Investigating Automatic Dominance Estimation in Groups From Visual Attention and Speaking Activity

Hayley Hung[1] Dinesh Babu Jayagopi[1,2] Sileye Ba[1]
Jean-Marc Odobez[1,2] Daniel Gatica-Perez[1,2]
(hhung,djaya,sba,odobez,gatica)@idiap.ch
[1] Idiap Research Institute, Martigny, Switzerland
[2] Ecole Polytechnique Federale de Lausanne (EPFL), Switzerland

ABSTRACT

We study the automation of the visual dominance ratio (VDR); a classic measure of displayed dominance in social psychology literature, which combines both gaze and speaking activity cues. The VDR is modified to estimate dominance in multi-party group discussions where natural verbal exchanges occur and other visual targets such as a table and slide screen are present. Our findings suggest that fully automated versions of these measures can estimate effectively the most dominant person in a meeting and can approximate the dominance estimation performance when manual labels of visual attention are used.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing indexing methods

General Terms

Algorithms, Human Factors

Keywords

meetings, dominance modeling, audio-visual feature extraction, visual focus of attention estimation

1. INTRODUCTION

Dominant interactive behaviour in groups occurs naturally whether the interactants are humans or animals. In particular, the encoding, or the act of communicating a nonverbal signal must comply with a code of conduct which is known by both the encoder and decoder of the interaction [10]. In animal societies, subordinate individuals tend to direct more visual attention towards the more dominant members of the group, when they themselves are not being observed [5]. On the other hand, dominant members monitor more freely, and evenly amongst all their subordinates [5]. Similarly, investigations into interpersonal dominance between humans have found that people with high status tend to receive more visual attention from others [5]. Understanding these constructs in automatic systems is desirable for tasks such as for improving task-oriented group effectiveness [13] or remote meeting scenarios [14].

We address dominance in this paper, which has been differentiated from influence or power in social psychology by Dunbar and Burgoon, who defined dominant behaviour as "expressive, relationally based communicative acts by which power is exerted and influence achieved" (p. 208) [6]. In 1982, Dovidio and Ellyson introduced the visual dominance ratio (VDR) for dyads, which is now considered in social psychology to be a classic measure of dominance [5]. The measure was motivated by two findings: the proportion of time someone spends looking at the other while speaking was proportional to levels of power or dominance, and the proportion of time spent looking at the other while listening was inversely related. The VDR was then defined as the ratio of these two measures, which was found to decode dominant behaviour reliably.

We investigate whether these concepts can be applied to an automated framework. Using both automatically generated audio cues from individual head-set sources and visual focus from video, we extend the idea of the VDR from dyadic to multi-party conversations. In social psychology [5] and ubiquitous computing [13, 14], there have been studies linked to observing gaze in dyads, triads or multi-party conversations for decoding nonverbal behaviour. In contrast, we investigate multi-party interactions where other visual targets such as a table, slide screen or whiteboard are present. Also, meetings can last between 15-35 minutes, which can lead to less intense interactions, compared to shorter conversational scenarios. It is also important to note that in addition to other visual targets, the distribution of focus targets can overlap significantly and the angular range of visual focuses for each participant varies according to their location in the room.

In terms of automated methods of estimating dominance, some methods have been proposed [7, 12]. The work of Otsuka et al. [12] is the closest to ours as it used audio-visual cues to estimate the visual focus of attention and interpersonal influence in four-person conversations. This work differs from others in method by which the gaze is estimated since high resolution images are not available to track eye-gaze robustly. Instead, the method uses estimates of a person's head pose to estimate visual focus using contextual cues about the conversational patterns during discussions. While some qualitative observations were made, the performance of their dominance measures were not evaluated systematically, and there were no extraneous visual targets.

ICMI'08, October 20–22, 2008, Chania, Crete, Greece.

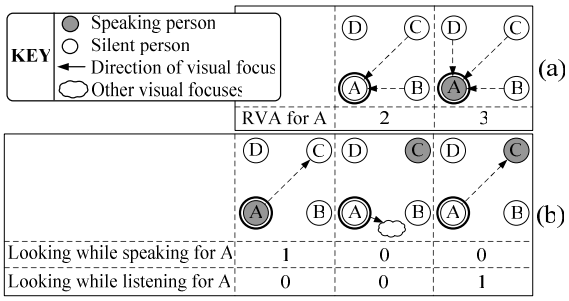

Figure 1: Example scorings of RVA and $MVDR$ for person A (highlighted node), at time t: (a) two examples of RVA; (b) three example scenarios for looking-while-speaking and looking-while-listening.

