IDIAP RESEARCH REPORT

Predicting the Dominant Clique in Meetings through Fusion of Nonverbal Cues

Dinesh Babu Jayagopi [a b] Hayley Hung [a]
Chuohao Yeo [a c] Daniel Gatica-Perez [a b]

IDIAP–RR 08-08

July 2008

[a] IDIAP Research Institute, Martigny, Switzerland
[b] Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
[c] Department of Computer Science, University of California, Berkeley

IDIAP Research Institute **www.idiap.ch**
Av. des Prés–Beudin 20 P.O. Box 592 1920 Martigny – Switzerland
Tel: +41 27 721 77 11 Fax: +41 27 721 77 12 Email: info@idiap.ch

IDIAP Research Report 08-08

Predicting the Dominant Clique in Meetings through Fusion of Nonverbal Cues

Dinesh Babu Jayagopi Hayley Hung Chuohao Yeo
Daniel Gatica-Perez

July 2008

Abstract. This paper addresses the problem of automatically predicting the dominant clique (i.e., the set of K-dominant people) in face-to-face small group meetings recorded by multiple audio and video sensors. For this goal, we present a framework that integrates automatically extracted nonverbal cues and dominance prediction models. Easily computable audio and visual activity cues are automatically extracted from cameras and microphones. Such nonverbal cues, correlated to human display and perception of dominance, are well documented in the social psychology literature. The effectiveness of the cues were systematically investigated as single cues as well as in unimodal and multimodal combinations using unsupervised and supervised learning approaches for dominant clique estimation. Our framework was evaluated on a five-hour public corpus of teamwork meetings with third-party manual annotation of perceived dominance. Our best approaches can exactly predict the dominant clique with 80.8% accuracy in four-person meetings in which multiple human annotators agree on their judgments of perceived dominance.

1 Introduction

Dominance is one of the most important dimensions in face-to-face social interactions. Dunbar et al. define dominance as a set of "expressive, relationally based communicative acts by which power is exerted and influence achieved" [2]. Establishment of a dominance hierarchy in a group happens within few minutes of interaction between unacquainted individuals, organizing the patterns and flow of interaction among group members. Recognizing dominance and related patterns automatically from audio-visual media have relevant applications as part of future tools for behavioral self-assessment in the workplace, for training of groups towards improvement of teamwork, but also for media content indexing (e.g. of TV group debates and interviews).

The human expression and perception of dominance have both been extensively studied in social psychology. The fundamental verbal and nonverbal cues related to dominance have also been established [2]. In particular, speaking activity and visual activity cues are both known to be significant nonverbal indicators of dominance. Initial attempts for automatic dominance analysis in group conversations have appeared only recently [5, 4, 3], often relying on (partly) hand-extracted cues [5], and targeting either the estimation of the most dominant person [3] or the individual classification of dominance levels [5]. Overall, many questions remain open in this research aspect of the emerging domain of computational analysis of social interaction.

This paper has two contributions. First, we address the novel problem of predicting the *dominant clique*, i.e., the set of dominant people, in face-to-face meetings using nonverbal cues. This problem has not been studied in previous work [5, 4, 3]. Small groups are often dynamic, and the presence of several actively involved people - competing for the floor or trying to guide the discussion - often leads to the display of dominant behavior by more than one person. Predicting the dominant clique is useful in scenarios where there is a power struggle or conflicting opinions, in which more than one dominant person is expected to be present. Also, in the context of a meeting retrieval system based on queries related to social behavior, automating the task of predicting the dominant clique could be a more robust preprocessing step than that of only predicting the most-dominant person [3] before human intervention. Picking the top few can be particularly relevant for larger groups.

The second contribution of our paper is a detailed study of nonverbal cues and dominance models for fully automatic prediction of the dominant clique. We study computationally efficient methods to derive a number of audio and visual activity cues (including activity duration, turns, and interruptions in both modalities), several of which have support from the social psychology literature. We investigate the performance of single-modality cues and fused cues, both single- and multimodal, using unsupervised and supervised learning methods, showing that multimodal fusion can be successful. We evaluate our models on 5 hours of meeting data manually annotated in terms of perceived dominance, which show the validity of our approach.

