Classification of Functional Attention in Video Meetings

Anastasia Kuzminykh
Cheriton School of Computer Science, University of Waterloo, Canada
akuzminykh@uwaterloo.ca

Sean Rintel
Microsoft Research
Cambridge, United Kingdom
serintel@microsoft.com

ABSTRACT
Participants in video meetings have long struggled with asymmetrical attention levels, especially when participants are distributed unevenly. While technological advances offer exciting opportunities to augment remote users' attention, the phenomenological complexity of attention means that to design attention-fostering features we must first understand what aspects of it are functionally meaningful to support. In this paper, we present a functional classification of observable attention for video meetings. The classification was informed by two studies on sense-making and selectiveness of attention in work meetings. It includes categories of attention accessible for technological support, their functions in a meeting process, and meeting-related activities that correspond to these functions. This classification serves as a multi-level representation of attention and informs the design of features aiming to support remote participants' attention in video meetings.

Author Keywords
video-mediated communication; meetings; attention; engagement; features

CCS Concepts
•Human-centered computing → HCI theory, concepts and models; Collaborative and social computing;

INTRODUCTION
Video communication is well known for the trouble people have with getting and paying attention [29, 53]. This is especially the case for hybrid work meetings in which participants are distributed unevenly over endpoints. Given that many mechanisms of situational awareness [32, 29] and social presence [34, 4] rely on visual cues, it seems especially galling that the value of the video channel ends up lost in translation [17, 24, 33, 66]. Attention's dual nature as both interactional and cognitive [39, 8] seems especially vulnerable to video communication's decades-long struggle with asymmetrical situational awareness [72, 19, 31].

CHI'20, April 25–30, 2020, Honolulu, HI, USA

ISBN 978-1-4503-6708-0/20/04…$15.00

DOI: https://doi.org/10.1145/3313831.3376546

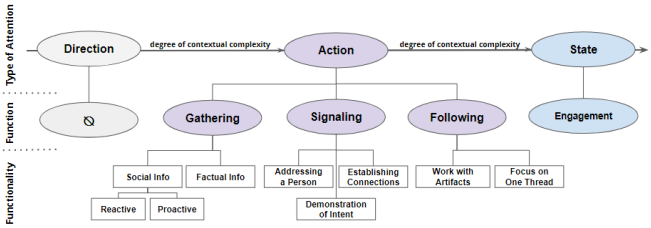

Figure 1. Classification of functional attention

Research on attention has ranged from gaze [28], through joint frame of reference [23] and situated awareness [27, 23], to broader concepts of engagement [3, 58]. There is a correspondingly extensive body of targeted solutions for providing situational cues for attention [70]. For example, systems have aimed to mitigate limitations of remote participation through tracking focus of attention [68] and eye-contact [18], selectively focusing bandwidth on participants' faces [75], supporting spatial arrangement similar to face-to-face meetings [77, 62], enabling remote users to point at objects within environments [48, 49], providing feedback on the conversation to participants [60], or immersive meeting environments [50, 22, 77, 41]. In sum, research on attention is very fractured [36]. Thus, there is a need for systematic understanding of what it actually means to support remote user's attention.

In this paper, we analyse participants' prioritization and sense-making of visual attention during video meetings at work. We focus on visual attention since video is proposed to support the many mechanisms of situational awareness and social presence rely on visual cues, because visual information is considered to be particularly important as a focus of attentional resources [16]. Our findings are drawn from two complementary studies. We first conducted semi-structured interviews to explore which aspects of visual attention contribute to participants' engagement in meetings, and how these aspects are affected by remote participation. To triangulate and further probe themes emerging from the interviews, we conducted a second, quasi-experimental, study in which we elicited participant descriptions of attention by asking them to narrate silent video footage of work meetings. These results provided us with more granular and procedural understanding of users' sense-making of the observable information on attention.

We synthesize our findings into a three-level functional classification of visual attention in meetings (see fig.1).

- The top level represents three *categories of attention* that a video meeting feature might support. First, *Attention as*

Direction – purely observable attention that constitutes a process, but not a recognized purpose (e.g. Alice is looking at Bob). Second, *Attention as Action*, which reflects attention processes with a recognized specific purpose (e.g. Alice is addressing Bob). Third, *Attention as State* – the overall sense of engagement, which reflects only purpose, but no particular process (e.g. Alice is attentive).

- The middle level represents *functions of attention*, i.e. recognizable purposes: There are three functions of *Attention as Action*: Gathering Information ('look to see'), Communicative Signaling ('look to be seen'), and Following Dynamic Processes ('look to be a part of a process'). A general Sense of Engagement ('be mentally there') is a function of *Attention as State*.

- Finally, in correspondence to each of three functions of visual Attention as Action, we elicited specific attention-related meeting activities, which we refer to as *functionalities*. Organized according to the corresponding attention functions, these functionalities form the bottom level of the classification.

Essentially, our model takes the amorphous concept of attention and develops a multi-level representation that focuses on what it means to **support** users' attention during a meeting. Attention actions are characterized by their procedural aspect (those amenable to technological augmentation) and functional aspect (common meeting activities) (fig.1). The intent of our model is provide a language that allows designers to be more deliberate when deciding and articulating what type of attention they wish to foster through their tools.

For example, consider the 'active speaker view' common in many video meeting systems (e.g. [14]). This feature uses audio to trigger enlarging or highlighting whoever is speaking at any given time and automatically switches the focus when a new person starts talking. According to our classification, this feature supports *Attention as Action* with the Information Gathering function. Another example is Companion Experiences for Microsoft Teams [55] which allows users to pair their computer and phone during a group call, for example, to present live mobile video or a photo while simultaneously maintaining the call on the computer. The Companion feature supports *Attention as Action* with the function of Process Following because it allows the presenting user to focus on an artifact in the meeting environment without losing social and situational awareness of the meeting in general. Finally, eye-tracking systems that correct remote users' perceived gaze direction [50], support *Attention as Direction*.

Our results reveal that the ability to perform purposeful attention actions significantly contributes to the overall sense of engagement of meeting participants. We suggest that *Attention as Action* yields the most promising initial direction for feature development. Arguably, though, the ultimate goal of features supporting user attention is to foster *Attention as State* – keeping remote participants generally engaged and mentally 'checked in'.

In the sections to follow, we first review previous research on attention in computer-mediated communication contexts, particularly highlighting visual attention issues. We then present our methodological approach and describe the design and results for each of our studies. Finally, we synthesize our findings as a functional classification of visual attention, discuss it in context of other theoretical conceptualizations of attention and situational awareness, and outline the implications for feature design.

PREVIOUS RESEARCH

Attention is above all a progressive process of selection that lies as much in our sense-making abilities as it does in our physical abilities. Attentional concepts in computer-mediated communication contexts have explored this notion of attention from a variety of standpoints.

Telecommunication studies of the 1970s considered the disadvantaged position of remote users to be a function of bandwidth of a medium's channels. Building on the concept of immediacy as a set of behaviors which increase involvement [44], Short et al.[64] introduced the concept of social presence to the context of telecommunication. They argued that social presence is a quality of medium itself and is enhanced by immediacy conveyed both verbally and non-verbally. From the 1980s onwards, explorations of social presence in video meetings argued that social presence depended not just on the medium itself (text, audio, or video) [47], but on interactions enabled and supported by this medium [12, 62, 69, 13]. Picard [51] suggested a focus on a medium's "affective channel capacity" - how much emotional information a channel can carry as compared to overall information. In the early 2000s research emphasis changed from a bandwidth orientation to a broader sense of social presence. For example, the Networked Minds Social Presence framework [4, 3] suggests that social presence is composed of three levels. A base, perceptual level is characterized by a sense of co-presence with the embodied other including observations of the other's identity, intentions, and attention. The middle subjective level is characterized by the psycho-behavioral accessibility of the other, sense of connection to other's intention, attention, and affective state. Finally, the top intersubjective level is characterized by a sense of behavioral engagement when actions are perceived interdependent with actions of the other.

