Predicting Dominance in Group Interaction Videos

Chongyang Bai, Dartmouth College

Joint work with Maksim Bolonkin, Norah Dunbar,
Judee Burgoon, Srijan Kumar,
Jure Leskovec, and V.S. Subrahmanian

Bai, Chongyang, et al. "Predicting the Visual Focus of Attention in Multi-Person Discussion Videos." *IJCAI*. 2019.
Bai, Chongyang, et al. "Predicting dominance in multi-person videos." *IJCAI*. 2019.

Dominance is Related to…

- Personal behaviors
  - Facial expressions
  - Vocalic attributes
- **Social interactions**
  - Look at / speak to / listen to each other

Problem Setup

- Define a player's dominance score as **perceived median dominance score** from other players

Problem Setup

- Input **frontal videos** of a group of players, we **predict**:
  - Most dominant person **(MDP)** in the group
  - **MDP-Distinct**: the group has a single most dominant person
  - Pairwise dominance prediction **(PDP):** who is more dominant out of 2 people
  - **PDP-Distinct**: when the difference of Dominance scores between 2 people is larger than 1

Challenges

- How to capture social interactions of a group in videos?

- How to learn ONE model for different groups?
  - People from different groups are not directly comparable
  - Different groups have different numbers of people

Our contribution

- How to capture social interactions in a group?
  - **1. Predict who looks at who**
  - **2. Dominance Rank features for dynamic interactions**
  - **3. Multi-modality prediction: visual, audio, and social interaction**

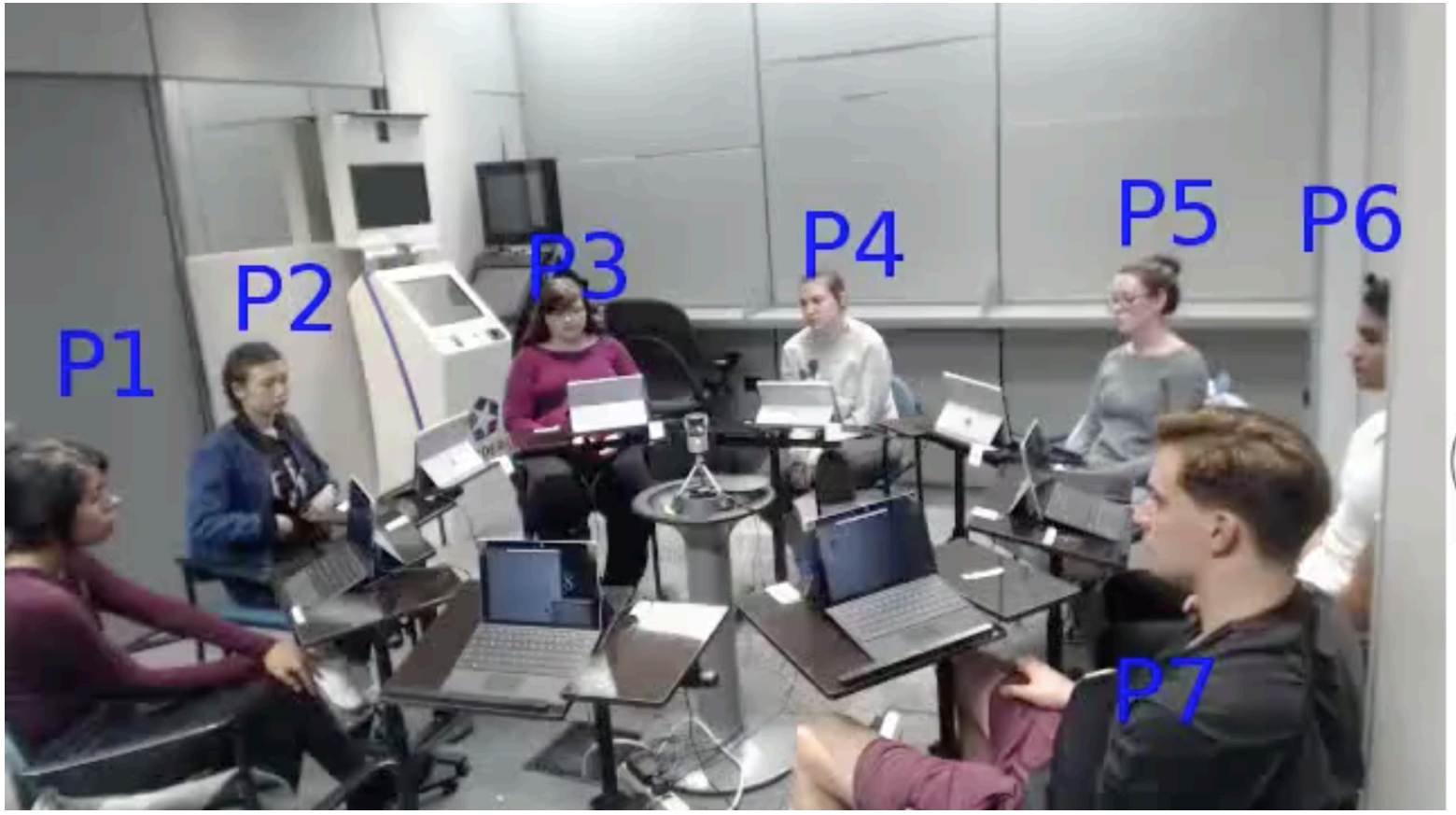
- How to learn ONE model for different groups?
  - **4. Group Dominance Prediction (GDP) Algorithm**

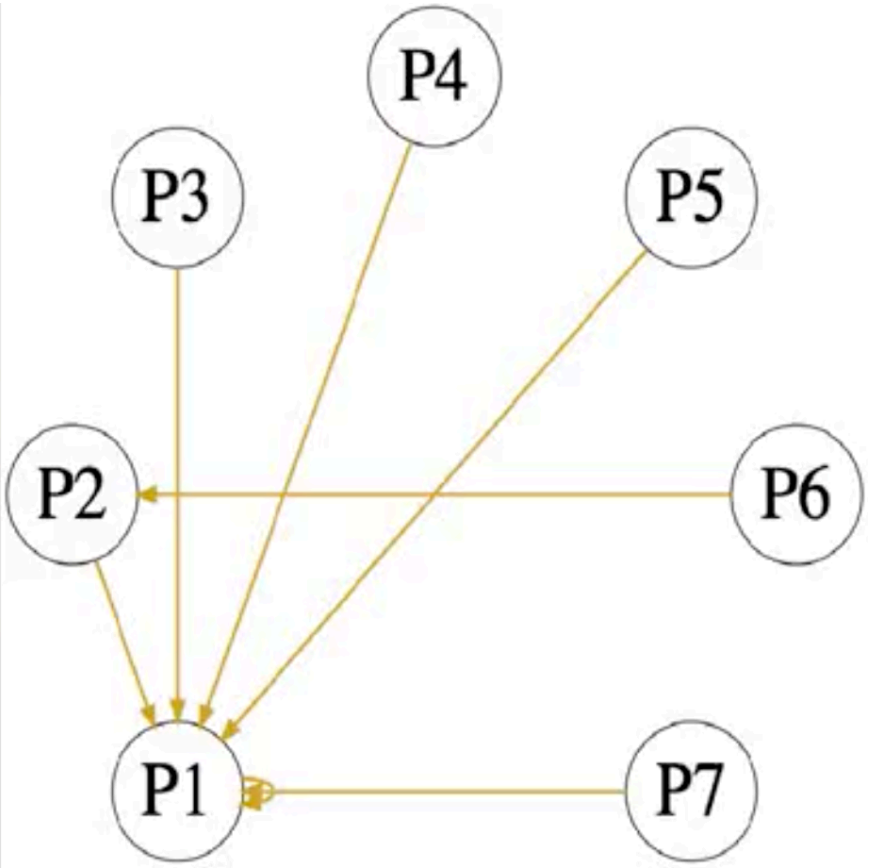
Our contribution

- How to capture social interactions in a group?
  - **1. Predict who looks at who**
  - 2. Dominance Rank features for verbal & non-verbal dynamic interactions
  - 3. Multi-modality prediction: visual, audio, and social interaction
- How to learn ONE model for different groups?
  - 4. Group Dominance Prediction (GDP) Algorithm

Predict who looks at who

C.Bai et al, Predicting the Visual Focus of Attention in Multi-person Discussion Videos, ***IJCAI'19***





Demo, network data and code at: https://home.cs.dartmouth.edu/~cy/icaf/

Why study who looks at whom?

- Conveys non-verbal interaction between people
- Identifies conversational interaction: speak, listen

Why study who looks at whom?

