Putting the Pieces Together: Multimodal Analysis of Social Attention in Meetings

Ramanathan Subramanian[1], Jacopo Staiano[1], Korina Kalimeri[1,2], Nicu Sebe[1],
Fabio Pianesi[2]
[1]University of Trento, Italy
[2]Foundazione Bruno Kessler, Trento, Italy
subramanian,staiano,sebe@disi.unitn.it, kalimeri,pianesi@fbk.eu

ABSTRACT

This paper presents a multimodal framework employing eye-gaze, head-pose and speech cues to explain observed social attention patterns in meeting scenes. We first investigate a few hypotheses concerning social attention and characterize meetings and individuals based on ground-truth data. This is followed by replication of ground-truth results through automated estimation of eye-gaze, head-pose and speech activity for each participant. Experimental results show that combining eye-gaze and head-pose estimates decreases error in social attention estimation by over 26%.

Categories and Subject Descriptors

H.1.2 [**User/Machine Systems**]: Human information processing; I.5.4 [**Pattern Recognition Applications**]: Computer vision

General Terms

Algorithms, Measurement, Human Factors

Keywords

social attention, eye-gaze, head-pose, meeting analysis

1. INTRODUCTION

Determining the direction of another person's attention is an important ability for humans. It not only provides salient information about the location of objects (food, predators), but also plays a fundamental role in many complex forms of social cognition such as visual perspective-taking, deception, empathy and the theory of mind [17], expression of intimacy and exercising of social control [5].

Gaze direction is an important cue for social attention and humans have evolved specialized neural mechanisms devoted to gaze processing [7]. However, it has been convincingly shown that there is more than just eye-gaze to visual attention; head and body orientation also significantly contribute towards deciphering another person's direction of attention [7]. While Perret *et al.* [11] developed an attentional model that integrates eye gaze, head and body directions in a hierarchical fashion, recent work [6] suggests these orientation cues are processed independently and combined so that one modulates the decision process concerning the others.

This paper investigates computational models of social attention by considering gaze direction, head orientation and speaking activity. We consider a number of hypotheses concerning social attention in meetings by analyzing results from ground-truth as well as automated analysis of four meeting videos from the 'Mission Survival II' corpus [9]. The following hypotheses are based on observations and presumptions stated in previous literature, but which have never been analyzed in great detail:

- H1: Attention is mostly given to the person sitting right in front of the observer. This hypothesis derives from an observation made in [14].
- H2: There exists a direct relationship between the verbal behavior of a person and the amount of attention he receives. This hypothesis derives from [10].
- H3: Use of eye-gaze in conjunction with head-pose improves accuracy of automated social attention estimation. This hypothesis directly derives from the above discussion.

We also attempt to characterize meetings and individuals based on the analysis of ground truth data. Finally, we describe automated methods to compute eye-gaze-cum-head-pose-based social attention, and replicate ground truth results by combining computed social attention estimates with speech data. To summarize, this is one of the first works to

1. Comprehensively analyze meeting videos by combining eye-gaze, head-pose and speech information. Past works have essentially focused on perfecting automated methods for computing visual social attention.
2. Automatically employ eye-gaze as a modality for estimating social attention using [15]. Previous works assume head-pose as the main indicator of social attention, mostly due to the difficulty in reliably computing eye-gaze. Experimental results show a significant increase in attention estimation accuracy when gaze cues are employed in conjunction with head-pose cues.

2. RELATED WORK

Social attention has been extensively investigated under the rubric of focus of attention (FOA) in meetings [2]. Pioneering work is described in [13], where subjects' FOA is

MM'10, October 25–29, 2010, Firenze, Italy.
Copyright 2010 ACM 978-1-60558-933-6/10/10 ...$10.00.

computed by combining head-pose information with *aprori* knowledge about the number of participants and their relative positions. Assuming the head-pose to be the main indicator of a person's direction of attention, the algorithm employs a Hidden Markov model (HMM) to map FOA estimates to real-world targets. This framework is extended to employ acoustic as well as visual cues in [14].

