Where the Action Could Be: Speakers Look at Graspable Objects and Meaningful Scene Regions when Describing Potential Actions

Gwendolyn Rehrig[1] (glrehrig@ucdavis.edu), Candace Peacock[1,2], Taylor Hayes[2], John Henderson[1,2], Fernanda Ferreira[1]

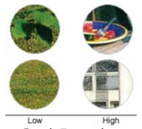
University of California, Davis [1]Department of Psychology [2]Center for Mind and Brain

What Guides Visual Attention During Speech?

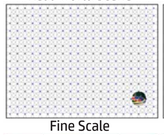
Background:

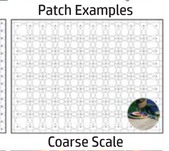
- **Meaning** vs. **saliency** maps (Henderson & Hayes, 2017)
  - Constructed meaning maps from mTurk ratings
  - Generated a spatial representation comparable to saliency maps (using GBVS, Harel et al., 2006)
  - Scene **meaning** explains variance in attention maps better than image **salience** does
  - Advantage of **meaning** replicates in scene and action description tasks (Henderson et al., 2018)
- The interactions we can perform on objects (*object affordances*) also influence visual attention (Malcolm & Shomstein, 2015; Castelhano & Witherspoon, 2016; Gomez & Snow, 2017)

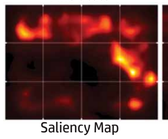

Real-world Scene

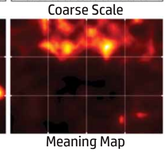


Patch Examples


Fine Scale


Coarse Scale


Saliency Map


Meaning Map

Research questions

- Do object affordances explain visual attention when describing a scene's potential actions?
  - Operationalized as **graspability**
  - ***Test:*** Compare **graspability** and attention maps
- Do meaning maps capture grasping affordances?
  - ***Test:*** Compare **meaning** and **grasp** maps

Feature Maps

Saliency maps: Physical salience calculated using Graph-Based Visual Salience (Harel et al., 2006)

Meaning maps: Crowdsourced scene patch ratings with scene context (Peacock et al., 2019)

- Exp. 1&2: 84 subjects, Exp. 3: 124 subjects

Grasp maps: Crowdsourced scene patch ratings with scene context

- Exp. 1&2: 84 subjects, Exp. 3: 128 subjects

Reach maps: Crowdsourced scene patch ratings with scene context

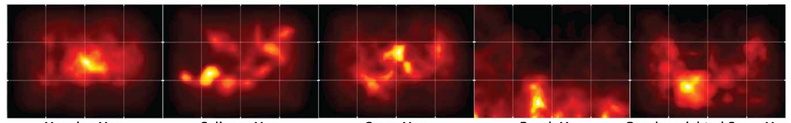
- Used to construct reach-weighted grasp maps (Exp. 3 only)
- Exp. 3: 212 subjects

*What do we mean by "**meaningful**"? We want you to assess how "meaningful" an image is based on how informative or recognizable you think it is.*

*What do we mean by "**graspable**"? We want you to assess how easily an object can be picked up or manipulated with one's hands.*

*What do we mean by "**reachable**"? We want you to assess how easily you could naturally reach what is shown in the scene patch, using your arms and hands only, if you were standing at the camera's viewpoint.*

Real-world Scene


Meaning Map | Saliency Map | Grasp Map | Reach Map | Reach-weighted Grasp Map

Experimental Paradigm

Subjects

- UC Davis undergraduates ($N_{Exp.1}$ = 30, $N_{Exp.2}$ = 40, $N_{Exp.3}$ = 40)
  - Native English speakers

Stimuli

- Real-world scenes ($N_{Exp.1}$ = 30, $N_{Exp.2}$ = 40, $N_{Exp.3}$ = 40)
  - Meaning, grasp, and saliency mapped
- Pseudorandom presentation order
  - No two scenes of the same type in a row (e.g., kitchen, living room)

Task

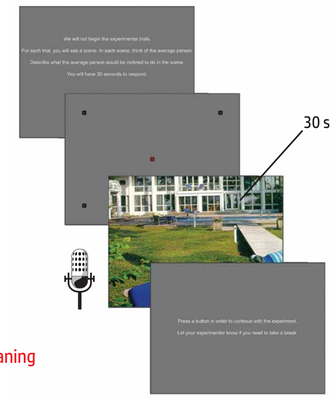
- Describe possible actions in the scene
  - 30 s viewing period
  - Simultaneously view and describe scene