Parallel research over similar periods emphasized the importance of shared environmental awareness. In video communication research this meant considering the need for more views of remote endpoints than the traditional face-to-face views, even as these extra views led to problems with establishing a joint frame of reference [23, 33, 43]. These problems stem as much from cognitive limitations as they do from interface issues [38]. The cognitive argument here is that while the environment is an extremely rich source of information, not all this information is equally important, and selectively attending to the information filtered as relevant helps communicators to establish shared knowledge [66] and avoid cognitive overload [33]. One of the cognitive behavioral mechanisms supporting and informing such prioritization involves directing more cognitive resources to targets of shared attention [66]. Shared or joint attention is a result of an ability and motivation to follow the direction of gaze or pointing actions [46] and develops

based on social awareness as signaled by physical proximity, head and body orientation, as well as dynamic gaze [52].

Gaze itself has been extensively studied as an attention mechanism [70, 36]. Research in social cognitive science shows that observers of complex scenes select the people's eyes as it allows observers to derive social information, more than bodies, background, or foreground objects [5, 6]. Further, people look at the eyes of others more frequently if the scene is highly social and includes several people engaged in joint activity [5, 6]. See Frischen et al. [21] for an overview of past research on the perception of gaze behavior and its effect on the observer. Eyes both collect information and communicate one's state [2] and attentional focus [21]. The dual nature of gaze has been adopted by modern cognitive science to explore naturalistic social attention [56]. For instance, in collaborative processes, the coupling of gaze patterns between interlocutors has been found to be causally related to the knowledge people share before a conversation and the information they later recall [54]. Gaze thus supports establishing common ground [1, 10], for example by disambiguating references [30] and helping to predict next task actions [76].

Video-mediated communication researchers have explored various non-verbal cues [68], with many focusing on detecting gaze [50] and eye-contact [18], or simulating gaze [48, 49, 73]. Gaze coordination has also been extensively explored in human-robot interaction [9, 45, 57, 58]. For example, people understand robot speech better when a robot's real time gaze behavior is similar to that of humans [67], and when a robot is interacting with multiple people, the visual focus of attention affects addressee recognition [63].

While gaze is tightly related to general and social attention processes, social cognitive studies shows the specifics of this relationship are not necessarily linear. Gaze might mean attention but attention does not always mean gaze. For instance, analyzing participant's viewing behavior during social interactions, Freeth and colleagues [20] demonstrated that people look at the partner's face significantly less when answering a question compared to when they are listening to the question being asked, which supported earlier research showing that averting one's gaze from other people can help one to think more effectively by reducing visual processing demands and cognitive load [15, 25]. Thus, while gaze indicates focus of visual attention it does not necessarily reflect focus of cognitive attention. Further, for gaze to play a signaling function [2] and invite joint attention in communication [59, 65] an interlocutor needs to recognize not only the focus of directed attention, but also the environmental context for such attention [27].

The range of different approaches to attention and technologies created to overcome attention problems suggests that designing corresponding features requires a better granular baseline for understanding what aspects and processes of attention should be augmented by technologies. Our research aims to identify the fundamental types of attention in meetings, accessible for technological support, and how they might be recognized during meetings. Thus, we explore the functional aspects of attention, i.e. how and why people pay and perceive attention.

METHOD

To probe participants' prioritization and sense-making of attention processes, we conducted two studies. First, we interviewed participants about attention in work meetings. We then quasi-experimentally elicited descriptions of videos of work meetings to triangulate themes emerging from the interviews.

We took a qualitative approach in both studies because our goal was to explore the way that participants understood attention and prioritized contextual cues. Interview-based research on video-mediated communication tends to report results that are well-suited to developing thematic glosses and evaluations. However, while people are generally good at providing narratives, they tend to be less successful at remembering the granular cues from which they weave these narratives. Additionally, thematic prioritization in self-reported information tends to relate to what is remembered. On the other hand, experimental procedures develop more replicable results that lend themselves to comparison and contrast, as well as tracing flows of action, but lack contextual richness and require deciding *a priori* on a subset of observable measures. Hence, for this work, we started with semi-structured interviews to gather open-ended accounts of meaningful attention, followed by a quasi-experimental study to act as a focus-check and expansion of those accounts. This data triangulation allowed us to cross-validate our observations between the two studies.

Study I Design

Through a series of semi-structured interviews [71], we explored users' experiences of participating in work meetings of various configurations, with a particular focus on attention processes and engagement. In total, we interviewed twelve participants (age 21-50, 5f,7m). All had a tertiary degree in diverse areas of knowledge, including engineering, science, and arts[1]. All participants had extensive experience in co-located, remote, and hybrid meetings (at least once a month for at least over a year), and each participant was a full-time employee at the time of the study.

The interviews were conducted one-on-one, in-person, and informed consent was obtained prior to the interview. Each audio recorded interview lasted approximately 40 minutes. We started by asking participants about their preferences for meeting configurations and what motivated those preferences. Through these discussions we also explored their practices and expectations for joining meetings in person and remotely, with video or audio-only channels available. We then concentrated on remote meeting participation experiences, asking participants to describe the strategies they use to understand a meeting environment and social dynamics during a meeting, strategies to recognize and support their own and other participants' attention and engagement, and obstacles related to attention dynamics that remote participants might experience.

After transcribing audio recordings of each interview, we performed incremental data analysis across participants using open and axial coding [11]. Each phase of coding was initially

[1]Note that individuals who engage in hybrid work meetings are often in professions that require tertiary degrees.

performed by the primary author, and then codes were discussed and refactored in consultation with the research team.

Study II Design

To validate and extend the granularity of the first study results, we conducted a second, quasi-experimental study, asking participants to narrate muted videos of work meetings, focusing on attention processes they observe. We recruited a new set of fifteen participants (21-36 y.o, 8f, 7m), intentionally excluding those who participated in the first study. We wanted to prevent a potential primacy bias and to support the external validity of the results by extending and diversifying the participants pool. While the goal of the first study was to generally explore participants' experiences regarding attention and engagement during work meetings, the second study focused on further understanding the prioritization and sense-making of purely visually observable cues of attention.

Participants were presented with six short video records (2 minutes) of real work meetings. Meeting scenarios included 4 to 6 meeting participants, either all co-located, or in a hybrid meeting configuration. The video clips reflected different combinations of the following meeting processes: joining and leaving a meeting environment, setting up a meeting environment, talking around a table, taking notes, using a whiteboard, using a projector screen, using smart devices. The videos were presented with muted audio, and participants were asked to narrate the visual information from the video to an experimenter sitting next to them. We chose the muted condition for two reasons. First, we wanted to eliminate potential biases on information interpretations based on the audio context. Second, we wanted participants to focus on the purely observable information and expected that the lack of audio context would ensure participants' focus on visual cues. At the beginning of the experiment participants practiced using an additional training video clip of a work meeting. After the experiment, we conducted a debrief interview asking participants about their experience and strategies for information descriptions and prioritization.

This second study was quasi-experimental in that we controlled the stimulus and required participants to follow a particular protocol, but we did not go so far as to develop and test hypotheses. Our analysis focused on participants' verbal descriptions of attention processes. From it, we developed a set of 'attention identifiers' – distinct meaningful elements of a description that signified an act or process of attention. Attention identifiers were most commonly denoted with a verb, examples of which included "to look", "to watch", "to listen", "to pay attention", "to direct attention", "to be attentive to", "to be engaged", "to talk to", "to have a conversation with", etc. In total, we identified and semantically analyzed 225 responses of observable visual attention from 90 descriptions of 6 videos.

