SNAG: Spoken Narratives and Gaze Dataset

Preethi Vaidyanathan[1], Emily Prud'hommeaux[2], Jeff B. Pelz[3], Cecilia O. Alm[3]
[1]LC Technologies, Inc. Fairfax, Virginia, USA, [2]Boston College, Boston, Massachusetts, USA
[3]Rochester Institute of Technology, Rochester, New York, USA

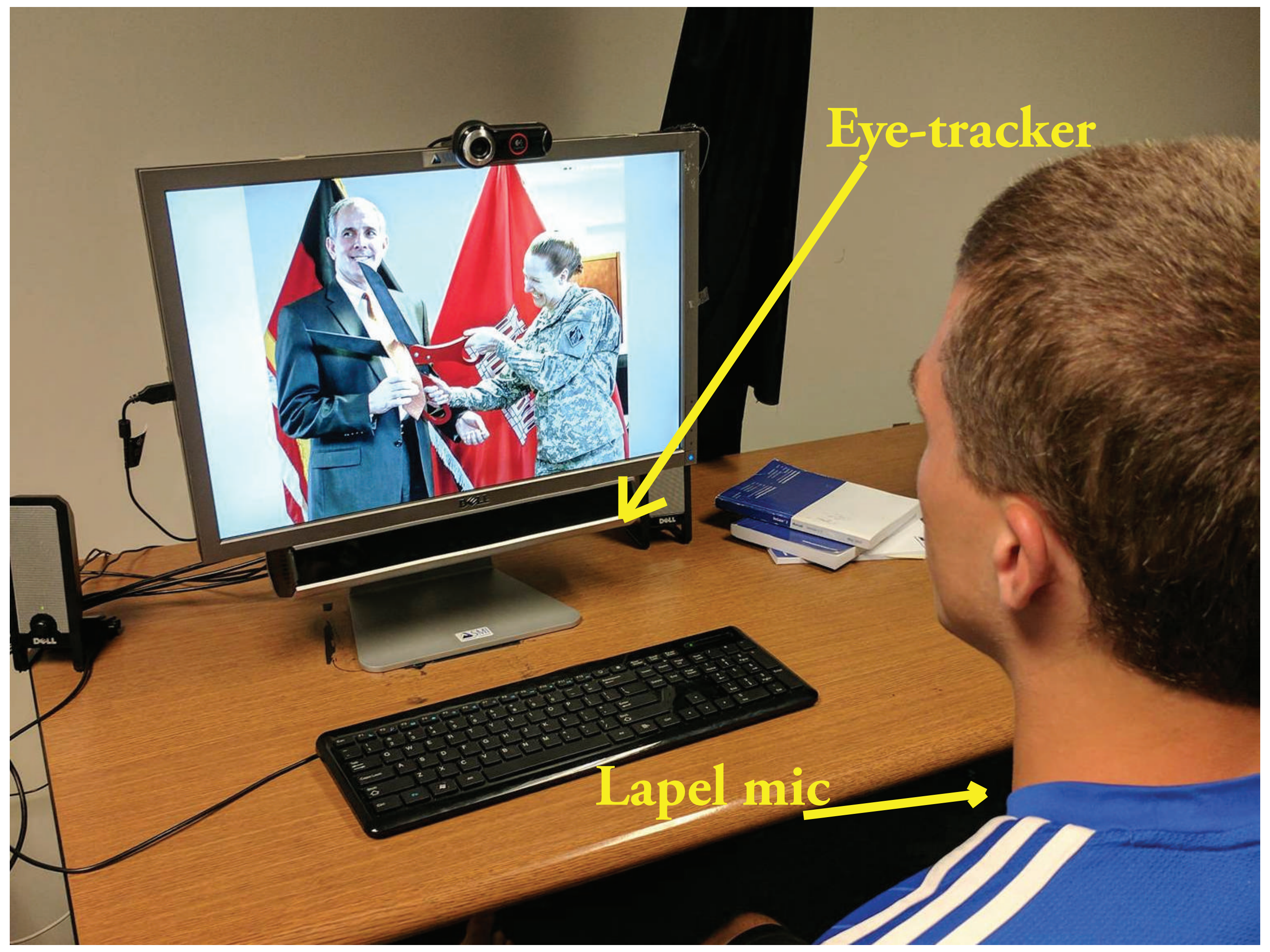
MULTIDISCIPLINARY VISION RESEARCH LABORATORY

Background

- Multimodal data useful to understand human perception
- No publicly available dataset with co-collected spoken narration and gaze information during naturalistic free viewing
- Unique multimodal dataset comprised of co-captured gaze and audio data, and transcriptions for the language and vision communities
- Application of SNAG to visual-linguistic annotation framework (Vaidyanathan et al. 2016) to label image regions

