# Dynamic Visualization System for Gaze and Dialogue Data

Jonathan Kvist[1], Philip Ekholm[1], Preethi Vaidyanathan[2], Reynold Bailey[3]
and Cecilia Ovesdotter Alm[3]
[1]*Department of Computer Science and Media Technology, Malmö University, Nordenskiöldsgatan 1, Malmö, Sweden*
[2]*Eyegaze Inc., 10363 Democracy Lane, Fairfax, VA 22030, U.S.A.*
[3]*Rochester Institute of Technology, 1 Lomb Memorial Drive, Rochester, NY 14623, U.S.A.*

Keywords: Eye-tracking, Dialogue, Multimodal Visualization.

Abstract: We report and review a visualization system capable of displaying gaze and speech data elicited from pairs of subjects interacting in a discussion. We elicit such conversation data in our first experiment, where two participants are given the task of reaching a consensus about questions involving images. We validate the system in a second experiment where the purpose is to see if a person could determine which question had elicited a certain visualization. The visualization system allows users to explore reasoning behavior and participation during multimodal dialogue interactions.

## 1 INTRODUCTION

AI systems collaborating with humans should understand their users. Having access to more than just one modality improves this interaction and can be less error-prone (Tsai et al., 2015; Kontogiorgos et al., 2018). However, since human signals are ambiguous, their interpretation is challenging (Carter and Bailey, 2012). By visualizing gaze and speech behaviors of people discussing the visual world we can better understand human reasoning. We present a visualization system incorporating gaze and speech from two interlocutors engaging in dialogue about images.

Visualizing data is helpful in many scenarios, e.g. eye movement data reveals how the user spread their attention (Blascheck et al., 2017). Popular visualizations such as heat maps and word clouds only cover one modality and are often presented as static images. In contrast, we present a multimodal dialogue visualization system that incorporates both gaze and speech data, enhancing the amount of information regarding how interlocutors reasoned. Multimodal dialogues are complex since they have temporal progression. Therefore, our system utilizes dynamic rendering to capture the evolution of gaze and dialogue.

To elicit multimodal data we conducted **Experiment I** where participants were paired and given the task to discuss and reach a verbal consensus on questions about images which were chosen based on *complexity* and *animacy* (see Figure 1). The participants' eye-movement and dialogue were recorded. This was followed by building a prototype visualization system and then evaluating and validating it in **Experiment II** by having participants view multimodal visualizations generated from Experiment I and attempting to identify the underlying questions that generated the visualization. The research questions studied were:

Figure 1: Images used in Experiment I (Top-left: simple inanimate; top-right: complex inanimate; bottom-left: simple animate; bottom-right: complex animate).

- RQ1: Does access to more modalities improve identifying the underlying question that elicited the dialogue shown (i.e. question recognition)?
- RQ2: Do the properties of stimuli, namely complexity, animacy, and question-type impact question recognition?



Kvist, J., Ekholm, P., Vaidyanathan, P., Bailey, R. and Alm, C.
Dynamic Visualization System for Gaze and Dialogue Data.
DOI: 10.5220/0008953201380145
In *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2020)*, pages 138-145
ISBN: 978-989-758-402-2

