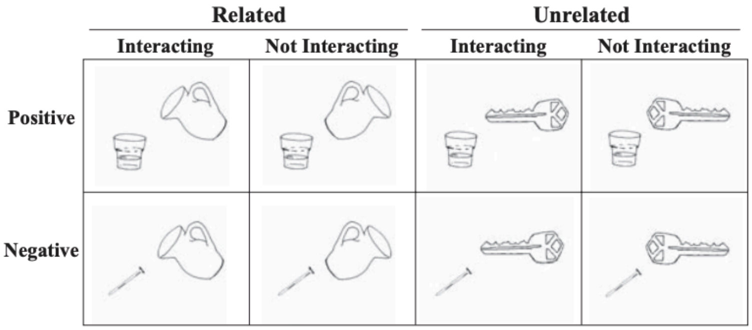
Role Recognition In Brief Displays: A Method for Naturalistic Images

Abstract

Humans are highly adept at extracting relational information from visual scenes. Research has shown that people can identify functional relationships between objects and are sensitive to relational roles between people (Green & Hummel, 2006; Papeo et al., 2017; Papeo et al., 2024). However, few studies have tested this ability using naturalistic images. Most studies rely on simplified, lab-generated stimuli. In this study, we aim to establish a method that can be used to determine whether humans can still extract agent-patient relational information from more realistic and complex images, such as news images.

Introduction/Background

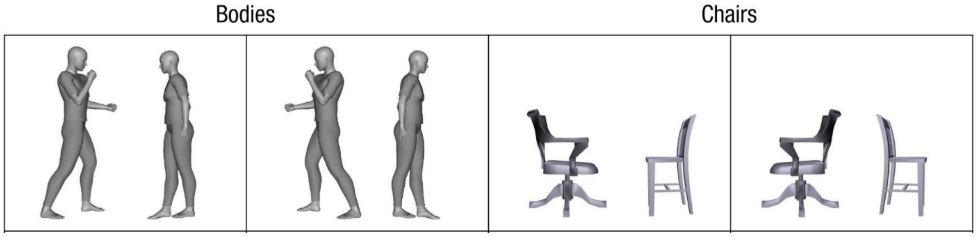
In a 2006 study, Green and Hummel investigated how people perceptually group interacting object pairs. Participants were briefly shown a target object and a distractor and asked whether the target matched a given label. In some trials, the target and distractor object were arranged to appear as if they were functionally interacting (ex. A pitcher 'pouring' into a glass). When the distractor was both semantically related to the target and interacting, participants identified the target more accurately than when the distractor was related and not interacting. When the distractor was unrelated and interacting, accuracy was lower than when the distractor was unrelated not interacting. These findings suggested that humans are attuned to functional relationships between objects based on visual input.

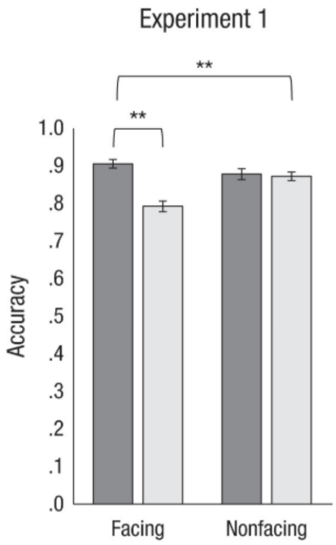


The label is "glass" in these examples.

Fig. 1. (Green & Hummel, 2006) In this example, the data showed that participants were more accurate in identifying the glass when it was paired with the pitcher pouring into it (row 1, column 1), compared to when the glass was paired with the pitcher pouring the wrong way (row 1, column 2). Additionally, participants were more accurate in identifying the glass when it was paired with the key inserting the wrong way (row 1, column 4), compared to when the glass was paired with the key inserting into it (row 1, column 3).

To extend this finding to relationships between humans, we refer to a study by Papeo et al. (2017), which examined how people perceive pairs of objects or bodies in either facing or non-facing orientations. Participants were asked to identify whether each pair depicted people, chairs, or plants. Body dyads facing each other, implying interaction, were recognized more accurately than body dyads facing away from each other, which appeared noninteracting. This result suggested that human recognition of functional relationships extends beyond objects to human bodies. Similar to the pitcher spout facing the glass, two facing bodies would be perceptually grouped together. However, the orientation of the chairs only affected recognition inconsistently. This supported the idea that this effect is driven by functional interactions between objects rather than their physical positioning.





