Fusing Dialogue and Gaze From Discussions of 2D and 3D Scenes

Regina Wang
Mills College
Oakland, California
rwang@alumnae.mills.com

Bradley Olson
University of Puget Sound
Tacoma, Washington
baolson@pugetsound.edu

Preethi Vaidyanathan
LC Technologies, Inc.
Fairfax, Virginia
pxv1621@rit.edu

Reynold Bailey
Rochester Institute of Technology
Rochester, New York
rjbvcs@rit.edu

Cecilia O. Alm
Rochester Institute of Technology
Rochester, New York
coagla@rit.edu

ABSTRACT

Conversation partners rely on inference using each other's gaze and utterances to negotiate shared meaning. In contrast, dialogue systems still operate mostly with unimodal question or command and response interactions. To realize systems that can intuitively discuss and collaborate with humans, we should consider other sensory information. We begin to address this limitation with an innovative study that acquires, analyzes, and fuses interlocutors' discussion and gaze. Introducing a discussion-based elicitation task, we collect gaze with remote and wearable eye trackers alongside dialogue as

ICMI '19 Adjunct, October 14–18, 2019, Suzhou, China

ACM ISBN 978-1-4503-6937-4/19/10.
https://doi.org/10.1145/3351529.3360661

Discussion Questions

Q1: Which 3 items would you choose to take with you to a desert island and why?
Q3: Which item do you think is the easiest to draw?
Q5: If you were to make a modern art sculpture out of 3 of these items, which would you choose?
Q9: If you were to tidy up this space, how would you re-organize the items?
Q10: How would you use one of these items in an innovative way?

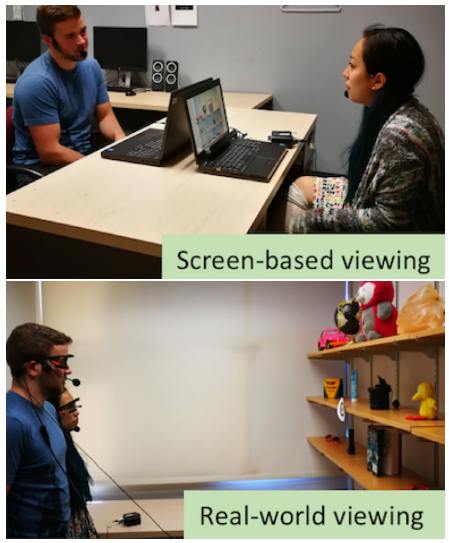

Figure 1: Data collection setup.

interlocutors come to consensus on questions about an on-screen 2D image and a real-world 3D scene. We analyze the visual-linguistic patterns, and also map the modalities onto the visual environment by extending a multimodal image region annotation framework using statistical machine translation for multimodal fusion, applying three ways of fusing speakers' gaze and discussion.

CCS CONCEPTS

• **Computing methodologies** → **Visual inspection**; • **Human-centered computing** → *Collaborative interaction.*

KEYWORDS

multimodal fusion, eye movements, spoken discussion, 2D and 3D scenes, gaze, dialogue

ACM Reference Format:
Regina Wang, Bradley Olson, Preethi Vaidyanathan, Reynold Bailey, and Cecilia O. Alm. 2019. Fusing Dialogue and Gaze From Discussions of 2D and 3D Scenes. In *Adjunct of the 2019 International Conference on Multimodal Interaction (ICMI '19 Adjunct), October 14–18, 2019, Suzhou, China.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3351529.3360661

INTRODUCTION

To truly enable human-machine collaboration, systems with team-centered inference capabilities must go beyond unimodal, language-based interaction and incorporate other sensing data, such as gaze [12, 15, 20]. We envision a human-centered system capable of complex joint reasoning in visual environments. Our research questions are:

RQ1: How do we devise data collection for pair-based visual-linguistic data in a reasoning scenario?
RQ2: How do questions about an image and viewing conditions (2D vs. 3D) impact gaze and dialogue?
RQ3: Can we adapt a monologue and description based image annotation framework for non-descriptive dialogue in visual environments?

Vaidyanathan et al.'s framework uses statistical machine translation to meaningfully map gaze and descriptive monologue. This framework has been applied to obtain image-region annotations for dermatology images [22], images with affective content [3, 6], and open-domain images [23]. In contrast we investigate fusing gaze and dialogue.

