Predicting Dominance in Multi-person Videos

Chongyang Bai[1], **Maksim Bolonkin**[1], **Srijan Kumar**[2], **Jure Leskovec**[2], **Judee Burgoon**[3], **Norah Dunbar**[4] and **V. S. Subrahmanian**[1]
[1]Dartmouth College
[2]Stanford University
[3]University of Arizona
[4]University of California Santa Barbara
{cy, mbolonkin}@cs.dartmouth.edu, {srijan, jure}@cs.stanford.edu,
judee@email.arizona.edu, ndunbar@comm.ucsb.edu, vs@dartmouth.edu

Abstract

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. We combine features not previously used for dominance prediction (e.g., facial action units, emotions), with a novel ensemble-based approach to solve these two problems. We test our models against four competing algorithms in the literature on two datasets and show that our results improve past performance. 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. Ablation testing shows that Dominance Rank features play a key role.

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 paper, 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.*, 2009] and [Aran and Gatica-Perez, 2010], 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 paper makes three novel contributions. First, we propose a family of *Dominance Rank (DR)* 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 four problems. Third, we propose the *Group Dominance Prediction* (GDP) algorithm to solve MDP-All and MDP-Distinct.

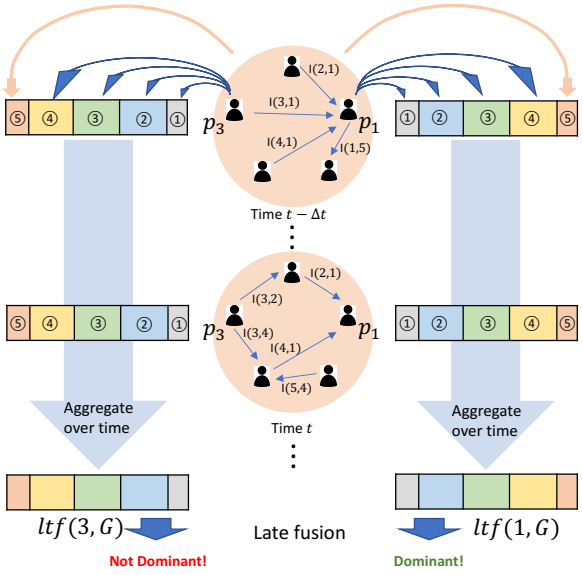

Figure 1: Our approach. In group G, for each player p at time t we have individual short-term features (1–4) and Dominance Rank features (5) based on the group interaction. We aggregate each kind of features over time to get long-term features for each player. Finally, we use the late fusion approach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

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 collected 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 [Sanchez-Cortes *et al.*, 2012], which shows small groups (3–4 people) involved in a winter survival task. The Resistance and 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). In contrast, the ELEA dataset involves a coop-

erative 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 1 depicts our approach to the four problems we study in this paper. 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 4) for player p during time slice t. The basic features fall into four categories: speech-related features [Escalera *et al.*, 2012], facial action unit features [Baltrusaitis *et al.*, 2018], emotion-related features from Amazon's Rekognition, and Mel-Frequency Cepstral Coefficient (MFCC) features [Davis and Mermelstein, 1980]. 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 [Perronnin *et al.*, 2010] and histograms. A similar aggregation is also applied to the Dominance Rank features to get a vector $ltf(p, G)$ of the long-term features for player p in game G.

We then develop predictive models based on each type of long-term feature 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 feature 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 preferable to treating them independently. This naturally sets up a classification problem where each player's ltf 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.

2 Related Work

Social scientists have studied dominance for years. [Dillard and Tusing, 2006] found that dominance is correlated with speaking rates and voice characteristics (frequency, amplitude, etc.). Visual cues like people looking at each other, body movements, gestures and facial expressions are also indicators of dominance in social interactions [Hall *et al.*, 2005; Dunbar and Burgoon, 2005]. Moreover, [Dovidio and Ellyson, 1982] studied the relationship between dominance and the combination of looking-while-speaking and looking-while-listening periods. As a result, most work on predicting dominance use features based on these social science findings. Some early prediction papers use discrete features based on binary speaking variables (for a given time segment, does the person speak in it or not). These features include statistics such as total speaking length, total speaking turns, and total successful interruptions [Jayagopi *et al.*, 2009; Aran and Gatica-Perez, 2010]. In addition, [Sanchez-Cortes *et al.*, 2012] use prosodic features such as energy and pitch variation. We differ from these features by not converting continuous speaking variables into binary ones. Additionally, we differ from these efforts because we use Mel-Frequency Cepstral Coefficient (MFCC) [Davis and Mermelstein, 1980] features, which are a richer representation of audio features than prosodic features.