Here, we investigate a model for visually dominant behaviour, grounded in findings from social psychology, using fully automatic audio and visual cues to assess whether the most dominant person can be reliably detected when there are varying degrees of agreement in human perceptions of dominance. We apply our experiments on two sets of data that contain 170 and 285 minutes of audio-visual data.

2. ESTIMATING VISUAL DOMINANCE

Dovidio and Ellyson [5] suggested that someone who receives more visual attention is perceived to be more dominant. The total received visual attention (RVA) for each participant i and their corresponding Visual Focus of Attention (VFOA), $f_t = (f_t^1, .., f_t^{M_p})$ at time t is defined as

$$RVA^i = \sum_{t=1}^{T} \sum_{j=1, j\neq i}^{M_p} \delta(f_t^j - i),\ i = 1, \ldots, M_p \ , \quad (1)$$

where T is the number of frames, $f_t^j \in \{1, \ldots, M\}$ where M is the number of focus targets, M_p is the number of participants ($M > M_p$), and $\delta(.)$ is the delta function such that $\delta(f_t^j - i) = 1$ when $f_t^j = i$. In our data, the focus targets were defined as the three other participants, the slide screen, the whiteboard, and the table. The table label was assigned whenever a person looked at the table or an object on it. For all gaze directed at other locations, an 'unfocused' label was also defined. Fig. 1 (a) shows two examples of different scenarios for participant A. They show that the VFOA of each participant on A is counted, regardless of whether A is speaking or not. We also encoded the ability of each person to 'grab' visual attention by considering the RVA feature in terms of events rather than frames.

Dovidio and Ellyson [5] defined the VDR between dyads as the proportion of time a person looks at the other while speaking divided by the time a person looks at the other while listening. It encodes the displayed dominance through either active or passive participation. We extend the VDR to a multi-party scenario ($MVDR$). The 'looking-while-speaking' feature is redefined as when a person who is speaking looks at any participant rather than at other objects in the meeting. Similarly, the 'looking-while-listening' case involves actively looking at any speaking participant while listening as shown in Fig. 1(b). Clearly other definitions of the MVDR are also possible. The $MVDR$ for person i is:

$$MVDR^i = \frac{MVDR_N^i}{MVDR_D^i} \ , \quad (2)$$

where the time that each participant spends looking at others while speaking is defined as

$$MVDR_N^i = \sum_{t=1}^{T} s_t^i \sum_{j=1, j\neq i}^{M_p} \delta(f_t^i - j) \ , i = 1, \ldots, M_p \ , \quad (3)$$

s_t^i is a binary vector containing the speaking status of each participant (speaking: 1, silence: 0). The time spent looking at a speaker while listening (i.e. not speaking) is defined as

$$MVDR_D^i = \sum_{t=1}^{T} (1 - s_t^i) \sum_{j=1, j\neq i}^{M_p} \delta(f_t^i - j) s_t^j . \quad (4)$$

3. CUE EXTRACTION

3.1 Speaking Activity

Speaking activity features are estimated from close-talk microphones attached to each meeting participant. At each time step, the audio energy of each participant, measured over a sliding window, is thresholded to define the participant's speaking status.

3.2 Visual Attention

We extend our recent work [2] to estimate the joint focus state of all participants. We rely on several features. The main one is the head pose of each participant. In the absence of direct eye observations due to low image resolution, head orientation can be used as a proxy for gaze estimation. However, as the same head pose can be used to gaze at different targets, the VFOA estimation was improved by modeling the relationship between people's VFOA their conversational events, and other contextual cues related to the group activity. This is reflected in the dynamical graphical model in Fig. 2, whose joint distribution is proportional to:

$$\prod_t p(f_t|f_{t-1}) p(f_t|e_t, a_t) p(e_t|a_t) p(\tilde{s}_t|e_t) p(o_t|f_t) \ , \quad (5)$$

we assumed that $p(f_t|f_{t-1}, e_t, a_t) \propto p(f_t|f_{t-1}) p(f_t|e_t, a_t)$; e_t denotes a conversational event; a_t, the time since the last slide change; $\tilde{s}_t$, the proportion of speaking time of all participants in a window; and o_t is the head pose of all people. These variables are described in more detail below.