Prediction of focus of visual attention in dynamic meeting scenes is discussed in [16]. Shifts of FOA in spontaneous situations are studied for 10 videos with 35 possible attention targets. The most probable target is identified by mapping head-pose to its most likely gaze angle counterpart, to achieve 57% correct recognition of the visual target. Another approach to recognizing social attention in meetings from head-pose modeled using a Gaussian mixture model (GMM) as well as HMM, is discussed in [1]. FOA targets are not restricted to participants alone, but to environmental targets (*e.g.*, projector) as well, and results of saccadic eye motion modeling are exploited to model head-pose given the upper-body pose and effective gaze target.

Recent research has focused on automatic analysis of social aspects such as meeting roles, with specific emphasis on *dominance*, which characterizes a person's status within a group and the power he/she has within it. A study on the usefulness of non-verbal audio-visual cues when employed individually or in combination, for automated dominance estimation is described in [3]. Another work that discusses dominance estimation from meetings is [4], where the visual dominance ratio (VDR) measure is employed for automated dominance computation.

Brief analysis of related literature shows that (1) While past works have investigated and considerably improved on the usage of features such as head-pose, an explicit analysis of social attention and multimodal cues to define **meeting characteristics** is missing. (2) Most works consider head pose as the main indicator of social attention, neglecting eye gaze primarily because of the difficulty in computing it. The next section describes the meeting videos used for analysis and the derivation of meeting characteristics upon analysis of the ground truth data.

3. GROUND-TRUTH ANALYSIS

3.1 'Mission Survival' meeting videos

We used data from the 'Mission Survival' corpus [9], a multimodal annotated collection of video and audio recordings in a lab setting. Each meeting consists of four participants seated around a table and engaged in the 'Mission survival' task, which is used in experimental and social psychology to elicit decision-making processes in small groups. The objective of the 'Mission Survival' task is to reach a consensus on how to survive a disaster scenario, *e.g.,* a plane crash in an uninhabited island. The group has to rank a number of (up to 15) items critical for survival, according to the participants. The consensus meeting scenario was chosen for the purpose of meeting dynamics analysis, which involves intensive engagement of the participants in order to reach an agreement, thus offering the possibility to observe a large set of social attitudes. All meetings are of 20-30 minutes duration, and recorded with four web cameras installed on the meeting table, while speech activity is recorded using close-talk microphones. Fig.1 shows an exemplar meeting scenario from the 'Mission Survival' data. Assuming that each participant directs his/her social attention targets included only the remaining three subjects, annotations were performed for the head-pose, eye-gaze and speech data to obtain the ground truth. Since the nature of the task involved choosing from a list of items, a 'self-attention' label, which denotes the state where a participant looks at the list provided to him/her, was also included in the annotation.

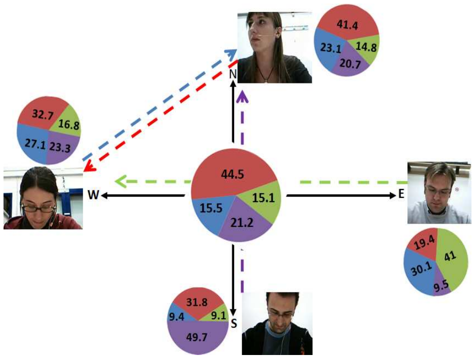


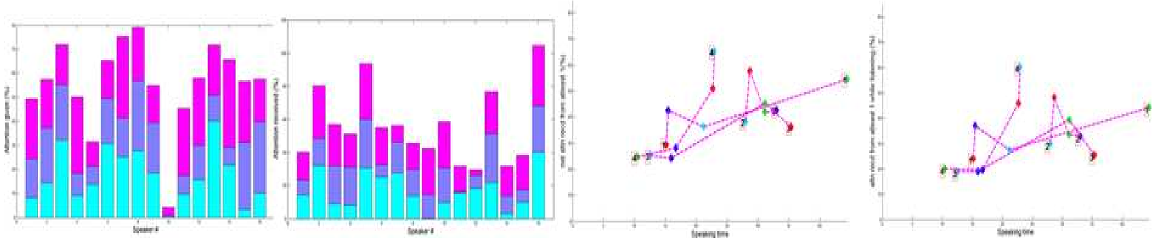
Figure 1: An exemplar meeting scene from the 'Mission Survival' dataset [9]. Color codes denoting subject locations are red (North), blue (West), violet (South) and green (East). The central pie-chart represents the distribution of speaking time, while pie-charts beside each participant denote the distribution of *attention given* by that subject to peers, including self-attention. Arrows denote direction of maximum *attention given* (excluding self-attention).

