ dimensional Fisher vector is the concatenation of all $\mathcal{G}_{\boldsymbol{\mu}_i}, \mathcal{G}_{\boldsymbol{\sigma}_i}$.

Histogram representation To enrich the statistical description of the temporal sequence, we compute the histogram vector of each individual feature. Suppose we take B bins, for dimension d, the histogram $\boldsymbol{h}_d \in \mathbb{R}^B$ describes the distribution of $\{\boldsymbol{f}_{1,d}, \ldots, \boldsymbol{f}_{T,d}\}$. The final histogram representation is the concatenation of $(\boldsymbol{h}_1, \ldots, \boldsymbol{h}_D)$. This is suitable for features with bounded, dense values (e.g. probabilities), and has been successfully applied in Chapters 5, 6.

Although the above methods can summarize the sequence of features, they ignore the temporal order and dynamics of features which neural models such as RNN and LSTM can do. However, since the sequences are too long, even LSTM cannot handle the long-term dependency. Below describes two alternative neural models we take.

Temporal Convolution Network (TCN) TCN is a successful adaptation of CNN to temporal data modeling [91]. It consists of several layers of 1D convolutions

+ max poolings. At layer l, denote $F^{(l)} = [\boldsymbol{f}_1^{(l)}, \ldots, \boldsymbol{f}_{T_l}^{(l)}] \in \mathbb{R}^{D_l \times T_l}$ as the feature representation, where $\boldsymbol{f}_t^{(l)}$ is the feature vector at time t. Note that $F^{(1)}$ is the input sequential features. Given the D_{l+1} one dimensional convolution filters $W_l = \{W_l^{(i)}\}_{i=1}^{D_{l+1}}$, each filter $W_l^{(i)} \in \mathbb{R}^{K_l \times T_l}$ with duration K_l, the representation of layer $l+1$, $F^{(l+1)} \in \mathbb{R}^{D_{l+1} \times T_{l+1}}$, is

$$F^{(l+1)} = \text{MaxPool}(g(W_l \circledast F^{(l)} + \boldsymbol{b}_l)), \tag{2.3}$$

where $\boldsymbol{b}_l$ is the bias, g is a non-linear activation function (e.g. ReLU). Note that the larger K_l is, the faster T_{l+1} decreases. Given the local temporal convolution filters, TCN can learn dynamic local interaction patterns with various durations (K_l) over the long-time span. For long sequences, the global summarization (i.e. receptive field covering the whole sequence) is achieved by stacking several layers with pooling and larger kernel durations. Moreover, TCN does not suffer from the computational dependency and memory issues of RNN or LSTM, since all computations are parallel and local. These make TCN a good fit for capturing the dyadic interactions over long videos. Chapter 6 and our work [150] have applied TCN successfully in group behavior prediction.

Transformer Transformer [144] is a multi-head self-attention model which can capture long-time dependency to be learned in temporal data and can be computed efficiently. We will quickly shed light on the key concept, the scaled dot-product attention, of the Transformer encoder used in Chapter 7 and leave the details in [144].

Assume we have a matrix (*value* matrix) $V = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_T]$ denoting a sequential feature vector (e.g., $V = F^{(1)}$ or its projection), the goal is to output at each timestamp $t \in \{1, \ldots, T\}$ the weighted average over V, where the weights are obtained

from the attention of query $\boldsymbol{q}_t$ towards the keys of all timestamps $\boldsymbol{k}_m, \forall 1 \leq m \leq T$, respectively. $Q = [\boldsymbol{q}_1, \ldots, \boldsymbol{q}_T]$ and $K = [\boldsymbol{k}_1, \ldots, \boldsymbol{k}_T]$ are called *query* and *key* matrices. Formally, the scaled dot-product attention can be defined as

$$\text{Attention}(Q, K, V) = \text{Softmax}(\frac{QK^T}{\sqrt{d_k}})V, \tag{2.4}$$

where d_k is the dimension of keys. The Softmax function serves as a normalization of weights over timestamps. Q, K, V are all linear projections of the encoder input. The Transformer encoder also captures local dynamics and can be computed efficiently. Yet different from TCN, it directly computes the long-time dependencies over all pairs of timestamps. In other words, a single attention layer has a global receptive field, and the model attends to the corresponding timestamps from the training process. In Chapter 7, we apply the Transformer encoder to one each of audio, visual and linguistic temporal input of debate videos to aggregate the temporal information.

Section 2.5

Multimodal fusion and prediction

2.5.1. Previous methods for group behavior prediction

On the one hand, early fusion concatenates the extracted features from all modalities to train a single ML model ([108, 109, 127]). In particular, Santos et al. [127] show that the feature-level fusion performs better than inference-level fusion for persuasion prediction. Despite the model simplicity, early fusion fails to align the heterogeneous data (e.g. text and audio), thus some modalities may not be fully exploited.

On the other hand, more approaches [75, 3, 4, 25, 107] use late fusion to make the final inference from inferences made by each model from each modality separately. The fusion strategies include averaging [107], voting [25], rank-based decision [125]

and so on.

We adopt an adaptive late fusion method in Chapters 5, 6. Formally, suppose we have N sources of prediction scores (e.g. classification probability) $S_1, \ldots, S_N$, including all modalities and prediction models, then the final score S is obtained by late fusion of all predictions:

$$S = \sum_{i=1}^{N} \alpha_i S_i$$

where $\sum_{i=1}^{N} \alpha_i = 1$. Values of the non-negative late fusion weights α_i are obtained by grid-search and cross-validation. Specifically, in each fold, we train N models in the training set, obtain the scores $S_i, i = 1, \ldots, N$ and search the optimal α_is in the validation set, finally apply the optimal α_is in the test set to get the final prediction S. The overall performance is averaged from test sets of all folds.

The late fusion methods can jointly make a decision from multiple source of inferences, whereas it disables the potential cooperation between modalities before predictions.

A few other methods take more advanced fusion approach. Beyan et al. [22] employ the multiple kernel learning for dominance prediction. It adapts the heterogeneous modalities by finding specific kernels, and learns the weights to average the kernel inferences together. However, the kernels are selected from pre-defined families which may limit the learning potential. Ray et al. [120] train a multi-level attentional fusion model from end to end for depression prediction, which combines the modalities at different levels by attention weights. Although powerful, the attentional model has many learnable parameters, thus requires lots of labeled training data.

2.5.2. General multimodal learning methods

In this section, we discuss other methods that better utilize the relationship between multiple modalities.

A body of multimodal learning methods defines the constraints between modalities in a latent space to capture their inter-relationships. Andrew et al. [2] extend Canonical Correlation Analysis by deep neural networks to maximize inter-modal correlations. Such correlation constraints have since been used in sentiment classification [54], emotion recognition [1] and semantic-visual embedding [59]. In addition to capturing the shared relationship, [111, 130, 146] try to extract the individual components of each modality through low-rank estimation. [77, 54] train auto-encoders to reconstruct a modality from itself and another modality. While these efforts provide important insights for creating multimodal embeddings, they do not show how to combine the learned embeddings for accurate prediction.

Another body of work explores architectures for fusing embeddings from modalities. Zadeh et al. [154] introduced bimodal and trimodal tensors via cross products to express inter-modal features. As cross products significantly increase the dimensionality of the feature space, [80, 20, 32] introduced bilinear pooling techniques to learn compact representations. Although these methods explicitly model inter-modal relationships, they introduce additional features that require larger networks to be learned for subsequent prediction tasks. In contrast, attention-based fusion [97, 70] learns the weighted sum of multimodal embeddings taking the prediction task into account. However, they require huge amounts of data to learn the optimal attention weights.

The third body is self-supervised learning, which utilizes the natural correspondence between modalities (e.g. guitar sound and guitar playing video) to pretrain a powerful model in large-scale unlabeled data, and finetunes the model on much smaller labeled datasets for specific tasks. Inspired by the huge success of BERT [46] in the NLP field, the multimodal BERT methods [141, 98, 92] design two kinds of pretraining tasks: (i) alignment prediction – predict whether the multimodal inputs are from the

same instance, and (ii) denoising autoencoder – mask a small portion of one modality input, and predict it from the rest of inputs. Note that these pretraining tasks don't require any human-labeled data. The models' backend is mostly the Transformer encoder [144] which fully exploits the inter-moal and inter-modal dependency. Another direction is the Multimodal Generative Pretrained Transformer (GPT)-based methods [95, 110], which is extended from the self-supervised autoregressive language generation model GPT-3 [26]. For group human behavior prediction, once we collect the large amount of related data (e.g. metting, interview, publish speech), the self-supervised learners can play a big role.

Our contributions We propose a supervised adaptive fusion framework and demonstrate its efficacy in persuasion prediction. On one hand, inspired by the first body of work, M2P2 encodes the primary embeddings to a shared space and enforces high correlation among the encoded embeddings. On the other hand, M2P2 computes a weighted concatenation of latent unimodal embeddings, where the weights are guided by the persuasiveness loss of each embedding through interactive training. These two innovations lead to a compact embedding that can be learned with a small dataset.

Chapter 3

Datasets

In this chapter, we introduce four video datasets on which we study group human behaviors: Resistance, ELEA, Qipashuo (new) and IQ2US.

Section 3.1

The Resistance dataset

The Resistance dataset [56] contains a set of videos depicting groups of 5-8 people playing a Mafia-style social game, called the Resistance[1]. Figure 1.2 depicts two views of videos captured. The upper left shows the overhead view, where all people are seated around a table, with a laptop placed in front of each of them. The laptop is used to capture their close-up front view videos, and to record their answers to the survey along time. The bottom of the figure shows the concatenation of each close-up view. The players are encouraged to interact with each other during the game.

3.1.1. The game

The game process is illustrated in Figure 3.1, and the major survey questions are listed in Table 3.1, readers can refer to [56] for more details.

[1] https://en.wikipedia.org/wiki/Resistance_(video_game_series)

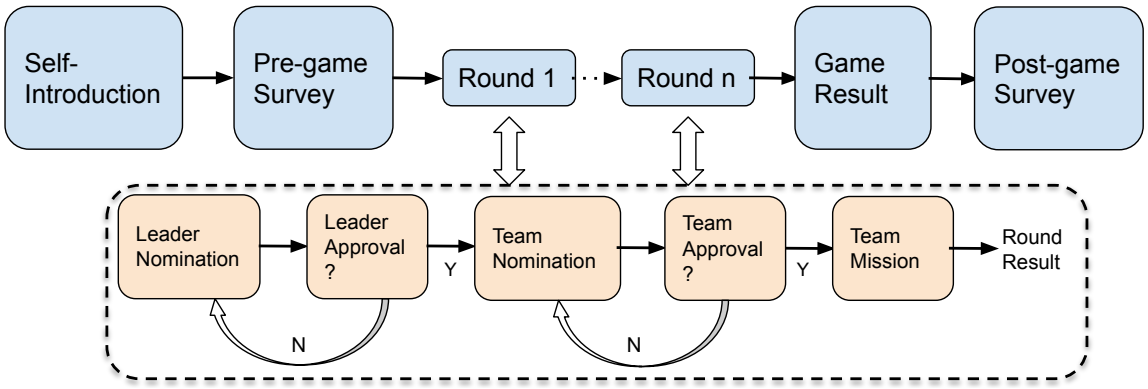


Figure 3.1: The process of Resistance game.

At the beginning, the players conduct an ice-breaking session, where each player makes a self-introduction, and one player is assigned asks a follow-up question to that player. Note that Chapter 4 proposes a automatic annotating method from this session to get rid of human labeling.

After that, the players answer several questions in the pre-game survey on their laptops in front. Each player is also secretly informed by the laptop that (s)he is either a "spy" or a member of the "resistance". Spies know who other spies are, but the resistance does not know any information. There are 2–3 spies in a game which proceeds in rounds (typically 3 to 8 in a game). The two teams form an adversarial setting.

Then the game proceeds with several rounds. Every round has three stages: leader nomination and election, nomination of team members by the leader, and team mission. In the leader nomination stage, players get nominated to serve as a leader. All players vote for or against the nominee. This stage is repeated up to three times until the team leader is elected. In the second stage of the round, the team leader nominates team members. Note that the team size varies according to the number of players and rounds to obtain the game fairness (Table 3.2). After a discussion, all players vote on approval or rejection of the proposed team. This stage

Table 3.1: Information collected in the Resistance game survey.

Survey Questions	When to answer
Please rate how much you trust each player from 1 to 5. 1 means least trustworthy, 5 means most trustworthy.	Pre-game, after even-numbered rounds
Please rate how dominant each player is from 1 to 5. 1 means least dominant, 5 means most dominant.	Pre-game, after even-numbered rounds
Please rate how nervous each player is from 1 to 5. 1 means least nervous, 5 means most nervous.	Pre-game, after even-numbered rounds
Please rate how likable each player is from 1 to 7. 1 means very unlikable, 7 means very likable.	Pre-game, after the last round
Please rate how much you think each player is a spy from 1 to 5. 1 means least suspicious, 5 means most suspicious.	After even-numbered rounds
Game role (spy or resistance). Secretely assigned to each player, and spies know who else are spies.	Pre-game
Approval (yes or no) of the nominated team leader.	Each round
Approval (yes or no) of the nominated team.	Each round
(For team members) Vote for mission failure or success.	Each round

is repeated up to three times or until the team is approved. In the first two stages, players first vote secretly and then publicly. The third stage consists of team members secretly voting for the success or failure of the mission. Note the spies want to fail the mission while hiding themselves in the resistance team. Again, the minimum number of votes to fail a mission depends on the number of players and round (Table 3.3). If the vote is in favor of the mission going forward, the resistance team collectively

Table 3.2: Team size for missions for each of the group size and rounds.

Number of players	R1	R2	R3	R4	R5	R6	R7	R8
6	3	3	4	4	4	5	5	5
7	3	3	4	4	5	4	4	4
8	3	3	4	4	5	5	5	5

Table 3.3: Number of fail votes required for mission failure for each of the group size and rounds.

Number of players	R1	R2	R3	R4	R5	R6	R7	R8
6	1	1	1	1	1	2	2	2
7	1	1	1	2	2	2	2	2
8	1	1	1	2	2	2	2	2

gets a point — if the vote goes the other way, the spies collectively get a point. Spies also score a point if players fail to elect a leader or approve the proposed team three times.

During each round, the voting for a team leader and a team are both conducted privately in the laptop and then publicly, while the mission voting only happens privately in the laptop by team members. The game facilitator announces the number of favoring voting and the number of non-favoring voting. After every two rounds and the final round, players conduct the survey on their laptops to answer the questions in Table 3.1.

Finally, a team (spies or resistance) with the highest score at the end of the game wins. Therefore spies have a natural incentive to get elected as team leaders and to get on mission teams. For the Resistance team, it is advantageous to identify spies as soon as possible and prevent them from getting on mission teams, which means spies need to make sure they are not discovered.

3.1.2. Dataset description

In the dataset, participants (N = 693; mean of ages = 22, standard deviation of ages = 3.75) were primarily college students, although some participants were recruited from the general public. Data collection took place at 8 public universities in the Southwestern US (9 games; n = 59), Western US (11 games; n = 67), Northeastern US (10 games; n = 74), Israel (10 games; n = 71), Singapore (12 games; n = 84), Fiji (14 games, n = 106), Hong Kong (15 games, n = 115), and Zambia (15 games, n = 117). Participants were recruited via email and advertisements on public message boards. The sample was 59% female and was ethnically diverse (although this varied by location), and the biggest groups were Asian (38%) and White (18%). They reported nationalities representing 41 different countries. Participants were required to be proficient English speakers.

To ensure fairness, the numbers of spies are 2, 2, 3, 3 in games of 5,6,7,8 players respectively. The mean and standard deviation are 3.28, 0.87 for dominance ratings and 2.93, 0.91 for nervousness ratings, respectively.

As such, there are N = 693 game videos in close-up views from 96 games, spanning from a minimum of 29 minutes to a maximum of 66 minutes with the average duration being 46 minutes. There are 2-8 rounds per game where each round is around 6 minutes on average. The information collected from the survey is used as labels of group human behavior (e.g. dominance, nervousness, game role), to train predictive models and conduct statistical analysis. Due to some data collection issues, the labels and videos available for different tasks can be different. We will describe the task-specific prediction problems, the number of videos and labels in Chapters 4,5,6 accordingly.

Section 3.2

The **ELEA** dataset

Figure 3.2: A screenshot of the ELEA data from [125].

The ELEA dataset is developed by Sanchez-Cortes et al. [125] to study emergent leadership as well as other human group behaviors. It is publicly available[2].

This dataset consists of videos of groups of people (3–4 persons in a group, 27 groups) participating in a winter survival task. Different from the Resistance game, it is a cooperative setting. The 3-4 participants sit at two sides of a square desk, and two cameras capture their close-up views as shown in Fig. 3.2. There are 102 participants and approximately 10 hours of videos.

The participants were given a list of 12 items and asked to rank their importance for survival in the hypothetical scenario of a plane crash in a winter forest. Participants needed to have a discussion and come up with a consensus. Each video lasts 15 minutes, and the discussion lasts 14.61 minutes, ranging form 8 to 19 minutes.

Videos are accompanied by survey results measuring participants' group behaviors in three ways:

- In-group scores where each participant scores each behavior of every participant

[2] `https://www.idiap.ch/dataset/elea`

from 1 to 5. The behaviors are leadership, dominance, competence, and likeness.

- External scores where each participant is rated by three independent observers not participating in the task from 1 to 5. This includes leadership and dominance.

- In-group ranks where the participants rank the dominance of each participant from 1 to the number of participants (3 or 4).

In Chapter 5, we use three measurements as the ground truth dominance for each player: Perceived dominance (PDom) is obtained by averaging the dominance scores from other players. Ranked dominance (RDom) is obtained by averaging the dominance ranks from other players. External dominance score is obtained by averaging dominance scores from all the external observers.

Extension: Nervousness annotation To study nervousness behavior, we also ask *external* observers to score the nervousness in the videos of ELEA. We randomly chose five-minute video segments of every group and assign it to three external trained observers. The rating instruction is the same as in the Resistance data: please rate each person in the video on a 5-point scale from 1 (complete calm and relaxation) to 5 (maximal nervousness and anxiety). Differently, the ratings in Resistance are made by the participants within the game.

Section 3.3

Debate datasets

We introduce two datasets collected from two distinguish debate TV shows, on which we study the persuasion behavior (Chapter 7).

3.3.1. The Qipashuo dataset

The Qipashuo dataset is collected from a popular Chinese debate TV show called Qipashuo[3]. A screenshot of the video is shown in Figure 3.3 (a).

(a) Screenshot of the Qipashuo debate videos.

(b) Screenshot of the Intelligence square US debate videos.

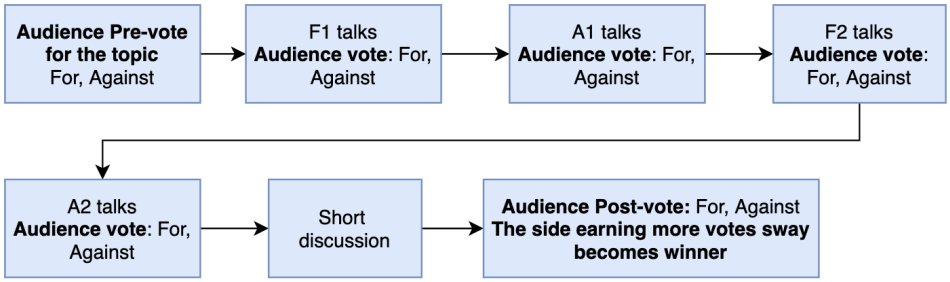
Figure 3.3: Screenshots of the Qipashuo and IQ2US datasets videos.

The debate pipeline is shown in Figure 3.4 (a) (only two rounds are drawn for simplicity). In each episode of the TV show, 100 audience members initially vote 'for' or 'against' a given debate topic. Debaters from 'for' and 'against' teams speak

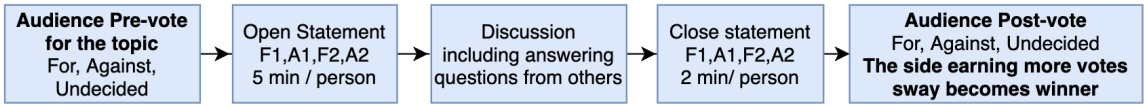
[3]An example can be found at https://youtu.be/P5ehhs0hpFI

alternately, and the audience can change their votes anytime. In general, there are 6–10 speech turns. Final votes are turned in after the last speaker. The winner is the team which has more votes at the end than at the beginning. For example, if the initial and final 'for' vs. 'against' votes are 30:70 and 40:60, respectively, then the 'for' team wins because they increased their votes from 30 to 40 (even though they still have fewer votes than the "against" team).

The videos capture the speakers, and the real-time audience vote ('for' vs. 'against') is shown occasionally. In total, we collect videos of 21 Qipashuo episodes with 205 speaking clips spanning a total of 582 minutes. Note that we also extract the transcripts from the video subtitles using the OCR technique. Details will be discussed in Chapter 7



(a) Debate flow in the Qipashuo (QPS) dataset.



(b) Debate flow in the Intelligence Square US (IQ2US) dataset.

Figure 3.4: Debate flows of the Qipashuo and IQ2US datasets. The Fn and An stands for the nth players in the 'For' and 'Against' team, respectively.

3.3.2. The IQ2US dataset

The IQ2US dataset comes from the popular American debate TV show, intelligence squared US[4]. A screenshot of it is shown in Figure 3.3 (b). The videos have been used

[4]www.intelligencesquaredus.org

by [25, 126, 157, 119, 148] to study persuasion. This dataset was originally collected by [25].

Figure 3.4 (b) shows the debate pipeline. Given a topic, the audience can only vote at the beginning and at the end of the debate, and the winner is determined in the same way as in Qipashuo. Note that we cannot use the same set of videos as [25], since they were interested in predicting the result of the whole debate, which doesn't require the transcripts to be aligned within shorter clips. Of the 100 episodes we collected, only 58 had transcripts that were correctly aligned with the visual modality at the minute level. Finally, we get 852 one-minute single-speaker clip instances from the 58 episodes — 428 of them belong to the winning side. As transcripts are available in the IQ2US data, no pre-processing is required for the language modality in this dataset.

Chapter 4

Visual Focus of Attention Prediction in Videos

Visual focus of attention in multi-person discussions is a crucial nonverbal indicator in tasks such as inter-personal relation inference, speech transcription, and deception detection. However, predicting the focus of attention remains a challenge because the focus changes rapidly, the discussions are highly dynamic, and the people's behaviors are inter-dependent. Here we propose ICAF (Iterative Collective Attention Focus), a collective classification model to jointly learn the visual focus of attention of all people. We evaluate ICAF on a annotated subset of the Resistance data containing 5 videos (35 people, 109 minutes, 7604 labels in all) of the popular Resistance game and a widely-studied meeting dataset with supervised prediction. ICAF outperforms the strongest baseline by 1%–5% accuracy in predicting the people's visual focus of attention. Further, we propose a lightly supervised technique to train models in the absence of training labels. We show that light-supervised ICAF performs similar to the supervised ICAF, thus showing its effectiveness and generality to previously unseen videos.

Section 4.1

Introduction

Given a group G of people, a person $P \in G$, and a short video clip v (1/3rd sec), the Visual Focus of Attention (VFOA) problem is to automatically predict who person P is looking at among all people in G in the video clip v. Solving the VFOA problem can provide profound insights into a number of factors, e.g., who is the dominant person in the group [65], who supports/opposes who in the group, who trusts/distrusts who in the group [83].

Figure 4.1(a) illustrates some of the challenges involved. First, even within a very short 1 second clip, a person may look at many people. The four frames shown in Figure 4.1(a) show the pictured subject looking at three people. Second, multi-person discussions are highly dynamic because many people may speak at the same time and the speakers change rapidly (Figure 4.1) — and as people often look at a speaker, solving VFOA requires the ability to rapidly estimate the VFOA. This is different from the structured meeting setting where there is one presenter. Third, non-verbal behaviors (e.g. eye rolling, head shaking) of people may influence another person's VFOA. Returning to Figure 4.1(a), one would expect people to look at the lady shown when she is speaking — however, their gaze may turn elsewhere if some unseen person makes a gesture. Alternatively, predicting the VFOA of person P might depend on predicting the VFOA of person P1 as both of them might be looking at the same person P2 who is speaking or gesturing. *In short, solving VFOA requires reasoning at the sub-second level and making rapid changes that take into account not only video of the person P whose gaze we are trying to predict, but also that of others.*

We address these challenges via a novel algorithm called ICAF (stands for Iterative Collective Attention Focus) which: (i) reasons at the 1/3 second level that prior

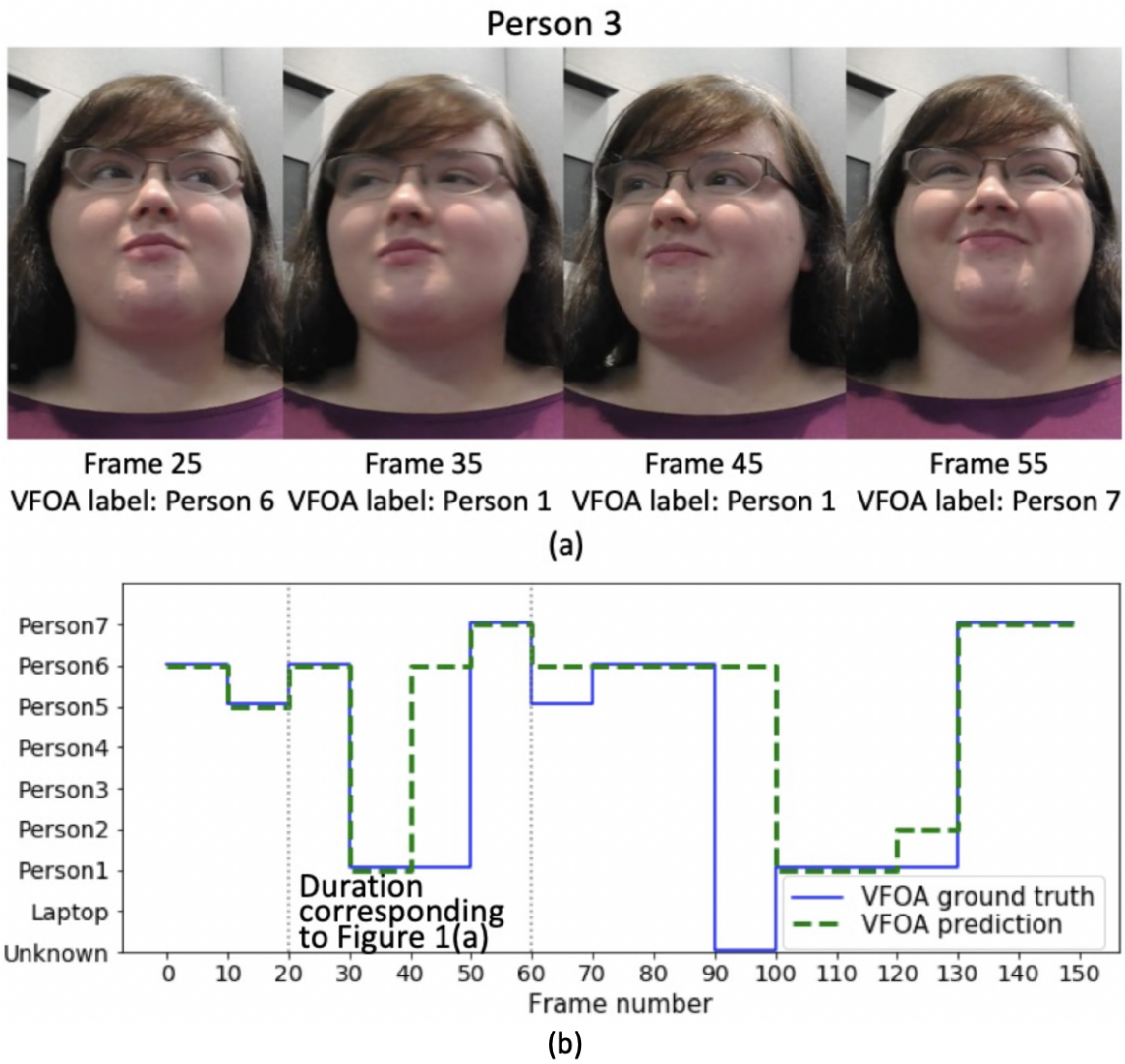


Figure 4.1: (a) An example of a person's (Person 3) Visual Focus of Attention (VFOA) in 4 frames out of a contiguous 4/3 second (40 frames) during a discussion. person 3's VFOA changes rapidly within this short time period, from looking at persons 6, 1, 1, 7, in frames 25, 35, 45, and 55, respectively. Note that even though the head pose in frames 25 and 55 are similar, the VFOA is different (6 vs 7) (b) Person 3's ground truth VFOA and predicted VFOA made by the proposed method, ICAF, of a 5-second discussion clip in which frames 20–60 correspond to Figure 1 (a). We observe that ICAF is able to efficiently predict the rapid change in VFOA.

research has established as the normal duration humans need to visually focus their attention [121], (ii) incorporates collective classification [129, 85] intuitions to capture the fact that where person P is looking might depend on where others are looking, and simultaneously assign VFOAs to all people rather than doing so independently, and (iii) ICAF iteratively builds a multi-layer network that captures the evolution of the collective classification. This captures the idea that predictions of who P is looking at depends on predictions of who others in the group are looking at. (iv) ICAF specifically captures the temporal dependency of VFOA, e.g. the conditional probability that P is looking at Q, given that she was looking at Q in the previous 1/3 sec. *To the best of our knowledge, no prior work on gaze estimation has considered using where others are currently looking and using this to arrive at a joint prediction as we do.*

We introduce a novel dataset (109 mins of video from 5 episodes of the Resistance game in 3 different countries with 35 people). The data was annotated with ground truth VFOA at the 1/3 second level (a huge task by itself leading to over 19,000 annotated 1/3 second clips). Resistance is an immensely popular, dynamic, animated (and sometimes very noisy) party game involving 5-8 people per game. The also well-known Mafia and Werewolf games are variants of Resistance.