- Conveys non-verbal interaction between people
- Identifies conversational interaction: speak, listen

Meetings

Social Games

Interviews

Our Task

- For every 1/3rd second, predict where every player looks at:
  - other players
  - frontal tablet

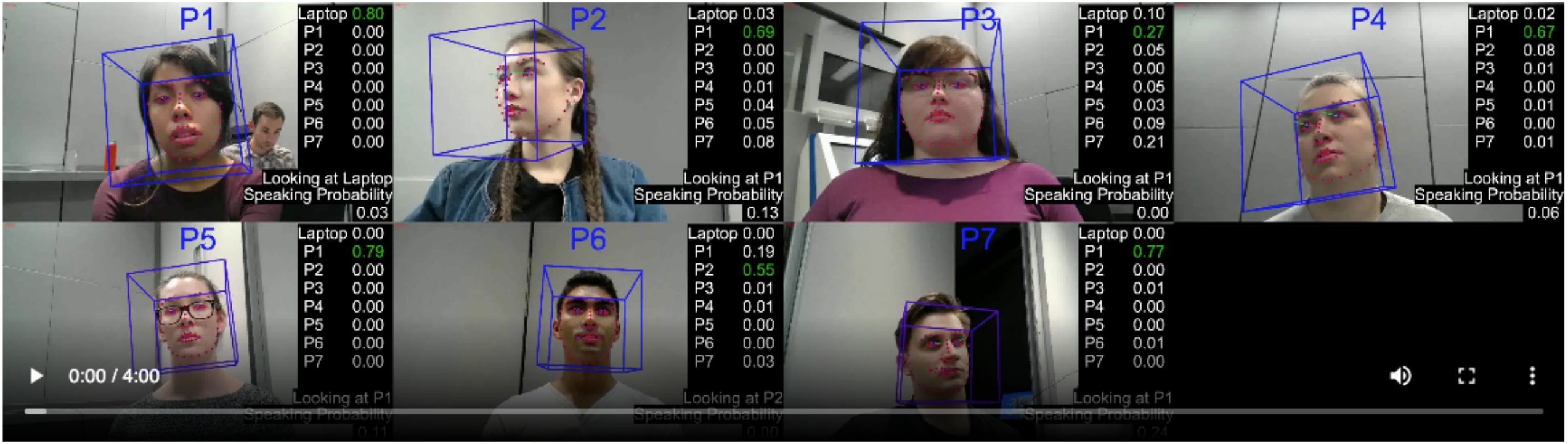
1/3rd second is the time it takes to visually focus attention (Rayner, 2009)

- A multi-class classification, with 5~8 classes

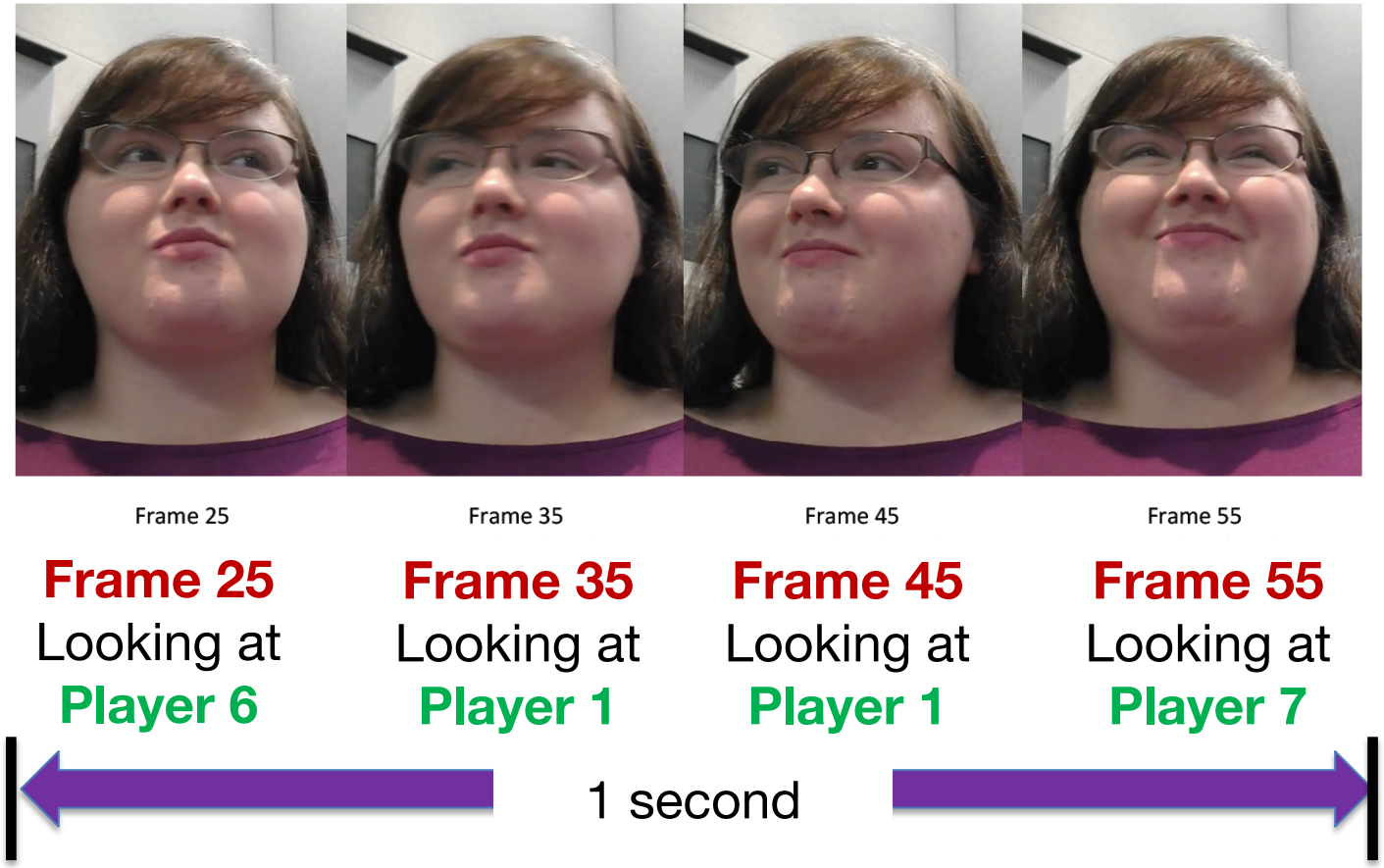
Demo: frontal view

https://home.cs.dartmouth.edu/~cy/icaf/

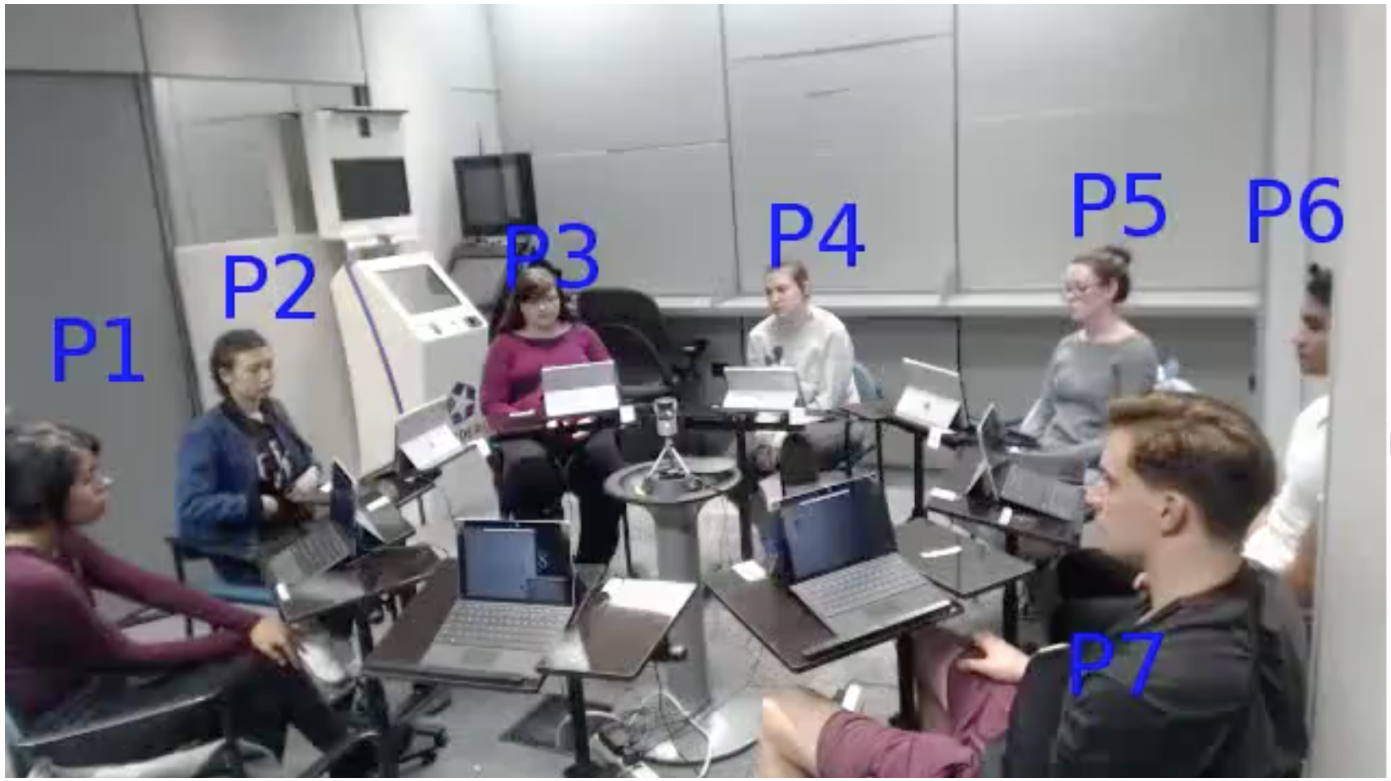
- Each person's look-at-whom probability and speaking probability is displayed, the look-at-whom network is visualized