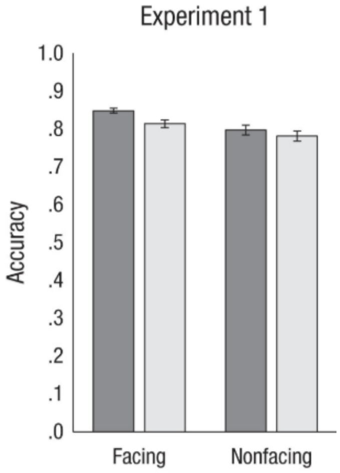


Fig. 2. (Papeo et al., 2017) Disregard the light grey bars; they represent data from a different experiment. The dark grey bars in the left graph indicate that facing bodies were more accurately recognized than nonfacing bodies. Although the dark grey bars in the right graph suggest a slight increase in recognition for facing chairs, subsequent experiments revealed that this effect was inconsistent.

These studies demonstrate that people are able to perceive functional relations between bodies. Building on this, a 2013 study by Hafri et al. showed that individuals can recognize agent-patient roles from brief visual displays, where the agent is the one initiating the action and the patient is the recipient. In Experiment 2A of the study, participants were shown a brief image of a male-female pair where one was performing an action on the other (e.g. ‘punching’), followed immediately by a scrambled mask to block further visual processing. Participants were then asked a question in the format, ‘Is the boy performing the action?’, and were prompted to answer yes or no. The results (see below) confirmed that event roles can be recognized from brief displays.

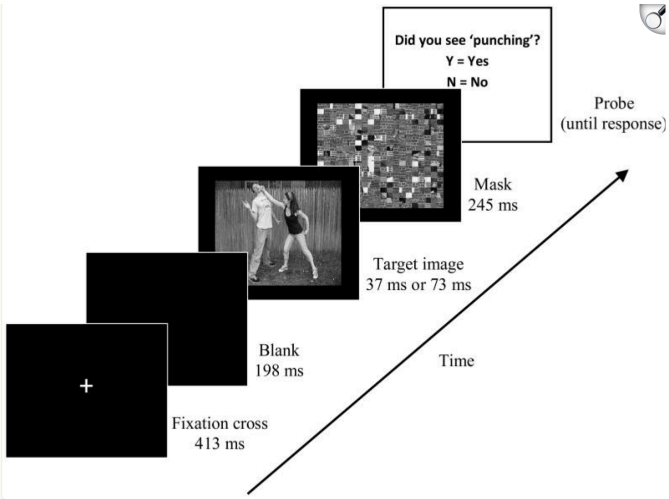
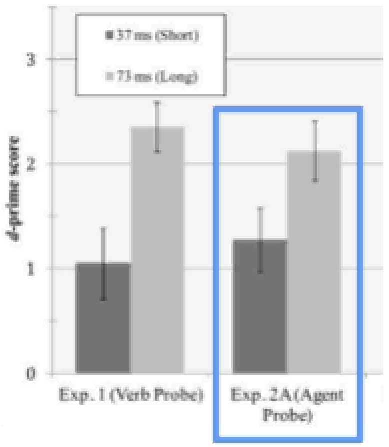
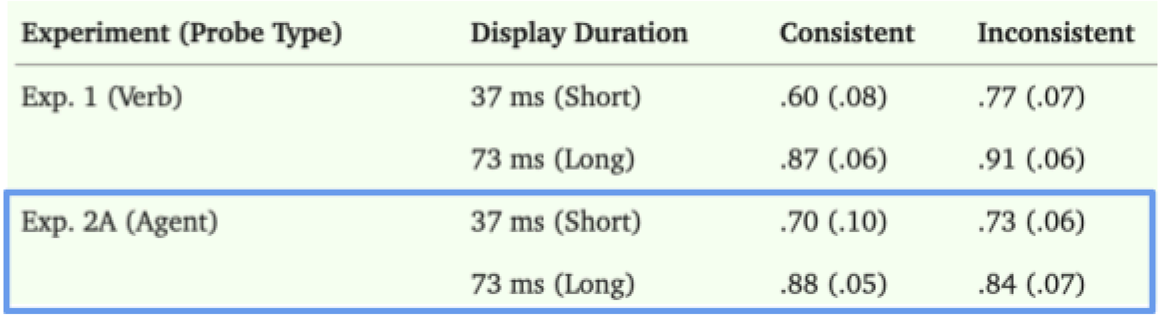