Measures

- Eye movements (EyeLink 1000+)
- Productions (Shure SM86 microphone)

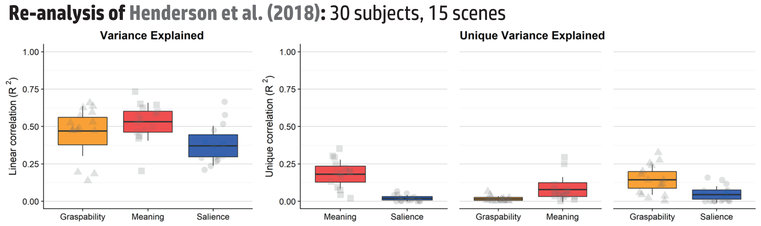
Predictions

- Meaning > salience (Henderson et al., 2018)
- If grasping affordances guide attention, grasp > meaning
- If meaning reduces to graspability, grasp = meaning



Experiment 1

Re-analysis of Henderson et al. (2018): 30 subjects, 15 scenes



Map correlations (R^2): Meaning and Graspability were highly correlated (M = 0.79)

Linear correlations (R^2): Feature maps and attention maps

- Meaning (M = 0.53) > Graspability (M = 0.47): $t(14) = 2.75$, p= 0.02, 95% CI = [0.01 0.11]
- Meaning (M = 0.53) > Salience (M = 0.37): $t(14) = 5.84$, $p < .0001$, 95% CI = [0.10 0.22]
- Graspability (M = 0.47) > Salience (M = 0.37): $t(14) = 2.63$, p= 0.02, 95% CI = [0.02 0.18]

Semipartial correlations (R^2): Unique correlations between feature maps and attention maps

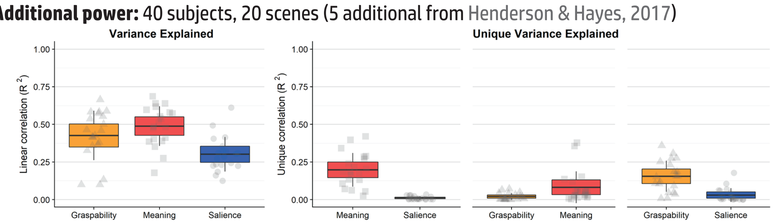
- Meaning (M = 0.08) > Graspability (M = 0.02): $t(14) = 2.74$, p = 0.02, 95% CI = [0.1 0.11]
- Meaning (M = 0.18) > Salience (M = 0.02): $t(14) = 5.84$, $p < 0.001$, 95% CI = [0.10 0.22]
- Graspability (M = 0.14) > Salience (M = 0.04): $t(14) = 2.63$, p = 0.02, 95% CI = [0.02 0.18]

Conclusion

- Graspability does not account for variance in attention as well as meaning

Experiment 2

Additional power: 40 subjects, 20 scenes (5 additional from Henderson & Hayes, 2017)



Map correlations (R^2): Meaning and Graspability were highly correlated (M = 0.81)

Linear correlations (R^2): Feature maps and attention maps

- Meaning (M = 0.49) > Graspability (M = 0.43): $t(19) = 2.59$, p = 0.02, 95% CI = [0.01 0.11]
- Meaning (M = 0.49) > Salience (M = 0.30): $t(19) = 7.18$, $p < .0001$, 95% CI = [0.13 0.24]
- Graspability (M = 0.43) > Salience (M = 0.30): $t(19) = 4.08$, $p < 0.001$, 95% CI = [0.06 0.19]

Semipartial correlations (R^2): Unique correlations between feature maps and attention maps

- Meaning (M = 0.08) > Graspability (M = 0.02): $t(19) = 2.59$, p = 0.02, 95% CI = [0.01 0.11]
- Meaning (M = 0.20) > Salience (M = 0.01): $t(19) = 7.19$, $p < 0.001$, 95% CI = [0.13 0.24]
- Graspability (M = 0.16) > Salience (M = 0.03): $t(19) = 4.09$, $p < 0.001$, 95% CI = [0.06 0.19]