Many studies have examined the relationship between gaze and spoken language and found that they were tightly linked [4, 5, 7, 16, 17, 25]. Researchers have observed that paired gaze was highly connected in discussions of a shared image [18] and that gaze plays a role in conversational turn management [2, 11]. Recently, Kontogiorgos et al. presented a multimodal corpus to investigate the distribution of gaze prior to referring expressions between speakers and listeners [12]. Matsuda et al.

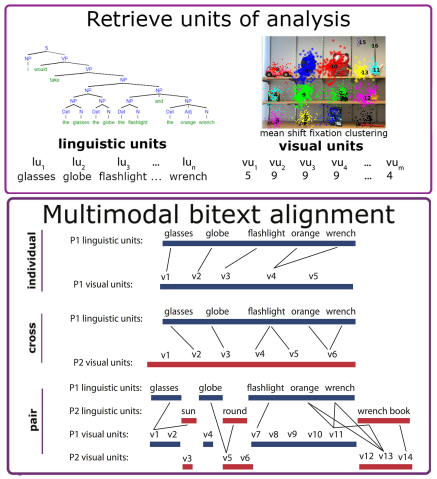

Figure 2: Visual and linguistic units are aligned with 3 fusion methods.

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
3D	6, 12	7, 6	10, 6	6, 7	7, 6	3, 1	7, 5	7, 12	7, 6	6, 3
2D	6, 12	3, 7	10, 6	7, 6	3, 7	7, 3	7, 9	3, 4	7, 6	3, 7

Figure 3: *Kitchen* scene AOIs (top) and top 2 gazed AOIs per question (bottom) indicated similar behaviors for 2D and 3D conditions. Four questions (green) had the same top 2 AOIs for 2D and 3D. Other questions had at least one match (blue). Only Q8 had no matches (red).

estimated user satisfaction in a tour by considering dialogue, facial expression and heartbeat [15]. These studies motivate our work in fusing interlocutors' gaze and speech.

DATA COLLECTION AND MULTIMODAL FUSION

Data were elicited from 32 fluent English speakers in 16 pairs. The experiment involved two scenes (*household* and *kitchen*), each with 12 items on shelves. All viewed one scene in 2D and the other in 3D, with distribution balanced. Subjects were instructed to discuss and come to consensus on 10 questions (5 are shown on page 2, left), asked in both conditions with order randomized. In the 2D condition, gaze was collected with remote SensoMotoric Instruments (SMI) RED250 eye trackers [10] as an image was viewed (Figure 1, top). A question was displayed on the screens, followed by the image as the pair discussed. In the 3D condition, subjects wearing Pupil Labs eye trackers [13] stood and viewed the scene (Figure 1, bottom). Questions were displayed on a nearby monitor. We used iMotions [9] to map the 3D gaze data onto a still image. For both conditions, the discussions were recorded. Two pairs' 3D *household* scene data was excluded from analysis due to data loss.

The audio recordings were transcribed with IBM Watson Speech to Text [8]. Nouns and adjectives were extracted from the transcripts following previously established protocol [21]. Words for advancing the experiment (*next*) were stop-listed along with any word uttered by fewer than three participants. Fixations of all subjects for a scene and eye tracker were combined and clustered with mean-shift clustering, which is effective in identifying regions of interest from fixations [19]. Fixations were encoded by cluster and data augmentation was performed as described by Vaidyanathan [21]. The resulting ordered sequences of words and encoded fixations form the *linguistic* and *visual units*, respectively (Figure 2, top). We align *linguistic* and *visual units* by varying their representation in three ways (Figure 2, bottom): (1) *individual* connects a dialogue partner's speech with their own gaze, (2) *cross* integrates a partner's speech with the other's gaze, while (3) *pair* concatenates both of their gaze and dialogue streams based on who is speaking for an interval. Since people do not look at objects at the same time as they mention its name [21, 22], we applied the Berkeley aligner which is based on statistical machine translation and treats the time-ordered linguistic and visual units as a parallel corpus of multimodal bitext [14]. The results are presented as an annotated image (Figure 6).

RESULTS AND DISCUSSION

Gaze Analysis: Gaze patterns changed with question, as previously observed [24]. Pair scanpaths often shared similar areas of focus to each other, while these areas varied between questions. Average fixation counts were generally lower for the last questions asked than for the first. Figure 3 shows the top 2 gazed areas of interest (AOIs) per question in the kitchen scene, and indicate similar gaze behaviors for 2D and 3D. Q1, Q3, Q4, and Q9 had the same top 2 AOIs for 3D and 2D. All other questions except Q8 shared at least one top AOI match. We applied the recurrence quantification

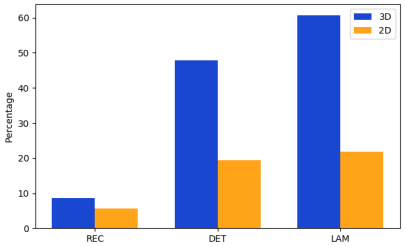

Figure 4: RQA measures are higher for 3D.