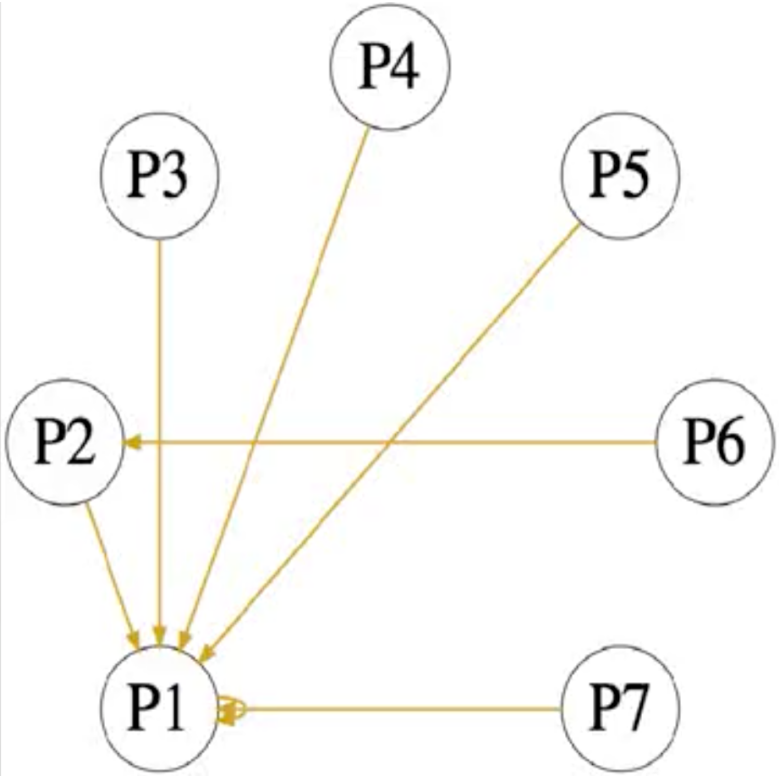


Challenge #1: Focus of attention **changes rapidly**



Challenge #2: One's focus of attention is **affected by others' verbal and non-verbal behavior**





Player 1 is speaking, so everyone except P6 looks at her

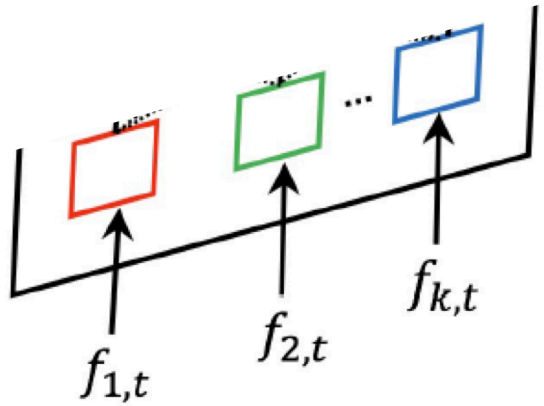
Model: Iterative Collective Attention Focus (**ICAF**)

- **Core idea:** where a player looks influences where others look and is influenced by where others look.
  - **Jointly** make the prediction for all players.

First model that uses where others are looking at the moment to come up with joint prediction of who-looks-at-whom

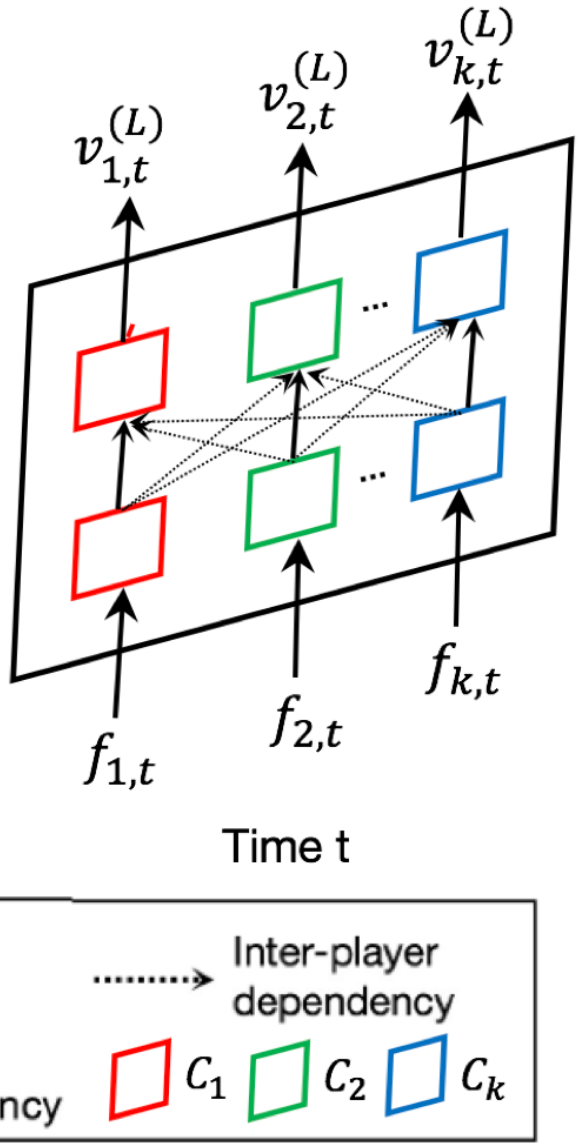
Input to classifiers

- K base classifiers for K people.
- At time t, base classifier C_i receives its raw input features $f_{i,t}$: **head pose, eye gaze vectors and speaking probabilities**