We experimentally show that ICAF outperforms several strong baselines in predicting people's next VFOA by over 1.3%, i.e. given a training video up to second t, we predict where each person looks at second $t+1/3$. Moreover, ICAF outperforms the best baseline between 1%–5% when predicting next k VFOAs. For example Figure 4.1(b) shows that even though Person 3 rapidly changes her VFOA during a 5 second multi-person discussion, ICAF predicts her VFOA correctly in 11 out of 14 points (78.6% accuracy), excluding a data point (frame 90-100) without VFOA. Finally, we experimentally show that both temporal dependency and collective classification

boost ICAF's performance.

Since getting ground truth labels is a tedious task, we create a lightly supervised version of ICAF that uses the speaker label to make predictions. We experimentally show that lightly supervised ICAF has similar performance to ICAF, showing the potential of using ICAF for previously unseen videos.

The demo, code, and predicted VFOA networks are available at: `https://www.cs.dartmouth.edu/~cy/icaf/`.

Section 4.2

Related work

Below we discuss the rich literature on predicting visual focus of attention and collective classification.

Predicting visual focus of attention. VFOA is determined by eye gaze. Due to the impracticality of tracking eye gaze in video (video resolution, eye visibility, etc.), many works use head pose as an approximation of VFOA and thus try to estimate head pose. For example, [138], [147] and [156] trained general head pose models from face image patches as input, and [7] employed particle filters to track head pose.

[140] experimentally proved that head pose is a good surrogate of VFOA in meeting scenarios. In real cases, however, only head pose can be misleading as head pose and VFOA may be different. Figure 4.1(a) shows an example in our dataset—while the player's head pose is similar in frames 25 and 55, her VFOA is different. In a different task of continuous gaze angle prediction, [6] took face images from close-up videos to train a Convolutional Neural Network(CNN) to estimate head pose, and further used estimated pose as well as the appearance around eyes to compute eye gaze. They showed that the fusion of head pose and eye gaze reduces mean prediction error of gaze rotation angles, and trained a participant-dependent model to further

boost performance.

Instead of modeling VFOA by static head pose, [131], [102] exploited the correlation between temporal head movement and eye gaze to predict VFOA, in which VFOAs are modeled by Gaussian distributions and their transitions probabilities. The method in [131] needs lots of parameters to be set. [102] proposed a temporal graphical model to efficiently track gaze and VFOA simultaneously, which handled cases when eyes cannot be detected. Although the algorithm is also evaluated in social interaction settings, they assumed the eye gazes of different people as independent, which is different from our assumption and method.

Ba et al. [9] used Gaussian Mixture Models (GMM) to model VFOA as a hidden state, with estimated head pose as observations. The GMM is further extended to a Hidden Markov Model (HMM) to incorporate temporal dependency of VFOA targets. Since the relationship between VFOA and head pose can vary according to individual habits, they used an unsupervised method to adapt Maximum A Posterior (MAP) parameters in their model. Inspired by [10] and [6], we train player-based models for VFOA targets prediction. Due to static seating and close-up cameras in our dataset, we directly use the OpenFace [18] library to extract both head pose and eye gaze.

In a group setting like a meeting or social game, people's VFOAs are influenced by each other due to visual and verbal communications. Stiefelhagen et al. [139] first used the prior that speakers usually draw people's attention to predict VFOA. They modeled VFOA predictions as a linear combination of the condition probability given gaze and condition probability given people speaking. The first term is estimated by GMM, and the second term is estimated using a neural network. Further, [7] took both speaking and visual active cues(gestures, movements, etc) as priors of VFOA, and modeled the probabilities by counting the frequency of people gazing these cues in training data. In contrast, to allow any nonlinear relationship between gaze and

speaker cues, we directly combine gaze and speaking features to jointly train our model.

Besides, Ba et al. [8] and their later work [10] further employed meeting activity context (such as slides updating), as well as a prior that people share VFOA, to predict visual focus of attention. Ba et al. [10] created a Bayesian model with a shared prior to incorporate similarity in participants' behavior, but this prior is constant and the same for all participants. In contrast, our proposed model ICAF adds the inter-player dependency directly, which enables the classifiers to learn the weights for other inputs and can change over time as behaviors shift during a video.

Another line of research lies on unsupervised learning of VFOA. [52] and [53] clustered visual focus of attention by low level Histogram of Gradient (HOG) features extracted from tracked face patches, and the parameter of VFOA transition probability is learned incrementally. The latter further extend the clustering to a dynamic HMM. They don't depend on any prior of participants and environments and can avoid intermediate error of estimating head pose. [21] similarly use head image histogram features, but also consider walking velocity as an observed dependency of gaze. They optimize a Conditional Random Field to estimate gaze in surveillance videos. We also introduce an unsupervised method to predict VFOA and show its efficacy by comparing with supervised results.

Collective classification. Collective classification methods are widely used in the graph mining tasks, such as node labeling [129, 85], link prediction [143] and a combination of both [23]. These methods are able to employ the correlated attributes of nodes/edges in a graph structure , thus train a collection of classifiers that are interdependent together.

Sen et al. [129] gave a brief introduction and experimentally comparison of 4 types of collective classification algorithms, iterative classification, Gibbs sampling,

loopy belief propagation and mean-field relaxation labeling. They further discussed various heuristics of constructing the features incorporating inter-dependent information, which they called relational features, and different ordering strategies of node feature update. In these methods, one group of nodes are usually modeled by a same classifier. [143] exploited the relational Markov network framework to build a joint probability distribution of links and related nodes. Parameters were trained to maximize link observation probability, then used to inference unknown link existence and types. However, these models are developed for static graphs and are not applicable to videos as they are temporal. Moreover, none of these models directly work on predicting the visual focus of attention from videos.

Section 4.3

Problem setup

We annotated a subset of the Resistance data involving the Resistance game[1] containing five games from five different locations—three from U.S.A., one from Israel, and one from Singapore. In each game, up to eight people are seated in an octagon layout (Figure 4.2). It has a total of 35 people whose goal is to identify deceptive people for additional financial reward. Each person has a tablet in front of them which records their activity. At the start of every game, all people introduce themselves, followed by several rounds of discussion where 2-3 people are deceptive and do not want to be identified by the other people whose goal is to unmask them. The people may not leave their seats. The discussions are emergent as there is no pre-determined presenter or leader.

We generated ground-truth labels for people's VFOA for every 10 frames (1/3 seconds in 30 frames per second videos), the time taken to register one's attention [121].

[1] https://en.wikipedia.org/wiki/The_Resistance_(game)

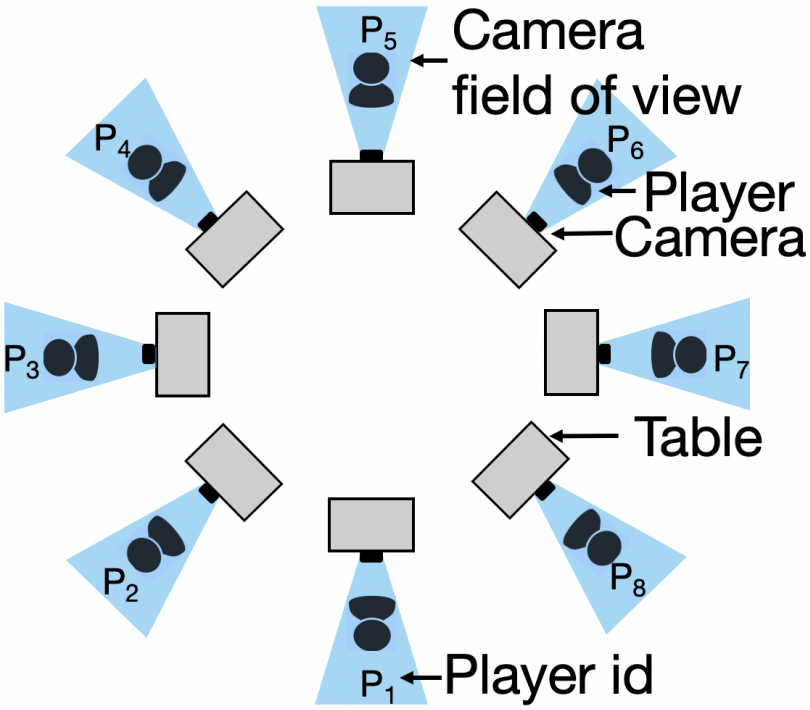


Figure 4.2: Data collection setup.

Video id	Number of seconds	10-frame segments	Number of labels
1	1062	3186	1086
2	896	2688	1541
3	1435	4305	1516
4	1984	5952	2060
5	1134	3402	1401
Total	6511	19533	7604

Table 4.1: The annotated VFOA dataset.

Figure4.1(a) is an example. An expert manually assigned one label for every 10 frame segment of each person. For each person, there are eight possible points of focus—one of the other 7 people and the tablet. A label is assigned if the person looks at the object (person or tablet) for the majority of the 10 frames, otherwise, an 'unknown' label is assigned. This results in a total of 7604 valid labeled segments. The 'unknown'-labeled segments are not used for training or testing.

We extract 3 clips from each game—the entire introduction round (where at most one person is speaking at a time), and two 5-second discussions (where multiple people are simultaneously speaking). This gives 6511 seconds of data in total for the 5 games. Table 4.1 shows the data distribution by game.

AMI corpus. We also used the widely-studied AMI meeting corpus [103], which is highly structured. In this dataset, we used closeup videos of 12 meetings with available VFOA annotation. Each meeting has 4 people and lasts 25 minutes on average. The VOFA targets are 4 people, table, whiteboard and slide screen.

4.3.1. Feature extraction

We extract two sets of features from the clips: face-based features and speaking probability features. As with face-based features, we extract the person's head pose angles and eye gaze vectors using OpenFace [18] since the tablet cameras can capture close-up video of each person.

Speaking prediction. We use visual information to predict if a person is speaking at an instance. First, we get 2-dimensional lip contour points $X^{(t)} = \{(x_i^{(t)}, y_i^{(t)}), i = 1, \ldots, n\}$ at frame t from OpenFace and normalize $X^{(t)}$ by its bounding box to avoid the influence of head movement. Second, we compute the gradient of point positions over time to capture mouth movement, which is $\vec{g}_i^{(t)} = (x_i^{(t)} - x_i^{(t-1)}, y_i^{(t)} - y_i^{(t-1)}), i =$

$1, \ldots, n$, and aggregate them as a frame feature vector $\vec{g}^{(t)}$. Third, we get feature $G^{(t)}$ by concatenating $(\vec{g}^{(t-s+1)}, \vec{g}^{(t-s+2)}, \ldots, \vec{g}^{(t)}, \ldots, \vec{g}^{(t+s)})$ around time t, in a window of size $2s$. This forms a sliding window over time. We use $G^{(t)}$ as a feature, and the introduction part of a game from this dataset to train a general speaking detection model **SP**. Finally, the speaking probability of a person at time t is given by $s = \mathbf{SP}(G^{(t)})$.

We do not create a new model for head pose angles or eye gaze vector extraction. Instead, we use these as inputs to our model to improve the predictions by using them collectively, instead of independently. ICAF takes the head-based features and speaking probability features as inputs.

Section 4.4

Methodology

Here we describe ICAF, the collective classification methods that incorporates inter-person dependencies and temporal consistency to jointly predict the VFOA of all people.

Let $\boldsymbol{f}_{i,t}$ denote the raw input feature vector of person $P_i \in \{P_1, \ldots P_k\}$ at time t. The raw input features for P_i include the head pose angles vector, the eye gaze vector and speaking probabilities vector $\vec{s} = (s_1, \ldots, s_{i-1}, 0, s_{i+1}, \ldots, s_k)$. Note that we don't use P_i's speaking probability s_i in $\vec{s}$, as P_i's speaking activity doesn't directly influence her VFOA. Let C_i denote the VFOA prediction model for P_i. ICAF builds separate models C_i for each person P_i. C_i outputs a vector $\boldsymbol{v}_{i,t}$, the probability distribution of person P_i's visual focus of attention at time t. This output vector specifies the probability that P_i's VFOA is person P_j (or the tablet) for each j. The ground truth label for person P_i at time t is denoted by $y_{i,t}$.

Figure 4.3 illustrates ICAF for k people and an L-layer network. Each person P_i

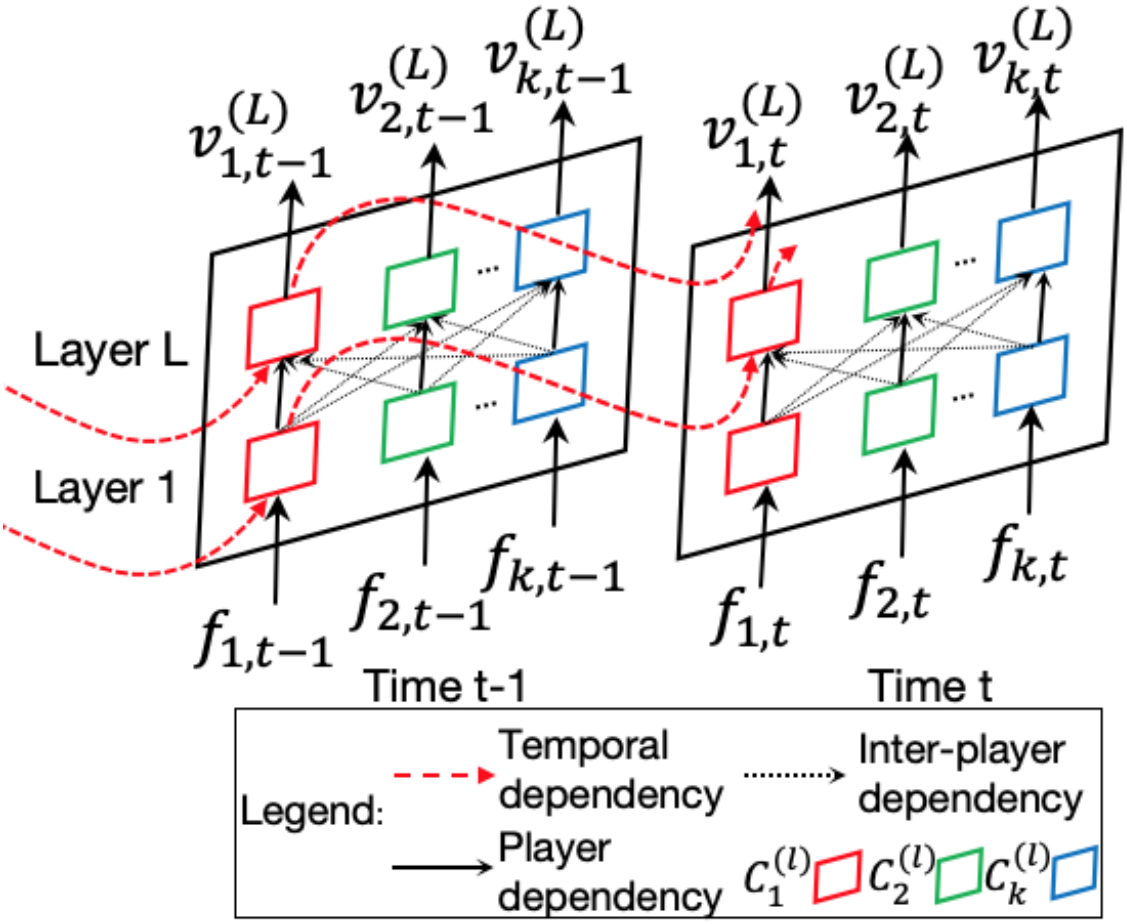


Figure 4.3: Architecture of the iterative collective classification model, ICAF. Each classifier C_i takes three inputs: output of its previous layer (person dependency), previous time (temporal dependency), and other people's output (inter-person dependency). Figure is best viewed in color.

has one classifier $C_i^{(l)}$ for each layer l. Raw features $\boldsymbol{f}_{i,t}$ are used as input for P_i at time t. The model has multiple layers $1, \ldots, L$ to add inter-person dependencies by using the output of other people's classifiers as input (shown in dotted lines). Each classifier also takes the previous timestep's output as input (shown in dashed lines only for C_1 for simplicity). The final output vectors are $\boldsymbol{v}_{i,t}^{(L)}$.

ICAF has three major inputs for each classifier $C_i^{(l)}$ at every time t and layer l as follows: (i) raw features $\boldsymbol{f}_{i,t}$ associated with P_i, (ii) inter-person dependencies $\boldsymbol{v}_{j,t}^{(l-1)}(j = 1, \ldots, k, j \neq i)$ incorporating the influence of the behavior of other people,

Algorithm 1: ICAF Model.

Input : Raw features $\boldsymbol{f}_{i,t}\ \forall i \in [1,\dots k], t \in [1,\dots T]$, Number of layers L.
Output: Predictions $\boldsymbol{v}_{i,t}^{(L)}$ of all people i at all times t

1 $\boldsymbol{v}_{i,0}^{(l)} = (\frac{1}{k+1}, \frac{1}{k+1}, .., \frac{1}{k+1})$
2 $\boldsymbol{v}_{i,t}^{(0)} = C_i^{(0)}(\boldsymbol{f}_{i,t})$
3 **for** $t \in [1,\dots T]$ **do**
  /* Operate on every time step t */
4 **for** $l \in [1,\dots L]$ **do**
    /* Process every layer l */
5   **for** $i \in [1,\dots k]$ **do**
      /* Update person P_i */
6      $S(V) = \sum_{j \in \{1,\dots k\}-\{i\}} \boldsymbol{v}_{j,t}^{(l-1)}$
7      $\boldsymbol{v}_{i,t}^{(l)} = C_i^{(l)}(\boldsymbol{f}_{i,t}, \boldsymbol{v}_{i,t}^{(l-1)}, \boldsymbol{v}_{i,t-1}^{(l-1)}, S(V))$
8   **end**
    /* Make prediction and save $C_i^{(l)}$ */
9 **end**
10 **end**
11 **return** $\boldsymbol{v}_{i,t}^{(L)}\ \forall i \in [1,\dots k], t \in [1,\dots T]$

and (iii) temporal consistency $\boldsymbol{v}_{i,t-1}^{(l-1)}$ enabling the model to make temporally consistent predictions. Together, this results in a collective classification model that makes predictions for all people. The overall algorithm of ICAF is shown in Algorithm 1.

4.4.1. Inter-person dependencies

In a multi-person discussion, the behavior of one person can influence the VFOA of others. Moreover, the behavior of people is highly correlated—when a person is speaking, other people are likely looking at him [10]. This mutual influence can be used to make accurate predictions.

We incorporate the person-to-person influence by adding explicit connections between their classifiers (lines 4–8 in Algorithm 1). In particular, for every person P_i's model C_i, we use the predictions of all other people's models $C_j, \forall j \in \{1,\dots,k\}-\{i\}$ as input. The resulting model is mutually-recursive. To solve this recursion, we un-

$$\boldsymbol{v}_{i,t}^{(l)} = C_i^{(l)}\Big(\underbrace{\boldsymbol{f}_{i,t}}_{\text{Raw input}}, \underbrace{\boldsymbol{v}_{i,t}^{(l-1)}}_{\text{Person input}}, \underbrace{\boldsymbol{v}_{i,t-1}^{(l-1)}}_{\text{Temporal input}}, \underbrace{\sum_{j\in\{1,\ldots k\}-\{i\}} \boldsymbol{v}_{j,t}^{(l-1)}}_{\text{Inter-person input}}\Big)$$

Figure 4.4: Final formulation of ICAF to output $\boldsymbol{v}_{i,t}^{(l)}$ of person i at time t on layer l.

fold the model for multiple layers so that the output of layer l is fed as input to layer $l+1$. This is shown as layers $1, \ldots L$ in Figure 4.3.

Thus, the input to person P_i's model $C_i^{(l)}$ at layer l is its output from layer $l-1$ and an aggregation of the set V of outputs from other people's models from layer $l-1$. The aggregation is a summation represented as $S(V)$, which is used as an input to the model (lines 6–7 in Algorithm 1).

To initialize for layer 1, let $\boldsymbol{v}_{i,t}^{(0)} = C_i^{(0)}(\boldsymbol{f}_{i,t})$, where $C_i^{(0)}$ is the classifier trained by only raw features of P_i, separately.

4.4.2. Temporal consistency

The VFOA of a person at time t is linked to her VFOA at time $t-1$. The temporal consistency component of ICAF explicitly incorporates this dependency by using the output of the predictions made during the last timestep for the person as an input. Specifically, the output $\boldsymbol{v}_{i,t-1}^{(l-1)}$ is an input to $C_i^{(l)}$. This is shown using the dashed lines in Figure 4.3 and in line 7 in Algorithm 1. For each layer l, we initialize $\boldsymbol{v}_{i,0}^{(l)}$ as a uniform probability distribution for VFOA targets.

The final formulation with all the components is shown in Figure 4.4. Overall, ICAF uses the real time inputs along with temporal and inter-person dependencies to jointly predict the visual focus of attention of all people.

Model	Accuracy
GMM(H,E)	0.716
HMM(H,E)	0.770
DBN(H,E,S)	0.800
G-DBN	0.782
GC(H,E,S)	0.756
PC(H,E,S)	0.818
ICAF	**0.831**

Table 4.2: Next VFOA Prediction: Table reports accuracy of ICAF and baselines using all features. Note that the best results of GC, PC, and ICAF are achieved by RF. H, E, and S denote head pose, eye gaze and speaking probability features, respectively. *All improvements of ICAF are statistically significant ($p < 0.05$).*

Section 4.5

Experiments

We conduct several experiments on Resistance and AMI datasets to show:

- ICAF outperforms all strong baselines by 1.3% in predicting VFOA in the next time step (i.e., 10 frames) with $p = 0.046$ by two-sample t-test.
- ICAF significantly outperforms the highest baseline by up to 5% when making predictions upto k time steps in the future ($p < 0.05$).
- Collective classification and temporal dependencies boost the performance of ICAF significantly.

4.5.1. Baselines

We compare with three sets of baselines that use head pose vector (H), eye gaze vector (E), and speaking probability vector(S) for predictions. The first set of baselines are [9, 10, 102], with comparable numbers of VFOA targets in similar settings. Specifically, GMM(H), GMM(H,E) use Gaussian Mixture Model with parameters from each individual [9]. HMM(H), HMM(H,E) uses Hidden Markov Model [9]. DBN(H,S), DBN(H,E,S) uses Dynamic Bayesian Network (DBN) incorporating conversational

dynamics and a shared constant focus prior [10]. Note that the screen activity feature is removed to adapt to our dataset. G-DBN uses DBN to track VFOAs and eye gaze simultaneously with people's global head poses as inputs [102]. In our dataset, people sit uniformly in a circle, so we convert their local head poses to global ones given poses of their cameras.

Further, we created two more sets of baselines using three sets of features H, (H,E) and (H,E,S). The second set of baselines trains one general classifier GC for all people by including the person index as input feature vector [10]. The last set of baselines trains a person-specific classifier PC for each person [6]. As in the case of GC, we create three baselines PC(H), PC(H,E), and PC(H,E,S).

4.5.2. Experimental setting

To get speaking probability features, we set the sliding window size as 30 frames (1 sec) and train a Random Forest speaking detection model **SP**. The training data uses people's introductions as speaking samples, and other people's introductions as non-speaking samples. The introductions were not drawn from our 5 video samples. We evaluate ICAF and baselines by respecting the temporal order of data. Instead of doing a k-fold cross-validation, we train the model for the first T data points and test on the $T + 1^{th}$ data point (each data point consists of 10 frames). T is varied from 96.3% to 99.9%, and the results are averaged. Note that we can not do a leave-one-game-out experiment [10] as the model needs to be trained on each person specifically. Recall that the data for each game is divided into three parts: an introduction round and two discussion rounds.

The introduction round clips are only used for training, and the temporal evaluation is done with the two discussion rounds. Both training and testing are at the frame level. Frame VFOA probabilities are further averaged over 10 frames as probabilities at each 10-frame segments. Given the generality of our model, we experiment with 4

classifiers: Random Forest (RF), Logistic Regression (LR), Linear SVM (LINSVM) and Gaussian Naive Bayes (NB). In all cases, ICAF has 3 layers. 70 trees are used in RF. All models are compared using the accuracy metric.

4.5.3. Next VFOA prediction

We compare ICAF with all baselines using all features. All models are trained on the first T data points and then used to predict the $T + 1^{th}$ data point. *Note that this means that we are predicting the visual focus of attention for each person 1/3 second into the future.* The features given to ICAF for every frame are the head pose vector (H), eye gaze vectors (E), and speaking probability vectors (S).

Table 4.2 shows the results. For fairness, we add eye gaze features (E) to baselines GMM, HMM and DBN. (i) Person-specific baseline models perform better than the corresponding general-classifier baselines using the same set of features. Specifically, PC(H,E,S) performs at least 6.2% better than GC(H,E,S). (ii) More importantly, ICAF performs between 1.3%–11.2% better than all baseline models. (iii) Indeed, it is 3% higher than state-of-the-art method DBN(H,E,S).

4.5.4. Longer-future predictions

We next evaluate the robustness of ICAF by predicting the $T + k^{th}$ data point while training only till the T^{th} data point. We vary k from 1 to 10, meaning that we predict who a person will look at between 0.3 and 3.3 seconds into the future.

Figure 4.5 shows the result. ICAF outperforms the best baseline by up to 5%. In fact, it is better than DBN(H,E,S) by 1.5%–5.7%. Moreover, ICAF is relatively stable as k increases, while some baselines drop rapidly. Specifically, ICAF's prediction accuracy varies only 7.5% over k, so it gives robust estimation of VFOA in the longer-term future.

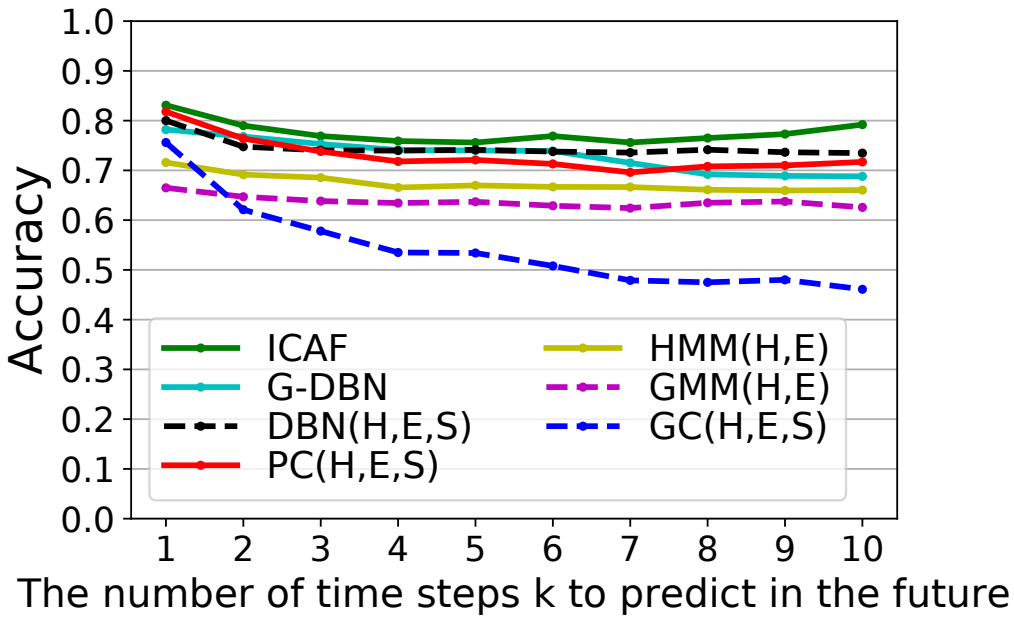


Figure 4.5: Longer-Future Prediction: Accuracy of predicting k steps to the future. ICAF is the highest over all time steps, and outperforms the best baseline by up to 5% ($p < 0.05$). Specifically, it outperforms state-of-the-art method, DBN(H,E,S) [10] by 1.5% – 5.7% .

4.5.5. Contribution of collective classification

Figure 4.6 compares the results of ICAF with and without the temporal and collective classification components. Note that ICAF without both components is equivalent to the baseline PC(H,E,S).

We observe that each of them boost the performance of ICAF from 0.2% to 5.3% w.r.t. all base classifiers. The combination of both components is important in ICAF: the performance of PC(H,E,S) is lower than ICAF without either of the components. Additionally, adding collective classification improves performance more than the temporal component alone. In addition, it is important to note that adding the collective component upon persons' speaking information can further boost the performance (from PC(H,E,S) of ICAF without the temporal component). Specifically, from 0.6% to 10.8%. Hence, the collective component of our model is able to cap-

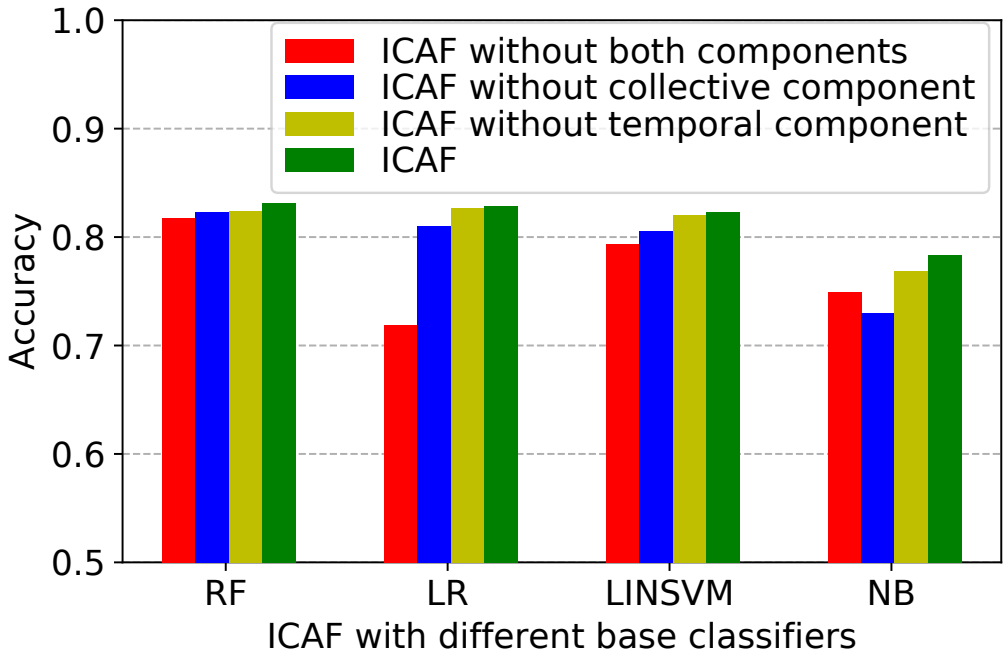


Figure 4.6: Contribution of collective classification: The performance drops when either the collective or the temporal components is removed and drastically when both are removed.

ture both the verbal activities, which draw persons' attentions, and the non-verbal activities, with the help of inter-dependent visual focus of attention. Therefore, both temporal and collective classification components of ICAF are essential, and the collective component is more critical for good predictions.