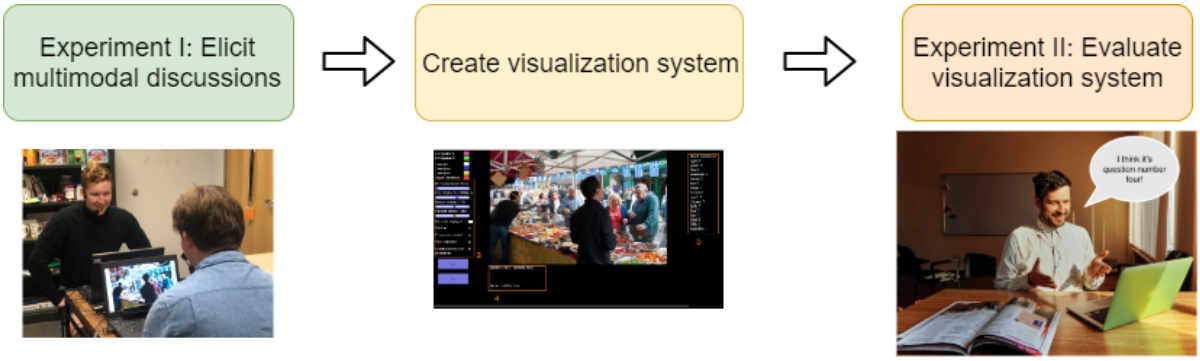


Figure 2: First Experiment I collected speech and gaze data. Pairs sat across each other and in front of eye trackers, wearing microphones to record their speech. Next, we developed a prototype of a multimodal visualization system using the data. Then, Experiment II evaluated the prototype by letting users interact with it and respond to questions.

We use question recognition as an objective measurement of visualization success.

## 2 PRIOR WORK

Eye-tracking is getting better, cheaper, more accessible, and could therefore let us understand more about user behavior and human reasoning. Eye tracking produces large amounts of data, making data analysis a challenge. Its analysis tend to involve quantitative statistical analysis or qualitative visualizations (Blascheck et al., 2014). The statistical techniques can give us information such as fixation count and average number of saccades, while visualizations can promote more holistic meaningful information by highlighting how user attention is distributed over the stimuli. Blascheck et al. (2017) explored how experts in the field analyzed eye tracking data, and they agreed that statistical analysis and exploratory visualizations both are important. Almost all experts used visualizations in their work, especially heat maps and scan paths. The authors suggested that more advanced techniques would be adopted if they were more widespread and straightforward to use.

Eye-tracking visualizations are categorized as either point-based or area-of-interest based (Blascheck et al., 2014). Regardless of the approach used, effective visualization frameworks should ideally allow for easy and customizable user interaction, provide methods to incorporate additional data modalities and be able to handle dynamic content, not just static 2D images (Blascheck et al., 2014; Blascheck et al., 2017; Stellmach et al., 2010; Blascheck et al., 2016; Ramloll et al., 2004). Interestingly, one of the experts in the Blascheck et.al. (2017) study mentioned that a visualization should be analyzed "not just as a means for the researcher to understand the data, but as a specific independent variable in which participants are shown the eye movement patterns of either themselves, or from other observers, and how they can (or cannot) understand and exploit this information." We leverage this idea as a way to assess the effectiveness of our visualization system.

Online collaborative learning tools have the potential to change the way people learn. Visualization of the information exchange can aid in the learning process (Bull and Kay, 2016). To improve such systems, it is important to analyze how users collaborate in them. Traditionally this might be done with a static 2D image after the session, but Koné et al. (2018) emphasize that it will be too late by then. They also point out that the temporal information from the collaboration is lost in such a visualization (Koné et al., 2018). The authors therefore argue for the need of a dynamic, real time visualization that includes the temporal context, e.g. by using animations. Another study reports on an approach that makes scan paths more easily comparable by mapping them as a function of time (Räihä et al., 2005). This is considered advantageous when the exact locations of fixations are less important than certain areas of interest.

Visualizations can be used to find relationships between modalities. A strong relationship has been observed between the dialogue and gaze patterns of two interlocutors (Sharma and Jermann, 2018). Wang et al. (2019) also found that the gaze from two interlocutors are linked in conversation; relationships that our system can highlight. They presented the users with both a 3D and 2D scene, where the results did not differ much (Wang et al., 2019). Our data collection is based on their work's 2D scene setup. During their study, they asked the participants questions which influenced conversation and gaze patterns.