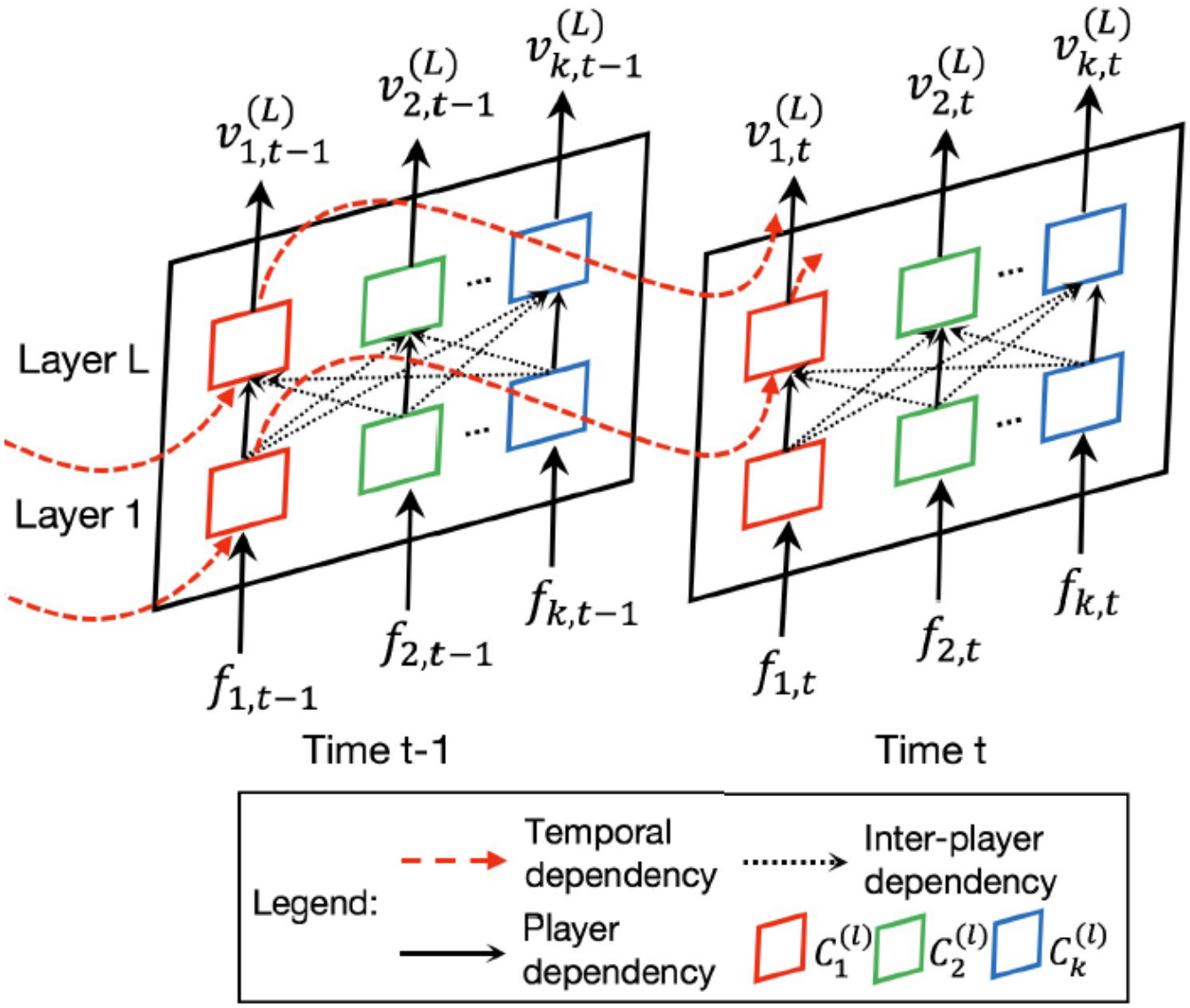
Collective Classification

- **Output from other players' classifiers** are used as additional inputs to the classifier



Temporal Component

- Each player's prediction time *t* depends on its **prediction at time *t-1***



Challenge #3: **Tedious and time-consuming annotation**

- Spend 40 hours to label total of ~2 hours of video
  - To observe eye gaze, we need to see frontal videos
  - To observe global position, we need to check the global video
- More labels are needed to learn a general model

Lightly supervised ICAF (LightICAF)

- **Intuition**:
  people most likely look at the speaker

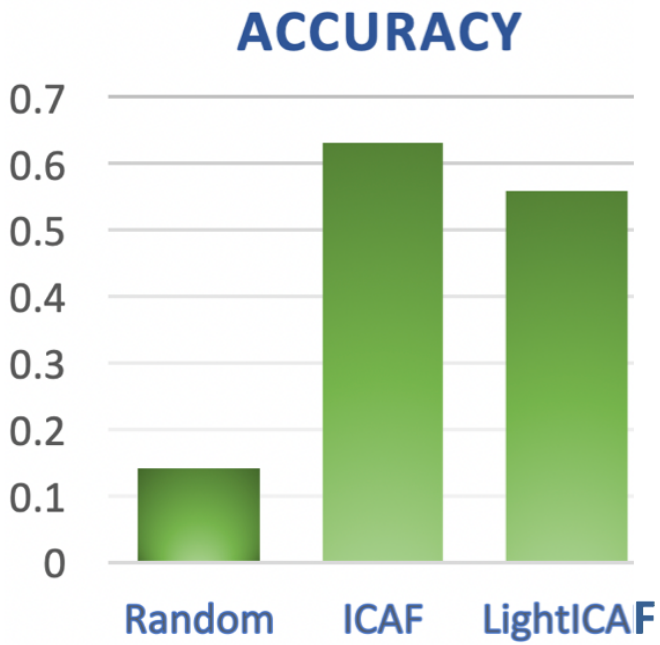
- **LightICAF:**
  - Predict continuous speaking segments
  - Use this segment as the label of all other players
  - Use the estimated label to train ICAF

Experiments: LightICAF

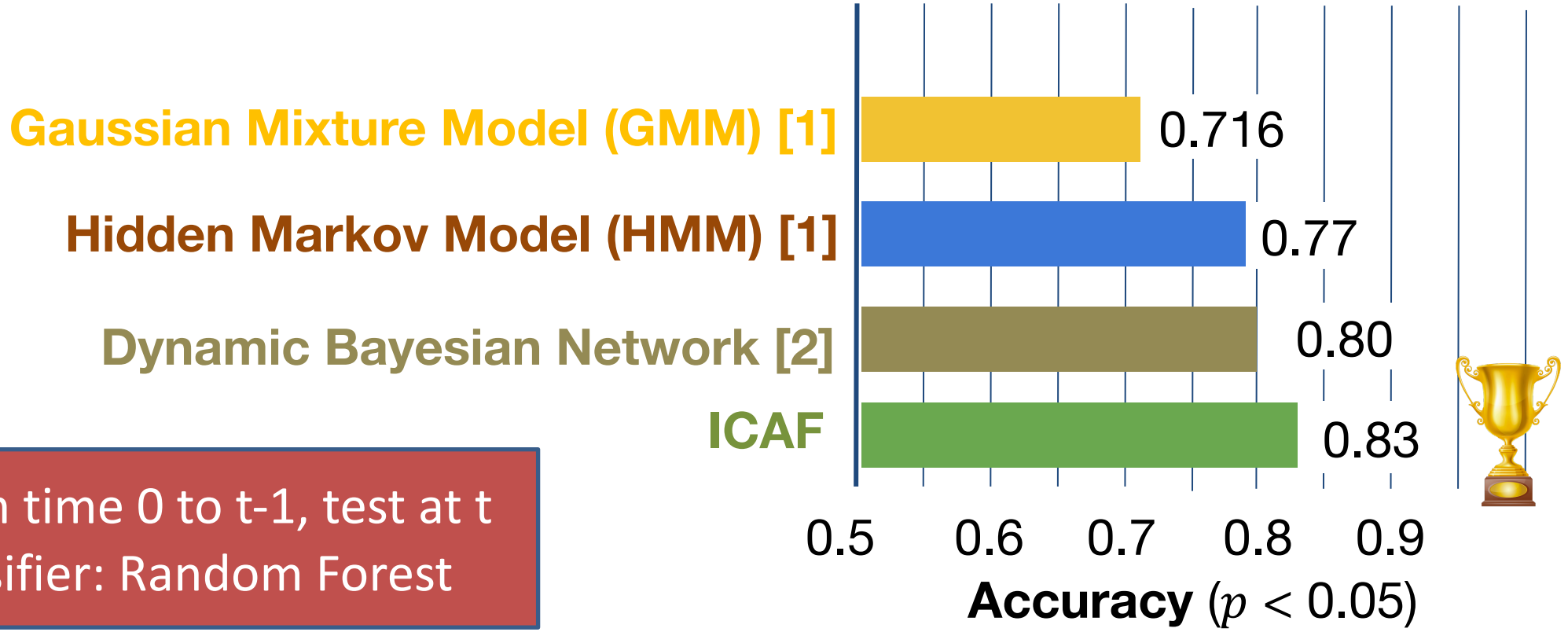
- ICAF performs slightly better than LightICAF
  - Multi-class classification with **5-8** classes.
  - Random baseline: 12.5%-20% accuracy

- **We have released 62 Look-at-whom networks (3M edges) generated by LightICAF**
  http://snap.stanford.edu/data/comm-f2f-Resistance.html

- Multi-cultural
  - Videos are from US, Zambia, HongKong, Israel, Singapore, and Fiji