4.5.6. Comparison with different features

We next explore the effects of different features on ICAF and baselines. Note that RF is used as the (base) classifier to obtain best results for GC, PC, and ICAF .

Table 4.3 shows the results for next VFOA prediction. First, for all models, eye gaze features E boost the predictions. It especially boosts [10, 9] by at least 13.5%. Second, speaking features S boost all models except for GC. These demonstrate that both E and S contribute to prediction of VFOA. Third, using features including E or S, ICAF outperforms all baselines.

Model	H	H,E	H,S	H,E,S
GMM	0.525	0.716	-	-
HMM	0.623	0.770	-	-
DBN	-	-	0.665	0.800
GC	**0.719**	0.799	0.731	0.756
PC	0.716	0.805	0.771	0.818
ICAF	0.718	**0.811**	**0.784**	**0.831**

Table 4.3: Comparison between different features: Both E and S boost the accuracy of all models except GC, and ICAF performs the best in 3 out of 4 cases. ($p < 0.05$)

Model	Static meetings	Dynamic meetings
[10]	0.556	0.520
ICAF	**0.568**	**0.538**

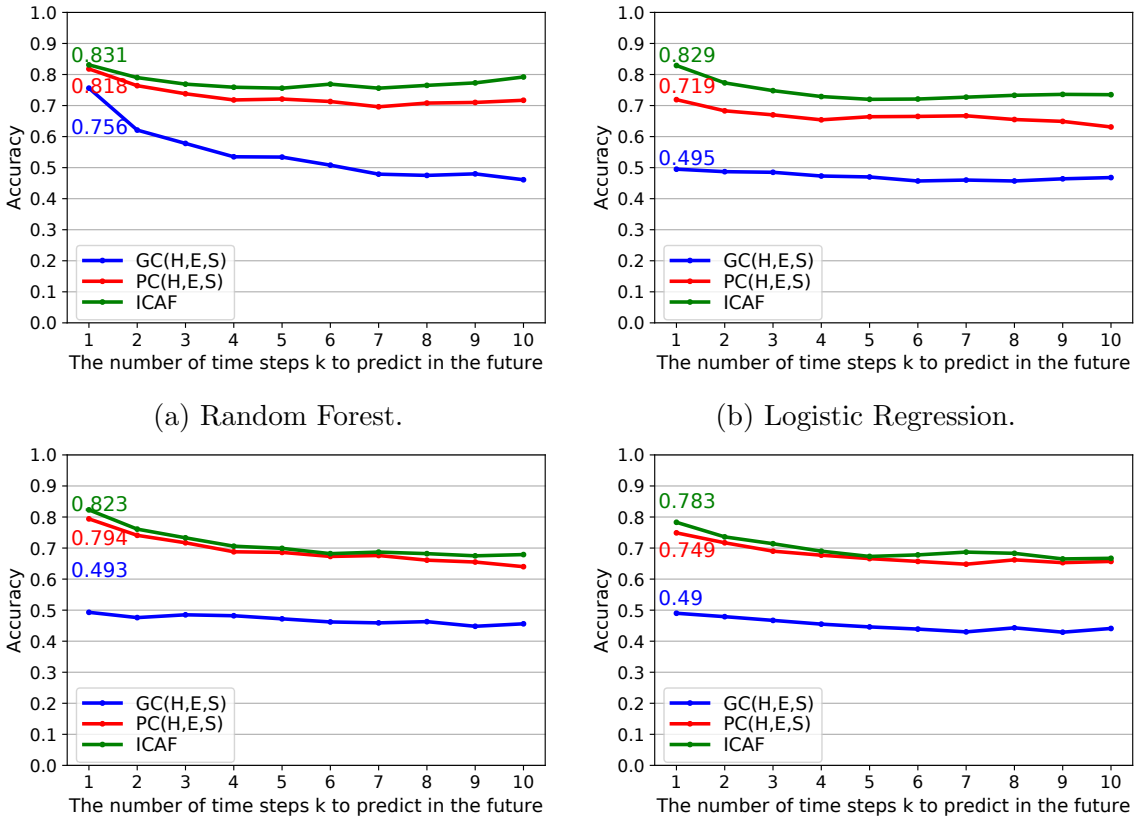
Table 4.4: AMI corpus experiments. Accuracy of the proposed model on static and dynamic meetings.

4.5.7. Comparison between different base classifiers

Here we explore performance of ICAF with different kinds of base classifiers: RF, LR, NB and LINSVM. In Figure 4.7 we compare ICAF with GC and PC in the cases of both next VFOA prediction ($k = 1$) and longer-future VFOA prediction ($k > 1$). The colored texts show the results for $k = 1$, where ICAF outperforms the corresponding best baseline by 1.3%-11%. For $k > 1$, it outperforms the best baseline by up to 5% with RF, 12% with LR, 3% with LINSVM, and 4% with NB. Thus, we observe the generality of ICAF.

4.5.8. AMI corpus experiments

We also conducted experiments on the AMI meeting corpus [103]. 8 meetings are dynamic, where people sit around a table and upto 1 person moves to the white-board/screen to present. 4 meetings are static, where all people remain seated. We use people's closeup videos to extract head pose, eye gaze, and speaking probability. We followed the leave-one-out protocol as in [10] and compare frame-based accuracy.



(a) Random Forest.

(b) Logistic Regression.

(c) Linear SVM.

(d) Naive Bayes.

Figure 4.7: Comparison between different (base) classifiers. In each subfigure, each of 3 colored numbers indicates the prediction accuracy of $k = 1$ in the same colored line.

Since the 4 seats over all meetings are fixed, we train seat-specific classifiers in ICAF.

Table 4.4 shows that ICAF outperforms [10] in both kinds of meetings.

Section 4.6

Lightly supervised VFOA prediction

A major challenge in VFOA prediction is the lack of labeled data for new videos. Annotating VFOA at a second or sub-second granularity is highly time-consuming and often not clean. We now propose to generate accurate VFOA predictions without ground truth labels. The proposed technique is general and can be used to train both

the baselines and ICAF.

The intuition is that people are highly likely to look at the person who is speaking if there is a single speaker [139]. Building on this intuition, we identify continuous clip segments where one person is speaking. This is done using the speaking prediction model **SP** described in Section 4.3.1. To reduce false positives, we further average over 10 frames' prediction probability around the current frame and use it as the final label to select single-speaker segments. We select two longest clips for each player as they are supposed to do it according to the introduction rule. For a segment where P_i is speaking, we assign i as the training label for all other people and the model is trained with it. To evaluate the effectiveness of this training method, we train all models using the introduction (by generating its speaker labels) and use the two discussion clips with the ground truth VFOA labels as test.

Figure 4.8 shows the results for all baselines and ICAF using RF as base classifier. Since the training labels are speaking labels, we remove speaking probability features from ICAF as well as baselines. Compared to random prediction of 14.4%, the lightly supervised training technique generates 41.2%-54.7% results. We also observe that ICAF performs better than the baselines. For comparison, Figure 4.8 shows the equivalent result with supervised training, where we train the models using the ground truth focus labels in the introduction round as well. We note that the lightly supervised prediction is comparable to supervised prediction, showing the effectiveness of the proposed training technique.

Using the light supervised ICAF method, we have extracted the who-look-at-whom networks from the videos of Resistance data, and released the networks at `https://www.cs.dartmouth.edu/~cy/icaf/` for future research.

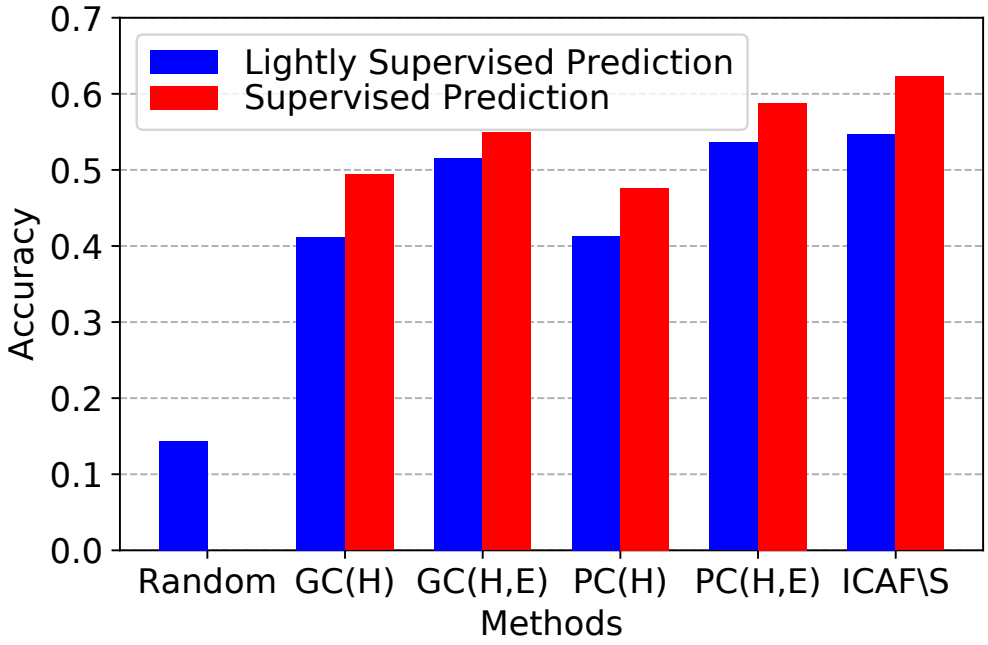


Figure 4.8: Lightly supervised predictions (in blue) and supervised predictions (in red): 'Random' denotes random prediction accuracy, and ICAF\S denotes ICAF without speaking feature.

Section 4.7

Conclusion

We showed that by explicitly incorporating inter-person dependencies and temporal consistency are crucial to accurately predict VFOA both in short-term future and long-term future. The ICAF model is, therefore, able to overcome the challenges of rapidly changing VFOA, high dynamics of the discussion, and person-person inter-dependencies. Moreover, the lightly supervised ICAF is crucial in making the model general to unseen videos. This opens doors to new research in efficient extraction of interaction networks from videos without any training labels.

Chapter 5

Dominance Prediction in Multi-person Videos

We consider the problems of predicting (i) the most dominant person in a group of people, and (ii) the more dominant of a pair of people, from videos depicting group interactions. We introduce a novel family of variables called Dominance Rank (DR). Combined with features not previously used for dominance prediction (e.g. facial action units, emotions), we develop an ensemble-based approach to solving these two problems. We test our models against 4 competing algorithms in the literature on two datasets and show that our results improve past performance. We show from 2.4% to 16.7% improvement in AUC compared to baselines on one dataset, and gain of 0.6% to 8.8% in accuracy on the other. Ablation testing shows that DR features play a key role.

Section 5.1

Introduction

The problem of identifying dominant people in a group setting is important for many applications. Businessmen in meetings with external partners or customers might wish to identify the key decision maker. Government delegations may be interested in identifying the most dominant person from the other side in a negotiation.

In this chapter, we study two problems: identifying the most dominant person (MDP problem) in a group-interaction video and identifying the more dominant person when looking at pairs of people in a group interaction (pairwise dominance prediction or PDP). Although the MDP problem has been previously studied in pioneering works by Jayagopi et al. [75] and Aran et al. [3], we are the first to study the PDP problem. We look at two variants of each of these problems (MDP-All and MDP-Distinct, PDP-All and PDP-Distinct). The chapter makes three novel contributions. First, we propose a family of *Dominance Rank* features, which captures the dynamics of interactions between participants in a group-interaction video. Second, we propose the *Dominance Ensemble Late Fusion* (DELF) algorithm that uses Dominance Rank in combination with several other features to solve all 4 problems. Third, we propose the *Group Dominance Prediction* (GDP) algorithm to solve MDP-All and MDP-Distinct.

We test the DELF and GDP algorithms on two datasets. Our first setting consists of audio-visual data of groups of people playing a variation of The Resistance game. We used a subset of Resistance data for 33 games involving 233 players with ground truth involving surveys on who is the most dominant. Each game involves 5–8 players. The data was collected from six sites (three in the US, one each in Israel, Zambia, and Singapore). The second dataset is the widely used ELEA dataset [125], which shows

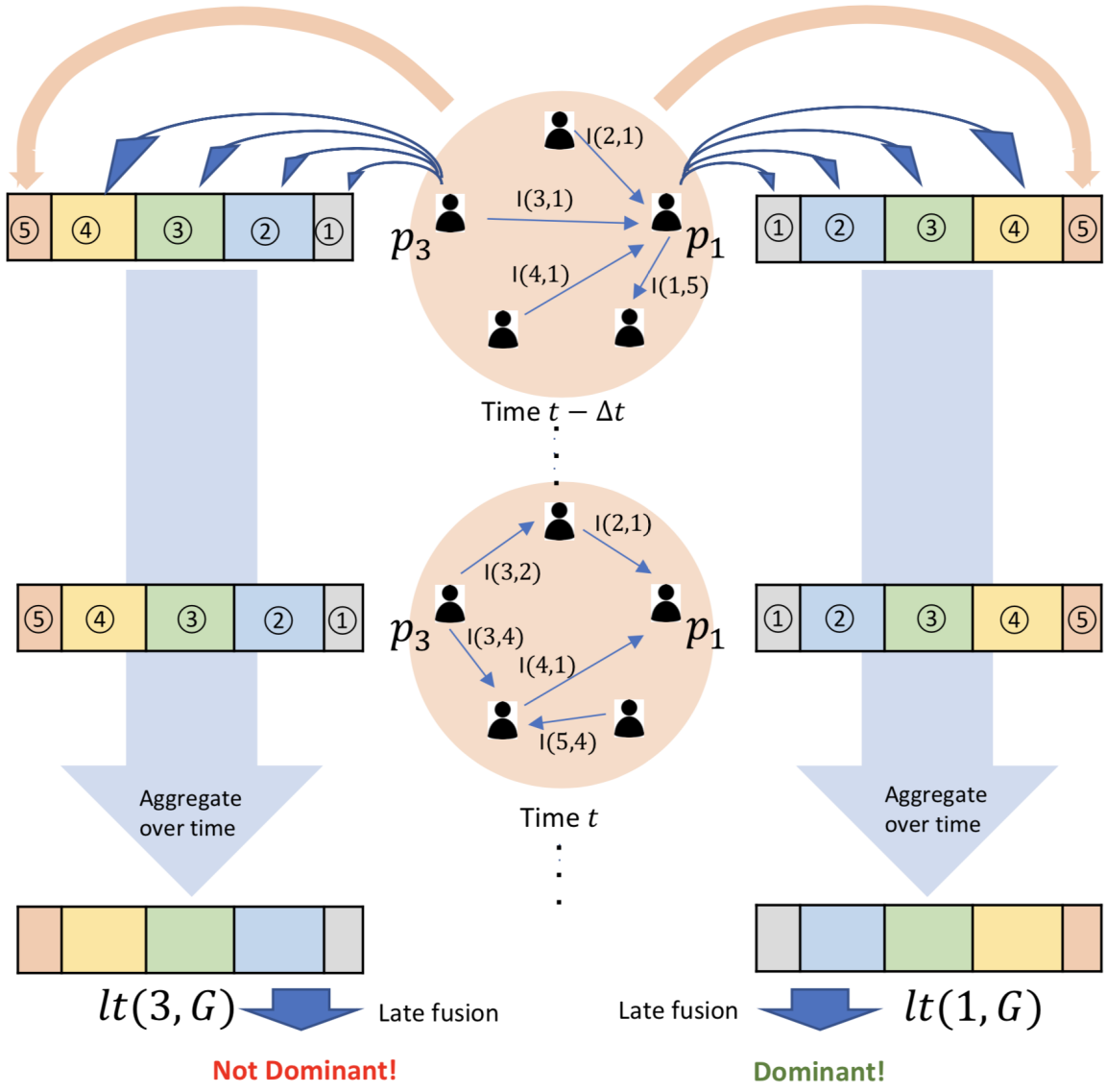


Figure 5.1: Our approach. In group G, for each player p at time t we have individual basic short-term features (1–4) and Dominance Rank features (5) based on the group interaction. We aggregate each kind of features over time to get long-term features for each player. Finally, we use an ensemble late fusion approach to make the final prediction.

small groups (3–4 people) involved in a winter survival task. The Resistanceand the ELEA datasets further differ in the nature of social interaction present in them. The former involves an adversarial situation and models a conflict between two groups (an informed group of spies and an uninformed group of resistance). Whereas in the ELEA dataset, there is a cooperative element as players wish to solve a common task. We test DELF and GDP both against each other and against several baselines and show that DELF beats out strong baselines from past work, and GDP beats out DELF. We should note that DELF is an improvement on past work, and hence all the excellent body of past work contributes to this algorithm.

Figure 5.1 depicts our approach to the four problems we study in this chapter. We first divide each game G's videos into equal time slices of length Δt seconds. For each player p, we then create a *basic short-term* feature vector $bst(p, t, g)$ showing the values of basic features (defined in Section 5.3) for player p during time slice t. The basic features fall into four categories: speech-related features [60], facial action unit features [19], emotion-related features from Amazon's Rekognition, and Mel-Frequency Cepstral Coefficient (MFCC) features [43]. We note that the emotion and MFCC features have never been used in prior work on dominance prediction. Based on $bst(p, t, G)$, we develop a novel set of *Dominance Rank* features, inspired by the PageRank algorithm, on top of basic features. We thus have five types of short-term features, all applicable to short video segments.

The ground truth dominance labels in both datasets are provided for an entire game. Therefore, we need to predict whether a player is the most dominant in a game as a whole (or more dominant than another player in the game as a whole) rather than in a short time segment Δt. For this, we associate a basic long-term feature vector $blt(p, G)$ that aggregates the features for the short-time slices into features for the game as a whole using Fisher vector encodings [118] and histograms. A similar

aggregation is also applied to the Dominance Rank features to get a vector $lt(p, G)$ of the long-term features for player p in game G.

We then develop predictive models based on each type of long-term features and develop an ensemble late fusion approach that merges the five predictive models to make a final prediction. We investigate the importance of each type of features in the ensemble predictor and show that Dominance Rank features play an important role. We re-emphasize that DR features build on top of basic features including ones proposed by others.

Finally, we also develop a *Group Dominance Prediction* (GDP) algorithm, which relies on the intuition that considering all players in the game at once is more beneficial than treating them independently. This naturally sets up a classification problem where each player's lt feature vectors are fed into the classifier for training, together with the one-hot encoding for players (most dominant player in that game or not). Because games can have 5–8 players, we associate with each game G, and each possible subset of 5 players in that game, the concatenation of the feature vectors of those 5 players, along with an indication of which player was the most dominant. We then learn a classifier on the resulting data.

Section 5.2

Dataset and task descriptions

Resistance-dominance data The Resistance dataset is described in Chapter 3. Every player is rated by every other player on an integer 1–5 scale (1 is not dominant at all, 5 is very dominant). We find the median score for each player and call it the ground truth perceived dominance score of the player in that round. The data contains 158 rounds with complete dominance labels in all.

ELEA data We also use the ELEA dataset ([125]) with 27 groups of videos, described in Chapter 3. We use three variants of dominance labels: Perceived dominance (PDom), Ranked dominance (RDom), and external dominance score (Chapter 3).

Most Dominant Player (MDP) The MDP problem is to find the most dominant player in a given round. This is a binary classification task with labels 1 if a player has the highest perceived dominance score among all the players in this round, and 0 otherwise. In our setting, however, more than one player in the group can have the highest dominance score. We therefore consider two instances of the problem: finding the most dominant players in all rounds (MDP-All), and finding the most dominant player in every *distinct* round (MDP-Distinct). A round is *distinct* if there is a single player with the highest dominance score.

Pairwise Dominance Prediction (PDP) We also consider the more fine-grained problem of pairwise comparison. The PDP-All problem takes two players in a game as an input and predicts which one has the higher dominance score. To pose this as a binary classification problem, we discard pairs with equal scores. The PDP-Distinct problem predicts which player in a pair is more dominant when the dominance scores of the players differ by 1 or more. We call such pairs of players *distinct pairs*.

Section 5.3

DELF and GDP algorithms

We have already provided a brief overview of our architecture in Section 5.1. We first describe our basic short-term features and then our Dominance Rank features (both denoted further as stf), followed by the extension of the short term features to the video as a whole.

5.3.1. Basic short-term features

Past work has shown that speech-related cues ([49, 22]) and gazing information ([65, 125]) are closely related to perceived dominance of a person. We also use facial expressions and emotions as additional signals for visual dominance ([55]). Our basic short-term features use audio-visual features from the frontal videos of players. While the use of these features is not novel, we note that emotion scores, facial action units, and MFCC features have never been used before for dominance prediction.

- *Speaking probability* $s_t(p_i)$ is an estimate of a probability that the player p_i is speaking during time interval t. This probability is estimated from the lip movement of the person for every $\Delta t = 0.33$ seconds [60].

- *Gazing probability* $g_t(p_i, p_j)$ is an estimate of the probability that player p_i looks at player p_j in time interval t ([10]). This probability is estimated for every $\Delta t = 0.33$ seconds according to Rayner et al. [121].

- *Facial Action Units* scores (FAUs) capture the intensity of 17 action units in the given frame. These values are produced with the OpenFace library [19].

- *Emotion scores* are the estimates of intensity of eight emotions and two facial traits (smile, open eyes) produced by Amazon's Rekognition.

- *Audio features* are represented by Mel-Frequency Cepstral Coefficients, which are widely used in audio analysis [43].

The concatenation of the above features yields a basic short term feature vector $bst(p, g, t)$ for each player p in game g's t'th time interval.

5.3.2. Dominance rank features

Previous research on dominance and leadership analysis shows that dyadic statistics are correlated with dominance ([4, 109, 125]). We propose a family of short-term

Table 5.1: Interaction functions for the Dominance Rank features and their Pearson (r) and Spearman (ρ) correlation with ground truth dominance scores. All correlation coefficients are significant with $p < 0.01$.

$\boldsymbol{I(p_i, p_j)}$	$\boldsymbol{r}$	$\boldsymbol{\rho}$
$G(p_i, p_j) - G(p_j, p_i)$	0.21	0.23
$G(p_i, p_j)/G(p_j, p_i)$	0.1	0.11
$LL(p_j, p_i) - LL(p_i, p_j)$	**0.49**	**0.53**
$LL(p_j, p_i)/LL(p_i, p_j)$	0.33	0.36
$LL(p_i, p_j)/LL(p_j, p_i)$	-0.26	-0.32
$LS(p_j, p_i)/LS(p_i, p_j)$	-0.16	-0.16
$LS(p_i, p_j) - LS(p_j, p_i)$	0.24	0.23
$LS(p_i, p_j)/LS(p_j, p_i)$	0.2	0.19
$LS(p_i, p_j)/LL(p_i, p_j)$	**0.50**	**0.52**
$LL(p_i, p_j)/LS(p_i, p_j)$	0.29	0.30

Dominance Rank (DR) features capturing the mutual interactions between players in the game. Suppose $I(p_i, p_j)$ is a function that returns a value capturing the interaction between players i and j (we will show several such functions shortly). The short-term Dominance Rank $R_{\text{dom}}(p_i)$ of a player p_i w.r.t. function I in a given time period t is defined by the following equation:

$$R_{\text{dom}}(p_i) = \frac{1-d}{N} + d \sum_{j \neq i} \frac{R_{\text{dom}}(p_j) I(p_i, p_j)}{N-1} \ , \tag{5.1}$$

where N is the number of players in the game, $I(p_i, p_j)$ is an interaction function, and $d \in [0, 1]$ is a damping factor. Damping factor d regulates the importance of the interaction function for the values of the Dominance Rank (the larger the d the more important role plays the interaction function). As d increases, the interaction function I plays an increasingly important role in determining the dominance of players, whereas when d is small, it plays a less important role. Although we note that Dominance Rank builds upon the idea of PageRank, unlike PageRank, R_{dom} is not one function, but a family of functions one for each possible interaction function I. Like PageRank, we set $d = 0.85$.

Computation of Dominance Rank Features. We compute the Dominance Rank the same way as the PageRank. We write Equation 5.1 in the matrix form:

$$\mathbf{R}_{dom} = \frac{1-d}{N}\mathbf{1} + \frac{d}{N-1}\mathbf{M}\mathbf{R}_{dom} \ , \tag{5.2}$$

where

$$\mathbf{R}_{dom} = [R_{dom}(p_1), \cdots, R_{dom}(p_N)]^\intercal \ ,$$
$$\mathbf{1} = [1, \cdots, 1]^\intercal, \text{ and}$$
$$\mathbf{M}_{ij} = \begin{cases} I(p_i, p_j), & \text{if } i \neq j. \\ 0, & \text{otherwise.} \end{cases}$$

The Equation 5.2 can be solved efficiently as a linear system of N equations with N unknowns. Since the number of players in a group N in the datasets is small (no more than eight), computing Dominance Rank for all players in the game takes constant time.

Interaction Functions. We define a family of interaction functions, each of which yields a different dominance rank function R_{dom} when used in Equation 5.1. We consider combinations of basic values defined on a slightly larger time period than basic features, representing interactions between players: S (speaking rate), G (gazing rate), LS (looking while speaking) and LL (looking while listening) defined as follows:

$$S(p_i) = \frac{1}{k}\sum_{t=t_1}^{t_2} s_t(p_i) \ , \tag{5.3}$$

$$G(p_i, p_j) = \frac{1}{k}\sum_{t=t_1}^{t_2} g_t(p_i, p_j) \ , \tag{5.4}$$

$$LS(p_i, p_j) = \frac{1}{k}\sum_{t=t_1}^{t_2} g_t(p_i, p_j) s_t(p_i) \ , \tag{5.5}$$

$$LL(p_i, p_j) = \frac{1}{k}\sum_{t=t_1}^{t_2} g_t(p_i, p_j) s_t(p_j) \ , \tag{5.6}$$

where k is the number of time slices of length Δt, on which we define speaking and gazing probability, that fit into the time period (t_1, t_2) for the Dominance Rank. In our experiments we exploit time periods of 1 and 5 seconds for Dominance Rank features, thus k equals to 3 or 15. Based on these values, we define a set of interaction functions (Table 5.1) representing how interaction between players may be connected to distribution of dominance in the group, e.g., if less dominant players look at more dominant players more often than the other way around (in rows 1–2). For example, the 1st one indicates how much does person p_j looks at p_i more than the opposite. These functions control the way and volume of transferring between people's dominance ranks during their interactions.

Normalized Dominance Ranks. To compare Dominance Rank (Equation 5.1) for players from different games, we normalize these values to be in $[0, 1]$. Table 5.1 lists some of the interaction functions we explored and the Pearson/Spearman Correlation Coefficients (r/ρ) of resulting Dominance Ranks with ground truth dominance scores. We recall that correlation coefficients lie in the $[-1, +1]$ interval. We see that some of the Dominance Rank Functions such as those associated with interaction functions $LL - LL$ and LS/LL (rows 3 and 9 respectively) demonstrate strong correlation with

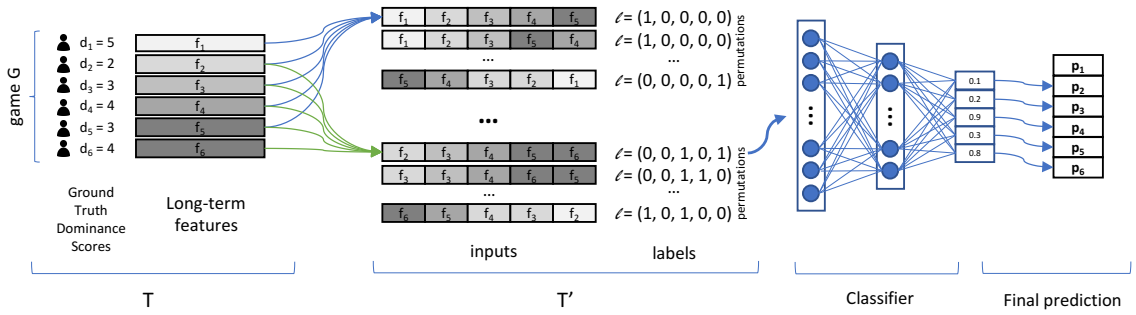


Figure 5.2: GDP algorithm. Given a dataset T, for every game G we form all possible groups of 5 players, form labels based on the players scores in every group, and concatenate long-term features for players to get group feature. We also augment the dataset with all possible permutations of the players. Then we train a model for the task of multilabel or multiclass classification on the new dataset T'. For the final prediction for a specific player we average predicted scores over all the groups and permutations where that player is present.

ground truth dominance scores. *Throughout the rest of this chapter, we will use the expression "short term features" (stf) to denote the set of basic short term features as well as normalized short term Dominance Rank features associated with any player in a game.*

5.3.3. Long term features

Since players in the game are instructed to score each other's dominance for only the round before the survey, to train models for our four classification tasks, we need to produce features representing whole rounds, which last 15 minutes on average. The features above, however, are extracted over a short-term period of time from 0.33 to 5 seconds. To aggregate these features we use two methods described below.