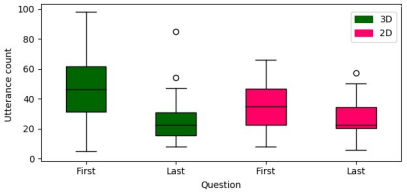

Figure 5: Elicited utterances decrease from first to last question.

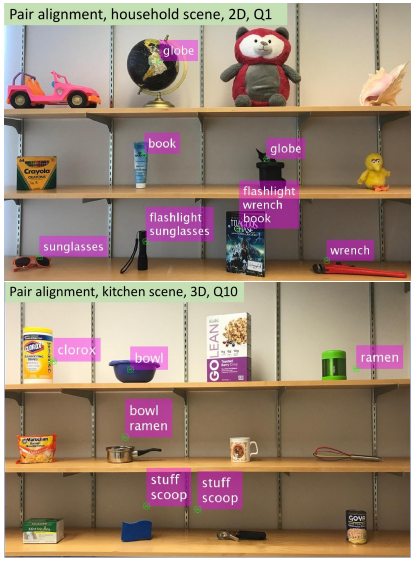

Figure 6: Each token matches its region (can maximally match one region).

analysis (RQA) on the gaze data [1]. 2D gaze data showed significantly ($p<0.05$) lower values of RQA measures compared to 3D (Figure 4). To rule out influence from eye tracker type, we analyzed the average number of fixations in both and found no significant difference. Jointly, these observations indicate that viewers repeated fixation patterns more often when viewing 3D scenes than 2D images.

Spoken Language Analysis: Average word token count and lexicon used varied by question as expected. Q3 frequently produced succinct conversations with few tokens as pairs quickly arrived at a consensus, whereas Q1 produced longer conversations. Similar trends were observed in discussion length, word type counts, and utterance counts but type/token ratio did not vary much. Welch's t-tests on questions' token counts between 2D and 3D scenes showed no significant differences. Similar to fixation counts, utterance counts tended to decrease for last questions. (Figure 5).

Multimodal Alignment Analysis: Figure 6 shows annotated images using the *pair* method for generating the bitext input. Each image shows the 5 most frequent linguistic units for that case with their top-2 most confidently aligned visual units. The *pair* approach for fusion, which leverages dialogue-based two-party reasoning the most, improves the region annotation over *individual* and *cross* approaches. In the top image from Q1, all five words label their objects. In the bottom image, *clorox*, as well as *bowl* and *scoop* label their objects, while *ramen* is associated with a pot and a thermos that functionally may be used to prepare or store the noodles. These examples demonstrate that the multimodal alignments capture framing provided by a scene or question. Tokens are fused with or around relevant, corresponding items even with modest quantities of dialogue data.

CONCLUSION

To address RQ1, we introduced an elicitation task for capturing multimodal discussion from two speakers-observers. Analyses confirm that question or scene frame the task to generate rich multimodal data, verifying our elicitation method. For RQ2, we found that viewing in 2D and 3D elicit similar data, though RQA measures indicated more frequently repeated fixation patterns in 3D. We also found that repeated viewing impacted the elicited amount of gaze and dialogue. To answer RQ3, we introduce an extension to multimodal alignment that goes beyond prior descriptive image region annotation to annotate images based on discussion-based dialogue and gaze, where questions frame how a scene is viewed and discussed. The *pair* method enhances image region annotation compared to other fusion methods. Future work would benefit from user evaluation of visual-linguistic annotated images and connective information found in verbs or prepositions to identify semantic links between objects.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Award No. IIS-1559889. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] Nicola C. Anderson, Walter F. Bischof, Kaitlin E.W. Laidlaw, Evan F. Risko, and Alan Kingstone. 2013. Recurrence quantification analysis of eye movements. *Behavior Research Methods* 45 (2013), 842–856.

[2] Dan Bohus and Eric Horvitz. 2010. Facilitating multiparty dialog with gaze, gesture, and speech. In *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*.

[3] Aliya Gangji, Trevor Walden, Preethi Vaidyanathan, Emily Prud'hommeaux, Reynold Bailey, and Cecilia O Alm. 2017. Using co-captured face, gaze, and verbal reactions to images of varying emotional content for analysis and semantic alignment. In *Proceedings of the AAAI Workshop on Human-Aware Artificial Intelligence*.