Data Collection

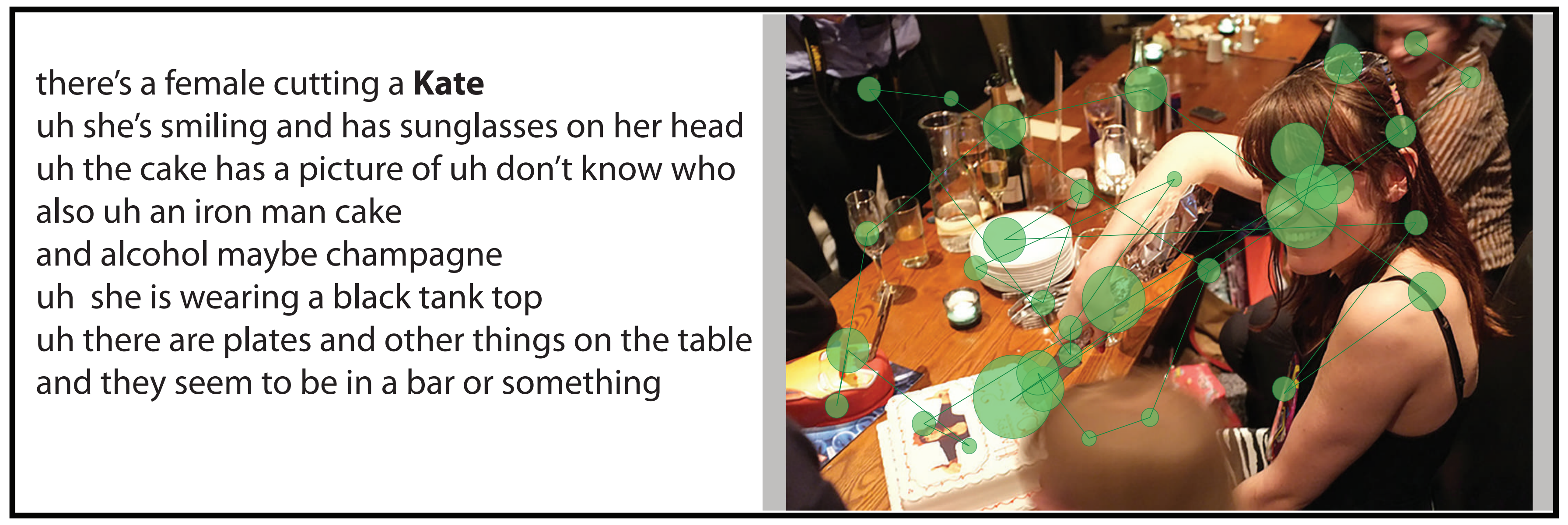
- 30 American English speakers, 18-25 yrs old, 13 female & 17 male
- 100 general-domain images selected from MSCOCO dataset
- DR-100MKII TASCAM with lapel microphone
- SMI Eye-Tracker RED250, remote eye tracker running at 250Hz
- Modified Master-Apprentice to elicit rich details
- "Describe the action in the images and tell the experimenter what is happening."