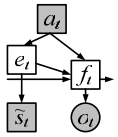

Figure 2: VFOA graphical model at time t. a_t: time since last slide transition, e_t: conversational events in a window, f_t: focus targets, $\tilde{s}_t$: proportion of speaking time in a window, o_t: head pose.

Conversational events e_t: a conversational event e_t at time t, denotes the hidden conversation type (silence/monologue /dialog/discussion) occurring over a time window, that affects the dynamics of the gaze and speech patterns. The estimation of this variable is mainly driven by the term $p(\tilde{s}_t|e_t)$, which models the likelihood of the speaking features of all participants, $\tilde{s}_t = (\tilde{s}_t^1, \ldots, \tilde{s}_t^{M_p})$, defined as the proportion of time they speak during the window. Since people tend to look at the current speaker [12], this contextual cue will also influence the estimation of people's VFOA through the term $p(f_t|e_t, a_t)$, which increases the prior probability of the VFOA targets corresponding to the people who are actively involved in the current conversational event.

Slide-screen activity a_t: People do not always gaze at the current speaker. During long monologues, they may listen to the speaker while looking elsewhere (usually at the table). Also, during slide presentations, people often look at the slide rather than at the speaker, especially right after

(a) (b) (c)

Figure 3: Evaluation setup: (a) Full view of meeting room; (b) example of input video image; (c) plan of the meeting room.

a slide change. In our model, the impact of this activity is taken into account by $p(f_t|e_t, a_t)$, by implicitly modulating the prior on looking at the current speaker(s) as a function of the duration a_t since the last slide change occurrence. Slide changes are estimated by thresholding visual activity features from the camera view in Fig. 3 (a). The prior on the conversational event $p(e_t|a_t)$ also depends on a_t.

Temporal modeling: This is introduced at the VFOA sequence level (term $p(f_t|f_{t-1})$), by enforcing temporal continuity of visual attention over time.

Head pose-VFOA relationship: The head pose for all participants $o_t = (o_t^1, \ldots, o_t^{M_p})$ is represented by the pan and tilt angles of each participant, which was estimated using the method described in [1], which jointly tracks, the head location and its pose using a Bayesian formulation solved through sampling techniques. To define the observation model $p(o_t|f_t)$, we assumed that people's head pose only depends on their given focus, i.e. $p(o_t|f_t) = \prod_{i=1}^{M_p} p(o_t^i|f_t^i)$. As proposed in [11], rather than relying on the common assumption that a person's head is oriented to gaze face-on to a visual target, we rely on findings in psychovision that model the contribution of the head orientation to actual eye-in-head gaze shifts, leading to a model of the relationship between a given VFOA target and the head pose.

Inference: The VFOA and conversational events are estimated using an iterative process. This alternatively maximizes the overall likelihood with respect to the VFOA (given the observation and the current conversational event estimates) and the conversational events (given observations and current VFOAs).

The VFOA recognition performance was 50% (all targets) and 40% (people only) where a random performance would be the reciprocal of the number of targets (14%). Details are omitted for space reasons but can be found in [3].

4. DATA AND ANNOTATIONS

We used meeting data from the publicly available AMI meeting corpus [4] to conduct our experiments. Audio-visual recordings were taken in a meeting room, with multiple cameras and microphones as shown in Figure 3. The room contains a table, slide screen and white board. For our experiments, we used data from headset microphones. There were two mounted cameras as shown in Fig. 3(c).

The dominance tasks and annotations were taken from [8], consisting of 11 meeting sessions selected from the AMI corpus, that lasted between 15 and 35 minutes. Each session was divided into 59 5-minute non-overlapping meeting segments. For each segment, 3 annotators ranked the participants from 1 (most) to 4 (least) indicating their perceived dominance. No initial definition of dominance was provided.

We used 2 data sets for our experiments. The first consists of 34 meetings where the annotators all agreed on the most dominant person. The second, included annotator variability where 57 meetings had majority agreement among the annotators. The self-reported confidence for the human judgments had an average score of 1.7 for the full agreement case and 1.9 for the majority case (where 1 represents the highest confidence and 7 represents the lowest). More details on the dominance annotations can be found in [8].

