Capturing, Analyzing, and Aligning Multimodal Responses to Affective Visual Content

David Nester (david.nester@emu.edu)
Nikita Haduong, Preethi Vaidyanathan, Emily Prud'hommeaux, Reynold Bailey, Cecilia O. Alm

Overview and Contributions

- Humans routinely extract important information from images and videos.
- Computers still have difficulty annotating important information in visual data in a human-like manner.
- We co-capture participants' gaze, language, and facial expressions as they describe positive and negative visual stimuli.
- Contributions:
  - We mapped gaze to speech with a multimodal alignment framework [1, 2], outperforming the baseline comparison.
  - Filtering words occurring once improved alignment performance and helped exclude ASR word errors.
  - We also explored patterns across modalities, for example the affect of linguistic tokens associated with stimulus valence.

Alignment

Example description: A woman is holding scissors...cutting his tie?...they're very large...both of them are smiling...there are some flags in the background...it's a fairly happy occasion

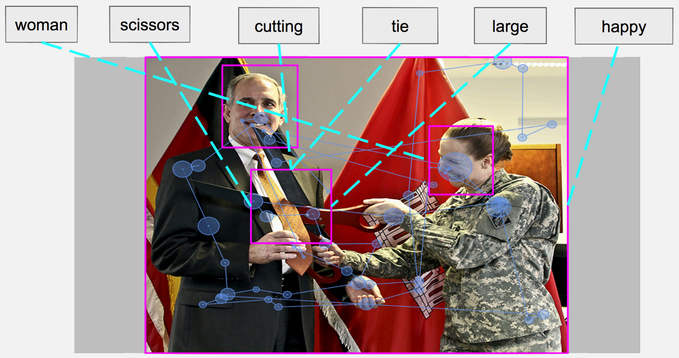


Figure 1: Example of mapping gaze to spoken words

- Berkeley Aligner usually aligns words between languages for machine translation
- Instead, we align linguistic units (nouns, adjectives) with visual units (clustered gaze regions)
- Baseline aligns linguistic and visual units temporally
- Both compared to manual reference alignments

$$AER = 1 - \frac{|A \cap S|}{|A + S|}$$

A = *Aligner Output Pairs*
S = *Reference Pairs*

Figure 2: Alignment performance metric

Figure 3: Clustered gaze regions

Experiment Setup

- 20 images and 20 short videos
- 10 positive and 10 negative each
- 21 subjects
- Task: *Describe the content of this image/video.*

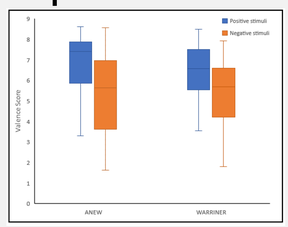
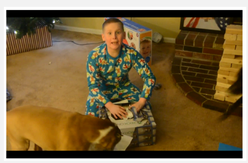


Figure 4: Words with positive valence were used more with positive stimuli, and vice versa.



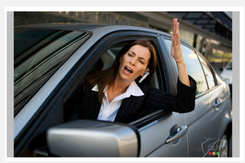


Figure 5: Examples of stimuli used - positive video and negative image

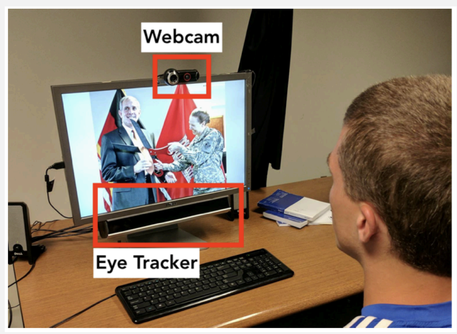


Figure 6: Observer in front of monitor with webcam and eye tracker. Lapel microphone used to record voice.

Acknowledgements

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.

Results

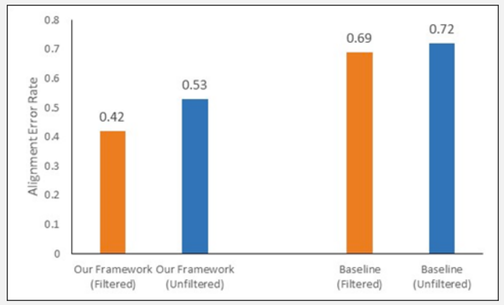
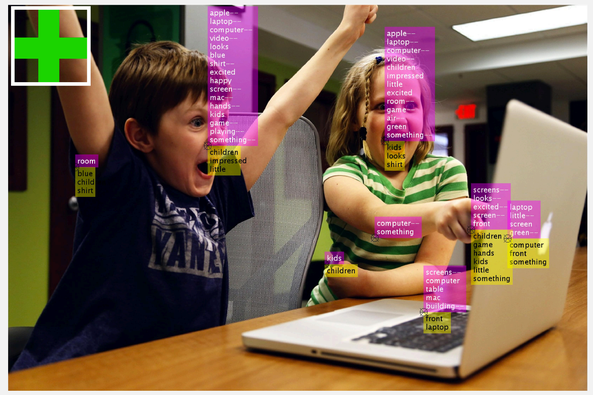


Figure 7: A comparison of average AER performance across images for alignment using unfiltered (blue) and filtered (orange) word lists.

a)

Unfiltered



b)

Filtered

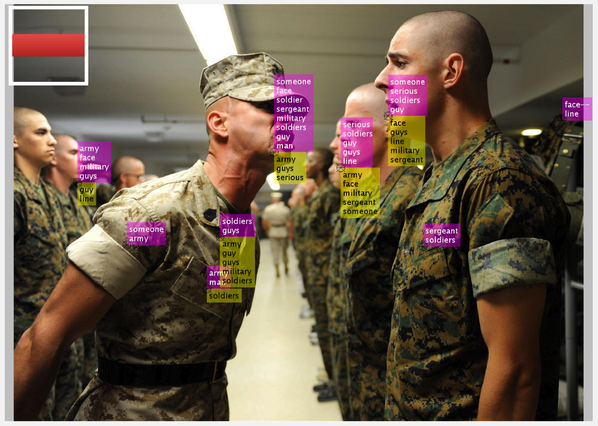


Figure 8 (a,b): Images with aligned linguistic units attached to locations of visual units. Magenta: aligned correctly, -- are incorrect. Yellow: not aligned but in reference list. Dominant valence of facial expressions depicted top left. Word list for a) is unfiltered while b) is filtered.

References

[1] Gangji, A.; Walden, T.; Vaidyanathan, P.; Prud'hommeaux, E.; Bailey, R.; Alm, C.; 2017. Using co-captured face, gaze, and verbal reactions to images of varying emotional content for analysis and semantic alignment.

[2] Vaidyanathan, P.; Prud'hommeaux, E.; Pelz, J. B.; Alm, C.O.; and Haake, A. R. 2016. Fusing eye movements and observer narratives for expert-driven image-region annotations.