Experiment: Next Focus of Attention

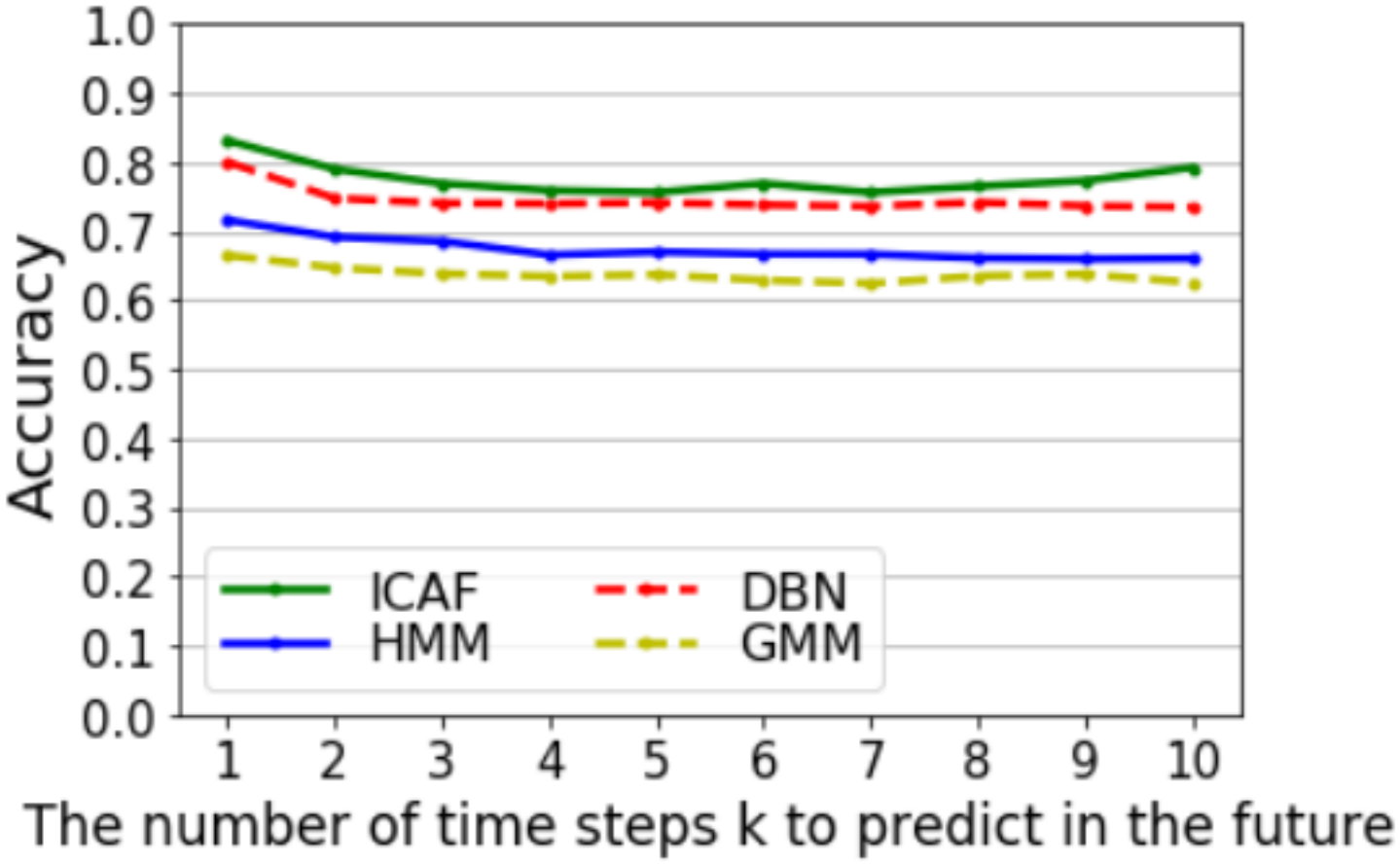


Train from time 0 to t-1, test at t
Base classifier: Random Forest

[1] Recognizing visual focus of attention from head pose in natural meeting. Ba et al., 2008

[2] Multiperson visual focus of attention from head pose and meeting contextual cues. Ba et al., 2011

Experiment: Future Focus Prediction



Train from time 0 to t-1, test at t+k
Base classifier: Random Forest

ICAF consistently performs better than others. 5% better than best baseline.

Our contribution

- How to capture social interactions in a group?
  - 1. Predict who looks at who
  - **2. Dominance Rank features for dynamic social interactions**
  - 3. Multi-modality prediction: visual, audio, and social interaction
- How to learn ONE model for different groups?
  - 4. Group Dominance Prediction (GDP) Algorithm

Family of Dominance Rank (DR) features

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

DR - PageRank weighted by interaction functions

$I(p_i, p_j)$ - interaction function

Represents how interactions between players influence the distribution of dominance in the group

N - number of players

d - damping factor

Basic Interactions

- S: speaking probabilities
- G: looking-at probabilities
- LS: looking while speaking
- LL: looking while listening

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

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

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

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

C. Bai et al, Predicting the visual focus of attention in multi-person videos, *IJCAI* 2019

Interaction functions I

- **Examples:**
  - $\boldsymbol{I(p_i, p_j) = LL(p_i, p_j) - LL(p_j, p_i)}$
    looking while listening probability difference
  - $\boldsymbol{I(p_i, p_j) = LS(p_i, p_j)/LL(p_i, p_j)}$
    (Visual Dominance Ratio, Dovidio and Ellyson*, 1982, Dunbar and Burgoon, 2005)*
    looking while speaking to looking while listening probabilities ratio

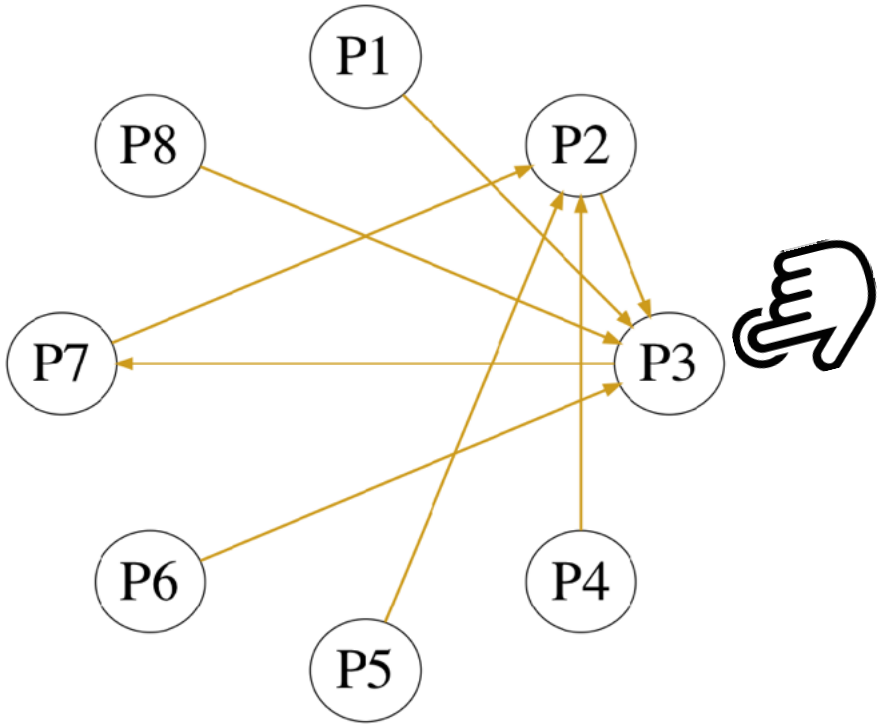
- S: speaking probabilities
- G: looking-at probabilities
- LS: looking while speaking
- LL: looking while listening

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

Interaction functions I

- **Examples:**
  - $I(p_i, p_j) = LL(p_i, p_j) - LL(p_j, p_i)$
    looking while listening probability difference

LL-LL graph



Interaction functions I

Pearson Correlation Coefficient

Spearman Correlation Coefficient

$I(p_i, p_j)$	r	ρ
$G(p_i, p_j) - G(p_j, p_i)$	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

- S: speaking probabilities
- G: looking-at probabilities
- LS: looking while speaking
- LL: looking while listening

