SNAG: Spoken Narratives and Gaze Dataset

Preethi Vaidyanathan[†‡], **Emily Prud'hommeaux**[‡°], **Jeff B. Pelz**[‡], **Cecilia O. Alm**[‡]
[†] LC Technologies, Inc., Fairfax, Virginia, USA
[‡] Rochester Institute of Technology, Rochester, New York, USA
[°] Boston College, Boston, Massachusetts, USA
{pxv1621,emilypx}@rit.edu, pelz@cis.rit.edu, coagla@rit.edu

Abstract

Humans rely on multiple sensory modalities when examining and reasoning over images. In this paper, we describe a new multimodal dataset that consists of gaze measurements and spoken descriptions collected in parallel during an image inspection task. The task was performed by multiple participants on 100 general-domain images showing everyday objects and activities. We demonstrate the usefulness of the dataset by applying an existing visual-linguistic data fusion framework in order to label important image regions with appropriate linguistic labels.

1 Introduction

In recent years, eye tracking has become widespread, with applications ranging from VR to assistive communication (Padmanaban et al., 2017; Holmqvist et al., 2017). Gaze data, such as fixation location and duration, can reveal crucial information about where observers look and how long they look at those locations. Researchers have used gaze measurements to understand where drivers look and to identify differences in experts' and novices' viewing behaviors in domain-specific tasks (Underwood et al., 2003; Eivazi et al., 2012). Numerous studies highlight the potential of gaze data to shed light on how humans process information, make decisions, and vary in observer behaviors (Fiedler and Glöckner, 2012; Guo et al., 2014; Hayes and Henderson, 2017; Brunyé and Gardony, 2017). Eye tracking has also long been an important tool in psycholinguistics (Cooper, 1974; Rayner, 1998; Richardson and Dale, 2005; Shao et al., 2013).

Co-collecting observers' gaze information and spoken descriptions of visual input has the potential to provide insight into how humans understand what they see. There is a need for public datasets containing both modalities. In this paper, we present the Spoken Narratives and Gaze dataset (SNAG), which contains gaze information and spoken narratives co-captured from observers as they view general domain images. We describe the data collection procedure using a high-quality eye-tracker, summary statistics of the multimodal data, and the results of applying a visual-linguistic alignment framework to automatically annotate regions of general-domain images, inspired by Vaidyanathan et al.'s (2016) work on medical images. Our main contributions are as follows:

1. We provide the language and vision communities with a unique multimodal dataset[1] comprised of co-captured gaze and audio data, and transcriptions. This dataset was collected via an image-inspection task with 100 general-domain images and American English speakers.

2. We demonstrate the usefulness of this general-domain dataset by applying an existing visual-linguistic annotation framework that successfully annotates image regions by combining gaze and language data.

2 Multimodal Data Collection

The IRB-approved data collection involved 40 university students who were native speakers of American English (10 were later removed), ranging in age from 18 to 25 years, viewing and describing 100 general-domain images. We sought out subjects who were speakers of American English in order to ensure reliable ASR output and a consistent vocabulary across subjects. Subjects consented to data release. The images

[1] https://mvrl-clasp.github.io/SNAG/

Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 132–137
Melbourne, Australia, July 15 - 20, 2018.

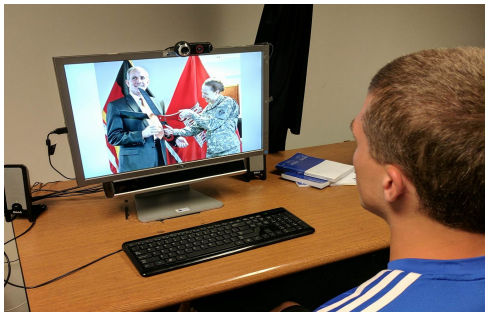

Figure 1: Data collection set-up. The eye tracker is under the display. The observer wears a lapel microphone connected to a TASCAM recorder.

