Analyzing Multimodal Data of Image Observers

Aliya Gangji, *Muhlenberg College*
Trevor Walden, Preethi Vaidyanathan, Reynold Bailey,
Emily Prud'hommeaux, Cecilia Ovesdotter Alm, *Rochester Institute of Technology*

Motivation

- Machine understanding and annotation of images is still a challenge.
- Emotions have not been extensively explored in image understanding.
- Integrating multimodal data remains a challenge for open domain images.

Contributions

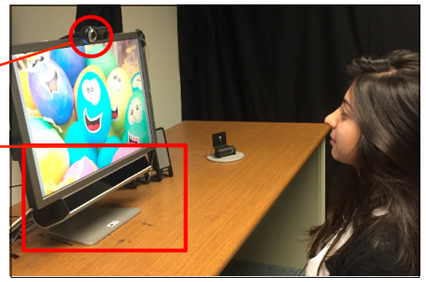
- Builds on work by Vaidyanathan et al. (2016) on multimodal alignment for dermatological images
- Application of multimodal alignment to open domain images
- Comparison of elicited language data based on prompt type and image valence

Multimodal data collection and examples

- 20 subjects
- 15 positive images
- 15 neutral images
- Logitech webcam
- SMI eye-tracker RED 250

Prompts

- *Describe prompt* - "Describe the following image."
- *Photographer prompt* - "What feeling was the photographer hoping to capture in the following image?"
- *Feel prompt* - "How does the following image make you feel?"

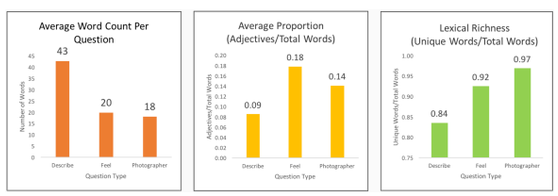


- Subjects were provided 3 prompts per image
- The prompt-image pairs appeared in a different random order for each subject

Mean Shift Fixation Cluster (MSFC)

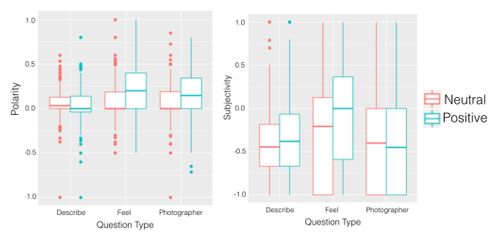
The colored and numbered areas are clusters identified by the MSFC Algorithm

Linguistic analysis



- Observers used more words when responding to the describe prompt
- Responses tended to have greater lexical richness and more adjectives when responding to the feel and photographer prompt

Sentiment analysis



- Observers responded more positively to feel and photographer prompt when positive images were shown (left)
- Subjectivity was highest for feel prompt when positive images were shown (right)

References and Software

Vaidyanathan, P., Prud'hommeaux, E., Pelz, J. B., Alm, C. O., And Haake, A. R. 2016. Fusing eye movements and observar narratives for expert drive image-region annotations. In Proceedings of the Ninth Biennial ACM Symposium of Eye Tracking Research & Applications, ACM, New York, NY, ETRA '16, 27-34.

- SMI Experiment Suite
- IBM ASR API
- Berkeley Parser
- Berkeley Aligner
- Affectiva Affdex SDK

Acknowledgments

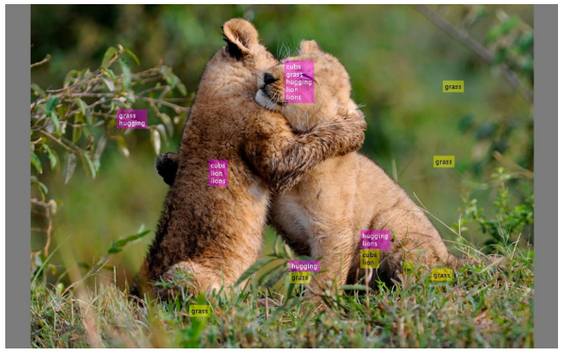
This material is based upon work supported by the National Science Foundation under Award No. IIS-1559889. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Multimodal alignment

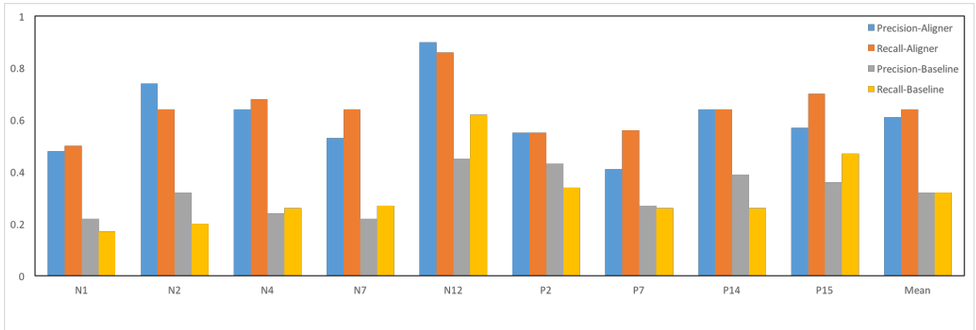
Eye movement data (Visual Units) and nouns and adjectives from subjects' transcribed narratives of the Description prompt (Linguistic Units) were fed through the bitext alignment pipeline established by Vaidyanathan et al.

Example: Uncorrected ASR
it sough to little mine cubs hugging

Example: Corrected ASR
it's looks like little lion cubs hugging



Each box of words is located at the center of its fixation cluster. Magenta boxes: words aligned to that cluster using our method. Yellow boxes: words in our reference alignment, not aligned by our method.



Alignment improves when linguistic units are constrained by frequency.

Comparison of alignment with vs. without manual correction of ASR:

- Improved recall over baseline for image N7 and image P2 was marginal
- Improvement was large for image P1
- Variability relates to the word error rate (WER) of the ASR, P1 had higher WER than N7 and P2

Conclusions and future work

- In some case, correct ASR output aids alignment quality more, and in others less. Our data suggests a link to ASR quality.
- Prompt type and image valence influence elicited language data
- Investigate ways that holistic information, such as responses to the Feel and Photographer prompts can be used to improve machine image understanding.
- Further investigate the difference in alignment performance with corrected and raw ASR output.