Dataset and tool available at:
https://mvrl-clasp.github.io/SNAG/

Multimodal Dataset

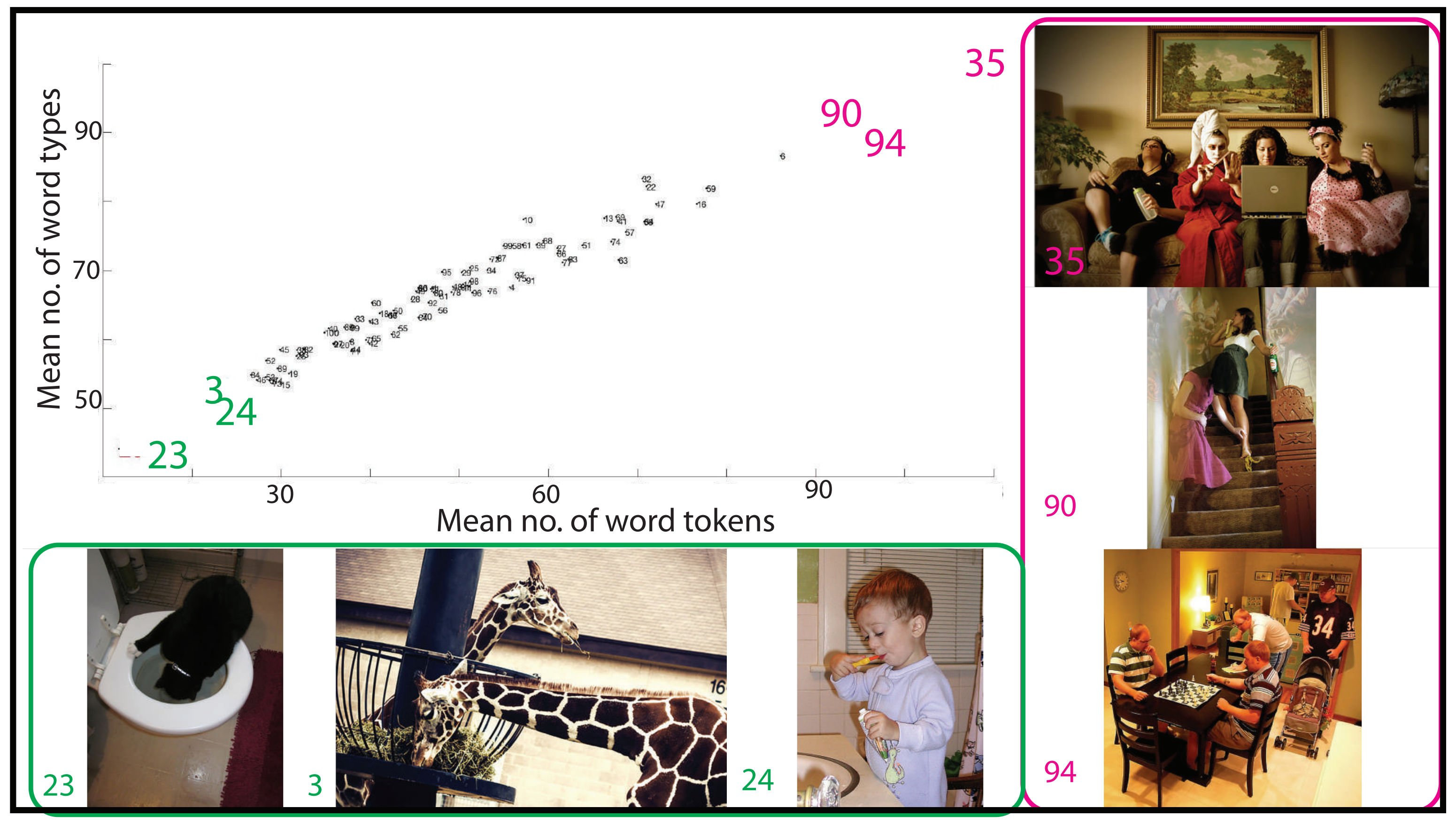
- Transcripts generated with IBM Watson STT (WER ~5%)
- Fixations represented using green circles, radius indicates fixation duration
- Green lines represent saccades