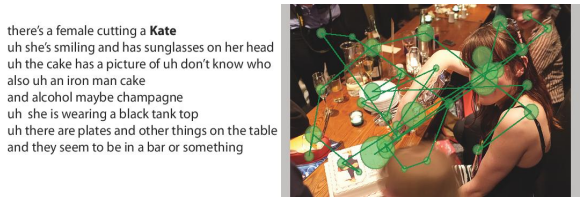

Figure 2: Example of multimodal data. *Left:* ASR transcript of a participant's spoken description. *Right:* Gaze data for the same observer overlaid on the image. Green circles show fixations, with radius representing fixation duration. Green lines connecting fixations represent saccades.

were selected from MSCOCO (Microsoft Common Objects in Context) (Lin et al., 2014), which totals over 300,000 images representing complex everyday scenes. The MSCOCO dataset was created by pooling images from sources such as Flickr and crowdsourcing them to obtain segments and captions (not used in this work). A researcher selected the images so that typically they depicted an event with at least one initiator of the event and one target of the action. Of the 100 images, 69 images clearly depict at least one event. The MSCOCO images vary in number of objects, scale, lighting, and resolution.

Gaze data was collected using a SensoMotoric Instruments (Sensomotoric Instruments, 2016) RED 250Hz eye-tracker attached under a display (Figure 1). The reported accuracy of the RED 250 eye-tracker is 0.5 degree. It is a non-intrusive and remote eye tracker that monitors the observer's gaze. Each image was presented to the observer on a 22-inch LCD monitor (1680×1050 pixels) located approximately 68 cm from the observer. We employed a double computer set-up with one computer used to present the image and the other used to run the SMI software iViewX and Experiment Center 2.3. After each stimulus, a blank gray slide was inserted to ensure that the gaze on the previous stimulus did not affect the gaze on the following stimulus. The blank gray slide was followed by a test slide with a small, visible target at the center with an invisible trigger area of interest. Using the test slide we could measure the drift between the location of the target at the center and the predicted gaze location over time that may have occurred due to the observer's movements. A validation was performed every 10 images and re-calibration was applied if the observer's validation error was more than one degree.

A TASCAM DR-100MKII recorder with a lapel microphone was used to record the spoken descriptions. To approximate the Master-Apprentice data collection method that helps in eliciting rich details (Beyer and Holtzblatt, 1997), observers were instructed to "describe the action in the images and tell the experimenter what is happening." Observers were given a mandatory break after 50 images and optional smaller breaks if needed to avoid fatigue. Observers were given a package of cookies along with a choice between entering into a raffle to win one of two $25 gift cards or receiving course credits. Observers were cooperative and enthusiastic.

3 Fixations, Narratives, and Quality

The SMI software BeGaze 3.1.117 with default parameters and a velocity-based (I-VT) algorithm was used to detect eye-tracking events. Figure 2 shows an example of the scanpath with fixations and saccades of an observer overlaid on an image. Of the original 40 observers, we removed one observer with partial data loss and nine observers whose mean calibration and validation error was greater than two standard deviations from the mean in the horizontal or vertical direction. The mean calibration accuracy (standard deviation) for the remaining subjects was $0.67(0.25)$ and $0.74(0.27)$ degrees for the x and y directions, respectively. One degree would translate to approximately 40 pixels in our set-up, therefore our mean calibration accuracy was roughly 27 pixels. For the remainder of this work, the corpus size is 3000 multimodal instances (100 images $\times$ 30 participants), with 13 female and 17 male

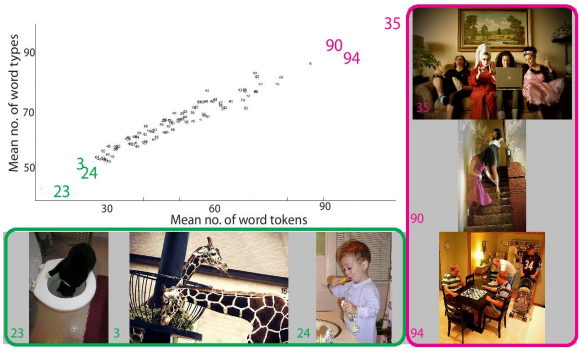

Figure 3: Scatter plot of mean word types vs. tokens per image. Example images have low (green) and high (magenta) type-token ratio.