The VFOA of each participant was labeled manually by 6 annotators who selected the gaze target of each participant from a list of items (three other participants, table, white-board, slide screen, unfocused). When a target person was standing in front of the slide screen or the white board, the annotators were asked to label the person as the focus of attention. More information about the annotation procedure for the VFOA targets can be found in [9].

5. RESULTS

Using the RVA and $MDVR$ measures defined in Eq.s 1 and 2 respectively, we estimated the most dominant person in each meeting for the two dominance data sets. We evaluated the performance of the RVA case for both frame and event-based cases. Also, to study the contribution of each element of the $MVDR$, we analyzed both the performance of the numerator and denominator separately as well as when combined. In each case except for the denominator, $MVDR_D$, defined in Eq.4, the person with the highest value was estimated to be the most dominant. For the case of the $MDVR_D$, the person with the smallest value was estimated to be the most dominant. The results for manual and automatically extracted cues are shown in Table 1.

	Meeting Classification Accuracy(%)			
	MostDom(Full)		**MostDom(Maj)**	
Method	**Manual**	**Auto**	**Manual**	**Auto**
RVA **(Time)**	58.8	67.6	52.6	61.4
RVA **(Events)**	70.6	38.2	61.4	42
$MVDR$	73.5	79.4	64.9	71.9
$MVDR_N$	79.4	70.6	70.1	63.2
$MVDR_D$	41.2	50	40.4	45.6
SL	85.3	85.3	77.2	77.2
Random	25			

Table 1: Percentage of correctly labeled meetings in the full and majority cases using manual and automatically estimated cues. SL: Speaking length.

Firstly, we considered the performance on the 34 meetings with full annotator agreement. We studied firstly the ideal case, where human annotations of speaking activity and VFOA were used. Using RVA events appeared to improve the performance compared to time (from 58.8% to 70.6%). Interestingly, this feature was quite discriminative, using just visual cues. The introduction of speaking activity features with the $MVDR$ appeared to improve the performance. Also, the $MVDR$ did not seem to perform as well as just the using the $MVDR_N$, which performed the best at 79.4%. $MVDR_D$ had the worst performance of 41.2%.

Using the automated estimates, the best performing feature was the $MVDR$ (79.4%). The RVA (Events) feature seemed to perform better in the manual rather than automatic case. This was probably because the estimates were smoothing out shorter events. The $MVDR_N$ feature seemed to perform worse compared to its manual counterpart. In contrast, the automatic case suggested a much better estimate using the $MVDR_D$ feature. Comparing these findings with the results of using the highest speaking length

(85.3%) to estimate the most dominant person, the automated VFOA estimation may have estimated the dominance person better than the manual case because the estimated VFOA dynamics are better correlated to the speaking activity compared to those of the ground truth. This could be due to the conditioning of the VFOA estimates on the conversational events. This is further suggested by the significant improvement in the performance of the VFOA estimation when speech priors were used in [2]. In terms of the decrease and increase in performance between $MVDR_N$ and $MVDR_D$, respectively, when we compare the manual to automated versions, we observe that while the VFOA estimates of a speaker may not be affected by their own speaking activity, those of a listener are clearly conditioned on the conversational events.

Finally, analyzing the results using the manual and automated dominance estimation results in Table 1 for the majority agreement data-set, there was a consistent drop in performance while the relative differences between feature types and also manual and automatic labels were similar.

Given the small number of data samples, most of the observed differences in performance are not significant at the 10% level.

The discriminative nature of each automated measure using the full-agreement data set was studied in more detail by plotting the distributions for each measure, for the most and non-most-dominant people (see Fig. 4). Histograms were accumulated for each measure (using the manually labeled visual focuses), which were normalized over the sum for all participants for each meeting. The highest separation between most dominant and non-most-dominant distributions is observed in the $MVDR$ case (b), while for (c,d), discriminating the distributions becomes progressively worse.