# 3 METHOD

This study included three steps as seen in Figure 2:

1. Experiment I - elicit multi-modal discussions
2. Prototype the multi-modal visualization system
3. Experiment II - evaluate the prototype

## 3.1 Experiment I - Eliciting Multi-modal Discussions

We conducted an IRB-approved data collection experiment with 20 participants, recruited at a university in the USA. The participants were paired up, and their remuneration for participation was 10 USD each. Half of the participants were female and the other half male, and in 8 of the 10 pairs a female participant was teamed with a male participant. Eighteen were 19-23 years old. The other two were 26-30. They first filled out a pre-survey, and were then seated in front of each other at a small table as shown in Figure 3. Each person was equipped with a head-worn microphone to record their speech, and we used a SensoMotoric Instruments RED 250Hz eye-tracker mounted on a laptop to track their gaze. To elicit different conversations and gaze patterns, their task was to discuss aloud and reach verbal consensus on a series of questions (shown in Figure 4) about four images (shown in Figure 1). Questions 1-3 were affective questions regarding the mood, feelings and subjective attitudes, while questions 4-6 were material questions about the subjective characteristics of the objects in the images. We included different types of questions to understand how visualization recognition may be impacted by more or less emotional answers.

Experiment I was performed 10 times with a total of 20 participants, and each experiment lasted for about 60 minutes. Data from two pairs had to be excluded due to misinterpretation of the task. Each pair elicited 24 conversations (4 images * 6 questions), meaning that we in total we got 192 conversations (8 pairs * 24 conversations).

Figure 3: Experiment set-up: Two interlocutors seated across each other with their respective microphones, eye trackers and laptops.

**Experiment I questions**

Q1. What activities does the owner of these objects or the main person in the picture enjoy?

Q2. How would you change the environment to make it more welcoming?

Q3. Describe an artwork that this image inspires you to create.

Q4. Pick an object or person in the picture explain why it does not belong there.

Q5. Which two items belong together in the picture?

Q6. Which non-living object is the oldest and which is the newest?

Figure 4: Experiment I asked three affective questions (Q1-Q3) and three material questions (Q4-Q6) to elicit discussions.

## 3.2 Visualization System

The visualization prototype was built using *Processing*, a Java-based sketchbook for drawing graphics, text, images, and more (Fry and Reas, 2014).

### 3.2.1 Constraints

To enable our system to be fully autonomous, we use Microsoft Azure Automatic Speech to Text (ASR) to transcribe the conversations (Microsoft, 2019). We computed the Word Error Rate (WER) for four conversations. The values ranged between 18-29%. We believe that the WER is high due to two reasons. First, both speakers had wearable microphones to record their speech, but they were seated at the same table which introduced some crosstalk. Second, we cannot verify whether the training data included dialogues as opposed to speech from a single person. The model's predictions may not be as accurate for dialogue data. To counteract this, we included the confidence level that the ASR system assigned to a particular utterance when presenting the transcription to the user. This gives the user information to decide for themselves whether to trust the transcription or not.


Figure 5: Overview of the prototype. **1:** Fixations and words spoken by a participant while looking in the surrounding area. **2:** Fixations and saccades in different colors corresponding to a speaker. **3:** Various settings. **4:** Machine-transcribed subtitles of the dialogue. **5:** Words that are mentioned frequently displayed along with their frequency count.

### 3.2.2 Co-reference Resolution

To extract more meaning from the utterances in the discussion, we applied co-reference resolution. Co-reference resolution is the task of determining if some expressions in the discussion refer back to the same entity, e.g. if *he* and *Daniel* refer to the same person (Soon et al., 2001). We use a pre-trained model, included in the AllenNLP processing platform (Gardner et al., 2018), which performs end-to-end neural co-reference resolution (Lee et al., 2017). This is applied to the text transcribed by Microsoft ASR and the resulting information is used by the *most frequently mentioned words table* in our system, so that an entity is updated with the correct number of mentions.