[4] Zenzi M. Griffin. 2004. Why look? Reasons for eye movements related to language production. In *The Interface of Language, Vision, and Action: Eye Movements and the Visual World*. Psychology Press, 213–248.

[5] Zenzi M. Griffin and Kathryn Bock. 2000. What the eyes say about speaking. *Psychological Science* 11, 4 (2000), 274–279.

[6] Nikita Haduong, David Nester, Preethi Vaidyanathan, Emily Prud'hommeaux, Reynold Bailey, and Cecilia Alm. 2018. Multimodal Alignment for Affective Content. In *Proceedings of the AAAI Workshop on Affective Content Analysis*.

[7] Jana Holsanova. 2008. *Discourse, Vision, and Cognition*. John Benjamins Publishing Company.

[8] IBM. 2019. Watson Text to Speech. https://www.ibm.com/watson/services/text-to-speech/

[9] iMotions. 2019. https://imotions.com/

[10] SensoMotoric Instruments. 2019. https://www.smivision.com/eye-tracking/products/software-for-eye-tracking/

[11] Kristiina Jokinen, Kazuaki Harada, Masafumi Nishida, and Seiichi Yamamoto. 2010. Turn-alignment using eye-gaze and speech in conversational interaction. In *Eleventh Annual Conference of the International Speech Communication Association*.

[12] Dimosthenis Kontogiorgos, Vanya Avramova, Simon Alexandersson, Patrik Jonell, Catharine Oertel, Jonas Beskow, Gabriel Skantze, and Joakim Gustafsson. 2018. A multimodal corpus for mutual gaze and joint attention in multiparty situated interaction. In *Language Resources and Evaluation Conference*.

[13] Pupil Labs. 2019. https://pupil-labs.com/

[14] Percy Liang, Ben Taskar, and Dan Klein. 2006. Alignment by agreement. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 104–111.

[15] Yuki Matsuda, Dmitrii Fedotov, Yuta Takahashi, Yutaka Arakawa, Keiichi Yasumoto, and Wolfgang Minker. 2018. Estimating User Satisfaction Impact in Cities using Physical Reaction Sensing and Multimodal Dialogue System. In *International Workshop on Spoken Dialogue Systems Technology*.

[16] Shaolin Qu and Joyce Y. Chai. 2008. Incorporating temporal and semantic information with eye gaze for automatic word acquisition in multimodal conversational systems. In *Proceedings of the ACL Conference on Empirical Methods in Natural Language Processing*. 244–253.

[17] Daniel C. Richardson and Rick Dale. 2005. Looking to understand: The coupling between speakers' and listeners' eye movements and its relationship to discourse comprehension. *Cognitive Science* 29, 6 (2005), 1045–1060.

[18] Daniel C. Richardson, Rick Dale, and Natasha Z. Kirkham. 2007. The art of conversation is coordination. *Psychological Science* 18, 5 (2007), 407–413.

[19] Anthony Santella and Doug DeCarlo. 2004. Robust clustering of eye movement recordings for quantification of visual interest. In *Proceedings of the ACM symposium on Eye tracking research & Applications*. 27–34.

[20] TJ Tsai, Andreas Stolcke, and Malcolm Slaney. 2015. Multimodal addressee detection in multiparty dialogue systems. In *Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing*. 2314–2318.

[21] Preethi Vaidyanathan. 2017. *Visual-Linguistic Semantic Alignment: Fusing Human Gaze and Spoken Narratives for Image Region Annotation*. Ph.D. Dissertation. Rochester Institute of Technology.

[22] Preethi Vaidyanathan, Emily Prud'hommeaux, Jeff B. Pelz, Cecilia Ovesdotter Alm, and Anne R. Haake. 2016. Fusing eye movements and observer narratives for expert-driven image-region annotations. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications*. 27–34.

[23] Preethi Vaidyanathan, Emily T. Prud'hommeaux, Jeff B. Pelz, and Cecilia O. Alm. 2018. SNAG: Spoken Narratives and Gaze Dataset. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, Vol. 2. 132–137.

[24] Alfred Yarbus. 1965. *Role of eye movements in the visual process.* Nauka Press, Moscow.

[25] Kiwon Yun, Yifan Peng, Dimitris Samaras, Gregory J. Zelinsky, and Tamara L. Berg. 2013. Studying Relationships between Human Gaze, Description, and Computer Vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 739–746.