Meeting in the Middle: The Interpretation Gap Between People and Machines

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

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

ABSTRACT

Effectively bridging the fields of HCI and AI requires operationalizing what human users treat as meaningful in the stream of environmental and content information. Research has yet to systematically address the significant gap between levels of granularity and interpretation of machine labels and of human comprehension. To illustrate the problem, we provide some preliminary results from our study on using machine vision to make work meetings more inclusive, particularly for visually impaired participants.

KEYWORDS

Machine vision, social attention, data labels, interpretation, user description

CHI'19 Extended Abstracts, May 4-9, 2019, Glasgow, Scotland, UK.

ACM ISBN 978-1-4503-5971-9/19/05.
DOI: https://doi.org/10.1145/3290607.XXXXXXX

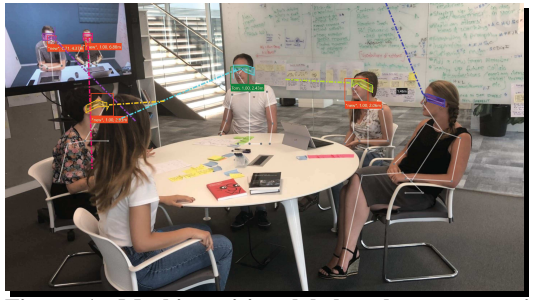

Figure 1: Machine vision labels relevant to social dynamics: people, poses, gaze direction.

INTRODUCTION

With this work we would like to bring to attention the problem of the gap between the degree of atomization of labels produced by machine recognition and the complexity of system output required for human comprehension. In other words, the output viewed by human users as elemental and objective is, in fact, often highly interpretive and requires several levels of algorithmic accumulation of basic recognized events. This gap becomes especially important when input and output are presented in different modalities, for example, when input for the system is visual and output for the user is auditory. For example, consider an image and a machine-produced label of a pose, which is an outline of a skeleton (see Fig.1); to verbally output the pose to a human user, the system would have to at least name the pose (i.e. "sitting", "standing", etc.), which is already an additional level of abstraction and labeling. Examples of commonly used applications that involve modality switch include voice-based virtual assistants and assistive technology for accessibility.

We illustrate the problem of the complexity gap between the machine and human basic event identification by providing some preliminary results from our study on using machine vision to make work meetings more inclusive, particularly for visually impaired participants. Recent advances in machine vision offer exciting opportunities to augment mediated situational awareness for users with restricted perceptional abilities by detecting visual social attention cues that might otherwise be missed. The environment is an extremely rich source of information, but not all this information is equally important. Selectively attending to the information filtered as relevant helps communicators to establish shared knowledge [1, 4] and avoid cognitive overload [2]. One of the cognitive behavioral mechanisms supporting and informing such prioritization involves directing more cognitive recourses to targets of shared attention [4]. We wanted to explore the possibilities for automatic recognition of relevant social attention in a stream of visual information. While basic visual information, such as pose and gaze direction (see Fig.1), can be easily recognized by machine vision, social attention is above all a progressive process of selection of cues that lies as much in human sense-making abilities as it does in their physical abilities. Thus, to teach a machine to recognize and properly output meaningful information about social dynamics, we needed to know how people prioritize and interpret this information.

Study I: How visually-impaired people experience social dynamics in meetings

Our study focused on the potential of automatic augmentation of experiences of visually impaired participants during work meetings. To understand the information needs and expectations of our potential users, we first conducted semi-structured interviews with ten visually-impaired employees from diverse types of organizations. We explored their practices of participating in work meetings in different settings, difficulties and strategies for picking up social information in the environment, and what additional information they would potentially find useful. Our findings suggest that a lot of visual signals, such as gestures and body pose, to derive the information of interest *are* potentially recognizable by machine vision. However, the degree of selectiveness and interpretation of this

"*It would be really helpful to know what people's body language was saying. So you know, bored, or engaged, or distrusted, or stressed.***"** [VIP2]

"*My preferences would be to know what others are actually doing during the meeting. Are they being aware of what I'm saying, are they attentive, do they show interest.***"** [VIP5]

"*I think anything actually that gave you evidence of body language might be useful. … And I think it would be very useful for someone like me to get a better knowledge of how people are visually expressing themselves rather than verbally expressing themselves***"** [VIP7]