Fisher-Vector features. Fisher vector (FV) is a bag-of-words model heavily used for object recognition in images [118]. Note that each round may have a different duration and hence the number of *bst* features can vary from round to round. Fisher Vectors aggregate the features of an arbitrarily long video into a fixed length encoding — we use 256-dimensional features for our experiments. The details can be found at

Section 2.4.

Histogram features. We compute a histogram feature for every short-term feature (both *bst* features and normalized short-term Dominance Rank features). For a player p_i in a game round G and a short-term feature stf, we have a set of all feature values for all short intervals over the round $\{stf(p_i, t_1, G), \ldots, stf(p_i, t_T, G)\}$. We build a histogram (formally described in Section 2.4) of short-term features $\mathcal{V}_{stf} = \langle v_1, v_2, \ldots, v_b \rangle$, where v_l are frequencies of values $stf(p_i, t_j, G)$ falling into the lth bin; b is the number of bins determined through cross-validation (on the training set alone).

5.3.4. Dominance Ensemble Late Fusion (DELF)

The best classifier for feature type i returns a *score* S_i denoting the probability of a subject being the most dominant player in the corresponding round. DELF then fuses the scores $S_1, \ldots, S_5$ by weighted late fusion described in Section 2.5. The best classifier for each of the five types of features is determined by exhaustive search through all possible combinations of classifiers.

5.3.5. Group Dominance Prediction (GDP)

We propose the Group Dominance Prediction (GDP) algorithm for solving MDP-All and MDP-Distinct. GDP's pseudo-code is shown as Algorithm 2 and also on Figure 5.2.

We reason that to determine the most dominant player in a game we need to compare players within that game to each other, therefore it should be beneficial to provide a classifier with features of all players in that game at once. In Resistance-dominance dataset, however, the numbers of players in games vary, which prevents us from building a single model with fixed feature length. GDP's goal is to develop a modified training set. The algorithm considers each game in turn and looks at all

Algorithm 2: GDP(T: training set, *ltf* : long term feature type)

1 $T' = \emptyset$
2 **foreach** *game* $G \in T$ **do**
3 $\quad G_5$ = set of all subsets containing 5 players from G
4 $\quad \Pi_5$ = set of all permutations of 5 elements
5 $\quad$ **foreach** $(i_1, i_2, i_3, i_4, i_5) \in G_5$ **do**
6 $\quad\quad$ **foreach** $\pi \in \Pi_5$ **do**
7 $\quad\quad\quad (j_1, j_2, j_3, j_4, j_5) = \pi(i_1, i_2, i_3, i_4, i_5)$
8 $\quad\quad\quad \text{Dom} = \arg\max_{p_{j_k}} \text{GT\_Dom\_Score}(p_{j_k})$
9 $\quad\quad\quad \text{input} = \text{concat}\,(ltf(p_{j_k}) \mid k = 1, \ldots, 5)$
$\quad\quad\quad$ `/* 𝟙_Dom(x) is indicator function */`
10 $\quad\quad\quad \text{label} = \text{concat}\,(\mathbb{1}_{\text{Dom}}(p_{j_k}) \mid k = 1, \ldots, 5)$
11 $\quad\quad\quad T' = T' \cup \{(\text{input}, \text{label})\}$
12 $\quad\quad$ **end**
13 $\quad$ **end**
14 **end**
15 Train a classifier on T'

possible subsets G_5 of 5 players in that game (the smallest possible number of players in any game). For each subset in G_5, GDP considers the maximal ground truth dominance score (Step 8). It then generates a new feature vector by concatenating the long-term feature vectors of the five players (Step 9) and then assigning a label of 1 to the most dominant players in the subset and a label of 0 to the others (Step 10). Furthermore, GDP considers all permutations of players to augment the dataset (Steps 6–7). This creates a new training set with feature vectors 5 times as long as before. GDP then trains a classifier (multilabel for MDP-All, multiclass for MDP-Distinct).

At the inference time, GDP performs the same procedure (forming subsets of 5 players and all permutations) with the validation set. Once all the binary predictions are made, to obtain the final probability of a player being most dominant in the game round, we average the predictions for this player for all groups and permutations where this player is present.

Table 5.2: Resistance-dominance Data — binary classification results. Table reports results of experiments with two groups of features: Dominance Ranks (DR) and Speaking probability, aggregated with Fisher Vector (FV) or Histograms, as well as DELF model. For Dominance Rank we use the LS/LL interaction function with timespan of 1 and 5 seconds. Details on DELF for each task are presented in Table 5.3. Baseline is adapted from Beyan et al. [22].

Feature	MPD-All			MDP-Distinct			PDP-All			PDP-Distinct		
	AUC	FPR	Acc.	AUC	FPR	Acc.	AUC	FPR	Acc.	AUC	FPR	Acc.
DELF	**0.791**	0.027	0.769	**0.894**	0.021	0.889	**0.874**	0.281	0.792	**0.949**	0.189	0.876
DR (LS/LL, 1 sec) + FV	0.754	0.056	0.761	0.855	0.017	0.89	0.77	0.281	0.694	0.832	0.235	0.741
DR (LS/LL, 1 sec) + Hist.	0.754	0.252	0.711	0.836	0.209	0.868	0.788	0.314	0.724	0.861	0.392	0.768
DR (LS/LL, 5 sec) + FV	**0.773**	0.064	0.761	**0.861**	0.167	0.868	0.771	0.328	0.695	0.835	0.28	0.74
DR (LS/LL, 5 sec) + Hist.	0.77	0.252	0.720	0.844	0.179	0.879	0.793	0.441	0.709	0.861	0.347	0.788
Speaking + FV	0.741	0.279	0.689	0.838	0.03	0.875	**0.853**	0.261	0.762	**0.92**	0.179	0.825
Speaking + Hist.	0.756	0.066	0.77	0.821	0.15	0.879	0.847	0.258	0.778	0.91	0.164	0.86
Baseline (speak.)	0.738	0.103	0.73	0.769	0.2	0.879	0.8	0.274	0.738	0.893	0.198	0.845
Baseline (comb.)	0.767	0.252	0.716	0.764	0.214	0.879	0.828	0.29	0.759	0.906	0.168	0.863

Section 5.4

Experiments on Resistance-dominance data

This section is organized as follows. We first show the results of applying DELF and single-ltf classifiers to four binary classifications tasks. We then show the results of an ablation study to determine the most important feature for our ensemble model. We also provide an analysis of how video length affects the predictive performance of models based on our proposed features. We examine the performance of two other choices for the Dominance Rank Interaction function. Finally, we demonstrate the performance of GDP algorithm.

Setup. We split the Resistance-dominance dataset into 10 folds by games. As each player appears in only one game, we always make predictions about the dominance of players in games that we have not seen before. Our classifier suite for binary prediction tasks consists of the 5 classifiers: k-Nearest Neighbors, Logistic Regression, Gaussian Naive Bayes, Linear SVM, and Random Forest. The tables below report the best results among these classifiers. Since our Resistance-dominance dataset is inherently imbalanced, we report the mean AUC over 10 folds and use it to compare models.

Excluded Feature	AUC
MDP-All	
All features present	0.790
FAU (AU15, AU20, AU25)	0.790
MFCC	0.775
DR (LS/LL, 5sec) + FV	**0.757**
Emotions (Angry, Surprised, Calm)	0.772
Speaking+Hist.	0.775
MDP-Distinct	
All features present	0.894
FAU (AU05, AU14, AU20)	0.888
MFCC	0.890
DR (LS/LL, 5sec) + FV	**0.849**
Emotions (Angry, Confused)	0.891
Speaking+FV	0.884
PDP-All	
All features present	0.874
FAU (AU15, AU20, AU25)	0.824
MFCC	0.867
DR (LS/LL, 5sec) + Hist.	0.866
Emotions (Smile, Angry, Surprised)	0.866
Speaking+ FV	**0.816**
PDP-Distinct	
All features present	0.949
FAU (AU14, AU15, AU25)	0.948
MFCC	**0.921**
DR (LS/LL, 1sec) + Hist.	0.934
Emotions (Happy, Angry, Calm)	0.945
Speaking + FV	0.949

Table 5.3: DELF ablation study. For every task we report the late fusion AUC. To assess the importance of every feature type, we exclude one feature type at a time and examine the AUC of DELF for the remaining feature types.

But we also report False Positive rate (FPR) and Accuracy (Acc.) as reported in past works ([22, 125, 108, 109, 4]).

5.4.1. Binary prediction with DELF

Table 5.2 shows the result of applying DELF to the four problems as well as single-ltf classifiers. We compare our models with two baselines adapted from the recent paper by Beyan et al. [22]: one model uses speaking features such as total number of speaking turns or number of times a player gets interrupted, the other model combines speaking features with gazing features such as number of times a player looks at other players.

DELF produces the best AUCs in all four tasks outperforming both baselines and out singe-ltf classifiers.

For each task, a single-ltf classifier (Dominance Rank or speaking-based feature) outperforms the baselines. In most cases, the improvement in AUC comes with reduced FPR and better accuracy than the baselines. We can see that Dominance Rank features prove to be more useful in the MDP task, while speaking-based features produce the highest AUCs on PDP among single-ltf features. We believe this happens because speaking-based features capture individual behavior of the player thus making it easier to compare two players, while Dominance Rank features capture the overall dynamics of the interaction of a player with all other players, which is useful for the most dominant player detection but introduces noise for pairwise comparison. Additionally, we found that features exclusively based on gazing information produce fairly poor results (not reported in the Table 5.2), which holds both for our features and baseline features.

We further note that "nice" instances of problems (MDP-Distinct and PDP-Distinct) are easier and get higher results, because difference in dominance between players in these cases is more prominent.

Ablation study. To assess the importance of each group of features used in DELF, we exclude features one at a time and perform another late fusion on the reduced group of features We see from Table 5.3 that DR features prove to be important for identifying the most dominant player — both for MDP-All and MDP-Distinct. For PDP-All and PDP-Distinct most value is provided by speaking-based features and MFCC respectively.

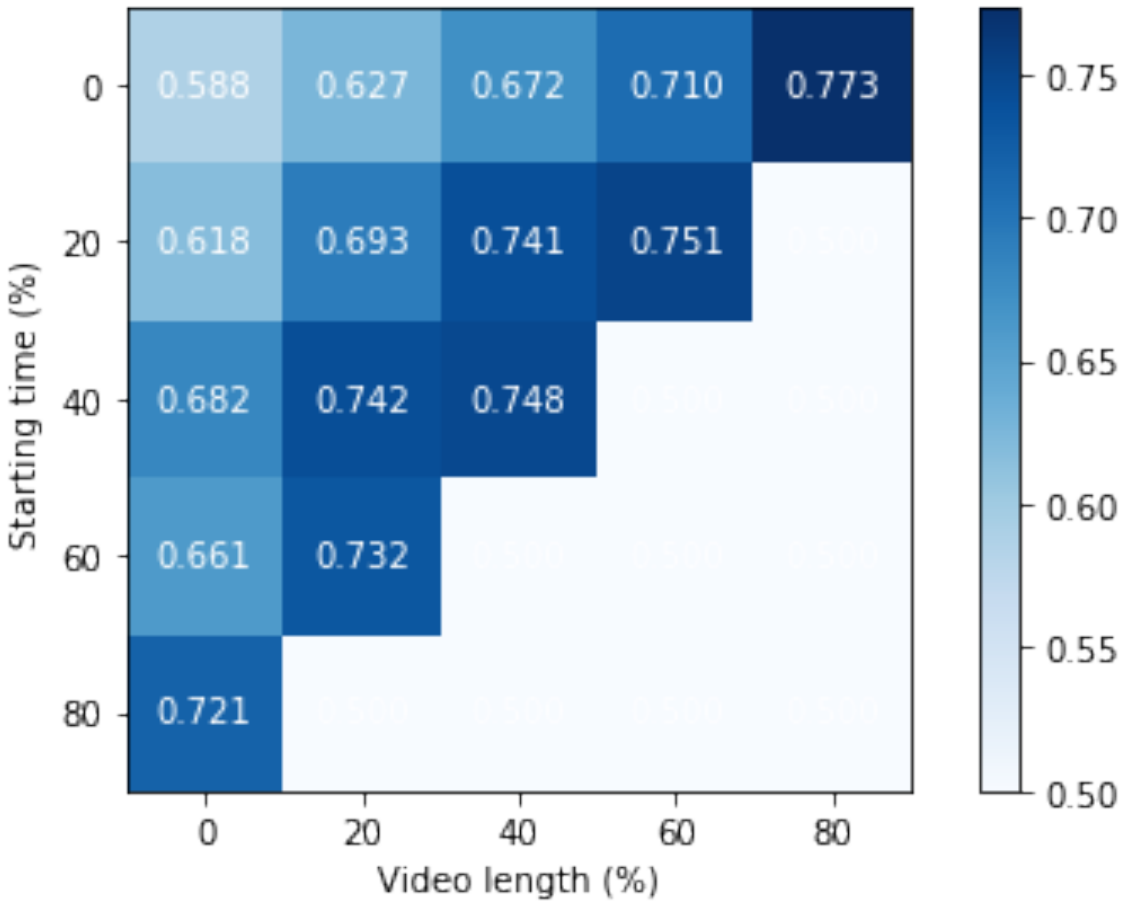


Figure 5.3: MDP-All: performance depending on the length of the video. For the best performing long-term feature (LS/LL, 5 sec + FV) AUC for the entire video is the highest, and for smaller portions of the video predictive performance drops. For any length of the video, parts closer to the end yield better AUC.

Predictions of our models depend on how large part of the game is considered and what part of the game is considered. In the Figure 5.3 we show how the LS/LL

Dominance Rank feature (best performing feature in Table 5.2) performs on MDP-All task when we process only a portion of each video from 20% to 100%. We also vary what portion of the video is processed. We found that considering only 20% of the video drops the predictive performance of our models for up to 0.2. Performance grows with increased video length reaching the highest result for the entire video. For the same length, however, it is usually advantageous to consider parts closer to the end of the game. The last 20% of the video sometimes can yield the performance very close to the classifiers trained on the entire video. We attribute this finding to the fact that ground truth labels used in our work are based on players' assessment of each other, which is collected after every two rounds of the game, and people tend to remember the most recent events better. Analysis shows, however, that for the best performance it's important to consider entire videos.

Interaction functions. In addition to Dominance Rank features w.r.t. the interaction function LS/LL, we examined two more interaction functions: $LL(p_j, p_i) - LL(p_i, p_j)$ and $LS(p_i, p_j) - LS(p_j, p_i)$. These functions show relatively high correlation with ground truth dominance scores (Table 5.1). For MDP-All these features yield the AUC of 0.755 and 0.748 respectively, showing the results close to the best single-ltf classifier in Table 5.2. For MDP-Distinct the AUCs are 0.795 and 0.847 respectively, which is higher than the corresponding baselines and on-par with the best single-ltf classifiers.

5.4.2. GDP algorithm performance

We tested GDP algorithm on the Resistance-dominance dataset. We used two classifiers: Multilayer Perceptron (MLP) with two layers, and Random Forest (RF) with 50 estimators. As shown in Table 5.4, GDP outperforms all the baselines as well as the various strong settings of DELF.

Table 5.4: GDP algorithm results. GDP in most cases improves over the corresponding single ltf binary prediction, as well as outperforming best DELF model.

Feature	Classif.	**AUC**	**FPR**	**Acc.**
MDP-All				
Speaking + FV	MLP	0.809	0.219	0.745
Speaking + FV	RF	**0.817**	0.133	0.77
DR (LS/LL, 5sec) + FV	MLP	0.783	0.222	0.733
DR (LS/LL, 5sec) + Hist.	MLP	0.772	0.157	0.746
MDP-Distinct				
Speaking + FV	MLP	**0.936**	0.048	0.917
Speaking + FV	RF	0.902	0.088	0.849
DR (LS/LL, 5sec) + FV	RF	0.878	0.071	0.878
DR (LS/LL, 5sec) + FV	MLP	0.85	0.065	0.889

Section 5.5

ELEA corpus experiments

We conducted further tests on the ELEA dataset [125] which is a widely used benchmark for modeling and detecting personal traits such as leadership and dominance. We use speaking and gazing labels provided with the dataset to produce Dominance Rank features. Every participant in the dataset has two dominance scores: perceived dominance (PDom) and ranked dominance (Rdom). We followed two protocols: (1) as in [109, 4, 108] we assign every participant a binary label by thresholding her dominance score by the median value, and (2) as in [125] we solve the MDP-All task, i.e., finding the most dominant participant in every group. As in these works we perform leave-one-game-out strategy when training and evaluating classifiers. Table 5.5 shows that our proposed Dominance Rank feature outperforms strong baselines in existing work.

We used the average dominance scores assigned to participants by the independent viewers not participating in the task as *human scores*. In Table 5.5 we can see that our proposed features outperform humans on the task of detecting participants who

Method	PDom	RDom
[109]	58.82	64.71
[4]	65.69	59.80
[108]	67.65	68.63
DR (LS/LL) + FV (ours)	**76.47**	67.65
DR (LS/LL) + Hist. (ours)	74.51	**71.57**
Human scores	68.63	—
Sanchez-Cortes et al. [125]	74.10	77.80
DR (LS/LL) + FV (ours)	**77.50**	**78.40**
DR (LS/LL) + Hist. (ours)	76.50	76.50
Human scores	**78.43**	—

Table 5.5: ELEA corpus experiments. Accuracy reported for the detection of dominant participants. Dominance is defined based on ranks (RDom) or scores (PDom). In rows 1–6 the median score is used as a threshold to assign labels, therefore random guess accuracy is close to 50%. Rows 7–10 report accuracy for MDP-All task.

are more dominant than others. Humans, however, are better at detecting the most dominant participant in a group, although our model achieves comparable accuracy.

Section 5.6

Conclusion and future work

We study two major problems: predicting the most dominant person in a group setting, as well as the more dominant of a pair of people. We develop a novel family of Dominance Rank features and develop two algorithms for these problems. The DELF algorithm uses past features (plus facial action unit, emotion, and MFCC features not previously used in dominance prediction), as well as dominance ranks combined with a late fusion approach and beats out past work in predictive accuracy — an ablation study additionally shows the Dominance Rank features to be the most important ones. The GDP algorithm proposes a way to expand and augment the dataset while retaining the group information. It beats out both past work and DELF on two tasks. But we note that both DELF and GDP use many well-known features from the past

literature to achieve these high AUCs.

One potential future work is the improvement of the GDP algorithm. When generating subsets of 5 players, a player in a larger group will appear in more subsets, which will cause the unbalance of training data. One can resolve this by subsampling the subsets or weighting the samples during training.

Chapter 6

Predicting Relative Nervousness from Group Interaction Videos

Given a video of a group of interacting humans, we solve three problems: (i) The Pairwise Nervousness Prediction (PNP) problem predicts if person A is more nervous than B even if the ground truth nervousness ratings for A and B are very close. (ii) The PNP-Distinct problem predicts if A is more nervous than B when the ground truth nervousness ratings for A and B differ by at least 1 on a 5 point scale. (iii) The Nervousness Change Prediction (NCP) problem predicts if A's nervousness rating increases or decreases compared to his previous rating in the same video. Compared to past work that looked at using emotions, facial action units (FAUs), and facial movements, we make two new contributions: (i) As social science theory suggests that A might be more nervous than B if B is more dominant, we define a new class of features called nervousness scores (NSs) from the audio-visual channel. NSs use dominance relationships between people, as well as gaze (who is looking at who), and speaker (who is speaking) information. (ii) We develop a novel Facial Emotion Graph Convolution Network (FE-GCN) together with an ensemble prediction architecture. Our results show that: (i) either NSs or FE-GCN generate the best performance in

head to head comparisons with five baselines based on past work, (ii) an ensemble that merges NSs and FE-GCN provides high quality results in terms of both F1-score and AUC compared to the five baselines, and (iii) the learned FE-GCN identifies landmarks that are highly relevant for nervousness prediction.

Section 6.1

Introduction

The ability to detect nervousness in group interactions has many applications. In a business negotiation, knowing that one party is nervous may suggest a negotiation strategy to the other party. In a social setting, understanding that one person is nervous may enable others to put that person at greater ease. Security agencies may use nervousness as a cue to determine whether a subject is suspicious.

In this chapter, we study the problem of predicting pairwise nervousness (Is person A more nervous than B? Is a person more nervous compared to his rating before?) in a social setting where subjects are involved in a group interaction (e.g. game or discussion). Past work by psychologists has suggested that nervousness is linked to the setting (e.g. is A speaking in public?) [137], the response of others (e.g. is person A listening to person B while expressing certain facial/body gestures?) [137], and dominance (what is the relative dominance of A relative to B?) [99]. We study this problem in 3 settings: (i) the Pairwise Nervousness Prediction (PNP) problem looks at all pairs (A, B) of people even if their nervousness ratings are near identical, (ii) the PNP-Distinct problem looks at all pairs of people where either A or B is clearly more nervous by a margin of 1 or more on a 5-point rating scale, and (iii) the Nervousness Change Prediction (NCP) problem, i.e. how a person's nervousness changes (increases/decreases) over time.

We leverage social science research, together with past work on emotion predic-

tion [93] and dominance prediction [13] to make several contributions. Our first contribution is the definition of a family of 54 new features called *nervousness scores (NSs)* that combine social science theory with interactions between people in the video. The NSs can be obtained from both audio and visual behaviors, which we call *ANS*s and *VNS*s respectively. Our second contribution is the development of a novel Facial Emotion based Graph Convolution Network (FE-GCN) that combines Graph Convolutional Networks (GCN) [82] and Convolutional Neural Networks (CNN) to generate facial embeddings based on facial landmarks. Unlike past work on GCNs and CNNs that require huge amounts of data for training, our FE-GCN can be trained even on the modest amount of data we have. We first predict nervousness using the ANS/VNS features and FE-GCN, and then combine these predictions using a late fusion step to generate results that have both high AUCs and F1-scores. We evaluate our methods on two datasets: the Resistance social game data from [13] and the ELEA "winter survival task" dataset from [125]. To better understand NS features, we also explore how different emotion categories (positive vs. negative) in NSs and different video content can influence the prediction performance.

Our experiments show that: (i) in head to head comparisons, one of our new techniques, i.e. ANS/VNS/FE-GCN yields the best results on all three problems beating five baselines and (ii) an ensemble that combines our ANS/VNS/FE-GCN techniques generates the overall best result, and (iii) the trained FE-GCN can identify the importance of facial landmarks that are relevant to nervousness.

Section 6.2

Dataset and tasks

We use two very different datasets in this chapter. The first one, Resistance-nervousness, is a subset of the Resistance data (Section 3.1) consisting of 25 videos involving 178

Dataset	# Games	# Players	Kendall W
Resistance	80	564	0.150
Resistance-1	16	112	0.301
Resistance-2	15	104	0.305
ELEA	27	102	0.332
ELEA-1	10	52	0.612
ELEA-2	9	46	0.640
ELEA-3	8	40	0.664

Table 6.1: Datasets statistics and inter-annotator agreements using Kendall's W coefficient. Note that the first rows of and ELEA are original datasets.

people, which contains complete nervousness ratings among players in each group. We use the median rating for each player as the ground truth perceived nervousness rating of the player in a round. The dataset contains 62 rounds in all. The average length of a round is 13.3 minutes. Second, we annotate the ELEA data (Section 3.2), and use the median rating as the perceived nervousness rating.

We assess inter-annotator agreement via the Kendall's W metric which turned out to be 0.15 for and 0.332 for ELEA. The agreement is low as judgment about nervousness is subjective. We therefore also created multiple subsets of both datasets where the inter-annotator agreement is higher (a Kendall W of at least 0.3 for and 0.6 for ELEA) and where a sufficiently large number of players is available to train and test on. Table 6.1 summarizes the two original datasets and five subsets with higher inter-annotator agreement.

Pairwise Nervousness Prediction (PNP) The PNP problem takes two players in a game (or one round in the Resistance dataset) as input and predicts which one has the higher nervousness rating. To pose this as a binary classification problem, we discard pairs with equal ratings. The PNP-Distinct problem predicts which player in a pair is more nervous when the nervousness ratings of the players differ by more than 1.

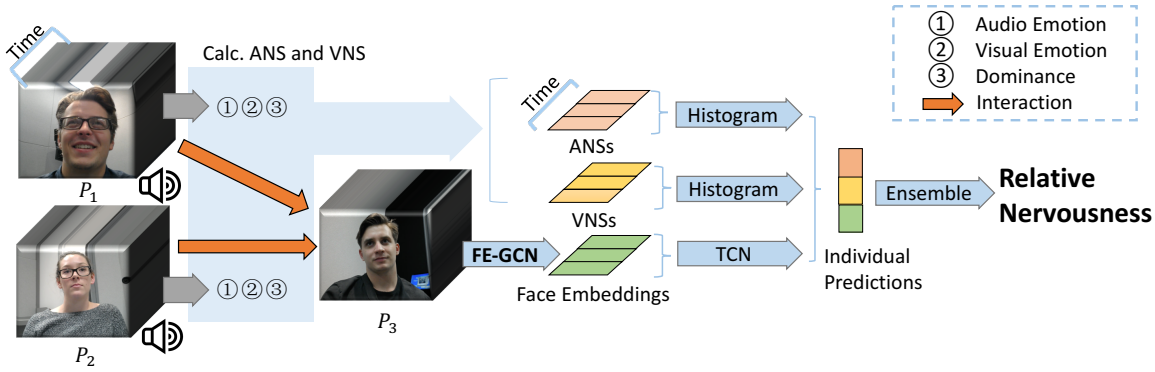


Figure 6.1: Nervousness Prediction Architecture (NPA) overview. Given input audio and video clips of several people (three in this example) interacting with each other, we first extract pairwise interactions (looking at each other, speaking), audio emotions, visual emotions, and dominance. To predict how nervous person P_3 is, we: (i) compute the audio nervousness scores (ANSs) and video nervousness scores (VNSs) from the interactions between P_1, P_2 and P_3 (details in Section 6.3.1), (ii) compute P_3's face embeddings using our FE-GCN methods (details in Section 6.3.2), and (iii) aggregate each of these 3 feature types over time and get three individual predictions and (iv) use an ensemble of these three models to get the final prediction. In pairwise tasks PNP and PNP-Distinct, for each kind of feature, we concatenate aggregated feature vectors of two players before making individual predictions.

Nervousness Change Prediction (NCP) Because multiple questionnaires can be filled out in the Resistance data (but not in the ELEA data), we also consider how subjects' nervousness levels change over time as the game proceeds. We compare the nervousness ratings of a player in two consecutive rounds and solve the binary classification problem of whether the rating increases or decreases from one round to the next. To keep the problem binary, we discard samples with the same rating in two consecutive rounds.

Section 6.3

Nervousness prediction architecture

In this section, we present the details of the Nervousness Prediction Architecture (NPA) shown in Figure 6.1. NPA takes a raw video clip v as input and extracts: (i)

speaker information (probability of each person being the speaker), (ii) gaze information (probability distribution of who each person is looking at - this could include "nobody" as an option) using the code from [13], (iii) four audio emotions (angry, sad, happy, neutral) extracted using [36] from the clips where speaking is predicted, and (iv) seven visual emotions (anger, disgust, fear, happy, sad, surprise, neutral) using the techniques and code from [93]. Note that these features are estimated for every Δt seconds ($\Delta t = 1$).

We use (i)–(iv) above together with results from social science linking emotions with nervousness [104], dominance with nervousness [48], and linking being the focus of visual attention with nervousness [137] in order to define a new class of features called *nervousness score (NS)*. Specifically, for each Δt time window, we have a set of video nervousness scores (or VNSs) for all video clips and audio nervousness scores (or ANSs) in any audio clips predicted to contain speech. For each short time window Δt, each NS feature f has a value $v_{f,t}$. Therefore, for a given video of length T (i.e. with time points $1, \ldots, T$, we have values $v_{f,1}, \ldots, v_{f,T}$ for a feature f. As different videos may vary in length, we generate a *histogram* (i.e. probability distribution) over feature values $v_{f,1}, \ldots, v_{f,T}$ in order to generate a fixed-length vector for f. Thus, each feature f has an associated histogram vector and the feature vector associated with a given video clip is the concatenation of the different histogram vectors.

Recent years have witnessed huge advances in the discovery of embedding features using neural nets [82, 145]. We therefore develop a novel model called Facial Emotion Graph Convolution Network (FE-GCN) to extract facial embedding features by combining GCNs with Temporal Convolutional Networks (TCN) [17] to generate predictions based on highly representative automatically learned features. These face embeddings intuitively capture far more complex facial features than hand-crafted ones [114, 63]. The advantage is that embeddings may capture low-level features that

humans may not imagine.

To solve the pairwise tasks (PNP and PNP-Distinct), we concatenate the feature vectors of two players. These concatenated feature vectors are then fed to one of the standard classifiers (in the case of baseline features and Nervousness Score features) or to the final layer (in case of FE-GCN + TCN).

Our Nervousness Prediction Architecture combines these three predictions to generate a final prediction using late fusion. Because the main novelty of our work involves: (i) Nervousness scores/ranks and (ii) FE-GCN and its combination with Temporal CNNs, the remainder of this section focuses on these components.

6.3.1. Nervousness Scores

In this section, we first define nervousness scores on videos and then extend it naturally to speech audios.