Other works extensively use visual features in the form of discrete variables. [Aran and Gatica-Perez, 2010] and [Jayagopi *et al.*, 2009] use statistics on overall visual activity (binary variable - person either moves or not). [Sanchez-Cortes *et al.*, 2012] and [Beyan *et al.*, 2018] analyze more fine-grained activity such as head and body movements, and gestures. In addition to these, a set of proposed methods use gaze-related features [Aran and Gatica-Perez, 2013; Okada *et al.*, 2015; Okada *et al.*, 2018].

Some past works attempt to capture dyadic or group-level information that might be relevant for the task. [Aran and Gatica-Perez, 2013] and [Okada *et al.*, 2018] mine co-occurring events in the sequence of visual and audio features of individual players. [Sanchez-Cortes *et al.*, 2012] use collective classification using speaking turns as weights on edges in the graph of the group.

We build upon these prior important works. The principal innovative class of representations that we introduce are Dominance Rank features that encode aspects of dynamic group interaction, building on basic audio-visual features. We validate our methods on two group-interaction datasets, and in ablation testing, we further show that Dominance Rank features turn out to be important for predicting dominance.

3 Dataset and Task Descriptions

Resistance dataset. Our team developed a Resistance dataset of videos of people participating in a variation of the role-playing party game The Resistance. Each game has 5–8 players secretly divided into two teams: spies and resistance members (approximately 40% of players are spies). Spies have full information about other players' roles, whereas members of resistance do not know other players' roles. Games proceed in rounds (called "missions") involving a nomination and election of a mission leader, nomination and approval of a mission team, and the mission itself. Spies and resistance members have opposite incentives in the game: spies want to get elected on as many missions as possible

in order to fail missions and thus earn points. Resistance members want to discover spies as early as possible and prevent spies from getting on mission teams. To achieve their goals, players on both teams need to be assertive and dominant, while spies also need to hide their true intentions. The data was collected in a variety of locations with different cultures (three locations in the USA, and one each in Israel, Zambia, and Singapore) and consisted of 33 games involving 233 players, each player appearing in exactly one game. Games typically last for 2–8 rounds. After every two rounds of the game, all players fill in a questionnaire, which asks them to rate other players' dominance in the past two rounds (we will refer to this period of time as "round" for simplicity). Therefore, 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. Our dataset contains 79 rounds in all.

ELEA dataset. In addition to the Resistance dataset, we used the ELEA dataset developed by [Sanchez-Cortes *et al.*, 2012]. This dataset contains videos of groups of people (3–4 persons in a group, 27 groups) participating in a winter survival task: the participants were given 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. Videos are accompanied by survey results measuring participants' dominance: perceived dominance (PDom), ranked dominance (RDom).

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 label 1 if the player has the highest perceived dominance score among all the players in this round, and 0 otherwise. In our setting, 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 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*.

4 DELF and GDP Algorithms

We have already provided a brief overview of our approach in Section 1. We first describe our short term traditional 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.

$I(p_i, p_j)$	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

Table 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 $p < 0.01$.

4.1 Basic Short-term Features

Past work has shown that speech-related cues [Dillard and Tusing, 2006; Beyan *et al.*, 2018] and gaze [Hall *et al.*, 2005; Sanchez-Cortes *et al.*, 2012] are closely related to perceived dominance of a person. We also use facial expressions and emotions as additional signals for visual dominance [Dunbar and Burgoon, 2005]. 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 player p_i is speaking during time interval t. This probability is estimated from the person's lip movement [Escalera *et al.*, 2012].
- *Gazing probability* $g_t(p_i, p_j)$ is an estimate of the probability that player p_i looks at player p_j for every $\Delta t = 0.33$ seconds [Ba and Odobez, 2011; Rayner, 2009].
- *Facial Action Units* scores (FAUs) capture the intensity of 17 action units in the given frame. These values are produced using OpenFace [Baltrusaitis *et al.*, 2018].
- *Emotion scores* are the estimates of intensity of eight emotions and two facial traits (smile, open eyes) produced by Amazon Rekognition.
- *Audio features* are represented by Mel-Frequency Cepstral Coefficients, which are widely used in audio analysis [Davis and Mermelstein, 1980].