STUDY I RESULTS

In this section, we present our interview findings (fig.2). We explore what functional aspects of attention contributed to people's sense of meeting engagement, and categorize them as Gathering Social Information, Following Dynamic Processes, and Communicative Signaling. For each function we describe the elicited activities relevant to the meeting process.

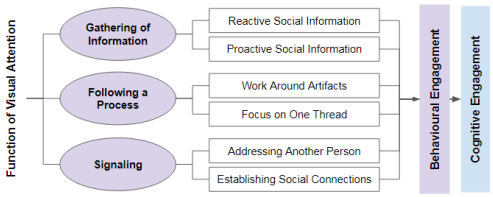

Figure 2. Results of the Interview Study (Study I)

Functions of Visual Attention in Engagement

Exploring the role of visual attention in users' sense of engagement, we first examined what information participants generally perceive as relevant and pay attention to during meetings. We then cross-analysed it with the information that participants perceive as more difficult to acquire in different meeting configurations. We found that, despite having a video channel, remote participation is perceived to affect user's access to three functions of visual attention (fig.2).

Gathering Social Information

The first function of visual attention affected by remote participation is Gathering Social Information, which refers to the ability to support social situational awareness. Even with a video channel, remote participation is perceived to constrain the user's ability to collect social information required for suitable behaviour adjustments.

In our analysis, we distinguish Gathering Social Information as reactive or proactive. Gathering *Reactive* Social Information indicates responses to one's actions and used to assess the social environment and required adjustments. Gathering *Proactive* Social Information indicates the social disposition in the environment and required to initiate one's action. Correspondingly, two themes emerged in participants' concerns regarding the accessibility of social information in remote participation: gathering of information on emotional reactions of others (reactive), and assessing social-behavioral cues to manage their own participation (proactive).

Limitations on Gathering *Reactive* Social Information affect the user's ability to assess the reception of their actions, and, thus, the ability to appropriately adjust their social behaviour:

> *"It's about being able to see the other person, their facial expressions, body language very clearly, which you can't always do even on the video. It is easier [in person] to get a feel for how the other person is reacting." [P12]*
> *"If I could see everyone, which we can't usually do, but let's say we could, so I can present and see how they are reacting, I don't know if I would need to be there in person." [P9]*

In addition to facial expressions and body language, this also includes more complex social-emotional cues, e.g. other participants' engagement and attention to one's actions.

> *"I'm looking for people nodding or smiling... I might rely on it subconsciously to see if people are engaged or not,*

to see whether people look interested or excited." [P4]
"During a Skype meeting I feel like it's difficult to say whether someone is paying attention." [P2]

Limitations on Gathering *Proactive* Social Information affect the remote participant's ability to appropriately initiate their action according to the current social environment. This, first, includes the information required to inform appropriate speech patterns – social cues on when is it appropriate to start talking:

"When remote, it's an additional challenge - you don't see cues that will let you say 'I'm gonna speak now'." [P9]
"In person you have all these non-verbal cues like you can see when someone is about to say something." [P10]

It also refers to gathering information on who is in the room, their social roles, and social dynamics between participants:

"It is also hard to figure out who is important, what's happening in the room." [P3]
"The problem is that you don't know who is looking at you. When you are there, you can move your head, see what's going on. On camera you have no idea." [P9]

Following Dynamic Processes

The second function of visual attention, Following Dynamic Processes, refers to the user's ability to dynamically direct their attention to follow the meeting's work. Two themes emerged in relation to this function: participation in work processes around environmental artifacts, and navigation within multiple conversation threads.

Following Dynamic Processes around environmental artifacts mainly included well established difficulties of working with whiteboards and projected presentations:

"I was joining remotely, others were in person. I was clearly sensing disadvantage. Because people were working on a white board and it was difficult to see." [P6]
"It's hard if there are slides or other materials... maybe I would like some indication of what's going on." [P1]

Another related challenge is the ability to follow one of multiple conversation threads, especially in hybrid settings:

"If I am remote and multiple threads of discussion develop in the meeting room, then it's like 'how do I get streamlined into one thread of discussion'." [P7]
"If a big meeting splits into several conversations, I just switch it off all together, because I can't join one of those conversations. But if I am in the room I can turn to my neighbour, have a subgroup." [P11]

Communicative Signaling

Lastly, remote participation affected the Communicative Signaling function of visual attention. While people are gathering and following, they also signal their reactions and intentions as part of communication, i.e. an act of visibly paying attention is itself social message to other meeting participants. The Communicative Signaling function contributes to a group communication dynamics. For instance, a constrained ability to use Communicative Signaling results in difficulties of directly addressing others:

*"Even just looking, talking to that person directly makes a huge difference. And it would be nice to know that people are speaking to *me*." [P2]*
"If somebody addresses me, in person they turn, face me. In a remote meeting they would have to turn to the camera, or say my name. All these additional decorations have to happen to ensure that I am included adequately." [P4]

Constraints on the Communicative Signaling function are also reflected in the user's reduced ability to manage their social interactions and, correspondingly, to appropriately participate in the meeting dynamics:

"In person, I make sure that I'm not just talking to one person or one direction in the room. You need to switch focus, talk to everyone. That's not something I can control remotely – I'm only looking at whatever I can see." [P3]
"When in person, you can look around the room, make contact with people who aren't talking. You know, make connections with other people. You can't do it remotely, you are only looking at one thing." [P5]

Similar to the gathering and following functions, the ability to use the Communicative Signaling function of attention contributes to the overall sense of engagement, both for a remote participant themselves, and for the other participants' perceptions:

"If a person is not trying to engage the crowd, then I lower my own engagement. I just kinda put it on the background. It's hard to engage with someone who isn't making an eye contact at least with someone." [P3]
"On Skype, you have video, you can notice when somebody is not paying attention... and you feel like people are watching you, so you can't stop paying attention." [P9]

Section Summary

To summarize, we explored the role of visual attention in a remote participant's engagement and found that people describe remote participation as affecting distinct functions: Gathering Social Information, Following Dynamic Processes, and Communicative Signalling. Each of these functions contributes both to people's general sense of engagement and is also associated with corresponding functionalities, including assessing reactions of others and initiating of one's own participation, working around artifacts, focusing on one of multiple discussion threads, addressing specific participants, and establishing social micro-connections.

STUDY II: NARRATION QUASI-EXPERIMENT

To further explore the granular prioritization and sense-making of attention processes, we conducted a second study in which we asked participants to narrate the attention processes they observed in silent videos of work meetings. This resultant attention descriptions were, of course, quite heterogeneous (fig.3), but we found a useful thematic distinction between purely descriptive and functionally interpreted attention. The results of second study validate our earlier findings on three functions of visual attention, and expand the set of related functionalities.

This section, first, presents our analysis of 'attention identifiers' – distinct descriptive elements signifying an act or process of attention. We then deepen into the participants' contextual interpretations of attention processes reflected in these identifiers, and describe the elicited types of interpreted attention, followed by the analysis of their interpreted functions.

Types of Attention Descriptions

We began by analyzing the differences in the attention identifiers used by participants to denote attention-related processes. Based on this analysis, we categorized attention identifiers as either descriptive or indicative (fig. 3). The descriptive identifiers were 'objectively' translations of visual information into a verbal form ("looking at an object"), representing narrations of purely observable attention. The indicative identifiers were triggered by the same visual information, but were formulations of the observer's interpretation of the social relevance of the visual cues.