Visual Nervousness scores. Messenger et al. [104] and Stein et al. [137] state that people are more nervous when negative attitudes are directed toward them. They are less nervous when others exhibit positive attitudes toward them. Therefore, when person u interacts with person v, we can use the facial expressions of u as a proxy for u's attitude towards v. Of the 7 visual emotions we extract, we use $PE = \{Happy\}$ and $NE = \{Sad, Angry, Disgusted, Surprised, Fearful\}$ (excluding "neutral") as the set of positive and negative emotions respectively. Given a short time window Δt, $s_t(u)$ is the probability of person u speaking and $p_{e,t}(u)$ is the probability of person u showing the emotion $e \in PE \cup NE$.

We capture the interaction between u and v during time window Δt with 3 forms of interaction functions $i_t(u, v)$:

- $g_t(u, v)$: probability of person u gazing at person v,
- $ls_t(u, v) = s_t(u)g_t(u, v)$: probability of person u looking at v while u is speaking,

- $ll_t(u, v) = s_t(v)g_t(u, v)$: probability of person u looking at v while v is speaking.

Social science theory [99] suggests that high dominance is linked with low nervousness and vice versa. Let $dom_t(u)$ represent the dominance of u during time window Δt. We consider the relative dominance of u w.r.t. v in time window Δt as $Rdom_t(u, v)$ in three possible forms:

- $Rdom_t(u, v) = dom_t(u)/dom_t(v)$. Here, the relative dominance of u w.r.t. v is the ratio of u's dominance to v's. Social science theory suggests that if this ratio is small, then v should not be nervous, while u would be nervous (and vice versa if this ratio is large).

- $Rdom_t(u, v) = dom_t(u)$. Here, we suggest that in an interaction between u and v, only u's dominance plays a role in v's nervousness.

- $Rdom_t(u, v) = 1/dom_t(v)$. This suggests that only v's dominance plays a role in v's behavior.

We consider two possible ways to define $dom_t(u)$. First, we can use the human-rated dominance score $ds(u)$, which is constant over two rounds in the Resistance data and over the whole game in the ELEA data. An inspection of the Resistance game and ELEA data showed negative Pearson Correlation Coefficients of -0.51 and -0.12 respectively ($p < 0.05$) between the human-rated dominance score of a player and nervousness rating. That said, using only dominance scores as features yields poor performance for nervousness prediction. However, as dominance can be dynamic and emergent [125], the use of $ds(u)$ to represent $dom_t(u)$ may not be sufficient. Moreover, a resulting system would not be an end-to-end automated system as it would require human input during the processing. We therefore rejected this option.

Instead, our second option builds upon the notion of Dominance Rank from [13]. Dominance Rank is a class of features which identify the relative dominance of each

person in a group by dynamic human interactions. Given a interaction function $I(u, v)$ between people u, v in a group of N, the dominance rank of u, $dr(u)$, is recursively defined as $\frac{1-d}{N} + d \sum_{v \neq u} \frac{dr(v)\dot{I}(u,v)}{N-1}$ ([13]). It is a form of PageRank weighted by $I(u, v)$. Various types of interaction function $I(u, v)$ are explored by [13]. Different interaction functions capture things such as the probability that u is looking at v while u is speaking, the probability that v is looking at u while u is speaking, and so forth.

We use $dr_t(u)$ to denote the dominance rank of u in the group during time Δt — we normalize it to ensure $\sum_u dr_t(u) = 1$. Specifically, we use two forms of dominance rank features $dr_t(u; LL)$ and $dr_t(u; LSLL)$ whose mean values over time are positively correlated with $ds(u)$ (cf. [13]). LL and $LSLL$ are two interaction functions. $LL(u, v) = ll(v, u) - ll(u, v)$ represents the relative looking-while-listen difference, and $LSLL(u, v) = ls(u, v)/ll(u, v)$ is the ratio between looking-while-speaking and looking-while-listening (cf. [13]).

We now combine the visual attitudinal information (positive/negative) of a person u towards a person v with the relative dominance and interaction between u and v by defining a class $NS_t(v)$ of visual nervousness scores of a person v as follows:

$$NS_t(v) = \alpha NS_{pos,t}(v) + (1 - \alpha) NS_{neg,t}(v) \ , \tag{6.1}$$

where α denotes the balance between the positive and negative attitudes of people interacting with v ($0 \leq \alpha \leq 1$),

$$NS_{pos,t}(v) = \frac{\sum_{e \in PE, u \neq v} Rdom_t(u, v) i_t(u, v) \left(1 - p_{e,t}(u)\right)}{|PE| \sum_{u \neq v} i_t(u, v)} \ ,$$

$$NS_{neg,t}(\frac{\sum_{e \in NE, u \neq v} Rdom_t(u, v) i_t(u, v) p_{e,t}(u)}{|NE| \sum_{u \neq v} i_t(u, v)} \ .$$

Intuitively, $NS_{pos,t}(v)$ summarizes the positive attitudes expressed by other peo-

Functions	Forms
$i_t(u,v)$	$g_t(u,v)$, $ls_t(u,v)$, $ll_t(u,v)$
$dom_t(u)$	$ds_t(u)$, $dr_t(u;LL)$, $dr_t(u;LSLL)$
$Rdom_t(u,v)$	$dom_t(u)$, $1/dom_t(v)$, $dom_t(u)/dom_t(v)$

Table 6.2: Different forms of nervousness scores. Note that $dr_t(u;LL)$ and $dr_t(u;LSLL)$ are two types of dominance ranks [13] which are positively correlated with the dominance score $ds(u)$.

ple toward v based on their interactions and relevance dominance, while $NS_{neg,t}(v)$ summarizes the negative ones.

Table 6.2 summarizes all the possible forms of functions in the definition of nervousness scores. Overall, we have 27 visual nervousness score (VNS) features.

Audio Nervousness Scores. Audio Nervousness Scores are computed in a similar manner. We let $PE = \{Happy\}$ be the set of positive audio emotions and $NE = \{Sad, Angry\}$ be the set of negative audio emotions. The three audio emotions are the most common emotional descriptors found in the literature ([29, 30]). Given a short speech clip Δt, we calculate the ANSs from Equation 6.1 by replacing the visual emotions with audio emotions thus obtaining 27 forms of ANSs.

6.3.2. FE-GCN

Since nervousness is a complex emotion/expression whose visual manifestation can vary dramatically from person to person, we explore the possibility of learning such models using embeddings. CNNs have recently been used to learn facial embeddings from images [113, 142] and videos [152, 35]. However, as convolutions only process local neighborhoods, they have to be stacked repeatedly (to create deep CNNs) in order to get non-local summaries of faces. Unfortunately, it is not feasible to learn such embeddings from the limited volume of data in our two nervousness datasets.

To solve these challenges, we propose a lightweight model called Facial Emotion-oriented Graph Convolution Networks (FE-GCN), to learn a facial emotion oriented

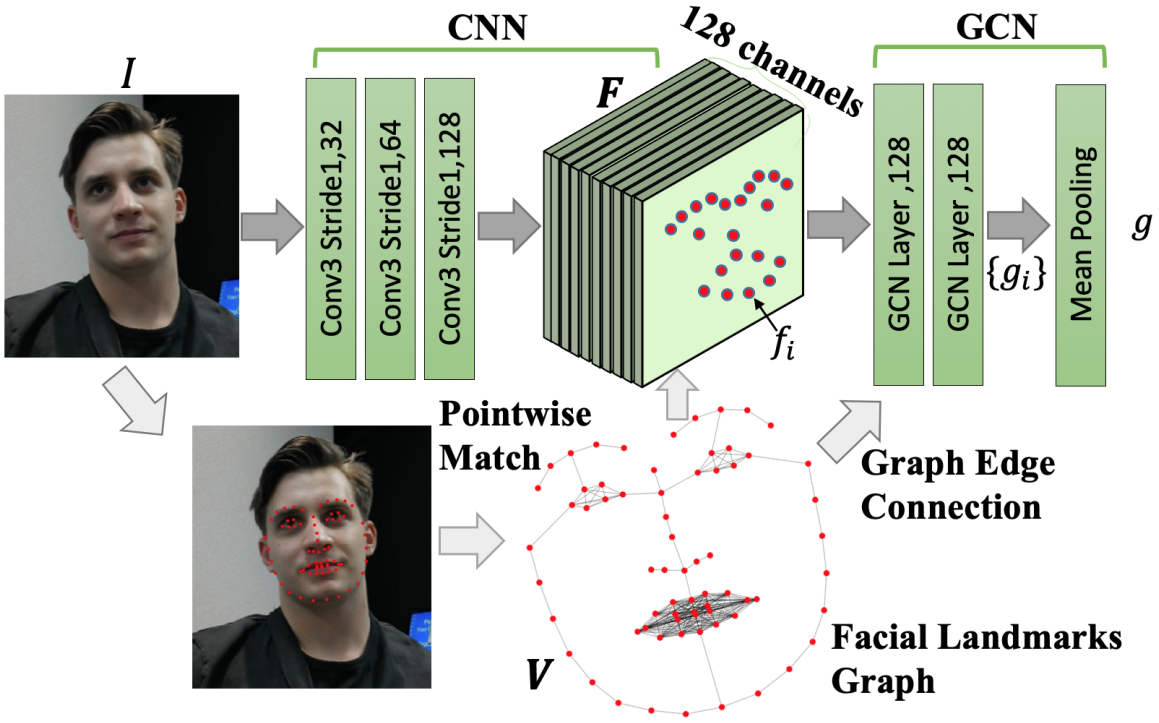


Figure 6.2: **FE-GCN Structure.** Given a face image I, the CNN layers extract local features and obtain the spatial size of output feature map $\boldsymbol{F}$. We use the locations of 68 landmarks and the feature map $\boldsymbol{F}$ to build a facial landmark graph. This graph is then fed into GCN layers to learn node representations $\{\boldsymbol{g_i}\}$ and apply mean pooling to get the final embedding $\boldsymbol{g}$.

embedding given a face image. Graph Convolution Networks (GCN) [82] are non-local networks that have significantly improved the quality of many prediction models, e.g. multi-label image recognition [37] and person re-identification [94]. Our FE-GCN model uses GCNs to learn long-range dependencies in a face from the graph built from facial landmarks. The local features of facial landmarks are learned by using a shallow CNN. *To the best of our knowledge, we are the first to leverage GCNs to learn facial embeddings.*

In order to learn representative embeddings, we first pre-train the FE-GCN on the SFEW2.0 dataset [47] consisting of high-resolution images depicting seven basic emotions. To fine-tune the model for our nervousness video datasets, we aggregate

the embeddings over time through the use of a lightweight Temporal Convolutional Network (TCN) [17]. FE-GCN is a lightweight model which only has 120K parameters compared to 25M in ResNet50 and 134M in VGGFace. Despite this, FE-GCN outperforms the VGGFace architecture on emotion recognition tasks on SFEW2.0. As emotion prediction is not a goal of this chapter, we omit the details.

Figure 6.2 shows our FE-GCN architecture. Given an input image $I \in \mathcal{R}^{H \times W \times C}$ where $H, W,$ and C denote height, width and channels resp.:

(a) We extract the set $V = \{V_i\} = \{(x_i, y_i)\}, i = 1 \dots N$ of I's facial landmarks (68 in all) using OpenFace [19] and build a facial landmark graph through the pre-defined weighted and undirected edges (as shown in Fig. 6.2). The (x_i, y_i)'s shown are coordinates in the image space. First, we create the following connected components: face profile, eyes, eyebrows, nose and mouth. Second, we add component-wise edges: eye-eyebrow, eye-profile, eye-nose, nose-mouth, and mouth-profile. Each edge is the shortest among all possible connections between components. The edges enable effective message passing between landmarks. Edge weights are set as the Euclidean distances between facial landmarks.

(b) The input I is then passed to three convolutional layers with kernel size 3×3, stride and padding 1. These layers preserve the spatial size of I and extract a low-level feature map $\boldsymbol{F} \in \mathcal{R}^{H \times W \times C'}$, which is used for point-wise matching with facial landmark coordinates. $C' = 128$ is the number of channels of $\boldsymbol{F}$.

(c) For each node V_i, we get the embedding $\boldsymbol{f}_i = \boldsymbol{F}(x_i, y_i) \in \mathcal{R}^{C'}$. We then feed the graph into two GCN layers to learn a non-local node representation $\boldsymbol{g}_i, \forall i$ from the global face structure.

(d) Finally, mean pooling is applied to get the final face embedding $\boldsymbol{g}$, i.e., $\boldsymbol{g} = \frac{1}{N} \sum_i^N \boldsymbol{g}_i$.

T	60	**70**	80	90	100
AUC	0.756	**0.775**	0.768	0.751	0.769

Table 6.3: Prediction AUC on ELEA with different numbers T of sampled frames for TCN. We randomly choose 20% games for testing and the rest for training.

The CNN in FE-GCN captures latent local information around each landmark, while the GCN captures a global representation by coalescing features between neighbouring facial landmarks, thus combining all the local descriptions generated by CNN into a global embedding that captures long-range interactions between landmarks.

Pre-training Procedure. After the mean-pooling layer, we add an output layer to FE-GCN, which learns to predict probabilities of the 7 emotions mentioned above. We train our network on the SFEW2.0 training set, generating a model that performs best on the validation set.

Fine-Tuning Procedure. Once the FE-GCN model is pre-trained on SFEW2.0, we remove the output layer and extract a sequence of facial embeddings from a video. The facial embeddings are fed to the Temporal Convolutional Network (TCN) to get a sequence of output embeddings. We feed the last output embedding $\boldsymbol{g}'$ to a fully connected (FC) layer to learn a predictive model of nervousness for the three problems addressed in this chapter. *Note that for the PNP and PNP-distinct tasks, given a pair of people's videos* (i, j)*, the final outputs* $\boldsymbol{g}'^{(i)}$ *and* $\boldsymbol{g}'^{(j)}$ *are concatenated and fed to the FC layer.* We illustrate the configuration of TCN in Section 6.4.

Sampling long videos for TCN. Since our videos last for 300–800 seconds (with one face embedding per second), it is impractical to feed the entire sequences to TCN. During each training epoch, we sample T frames ($60 \leq T \leq 100$) uniformly at random as the input of TCN. The average prediction from multiple sampled sequences is used during the test stage. In practice, we find that $T = 70$ yields the best results (Table 6.3).

Section 6.4

Experimental results

6.4.1. Experiment setup

We split both datasets into 10 folds by games. As each player appears in only one game, we always make predictions about players never seen before. Since results vary for different classifiers and features, we report the best results from seven classifiers: k-Nearest Neighbor, Logistic Regression, Gaussian Naive Bayes, Linear SVM, and Random Forest. We test our proposed ANS and VNS feature types as well as 3 baseline feature types with each of these 5 classifiers. We also train and test the FE-GCN + TCN. We report the best results for each method in Table 6.4.

We compare our method with seven baselines that lie in two categories. All baselines are evaluated in the same setup as our methods.

FE-GCN and TCN configurations We use Batch Normalization [73] following each convolution layer, and dropout rate 0.5 in each GCN layer. ReLU activation is applied after each layer except for the output. We use 32, 64, and 128 channels respectively for the 3 convolutional layers. The node embedding dimensions of the two GCN layers are both 128 and the other hyper-parameters are the same as in [82]. We use the Adam optimizer [81] with default settings to train the cross-entropy loss. We pre-train the FE-GCN on SFEW2.0 for 100 epochs with a batch size of 96. We adopt 6 layers of TCN, with kernel sizes 3, dilation factors 2,4,8,16,32,64, and channel sizes $128, 128, 96, 96, 64, 64$ respectively. Under this setting, the receptive field of the output layer of the TCN is 127. Hence, the last timestep of the output sequence aggregates the whole input sequence.

6.4.2. Baselines

High-level features baselines extract features related to nervousness and use them to train classifiers. Specifically, Giannakakis et al. [63] extract facial cues from images including aperture change, blink rates, pupil variation, head movements, non-speaking mouth movements, and heart rate estimated from the facial color change frequency. Hung et al. [72] extract speaking cues such as numbers of interruptions and turns. Jayagopi et al. [76] consider non-verbal activities such as visual activity turns estimated from visual motion vectors and speaking turns estimated from audio signals. Bai et al. [13] consider facial expressions by computing histogram features of facial action units (FAUs) and emotions. All these features are fed into the same classifiers as our new methods and trained in the same manner.

Neural network baselines. We also consider two neural network baselines: VGGFace+TCN and ResNet50+TCN, which replace the FE-GCN with the state-of-art VGGFace [113] and ResNet50 [66] architectures respectively, and are both combined with TCN for nervousness prediction. The last dense layers of both networks are removed. For fairness, both VGGFace and ResNet50 are pre-trained and fine-tuned in the same way as FE-GCN. These two baselines serve as a direct evaluation of our FE-GCN embeddings.

We now report the results of eight sets of experiments. Experiments A–C provide detailed AUC and F1-score comparison among different methods. Experiments D-E explore the NS features: the impact of emotion categories in NSs (positive vs. negative), and the impact of different video content to compute NSs. Experiment F visualizes the relevant facial landmarks and faces learned by FE-GCN with regard to nervousness. Experiment G compares the importance of individual ANS and VNS features (cf. Table 6.2). Experiment H compares the prediction AUC in datasets with different annotation agreements (cf. Table 6.1).

6.4.3. Head to head feature comparisons

Table 6.4 summarizes the best results in terms of AUC and F1-score for our new approaches and the baselines. The results show that the new techniques introduced in this chapter provide the best performance on all tasks. *All improvements over baselines are statistically significant based on a t-test (p-val < 0.01).*

	Dataset	Resistance		Resistance		Resistance		ELEA	
	Task	PNP		PNP-Distinct		NCP		PNP	
	Methods	F1	AUC	F1	AUC	F1	AUC	F1	AUC
Ours	**ANSs + Hist.**	0.596	0.635	**0.746**	0.723	**0.688**	**0.724**	0.640	0.623
	VNSs + Hist.	0.624	0.668	0.733	**0.765**	0.568	0.667	0.622	0.760
	FE-GCN + TCN [17]	0.633	**0.681**	0.742	0.744	0.657	0.634	**0.710**	**0.802**
	Late Fusion	**0.678**	**0.703**	**0.773**	**0.807**	0.681	**0.733**	0.664	**0.813**
Baselines	Facial Cues [63]	0.520	0.535	0.531	0.469	0.420	0.587	0.568	0.580
	Speaking Cues [72]	0.526	0.532	0.538	0.598	0.521	0.573	0.596	0.603
	Non-verbal Activities [76]	0.561	0.584	0.585	0.612	0.594	0.607	0.589	0.610
	FAU + Hist. [13]	0.589	0.656	0.672	0.707	0.481	0.632	0.674	0.763
	Emotion + Hist. [13]	0.592	0.649	0.658	0.687	0.579	0.605	0.557	0.749
	ResNet50 [66] + TCN [17]	**0.676**	0.522	0.573	0.633	0.583	0.617	0.674	0.758
	VGGFace [113] + TCN [17]	0.590	0.621	0.696	0.695	0.599	0.511	0.650	0.724

Table 6.4: Nervousness prediction comparison. We compare the F1 scores and AUCs for all methods in all datasets and tasks. The top four lines represent our new methods and an ensemble of them, while the other seven lines present the baseline approaches. Note the underscored bold numbers are the best in each column, and bold-only numbers are the second best. In all cases, our best methods outperform all the baselines. All of these improvements are statistically significant via a Student t-test (p-val < 0.01).

(a) **PNP:** In this task (for the Resistance dataset) where two subjects may have very close nervousness ratings, FE-GCN yields the best performance with AUC=68.1% which beats the best baseline which has AUC= 65.6%. For F1-score, the best baseline yields the best F1-score of 67.6% (our best algorithm yields an F1-score of 63.3%).

(b) **PNP-Distinct:** Using VNSs on the Resistance data yields an AUC of 76.5% while ANS yields an F1 of 74.6%, which handily beat the best baselines which

have AUC=70.7% and F1= 69.6%. All of our methods beat all baselines in terms of AUC and F1-score.

(c) **NCP:** This could only be applied on the Resistance dataset. ANS features produced the best results with 72.4% AUC and 68.8% F1. Again, the best baselines only achieve 63.2% AUC and F1=59.9% .

(d) **ELEA:** On this data, FE-GCN yields the best results for the PNP task with AUC and F1-scores of 80.2% and 71% respectively. This approach outperforms the best baseline which has AUC=76.3% and F1=67.4%.

6.4.4. Ensemble prediction performance

Figure 6.1 shows that NPA generates an individual prediction based on each of the three types of features. These predictions are then combined using late fusion. If a binary prediction using one of the above three predictors returns class i with probability p_i, then we combine the predictions linearly as $\Sigma_{i=1}^{3} w_i p_i$ (where each $w_i \in [0, 1]$ and $\Sigma_{i=1}^{3} w_i = 1$) to compute an overall probability. We use a grid search over the space of possible values to find the best w_i's value. The best w_i learned on the training and validation sets are used in the predictions on the test set (so in particular, the test set was never used when computing the w's).

The result of late fusion is compared with all methods in Table 6.4. We see that our NPA architecture performs well overall — not surprisingly, it performs better on the PNP-Distinct Task than the other two tasks where differences might be very small.

6.4.5. Ablation study

We also performed ablation testing in which each of the three classes of features was removed one at a time in order to assess the importance of that class of features in

Excluded Method	F1	AUC
Resistance& PNP		
ANS + Hist.	0.695	0.701
VNS + Hist.	0.679	0.694
FE-GCN + TCN	**0.644**	**0.689**
Resistance& PNP-Distinct		
ANS + Hist.	0.776	0.795
VNS + Hist.	**0.696**	**0.780**
FE-GCN + TCN	0.763	0.790
Resistance& NCP		
ANS + Hist.	0.627	**0.667**
VNS + Hist.	0.626	0.701
FE-GCN + TCN	**0.614**	0.701
ELEA& PNP		
ANS + Hist.	0.711	0.802
VNS + Hist.	0.703	0.772
FE-GCN + TCN	**0.660**	**0.765**

Table 6.5: *Experiment C: Ablation study.* For each dataset and task, **we report the performance of the ensemble after excluding one of individual methods. We highlight the lowest F1 and AUC scores to indicate the most important method in each case.**

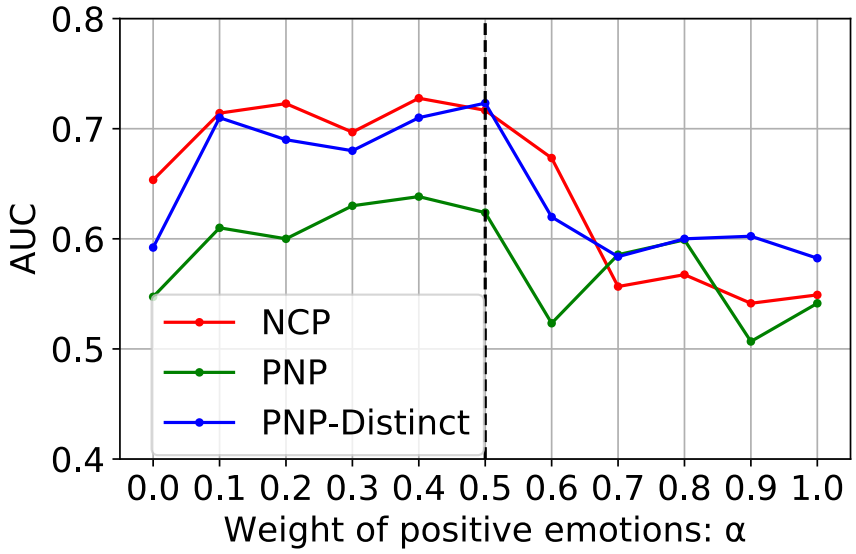
making the overall NPA ensemble prediction. Table 6.5 shows the result — note that the most important class of features causes predictive performance to drop the most, so the lowest numbers are the ones that indicate the most significant feature types. We observe that FE-GCN is the most important predictor in the PNP tasks for both the Resistance and ELEA datasets, while ANS/VNS features are the most significant ones in the other two tasks in terms of AUC.

6.4.6. Emotion impact for nervousness scores

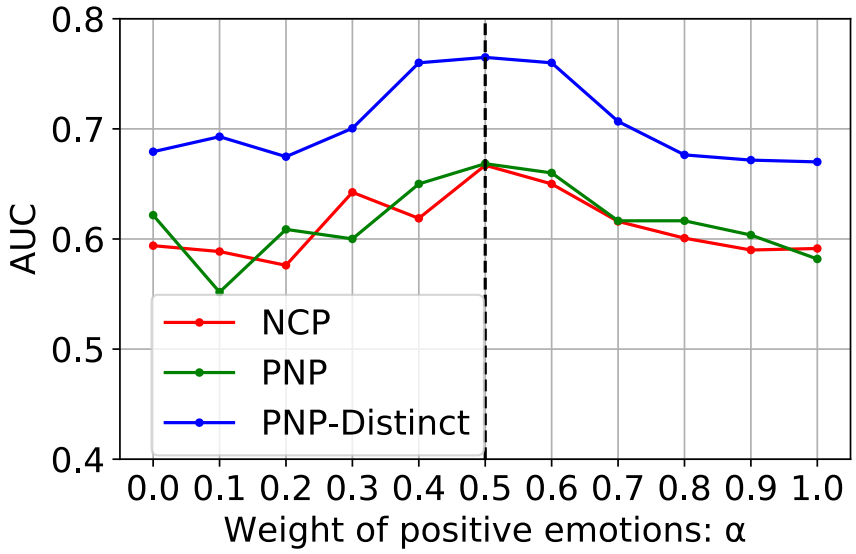
As shown in Equation 6.1, nervousness scores are a weighted sum of the influence from positive emotions and negative emotions. We vary the weight α for positive emotions from 0 to 1 with step 0.1, and evaluate the prediction AUC for all tasks in Resistance dataset. Figs. 6.3a and 6.3b show the results for ANS and VNS features respectively. For ANSs, we observe that the AUCs for small αs are higher than those for larger αs, indicating that the negative emotions (small α) are more important in speech audio for nervousness prediction. For VNSs, we see that the high AUCs are shown in the center of Fig. 6.3b, meaning that both types of emotions are needed for accurate nervousness prediction.

6.4.7. Change in prediction performance based on video start time and length

In this experiment, we vary both the start time and video length to explore the AUC of VNS features on PNP and PNP-Distinct tasks on the Resistance dataset. Figure 6.4 shows the relative change in performance compared to the performance of the models on the whole video. Figure 6.4(a) shows that for PNP, irrespective of where we start, we should make use of as much video as possible. Figure 6.4(b) suggests that using 60% of the video starting either at the beginning or after 20% of the video has elapsed or after 40% of the video has elapsed generates near optimal results.

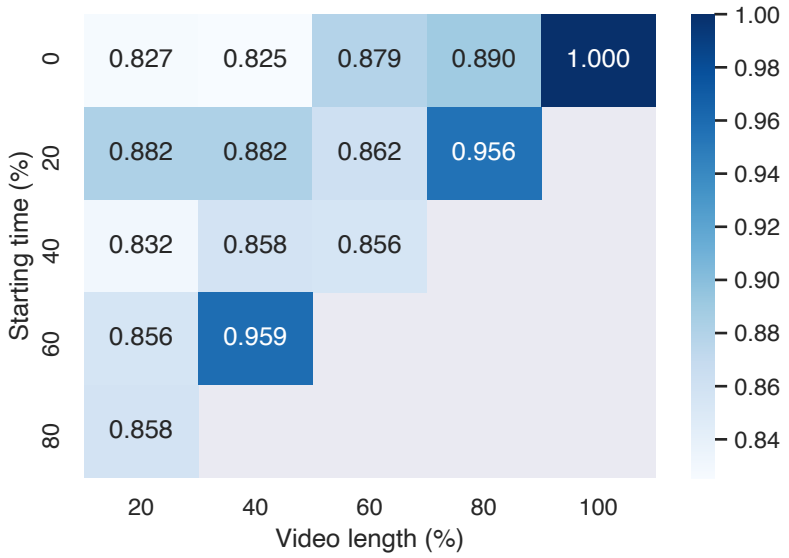


(a) ANS features.

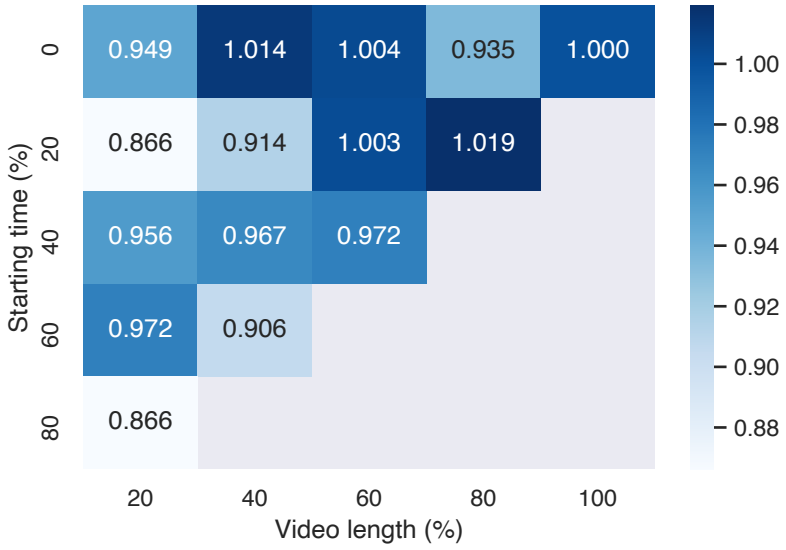


(b) VNS features.

Figure 6.3: *Experiment D: Emotion Impact for Nervousness Scores.* We vary α (cf. Equ. 6.1), the weight of positive emotions for nervousness scores. Figures show the best AUC on the three tasks with the Resistance dataset as α changes. For ANS features (top), the negative emotions are more important for predicting nervousness. For VNS features (bottom), both positive and negative emotions are needed to make an accurate prediction.



(a) PNP.



(b) PNP-Disdinct.