Conclusion

- Graspability again does not account for variance in attention as well as meaning

Experiment 3: Motivation

Experiment 1 and 2 Limitations

- Stimuli were not optimal
  - Did not contain many graspable objects
  - Did not depict reachable spaces (Josephs & Konkle, 2019)
- Task instruction was to imagine actions the average person would do in the scene
- Limitations may have downweighted the role of grasping object affordances

Experiment 3

- 20 novel scenes
  - All show graspable objects within reach of the camera's viewpoint
- Task instructions
  - Instruction changed to ask what the subject would do in the scene
- 40 subjects

Experiment 3: Results

Linear correlations (R^2, *M*)

- Meaning = Graspability (.36)
- Meaning (.36) > Salience (.28)*
- Graspability (.36) > Salience (.28)*
- Meaning (.36) > Reach-weighted (.29)*
- Graspability (.36) > Reach-weighted (.29)*
- Reach-weighted (.29) ~= Salience (.28)

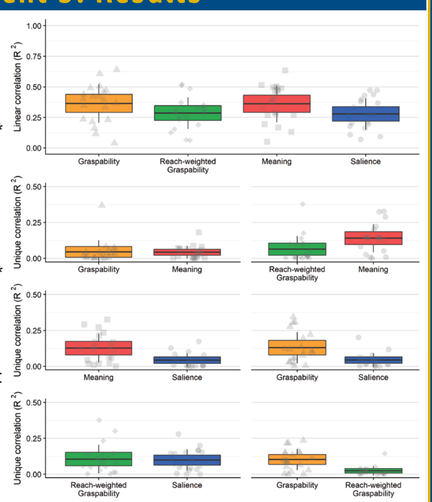
Semipartial correlations (R^2, *M*)

- Meaning (.04) ~= Graspability (.05)
- Meaning (.13) > Salience (.04)*
- Graspability (.13) > Salience (.04)*
- Meaning (.14) > Reach-weighted (.06)*
- Graspability (.10) > Reach-weighted (.02)*
- Reach-weighted (.11) ~= Salience (.10)

* paired t-test $p < .05$

Early fixations

- Results for early fixations were consistent with the full trial period *except*:
  - Reach-weighted graspability outperformed graspability and meaning during the first two fixations only



Conclusions

Interpretation of Experiment 3 Results

- Graspability and meaning accounted for comparable variance in attention maps for new scenes when multiple graspable objects are present both in reach of the camera's viewpoint and beyond
- Reach-weighted graspability accounted for variance in attention maps well early on only
  - May reflect foreground bias (Fernandes & Castelhano, 2019) or center bias (Tatler, 2007; Hayes & Henderson, 2019)

Conclusions

- Counter to our predictions, graspability did not outperform meaning when speakers described possible actions in a scene. For optimal stimuli, graspability and meaning explained variance in attention maps equally well, suggesting graspability is co-extensive with meaning (Altmann & Kamide, 2007)
- Counter to accounts of language production (e.g., Gleitman et al., 2007), image salience does not determine what speakers look at when describing a scene
  - Psycholinguists studying vision-language interactions should quantify image salience
- Visual cognition optimizes for the information available, pushing attention to locations with information that is most task-relevant
  - Sensitive to grasping affordances when there are graspable objects in the scene within reach
  - Uses more general scene meaning otherwise

Discussion & Acknowledgments

Limitations

- Used 2D stimuli and task
  - Grasping affordances likely more relevant in 3D space
- Only measured grasping affordances
  - Other object affordances (e.g., sitting on a chair) may have been important

Future work

- Construct interact maps to capture broader representation of object affordances
- Verbal descriptions
  - Analyze the verbs subjects use to describe possible actions
  - Determine which objects (and associated verbs) were mentioned first

Thank you to our research assistants for their hard work, and to members of the Ferreira Lab and the Visual Cognition Lab for their help and feedback

For additional results and details, see Rehrig, Peacock, Hayes, Henderson, & Ferreira (2020)

This research was supported by the National Eye Institute of the National Institutes of Health under award number R01EY027792 awarded to John M. Henderson, and National Science Foundation grant BCS-1650888 awarded to Fernanda Ferreira