Dominance Rank Features

Previous research on dominance and leadership analysis shows that dyadic statistics are correlated with dominance ([Aran and Gatica-Perez, 2013; Okada *et al.*, 2018; Sanchez-Cortes *et al.*, 2012]). We propose a family of short-term 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 p_i and p_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.

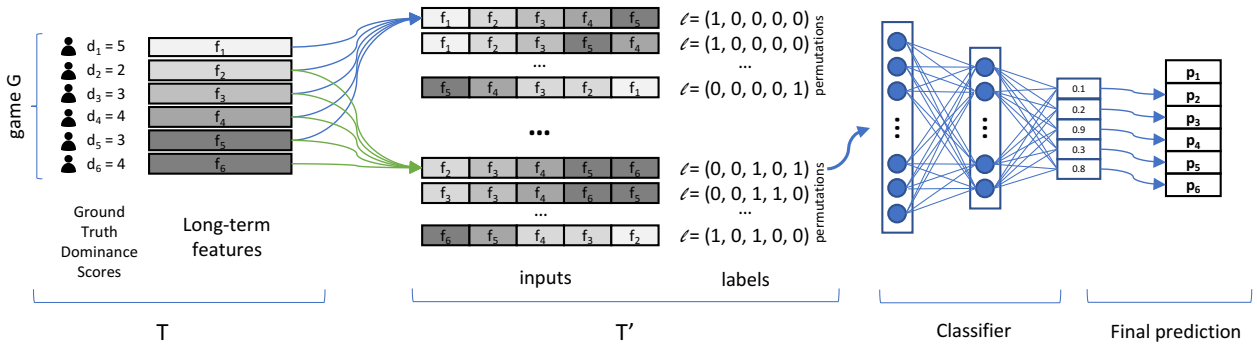

Figure 2: GDP algorithm. Given a dataset T, for every game G we form all possible groups of 5 players, form labels based on the players scores in every group, and concatenate long-term features for players to get group feature. We also augment the dataset with all possible permutations of the players. 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.

function I in a given time period t is defined as:

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

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). 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 $d = 0.85$. We compute the Dominance Rank the same way as the PageRank: by building a system of N equations with N unknowns and iteratively solving it.

Interaction Functions. We define a family of interaction functions, each of which yields a different dominance rank function R_{dom} when used in Equation 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:

$$S(p_i) = \frac{1}{k} \sum_{t=t_1}^{t_2} s_t(p_i) , \quad (2)$$

$$G(p_i, p_j) = \frac{1}{k} \sum_{t=t_1}^{t_2} g_t(p_i, p_j) , \quad (3)$$

$$LS(p_i, p_j) = \frac{1}{k} \sum_{t=t_1}^{t_2} g_t(p_i, p_j) s_t(p_i) , \quad (4)$$

$$LL(p_i, p_j) = \frac{1}{k} \sum_{t=t_1}^{t_2} g_t(p_i, p_j) s_t(p_j) , \quad (5)$$

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 use time periods of 1 and 5 seconds for Dominance Rank features, thus k is 3 or 15. Based on these values, we define a set of interaction functions (Table 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). To compare Dominance Rank (Equation 1) for players from different games, we normalize these values to be in $[0, 1]$. Table 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 11 respectively) demonstrate strong correlation with ground truth dominance scores.

4.2 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 [Perronnin *et al.*, 2010]. Note that each round may have a different duration and hence the number of *stf* 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.

Histogram features. We compute a histogram feature for every *stf* 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 $\{stf(p_i, t_1, G), \ldots, stf(p_i, t_T, G)\}$ of all feature values for all short intervals over the round. We build a histogram of short-term features $\mathcal{V}_{stf} = \langle v_1, v_2, \ldots, v_b \rangle$, where v_l are fre-

quencies 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).

4.3 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 late fusion as

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

where $\sum_{i=1}^{5} \alpha_i = 1$. The values of late fusion weights α_i are obtained by grid-search and cross-validation on the training set alone. The best classifier for each of the five types of features is determined by exhaustive search through all possible combinations of classifiers.

4.4 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 1 and also on Figure 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. But in the Resistance dataset, 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 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 test 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.