### 3.2.3 Creation of the Visualization

The prototype system is made up of four major components, which correspond to the different levels of dialogue modalities in the visualization system seen in Table 1: gaze + word tokens (M1), gaze + word tokens + subtitles (M2), gaze + word tokens + subtitles + conversation playback (M3), gaze + word tokens + subtitles + conversation playback + access to all settings (M4). Previous research identified the importance of handling the temporal progression of conversations, meaning that our system had to be dynamic. Different users would potentially also use the system for different tasks. Therefore, we wanted to give the user access to as many settings as possible from the interface. Figure 5 shows a screenshot of the dynamic visualization prototype. The key parts of the system are numbered and explained below:

1. **Fixations** - represented by either a green or pink circle, depending on which interlocutor generated it. The size of the circle is proportional to the fixation duration.
2. **Saccades** - represented by a line between two fixations. As new saccades are displayed, the oldest one present on the screen is removed to make the visualization less cluttered.
3. **Settings** - the user can access the settings of the visualization in the panel on the left side of the system. A legend in the upper-left corner shows the different color codes used. Additionally, the

Table 1: Modalities active for each iteration during Experiment II. The evaluators had access to overlaid gaze and words at all times but subtitles, audio, and the settings were introduced as the experiment progressed. Stopwords were not shown per default.

| Modalities | Gaze | Words | Subtitles | Audio | Settings |
|---|---|---|---|---|---|
| M1 | Yes | Yes | | | |
| M2 | Yes | Yes | Yes | | |
| M3 | Yes | Yes | Yes | Yes | |
| M4 | Yes | Yes | Yes | Yes | Yes |

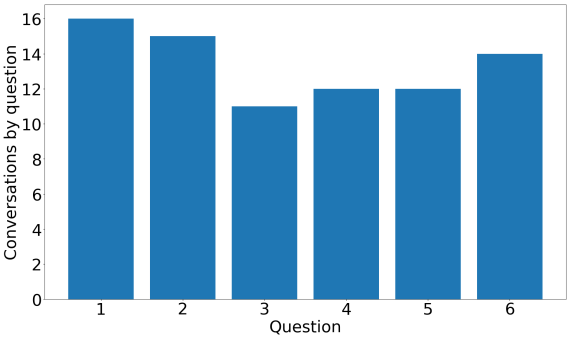


Figure 6: Number of times the different questions appeared in Experiment II.

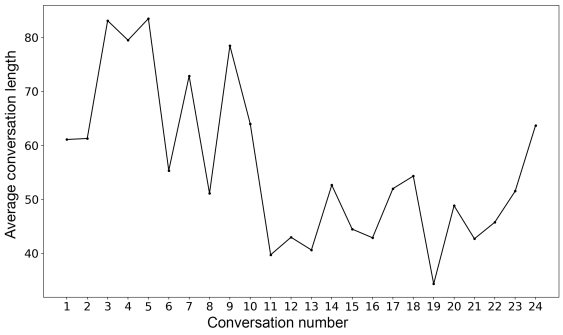


Figure 7: The conversation length decreased as the experiment progressed indicating growing scene familiarity.

following settings can be accessed:

- *Minimum fixation length* (sliding bar)
- *Word display time* (sliding bar)
- *Fixation scaling* (sliding bar)
- *Playback speed* (sliding bar)
- *Saccades displayed* (number input)
- *Subtitles* (toggle)
- *Conversation sound* (toggle)
- *Show stopwords* (toggle) - display and count words that are recognized as stopwords. The Natural Language Toolkit (Bird et al., 2009) was used for filtering them out.
- *Update word count by co-reference* (toggle) - update word count for entities discovered during co-reference resolution.

4. **Subtitles** - subtitles from the conversation in real time as generated by the Microsoft ASR.
5. **Most mentions** - the words that have been mentioned the most in the conversation.

### 3.3 Experiment II - Evaluation of the Multimodal Visualization System

In order to evaluate the effectiveness of our visualization system, we conducted a second experiment which utilized the data elicited from Experiment I. For this experiment, we included ten participants referred to as **evaluators** to distinguish them from the participants of Experiment I. All evaluators were recruited from the same university as in Experiment I. Six were female and four were male, and their remuneration was 10 USD. Nine evaluators were between 19 and 29 years old and one was above 40.