Descriptive Attention Identifiers

Descriptive attention identifiers were 'objective' narratives of visual information, simply indicating the direction of gaze or body orientation (e.g. *'now both guys are looking into the phone'* [V6P10]). They were 'objective' in the sense that they did not contain any information on the purpose of the corresponding attention processes, but only registered their occurrence.

For example, in the description: *"The guy with the long hair is looking at the guy with the Coke"* [V3P10], the descriptive attention identifier "looking at" purely denotes that the gaze of "the guy with the long hair" is directed towards "the guy with the Coke". For further illustration consider the following examples with descriptive attention identifiers: *"Man number 2 is **looking at** the whiteboard"* [V4P6]; *"Person 2 is laughing and their **attention was just directed** to the screen"* [V1P4]. The type of observable attention signified using descriptive identifiers is purely registrable, thus it relates the function of Attention as Direction.

Indicative Attention Identifiers

In contrast to the 'objective' nature of descriptive identifiers, indicative attention identifiers include narrators' contextual interpretations of observable information (e.g. *"people still seem to be paying quite close attention to the speaker"* [V3P7]). Thus, the indicative identifiers 'indicate' the functional aspect of the corresponding attention process, according to the narrator's interpretation. For example, the description *"he is getting involved in this conversation with the person on Skype"* [V6P1] narrates the scene of a man looking at the screen displaying another person. In this description, the indicative attention identifier *"getting involved"* is formed based on the visual information which is contextually interpreted as engaging into a Skype conversation. While indicative identifiers are still triggered by atomic visual stimuli (recall that participants were presented with muted videos), the corresponding attention descriptions are produced as declarative as opposed to purely descriptive: *"They are **listening** to the woman... **listening intently**."* [V2P5]; *"The women is **responding** to what the guy was saying"* [V5P14].

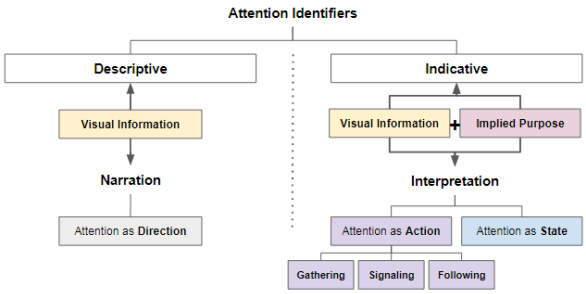

Figure 3. Results of the Narration Experiment Study (Study II)

Participants would also commonly augment descriptions of attention processes by interpreting related social dynamics: *"The person who is videoconferencing is speaking because everybody is looking at the screen to listen to this person"* [V4P8]; *"She must be explaining something because everyone is looking at her and nodding as well"* [V2P9].

This tendency to augment descriptions with contextual interpretations was further reflected in data from debrief interviews:

> *"I interpret the gesture as having some meaning in a sort of conversational dialog, back and forth. It was both the captured motion and then the meaning of the motion that I was interpreting." [DebriefP3]*
> *"I also tried sometimes to add additional features like 'they are doing some actions' but not all the actions, I tried to find the leading person every time, who is leading, who is speaking, who is drawing - because I find that that's how the pipeline of the meeting is happening." [DebriefP12]*

Participants omitted the functional augmentations of attention descriptions if the communication context was missing or unclear. In this case, they would resort to the descriptive identifiers to merely capture the direction of attention.

> *"That could be tricky because sometimes it's not clear whether there is something currently being spoken about which warns them to look at something." [DebriefP6]*

The 'interpretive' nature of indicative identifiers constitutes the functional aspects of the observed attention processes – their perceived contextual meaning.

Types of Functionally Interpreted Attention

Indicative attention descriptions showed that corresponding contextual interpretations included either descriptions of attention as general cognitive state (e.g. *"they seem to be quite engaged"* [V5P12]), or descriptions of attention as a purposeful action (e.g. *"the blond woman is listening to the conversation he is having with the woman with dark hair"* [V5P10]).

While descriptions of attention as a cognitive state were triggered by visual information, they typically did not include information on related behavioural displays. For example, in the description *"They seem to be quite engaged"* [V5P12], the indicative attention identifier "to be engaged" describes the overall state rather than a specific behaviour. Moreover, these descriptions do not specify a direction or a target of attention:

"Otherwise people seem to be pretty attentive" [V3P7]; *"The man is not paying attention anymore"* [V5P10].

In contrast to the descriptions of attention as a cognitive state, descriptions of Attention as Action include a particular direction and a specified target of an attention process. For example, in the description *"person 4 is talking again to persons 1 and 2"* [V5P4], the indicative attention identifier "talking to" implies a direction of the attention act toward a target – "persons 1 and 2". Similarly, the following examples of Attention as Action include a target, denoted in both cases by a pronoun "him": *"The guy in glasses is talking and everyone is listening to him"* [V3P10]; *"Now person 2 on the right is talking and the other two are paying attention to him"* [V1P1].

Furthermore, descriptions of Attention as Action were formed based on an interpreted purpose of observed actions. For example, *"listening"* is a contextual interpretation of atomic visual information, such as *"looking in a direction"* of someone who is speaking. For more illustration, consider the following descriptions: *"Really it's two people talking to each other and other two are just kind of watching them"* [V2P7]; *"The blond woman is just listening to the conversation he is having with the woman with dark hair"* [V5P10].

Informed by the first study results, we analyzed these inferred intentions in terms of the three functions of our model: actions for Gathering Information, actions intended for Communication Signalling, and actions for Following Dynamic Processes.

Gathering Information

The attention actions for Gathering Social Information collect can be described as 'look to see': e.g. *"Man 1 is checking the screen of the smart device used by man 2"* [V6P6]; *"The person who is standing kind of looks at what he'd drawn"* [V4P7]. While the interview data from the first study revealed the theme of Gathering Social Information (sec.4.1.1), the quasi-experimental data extends those results by demonstrating that all types of environmental information are noticed and appears in description of the attention processes. e.g. *"They are talking about something... yeah... the two guys at the back don't, they are just listening"* [V2P2].

Communicative Signalling

If actions for Gathering Information can be described as 'look to see', then actions for Communicative Signalling are the converse: 'look to be seen'. These actions are performed to be acknowledged by other meeting participants, e.g. *"The woman is talking. She is addressing both men."* [V1P6]. Communicative Signalling typically refers to either displaying attention to a process (e.g. *"person 2 on the right is talking and the other two are paying attention to him"* [V1P1]), or to directing speech to specific meeting participant (e.g. *"Man 3 is addressing everyone at the table, making eye-contact with everyone"* [V3P6]).

Our interview results identified two main functionalities for Communicative Signalling: establishing social micro-connections with other meeting participants, and directly addressing specific people (sec.4.1.3). Again, While validating these results, the quasi-experimental data also expands them by revealing a functionality of demonstrating intent, when a meeting participant displays their engagement in the process, e.g. *"he doesn't seem to be acknowledging the conversation, he is kinda writing something down in his notebook"* [V5P4].

Following Dynamic Processes

The third type of inferred purposes of attention actions is Following Dynamic Processes, describable as 'look to be a part of [a process]'. These actions combine the Gathering Information and Communicative Signaling purposes with an addition of a dynamic target of the action (e.g. adding information on a whiteboard). For example, *"They are adding drawings to the board, looking at it intently. And discussing the best way to do this"* [V4P11]; *"Everyone is very interested in what this person is about to do on the board"* [V6P3]; *"he is talking to the guy who is drawing, instructing him on how to draw stuff or something like that"* [P2V6]. Additionally, these actions require a person to be able to dynamically switch between a complex combination of targets, including social and environmental targets, e.g. *"they sometimes refer to the screen and say something, discuss something that's displayed on the screen"* [V1P14].