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

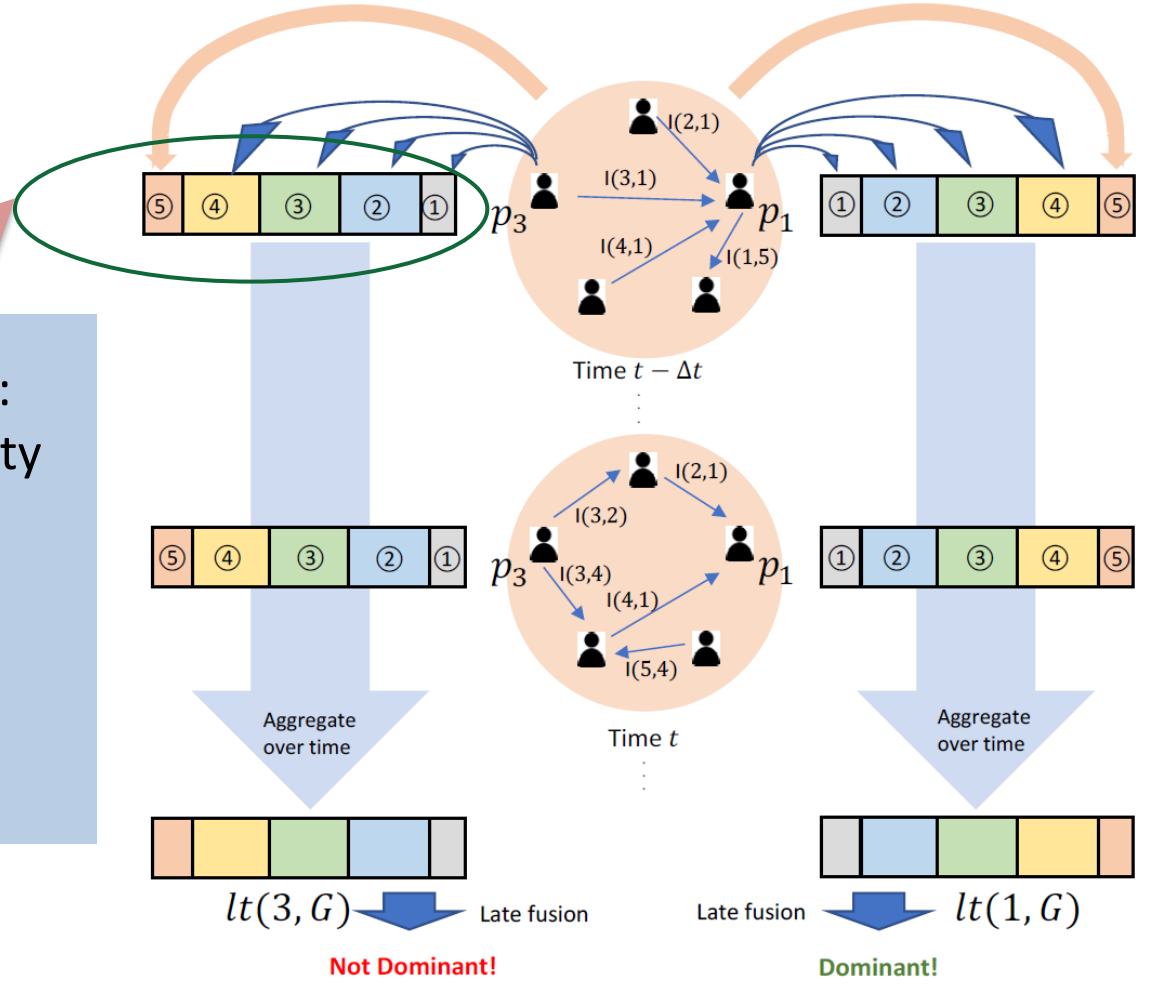
Our contribution

- How to capture social interactions in a group?
  - 1. Predict who looks at who
  - 2. Dominance Rank features for verbal & non-verbal dynamic interactions
  - **3. Multi-modality prediction: visual, audio, and social interaction**
- How to learn ONE model for different groups?
  - 4. Group Dominance Prediction (GDP) Algorithm

Dominance Ensemble Late Fusion (DELF)

Multi-modal Features:

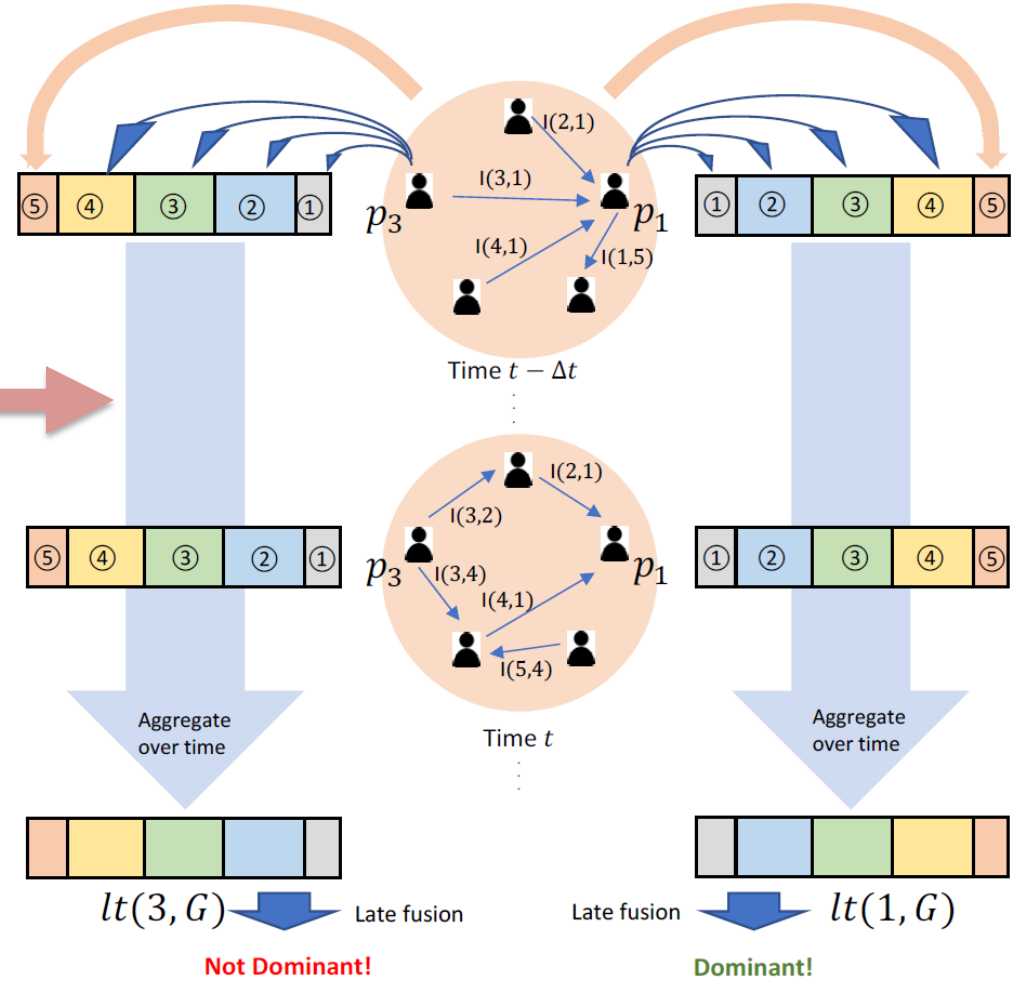
1. Speaking probability
2. Facial Action Units
3. Emotions
4. Audio features (MFCC)
5. **Dominance Rank**



Dominance Ensemble Late Fusion (DELF)

Feature aggregation, normalize feature dimension for varied video length

1. Fisher Vector
2. Histograms

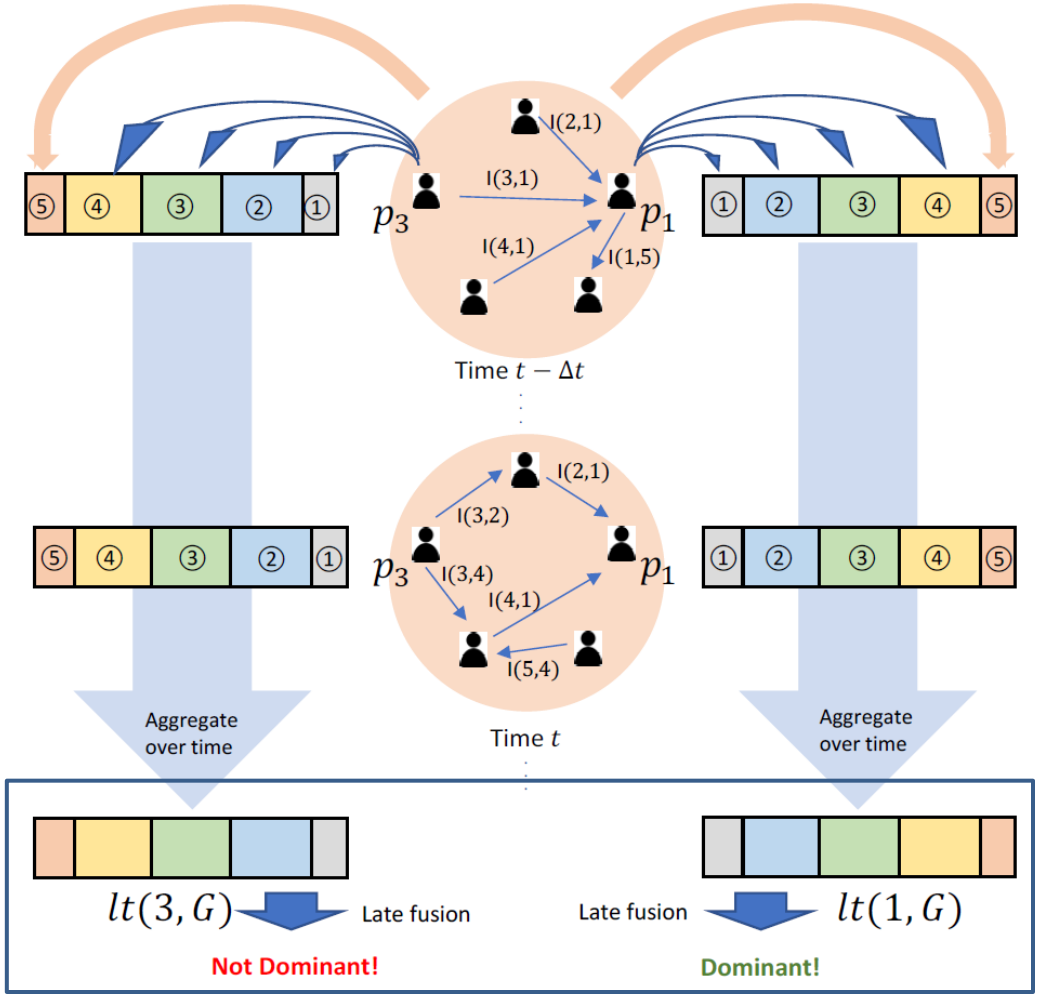


Dominance Ensemble Late Fusion (DELF)

Multimodal fusion:

$$S = \sum_{i=1}^{5} \alpha_i S_i ,$$

S_i are dominant probabilities predicted by individual features.