Figure 6.4: *Experiment E: Relative change in performance for two tasks on the Resistance data Using VNS features.* We vary the starting time and the length of the video clip used and assess how much performance changes compared to the whole video. Performance depends on which part of the video is used.

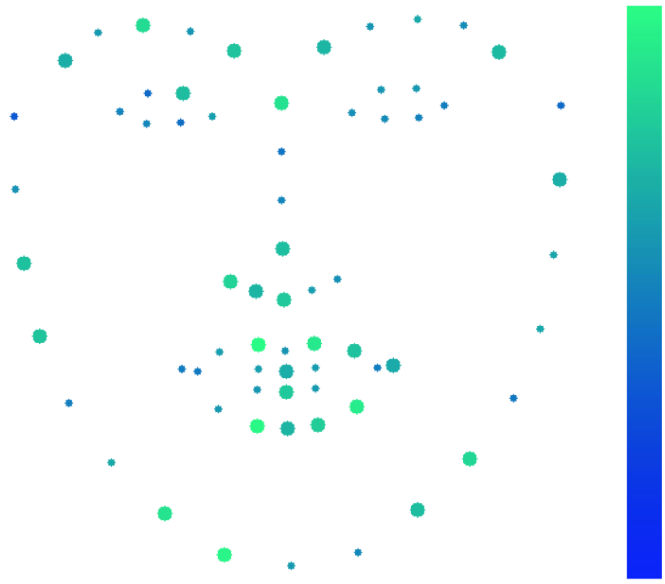


Figure 6.5: *Experiment F: Visualization of the Gradients of Facial Landmarks.* The lighter the color is and the bigger the point is, the bigger the gradients are.

Dataset **Task**	Resistance PNP	Resistance PNP-Distinct	Resistance NCP	ELEA PNP
First	V: $ls(u,v), ds(u), 1/dom(v)$	V: $ls(u,v), ds(u), 1/dom(v)$	A: $ll(u,v), ds(u), dom(u)/dom(v)$	V: $ll(u,v), ds(u), dom(u)/dom(v)$
Second	V: $g(u,v), ds(u), 1/dom(v)$	V: $ls(u,v), ds(u), dom(u)$	A: $g(u,v), ds(u), dom(u)$	V: $g(u,v), ds(u), 1/dom(v)$
Third	V: $g(u,v), ds(u), dom(u)$	V: $g(u,v), ds(u), 1/dom(v)$	A: $g(u,v), ds(u), 1/dom(v)$	V: $g(u,v), ds(u), dom(u)$

Table 6.6: *Experiment G: Importance of individual Nervousness Scores.* The top-3 important individual nervousness scores for each task are reported. A stands for ANS and V stands for VNS. The meaning for each function is defined in Section 6.3.1.

6.4.8. FE-GCN nervousness landmark visualization and face retrieval

We analyze the trained FE-GCN + TCN model by visualizing the importance of facial landmarks and retrieving the most relevant faces. In the PNP task (on Resistance data), we computed the gradients of the model output towards the facial landmark pixel intensities from all face images. Larger gradients indicate that changing the pixel intensities will influence the prediction output more [132, 136, 155], which suggests high relevance for nervousness prediction. We conduct two visualization experiments below.

(i) For each of the 68 facial landmarks, we computed the average L1 norm of the gradient vectors over all the images in the dataset. Figure 6.5 shows the heatmap of the average gradient L1 norms, where lighter colors and bigger points indicate larger gradient values. *We observe that the landmarks in the mouth-nose and chin regions are the most relevant for predicting nervousness.*

(ii) Next, for each video, we retrieve the top images sorted by the $L1$ norms of the landmark gradient vectors. The landmarks (and faces of selected images) are assumed to have large responses to nervousness. Figure 6.6 shows two sample pairs of players in two games. The numbers on the left show the ground truth perceived nervousness ratings (a higher score means more nervous). The color bars on the right show the norm of the gradient (the light, the better). We observe that faces with a rating of 4 usually don't smile (row 1, columns 2–4 and row 2), pout (row 3, column 3), or rest the chin (row 3, column 4), while faces with low ratings are usually happy (row 2) and relaxed (row 3, column 2–4). In addition, the heatmaps of landmark gradients vary dramatically between highly nervous and less nervous faces, indicating that different nervousness levels respond to different landmarks.

6.4.9. Importance of individual nervousness scores

As defined in Table 6.2, there are 54 types of individual video and audio nervousness scores (27 of each). However our experiments show that no single feature by itself yields good results: the best individual feature only yields 0.66 AUC. We therefore evaluate individual feature importance as follows. For a given prediction problem, we pick the top 10 results of all feature combinations for this task sorted by AUC, and use the frequency with which that feature appears as its importance. Table 6.6 shows the top-3 most important features for each task. We observe that (i) all types of interactions (look-at, look-while-listen and look-while-speak) are the most important, and (ii) individual visual features are more important than audio features, and (iii) dominance of the subject is less important than dominance of others in the group.

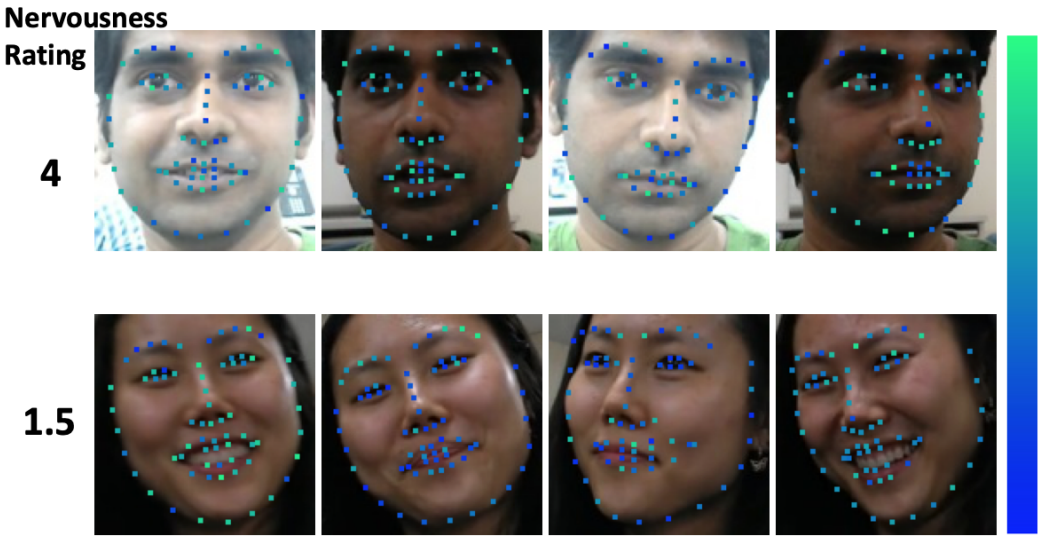
6.4.10. Prediction on data with different annotation agreements.

This experiment explores the nervousness prediction results on datasets with higher annotator agreements as defined in Table 6.1. Since some sub-datasets have less than 10 games, we randomly split all sub-datasets into 5 folds for cross-validation. Figure 6.7 (a)–(d) shows the prediction AUC of our four methods as well as the best performing baselines on each of the four tasks. The x-axis of each figure are the datasets ordered by the annotation agreements, with the leftmost being the original data (lowest agreements). We observe that the proposed FE-GCN+TCN (green lines) perform consistently well on different sub-datasets compared to the baselines (dotted lines). Not surprisingly, late fusion, which incorporates FE-GCN+TCN, achieves the best AUC in all cases. The datasets with higher annotation agreement do not always lead to better predictions, however. The reason might be that the training sets become too small (only 19%–37% of the original data) to enable well-trained models.

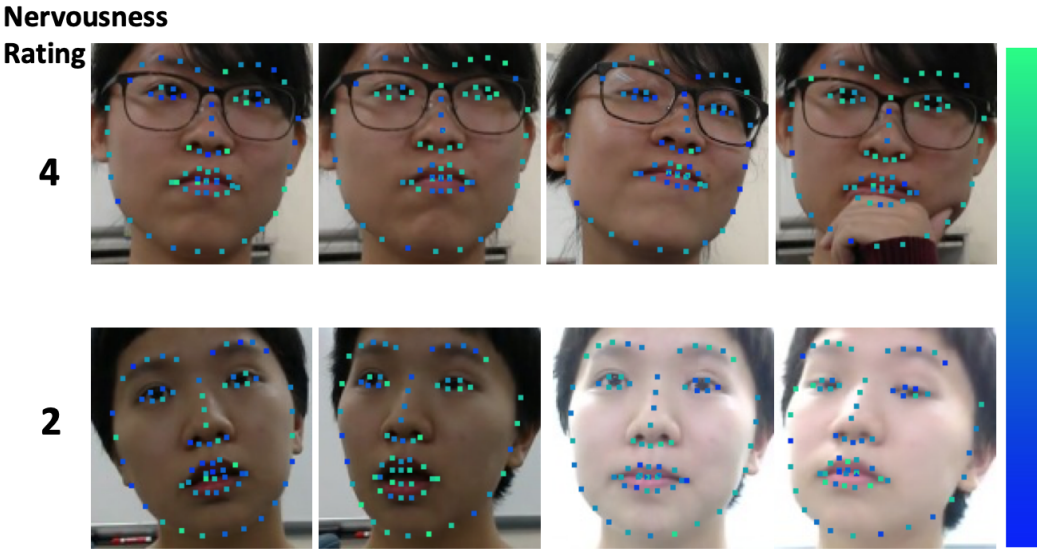
Section 6.5

Conclusion

To the best of our knowledge, NPA is the first framework to predict nervousness of subjects in group interaction videos. We introduce a new class of features called Nervousness Scores based on social science theory. We propose a novel combination of CNNs and GCNs called Facial Emotion Graph Convolution Network (FE-GCN) that generates facial embedding based features. We show that our methods beat five baselines in head to head testing and that our overall framework shows good performance on three nervousness related problems and two datasets.



(a) Game 1.



(b) Game 2.

Figure 6.6: *Experiment F: FE-GCN nervousness face retrieval and facial landmark gradients visualization on the Resistance data.*

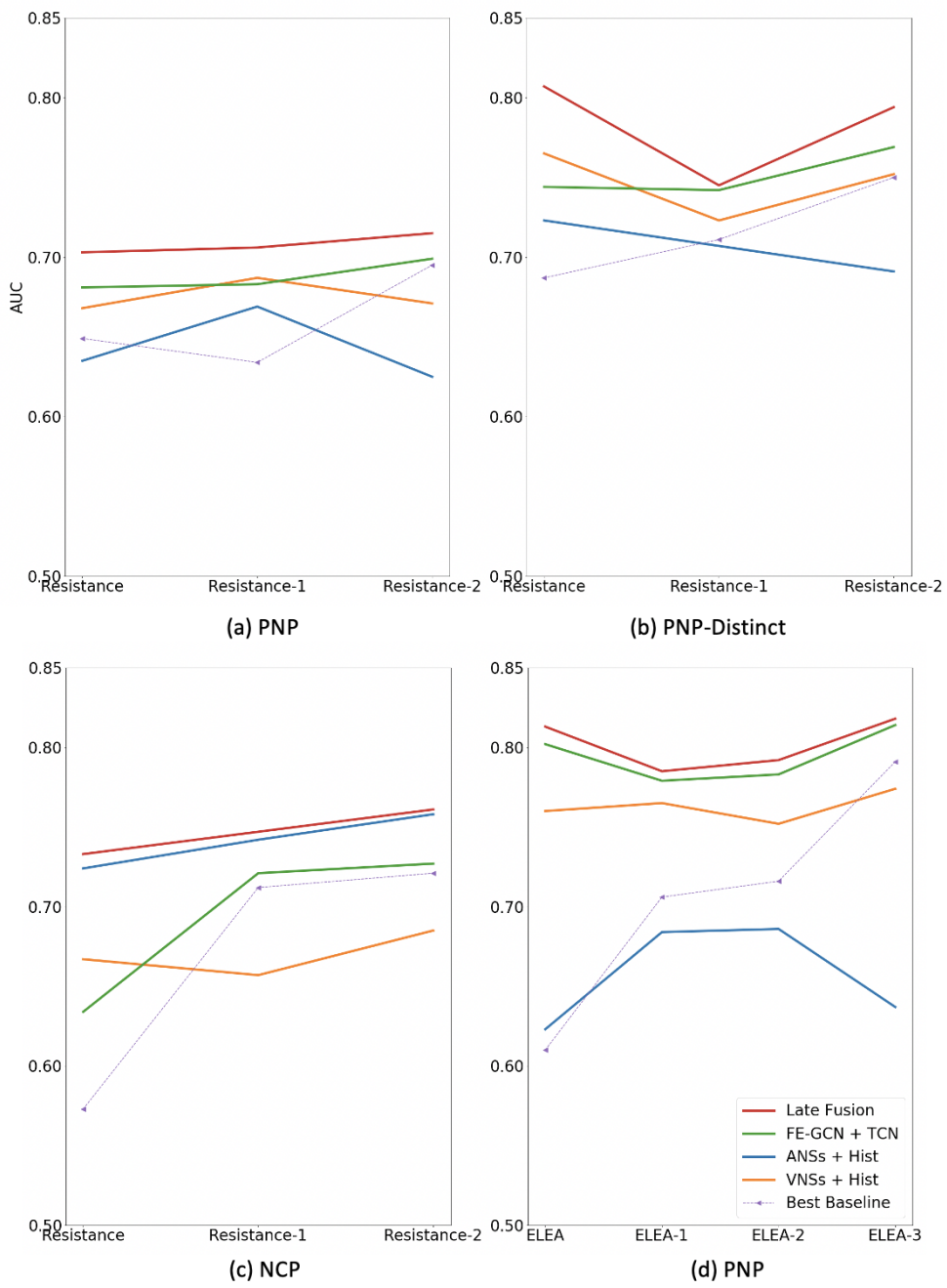


Figure 6.7: *Experiment H: Prediction on data with different annotation agreements.* Note that solid lines are our proposed methods, while dotted lines are the best performing baselines: Emotion + Hist. [13] for (a)–(b), Speaking Cues [72] for (c), and Non-verbal Activities [76] for (d). In each figure, the annotation agreement increases and the dataset size decreases from left to right.

Chapter 7

Adaptive Multimodal Fusion for Persuasion Prediction

Debates are ubiquitous. Politicians engage in debates over various policies. Companies debate the pros and cons of legislation. Students debate the pros and cons of modifications to grading systems. We develop M2P2, a system that uses multi-modal data such as acoustic, visual, language, and debate metadata to predict persuasiveness in a given debate. M2P2 considers two prediction tasks: Debate Outcome Prediction (DOP) problem (predict who wins / loses) and the Intensity of Persuasion Prediction (IPP) problem which predicts the number of votes for the position of a speaker after he speaks as compared to the number of votes before. M2P2 has several novelties: an *alignment* module that extracts shared information between modalities and a *heterogeneity* module that adaptively learns the weights of different modalities with guidance from three separately trained unimodal reference models. We test M2P2 on two debate video datasets, which significantly outperforms 3 recent baselines.

Section 7.1

Introduction

Controversial topics (e.g. foreign policy, immigration, national debt, privacy issues) engender much debate amongst academics, businesses, and politicians. Speakers who are persuasive often win such debates. Given videos of discussions between two participants, the goal of this work is to provide a fully automated system to solve two persuasion related problems. The Debate Outcome Prediction problem (DOP) tries to determine which of two teams "wins" a debate. Suppose the two teams are A and B and suppose bef_A, bef_B denote the number of supporters for A and B's positions respectively before the debate and aft_A, aft_B denote the same after the debate. Hence, $bef_A + bef_B = n = aft_A + aft_B$. In the DOP problem, we say that team A (resp. team B) wins the debate if $bef_A < aft_A$ (resp. $bef_B < aft_B$). We say a speaker is a winner if s/he belongs to the winning team. The Intensity of Persuasion Problem (IPP) tries to predict the increase (or decrease) in the number of votes of each speaker (as opposed to a team). We use the same notation as before but assuming we have two speakers S_1, S_2. The intensity of speaker X's persuasiveness is $\frac{aft_X - bef_X}{n}$ for $X \in \{S_1, S_2\}$. It is clear that both these problems are important. In a business meeting, it might be important to win (DOP), but in other situations, peeling away support for an opponent might be important (IPP). *The more support a speaker can peel away from the opponent, the more persuasive s/he is.*

Solving DOP and IPP using video data alone can pose many challenges. In this work, we test our M2P2 algorithm against two datasets, the IQ2US dataset[1] from a popular US debate TV show and the Qipashuo dataset from the popular Chinese TV show Qipashuo[2]. Real-world videos such as these come with three broad properties:

[1] https://www.intelligencesquaredus.org
[2] https://www.imdb.com/title/tt4397792/

(i) as we can see in Figure 7.2b, the detected language can be very noisy — this must be accounted for, (ii) as we can see from Figure 7.2a, there can be considerable noise in the video modality as well — for instance, a man's face might be shown in the video while a woman is speaking and these kinds of audio-video mismatches must be addressed, (iii) but in some cases — as shown in Figure 7.1, the modalities might be nicely aligned where the audio, language, and video modalities are all correct and the speaker's speech and visual signals are aligned. The problem of identifying these types of mismatches poses a major challenge in building a single model to predict both DOP and IPP.

Though we are not the first to take on the DOP problem, we are the first to solve IPP. DOP has been addressed by [25, 107, 127] who use multimodal sequence data to predict who will win a debate. However, these efforts do not address all the three challenges described above. To the best of our knowledge, there is no existing dataset that addresses the IPP problem and there are no algorithms to solve the IPP problem. *In this work, we develop a novel algorithm called* M2P2 *and show that* M2P2 *improves upon past solutions to DOP by 2%–3.7% accuracy (statistically significant with a p-value below 0.05) and beats adaptations of past work on DOP to the IPP case by over 25% MSE (statistically significant with* $p < 0.01$*).*

When all three modalities (audio, video, language) agree, then that "common" information must be correctly captured by a predictive model. In this case, we say that the modalities are *aligned*. However, there can be cases where some modalities suggest one thing, while the other(s) suggest something different. In this case, we say the modalities are *heterogeneous*. Our solution, M2P2, captures both aspects and also learns how to weight the two aspects in order to maximize prediction accuracy. M2P2 first leverages the Transformer encoder structure [144] to project the three modalities into three latent spaces. To combine the information from the latent spaces, the

Figure 7.1: *In multimodal content, the modalities are semantically aligned.* This example shows a case where the visual modality (facial expressions) and the language modality (the content of the speech) are closely aligned.

(a) *There are cases where the visual modality is noisy, while the language modality is clean.* In 4 consecutive frames when the woman is speaking, the face of a man appears (see frames 2 and 3) and the man's face is incorrectly assumed to be the woman's. The language modality, however, is correct.

Detected Chinese:
因为我燕麦奶燕奶喝呀

(b) *There are cases where the language modality can be noisy, while the visual modality is clean.* We use Baidu's off-the-shelf OCR detector to extract the Chinese transcripts from the video frames. In the video frame (the right side of the figure), the transcripts extracted by the OCR system (the left side) are incorrect due to the milk ads shown.

Figure 7.2: *Individual modalities can be noisy.* Here we show examples where the visual or the language modality are wrong. M2P2 learns to down-weight the noisy modalities.

model then devises two major modules: *alignment* and *heterogeneity*.

The *alignment* module learns to highlight the shared, aligned information across

modalities. It enforces an alignment loss in the loss function as a regularization term during training. This ensures that there is relatively little discrepancy between the latent embeddings of different modalities when they are aligned.

The *heterogeneity* module first learns the weights of modality-specific information and applies weighted fusion to harden the model against noisy modalities (cf. Figure 7.2). M2P2 uses a novel interactive training procedure to learn the weights from three separately trained reference models, each corresponding to a single modality. Intuitively, a modality with smaller unimodal loss should be assigned a higher weight in the multimodal model. Finally, the outputs of both modules are combined with the debate meta-data for persuasion prediction.

We evaluate M2P2 on the IQ2US and Qipashuo datasets. IQ2US was first used by [25] to evaluate the DOP problem. The IQ2US dataset only has the final debate outcomes, without any labels about how persuasive each speaker is during the debate. Hence, IQ2US cannot be used to evaluate IPP. To this end, we created a new dataset Qipashuo, based on an extremely popular Chinese entertainment debate TV show called Qipashuo[2]. In Qipashuo, the audience provides real-time votes before and after each speaker in order to gauge how persuasive the speaker is. Qipashuo therefore provides a direct measure of each speaker's persuasiveness for training and evaluation. We use the IQ2US dataset for the DOP problem and the Qipashuo dataset for IPP problem.

The code of M2P2 can be found at `https://shorturl.at/nqsyT`. M2P2 outperforms baselines based on three recent works [25, 107, 127] which were originally designed to predict debate outcomes (or other related problem scenarios). We also conduct ablation studies and visualize our results to show the effectiveness of different novel components in M2P2.

To summarize, we make the following contributions:

- To the best of our knowledge, M2P2 is the first to solve the IPP problem.
- We design a novel adaptive fusion learning framework to solve the IPP and DOP problems.
- We curate a new dataset Qipashuo from the well-known Chinese debate TV show Qipashuo. Qipashuo will be a strong benchmark for future work on persuasion prediction as well as multimodal learning.
- M2P2 outperforms reasonable baselines adapting recent papers [25, 107, 127] by 25% in IPP and 3% in DOP problems — and these results are statistically significant.

Section 7.2

The M2P2 framework

Figure 7.3 shows an overview of our M2P2 architecture with a brief description of its major components. Note that the key novelties of this work are the two novel modules (i.e., the alignment module and the heterogeneity module shaded in yellow in Figure 7.3) that constitute the adaptive fusion framework (Section 7.2.3) [3].

7.2.1. Generating primary input embeddings

Given a video clip, we respectively represent the acoustic, visual and language input as $X_A \in \mathcal{R}^{T_A}$, $X_V \in \mathcal{R}^{(H\times W\times C)\times T_V}$, $X_L \in \mathcal{R}^{D\times T_L}$, where T_A, T_V, T_L are respectively the lengths of the audio signal, face sequence, and word sequence. H, W, C are the height, width and the number of channels of each image, and D is the length of our dictionary of words. In addition, we also use two debate meta-data features: the

[3] Our proposed adaptive fusion framework has the potential of being broadly utilized in other multimodal learning tasks. We leave that exploration for future work.

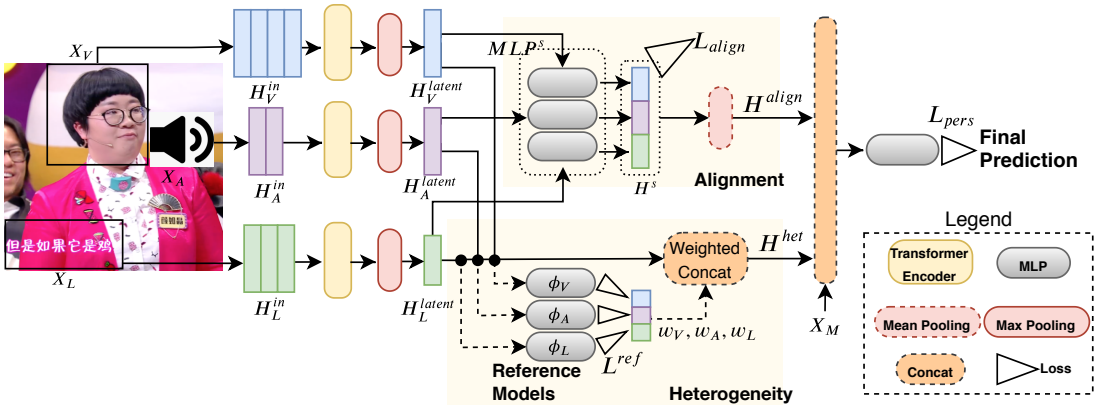


Figure 7.3: M2P2 architecture. First, audio, face and language sequences are extracted from a video clip and fed to three separate modules to get primary input embeddings. Second, each of these embeddings is fed to a Transformer encoder [144] followed by a max pooling layer, which yields the latent embeddings. Third, the latent embeddings are fed to the alignment and heterogeneity modules to generate the embeddings H^{align} and H^{het}. Last, we concatenate H^{align}, H^{het} and the debate meta-data X_M which is fed to an MLP for persuasiveness prediction. The latent embeddings interact with two procedures alternately: optimize the alignment loss L_{align} and persuasiveness loss L_{pers}, and learn weights through 3 reference models.

number of votes before a speech and the length of the speech. We generically denote the debate meta-data as a vector $X_M \in \mathcal{R}^{d_M}$, where $d_M = 2$.

We first extract features from the three modalities, then add a fully-connected (FC) layer for each modality to obtain low dimensional primary input embeddings. The generated primary input embeddings are depicted as multi-dimensional bars (as a symbol of vector sequences) in Figure 7.3. Here we describe the detailed feature extraction components.

Feature extraction from the acoustic modality. For each audio clip, we use Covarep [44] to extract MFCCs[4], Glottal source parameters, pitch-related features, and features using the Summation of Residual Harmonics method [51]. These features capture human voice characteristics from different perspectives and are all shown to be relevant to emotions [62]. These 73 dimensional features are averaged over every

[4]The energy-related 0th coefficient is excluded

half second.

Feature extraction from the visual modality. Since the speakers in both datasets can be highly dynamic and occluded, we capture only their faces as Brilman et al. [25] did to reduce noisy input. The details of face detection and recognition are in Section 7.3. Given each facial image, we use the VGG19 architecture [133] pre-trained on the Facial Emotion Recognition FER2013 dataset[5] and extract the 512 dimensional output before the last FC layer as the face features.

Feature extraction from the language modality. We use the Jieba[6] Chinese text segmentation library to segment Chinese sentences (utterances) into words. We use the Tencent Chinese embedding corpus [135] to extract 200 dimensional word embeddings. In the case of English, we extract 64 dimensional Glove word embeddings [116] trained from all transcripts from the IQ2US debates.

All features are passed to a learnable FC+ReLU layer which converts the initial features into *primary input embeddings*. The primary input embeddings thus obtained for each of the three modalities are respectively $H_A^{in} \in \mathcal{R}^{d_{in} \times T'_A}, H_V^{in} \in \mathcal{R}^{d_{in} \times T'_V}, H_L^{in} \in \mathcal{R}^{d_{in} \times T'_L}$, where $d_{in} = 16$ is the row-dimension of the primary input embeddings, which is same across different modalities. T'_A, T'_V, T'_L denote the sequence lengths of the modalities, where $T'_V = T_V, T'_L = T_L$. Note that in our primary input embeddings, the timestamps of the acoustic, visual, and language modality respectively represent a short time window, a frame, and a word.

[5]https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge/overview

[6]https://github.com/fxsjy/jieba

7.2.2. Generating compact latent embeddings of modalities with Transformers

To get a compact representation of the primary input embeddings for each modality, we aggregate the sequence of features into a single representation vector using one Transformer encoder per modality. Transformer encoders have been shown to outperform many other deep architectures, including RNNs, GRUs, and LSTMs in many sequential data processing tasks in computer vision [149] and natural language processing [46]. The multi-head self-attention mechanism of Transformer better memorizes the long-term temporal dynamics [144].

With the Transformer encoder, the primary input embedding $H_m^{in}, m \in \{A, V, L\}$ of each modality is respectively transformed into a representation as:

$$H_m^{trans} = \text{TransformerEncoder}(H_m^{in}), \tag{7.1}$$

where $H_m^{trans} \in \mathcal{R}^{d_{trans} \times T'_m}$, and $d_{trans} = 16$ is the dimension of the latent space after the Transformer encoder.

To convert arbitrary length time sequences into standardized latent embedding vectors $H_m^{latent} \in \mathcal{R}^{d_{trans} \times 1}$, we additionally use a max pooling layer:

$$H_m^{latent} = \text{MaxPool}(H_m^{trans}). \tag{7.2}$$

H_m^{latent} intuitively captures the maximum activation over the time sequence along each dimension of d_{trans}.

7.2.3. Balancing shared and heterogeneous information with adaptive fusion

As mentioned earlier, there are two conflicting aspects of multimodal data. First, data from different modalities within the same time frames may sometimes be highly aligned (i.e., have shared information). Second, different modalities may sometimes contain diverse cues which may not be equally important for prediction. To balance the aligned and heterogeneous multimodal information, we propose a novel adaptive fusion framework consisting of two key modules: an *alignment* module and a *heterogeneity* module (shaded in yellow in Figure 7.3).

Alignment module. To extract information shared across different modalities, we first use a *shared* multi-layer perceptron (MLP^s) to project the latent embeddings of each modality $m = A, V, L$ into the same latent space:

$$H_m^s = \text{MLP}^s(H_m^{latent}) \tag{7.3}$$

Here, $H_m^s \in \mathcal{R}^{d_s}$, where $d_s = 16$ is the dimension of the shared projection space. MLP^s is shown as three rounded grey boxes in Figure 7.3.

Inspired by existing multimodal representation learning work [2, 54], we use three cosine loss terms $1 - cos(H_m^s, H_n^s)$ $(\forall m, n = A, V, L, m \neq n)$ across the modalities to measure the alignment of modalities in the shared projection space:

$$\begin{aligned} \mathcal{L}_{align} &= 1 - cos(H_A^s, H_V^s) + \\ & 1 - cos(H_A^s, H_L^s) + 1 - cos(H_V^s, H_L^s) \end{aligned} \tag{7.4}$$

During training, the alignment loss will be added to the entire prediction loss function as a regularization term to penalize lack of alignment between the 3 modalities in the projected space.

After the shared MLP layer, the regularized embeddings H_m^s are in the same latent space. We apply mean pooling to average the three embeddings:

$$H^{align} = \text{MeanPool}(H_A^s, H_V^s, H_L^s) \ , \tag{7.5}$$

$H^{align} \in \mathcal{R}^{d_s}$ now contains shared information from all modalities.