Fig. 3. (Hafri et al., 2013) Trial structure for the experiments. Experiment 2A had the same structure, except the question was ‘Is the boy/girl performing the action?’ All target images were staged and photographed by the experimenter.





Experiment (Probe Type)	Display Duration	Consistent	Inconsistent
Exp. 1 (Verb)	37 ms (Short)	.60 (.08)	.77 (.07)
	73 ms (Long)	.87 (.06)	.91 (.06)
Exp. 2A (Agent)	37 ms (Short)	.70 (.10)	.73 (.06)
	73 ms (Long)	.88 (.05)	.84 (.07)

Fig. 4. (Hafri et al., 2013) See results enclosed in blue rectangles. The accuracy in the table for the Agent probe is above chance (0.5) for both the Short and Long durations of both the Consistent and Inconsistent trials, where Consistent means the person being asked about is the agent and vice versa. The d-prime score graph shows that participants were responsive to the signal in both the Long and Short durations, but significantly more in the Long durations.

Our study aims to investigate whether humans can recognize role relations in more naturalistic images, extending the results from Hafri et al. (2013). This is a pertinent question because real world images introduce complexities absent from controlled lab stimuli. Unlike staged experimental images, real images often include more than two people, off-center compositions, and ambiguous or obscured body postures, all of which make it more difficult to determine agent-patient relationships.

However, before attempting to replicate Hafri et al.'s (2013) experiment with naturalistic stimuli, we must first develop a reliable method for referring to individuals in the real images. In the original study, participants were shown images containing only a boy and a girl, allowing for simple questions like 'Is the boy performing the action?' With realistic images there may be a greater number and variety of individuals, making such simple labels inadequate. To address this, we introduce and test a new method of referring to the individuals in the images–using visual markers (e.g. crosses) to clearly indicate the person being referenced.

Experiment 1a: Recognition of Event Roles Using a Cross to Refer to the Person

We modify the procedure used by Hafri et al. (2013) to test if we can replicate their results using our revised method of posing the question. The following section outlines Hafri et al.'s (2013) original methodology, along with our modifications.

Methods

Subjects

We initially recruited 63 participants, however, data from the first 16 individuals had to be excluded due to a misunderstanding caused by the wording in the trial. Specifically, the trial prompt used the phrase 'Is the person at the cross taking the action?' Several participants (n=3) asked for clarification because they interpreted 'taking the action' as receiving the action rather than performing it. Upon reviewing Hafri et al.'s (2013) original phrasing, we realized we had mistakenly changed the wording, so we revised the wording to the original 'performing the action' phrasing for subsequent participants. All participants were undergraduate students enrolled in a psychology course at UCLA or other adults from the UCLA community. All participants received course credit for their participation. Additionally, one participant self-reported an age of 2 years old and was therefore dropped from the dataset.

Stimuli and Apparatus

We are grateful to Hafri et al. for sharing the image set they created and used in their study. The experimental stimuli were photographs each depicting two people (one male, one female) engaging in an action with a clearly defined agent and patient. There were 16 distinct actions, each represented by four images to account for both role and gender swaps as follows: male agent on the left, male agent on the right, female agent on the left, female agent on the right. This created a total of 64 unique images.

Each image was a 640x480-pixel color image. A 38x38-pixel cross was manually placed at the center of each person's face (both agent and patient) to create the prompt cross images.

Stimuli were displayed on a CRT monitor with a refresh rate of 75 Hz.

Design

The study used a within-subjects design. Each participant experienced both role conditions and both durations. The role condition was whether the role at the cross was the agent or patient. The duration condition was either short (40 ms) and long (80 ms). The dependent variable was the mean accuracy for role recognition and d' prime scores.