Thus, the quasi-experimental data supports one of the two themes on the ability to direct remote participant's attention from study I: obstacles to participating in work processes involving environmental artifacts (sec.4.1.2). The theme of navigating within multiple conversations did not appear in the quasi-experimental narrations, probably due to silent nature of the video stimuli.

Inter-dependency of Functional Types of Attention

Finally, we explored the inter-dependencies between descriptions of one's attention as cognitive state and attention as a purposeful actions. Validating the first study's results, the second study debrief interviews also suggested that cognitive engagement is significantly influenced by one's ability to perform specific attention actions. This, for example, includes the ability to gather information on other participants' attention-related actions, and to noticeably (behaviourally) participate in the meeting process:

> *"what was important to the members of the meeting – what they would be looking at, carrying about. Rather than details on what's someone is not doing or what is not being looked at - it's not relevant. The contribution to the meeting process is critical, the ones who don't speak or appear to be paying attention tend to be kind of ignored." [DebriefP6]*

The ability to noticeably participate in the meeting process also affects the social inclusion of the participant:

> *"If someone doesn't say anything, doesn't do anything, you can just forget them." [DebriefP2]*
> *"Maybe someone walked into the room but they were not noticed as a new player by the people who are already in the meeting – like someone else is already having a conversation and a new person came in and they didn't immediately acknowledged that person coming in - then it wasn't as important, as if that person came in and started talking." [DebriefP4]*

Section Summary

The analysis of attention descriptions allowed us to discriminate between three types of noticeable attention in the user's sense-making: Attention as Direction, Attention as Action, and Attention as State. We categorized the interpreted intentions associated with attention actions as Gathering Information, Communicative Signaling, and Following Dynamic Processes, which validated our earlier results of three functions of visual attention. Finally, the results from both our studies suggest that the ability to perform specific attention actions and, thus, the accessibility of corresponding attention functions contribute to the participant's overall sense of engagement in a meeting.

To highlight these commonalities and to operationalize them for feature design in video meetings, in the following section we synthesize our findings as a three-level functional classification of observable attention.

FUNCTIONAL CLASSIFICATION OF ATTENTION

Based on cross-validation of the results from the interview study and the analysis of the narration experiment data (fig. 2 and 3), we constructed a three-level classification of functional types of observable attention in work meetings (fig.1). The top level represents functional types of attention, organized according to the degree of the contextual complexity required to recognize and interpret each type of attention. The second level represents functions of attention types, affected in video meetings. The bottom level organizes the functionalities of attention actions – tasks and activities within a work meeting, – corresponding to the second-level functions of attention.

Level 1: Contextual Complexity Spectrum

First, the classification includes three types of attention, accessible for video meeting feature support. These types of attention are organized on a spectrum left to right, from the least to the most contextually complex for recognition.

Attention as Direction represent the contextually simplest type of noticeable attention processes without a recognized purpose. In fact, both our studies demonstrate that Attention as Direction cannot be fully considered as a functional type of attention. Instead, it encompasses purely observable attention processes, which, in turn, can be used to infer social meanings for either Attention as Action, or Attention as State.

For instance, the interview participants acknowledged that they considered the visible information as Attention as Direction, but also articulated that these visual cues are not fully informative for assessment of one's attention actions and engagement:

> *"If it feels like I am looking at you and I'm like 'no, I'm not, I just got an email on the screen. So you only see if they are looking at the screen or not." [P2]*
> *"because camera is over here and they are looking at the screen, so I can't tell whether they are actually looking at me or at something completely different." [P9]*

Similarly, the second study shows that descriptions of Attention as Direction might, in fact, be incomplete descriptions of Attention as Action, when contextual information required for interpretations is missing or unclear (sec.5.1).

Attention as Action refers to the attention processes with specified direction and target, as well as recognizable functional meaning (purpose). For instance, in our interview data, Attention as Action would be described as observable but interpreted actions: *"people are looking at you, listening to you"* [P1]; *"you can know by looking at someone who is making eye contact when talking"* [P3]; *"them reading their phone"* [P9]. We positioned Attention as Action in the middle of the contextual complexity spectrum, between Attention as Direction and Attention as State. While Attention as Action augments Attention as Direction with the functional meaning of the processes, it is less contextually complex than Attention as State. First, structurally, descriptions of Attention as Action include a direction and a target of an attention process, which makes them easier to recognize in the environment, and second, because it causally affects Attention as State – the ability to perform attention actions contribute to the overall sense of engagement.

Lastly, the most contextually complex type of attention is *Attention as State*, which reflects the overall sense of one's engagement in the meeting process. In other words, when we casually talk about paying attention to the meeting and being engaged, we generally refer to the Attention as State. For example, in our interview data, Attention as State would be described as general engagement or *"directing all my attention [to the meeting]"* [P11], while the lack of Attention as State would be described as *"I don't have 100% of my attention available"* [P7]; *'I check out mentally'* [P10], or *"put [the meeting] on the background"* [P3].

Our classification positions the functional type of Attention as State on the far right end of the contextual complexity spectrum for several reasons. First, Attention as State seems to be directly affected by the accessibility of Attention as Action. Second, the structural analysis of corresponding attention descriptions shows that descriptions of Attention as State do not include a specified direction or target of an attention process (e.g. being *"pretty attentive'* [V3P7]). Thus, Attention as State reflects only purpose, but no particular process associated with it, which makes it the most complex to recognize and support in the environment.

Levels 2 and 3: Functions and Functionalities

Since Attention as Direction refers to purely observable attention that constitutes a process, but not a recognized purpose, the corresponding second and third levels of the classification are empty (fig.1, left). Attention as State only has a general purpose of being attentive and engaged (fig.1, right). Finally, Attention as Action (fig.1, middle) augments Attention as Direction by having both a process and a purpose. These types are represented on the second level of the classification and reflect three identified functions of attention: Gathering Information, Communicative Signaling, and Following Dynamic Processes. Each of these functions corresponds to a set of functionalities within a work meeting – specific tasks and activities performed through the attention actions, which form the third level of the classification.

Gathering Function

First functional type of attention actions includes the actions directed to the passive collection of information from the environment. In the interview results, information gathering mostly described collecting information for social situation awareness: either reactive or proactive social information. The experiment data, however, allowed us to expand this category by including the collection of 'factual' information as well.

The functionality of passively Gathering Reactive Social Information refers to the ability to assess the reception of one's communication action by the other meeting participants, e.g. *"If you are doing a presentation for a client, you want to know how they are receiving things, when they are losing attention, when they do take notes."* [P5]. The functionality of passively Gathering Proactive Social Information reflects the ability to assess the social situation to appropriately initiate one's actions, e.g. *"Cues as to when people look like they can be interrupted, it's a bit harder when you are doing it remotely."* [P4]. Finally, the functionality of Gathering 'Factual' Environmental Information refers to one's ability to passively collect the non-social, content-related information during the meeting process (e.g. *"checking the screen"* [V6P6], *"just listening"* [V2P2]).

Communicative Signaling

The second functional type of attention actions includes actions performed to be acknowledged by others. Both the interview results and the narration experiment data identified two main functionalities within the signaling function of visual attention: establishing social micro-connections with other meeting participants, and directly addressing specific people. The analysis of the quasi-experimental data allowed to add the functionality of demonstrating one's engagement in the process, further supported by revisiting the interview data.