Experiment: DELF performance (AUC)

Our Results

Features	MDP-All	MDP-Distinct	PDP-All	PDP-Distinct
DELF	**0.791**	**0.894**	**0.874**	**0.949**
DR (LS/LL, 1 sec) + FV	0.754	0.855	0.77	0.832
DR (LS/LL, 1 sec) + Hist.	0.754	0.836	0.788	0.861
DR (LS/LL, 5 sec) + FV	**0.773**	**0.861**	0.771	0.835
DR (LS/LL, 5 sec) + Hist.	0.770	0.844	0.793	0.861
Speaking + FV	0.741	0.838	**0.853**	**0.92**
Speaking + Hist.	0.756	0.821	0.847	0.91
Baseline (speak.)	0.738	0.769	0.800	0.893
Baseline (comb.)	0.767	0.764	0.828	0.906

Single features:

(a) Dominance rank features gave the best AUC for MDP tasks

(b) Speaking probability features achieved the best AUC for PDP tasks

(c) Better than baseline [Beyan et al. 2018]

Experiment: DELF performance (AUC)

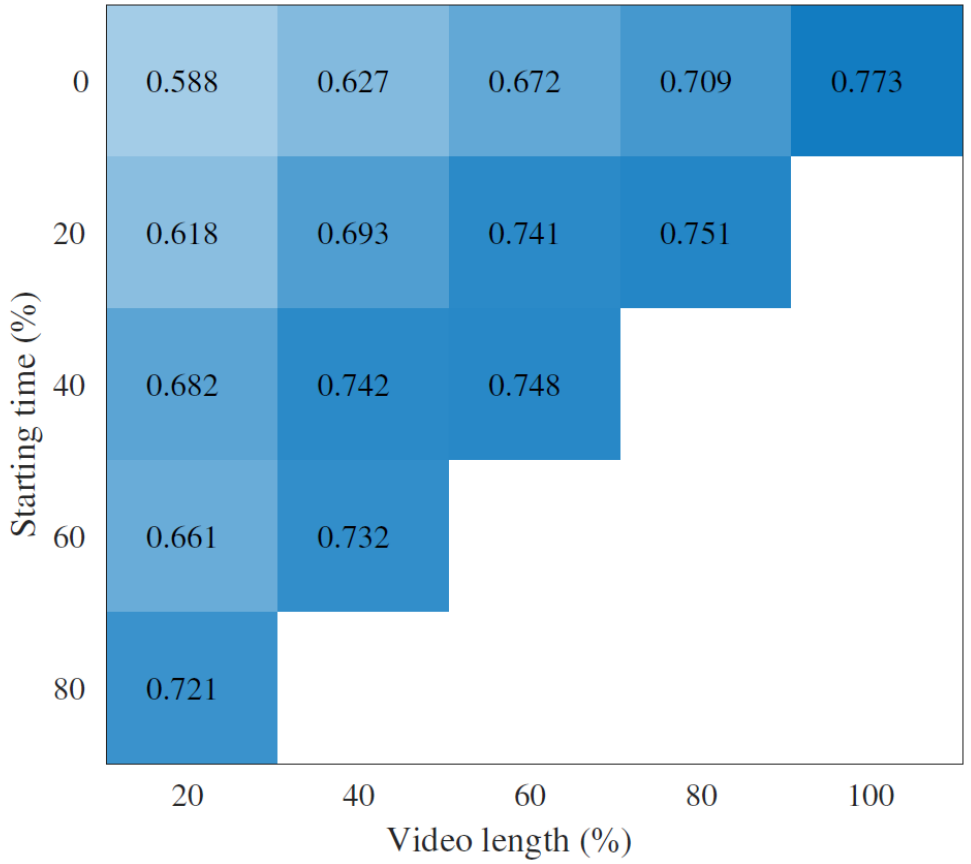
Our Results

Features	MDP-All	MDP-Distinct	PDP-All	PDP-Distinct
DELF	**0.791**	**0.894**	**0.874**	**0.949**
DR (LS/LL, 1 sec) + FV	0.754	0.855	0.77	0.832
DR (LS/LL, 1 sec) + Hist.	0.754	0.836	0.788	0.861
DR (LS/LL, 5 sec) + FV	**0.773**	**0.861**	0.771	0.835
DR (LS/LL, 5 sec) + Hist.	0.770	0.844	0.793	0.861
Speaking + FV	0.741	0.838	**0.853**	**0.92**
Speaking + Hist.	0.756	0.821	0.847	0.91
Baseline (speak.)	0.738	0.769	0.800	0.893
Baseline (comb.)	0.767	0.764	0.828	0.906

Multimodal fusion:
DELF achieved the highest AUC overall for all 4 tasks.

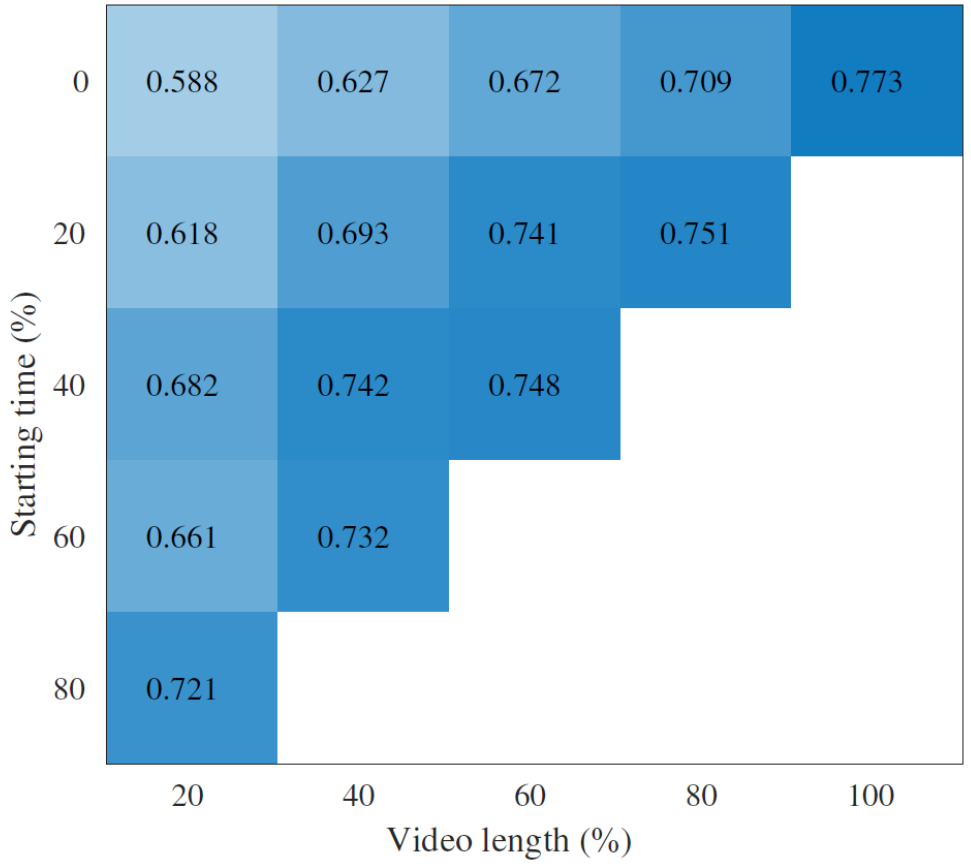
Performance (AUC) depends on the length of the video used

- Task: MDP-All
- Control two variables:
  - X: percentage of video length
  - Y: percentage of video starting time



Performance (AUC) depends on the length of the video

- Task: MDP-All
- Control two variables:
  - X: percentage of video length
  - Y: percentage of video starting time



For any length, videos closer to the end yield better prediction AUC.

The entire video gives the highest prediction result.