2 Data and Annotation

We used publicly available data recorded from the Augmented Multi-party Interaction (AMI) meeting corpus [1]. Each meeting has 4 participants. Each participant was assigned a distinct roles from 'Project Manager', 'User Interface specialist', 'Marketing Expert', and 'Industrial Designer'. The team was required to design a remote control. Audio data was recorded using 4 head-set microphones, 4 lapel microphones, and 2 circular microphone arrays. Video data was recorded using 4 close-view cameras, 2 side-view (Right and Left) cameras, and a center-view camera. A snapshot of all the the camera views is shown in Figure 1.

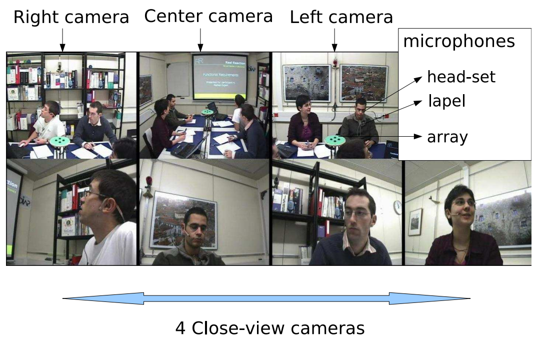

Figure 1: Example of the seven camera views available in the AMI meeting room.

From the AMI meeting data, 11 meetings were chosen and divided into 5 minute segments making a total of 59 meeting segments (simply called meetings from here on for convenience). Each of these meetings were annotated by three different annotators to enable a majority consensus. 21 participants, 14 men and 7 women, with varying cultural backgrounds were used to annotate the meetings. The annotators ranked the members in the meeting according to their perceived level of dominance. No prior definition of dominance was given to the annotators. After the annotations, we looked at inter-annotator agreement to select the data for the dominant clique task. For a clique of size equal to two, which partitions each meeting into two two-person subsets, we found that there was *full agreement* in 26 meetings (where all three annotators agree on the dominant clique), and that there was *majority agreement* in 57 meetings (where at least two of the three annotators agreed on the clique). These two data sets are used in experiments to investigate the effect of an increasing degree of variation in human perception of dominance on automatic prediction.

3 Our Framework

Our work consists of two functional blocks, namely nonverbal activity cue extraction, and dominant-clique prediction, described in the following subsections.

3.1 Nonverbal activity cues

Various vocalic and kinesic cues [2] have been found relevant to display and perception of dominance in the social psychology literature. Vocalic cues include speaking length [6], speech energy, speech tempo, pitch [2], and interruptions [7]. Kinesic cues include body movement, posture and gestures, facial expressions, and eye gaze [2]. Dominant people are usually more active than non-dominant people, moving more and with a wider range of motion. In our work, as an alternative to extracting some of the above cues, which can be computationally expensive, we focus on *speech and visual activity*-based cues that are easily extracted and cheap. We used the audio data from the headset microphones and the video data from the close-view cameras. The cues, which substantially extend the cues studied in our previous work [3], are described below.

Vocalic cues. Using the personal headset microphones, we first computed two time-series for each participant:

Speaking energy: We compute real-valued speaker energy for each participant using the root mean square amplitude of the audio signal over a sliding time window for each audio track. A window of 40 ms was used with a 10 ms time shift. This cue follows the findings in psychology that high speaker energy is a manifestation of dominant behavior [2].

Speaking status: A binary variable was computed by thresholding the speaker energy values. This indicates the speaking / silent status of each participant at each time step.

The two time series are then used to compute cues accumulated over the entire meeting:

Total Speaking Energy (TSE): Speaking energy accumulated over the duration of the meeting.

Total Speaking Length (TSL): Total time that a person speaks [6] according to their binary speaking status.

Total Speaker Turns (TST): A speaker turn is the time interval for which a person's speaking status is active.

Total Successful Interruptions by the speaker (TSI): This cue encodes the hypothesis that dominant people interrupt others more often [7]. The feature is defined by the cumulative number of times that a speaker starts talking and keeps the floor before the current speaker finishes his turn, i.e. only successful interruptions are counted.

Speaker Turn Duration Histogram (SDHist): We considered 11 bins, such that 10 bins were equally spaced at 1-second intervals, and the last bin included all turns in length greater than 10 seconds for every participant. The bins were chosen in this way to primarily distinguish short turns (some of which are likely to be backchannels) from monologues.