3.2 Inferences from ground-truth

Since eye-gaze is the most reliable social attention cue, we analyze eye-gaze and speech ground-truth data to derive inferences in this section. Fig.1 presents the distribution of social attention and speech activity for a meeting. Let A_i^j denote the *attention given* by subject i to j. Conversely, A_j^i denotes *attention received* by subject i from j. A_i^j and A_j^i may be expressed in minutes or as percentages. Henceforth, $i, j \in L, O, R$, where L, O and R denote the person located at the *left*, *opposite* and *right* respectively, with respect to the reference. Also, let $A^i = \sum_{\forall j, j \neq i} A_j^i$. denote the *overall attention received* by subject i. Likewise, $A_i = \sum_{\forall j, j \neq i} A_i^j$ denotes *overall attention given* by subject i to his peers.

3.2.1 Validation of H1

Fig.2(a,b) present the distribution of A_i^j and A_j^i to subjects seated to the left, right and directly opposite, respectively. Evidently, the distribution is not biased as observed in [14], where the authors note that the person in front gets almost twice as much attention as the persons on either side. Across all four meetings, we find that the proportions of A_i^L, A_i^R and A_i^O are 17.6%, 16.3% and 21.9% respectively, implying that the likelihood of a subject giving/receiving attention to/from each of the other group members is roughly equal. Therefore, on the basis of the observations made from ground-truth data, we reject ***hypothesis H1, i.e., the person located directly opposite (to the reference subject) does not receive/give significantly more attention***.



(a) (b) (c) (d)

Figure 2: (a,b) present bar graphs denoting distribution of A_i^j and A_j^i respectively with the bottom (sky-blue), middle (sea-blue) and top (pink-purple) shades respectively denoting $j = L,\ R,\ O$ for 16 participants. (c,d) denote plots of *Overall attention received* (A^i) vs *speaking time* (ST_i) and *Attention received while listening* ($A^i_{(l)}$) vs ST_i for the four meetings. The speakers at the East, North, West and South are denoted by points marked in red, green, blue and cyan respectively. All measures are expressed as percentages.

Table 1: Social attention from (a) ground-truth and (b) automated analysis with head-pose only (HP) and eye-gaze + head-pose ($HP + EG$) information.

			(a)						(b)		
Meeting #	Subject #	ST_i (%)	A^i (%)	$A^i_{(l)}$ (%)	A_i (%)	AQ_i	ST_i (%)	A^i (HP)	A^i (HP+EG)	A_i (HP)	A_i (HP+EG)
1	1	15.2	29.5	24.1	49.3	0.6	15.7	33.6	33.4	49.3	49.9
	2	44.5	54.6	44.3	57.4	0.95	43	46.6	54.4	56.8	57.5
	3	15.4	42.6	37	71.9	0.59	18.2	29.8	33.5	71.1	71.9
	4	21.2	36.3	27.6	50.2	0.72	20.5	37	40.4	49.9	50.2
2	1	28.7	57.7	48.2	31.5	1.83	29.3	43.9	46.1	57	55.2
	2	31.1	42	33.6	65.2	0.64	32.5	35.9	39.3	69.7	63.7
	3	33	42.7	32.9	75.3	0.57	33.2	43.9	45.6	90.8	82.3
	4	28.2	38.2	29.9	78.9	0.48	28	24.4	30.4	63.6	83.9
3	1	35.4	36.1	25.4	55	0.66	32.8	38.6	40.8	47.5	56.8
	2	31.2	45.4	39.4	4.3	10.7	30.6	39.4	42.8	8.4	7.1
	3	15.9	24.3	19.2	45.5	0.53	14.7	21.3	21.8	38.4	45.9
	4	12.5	25.3	19	58	0.44	13.6	27.6	28	66.8	64.2
4	1	22.7	51	45.9	71.8	0.71	21.3	43.6	45.6	71.9	72.4
	2	10.4	24.7	20.1	65.8	0.38	13.8	31.9	28	66.1	66.5
	3	16.7	28.1	19.6	56.6	0.5	18.1	26	25.5	56.7	57.1
	4	23	65.3	60.4	57.5	1.14	22.2	55.8	58	58.4	58.8