"*I'd like to know facial expressions and gestures. [If they are] confused and concerned about the situation… If someone is nodding encouragement to know if you have support… And I think "someone is taking notes", "someone is reading notes", "someone is looking at the window" is fine.***"** [VIP4]

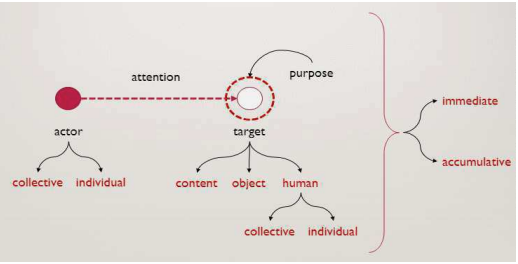

Figure 2: Structural model of human-produced descriptions of social attention.

information, expected by participants, opens further significant challenges, that are considerably more complex to address (See VIP2 and VIP 5 comments in the sidebar to the left). It should be noted, that the information of interest, articulated by our participants, was generally perceived as observable and elemental as opposed to interpretive; participants would commonly specify that they would prefer to receive this "objective" information (See VIP7 and VIP4 comments in the sidebar to the left) to make interpretations on their own.

Study II: How people narrate the social attention cues of a meeting

Given the high level of abstraction and interpretability of expected output that we received in the interview study, we were then interested in modelling the mechanisms of identifying 'objective' observable social information in human-produced descriptions. To collect such descriptions we conducted a quasi-experimental study in which 15 participants were presented with six ~2 min video records of work meetings with 4 to 6 meeting participants. Video clips were presented with muted audio and participants were asked to narrate the visual information on the video to an (imaginary) visually impaired person with focus on social dynamics. We chose to exclude the audio context to ensure participants' focus on visual cues such as body language and facial expressions.

We found that while participants' descriptions were granular enough for us to develop a structural model of social attention description (see Fig.2), the elements of descriptions still often had an accumulative and interpretive nature. For example, descriptions identified an **actor** element as a source of a 'attention action' (the 'who' of an attention description), differentiated as individual and collective. An **individual actor** designates a single person as a source of attention action (*"one of the people on the right is looking at their computer" [P1V1]*) whereas a **collective actor** designated a combined source of attention action that can be described using either plural or singular nouns or pronouns (*"everybody looks at him" [P2V3]*). Another key element was the **target** element ('what' attention is directed to in an attention description). We distinguished several types of targets – content, object, and human. As with an actor element, a human target may be individual or collective, and there are significant differences in the description of dynamic recognition process for these two types. Consider the following: the individual target (*"Now there is **a guy on the left** who is talking and people's attention is on **him**" [P1V3]*) might be recognized from a single frame, a static image; however, the collective target (*"The lady is sitting and listening to **both of them**" [P13V1]*) requires a combination of attention acts with individual human targets formed into a single recognized attention action event. Furthermore, we found an **attention identifier** element, signifying the act of attention, described using a large variety of interpretive verbs, including "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". These verbs interpretation action well beyond the purely visual information from which they were derived.

The model we created is driven by two motivations: the needs of machine perception engineering and the specifics of an output to a user when the target is recognized. While recognition of accumulative acts and collective targets presents numerous challenges, they play a significant role in reducing the

user's cognitive overload when presenting output information. Thus, it is important that research continues to explore further the mechanisms of accumulative descriptions and formulate them as machine recognition rules.

CONCLUSION

We are still at the very early stages of understanding how to identify the appropriate degree of interpretability to translate machine recognition into output comprehensible for human users [3]. We believe that this question should be addressed through active collaboration of AI and HCI researchers. Rapidly growing capabilities of machine recognition require extensive work from HCI to inform AI development by articulating the systematic requirements necessary for applying their advances within human-oriented technology. To conclude, we strongly encourage specialists from both fields to start considering the degrees of minimal required interpretability in AI systems to advance the possibilities of systems applications.

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] 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.

[3] Cecily Morrison, Kit Huckvale, Bob Corish, Richard Banks, Martin Grayson, Jonas Dorn, Abigail Sellen, and Sân Lindley. "Visualizing Ubiquitously Sensed Measures of Motor Ability in Multiple Sclerosis: Reflections on Communicating Machine Learning in Practice." ACM Transactions on Interactive Intelligent Systems (TiiS) 8, no. 2 (2018): 12.

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