Figure 4: RegionLabeler GUI (released with dataset) used to acquire reference alignments. Annotator draws borders around regions and checks off linguistic units.

participants.

The speech recordings for the 3000 instances were machine-transcribed using the cloud-based IBM Watson Speech-to-Text service, an ASR system accessible via a Websocket connection[2]. Figure 2 (left panel) shows example ASR output, which is accurate other than the substitution of *Kate* for *cake*. IBM Watson reports timestamps for each word, and those timestamps are included in the released dataset. Additionally, all spoken descriptions for a subset of 5 images were manually corrected using Praat (Boersma, 2002) in order to verify the quality of the ASR output. We found the word error rate (WER) to be remarkably low (5%), demonstrating the viability of using ASR to automate the transcription of the narratives. The ASR and manually corrected transcriptions are included in the dataset.

A descriptive analysis of the gaze and narratives shows that the average fixation duration across the 30 participants was 250 milliseconds and the average narrative duration was about 22 seconds. The transcribed narratives were segmented into word tokens using the default NLTK word tokenizer. Various measures for the first-order analysis of the narratives were then calculated. The mean number of tokens and the average duration of narratives together indicate that on average observers uttered 2.5 words per second. The mean type-token ratio was 0.75, suggesting that there is substantial lexical diversity in the narratives, which demonstrates the richness of the dataset. Figure 3 shows a scatter plot for the mean number of word types against the mean number of word tokens for the 100 images across 30 participants. The plot illustrates that a larger number of tokens typically results in a larger number of types. Images 23, 3, and 24, highlighted in green, have fewer mean word tokens and types than images 35, 90, and 94, highlighted in magenta. For this dataset, this may be due to the number of significant objects in the images where a significant object is defined as an object that occupies a significantly large area of the image. Images 23, 3, and 24 have on average two objects while images 35, 90, and 94 have more than two.

[2]https://www.ibm.com/watson/services/speech-to-text/

4 Application to Multimodal Alignment

We examine the usefulness of our general-domain dataset on image-region annotation, adapting the framework given by Vaidyanathan et al. (2016).

Linguistic units: We process the narratives in order to extract nouns and adjectives, which serve as the linguistic units. Additionally, we remove word tokens with a frequency of 1 in order to reduce the impact of speech errors and one-off ASR errors.

Visual units: To encode fixations into meaningful regions similar to Vaidyanathan et al. (2016) we apply mean shift fixation clustering (MSFC). We also use modified k-means and gradient segmentation (GSEG). Modified k-means uses the number of clusters obtained from MSFC as the value of k instead of 4 as in the original framework. GSEG uses color and texture with region merging to segment an image (Ugarriza et al., 2009). The outputs of the three clustering methods are shown in Figure 5. The rest of the alignment framework, including using the

Figure 5: Example region annotations. Top-left: Reference alignments. Alignment output using: top-right: MSFC; bottom-left: modified *k*-means; and bottom-right: GSEG. Correct alignments in pink. Misalignments and labels not belonging to reference alignments in yellow.

Berkeley aligner (Liang et al., 2006), remained the same.

Reference alignments: Both SURE and POSSIBLE (Och and Ney, 2003) reference alignments were prepared using RegionLabeler, a GUI (Figure 4) to allow evaluation of the resulting multimodal alignments. With this tool, an annotator drew borders of image regions and selected the associated linguistic units.

Baseline alignments: For comparison, we use the baselines proposed by Vaidyanathan et al. (2016): *simultaneous* which assumes that the observers utter the word corresponding to a region at the exact moment their eyes fixate on that region, and *1-second delay* which assumes that there is a 1-second delay between a fixation and the utterance of the word corresponding to that region.