Algorithm 1: GDP(T: training set, ltf : long term feature type)

1 $T' = \emptyset$
2 **foreach** *game* $G \in T$ **do**
3 G_5 = set of all subsets containing 5 players from G
4 Π_5 = set of all permutations of 5 elements
5 **foreach** $(i_1, i_2, i_3, i_4, i_5) \in G_5$ **do**
6 **foreach** $\pi \in \Pi_5$ **do**
7 $(j_1, j_2, j_3, j_4, j_5) = \pi(i_1, i_2, i_3, i_4, i_5)$
8 $\text{Dom} = \arg\max_{p_{j_k}} \text{GT\_Dom\_Score}(p_{j_k})$
9 $\text{input} = \text{concat}\,(ltf(p_{j_k}) \mid k = 1, \ldots, 5)$
/* $\mathbb{1}_{\text{Dom}}(x)$ is indicator function */
10 $\text{label} = \text{concat}\,(\mathbb{1}_{\text{Dom}}(p_{j_k}) \mid k = 1, \ldots, 5)$
11 $T' = T' \cup \{(\text{input}, \text{label})\}$
12 **end**
13 **end**
14 **end**
15 Train a classifier on T'

5 Experimental Results with Resistance Data

Setup. We split the Resistance 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 dataset is inherently imbalanced, we report the mean AUC over 10 folds and use it to compare models. But we also report False Positive rate (FPR) and Accuracy (Acc.) as reported in past works [Beyan *et al.*, 2018; Sanchez-Cortes *et al.*, 2012; Okada *et al.*, 2015; Okada *et al.*, 2018; Aran and Gatica-Perez, 2013].

5.1 Binary Prediction with DELF

Table 2 shows the result of applying DELF and single ltf-classifiers to the four problems. We compare our models with two baselines adapted from the recent paper by Beyan et al. [Beyan *et al.*, 2018]: 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 our single-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 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 gaze information produce poor results (not reported in the Table 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 the difference in dominance between players is more prominent.

	MPD-All			MDP-Distinct			PDP-All			PDP-Distinct		
Features	**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.	0.770	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	0.030	0.875	**0.853**	0.261	0.762	**0.92**	0.179	0.825
Speaking + Hist.	0.756	0.066	0.770	0.821	0.150	0.879	0.847	0.258	0.778	0.91	0.164	0.860
Baseline (speak.)	0.738	0.103	0.730	0.769	0.200	0.879	0.800	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	0.290	0.759	0.906	0.168	0.863

Table 2: Resistance Dataset: 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 feature with timespan of 1 and 5 seconds. Details on DELF for each task are presented in Table 3. Baseline is adapted from [Beyan *et al.*, 2018].

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

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 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.

We see from Table 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.

Predictions of our models depend on the size of the game portion considered and what part of the game is considered. Figure 3 shows how the LS/LL Dominance Rank feature (best performing feature in Table 2) performs on MDP-All task when we process only a portion of each video (varied from 20% to 100%). We found that considering only 20% of the video drops the predictive performance of our models up to 0.2. Performance grows with increased video length reach-

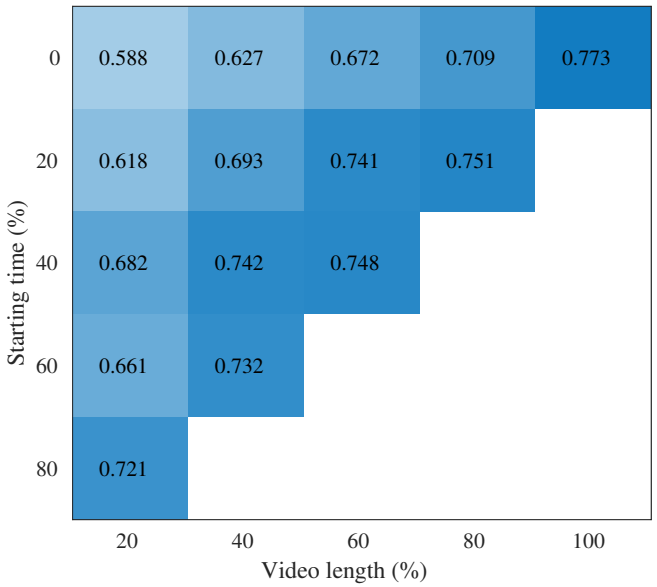

Figure 3: MDP-All: performance depending on the length of the video portion. For the best performing long-term feature (LS/LL, 5 sec + FV) AUC for the entire video is the highest, and for smaller portions of the video predictive performance drops. For any length of the video, parts closer to the end yield better AUC.

ing 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 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 recent events better. The analysis shows, however, that for the best performance it's important to consider the entire video.