We then observed the measures $MVDR_N$ and $MVDR_D$, as a proportion of the total meeting time. Dovidio and Ellyson [5] found that in their dyadic interaction conditions, the most dominant person tended to have a numerator and denominator for the VDR of 55% and 40% respectively while for our meetings, the numbers were noticeably lower at 20% and 13% on average, though the maximum values went up to 58% and 45%. For the non-most-dominant, Dovidio and Ellyson found that the amount of look-speak to look-listen were 40%-60% and 25%-75% respectively while for the automatic case, the values were 0%-44% and 0%-57% respectively. This is not surprising as there are more interlocutors, so all the speakers talk for a smaller proportion of the total time on average. They could also look at other objects in the room such as the table or slide screen. The most dominant and non-most dominant person received respectively, the visual focus on average 15% and 8% of the total time.

6. CONCLUSION

Our work shows that extending Dovidio and Ellyson's measures of dominance to the group case was indeed effective. Our study also suggests that while audio cues are very strong, visual attention is also quite discriminant and could be useful in the absence of audio sensors. However, we have yet to discover other features that are jointly complementary. A more in-depth study of modifications to the VDR to the multiparty case is reserved for future work.

Acknowledgments

This research was partly funded by the U.S. VACE program, the EU project AMIDA and the Swiss NCCR IM2.

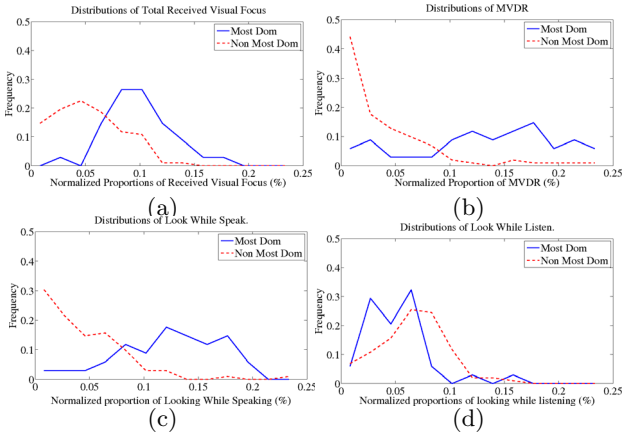

Figure 4: Distributions of each accumulated feature for the most and non-dominant participants. (a) RVA; (b) $MVDR$; (c) $MVDR_N$; (d) $MVDR_D$.

7. REFERENCES

[1] S. O. Ba and J.-M. Odobez. A Rao-Blackwellized mixed state particle filter for head pose tracking. In *Proc. ACM-ICMI-MMMP*, pages 9–16, 2005.

[2] S. O. Ba and J.-M. Odobez. Multi-party focus of attention recognition in meetings from head pose and multimodal contextual cues. In *Proc. ICASSP*, 2008.

[3] S. O. Ba and J.-M. Odobez. Multi-Person Visual Focus of Attention from Head Pose and Meeting Contextual Cues *Idiap Research Report, IDIAP-RR-08-47*

[4] J. Carletta et al. The AMI meeting corpus: A pre-announcement. In *Proc. MLMI*, 2005.

[5] J. F. Dovidio and S. 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, June 1982.

[6] N. E. Dunbar and J. K. Burgoon. Perceptions of power and interactional dominance in interpersonal relationships. *Journal of Social and Personal Relationships*, 22(2):207–233, 2005.

[7] H. Hung et al. Using audio and video features to classify the most dominant person in a group meeting. In *Proc. ACM Multimedia*, 2007.

[8] D. Jayagopi et al. Modeling dominance in group conversations using nonverbal activity cues. *IEEE Transactions on Audio, Speech and Language Processing, accepted for publication.*

[9] N. Jovanovic. To Whom It May Concern - Addressee Identification in Face-to-Face Meetings, *PhD thesis*, University of Twente, 2007.

[10] S. R. Langton, et al. Do the eyes have it? cues to the direction of social attention. *Trends in Cognitive Sciences*, 4(2):50–59, February 2000.

[11] J.-M. Odobez and S. Ba. A cognitive and unsupervised MAP adaptation approach to the recognition of focus of attention from head pose. In *Proc. ICME*, 2007.

[12] K. Otsuka, et al. Quantifying interpersonal influence in face-to-face conversations based on visual attention patterns. In *Proc. ACM CHI Extended Abstracts*, 2006.

[13] J. Sturm et al. Influencing social dynamics in meetings through a peripheral display. In *Proc. ICMI*, 2007.

[14] R. Vertegaal and Y. Ding. Explaining effects of eye gaze on mediated group conversations: amount or synchronization? In *Proc. ACM CSCW*, 2002.