5 Results and Discussion

We calculated average precision, recall, and AER for alignments and compared them against the baselines following Och and Ney (2003).

The two baselines performed similarly. Table 1 shows that the alignment framework performs better than either baseline. MSFC yields the highest recall and lowest AER with an absolute improvement of 0%, 19%, and 10% for precision, recall and AER, over the 1-second delay baseline. Modified *k*-means achieves higher precision with an absolute improvement of 6%, 14%, and 14% over baseline. GSEG performed with less success.

Figure 5 visually compares reference and obtained alignments. Most words are correctly aligned. MSFC correctly aligns labels such as *cake* and *plates*, yielding higher recall. It aligns some labels such as *plates* to incorrect regions, explaining the lower precision. All methods erroneously assign labels not grounded to any region but representing the perspective of the photographer, such as *camera*, to regions in the image, which lowers precision.

6 Related Work

There are publicly available datasets that provide gaze data with no language data (Krafka et al., 2016; Borji and Itti, 2015; Wilming et al., 2017) for tasks such as image saliency or driving. Vasudevan et al. (2018b) collected a dataset in which crowdworkers viewed objects in bounding boxes and read aloud pre-scripted phrases describing those objects. Although their dataset consists of spoken language, it lacks co-collected gaze data and uses a bounding box to highlight an object as opposed to allowing the observer to view the image freely. A more recent study describes the collection of a dataset in which crowdworkers were instructed to draw bounding boxes around objects in videos and provide written phrases describing these objects

	MSFC			Modified k-means		
	Precision	Recall	AER	Precision	Recall	AER
Simultaneous	0.42	0.30	0.65	0.49	0.17	0.74
1-second delay	**0.43**	0.31	0.64	**0.50**	0.17	0.74
Alignment framework	**0.43**	**0.50**	**0.54**	**0.56**	**0.31**	**0.60**

Table 1: Average alignment performance across images. MSFC provides the best recall and lowest AER, and modified k-means the best precision. In all cases, the alignment framework yields stronger results than either of the timing-based baselines.

(Vasudevan et al., 2018a). In a separate task, crowdworkers were asked to view those same videos and to gaze within the bounding boxes for each object while face data was recorded. The authors infer gaze using the recorded face data. None of these datasets involves simultaneous visual-linguistic capture of spoken narration or precision eye-tracking equipment during naturalistic free viewing. Ho et al. (2015) provide a dataset that consists only of gaze and speech time stamps during dyadic interactions. The closest dataset to ours is the multimodal but non-public data described by Vaidyanathan et al. (2016).

7 Conclusions

The SNAG dataset is a unique and novel resource that can provide insights into how humans view and describe scenes with common objects. In this paper, we use SNAG to demonstrate that multimodal alignment does not depend on expert observers or image type, with comparable results to Vaidyanathan et al. (2016) for dermatological images. SNAG could also serve researchers outside NLP, including psycholinguistics. Spontaneous speech coupled with eye-tracking data could be useful in answering questions about how humans produce language when engaging with visual tasks. Parallel data streams can, for example, help in investigating questions such as the effects of word complexity or frequency on language formation and production. It might also aid in studies of syntactic constructions and argument structure, and how they relate to visual perception. Qualitative analysis of our transcripts indicates that they contain some emotional information in the form of holistic comments on the overall affect of the images, which could be helpful in affective visual or linguistic computing tasks. Future work could co-collect modalities such as facial expressions, galvanic skin response, or other biophysical signals with static or dynamic visual materials.

Acknowledgments

We thank Tommy P. Keane for his assistance in developing the image annotation software.

References

Beyer, H. and Holtzblatt, K. (1997). *Contextual Design: Defining Customer-Centered Systems*. Elsevier.

Boersma, P. (2002). Praat, a system for doing phonetics by computer. *Glot International*, 5(9/10):341–345.