Heterogeneity module. Another key observation discussed in Section 7.1 is that different modalities may contain diverse information, and therefore make unequal contributions to the final prediction of persuasiveness (e.g., due to the noisy data from certain modalities as shown in Figure 7.2). We therefore propose a novel heterogeneity module which utilizes an interactive training procedure (Algorithm 3) to learn weights for different modalities.

Intuitively, the importance of each modality should be inversely proportional to the "error" caused by the modality. To estimate this error term, we create three unimodal MLP reference models (represented as dashed arrows and rounded grey boxes at the central bottom of Figure 7.3) parameterized by ϕ_A, ϕ_V, ϕ_L for the acoustic, visual, and language modalities respectively. Each unimodal MLP takes the compact latent embedding H_m^{latent} generated by the Transformer encoder as input and generates a unimodal prediction $\hat{Y}_m^{ref}$ for each modality $m = A, V, L$:

$$\hat{Y}_m^{ref} = \text{MLP}_m^{ref}(\phi_m; H_m^{latent}) \ . \tag{7.6}$$

We use T_{val} to denote the validation set, $Y_{val} \in \mathcal{R}^{|T_{val}|}$ are the labels, and $\hat{Y}_{m,val}^{ref} \in \mathcal{R}^{|T_{val}|}$ are the predictions made by the unimodal reference model for modality m.

The reference models (ϕ_m's) are updated using the following Mean Squared Error

(MSE) loss alone:

$$\mathcal{L}_m^{ref} = \frac{\left\| Y_{val} - \hat{Y}_{m,val}^{ref} \right\|_2^2}{|T_{val}|} \tag{7.7}$$

After several epochs of training ϕ_m's, we are able to obtain a converged MSE loss of each reference model. We then use the updated reference model to estimate the prediction errors by $\mathcal{L}_m^{ref}$. $\mathcal{L}_m^{ref}$ is used to guide the weights w_m of latent embeddings H_m^{latent} $(m = A, V, L)$ to be concatenated in the heterogeneity module:

$$H^{het} = w_A H_A^{latent} \oplus w_V H_V^{latent} \oplus w_L H_L^{latent}. \tag{7.8}$$

w_A, w_V, w_L are scalars incrementally updated over epochs:

$$w_m = \alpha w_m + (1 - \alpha)\tilde{w}_m, \tag{7.9}$$

where $\alpha \in (0, 1)$ controls the rate of update, and $\tilde{w}_m$ is obtained using the following softmax function of the reference model validation losses:

$$\tilde{w}_m = \frac{\exp\{-\beta \mathcal{L}_m^{ref}\}}{\sum_{m'=A,V,L} \exp\{-\beta \mathcal{L}_{m'}^{ref}\}}, \forall m = A, V, L \tag{7.10}$$

$\beta > 0$ is a scaling factor. Since $\sum_m \tilde{w}_m = 1$, combining Equation (7.9), it is guaranteed that $\sum_m w_m = 1$.

Adaptive fusion with interactive training. The representations obtained from the alignment module (H^{align}) and the heterogeneity module (H^{het}) are then concatenated together with the debate meta-data X_M and fed into a final MLP layer to make the final prediction $\hat{Y}$:

$$\hat{Y} = f(\theta; X_A, X_V, X_L, X_M) = \text{MLP}(H^{align} \oplus H^{het} \oplus X_M) \tag{7.11}$$

Algorithm 3: M2P2 interactive training procedure.

Input: Training dataset T, validation datset T_{val}; Number of epochs n and N

Output: Multi-modality model $f(\theta; X_A, X_V, X_L, X_M)$, modality weights w_m $(\forall m = A, V, L)$

```
1  Initialize three unimodal reference models φ_m(∀m = A, V, L) and θ;
2  Initialize w_A = w_V = w_L = 1/3;
3  % Master Procedure Start
4  for epoch=1,...,N do
5  |  Update θ with loss function Equation (7.12);
6  |  Get latent embeddings H_m^latent, ∀m = A, V, L;
7  |  % Slave Procedure Start
8  |  for epoch=1,...,n do
9  |  |  Update φ_m, ∀m = A, V, L with loss function in Equation (7.7);
10 |  end
11 |  % Slave Procedure End
12 |  Get reference model losses L_m^ref, ∀m = A, V, L;
13 |  Update modality importance weights w_m, ∀m = A, V, L using
   |   Equations (7.9)-(7.10);
14 end
15 % Master Procedure End
16 return θ, w_m(∀m = A, V, L)
```

where θ is the set of parameters of the M2P2 model excluding the reference model parameters ϕ_m.

To train the M2P2 model, we have two loss terms: a novel alignment loss $\mathcal{L}_{align}$, and a persuasiveness loss term $\mathcal{L}_{pers}$. In the case of the IPP problem, $\mathcal{L}_{pers}$ is the MSE loss. In the case of DOP, we use cross-entropy loss for the binary classification.

The total loss function is a weighted combination:

$$\mathcal{L}_{final} = \mathcal{L}_{pers} + \gamma \mathcal{L}_{align}, \tag{7.12}$$

where γ is a weight factor.

The entire training proceeds in a master-slave manner, as shown in Algorithm 3. In each epoch of the master training procedure (Lines 4 to 14), we use the total

loss function in Equation (7.12) to update the parameters θ of the main M2P2 components. The weights w_A, w_V, w_L of the 3 modalities are obtained using reference models ϕ_m, and their losses $\mathcal{L}_m^{ref}$ are then updated in the slave procedure. In each epoch of the slave procedure (Lines 8 to 10), we take the latent embeddings from the master procedure as input and update the reference models with the loss function in Equation (7.7). We then obtain the weights w_A, w_V, w_L of different modalities in the heterogeneity module.

Section 7.3

Data preprocessing

7.3.1. Qipashuo dataset

The dataset is described in Section 3.3.2. We extracted the transcripts from the video subtitles. To sufficiently preprocess the videos for subtitle extraction, we took the following steps. First, we sampled 2 frames per second and binarize the images with a threshold 0.6, which can avoid the influence from various colors of subtitles in videos. Second, we cropped the subtitles by a fixed bounding box since the position of subtitles is fixed in all the videos. Third, we clustered the binarized images into buckets such that any two binarized images in the same bucket are identical on 90% or more pixels. We then randomly selected one of these images to represent the cluster. This helps reduce noise (e.g. from advertisements displayed on the image). Finally, the surviving binary images were fed into an OCR API to get accurate transcripts. We used Baidu's off-the-shelf pre-trained OCR API[7], so no extra data is needed for training.

If we take each speaking clip as a train/test instance, there would be a total of 205 data points. This paucity of information poses a huge challenge for machine learning.

[7]https://ai.baidu.com/tech/ocr

We therefore segment each speaking clip into clips of 50 utterances each according to the transcript we extract above. Note that 50 is the smallest number of utterances in any speaking clip of our dataset. Moreover, note that these "sub-clips" of 50 utterances yield a temporal sequence whose temporal dynamics can be important. The labels are shared for segments extracted from the same clip. This trick yields 2,297 such segments which are used as train/test instances in our evaluation.

As the speakers are highly dynamic and often occluded, we only use speakers' faces as the visual input. We extract 2 frames per second from videos and use Dlib[8] for face detection and recognition. The recognition is based on one pre-annotated profile for each speaker and is only needed for training. To further reduce false positives (i.e., extracting the face of the non-speakers), we first use the model from [15] to remove faces in the image that are not speaking, and then use the method from [100] for face tracking.

7.3.2. IQ2US dataset

In the IQ2US data (Section 3.3.1), no pre-processing is required for the language modality. For the visual modality, we use the same procedures as in the Qipashuo dataset to extract the face image sequences of the speakers. Since there are no intermediate votes in IQ2US, we only predict the debate outcome (i.e. whether a single-speaker clip instance belongs to the winning team).

Section 7.4

Experimental evaluations

Our experiments assess the performance of M2P2 on the DOP and IPP tasks. Specifically:

[8] http://dlib.net

Fold	1	2	3	4	5	6	7	8	9	10	Average
Brilman et al. [25]	0.009	0.011	0.016	0.017	0.030	0.018	0.020	0.012	0.013	0.018	0.016
Nojavanasghari et al. [107]	0.007	0.015	0.019	**0.011**	0.027	**0.014**	0.020	0.012	0.020	0.015	0.016
Santos et al. [127]	0.025	0.019	0.018	0.019	**0.018**	0.017	0.029	0.016	0.024	0.018	0.020
M2P2 (ours)	**0.006**	**0.010**	**0.015**	0.015	0.020	0.015	**0.012**	**0.009**	**0.009**	**0.013**	**0.012**
dec. %	14.2	9.1	6.3	-36.4	-11.1	-7.1	40.0	25.0	30.8	13.3	25.0

Table 7.1: MSE for each test fold of different approaches to solving the Intensity of Persuasion Prediction (IPP) on the Qipashuo Dataset. The last row shows the MSE decrease percentage of M2P2 compared to the best baseline in each fold. On average, M2P2 achieves a lower MSE than the baselines by at least 25%. Results are statistically significant with p-val < 0.01. Note that the vote scores we predict range from 0 to 1.

Method	DOP (Accuracy)	IPP (MSE)
Brilman et al. [25] (early fusion)	0.614	0.016
Nojavanasghari et al. [107] (late fusion)	0.615	0.016
Santos et al. [127] (early fusion)	0.598	0.020
M2P2 (proposed method)	**0.635**	**0.012**

Table 7.2: Prediction accuracy for Debate Outcome Prediction in IQ2US dataset. Our M2P2 is 2%–3.7% better than baselines. Results are statistically significant with p-val < 0.05.

(a) (IPP) We predict the change of number of votes after a speech by a debater — this is done on the Qipashuo dataset;

(b) (DOP) We predict whether a clip in which a debater is speaking is part of the winning team of the debate — this is done on the IQ2US dataset;

In addition, we also conduct an ablation study that assesses the contributions of different components of M2P2. Moreover, we assess the importance of different modalities as well as time frames using the Qipashuo dataset. Finally, we compare the results of different ways of encoding the linguistic inputs.

7.4.1. Experimental settings

Qipashuo uses a 10-fold rolling window prediction. Specifically, we construct 10 se-

quences of consecutive episodes of the show. For instance, if $E_1, \ldots, E_k$ represent the set of all Qipashuo episodes, then one sequence would be $Seq_k = E_1, \ldots, E_k$, another would be $Seq_{k-1} = E_1, \ldots, E_{k-1}$. For any such sequence $Seq_i = E_1, \ldots, E_i$, we set E_i as the test episode (i.e. the episode on which we make predictions). We learn a model from the first $i-3$ episodes $E_1, \ldots, E_{i-3}$ and identify the best parameters for our model by using episodes E_{i-2}, E_{i-1} as the validation set. As the same subject can occur in multiple episodes of Qipashuo in order to avoid information leakage from training to test data, we do not train a model from E_i to predict $E_{j,j<i}, \forall i, j$.

For IQ2US, 10-fold cross validation is used since a debater can only appear in one episode. The initial vote score and speaking length features are normalized to $(0, 1]$.

Denote FCn as a fully-connected layer that outputs n-dimension vectors. The MLPs in the reference models and final multimodal prediction model are all configured as FC16+ReLU, FC8+ReLU, and FC1+Sigmoid. The shared MLP in *alignment module* is FC16+ReLU. M2P2 uses Batch Normalization [74] right after each of the FC layers for input embeddings, and uses 0.4 as dropout [68] after all FC16 layers. For the Transformer encoder, we use a single layer with 4 heads, where the input, hidden, and output dimension are all 16. We use the Adam [81] optimizer with a weight decay of 10^{-5}. The numbers of epochs in Algorithm 1 is $N = 200$ and $n = 10$. The learning rate lr, alignment loss weight γ, update scalar α, scaling factor β are finalized by grid search. We ended up using $lr = 0.001, \gamma = 0.1, \alpha = 0.5, \beta = 50$ as these yield the best results on the validation sets.

7.4.2. Comparison with baselines

We compare both tasks with the following multimodal persuasion prediction baselines: early fusion + SVM [25], deep multimodal late fusion [107], and early fusion + LSTM [127]. Brilman et al. [25] extract audio, visual and linguistic features from IQ2US debate videos and concatenate these features, which are fed into an SVM for

classification. Although [25] also solves the DOP problem on the IQ2US dataset, it is different from our work in that (i) the used episodes are different (see Section 7.3.2 and (ii) it uses long video inputs (9–36 minutes) of all debates while we only use a short speaking clip (1 minutes) of a single speaker. Thus, for fair comparison, we implemented their method and ran experiments in our data. Nojavanasghari et al. [107] first feed features of each modality to a neural network to get predictions of the modality, then uses a fusion neural network to combine the modality-based predictions. Santos et al. [127] model the temporal dynamics by using an LSTM on the concatenated features from all modalities.

In the case of the IPP problem, we adapt the first baseline by modifying it to use an SVM regressor (rather than an SVM classifier). For the other two baselines, we use MSE loss to train the models. For fairness, we also allow the baselines to use the two debate meta-data features. The results comparing M2P2 on IPP and DOP with past approaches are shown in Tables 7.1 and 7.2, respectively.

IPP Problem Table 7.1 shows the MSE obtained by different approaches in each fold and the average on the Qipashuo dataset. Note that the vote scores we predict are normalized to lie in the $[0, 1]$ interval. The last line of Table 7.1 shows the decrease percentage of MSE which is defined as *dec.* = 1-MSE(M2P2)/MSE(the best baseline). For instance, from the first column of Table 7.1, we see that the percentage decrease is $1 - \frac{0.006}{0.007} \approx 0.14$ representing a 14% decrease of MSE generated by M2P2 compared to the best of the baselines. In the case of IPP, we see that on average, M2P2 yields a 25% decrease of MSE compared with the best baseline which is statistically significant via a Student t-test (p-val < 0.01). Moreover, M2P2 is more robust and performs better than all baselines in 7 out of 10 folds.

DOP Problem Table 7.2 shows the average prediction accuracy over 10 folds on the DOP problem w.r.t. the IQ2US dataset. It is clear that M2P2 achieves 2%–3.7% higher average accuracy than the baselines, the improvement is statistically significant (p-val < 0.05). These make M2P2 the best performing system for both the IPP and the DOP problems.

Method	MSE
M2P2 without alignment loss	0.018
M2P2 without reference models	0.015
M2P2-LSTM	0.032
M2P2-Acoustic (unimodal)	0.017
M2P2-Visual (unimodal)	0.019
M2P2-Language (unimodal)	0.016
M2P2	**0.012**

Table 7.3: Ablation study results. All improvements are statistically significant (p-val < 0.01). The methods from top to bottom are: M2P2 without correlation losses, M2P2 without reference models, M2P2 with LSTM layer instead of Transformer Encoder and max pooling, M2P2 with only acoustic modality, only visual modality, and only language modality.

7.4.3. Ablation study

To measure the contributions of the different components of M2P2, we create four methods, each with one component removed from M2P2 :

- M2P2 without the alignment loss.
- M2P2 without reference models. The latent embeddings are concatenated by equal weights 1/3.
- M2P2-LSTM. The Transformer encoder and max pooling layer are replaced by a 1-layer LSTM.

- M2P2-unimodal. We input a single modality without alignment loss and latent embedding concatenation. That is, the latent embedding is directly concatenated with the debate meta-feature and fed to the final MLP.

IPP Problem Table 7.3 shows the average MSE obtained on the Qipashuo dataset for both M2P2 and the 4 methods above. First, according to rows 1,2 and the last row, we find that if M2P2 does not use the alignment module and reference models in the heterogeneity module, the MSE increases from 0.012 to 0.018 and 0.015 respectively. This is statistically significant (p-val < 0.01) and hence shows the power of both proposed adaptive fusion modules in Section 7.2.3. Second, we observe the power of the Multihead-attention Transformer encoder to handle long sequences, as the M2P2-LSTM model achieves the worst MSE amongst all methods. Third, we observe from rows 4-6 that the language modality is the most important in the prediction task, while the acoustic and visual modalities are less important. This observation is consistent with the modality concatenation weights that will be shown in the following subsection.

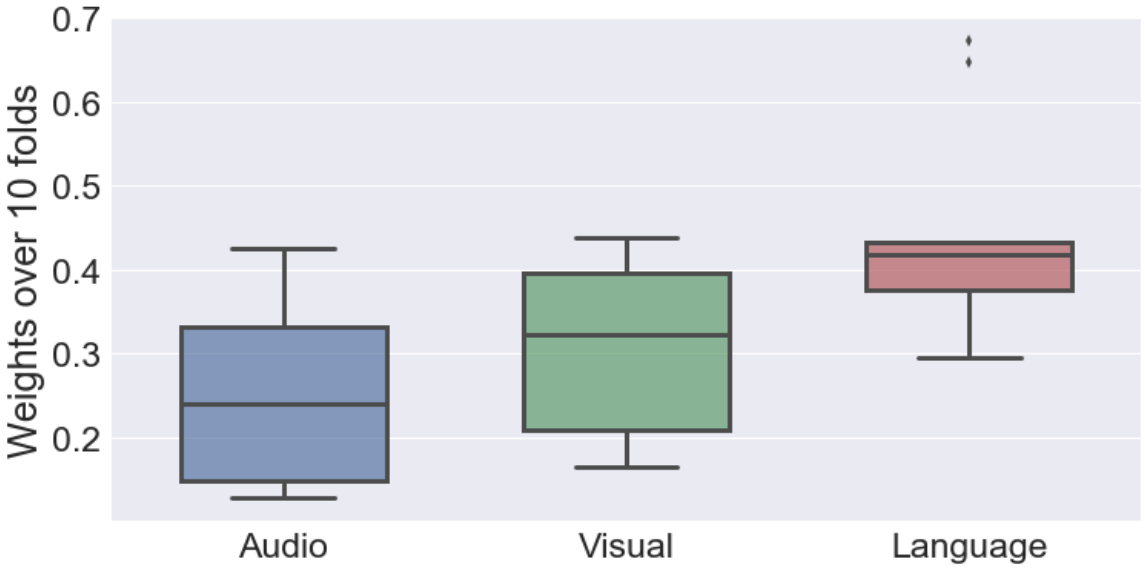


Figure 7.4: Modality weights in the heterogeneity module.

7.4.4. Visualization of prediction

In this experiment, we show (1) the importance of modalities through their learned weights (cf. Equation (7.8)), and (2) the examples of learned temporal attention weights from different modalities.

Modality weights We report the modality weights in the heterogeneity module of the trained M2P2 in all folds of Qipashuo dataset. Figure 7.4 shows box plots for the three modalities. The language modality is the most important and robust over all folds with a median weight of 0.42, while the median weights of acoustic and visual modalities are 0.23 and 0.32 respectively.

Temporal attention weights We visualize the temporal attention weights of two sample sequences of visual (Figure 7.5) and language (Figure 7.6) modalities. For each timestamp, we average the attention weights of all timestamps and all heads towards it, as its attention weight. In Figure 7.5 (top), the man's face is not detected correctly in frames 3 and 6 – and we see that M2P2 assigns near-zero attention weights to both frames, suggesting that these frames should be ignored. Moreover, the happy expression in frame 2 gets a high attention weight. The woman below gets high attention weights when she actively talks to someone (frames 2,4,5). In Figure 7.6, we notice that reasonable keywords like 'wear', 'shackle', 'passive', and 'hold' also get high attention weights. Therefore, our M2P2 captures the meaningful long-range temporal dynamics with the help of Transformer Encoder.

Figure 7.5: Temporal attention of visual modality – color coded as blue. Darker color implies higher attention weight.

If I always keep helping her , work out everything at clutch time , would it wear her a shackle so that she can't achieve her own life value ?

I've got self-control, but I have slightly passive attitude towards humanity . The question is – can I always be myself over time ?

Figure 7.6: Temporal attention of language modality – color coded as red. Darker color implies higher attention weight. The original Chinese transcripts are translated to English.

Section 7.5

Discussion

7.5.1. Text encoder comparison for linguistic inputs

In M2P2 , the sequence of word embeddings is used as the sequence input to the Transformer encoder. Another way is to encode each sentence to an embedding and feed the sequence of sentences to the Transformer encoder. We have conducted experiments to compare these two methods. To get English sentence embeddings in IQ2US, we employ the pre-trained Universal Sentence Encoder [33] in TFHub[9]. For Chinese sentences in Qipashuo we train an LSTM to get 128-dimensional sentence embeddings. We replace word embeddings with sentence embeddings and conduct the experiments in both datasets. As a result, the accuracy in IQ2US is 0.623 (1.4% worse than M2P2) and the mean squared error in Qipashuo is 0.014 (20% worse than M2P2). Thus, the fine-grained word-level embeddings are better than sentence-level embeddings. The word order and semantic meaning is already captured by the word-level embeddings.

7.5.2. Heterogeneity module *vs.* attention mechanism

Intuitively, the heterogeneity module in M2P2 aims to learn the modality-wise importance from data. An alternative is to use the attention mechanism to attend the model to different modalities. However, the attention mechanism introduces extra amount of trainable parameters into M2P2 . In our early experiments, this resulted in worse results due to the small dataset (2297 and 805 data points for Qipashuo and IQ2US resp.). On the contrary, the parameters introduced by heterogeneity module are independent from the rest of M2P2 model, which fuses modalities and achieves better prediction results.

[9] https://tfhub.dev/google/universal-sentence-encoder/1

Section 7.6

Conclusion and future work

In this work, we have solved two problems. First, we provide a solution to the Debate Outcome Prediction (DOP) problem that improves on past work by 2%–3.7%. Though these numbers are not huge, they are statistically significant. Second, we are the first to pose and solve the Intensity of Persuasion Prediction (IPP) problem. We show that we are able to beat baselines built on top of past solutions to IPP by 25% on average. Our proposed M2P2 framework leverages both the common and modality-specific information contained in multimodal sequence data (audio, video, language), while learning to focus attention on the meaningful part of the data. Moreover, our newly created Qipashuo dataset provides a valuable new asset for future research.

However, there is ample scope for future work. First, we do not provide any theoretical guarantees on the convergence of modality weights. Second, more scalable methods to capture cross-modal interaction would be very valuable. Third, one may inspect if the interactive training procedure has overpower issues of one over another and improve further.

It is important to note that the adaptive fusion technique in M2P2 can be generalized to other multimodal sequence prediction problems such as video question answering and video sentiment analysis. We leave this exploration for future work. In other future work, we plan to conduct semantic-level studies to gain knowledge of the persuasive attributes (e.g. are high pitch, positive sentiment, attractive faces more persuasive?). One can also explore richer primary input modality embeddings (e.g. body pose, context-related word embeddings).

Chapter 8

Conclusion and future work

In the final chapter, we give a complete picture of our main contributions and findings across all chapters, and point out several prominent future directions to be explored.

Section 8.1 Conclusion

This thesis proposed several predictive models of group human behaviors on videos and analyzed the cues characterizing different behaviors. Accurate identification of the behaviors is great needed by companies and governments in situations like decision making, consulting, security check and marketing. To sum up, we have studied the following aspects:

(a) building dynamic non-verbal interaction networks (e.g. looking at, talking to) from videos of a group of people,

(b) defining informative features involving group interactions and multiple modalities, inspired from social science findings of human behaviors,

(c) developing predictive models to consider group-level influence and fuse multimodal features,

(d) analyzing behavior-specific patterns from the models,

In chapter 2, we summarized the factors relating the dominance, nervousness and persuasion behaviors, including visual (e.g. emotions, gazes), vocal (e.g. pitch, volume) linguistic (e.g. semantic) and interaction (e.g. interruption) cues. We then summarized the existing computational efforts involving feature engineering, temporal aggregation, group influence modeling, and multimodal fusion.

In chapter 3, we introduced four video datasets we used to train and evaluate models and study group behaviors. We employed the Resistance dataset [56] where people played the Resistance social game in an adversarial setting, and the ELEA dataset [124] recording a cooperative game played by a group. The two datasets have similar labels such as dominance, like/dislike, nervousness. We also introduced two debate datasets, Qipashuo (in Chinese) and IQ2US (in English) on which we study the persuasion behavior.

Chapter 4 proposed an algorithm to extract the non-verbal interaction (who looks at who) from group interaction videos. It also proposed a lightly-supervised version of this algorithm, which generalizes the prediction towards unseen videos using the prior that people usually look at the single speaker. It further developed an accurate model to predict the speaking behavior from mouth movements. Both the look-at and speaking behaviors are the foundation of building more complex interactions. Finally, a face-to-face dynamic communication network dataset is released for further research, which contains 62 networks, $\sim$3M edges.

Chapter 5 developed methods to predict (i) the most dominant person in a group and (ii) the more dominant person in a pair. We proposed the dominance rank features and two models. The dominance ranks capture the relative dominance from various types of interactions. The DELF model fuses multiple modalities and achieves at least 0.79 AUC on all tasks, and the GDP model (for problem (i)) improves the it to

0.82 by data augmentation and group-level prediction. We found that the dominance rank features and speaking histogram features play a key role in problem (i) and (ii) respectively.

In chapter 6, we developed a hybrid system combing feature engineering and end-to-end representation learning. On one hand, we proposed the nervousness score features which consider the non-verbal interactions together with audiovisual emotions and relative dominance ranks between people. On the other hand, we designed the FE-GCN + TCN model to learn face emotional embeddings from the dynamics of the face landmarks. Our system achieves 0.7 to 0.81 AUC on four tasks and two datasets. We found that (i) the visual emotions are more important than audio emotions in the nervousness scores for accurate prediction, and the negative audio emotions expressed to a person have more impact on his/her nervousness than the positive ones, (ii) the speak-to and listen-to interactions are more important than the look-at interaction for prediction, indicating that the speaking behavior plays a key role, and (iii) the landmarks in mouth-nose and chain regions are indicators for nervousness.

Chapter 7 came up with a multimodal adaptive fusion framework M2P2 and demonstrated its efficacy on persuasion prediction. The framework consists of an alignment module to project the multimodal inputs into one latent space, and a heterogeneity module to learn the modality importance adaptively through the guidance of three single-modal models. As a result, M2P2 achieves 0.64 accuracy on the debate outcome persuasion and 0.012 MSE on the intensity prediction of persuasion, beating all three previous baselines. Our heterogeneity module shows that order of modality importance is: language, video and audio.

Section 8.2

Future work

We discuss several significant research directions which will increase our understandings of group human behaviors as well as further improve and generalize our prediction models.

8.2.1. Better understanding of group human behaviors

The context of verbal interaction is essential for understanding human behaviors, yet it is non-trivial to extract such interaction automatically. One potential way is to use Automatic Speech Recognition (ASR) techniques ([158, 128, 11]) to convert audio to transcripts, and combine with the look-at information predicted by our model (chapter 4). Once the verbal interaction is included, further research can explore the factors (e.g. sentiments, key words, topics) among communications that characterize group human behaviors. What's more, it could be helpful to extract features or word embeddings of such texts and feeding into the existing models.

Another meaningful direction is to study the relationship between human behaviors and gender, ethnicity, and culture. Social scientists have conducted such studies on deception detection [28] and trustworthiness [24]. As the Resistance data contains this information, we can divide the data and apply the existing models separately. The separate models can further characterize the behaviors (e.g. do male and female behave differently when being dominant?), and might improve the prediction performance since the variance of samples is reduced. However, a challenge raised by this is the much smaller available training dataset, which might be resolved by pre-training (next subsection).

8.2.2. Model generalization

Unified framework In most existing work, a model is only used to make predictions by hand-crafted features for one kind of behavior ([106, 107, 13]), which limits the usage of the model. As human behaviors are usually inter-correlated (e.g. nervousness, leadership, and dominance, chapter 6), a unified framework that simultaneously or collectively predicts all of them might be beneficial that: (i) the correlation can be exploited by collective classification, (ii) multiple datasets of different behaviors can be combined to train the framework in a multi-task manner, which will increase the training samples, and (iii) the general representation can be applied to multiple behaviors, and domain-specific knowledge may not be required.

One related concurrent work is Wang et al. [150]. They build a general graph neural network based model to capture dynamic interactions and show its success in predicting dominance, nervousness and deception.

Pre-training The self-supervision manner of pre-training and fine-tuning has shown great success in NLP [46], image-text learning [141, 98], and audio-visual learning [5]. Such methods employ the self-correspondence of the data, such as an image and the text describing it or a guitar video and its sound, to pre-train a large powerful model and fine-tune it with annotated labels in much smaller datasets. Future research can collect large-scale dataset of people interacting with each other, and design specific pre-training tasks. Priors such as audio-visual correspondence enable the model to learn the multimodal representation of the cues such as emotions and semantics. The pre-trained model can then be fine-tuned on Resistance, ELEA or IQ2US data for group human behavior predictions.

Less constrained settings Currently, most proposed models take frontal view videos as input. Although less noisy, the applications are limited – close-up cameras

are needed to capture each individual. Moreover, the frontal view videos make it more difficult to capture the group interaction [16]. Recently, researchers try to predict social relationship from single-view videos ([96, 86]) with annotated labels. Using such videos, future work can build the interaction through the geometry of the body, head and eyes (e.g. [101]) and further predict the group human behaviors.

Bibliography

[1] Gustavo Aguilar, Viktor Rozgic, Weiran Wang, and Chao Wang, *Multimodal and multi-view models for emotion recognition*, Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (Florence, Italy), Association for Computational Linguistics, July 2019, pp. 991–1002.

[2] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu, *Deep canonical correlation analysis*, International conference on machine learning, 2013, pp. 1247–1255.

[3] Oya Aran and Daniel Gatica-Perez, *Fusing audio-visual nonverbal cues to detect dominant people in group conversations*, 2010 20th International Conference on Pattern Recognition, Aug 2010, pp. 3687–3690.

[4] Oya Aran and Daniel Gatica-Perez, *One of a kind: inferring personality impressions in meetings*, 2013 International Conference on Multimodal Interaction, ICMI '13, Sydney, NSW, Australia, December 9-13, 2013 (Julien Epps, Fang Chen, Sharon L. Oviatt, Kenji Mase, Andrew Sears, Kristiina Jokinen, and Björn W. Schuller, eds.), ACM, 2013, pp. 11–18.