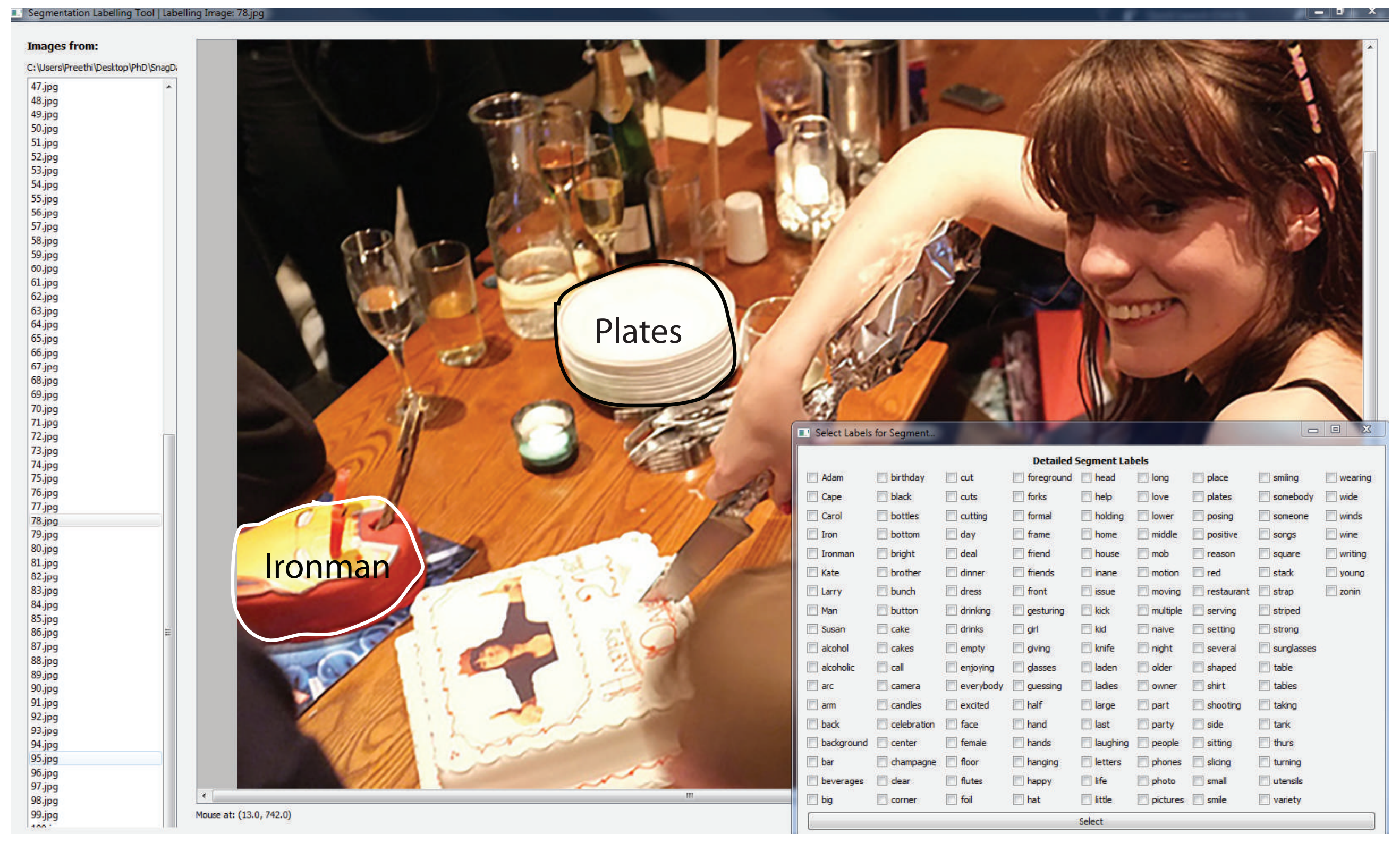
there's a female cutting a **Kate**
uh she's smiling and has sunglasses on her head
uh the cake has a picture of uh don't know who
also uh an iron man cake
and alcohol maybe champagne
uh she is wearing a black tank top
uh there are plates and other things on the table
and they seem to be in a bar or something

ASR transcript | Eye movements

- Wide range of type-token ratio corresponds to range of image complexity.
- Overall mean type-token ratio (0.75) shows substantial lexical diversity.