Total Speaker Turns without Short Utterances(TSTwoSU): The cumulative number of turns that a speaker takes where each speaker turn duration is longer than one second. The goal is to retain only those turns that are most likely to correspond to 'real' turns, eliminating all short utterances that are, in many cases, likely to be backchannels.

Kinesic cues. Following our recent work [3], we leverage the fact that meeting videos are already in compressed form to extract visual activity features efficiently. These features are generated from compressed-domain information such as motion vectors and block DCT coefficients that are accessible at almost zero cost from compressed video [8]. Each video is compressed by a MPEG-4 encoder with a group-of-picture (GOP) size of 250 frames and a GOP structure of I-P-P-..., where the first frame in the GOP is Intra-coded, and the rest of the frames are predicted frames.

If the participant is visible in the close-up view, we measure his visual activity by using either or both of the motion vector magnitude and the residual coding bitrate. The two methods differ in the information that they capture. While the motion vectors capture the rigid body motion like translation, bitrate often captures non-rigid motion. To meaningfully compare motion vector magnitudes and residual coding bitrate, we need to normalize the quantities [3]. If a participant is not detected in a frame of the close-up view, he is assumed to be presenting at the projection screen, and so is assumed to be visually active.

In an analogous fashion to the audio cues, we define a number of visual activity cues from the raw motion values:

Visual Activity: A binary variable computed that indicates whether a participant is visually active at each time step. Three variations were tested, based on Motion Vectors, Residual Coding Bitrate, and the combination (average) of both features.

Total Visual Activity Length (TVL), Total Visual Activity Turns (TVT), Total Visual Activity Interruptions (TVI), Visual Activity and *Turn Duration Histogram (VDHist)*: All these cues are the visual counterpart of the audio cues defined previously.

Both audio and visual cues are finally subsampled at a frame rate of 5 fps.

3.2 Dominant-clique prediction models

We investigate two dominance models, one unsupervised and another one supervised, to predict the dominant clique. Assuming that the clique size is equal to two (a reasonable assumption for the small groups depicted in the AMI data), the task is to predict, for each meeting, the top two most dominant individuals. Rather than being exhaustive about models, our goal is to analyze the behavior of the nonverbal cues described previously when used with different prediction strategies.

The unsupervised model accumulates single audio or video cues over the entire duration of the meeting and declares the people providing the largest two accumulated values as the dominant clique. The hypothesis is that dominant people speak, move, or grab the floor more than others [2, 6]. This model has the advantages of being extremely efficient and does not require training data. We evaluate the model by comparing the labels of the two most dominant people estimated automatically with those of the ground truth (as described in Section 4).

Our experiments with single features showed that some nonverbal cues were complementary in nature. We use support vector machine (SVM) classifiers to perform early fusion of both audio and visual cues. We carried out experiments to explore the best audio-only, video- only, and audio-visual combinations. For training, we use a two-class SVM to discriminate between the 'dominant' and 'non-dominant' clique classes in each meeting. For each test case, the SVM score produced for each person's features are ranked. The rankings are then used to determine the top two participants, by considering the two points which are furthest from the class boundary. In all experiments, the model was evaluated using a leave-one-out approach for each combination of input cues.

4 Results

4.1 Full-agreement data set

For this task, there were 26 meetings (44% of the total data). Firstly, we considered each modality separately for both the supervised and unsupervised approaches. We further used the supervised approach to do multi-cue fusion. We considered two evaluation criteria, where hard (EvH) or soft (EvS) scoring criteria were used. Let N denote the total number of meetings and let n be the number of times the automatically predicted clique is correct (i.e., the two people in the clique are correct), and m be the number of times when only one out of the two predicted people is correct. EvH computes the classification accuracy as n/N, and EvS computes accuracy as $(n + 0.5 * m)/N$. In other words, while EvH makes a decision on the full clique, EvS makes a decision based on individuals.