[5] Relja Arandjelovic and Andrew Zisserman, *Look, listen and learn*, Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 609–617.

[6] Stylianos Asteriadis, Kostas Karpouzis, and Stefanos Kollias, *Visual focus of attention in non-calibrated environments using gaze estimation*, International Journal of Computer Vision **107** (2014), no. 3, 293–316.

[7] Sileye O Ba, Hayley Hung, and Jean-Marc Odobez, *Visual activity context for focus of attention estimation in dynamic meetings*, Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on, IEEE, 2009, pp. 1424–1427.

[8] Sileye O Ba and Jean-Marc Odobez, *Multi-party focus of attention recognition in meetings from head pose and multimodal contextual cues*, Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on, IEEE, 2008, pp. 2221–2224.

[9] Sileye O. Ba and Jean-Marc Odobez, *Recognizing visual focus of attention from head pose in natural meetings*, IEEE Trans. Syst. Man Cybern. Part B **39** (2009), no. 1, 16–33.

[10] Sileye O Ba and Jean-Marc Odobez, *Multiperson visual focus of attention from head pose and meeting contextual cues*, IEEE Transactions on Pattern Analysis and Machine Intelligence **33** (2011), no. 1, 101–116.

[11] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli, *wav2vec 2.0: A framework for self-supervised learning of speech representations*, 2020.

[12] Chongyang Bai, Maksim Bolonkin, Judee K Burgoon, Chao Chen, Norah Dunbar, Bharat Singh, VS Subrahmanian, and Zhe Wu, *Automatic long-term deception detection in group interaction videos*, 2019 IEEE International Conference on Multimedia and Expo, ICME 2019, IEEE Computer Society, 2019, pp. 1600–1605.

[13] Chongyang Bai, Maksim Bolonkin, Srijan Kumar, Jure Leskovec, Judee Burgoon, Norah Dunbar, and V. S. Subrahmanian, *Predicting dominance in multi-person videos*, Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19, International Joint Conferences on Artificial Intelligence Organization, 7 2019, pp. 4643–4650.

[14] Chongyang Bai, Haipeng Chen, Srijan Kumar, Jure Leskovec, and VS Subrahmanian, *M2p2: Multimodal persuasion prediction using adaptive fusion*, 2020.

[15] Chongyang Bai, Srijan Kumar, Jure Leskovec, Miriam Metzger, Jay F. Nunamaker, and V. S. Subrahmanian, *Predicting the visual focus of attention in multi-person discussion videos*, Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19, International Joint Conferences on Artificial Intelligence Organization, 7 2019, pp. 4504–4510.

[16] Chongyang Bai, Srijan Kumar, Jure Leskovec, Miriam Metzger, Jay F Nunamaker Jr, and VS Subrahmanian, *Predicting the visual focus of attention in multi-person discussion videos.*, IJCAI, 2019, pp. 4504–4510.

[17] Shaojie Bai, J Zico Kolter, and Vladlen Koltun, *An empirical evaluation of generic convolutional and recurrent networks for sequence modeling*, 2018.

[18] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L. P. Morency, *Openface 2.0: Facial behavior analysis toolkit*, 2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018), May 2018, pp. 59–66.

[19] Tadas Baltrusaitis, Amirali Bagher Zadeh, Yao Chong Lim, and Louis-Philippe Morency, *Openface 2.0: Facial behavior analysis toolkit*, 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), 2018, pp. 59–66.

[20] Hedi Ben-Younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome, *Mutan: Multimodal tucker fusion for visual question answering*, Proceedings of the IEEE international conference on computer vision, 2017, pp. 2612–2620.

[21] Ben Benfold and Ian Reid, *Unsupervised learning of a scene-specific coarse gaze estimator*, Computer Vision (ICCV), 2011 IEEE International Conference on, IEEE, 2011, pp. 2344–2351.

[22] 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** (2018), no. 2, 441–456.

[23] Mustafa Bilgic, Galileo Mark Namata, and Lise Getoor, *Combining collective classification and link prediction*, Data Mining Workshops, 2007. ICDM Workshops 2007. Seventh IEEE International Conference on, IEEE, 2007, pp. 381–386.

[24] Béla Birkás, Milena Dzhelyova, Beatrix Lábadi, Tamás Bereczkei, and David Ian Perrett, *Cross-cultural perception of trustworthiness: The effect of ethnicity features on evaluation of faces' observed trustworthiness across four samples*, Personality and Individual Differences **69** (2014), 56–61.

[25] Maarten Brilman and Stefan Scherer, *A multimodal predictive model of successful debaters or how i learned to sway votes*, Proceedings of the 23rd ACM international conference on Multimedia, ACM, 2015, pp. 149–158.

[26] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al., *Language models are few-shot learners*, 2020.

[27] Judee K Burgoon, Thomas Birk, and Michael Pfau, *Nonverbal behaviors, persuasion, and credibility*, Human communication research **17** (1990), no. 1, 140–169.

[28] Judee K Burgoon, Dimitris Metaxas, Jay F Nunamaker, and Saiying Tina Ge, *Cultural influence on deceptive communication*, Detecting Trust and Deception in Group Interaction, Springer, 2021, pp. 197–222.

[29] Carlos Busso, Zhigang Deng, Serdar Yildirim, Murtaza Bulut, Chul Min Lee, Abe Kazemzadeh, Sungbok Lee, Ulrich Neumann, and Shrikanth Narayanan, *Analysis of emotion recognition using facial expressions, speech and multimodal information*, Proceedings of the 6th international conference on Multimodal interfaces, 2004, pp. 205–211.

[30] Carlos Busso and Shrikanth S Narayanan, *Interrelation between speech and facial gestures in emotional utterances: a single subject study*, IEEE Transactions on Audio, Speech, and Language Processing **15** (2007), no. 8, 2331–2347.

[31] Vicente E Caballo, Isabel C Salazar, María Jesús Irurtia, Benito Arias, and Stefan G Hofmann, *Measuring social anxiety in 11 countries*, 2010.

[32] Remi Cadene, Hedi Ben-Younes, Matthieu Cord, and Nicolas Thome, *Murel: Multimodal relational reasoning for visual question answering*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 1989–1998.

[33] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil, *Universal sentence encoder for English*, Proceedings of the 2018 Conference on Empirical Methods in Natural Language

Processing: System Demonstrations (Brussels, Belgium), Association for Computational Linguistics, November 2018, pp. 169–174.

[34] Haipeng Chen, Rui Liu, Noseong Park, and V.S. Subrahmanian, *Using twitter to predict when vulnerabilities will be exploited*, Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (New York, NY, USA), KDD '19, Association for Computing Machinery, 2019, p. 3143–3152.

[35] J. Chen, V. M. Patel, and R. Chellappa, *Unconstrained face verification using deep cnn features*, 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), 2016, pp. 1–9.

[36] Mingyi Chen, Xuanji He, Jing Yang, and Han Zhang, *3-d convolutional recurrent neural networks with attention model for speech emotion recognition*, IEEE Signal Processing Letters **25** (2018), no. 10, 1440–1444.

[37] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo, *Multi-label image recognition with graph convolutional networks*, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2019.

[38] Mayo Clinic, *Social anxiety disorder (social phobia)*, Aug 2017.

[39] Meredith E Coles, Cynthia L Turk, Richard G Heimberg, and David M Fresco, *Effects of varying levels of anxiety within social situations: Relationship to memory perspective and attributions in social phobia*, Behaviour Research and Therapy **39** (2001), no. 6, 651–665.

[40] Ronald E Cromwell and David H Olsen, *The bases of conjugal power*, Power in families (Ronald E Cromwell and David H Olsen, eds.), Sage, Oxford, 1975, pp. 217–232.

[41] Mohamed Dahmane and Jean Meunier, *Emotion recognition using dynamic grid-based hog features*, 2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG), IEEE, 2011, pp. 884–888.

[42] John A Daly, Anita L Vangelisti, and Samuel G Lawrence, *Self-focused attention and public speaking anxiety*, Personality and Individual Differences **10** (1989), no. 8, 903–913.

[43] Steven B. Davis and Paul Mermelstein, *Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences*, IEEE Transactions on Acoustics, Speech, and Signal Processing **28** (1980), no. 4, 357–366.

[44] Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer, *Covarep—a collaborative voice analysis repository for speech technologies*, 2014 IEEE international conference on acoustics, speech and signal processing (icassp), IEEE, 2014, pp. 960–964.

[45] Bella M. DePaulo and Gregory W Swaim, *23 lying and detecting lies in organizations*, Impression management in the organization (2013), 377.

[46] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, *BERT: Pre-training of deep bidirectional transformers for language understanding*, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (Minneapolis, Minnesota), Association for Computational Linguistics, June 2019, pp. 4171–4186.

[47] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, *Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark*, 2011 IEEE

International Conference on Computer Vision Workshops (ICCV Workshops), Nov 2011, pp. 2106–2112.

[48] Corine Dijk, Arnold AP van Emmerik, and Raoul PPP Grasman, *Social anxiety is related to dominance but not to affiliation as perceived by self and others: a real-life investigation into the psychobiological perspective on social anxiety*, Personality and Individual Differences **124** (2018), 66–70.

[49] James Price Dillard and Kyle James Tusing, *The sounds of dominance: Vocal precursors of perceived dominance during interpersonal influence*, Human Communication Research **26** (2006), no. 1, 148–172.

[50] John F. Dovidio and Steve L. Ellyson, *Decoding visual dominance: Attributions of power based on relative percentages of looking while speaking and looking while listening*, Social Psychology Quarterly **45** (1982), no. 2, 106–113.

[51] Thomas Drugman and Abeer Alwan, *Joint robust voicing detection and pitch estimation based on residual harmonics*, Twelfth Annual Conference of the International Speech Communication Association, 2011.

[52] Stefan Duffner and Christophe Garcia, *Unsupervised online learning of visual focus of attention*, Advanced Video and Signal Based Surveillance (AVSS), 2013 10th IEEE International Conference on, IEEE, 2013, pp. 25–30.

[53] ———, *Visual focus of attention estimation with unsupervised incremental learning*, IEEE Transactions on Circuits and Systems for Video Technology **26** (2016), no. 12, 2264–2272.

[54] Sri Harsha Dumpala, Rupayan Chakraborty, and Sunil Kumar Kopparapu, *Audio-visual fusion for sentiment classification using cross-modal autoencoder*,

32nd Conference on Neural Information Processing Systems (NIPS 2018), 2019, pp. 1–4.

[55] Norah Dunbar and Judee Burgoon, *Perceptions of power and interactional dominance in interpersonal relationships*, Journal of Social and Personal Relationships **22** (2005), 207–233.

[56] Norah E Dunbar, Bradley Dorn, Mohemmad Hansia, Becky Ford, Matt Giles, Miriam Metzer, Judee K Burgoon, Jay F Nunamaker, and Subrahmanian VS, *Dominance in groups: How dyadic power theory can apply to group discussions*, Detecting Trust and Deception in Group Interaction (VS Subrahmanian, Judee K Burgoon, and Norah E Dunbar, eds.), Springer, 2021, pp. 56–73.

[57] Norah E Dunbar, Matthew L Jensen, Judee K Burgoon, Katherine M Kelley, Kylie J Harrison, Bradley J Adame, and Daniel Rex Bernard, *Effects of veracity, modality, and sanctioning on credibility assessment during mediated and unmediated interviews*, Communication Research **42** (2015), no. 5, 649–674.

[58] Nicholas D Duran, Rick Dale, Christopher T Kello, Chris NH Street, and Daniel C Richardson, *Exploring the movement dynamics of deception*, Frontiers in psychology **4** (2013), 140.

[59] Martin Engilberge, Louis Chevallier, Patrick Pérez, and Matthieu Cord, *Finding beans in burgers: Deep semantic-visual embedding with localization*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3984–3993.

[60] Sergio Escalera, Xavier Baró, Jordi Vitria, Petia Radeva, and Bogdan Raducanu, *Social network extraction and analysis based on multimodal dyadic interaction*, Sensors **12** (2012), no. 2, 1702–1719.

[61] Corneliu Florea, Laura Florea, Mihai-Sorin Badea, Constantin Vertan, and Andrei Racoviteanu, *Annealed label transfer for face expression recognition.*, BMVC, 2019, p. 104.

[62] Sayan Ghosh, Eugene Laksana, Louis-Philippe Morency, and Stefan Scherer, *Representation learning for speech emotion recognition.*, Interspeech, 2016, pp. 3603–3607.

[63] G Giannakakis, Matthew Pediaditis, Dimitris Manousos, Eleni Kazantzaki, Franco Chiarugi, Panagiotis G Simos, Kostas Marias, and Manolis Tsiknakis, *Stress and anxiety detection using facial cues from videos*, Biomedical Signal Processing and Control **31** (2017), 89–101.

[64] Ivan Habernal and Iryna Gurevych, *Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional lstm*, Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2016, pp. 1589–1599.

[65] Judith A Hall, Erik J Coats, and Lavonia Smith LeBeau, *Nonverbal behavior and the vertical dimension of social relations: A meta-analysis.*, Psychological bulletin **131** (2005), 898–924.

[66] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, *Deep residual learning for image recognition*, Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[67] Richard G Heimberg, Gregory P Mueller, Craig S Holt, Debra A Hope, and Michael R Liebowitz, *Assessment of anxiety in social interaction and being observed by others: The social interaction anxiety scale and the social phobia scale*, Behavior therapy **23** (1992), no. 1, 53–73.

[68] Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov, *Improving neural networks by preventing co-adaptation of feature detectors*, 2012.

[69] Carolyn R Hodges-Simeon, Steven JC Gaulin, and David A Puts, *Different vocal parameters predict perceptions of dominance and attractiveness*, Human Nature **21** (2010), no. 4, 406–427.

[70] Chiori Hori, Takaaki Hori, Teng-Yok Lee, Ziming Zhang, Bret Harsham, John R Hershey, Tim K Marks, and Kazuhiko Sumi, *Attention-based multimodal fusion for video description*, Proceedings of the IEEE international conference on computer vision, 2017, pp. 4193–4202.

[71] Xinyue Huang and Adriana Kovashka, *Inferring visual persuasion via body language, setting, and deep features*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2016, pp. 73–79.

[72] Hayley Hung, Yan Huang, Gerald Friedland, and Daniel Gatica-Perez, *Estimating dominance in multi-party meetings using speaker diarization*, IEEE Transactions on Audio, Speech, and Language Processing **19** (2010), no. 4, 847–860.

[73] Sergey Ioffe and Christian Szegedy, *Batch normalization: Accelerating deep network training by reducing internal covariate shift*, 2015.

[74] Sergey Ioffe and Christian Szegedy, *Batch normalization: Accelerating deep network training by reducing internal covariate shift*, International conference on machine learning, PMLR, 2015, pp. 448–456.

[75] 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** (2009), no. 3, 501–513.

[76] 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** (2009), no. 3, 501–513.

[77] Dae Ung Jo, ByeongJu Lee, Jongwon Choi, Haanju Yoo, and Jin Young Choi, *Cross-modal variational auto-encoder with distributed latent spaces and associators*, 2019.

[78] Ralph Henry Johnson and J Anthony Blair, *Logical self-defense*, Idea, 2006.

[79] Jungseock Joo, Weixin Li, Francis F Steen, and Song-Chun Zhu, *Visual persuasion: Inferring communicative intents of images*, Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 216–223.

[80] Jin-Hwa Kim, Kyoung Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang, *Hadamard product for low-rank bilinear pooling*, 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings, 2017.

[81] Diederik P. Kingma and Jimmy Ba, *Adam: A method for stochastic optimization*, 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (Yoshua Bengio and Yann LeCun, eds.), 2015.

[82] Thomas N Kipf and Max Welling, *Semi-supervised classification with graph convolutional networks*, 2016.

[83] Mark L Knapp, Judith A Hall, and Terrence G Horgan, *Nonverbal communication in human interaction*, Cengage Learning, 2013.

[84] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras, *Deap: A database for emotion analysis; using physiological signals*, IEEE transactions on affective computing **3** (2011), no. 1, 18–31.

[85] Xiangnan Kong, Philip S Yu, Ying Ding, and David J Wild, *Meta path-based collective classification in heterogeneous information networks*, Proceedings of the 21st ACM international conference on Information and knowledge management, ACM, 2012, pp. 1567–1571.

[86] Anna Kukleva, Makarand Tapaswi, and Ivan Laptev, *Learning interactions and relationships between movie characters*, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 9849–9858.

[87] Srijan Kumar, Chongyang Bai, VS Subrahmanian, and Jure Leskovec, *Deception detection in group video conversations using dynamic interaction networks*, ICWSM, 2020.

[88] Srijan Kumar, Bryan Hooi, Disha Makhija, Mohit Kumar, Christos Faloutsos, and VS Subrahmanian, *Rev2: Fraudulent user prediction in rating platforms*, Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, 2018, pp. 333–341.

[89] Georgia Laretzaki, Sotiris Plainis, Ioannis Vrettos, Anna Chrisoulakis, Ioannis Pallikaris, and Panos Bitsios, *Threat and trait anxiety affect stability of gaze fixation*, Biological psychology **86** (2011), no. 3, 330–336.

[90] John Laver and Peter Trudgill, *Phonetic and linguistic markers in speech*, Social markers in speech **1** (1979), 32.

[91] Colin Lea, Rene Vidal, Austin Reiter, and Gregory D Hager, *Temporal convolutional networks: A unified approach to action segmentation*, European Conference on Computer Vision, Springer, 2016, pp. 47–54.

[92] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, Daxin Jiang, and Ming Zhou, *Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training.*, AAAI, 2020.

[93] Shan Li, Weihong Deng, and JunPing Du, *Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild*, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017, pp. 2584–2593.

[94] Yaoyu Li, Hantao Yao, Lingyu Duan, Hanxing Yao, and Changsheng Xu, *Adaptive feature fusion via graph neural network for person re-identification*, Proceedings of the 27th ACM International Conference on Multimedia, 2019, pp. 2115–2123.

[95] Zekang Li, Zongjia Li, Jinchao Zhang, Yang Feng, and Jie Zhou, *Bridging text and video: A universal multimodal transformer for video-audio scene-aware dialog*, 2021.

[96] Xinchen Liu, Wu Liu, Meng Zhang, Jingwen Chen, Lianli Gao, Chenggang Yan, and Tao Mei, *Social relation recognition from videos via multi-scale spatial-temporal reasoning*, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 3566–3574.

[97] Xiang Long, Chuang Gan, Gerard De Melo, Xiao Liu, Yandong Li, Fu Li, and Shilei Wen, *Multimodal keyless attention fusion for video classification*, Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[98] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee, *Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks*, Advances in Neural Information Processing Systems, 2019.

[99] Jon K Maner, Saul L Miller, Norman B Schmidt, and Lisa A Eckel, *Submitting to defeat: Social anxiety, dominance threat, and decrements in testosterone*, Psychological Science **19** (2008), no. 8, 764–768.

[100] M. J. Marin-Jimenez, V. Kalogeiton, P. Medina-Suarez, and A. Zisserman, *LAEO-Net: revisiting people Looking At Each Other in videos*, International Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[101] Manuel J Marin-Jimenez, Vicky Kalogeiton, Pablo Medina-Suarez, and Andrew Zisserman, *Laeo-net: revisiting people looking at each other in videos*, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 3477–3485.

[102] Benoît Massé, Silèye Ba, and Radu Horaud, *Tracking gaze and visual focus of attention of people involved in social interaction*, 2017.

[103] Iain McCowan, Jean Carletta, W Kraaij, S Ashby, S Bourban, M Flynn, M Guillemot, T Hain, J Kadlec, V Karaiskos, et al., *The ami meeting corpus*, Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research, vol. 88, 2005, p. 100.

[104] Michelle Messenger, Mark Onslow, Ann Packman, and Ross Menzies, *Social anxiety in stuttering: Measuring negative social expectancies*, Journal of fluency disorders **29** (2004), no. 3, 201–212.

[105] Amanda S Morrison and Richard G Heimberg, *Social anxiety and social anxiety disorder*, Annual review of clinical psychology **9** (2013), 249–274.

[106] Philipp Müller and Andreas Bulling, *Emergent leadership detection across datasets*, 2019.

[107] Behnaz Nojavanasghari, Deepak Gopinath, Jayanth Koushik, Tadas Baltrušaitis, and Louis-Philippe Morency, *Deep multimodal fusion for persuasiveness prediction*, Proceedings of the 18th ACM International Conference on Multimodal Interaction, ACM, 2016, pp. 284–288.

[108] Shogo Okada, Oya Aran, and Daniel Gatica-Perez, *Personality trait classification via co-occurrent multiparty multimodal event discovery*, Proceedings of the 2015 ACM on International Conference on Multimodal Interaction (New York, NY, USA), ICMI '15, 2015, pp. 15–22.

[109] Shogo Okada, Laurent Son Nguyen, Oya Aran, and Daniel Gatica-Perez, *Modeling dyadic and group impressions with inter-modal and inter-person features*, 2018.

[110] OpenAI, *Clip: Connecting text and images - openai*, `https://openai.com/blog/clip/`, 2021.

[111] Yannis Panagakis, Mihalis A Nicolaou, Stefanos Zafeiriou, and Maja Pantic, *Robust correlated and individual component analysis*, IEEE transactions on pattern analysis and machine intelligence **38** (2015), no. 8, 1665–1678.

[112] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman, *Deep face recognition*, British Machine Vision Conference, 2015.

[113] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman, *Deep face recognition*, 2015.

[114] Matthew Pediaditis, Giorgos A. Giannakakis, Franco Chiarugi, Dimitris Manousos, Anastasia Pampouchidou, Eirini Christinaki, Galateia Iatraki, Eleni Kazantzaki, Panagiotis G. Simos, Kostas Marias, and Manolis Tsiknakis, *Extraction of facial features as indicators of stress and anxiety*, 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (2015), 3711–3714.

[115] James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn, *The development and psychometric properties of liwc2015*, 2015.

[116] Jeffrey Pennington, Richard Socher, and Christopher D. Manning, *Glove: Global vectors for word representation*, Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1532–1543.

[117] Richard M Perloff, *The dynamics of persuasion: Communication and attitudes in the twenty-first century*, Routledge, 2020.

[118] Florent Perronnin, Jorge Sánchez, and Thomas Mensink, *Improving the fisher kernel for large-scale image classification*, Proceedings of the 11th European Conference on Computer Vision: Part IV (Berlin, Heidelberg), ECCV'10, Springer-Verlag, 2010, pp. 143–156.

[119] Peter Potash and Anna Rumshisky, *Towards debate automation: a recurrent model for predicting debate winners*, Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017, pp. 2465–2475.

[120] Anupama Ray, Siddharth Kumar, Rutvik Reddy, Prerana Mukherjee, and Ritu Garg, *Multi-level attention network using text, audio and video for depression prediction*, Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop, 2019, pp. 81–88.

[121] Keith Rayner, *Eye movements and attention in reading, scene perception, and visual search*, The quarterly journal of experimental psychology **62** (2009), no. 8, 1457–1506.

[122] Stephen D Reese, Oscar H Gandy Jr, and August E Grant, *Framing public life: Perspectives on media and our understanding of the social world*, Routledge, 2001.

[123] Md Sahidullah and Goutam Saha, *Design, analysis and experimental evaluation of block based transformation in mfcc computation for speaker recognition*, Speech communication **54** (2012), no. 4, 543–565.

[124] Dairazalia Sanchez-Cortes, Oya Aran, Dinesh Babu Jayagopi, Marianne Schmid Mast, and Daniel Gatica-Perez, *Emergent leaders through looking and speaking: from audio-visual data to multimodal recognition*, Journal on Multimodal User Interfaces **7** (2013), no. 1, 39–53.

[125] 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** (2012), no. 3, 816–832.

[126] Pedro Bispo Santos, Lisa Beinborn, and Iryna Gurevych, *A domain-agnostic approach for opinion prediction on speech*, Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES), 2016, pp. 163–172.

[127] Pedro Bispo Santos and Iryna Gurevych, *Multimodal prediction of the audience's impression in political debates*, Proceedings of the 20th International Conference on Multimodal Interaction: Adjunct, ACM, 2018, p. 6.

[128] George Saon, Gakuto Kurata, Tom Sercu, Kartik Audhkhasi, Samuel Thomas, Dimitrios Dimitriadis, Xiaodong Cui, Bhuvana Ramabhadran, Michael Picheny, Lynn-Li Lim, et al., *English conversational telephone speech recognition by humans and machines*, 2017.

[129] Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Galligher, and Tina Eliassi-Rad, *Collective classification in network data*, AI magazine **29** (2008), no. 3, 93.

[130] Amir Shahroudy, Tian-Tsong Ng, Yihong Gong, and Gang Wang, *Deep multimodal feature analysis for action recognition in rgb+ d videos*, IEEE transactions on pattern analysis and machine intelligence **40** (2017), no. 5, 1045–1058.

[131] Samira Sheikhi and Jean-Marc Odobez, *Investigating the midline effect for visual focus of attention recognition*, Proceedings of the 14th ACM international conference on Multimodal interaction, ACM, 2012, pp. 221–224.

[132] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman, *Deep inside convolutional networks: Visualising image classification models and saliency maps*, 2013.

[133] Karen Simonyan and Andrew Zisserman, *Very deep convolutional networks for large-scale image recognition*, International Conference on Learning Representations, 2015.

[134] Denise Haunani Solomon and Michael E Roloff, *Power and interpersonal communication*, Power in Close Relationships (2019), 241–260.

[135] Yan Song, Shuming Shi, Jing Li, and Haisong Zhang, *Directional skip-gram: Explicitly distinguishing left and right context for word embeddings*, Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers) (New Orleans, Louisiana), Association for Computational Linguistics, June 2018, pp. 175–180.

[136] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller, *Striving for simplicity: The all convolutional net*, 2014.

[137] Murray B Stein, Laine J Torgrud, and John R Walker, *Social phobia symptoms, subtypes, and severity: findings from a community survey*, Archives of general psychiatry **57** (2000), no. 11, 1046–1052.

[138] Rainer Stiefelhagen, Michael Finke, Jie Yang, and Alex Waibel, *From gaze to focus of attention*, International Conference on Advances in Visual Information Systems, Springer, 1999, pp. 765–772.

[139] Rainer Stiefelhagen, Jie Yang, and Alex Waibel, *Modeling focus of attention for meeting indexing based on multiple cues*, IEEE Transactions on Neural Networks **13** (2002), no. 4, 928–938.

[140] Rainer Stiefelhagen and Jie Zhu, *Head orientation and gaze direction in meetings*, CHI'02 Extended Abstracts on Human Factors in Computing Systems, ACM, 2002, pp. 858–859.

[141] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai, *Vl-bert: Pre-training of generic visual-linguistic representations*, International Conference on Learning Representations, 2019.

[142] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, *Deepface: Closing the gap to human-level performance in face verification*, 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1701–1708.

[143] Ben Taskar, Ming-Fai Wong, Pieter Abbeel, and Daphne Koller, *Link prediction in relational data*, Advances in neural information processing systems, 2004, pp. 659–666.

[144] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, *Attention is all you need*, Advances in neural information processing systems, 2017, pp. 5998–6008.

[145] Raviteja Vemulapalli and Aseem Agarwala, *A compact embedding for facial expression similarity*, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2019.

[146] Sunny Verma, Chen Wang, Liming Zhu, and Wei Liu, *Deepcu: integrating both common and unique latent information for multimodal sentiment analysis*, Proceedings of the 28th International Joint Conference on Artificial Intelligence, AAAI Press, 2019, pp. 3627–3634.

[147] Michael Voit and Rainer Stiefelhagen, *Deducing the visual focus of attention from head pose estimation in dynamic multi-view meeting scenarios*, Proceedings of the 10th international conference on Multimodal interfaces, ACM, 2008, pp. 173–180.

[148] Lu Wang, Nick Beauchamp, Sarah Shugars, and Kechen Qin, *Winning on the merits: The joint effects of content and style on debate outcomes*, Transactions of the Association for Computational Linguistics **5** (2017), 219–232.

[149] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He, *Non-local neural networks*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7794–7803.

[150] Yanbang Wang, Pan Li, Chongyang Bai, and Jure Leskovec, *Tedic: Neural modeling of behavioral patterns in dynamic social interaction networks*, 2021.

[151] Zhe Wu, Bharat Singh, Larry S Davis, and VS Subrahmanian, *Deception detection in videos*, Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[152] Jiaolong Yang, Peiran Ren, Dongqing Zhang, Dong Chen, Fang Wen, Hongdong Li, and Gang Hua, *Neural aggregation network for video face recognition*, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017.

[153] M Kâzım Yazici, Başaran Demir, Nilgün Tanriverdi, E Karaagaoglu, and Perin Yolac, *Hamilton anxiety rating scale: interrater reliability and validity study*, Turk Psikiyatri Derg **9** (1998), no. 2, 114–117.

[154] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency, *Tensor fusion network for multimodal sentiment analysis*, Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (Copenhagen, Denmark), Association for Computational Linguistics, September 2017, pp. 1103–1114.

[155] Matthew D Zeiler and Rob Fergus, *Visualizing and understanding convolutional networks*, European conference on computer vision, Springer, 2014, pp. 818–833.

[156] Honggang Zhang, Lorant Toth, Jun Guo, Jie Yang, et al., *Monitoring visual focus of attention via local discriminant projection*, Proceedings of the

1st ACM international conference on Multimedia information retrieval, ACM, 2008, pp. 18–23.

[157] Justine Zhang, Ravi Kumar, Sujith Ravi, and Cristian Danescu-Niculescu-Mizil, *Conversational flow in oxford-style debates*, Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016, pp. 136–141.

[158] Yu Zhang, James Qin, Daniel S Park, Wei Han, Chung-Cheng Chiu, Ruoming Pang, Quoc V Le, and Yonghui Wu, *Pushing the limits of semi-supervised learning for automatic speech recognition*, 2020.