The functionality of establishing social micro-connections allows meeting participants to *"make contact with other people who aren't talking"* [P5] as well as to gain attention of others, e.g. *"In person we rely on quite a lot eye contact and people on calls, they tend to do a lot to make sure that they grab the attention of the one person that they specifically need."* [P2]. The functionality of addressing a person reflects one's ability to direct their communication message to a specific subset of meeting participants: *"It would be great if you could direct your conversation towards a certain group of people... like 'I want to talk to A and B'. But yeah... that's not really possible."* [P11]. Demonstrating one's engagement refers to one's ability to purposefully display their interest and involvement in the process, e.g. *"you can know if they are paying attention by looking at who is making eye contact when talking"* [P3].

Following Dynamic Processes

Finally, attention actions with the purpose of Following Dynamic Processes combine information gathering and signaling functions with addition of the dynamic nature of the target. This means that Following Dynamic Processes is directed to the dynamic changes within the immediate environment, which consequently allows a participant to affect the environment if needed, thus, reflecting an ability to contribute.

Both our interview results and the results of the narration experiment showed that in the meeting context, functionalities of Following Dynamic Processes often revolves around being able to follow the work on a whiteboard or screen (e.g. *"he is talking to the guy who is drawing, instructing him on how to draw stuff or something like that"* [P2V6]), or to follow a specific conversation when a discussion splits: *"If somebody is talking, trying to convey something and then somebody mixed into that whispering. It's not distracting to the group, it's just distracting to you [when remote]".* [P9].

DISCUSSION

In this paper, we address the need for systematic understanding of how attention processes might be supported through video meetings features. The consolidated results of two studies on prioritization and sense-making of attention information allowed us to construct a three-level classification of functional types of observable attention in meeting communication. The classification shows that attention processes that might be accessible for the feature support fall into one of three categories – Attention as Direction (only process), Attention as Action (process and purpose), and Attention as State (only purpose) – some of which have corresponding functions and functionalities.

The first category of visual attention that a video meetings feature might support is Attention as Direction, which simply denotes the observable information on a direction of one's gaze and pose. Attention as Direction presents a fairly straightforward entry point for developing new video meeting features, for example, due to the possibility for direct automatic recognition [48, 49, 73]. However, while this information might be used to infer that one is 'paying attention', it is considered to be not particularly informative on its own. For example, a specific instance of Attention as Direction might either reflect an 'accidental' target of averted gaze (e.g., Alice is looking at the desk while listening to Bob) or a partially perceived attention processes when there is not enough context to infer the purpose of an attention act. Thus, although this type is the simplest for technological support, Attention as Direction can not be fully considered as a functional type of attention.

Attention as Action and Attention as State, however, differ from Attention as Direction by augmenting this pure visual information with a contextually interpreted function of corresponding attention processes. The implied purpose of attention processes can be compared to the notion of a goal in goal-directed actions in the activity theory [74]. Within a three-level structural model of activity, Leontiev [42] differentiates between object-oriented activity, goal-directed actions, and basic operations. While an object is a motivation for the complex activity, goals of actions are more temporary and individually focused. Furthermore, Attention as Direction can be loosely compared to the Leontiev's level of operations that do not have their own goals [40]. For in-depth analysis on activity theory see [40, 74]. Similarly, describing pointing as situated practice, Goodwin shows that an act of pointing becomes interactive through an association with a purpose recognized in the context of the communicative situation [27]. Goodwin uses the term 'activity framework' to refer to a can-

didate target connected to the system of recognizable activities within which that target is embedded. This approach suggests that as well as the object that one is pointing at (as in Attention as Direction), the possible operationalized implication of this pointing act (as in Attention as Action) should be considered part of the meaningful communicative practice.

The differences between Attention as Action and Attention as State also find reflection in previous research. For example, Shteynberg [66] highlights the distinction between the psychological state of shared attention and the activity of attending together with another social agent, noting that while these phenomena can affect each other, they might also exist independently. The interdependent implications of awareness as a state and of the 'shape' of actions related to participation in the process are also discussed by Heath et al. [33].

Furthermore, our results demonstrated that the accessibility of means for the efficient behavioural engagement (performance of Attention as Action) contributes to the participant's overall sense of engagement. We found that these effects seem to be precipitated by the limitations that remote participation puts on the three functions of visual attention.

While the Gathering Information function might be the most intuitive, the importance of the Communicative Signaling function of attention is also reflected in previous research. In Biocca's words, "social presence is not just sense of the other, but it is very much a sense of the others of 'me'" [3]. Similarly, Goodwin [26] has previously acknowledged that for an action to become social it should include not only the party producing the action, but the recognition and understanding of this action by others present in the social situation. Analyzing the literature on Social Attention, Salley and Colombo [59] offer a framework of conceptual approaches based on the functions of social attention process. They distinguish social attention as the social behavior enabling interaction with the social world, social attention as social motivation to attend to and engage with the social world, and social attention as basic visual attention to the social world. Within this framework, attention as social behaviour can be mapped to the Communicative Signaling function in our classification, as it outlines the communication aspect of attention, whereas the basic visual attention approach can be mapped to the Information Gathering function, since it refers to inert acquisition of information.

The central characteristic of the Following Dynamic Processes function is that this attention is paid to the dynamic changes within the immediate environment which consequently allows a participant to affect the environment if needed. Previous research has recognized the importance of responsiveness to other's activity within the mediated environment [3], as and the relevance of the dynamic awareness of the local environment as it often contains "a diverse and shifting display of different forms of information which are more or less relevant to the activities in which participants engage" [33]. This is partially supported in video meeting systems that enabling users to refer to and point at objects and artifacts within each other's remote environment [43, 48, 49].

Leveraging the Attention Classification in Design

Our classification revolves around procedural aspects of attention. The categories are not mutually exclusive and instead build upon each other (e.g. Attention as Action must include Attention as Direction). This structure aids in system design by providing a language that allows a developer to identify what type of attention and what specific capabilities a particular feature might choose to enable. In other words, instead of developing video meeting features intended to foster attention 'in general', the classification allows developers to identify and design targeted features to deal with specific attention-related problems. For example, features aiming to support the Information Gathering function of Attention as Action should involve timely augmentation of information streams, e.g. narrating, zooming-in, or focusing system's view or bandwidth (e.g. [7]). Supporting the Communicative Signaling function of Attention as Action necessarily requires some form of notification about one's actions delivered to other meeting participants (e.g. [37]). Augmenting the Following Dynamic Processes function of Attention as Action should include the remote participant's ability to split system views, signal their intentions, and selectively contribute to simultaneous processes (e.g. [55]). Supporting Attention as Direction, on the other hand, would only require augmenting information on one's gaze and/or pose direction, e.g. providing gaze heatmaps or projecting gaze directions on the video stream (e.g. [61]). Finally, directly augmenting Attention as State is challenging since it does not constitute any particular process, however, once achieved, Attention as State can be supported, for example, by eliminating distractions in remote participant's environment (e.g. [35]).

Furthermore, the presented classification informs how attention can be deployed in different ways by participants with different abilities in different meeting configurations.

CONCLUSION

In this paper we have presented an empirically grounded functional classification of observable attention in video meetings. The classification includes categories of attention accessible for technological support, their functions in a meeting process, and meeting-related activities that correspond to these functions. We demonstrate that a particular type of attention – Attention as Action – provides special interest for technological feature development due to its structure and effects on participants' engagement. Designing a one-size-fits-all solution to 'fixing' attention asymmetries in video meetings is challenging because of the astonishing range of idiosyncratic conditions and personal needs. We hope that the granular model presented here could inform the design of systems that can capture a wide range of behaviors and correspondingly balance personalized attention processes with group attention needs that could be contextually specified.

ACKNOWLEDGEMENTS

We thank participants in our study, and Dr. Edward Lank for editorial suggestions. The research in this paper was reviewed and approved by the institutional Office of Research Ethics.