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 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 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.

5.2 GDP Algorithm Performance

We tested GDP algorithm on the Resistance dataset. We used two classifiers: Multilayer Perceptron (MLP) with two layers, and Random Forest (RF) with 50 estimators. As shown in Table 4, GDP outperforms all the baselines as well as the various strong settings of DELF.

6 ELEA Corpus Experiments

We conducted further tests on the ELEA dataset [Sanchez-Cortes *et al.*, 2012] 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 [Okada *et al.*, 2018; Aran and Gatica-Perez, 2013; Okada *et al.*, 2015] we assign every participant a binary label by thresholding her dominance score by the median value, and (2) as in [Sanchez-Cortes *et al.*, 2012] we solve the MDP-All task, i.e., finding the most dominant participant in every group. As in these works we use the leave-one-game-out method for training and testing classifiers. Table 5 shows that our proposed Dominance Rank feature outperforms strong baselines in existing work.

Feature	Classif.	AUC	FPR	Acc.
MDP-All				
Speaking + FV	MLP	0.809	0.219	0.745
Speaking + FV	RF	**0.817**	0.133	0.770
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	0.850	0.065	0.889

Table 4: GDP algorithm results. GDP in most cases improves over the corresponding single *ltf* binary prediction, as well as outperforming best DELF model.

Method	PDom	RDom
[Okada *et al.*, 2018]	58.82	64.71
[Aran and Gatica-Perez, 2013]	65.69	59.80
[Okada *et al.*, 2015]	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.*, 2012]	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: 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.

We used the average dominance scores assigned to participants by the independent viewers not participating in the task as *human scores*. In Table 5 we can see that our proposed features outperform humans on the task of detecting participants who are more dominant than others. Humans, however, are better at detecting the most dominant participant in a group, although our model achieves comparable accuracy.

7 Conclusion

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.

Acknowledgements

This work was funded by ARO Grant W911NF1610342.
Role of Authors. Authors Burgoon and Dunbar designed the Resistance-style game, designed how the game would be run face to face, and collected the Resistance data. The remaining authors designed the feature extraction and machine learning algorithms and software, and designed/ran all experiments.

References

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

[Aran and Gatica-Perez, 2013] Oya Aran and Daniel Gatica-Perez. One of a kind: Inferring personality impressions in meetings. In *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, ICMI '13, pages 11–18, New York, NY, USA, 2013.

[Ba and Odobez, 2011] 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(1):101–116, 2011.

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

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

[Davis and Mermelstein, 1980] 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(4):357–366, August 1980.

[Dillard and Tusing, 2006] James Price Dillard and Kyle James Tusing. The sounds of dominance: Vocal precursors of perceived dominance during interpersonal influence. *Human Communication Research*, 26(1):148–172, 01 2006.

[Dovidio and Ellyson, 1982] 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(2):106–113, 1982.

[Dunbar and Burgoon, 2005] Norah Dunbar and Judee Burgoon. Perceptions of power and interactional dominance in interpersonal relationships. *Journal of Social and Personal Relationships*, 22:207–233, 04 2005.

[Escalera *et al.*, 2012] Sergio Escalera, Xavier Baró, Jordi Vitria, Petia Radeva, and Bogdan Raducanu. Social network extraction and analysis based on multimodal dyadic interaction. *Sensors*, 12(2):1702–1719, 2012.

[Hall *et al.*, 2005] Judith A Hall, Erik J Coats, and Lavonia Smith LeBeau. Nonverbal behavior and the vertical dimension of social relations: A meta-analysis. *Psychological bulletin*, 131:898–924, 12 2005.

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

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

[Okada *et al.*, 2018] Shogo Okada, Laurent Son Nguyen, Oya Aran, and Daniel Gatica-Perez. Modeling dyadic and group impressions with inter-modal and inter-person features. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2018.

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

[Rayner, 2009] Keith Rayner. Eye movements and attention in reading, scene perception, and visual search. *The quarterly journal of experimental psychology*, 62(8):1457–1506, 2009.

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