Each evaluator first answered the same demographic pre-survey as in Experiment I. They were then seated in front of a screen that displayed the visualization prototype. They were presented with eight visualizations in total, all with data elicited in Experiment I. We chose the eight visualizations based on three factors: (1) underlying question, (2) image, and (3) pair from Experiment I. Each evaluator was presented with each question once and two additional questions selected at random. Each image appeared twice, and the conversations were chosen from at least seven different pairs from Experiment I. A total of 80 multimodal visualizations were presented to evaluators (10 participants * 8 visualized conversations). Every selected visualization was presented to an evaluator four times. Each time, the access to modalities or functionalities in the interface increased, as can be seen in Table 1. First, the evaluator only had access to the word tokens and the gaze points in the system (M1). Then the subtitles were added (M2), as transcribed by Microsoft ASR. In the third run the evaluator wore earphones and the actual conversation was played back along with the visualization (M3). Lastly, the evaluator got full access to all the settings in the interface (M4). After each introduction of modality or functionality, the evaluator were asked three questions:

1. Which question generated the visualized gaze and spoken language? (Menu with questions)
2. How confident are you about this answer? (Menu with 1-10, where 0 means not confident and 10 means extremely confident)
3. Why did you choose that question? (Long answer text)

For the drop down menu in question 1, the evaluator could choose from the questions in Figure 4. After M1 to M4 had been covered for one visualization, the evaluator was also asked to give a free text answer to the question: *what is your perception about this visualization?*

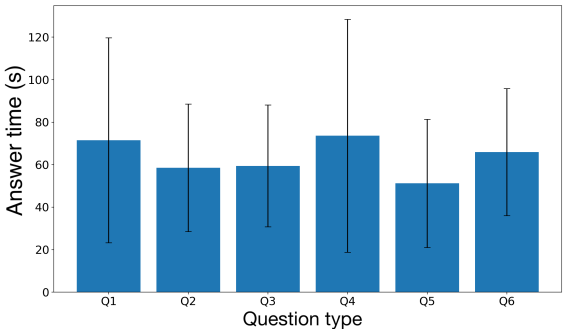


Figure 8: The question type (affective (Q1-Q3) or material (Q4-Q6) all had similar conversation length.

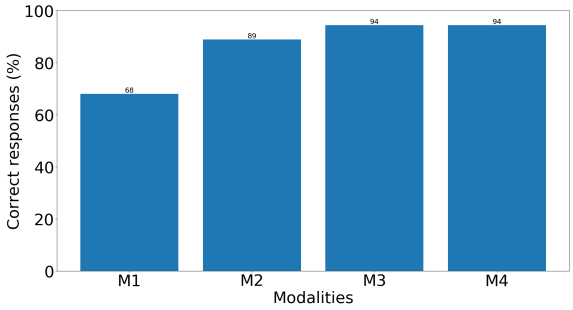


Figure 9: Accuracy, measured as how often the evaluators in Experiment II managed to correctly identify which question elicited the multimodal visualization, increased when providing more forms of data. M1 = Gaze data + word tokens, M2 = M1 + machine-transcribed subtitles, M3 = M2 + dialogue sound, M4 = M3 + ability to customize settings. A one-way ANOVA resulted in a p-value $< 0.001$.

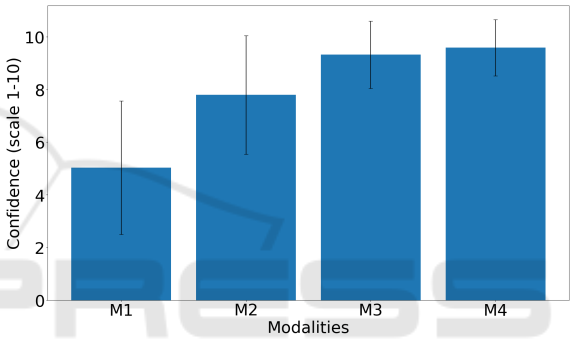


Figure 10: Average self-reported confidence increases as more modalities/functionalities are introduced over M1 through M4. This follows the same pattern as the response accuracy, as can be seen in Figure 9. A one-way ANOVA was used as hypothesis-test with a p-value of 0.04.