REFERENCES

[1] Manabu Arai, Ellen Gurman Bard, and Robin Hill. 2009. Referring and gaze alignment: Accessibility is alive and well in situated dialogue. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 31.

[2] Michael Argyle and Mark Cook. 1976. Gaze and mutual gaze. (1976).

[3] Frank Biocca and Chad Harms. 2002. Defining and measuring social presence: Contribution to the networked minds theory and measure. *Proceedings of PRESENCE* 2002 (2002), 7–36.

[4] Frank Biocca, Chad Harms, and Jenn Gregg. 2001. The networked minds measure of social presence: Pilot test of the factor structure and concurrent validity. In *4th annual international workshop on presence, Philadelphia, PA*. 1–9.

[5] Elina Birmingham, Walter F Bischof, and Alan Kingstone. 2008a. Gaze selection in complex social scenes. *Visual Cognition* 16, 2-3 (2008), 341–355.

[6] Elina Birmingham, Walter F Bischof, and Alan Kingstone. 2008b. Social attention and real-world scenes: The roles of action, competition and social content. *The Quarterly Journal of Experimental Psychology* 61, 7 (2008), 986–998.

[7] Jeremy Birnholtz, Lindsay Reynolds, Eli Luxenberg, Carl Gutwin, and Maryam Mustafa. 2010. Awareness beyond the desktop: exploring attention and distraction with a projected peripheral-vision display. In *Proceedings of Graphics Interface 2010*. Canadian Information Processing Society, 55–62.

[8] Ingar Brinck. 2001. Attention and the evolution of intentional communication. *Pragmatics & Cognition* 9, 2 (2001), 259–277.

[9] Allison Bruce, Illah Nourbakhsh, and Reid Simmons. 2002. The role of expressiveness and attention in human-robot interaction. In *Robotics and Automation, 2002. Proceedings. ICRA'02. IEEE International Conference on*, Vol. 4. IEEE, 4138–4142.

[10] Herbert H Clark and Deanna Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition* 22, 1 (1986), 1–39.

[11] Juliet Corbin and Anselm Strauss. 1990. Grounded theory research: Procedures, canons and evaluative criteria. *Zeitschrift für Soziologie* 19, 6 (1990), 418–427.

[12] Guoqiang Cui, Barbara Lockee, and Cuiqing Meng. 2013. Building modern online social presence: A review of social presence theory and its instructional design implications for future trends. *Education and information technologies* 18, 4 (2013), 661–685.

[13] Richard L Daft and Robert H Lengel. 1986. Organizational information requirements, media richness and structural design. *Management science* 32, 5 (1986), 554–571.

[14] Madeleine Dean. 2019. Skype now supports Speaker View in group calls. (3 2019). Retrieved September 07, 2019 from `https://windowsreport.com/skype-speaker-view-group-calls/`.

[15] Gwyneth Doherty-Sneddon and Fiona G Phelps. 2005. Gaze aversion: A response to cognitive or social difficulty? *Memory & cognition* 33, 4 (2005), 727–733.

[16] John V Draper, David B Kaber, and John M Usher. 1998. Telepresence. *Human factors* 40, 3 (1998), 354–375.

[17] Mica R Endsley. 1995. Measurement of situation awareness in dynamic systems. *Human factors* 37, 1 (1995), 65–84.

[18] Florian Eyben, Felix Weninger, Lucas Paletta, and Björn W Schuller. 2013. The acoustics of eye contact: detecting visual attention from conversational audio cues. In *Proceedings of the 6th workshop on Eye gaze in intelligent human machine interaction: gaze in multimodal interaction*. ACM, 7–12.

[19] Kathleen E Finn, Abigail J Sellen, and Sylvia B Wilbur. 1997. *Video-mediated communication*. L. Erlbaum Associates Inc.

[20] Megan Freeth, Tom Foulsham, and Alan Kingstone. 2013. What affects social attention? Social presence, eye contact and autistic traits. *PloS one* 8, 1 (2013), e53286.

[21] Alexandra Frischen, Andrew P Bayliss, and Steven P Tipper. 2007. Gaze cueing of attention: visual attention, social cognition, and individual differences. *Psychological bulletin* 133, 4 (2007), 694.

[22] Samratul Fuady, Masato Orishige, Haoyan Li, Hironori Mitake, and Shoichi Hasegawa. 2016. Natural Interaction in Asymmetric Teleconference using Stuffed-toy Avatar Robot.. In *ICAT-EGVE*. 93–98.

[23] William W Gaver, Abigail Sellen, Christian Heath, and Paul Luff. 1993. One is not enough: Multiple views in a media space. In *Proceedings of the INTERACT'93 and CHI'93 Conference on Human Factors in Computing Systems*. ACM, 335–341.

[24] Darren Gergle, Robert E Kraut, and Susan R Fussell. 2013. Using visual information for grounding and awareness in collaborative tasks. *Human–Computer Interaction* 28, 1 (2013), 1–39.

[25] Arthur M Glenberg, Jennifer L Schroeder, and David A Robertson. 1998. Averting the gaze disengages the environment and facilitates remembering. *Memory & cognition* 26, 4 (1998), 651–658.

[26] Charles Goodwin. 2000. Action and embodiment within situated human interaction. *Journal of pragmatics* 32, 10 (2000), 1489–1522.

[27] Charles Goodwin. 2003. Pointing as situated practice. In *Pointing*. Psychology Press, 225–250.

[28] David M Grayson and Andrew F Monk. 2003. Are you looking at me? Eye contact and desktop video conferencing. *ACM Transactions on Computer-Human Interaction (TOCHI)* 10, 3 (2003), 221–243.

[29] Carl Gutwin and Saul Greenberg. 2002. A descriptive framework of workspace awareness for real-time groupware. *Computer Supported Cooperative Work (CSCW)* 11, 3-4 (2002), 411–446.

[30] Joy E Hanna and Susan E Brennan. 2007. Speakers' eye gaze disambiguates referring expressions early during face-to-face conversation. *Journal of Memory and Language* 57, 4 (2007), 596–615.

[31] Steve Harrison. 2009. A brief history of media space research and mediated life. In *Media Space 20+ Years of Mediated Life*. Springer, 9–16.

[32] Christian Heath and Paul Luff. 1991. Disembodied conduct: communication through video in a multi-media office environment. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 99–103.

[33] Christian Heath, Marcus Sanchez Svensson, Jon Hindmarsh, Paul Luff, and Dirk Vom Lehn. 2002. Configuring awareness. *Computer Supported Cooperative Work (CSCW)* 11, 3-4 (2002), 317–347.

[34] Jim Hollan and Scott Stornetta. 1992. Beyond being there. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 119–125.

[35] Eric Horvitz, Edward Cutrell, and Mary Czerwinski. 2001. Notification, disruption, and memory: Effects of messaging interruptions on memory and performance. In *Proceedings of INTERACT*. 263.

[36] Eric Horvitz, Carl Kadie, Tim Paek, and David Hovel. 2003. Models of attention in computing and communication: from principles to applications. *Commun. ACM* 46, 3 (2003), 52–59.

[37] Eric Horvitz and Tim Paek. 2001. Harnessing models of users' goals to mediate clarification dialog in spoken language systems. In *International Conference on User Modeling*. Springer, 3–13.

[38] Bernardo A Huberman and Fang Wu. 2007. The economics of attention: maximizing user value in information-rich environments. In *Proceedings of the 1st international workshop on Data mining and audience intelligence for advertising*. ACM, 16–20.

[39] Rodney H Jones. 2005. 14 Sites of engagement as sites of attention: Time, space and culture in electronic discourse. (2005).

[40] Victor Kaptelinin and Bonnie A Nardi. 2006. *Acting with technology: Activity theory and interaction design*. MIT press.