3.2.2 Validating H2

The striking resemblance between the plots for the overall *attention received* (A^i) and *attention received* while *listening* ($A^i_{(l)}$) with respect to the speaking time (ST_i), can be seen from Fig.2(c,d) (values in Table 1(a)). For ease of comparison, both plots have been obtained with identical (x, y) scale. An examination of the plots reveal that the *attention received*, in general, increases with *speaking time*, and that *the speaking activity of a subject influences the amount of attention received by a subject,* ***even when the subject is not speaking***.

Statistically, the correlation between speaking time and attention received is 0.584, which is significant with $p < 0.01$. This corresponds to a coefficient of determination R^2 of 0.341, meaning that speaking time explain 34.1% of the variance in attention received. To conclude, based on the observations made from empirical evidence, ***we validate hypothesis H2, i.e., the overall attention received is influenced by the amount of speech activity***.

3.2.3 Characterizing meetings and persons

Considering the *speaking time* and the overall *attention received* as two dimensions for analysis (Table 2), Meeting 2 presents an interesting case where both ST_i and A^i have low variation, showing that all the participants contribute equally while receive roughly equal attention- this corresponds to the *ideal meeting scenario.* Meetings 1 and 3 are cases where the variance along A^i is low and the variance along ST_i is high. Meeting 4 is the opposite: the variance in A^i is high while the speech activity for the various subjects is not very different; *i.e.* someone receives more attention than others, but the speech activity is almost identical across the group. We hypothesize that this corresponds to a group with an established leadership, while the leadership remains undecided in Meetings 1 and 3.

Finally, we define the Attention Quotient for a subject, denoted by AQ_i, as the ratio of the overall attention received to the overall attention given by the individual, *i.e.*, $AQ_i = \frac{A^i}{A_i}$ (Table 1(a)). It has been convincingly shown

Table 2: Characterization of meetings based on the variance in speaking time (ST_i) and overall attention received (A^i)

High ST_i,*High* A^i	*High* ST_i,*Low* A^i Meetings 1,3
Low ST_i,*High* A^i Meeting 4	*Low* ST_i,*Low* A^i Meeting 2 (Ideal meeting)

that meeting behavior can be strongly correlated with one's personality [8]. The 'Mission Survival' data also contains annotated ground-truths for the *Extraversion* and the *Locus of Control* personality traits. *Extroversion* is associated with assertive and highly outgoing personalities while the *Locus of control* (LOC) refers to an individual's nature to be self-determined and undeterred by external factors. In accordance with social psychology literature, we observe a positive correlation between AQ and the *Extraversion* and LOC traits.

4. AUTOMATED SOCIAL ATTENTION ESTIMATION

In order to validate the ground truth analysis presented before, we also performed automated analysis of the social attention. The long-term spectral divergence algorithm [12] is used to discriminate between speaking/non speaking regions, while the head-pose-cum-eye center estimation algorithm [15], is employed to estimate the point-of-gaze.The gaze estimation involves integration of a cylindrical head model-based pose estimation and an isophote-based eye-center locator to overcome shortcomings in both systems.

Results from automated analysis are presented in Table 1(b), based on which we validate **H3:** ***The use of eye-gaze in conjunction with head-pose improves accuracy of automated social attention estimation***.