## 4 RESULTS AND DISCUSSION

**Conversation Duration:** We analyzed the duration of conversations in Experiment I. Figure 7 shows that average conversation length decreased as the experiment progressed. This could be due to growing scene familiarity. Also, since images and questions were repeated multiple times, fatigue and boredom could also have played a part. Interestingly, the duration of a conversation did not affect the ability of the evaluator in accurately identifying the underlying question, i.e. question recognition. Figure 8 shows that the conversation duration is not affected substantially by the type of question asked.

**Effectiveness of the Visualization:** We received mixed responses from evaluators in Experiment II. One evaluator described the visualization as *very confusing*, whereas another evaluator said *think its clear where circles with lines showing which objects take out*. Many participants mentioned that the extra modalities (subtitles, sound) made it easier to figure out the underlying question. However, they highlighted that one of the more helpful tools was the list of most frequently mentioned words (Figure 5). This list was available throughout M1 to M4. Certain questions have a tendency to generate certain keywords, which likely aided the evaluators in correctly identifying the underlying question. We further note that a majority of the evaluators did not change any settings in M4 even though they could, indicating that the default settings were well calibrated.

**Modalities and Question Recognition (RQ1):** Our results show that as we increase the modality level i.e. provide more information to the evaluator in Experiment II, their ability to recognize the underlying questions increased. This is evident from Figure 9. With only word tokens and gaze data (M1), 68% of the time evaluators were able to correctly recognize the underlying question, which is far above random guess between the six options. When provided with added Microsoft ASR subtitles (M2), a significantly higher share (89%) of the evaluators successfully recognized the question. This could be due to that the subtitles encapsulate syntactic structure and context, which helps the evaluator identify what the conversation was about. Instead of only having access to word tokens and no syntactic structure, the evaluator can now follow the unfolding of the conversation. With added access to the audio of the conversation (M3), 94% of the evaluators successfully recognized the question. Listening to the conversation gave access to prosody and voice inflection, providing the evaluator more information about the conversation that subtitles cannot straightforwardly capture, for instance sarcasm. Also, the conversation audio playback is not reliant on the ASR system to provide the correct output, which is likely important due to its higher WER. The last modality (M4) i.e. access to the various settings did not impact the share of evaluators that correctly recognized the underlying question.

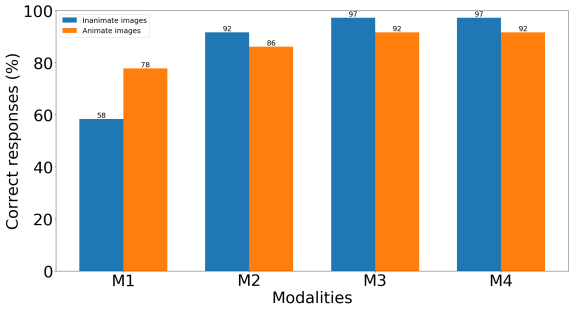


Figure 11: No significant difference in question recognition depending on the animacy of the visual content. A one-way ANOVA resulted in a p-value of 0.88.

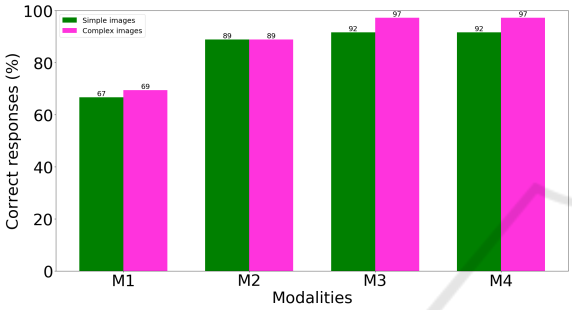


Figure 12: No significant difference in question recognition depending on the visual complexity of the image. A one-way ANOVA resulted in a p-value of 0.46.