Audio cues. Table 1 shows the results obtained using audio cues. Using the unsupervised model with single features, the total speaking length (TSL) and total number of speaker turns with short turns removed (TSTwoSU) were most effective in classifying the dominant clique (best accuracies of 84.6% and 73.1% for EvS and EvH, respectively). Clearly, EvH imposes harder constraints on the evaluation. Social psychology literature does support the results that speaking time is a very strong cue for dominance perception [6]. The total speaking energy (TSE) also performed well. All cues performed considerably better than chance (which would result in a classification accuracy of 50% for EvS and 16.6%for EvH). Comparing to our recent work [3], predicting the dominant-clique is a more difficult task than predicting the top person only.

Features	Class. Acc.(%)	
	EvS	EvH
Random Guess	50.0	16.6
Unsupervised		
Total Speaking Length (TSL)	**84.6**	69.2
Total Speaking Energy (TSE)	82.7	69.2
Total Speaker Turns (TST)	80.8	61.5
Total Speaker Interruptions (TSI)	71.2	53.8
Total Speaker Turns without Short Utterances (TSTwoSU)	**84.6**	**73.1**
Supervised		
Speaker Turn Distribution Histogram (SDHist)	82.7	69.2
TSE, TST	**86.5**	**76.9**
TSL, TSE, TST	**86.5**	**76.9**

Table 1: Performance of **audio** cues for predicting the dominant clique with **full**-agreement data.

The results with the supervised model trained on multiple audio cues are also shown in Table 1 with a selection of the best performing feature combinations. We first observe that the Speaker Turn

Duration Histogram (SDHist) performed as well as the simple speaking length when using EvH. (as discussed before, a coarse proxy for backchannels). We observe that although TSE and TST are not the best single features, their combination yielded a 1.9% accuracy improvement for EvS and 7.7% accuracy improvement for EvH .

Visual cues. Table 2 shows the results obtained with visual cues. Though we experimented with the three options to compute visual activity described before (motion vectors, residual bitrate, and their combination), we only report the results using the combination (which was one of the two best performing options) for space reasons.

Regarding single cues in the unsupervised setting, the total visual activity length (TVL), which quantifies how much people move, and total motion turns (TVT), which quantifies how often people move (removing the very short turns that we assume to be noise) performed relatively poorly when using EvH but better using EvS. Compared to single audio cues, the best results with single visual cues degrade by 11.5% using EvS (from 84.6% to 73.1%).

Combining these cues using SVM early fusion improved the performance like in the audio case. Overall, the best achieved performance with visual cues and supervised learning is 12.3% worse than the corresponding best performance for audio cues (86.5% vs. 78.8%, using EvH, compare Tables 1 and 2).

Features	Class. Acc.(%)	
	EvS	EvH
Unsupervised		
Total Visual Activity Length (TVL)	**73.1**	46.2
Total Visual Activity Turns (TVT)	**73.1**	**50.0**
Total Visual Activity Interruptions (TVI)	65.4	34.6
Supervised		
VDHist	75.0	50.0
TVL, TVT	73.1	50.0
VDHist, TVL	**78.8**	**57.7**
VDHist, TVL, TVT	76.9	53.8

Table 2: Performance of **visual** cues for predicting the dominant clique with **full**-agreement data.

Audio-visual fusion. A selection of results obtained with audio-visual cues and the supervised approach are shown in Table 3. Again, for the visual activity features, we use the combination option. In this case, it is intersting to see that audio-visual fusion yields a slight improvement (3.9% for both EvS and EvH) in classification performance compared to the best audio cues, corresponding to the feature combination of (TSE, TST, VDHist), reaching 80.8% performance for the exact prediction of the clique, and 90.4% for the prediction of the clique individuals.

Features	Class. Acc.(%)	
	EvS	EvH
TSL, TVL	78.8	57.7
TSE, TVL	75	50.0
TST, TVT	80.8	61.5
TSL, TVL, TVT	80.8	61.5
SDHist, TVL	76.9	57.7
TSE, TST, TVL	84.6	69.2
TSE, TST, VDHist	**90.4**	**80.8**
TSE, TST, TVT	86.5	73.1

Table 3: Performance of **audio-visual** cues for predicting the dominant clique, **full**-agreement data.