Our contribution

- How to capture social interactions in a group?
  - 1. Predict who looks at who
  - 2. Dominance Rank features for verbal & non-verbal dynamic interactions
  - 3. Multi-modality prediction: visual, audio, and social interaction

- How to learn ONE model for different groups?
  - **4. Group Dominance Prediction (GDP) Algorithm**

Group dominance prediction (GDP)

- **Motivation**:
  - To find the most dominant player it is desirable to compare all the players in the game
  - We only care about the most dominant player
- **Challenges:**
  - Different games have different numbers of players
  - Small number of games

Group dominance prediction (GDP)

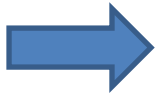
- **Solution**:
  - Form a new dataset by combining groups of 5 players

Train	Game 1	p1	p2	p3	p4	p5	p6		
	Game 2	p1	p2	p3	p4	p5	p6	p7	p8
	Game 3	p1	p2	p3	p4	p5	p6	p7	
	...	...	...	...	...	...	...	...	...
	Game 71	p1	p2	p3	p4	p5			
Test	Game 72	p1	p2	p3	p4	p5	p6		
	...	...	...	...	...	...	...	...	...
	Game 79	p1	p2	p3	p4	p5	p6	p7	

Group dominance prediction (GDP)

- **Solution**:
  - Form a new dataset by combining groups of 5 players
  - Augment the data by considering all permutations of players in the group

Train	Game 1	p1	p2	p3	p4	p5	p6		
	Game 2	p1	p2	p3	p4	p5	p6	p7	p8
	Game 3	p1	p2	p3	p4	p5	p6	p7	
	…	…	…	…	…	…	…	…	…
	Game 71	p1	p2	p3	p4	p5			
Test	Game 72	p1	p2	p3	p4	p5	p6		
	…	…	…	…	…	…	…	…	…
	Game 79	p1	p2	p3	p4	p5	p6	p7	



Train	Game 1_1	p1	p2	p3	p4	p5
	Game 1_2	p1	p2	p3	p4	p6
	Game 1_3	p1	p2	p3	p5	p6
	Game 1_4	p1	p3	p4	p5	p6
	Game 1_5	p2	p3	p4	p5	p6
	Game 1_6	p5	p4	p3	p2	p1
	…					
	Game_ 71_1					
	…					
	Game 71_120					
Test	Game 72_1	p1	p2	p3	p4	p5
	Game 72_2	p1	p2	p3	p4	p6
	Game 72_3	p1	p2	p3	p5	p6
	Game 72_4	p1	p3	p4	p5	p6
	Game 72_5	p2	p3	p4	p5	p6
	Game 72_6	p5	p4	p3	p2	p1
	…					

permutation

Group dominance prediction (GDP)

- **Solution**:
  - Form a new dataset by combining groups of 5 players
  - Augment the data by considering all permutations of players in the group

Train a model to infer probabilities of each player in each augmented games

Train	Game 1_1	p1 p2 p3 p4 p5
	Game 1_2	p1 p2 p3 p4 p6
	Game 1_3	p1 p2 p3 p5 p6
	Game 1_4	p1 p3 p4 p5 p6
	Game 1_5	p2 p3 p4 p5 p6
	Game 1_6	p5 p4 p3 p2 p1
	...	
	Game_ 71_1	
	...	
	Game 71_120	
Test	Game 72_1	p1 p2 p3 p4 p5
	Game 72_2	p1 p2 p3 p4 p6
	Game 72_3	p1 p2 p3 p5 p6
	Game 72_4	p1 p3 p4 p5 p6
	Game 72_5	p2 p3 p4 p5 p6
	Game 72_6	p5 p4 p3 p2 p1
	...	

Group dominance prediction (GDP)

- **Solution**:
  - Form a new dataset by combining groups of 5 players
  - Augment the data by considering all permutations of players in the group

Train	Game 1_1	p1 p2 p3 p4 p5
	Game 1_2	p1 p2 p3 p4 p6
	Game 1_3	p1 p2 p3 p5 p6
	Game 1_4	p1 p3 p4 p5 p6
	Game 1_5	p2 p3 p4 p5 p6
	Game 1_6	p5 p4 p3 p2 p1
	...	
	Game_ 71_1	
	...	
	Game 71_120	
Test	Game 72_1	p1 p2 p3 p4 p5
	Game 72_2	p1 p2 p3 p4 p6
	Game 72_3	p1 p2 p3 p5 p6
	Game 72_4	p1 p3 p4 p5 p6
	Game 72_5	p2 p3 p4 p5 p6
	Game 72_6	p5 p4 p3 p2 p1
	...	

During test stage
- predict the most dominant player for each augmented game.
- average the predictions of a player from all the augmented games he appeared

Group dominance prediction (GDP)

- **Solution**:
  - Form a new dataset by combining groups of 5 players, consider MDP task for those groups
  - Augment the data by considering all permutations of players in the group

Experiment: GDP in MDP task

Feature	Classif.	AUC
	MDP-All	
Speaking + FV	MLP	0.809
Speaking + FV	RF	**0.817**
DR (LS/LL, 5sec) + FV	MLP	0.783
DR (LS/LL, 5sec) + Hist.	MLP	0.772
	MDP-Distinct	
Speaking + FV	MLP	**0.936**
Speaking + FV	RF	0.902
DR (LS/LL, 5sec) + FV	RF	0.878
DR (LS/LL, 5sec) + FV	MLP	0.850

Features	MDP-All	MDP-Distinct
DELF	**0.791**	**0.894**

In both MDP tasks, the GDP algorithm outperforms our ensemble DELF model

ELEA dataset

https://www.idiap.ch/dataset/elea

Video content:

- Cooperative setting
- Winter survival task:
  3-4 people in a group,
  decide to arrange items in
  the order of importance

Labels:

- In-group ratings of
  dominance scores (PDom)

Human study:

- Out-of-group ratings by
  independent humans

Experiment in ELEA dataset

Task 1: predict who are more dominant than others (compare to median score)

DR features outperform baselines and human study predictions

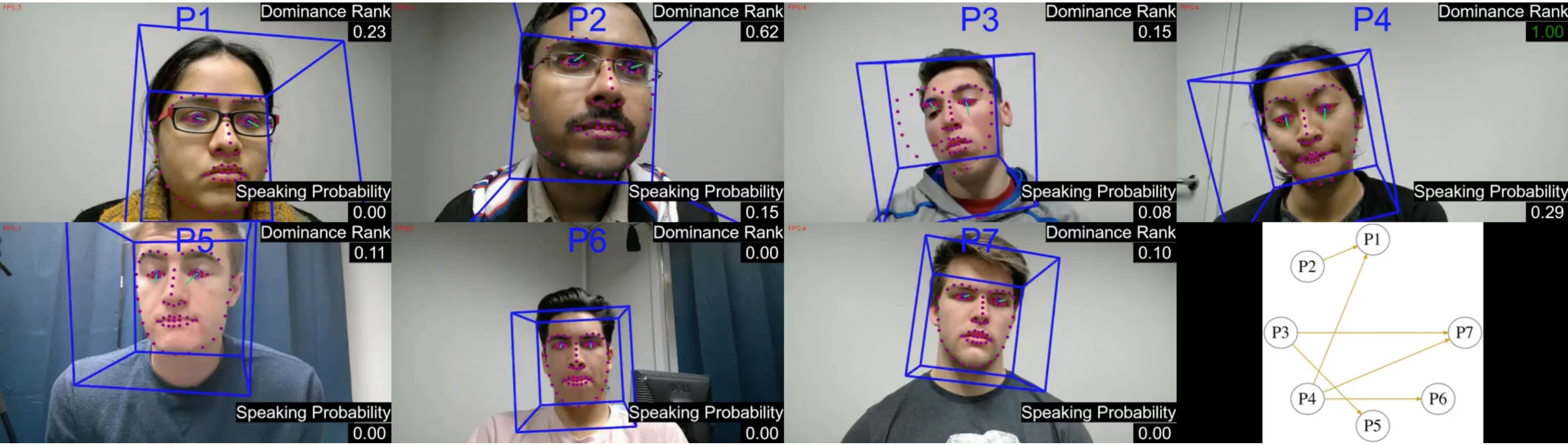
Method	PDom
[Okada *et al.*, 2018]	58.82
[Aran and Gatica-Perez, 2013]	65.69
[Okada *et al.*, 2015]	67.65
DR (LS/LL) + FV (ours)	**76.47**
DR (LS/LL) + Hist. (ours)	74.51
Human scores	68.63
[Sanchez-Cortes *et al.*, 2012]	74.10
DR (LS/LL) + FV (ours)	**77.50**
DR (LS/LL) + Hist. (ours)	76.50
Human scores	**78.43**

Experiment in ELEA dataset

Method	PDom
[Okada *et al.*, 2018]	58.82
[Aran and Gatica-Perez, 2013]	65.69
[Okada *et al.*, 2015]	67.65
DR (LS/LL) + FV (ours)	**76.47**
DR (LS/LL) + Hist. (ours)	74.51
Human scores	68.63
[Sanchez-Cortes *et al.*, 2012]	74.10
DR (LS/LL) + FV (ours)	**77.50**
DR (LS/LL) + Hist. (ours)	76.50
Human scores	**78.43**

Task 2: predict the most dominant person
DR features outperform baselines, but human scores are slightly better

Demo



Demo at: http://home.cs.dartmouth.edu/~cy/dom/

Conclusion

- Predict who looks at who in group interaction videos
  - Release the interaction network dataset
- Study two classes of dominance-related problems
  - Most dominant person
  - more dominant person
- Propose a novel family of **Dominance Rank** features
- Develop **DELF** model and **GDP** algorithm
- Beat baselines in Resistance and ELEA dataset

Q & A

Dataset, demo and code:
http://home.cs.dartmouth.edu/~cy/icaf/
http://home.cs.dartmouth.edu/~cy/dom/

Contact:
cy@cs.dartmouth.edu
www.cs.dartmouth.edu/~cy