[41] Annica Kristoffersson, Silvia Coradeschi, and Amy Loutfi. 2013. A review of mobile robotic telepresence. *Advances in Human-Computer Interaction* 2013 (2013), 3.

[42] Aleksei N Leont'ev. 1974. The problem of activity in psychology. *Soviet psychology* 13, 2 (1974), 4–33.

[43] Paul Luff, Christian Heath, Hideaki Kuzuoka, Jon Hindmarsh, Keiichi Yamazaki, and Shinya Oyama. 2003. Fractured ecologies: creating environments for collaboration. *Human-Computer Interaction* 18, 1 (2003), 51–84.

[44] Albert Mehrabian and others. 1971. *Silent messages*. Vol. 8. Wadsworth Belmont, CA.

[45] AJung Moon, Daniel M Troniak, Brian Gleeson, Matthew KXJ Pan, Minhua Zheng, Benjamin A Blumer, Karon MacLean, and Elizabeth A Croft. 2014. Meet me where i'm gazing: how shared attention gaze affects human-robot handover timing. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*. ACM, 334–341.

[46] Peter Mundy and Lisa Newell. 2007. Attention, joint attention, and social cognition. *Current directions in psychological science* 16, 5 (2007), 269–274.

[47] Brid O'Conaill, Steve Whittaker, and Sylvia Wilbur. 1993. Conversations over video conferences: An evaluation of the spoken aspects of video-mediated communication. *Human-computer interaction* 8, 4 (1993), 389–428.

[48] Mai Otsuki, Taiki Kawano, Keita Maruyama, Hideaki Kuzuoka, and Yusuke Suzuki. 2017. ThirdEye: Simple Add-on Display to Represent Remote Participant's Gaze Direction in Video Communication. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 5307–5312.

[49] Mai Otsuki, Keita Maruyama, Hideaki Kuzuoka, and Yusuke Suzuki. 2018. Effects of Enhanced Gaze Presentation on Gaze Leading in Remote Collaborative Physical Tasks. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 368.

[50] Ye Pan and Anthony Steed. 2012. Preserving gaze direction in teleconferencing using a camera array and a spherical display. In *2012 3DTV-Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON)*. IEEE, 1–4.

[51] Rosalind W Picard. 1997. Affective computing. *The MIT Press, Cambridge (MA)* 167 (1997), 170.

[52] Aina Puce and Bennett I Bertenthal. 2015. *The Many Faces of Social Attention: Behavioral and Neural Measures*. Springer.

[53] Irene Rae, Gina Venolia, John C Tang, and David Molnar. 2015. A framework for understanding and designing telepresence. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*. ACM, 1552–1566.

[54] Daniel C Richardson, Rick Dale, and John M Tomlinson. 2009. Conversation, gaze coordination, and beliefs about visual context. *Cognitive Science* 33, 8 (2009), 1468–1482.

[55] Sean Rintel. 2018. Mobile Sharing and Companion Experiences for Microsoft Teams Meetings. (3 2018). Retrieved September 07, 2019 from https://www.microsoft.com/en-us/garage/wall-of-fame/companionexperiences/.

[56] Evan F Risko, Daniel C Richardson, and Alan Kingstone. 2016. Breaking the fourth wall of cognitive science: Real-world social attention and the dual function of gaze. *Current Directions in Psychological Science* 25, 1 (2016), 70–74.

[57] Ben Robins, Paul Dickerson, Penny Stribling, and Kerstin Dautenhahn. 2004. Robot-mediated joint attention in children with autism: A case study in robot-human interaction. *Interaction studies* 5, 2 (2004), 161–198.

[58] Hanan Salam and Mohamed Chetouani. 2015. A multi-level context-based modeling of engagement in human-robot interaction. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, Vol. 3. IEEE, 1–6.

[59] Brenda Salley and John Colombo. 2016. Conceptualizing social attention in developmental research. *Social Development* 25, 4 (2016), 687–703.

[60] Samiha Samrose, Ru Zhao, Jeffery White, Vivian Li, Luis Nova, Yichen Lu, Mohammad Rafayet Ali, and Mohammed Ehsan Hoque. 2018. CoCo: Collaboration Coach for Understanding Team Dynamics during Video Conferencing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 4 (2018), 160.

[61] Ted Selker, Andrea Lockerd, and Jorge Martinez. 2001. Eye-R, a glasses-mounted eye motion detection interface. In *CHI'01 extended abstracts on Human factors in computing systems*. ACM, 179–180.

[62] Abigail J Sellen. 1995. Remote conversations: The effects of mediating talk with technology. *Human-computer interaction* 10, 4 (1995), 401–444.

[63] Samira Sheikhi, Dinesh Babu Jayagopi, Vasil Khalidov, and Jean-Marc Odobez. 2013. Context aware addressee estimation for human robot interaction. In *Proceedings of the 6th workshop on Eye gaze in intelligent human machine interaction: gaze in multimodal interaction*. ACM, 1–6.

[64] John Short, Ederyn Williams, and Bruce Christie. 1976. The social psychology of telecommunications. (1976).

[65] Garriy Shteynberg. 2014. A social host in the machine? The case of group attention. *Journal of Applied Research in Memory and Cognition* 3, 4 (2014), 307–311.

[66] Garriy Shteynberg. 2015. Shared attention. *Perspectives on Psychological Science* 10, 5 (2015), 579–590.

[67] Maria Staudte and Matthew W Crocker. 2009. Visual attention in spoken human-robot interaction. In *Human-Robot Interaction (HRI), 2009 4th ACM/IEEE International Conference on*. IEEE, 77–84.

[68] Rainer Stiefelhagen, Jie Yang, and Alex Waibel. 2001. Estimating focus of attention based on gaze and sound. In *Proceedings of the 2001 workshop on Perceptive user interfaces*. ACM, 1–9.

[69] Roel Vertegaal. 1999. The GAZE groupware system: mediating joint attention in multiparty communication and collaboration. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. ACM, 294–301.

[70] Roel Vertegaal, Jeffrey S Shell, Daniel Chen, and Aadil Mamuji. 2006. Designing for augmented attention: Towards a framework for attentive user interfaces. *Computers in Human Behavior* 22, 4 (2006), 771–789.

[71] Robert S Weiss. 1995. *Learning from strangers: The art and method of qualitative interview studies*. Simon and Schuster.

[72] Christopher D Wickens, Justin G Hollands, Simon Banbury, and Raja Parasuraman. 2015. *Engineering psychology and human performance*. Psychology Press.

[73] Bin Xu, Jason Ellis, and Thomas Erickson. 2017. Attention from Afar: Simulating the Gazes of Remote Participants in Hybrid Meetings. In *Proceedings of the 2017 Conference on Designing Interactive Systems*. ACM, 101–113.

[74] Lisa C Yamagata-Lynch. 2010. Understanding cultural historical activity theory. In *Activity systems analysis methods*. Springer, 13–26.

[75] Jie Yang, Leejay Wu, and Alex Waibel. 1996. *Focus of attention in video conferencing*. Technical Report. CARNEGIE-MELLON UNIV PITTSBURGH PA SCHOOL OF COMPUTER SCIENCE.

[76] Weilie Yi and Dana Ballard. 2009. Recognizing behavior in hand-eye coordination patterns. *International Journal of Humanoid Robotics* 6, 03 (2009), 337–359.

[77] Cha Zhang, Qin Cai, Philip A Chou, Zhengyou Zhang, and Ricardo Martin-Brualla. 2013. Viewport: A distributed, immersive teleconferencing system with infrared dot pattern. *IEEE MultiMedia* 20, 1 (2013), 17–27.