Borji, A. and Itti, L. (2015). Cat2000: A large scale fixation dataset for boosting saliency research. *arXiv preprint arXiv:1505.03581*.

Brunyé, T. T. and Gardony, A. L. (2017). Eye tracking measures of uncertainty during perceptual decision making. *International Journal of Psychophysiology*, 120:60–68.

Cooper, R. M. (1974). The control of eye fixation by the meaning of spoken language: A new methodology for the real-time investigation of speech perception, memory, and language processing. *Cognitive Psychology*, 6(1):84–107.

Eivazi, S., Bednarik, R., Tukiainen, M., von und zu Fraunberg, M., Leinonen, V., and Jääskeläinen, J. E. (2012). Gaze behaviour of expert and novice microneurosurgeons differs during observations of tumor removal recordings. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, pages 377–380. ACM.

Fiedler, S. and Glöckner, A. (2012). The dynamics of decision making in risky choice: An eye-tracking analysis. *Frontiers in Psychology*, 3:335.

Guo, X., Li, R., Alm, C., Yu, Q., Pelz, J., Shi, P., and Haake, A. (2014). Infusing perceptual expertise and domain knowledge into a human-centered image retrieval system: A prototype application. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, pages 275–278. ACM.

Hayes, T. R. and Henderson, J. M. (2017). Scan patterns during real-world scene viewing predict individual differences in cognitive capacity. *Journal of Vision*, 17(5):23–23.

Ho, S., Foulsham, T., and Kingstone, A. (2015). Speaking and listening with the eyes: Gaze signaling during dyadic interactions. *PloS One*, 10(8):e0136905.

Holmqvist, E., Thunberg, G., and Dahlstrand, M. P. (2017). Gaze-controlled communication technology for children with severe multiple disabilities: Parents and professionals perception of gains, obstacles, and prerequisites. *Assistive Technology*, 0(0):1–8. PMID: 28471273.

Krafka, K., Khosla, A., Kellnhofer, P., Kannan, H., Bhandarkar, S., Matusik, W., and Torralba, A. (2016). Eye tracking for everyone. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Liang, P., Taskar, B., and Klein, D. (2006). Alignment by agreement. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 104–111.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, pages 740–755.

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Padmanaban, N., Konrad, R., Stramer, T., Cooper, E. A., and Wetzstein, G. (2017). Optimizing virtual reality for all users through gaze-contingent and adaptive focus displays. *Proceedings of the National Academy of Sciences*.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124(3):372–422.

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

Sensomotoric Instruments (2016). Sensomotoric Instruments. https://www.smivision.com/.

Shao, Z., Roelofs, A., and Meyer, A. (2013). Predicting naming latencies for action pictures: Dutch norms. *Behavior Research Methods*, 46:274–283.

Ugarriza, L. G., Saber, E., Vantaram, S. R., Amuso, V., Shaw, M., and Bhaskar, R. (2009). Automatic image segmentation by dynamic region growth and multiresolution merging. *IEEE Transactions on Image Processing*, 18(10):2275–2288.

Underwood, G., Chapman, P., Brocklehurst, N., Underwood, J., and Crundall, D. (2003). Visual attention while driving: sequences of eye fixations made by experienced and novice drivers. *Ergonomics*, 46(6):629–646.

Vaidyanathan, P., Prud'hommeaux, E., Alm, C. O., Pelz, J. B., and Haake, A. R. (2016). Fusing eye movements and observer narratives for expert-driven image-region annotations. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, pages 27–34. ACM.

Vasudevan, A. B., Dai, D., and Van Gool, L. (2018a). Object referring in videos with language and human gaze. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Vasudevan, A. B., Dai, D., and Van Gool, L. (2018b). Object referring in visual scene with spoken language. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*.

Wilming, N., Onat, S., Ossandn, J. P., Acik, A., Kietzmann, T. C., Kaspar, K., Gameiro, R. R., Vormberg, A., and Knig, P. (2017). An extensive dataset of eye movements during viewing of complex images. *Scientific Data*, (4).