This can also be seen in Figure 10, which shows that the overall self-reported confidence level corresponds well with the actual overall question recognition accuracy. We can see that during M3 (gaze, word tokens, and audio), evaluators were almost certain they had the correct answers. This gave them little to no incentive to change any of the settings and improve upon the prior answers. We observed that many evaluators did not change any settings at all.

In summary, our results show that gaze and word tokens were displayed in a manner that was helpful for recognizing the underlying question. However, context around the word tokens, whether provided via subtitles or audio, appear to be a key piece of information in understanding the focal point of a given conversation. This means that a useful visualization benefits from multimodality, but also from structurally sound content. The user is helped by more than just extracted word tokens. It would help to investigate the usefulness of phrases as compared to word tokens and to measure how much of the context around a word is required to understand the conversation.

**Stimulus, Questions, and Question Recognition (RQ2):** We observe that animacy of the visual content might impact the ability to accurately recognize the questions to some degree (Figure 11), however, this is not statistically significant. Similarly, the complexity of the image (Figure 12) or type of question as affective or material (Figure 13) did not seem to affect the ability of the evaluator to correctly recognize the underlying question. This is advantageous as a single framework of the visualization can be used across images or question types.

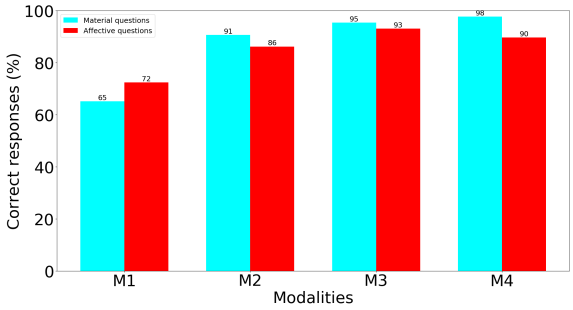


Figure 13: No significant difference in question recognition depending on the type of question. A one-way ANOVA resulted in a p-value of 0.73.

## 5 CONCLUSIONS

This work provides insight into how data from multiple modalities can be jointly displayed in a visualization system for qualitative analysis and research purposes. Our visualization system is dynamic and allows a user to visualize both gaze and speech data over time. We go one step further by providing multiparty gaze and dialogue data visualization which will be highly useful in studies aimed at understanding human-human interaction and behavior. Such a visualization system can also be used for remote mentoring by an instructor. In this work we exemplify this feature by showing the data from a pair of subjects but this can be extended to more than two speakers e.g. a group of people looking at a visual environment and discussing it. We used 2D eye trackers but based on Wang et al's (2019) findings of few differences in multimodal 2D and 3D scene understanding, we do not necessarily anticipate major changes if moving to 3D actual visual environments.

Our results show that access only to gaze and word tokens is helpful in understanding the focal point of a conversation (via question recognition) to some degree. However, access to the context of the language structure significantly increases the users ability to identify the focal point. This means that a multimodal visualization system that simply displays words is not perfect and that transcriptions in the form of subtitles or spoken conversation are useful. The fact that the evaluators found our visualization system (with all modalities) useful, based on their response accuracy,

despite the type of image or question and even image complexity indicates that the visualization system can be used more broadly for understanding dialogues. This is important for face-to-face collaborative applications such as a group of students collaborating on a task in a classroom. In the future we would like to analyze the benefits of providing only gaze information without the most frequent words and compare it to other modalities, isolated and combined.

Our visualization system is flexible and can be expanded to include more features. For example, quantitative metrics such as mean and standard deviation for fixation duration and type-token-ratio can also be added in the future. In addition, if more human generated data is elicited, it could be added into the system as new user features. Currently, the system displays the static image that was used in Experiment I but it can be extended to display a dynamic stimulus or 3D real-world scenes. This is challenging and will be particularly helpful for researchers who want to analyze data from wearable eye trackers.