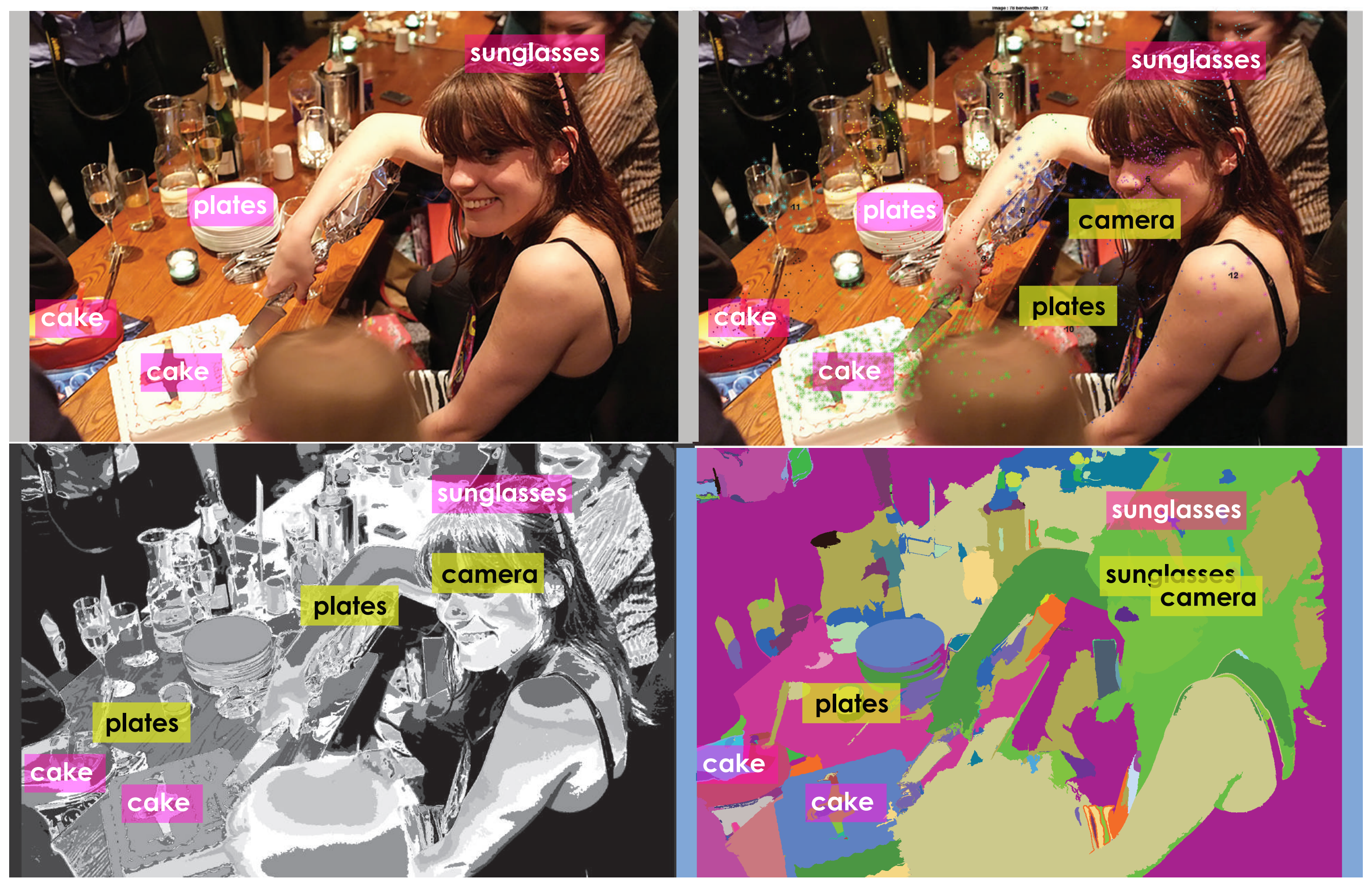
RegionLabeler: Image Annotation Tool



Labeling Images via Multimodal Alignment

- Alignments generated via Berkeley aligner used for machine translation
- Alignments from framework compared against 1-sec delay baseline
- Best AER=0.54 using MSFC vs. baseline AER=0.64

Reference alignments | Mean shift fixation clustering (MSFC)



Adaptive *k*-means | Gradient segmentation (GSEG)

Conclusions and Future Work

- Unique and novel resource for understanding how humans view and describe scenes with common objects.
- It can serve researchers in computer vision, computational linguistics, psycholinguistics, and others.
- Visual-linguistic alignment framework independent of the type of images or expert observers.
- Co-collect modalities such as facial expressions, galvanic skin response, or other biophysical signals with static and dynamic visual materials.

References

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 and Research Applications*, pg 27-34, ACM.