We use A^i and A_i estimates to evaluate the accuracy of automated analysis against the ground truth. For A_i, the mean euclidian error between the estimates obtained when using the head pose only (HP) and eye-gaze with head-pose ($HP+EG$) are 7.62 and 5.33 respectively. This corresponds to an error reduction of 30%. For A^i, the corresponding errors are 6.28 and 4.59, yielding an improvement of 26.9%. Therefore, based on the automated social attention estimation results, ***we corroborate H3, i.e, employing eye-gaze along with head-pose cues improves accuracy of automated social attention estimation***.

5. CONCLUSION

We have proposed a multimodal framework to analyze social attention in meeting scenarios. To the best of our knowledge, this is one of the first attempts at: (i) simultaneously characterizing both meetings as well as participants by means of multimodal cues, and (ii) explicitly employing eye-gaze as a modality to estimate social attention.

We believe that our results will pave the way for future research connecting social attention with meeting roles and personality traits. To this end, more research is needed to determine the exact relationship between the attention received and given by a person, speech and postural activity, and personality.

6. REFERENCES

[1] S. O. Ba and J.-M. Odobez. Recognizing visual focus of attention from head pose in natural meetings. *Trans. Sys. Man Cyber. Part B*, 39(1):16–33, 2009.

[2] D. Gatica Perez. Automatic nonverbal analysis of social interactions in small groups: A review. *Image and Vision Computing*, 27(12), 2009.

[3] H. Hung and D. Gatica Perez. Identifying dominant people in meetings from audio-visual sensors. In *IEEE Face and Gesture Recognition*, pages 1–6, 2008.

[4] H. Hung, D. B. Jayagopi, S. Ba, J.-M. Odobez, and D. Gatica-Perez. Investigating automatic dominance estimation in groups from visual attention and speaking activity. In *ICMI '08*, pages 233–236, 2008.

[5] C. Kleinke. Gaze and eye contact: A research review. *Psychological Review*, 100:78–100, 1986.

[6] S. Langton. The mutual influence of gaze and head orientation in the analysis of social attention direction. *Q.J. Experimental Psychology*, 53A(3):825–845, 2000.

[7] S. Langton, R. Watt, and V. Bruce. Do the eyes have it? Cues to the direction of social attention. *Trends in Cognitive Sciences*, 4:50–59, 2000.

[8] B. Lepri, N. Mana, A. Capelletti, F. Pianesi, and M. Zancanaro. Modelling personality of participants during group interactions. In *UMAP*, 2009.

[9] N. Mana, B. Lepri, P. Chippendale, A. Cappelletti, F. Pianesi, P. Svaizer, and M. Zancanaro. A multimodal annotated corpus of consensus decision making meetings. *Language Resources and Evaluation*, 41(3):409–429, 2007.

[10] A. Pentland. Social aware computation and communication. *IEEE Computer*, 38(3):33–40, 2005.

[11] D. Perrett and N. Emery. Understanding the intentions of others from visual signals: neurophysiological evidence. *Cashiers de Psychologie Cognitive*, 13:683–694, 1994.

[12] J. Ramirez, J. C. Segura, d. l. T. A. Benitez, C., and A. Rubio. Efficient voice activity detection algorithms using long-term speech information. *Speech Communication*, 42(3-4):271–287.

[13] R. Stiefelhagen, J. Yang, and A. Waibel. Modeling focus of attention for meeting indexing. In *ACM MM*, pages 3–10, 1999.

[14] R. Stiefelhagen, J. Yang, and A. Waibel. Modeling focus of attention for meeting indexing based on multiple cues. *IEEE Transactions on Neural Networks*, 13:928–938, 2002.

[15] R. Valenti, Z. Yucel, and T. Gevers. Robustifying eye center localization by head pose cues. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 612–618, 2009.

[16] M. Voit and R. Stiefelhagen. Deducing the visual focus of attention from head pose estimation in dynamic multi-view meeting scenarios. In *ICMI '08*, pages 173–180, 2008.

[17] A. Whiten. Evolutionary and developmental origins of the mindreading system. In *Evolution and Development.* Lawrence Erlbaum, 1997.