A closer look at the data set with full agreement among the annotators w.r.t. the dominant clique, shows that 73% of the meetings had full agreement on the specific ordering as well, i.e. on the most dominant person as well as the second most dominant person. In this subset, we found that almost all of the cues predicted the most dominant person as one of the members of the clique. The errors, if any, were due to the second most dominant person. This highlights, on one hand, the relevance of the studied cues for the task considered, and on the other, the higher difficulties in correctly predicting multiple dominant people compared to only one, as done in most existing work. Related demos are available at `http://www.idiap.ch/~djaya/DominantClique/`.

4.2 Majority-agreement data set

For this task we assume the ground truth is the dominant clique on which the majority of annotators (i.e., at least two out of the three annotators for each meeting) agree. As mentioned previously, there were 57 meetings in this data set (96% of all data). This data inherently has higher variability with respect to human perception of dominance, and this fact, added to the combinatorial nature of the dominant-clique task, produces a much more difficult task. In this data set, there are many cases in which 3 of the meeting participants belong to the dominant clique, depending on what specific annotator's judgment is taken into consideration.

The evaluation of this task is therefore aimed at analyzing the performance of cues and models in much more challenging conditions. A selection of results are shown in Table 4. Overall, there was a clear drop in performance for both evaluation criteria EvS and EvH. Furthermore, unlike the full agreement case, the feature combinations do not improve performance.

Features	Class. Acc.(%)	
	EvS	EvH
Unsupervised		
TSL	**71.1**	**45.6**
TSTwoSU	69.3	40.4
TVL	62.3	29.8
TVT	63.2	33.3
Supervised		
TSL, TSE, TST	**70.2**	**43.9**
SDHist, TSE, TST, TSI	69.3	42.1
VDHist, TVL	66.7	35.1
TSE, TST, VDHist	**71.1**	**43.9**

Table 4: Performance of **audio-visual** cues, **majority**-agreement data.

5 Conclusion

In this paper we addressed a new task, unlike the most and least dominant person task, namely the automatic prediction of the dominant clique in group meetings. Our study has investigated how the task of predicting the dominant clique can be affected by different audio and video cues, annotator variability and the estimation method involved. We employed automatic, easily computable nonverbal activity cues for doing the prediction, and studied both unsupervised and supervised approaches. Though the audio modality performed better than the video modality, the visual activity cues were in some cases relatively effective. The total speaking length performed well as a single feature, as observed in the social psychology literature. Overall, the audio-visual cue combination of total speaking energy, total speaker turns, and visual activity turn duration histogram performed the best, for the full-agreement case. For the majority agreement case, the relative rankings of the best performing cues were almost similar and we observed a consistent decrease in prediction performance. This shows that, while the investigated cues were relatively consistent for the two subtasks considered, the estimation of the dominant clique is a challenging problem.

6 Acknowledgments

This research was partly supported by the US VACE program, the EU project AMIDA, the Swiss NCCR IM2, and by Singapore (A*STAR).

References

[1] J. Carletta et al. "The AMI meeting corpus: A pre-announcement," *Proc. MLMI Workshop, Edinburgh,UK,Jul. 2005*

[2] N.E. Dunbar et al. "Perceptions of power and interactional dominance in interpersonal relationships," *Journal of Social and Personal Relationships, 22(2):207-233,2005.*

[3] H. Hung et al. "Using Audio and Video Features to Classify the Most Dominant Person in a Group Meeting," *Proc. ACM MM, Augsburg, Sep.2007.*

[4] K. Otsuka et al. "Quantifying interpersonal influence in face-to-face conversations based on visual attention patterns," *Conference on Human Factors in Computing Systems,pages 1175-1180, 2006.*

[5] R.J. Rienks and D. Heylen. "Automatic dominance detection in meetings using easily detectable features," *Proc. MLMI Workshop, Edinburgh,UK,Jul. 2005*

[6] M. Schmid Mast. "Dominance as expressed and inferred through speaking time: A meta-analysis," *Human Communication Research, 28(3):420-450, Jul.2002.*

[7] L. Smith-Lovin and C. Brody. "Interruptions in Group Discussions: The Effects of Gender and Group Composition," *American Sociological Review. 54(3):424-435,Jun.1989.*

[8] H. Wang et al. "Survey of compressed-domain features used in audio-visual indexing and analysis," *Journal of Visual Comm. and Image Representation, 14(2):150-183,2003.*