Finally, we have shown that our discussion-based multimodal data elicitation method can capture multiparty reasoning behavior in visual environments. Our framework is an important step toward meaningfully visualizing and interpreting such multiparty multimodal data.

## ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation under Award No. IIS-1851591. 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

Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media, Inc., 1st edition.

Blascheck, T., John, M., Kurzhals, K., Koch, S., and Ertl, T. (2016). Va2: A visual analytics approach for evaluating visual analytics applications. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):61–70.

Blascheck, T., Kurzhals, K., Raschke, M., Burch, M., Weiskopf, D., and Ertl, T. (2014). State-of-the-art of visualization for eye tracking data. In *EuroVis*.

Blascheck, T., Kurzhals, K., Raschke, M., Burch, M., Weiskopf, D., and Ertl, T. (2017). Visualization of eye tracking data: A taxonomy and survey. *Computer Graphics Forum*, 36(8):260–284.

Bull, S. and Kay, J. (2016). Smili: a framework for interfaces to learning data in open learner models, learning analytics and related fields. *International Journal of Artificial Intelligence in Education*, 26(1):293–331.

Carter, S. and Bailey, V. (2012). *Facial Expressions : Dynamic Patterns, Impairments and Social Perceptions*. Hauppauge, N.Y. : Nova Science Publishers, Inc. 2012.

Fry, B. and Reas, C. (2014). *Processing: A Programming Handbook for Visual Designers and Artists*. The MIT Press.

Gardner, M., Grus, J., Neumann, M., Tafjord, O., Dasigi, P., Liu, N. F., Peters, M., Schmitz, M., and Zettlemoyer, L. (2018). AllenNLP: A deep semantic natural language processing platform. pages 1–6.

Koné, M., May, M., and Iksal, S. (2018). Towards a dynamic visualization of online collaborative learning. In *Proceedings of the 10th International Conference on Computer Supported Education - Volume 1: CSEDU,*, pages 205–212. INSTICC, SciTePress.

Kontogiorgos, D., Avramova, V., Alexanderson, S., Jonell, P., Oertel, C., Beskow, J., Skantze, G., and Gustafson, J. (2018). A multimodal corpus for mutual gaze and joint attention in multiparty situated interaction. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, Miyazaki, Japan. European Languages Resources Association (ELRA).

Lee, K., He, L., Lewis, M., and Zettlemoyer, L. (2017). End-to-end neural coreference resolution. pages 188–197.

Microsoft (2019). Microsoft Azure Speech to Text. Accessed: 2019-12-17.

Räihä, K.-J., Aula, A., Majaranta, P., Rantala, H., and Koivunen, K. (2005). Static visualization of temporal eye-tracking data. In *IFIP Conference on Human-Computer Interaction*, pages 946–949. Springer.

Ramloll, R., Trepagnier, C., Sebrechts, M., and Beedasy, J. (2004). Gaze data visualization tools: Opportunities and challenges. *2010 14th International Conference Information Visualisation*, 0:173–180.

Sharma, K. and Jermann, P. (2018). Gaze as a proxy for cognition and communication. *2018 IEEE 18th International Conference on Advanced Learning Technologies (ICALT), Advanced Learning Technologies (ICALT), 2018 IEEE 18th International Conference on, ICALT*.

Soon, W. M., Ng, H. T., and Lim, D. C. Y. (2001). A machine learning approach to coreference resolution of noun phrases. *Comput. Linguist.*, 27(4):521–544.

Stellmach, S., Nacke, L. E., Dachselt, R., and Lindley, C. A. (2010). Trends and techniques in visual gaze analysis. *CoRR*, abs/1004.0258.

Tsai, T. J., Stolcke, A., and Slaney, M. (2015). Multimodal addressee detection in multiparty dialogue systems. In *Proc. IEEE ICASSP*, pages 2314–2318. IEEE - Institute of Electrical and Electronics Engineers.

Wang, R., Olson, B., Vaidyanathan, P., Bailey, R., and Alm, C. (2019). Fusing dialogue and gaze from discussions 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.*