Participants first completed four practice trials using two images (not included in the test set). The practice trials had two long-duration (200 ms) and two short-durations (80 ms) presentations.

Participants then completed 64 test trials. Images from the same action category or featuring the same actor pair were never presented consecutively. At least 7 other images were placed between any two images that came from the same event category. Each trial was structured as follows: a fixation dot, a blank screen, the stimulus images, another blank screen, and then the yes/no prompt question with a cross at the location of either the agent/patient's face in the previous image. The mask was removed to help participants remember which person was at the location of the prompt cross.

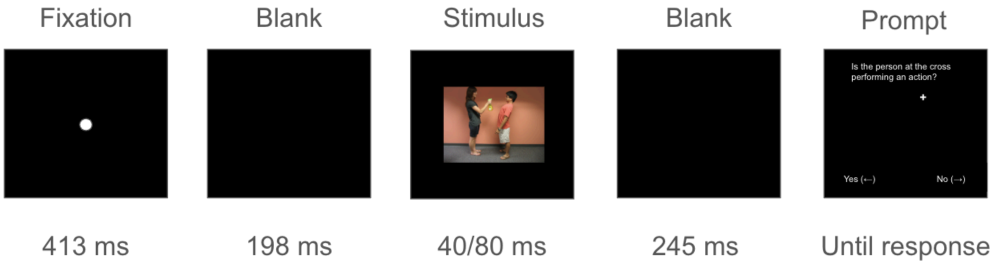


Fig. 5. (Courtesy of Yiling Yun) Trial structure as described above.

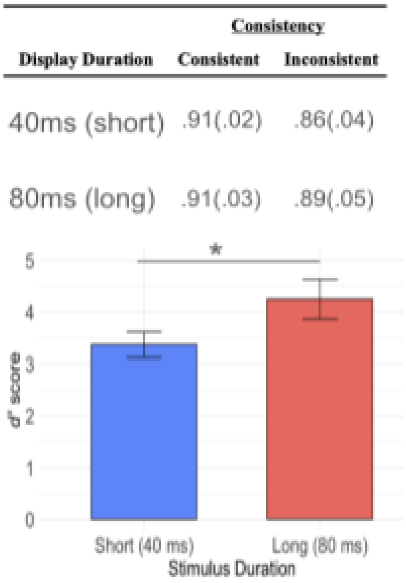
Procedure

Following informed consent, subjects were tested individually in a dark, quiet room. They were seated in front of the computer monitor and were given verbal instructions not to turn on the lights or exit the program. Participants were informed they would be viewing images and answering a yes or no question after each image using the keyboard.

A research assistant guided each participant through a practice trial, without disclosing any correct answers, after which participants completed the test trials independently. At the end of the session, participants answered a brief post-experiment survey assessing how seriously they engaged with the trials, technical difficulties, and any other comments.

Results and Discussion

Accuracy scores were above chance for both the Consistent and Inconsistent trials, across both Short and Long duration conditions. There was a significant difference in accuracy between Consistent and Inconsistent trials in the Short duration condition but no significant difference for the Long duration condition. Accuracy is significantly higher when the cross is Consistent (i.e. placed on the agent) for the Short duration condition. D-prime scores were reliably above zero for both durations, indicating that participants were able to discriminate between trial types.



	Consistency	
Display Duration	Consistent	Inconsistent
40ms (short)	.91(.02)	.86(.04)
80ms (long)	.91(.03)	.89(.05)

Fig. 6. (Courtesy of Yiling Yun) Results as described.

Additionally, overall accuracy and sensitivity in our experiment were higher than those reported in Hafri et al. (2013) (see Fig. 4). One possible explanation for this is that using a visual cue like the cross to refer to the person may be processed faster than relying on a conceptual gender-based cue, which may require more mental processing. However, the higher accuracy could also be due to the absence of a mask in our design, which may have allowed participants to retain a clearer mental image of the trial stimuli.

Some future steps could include rerunning the experiment with a mask to assess whether participants can still perform the task under more perceptually constrained conditions. Alternatively, we could find a way to modify the question format that allows for using a mask, while still making it clear to the participants which person in the image we are referring to. This would allow us to make stronger claims about the perceptual mechanisms behind our results.

Future experiments in this study will test the effects of inversion and eventually, attempt to replicate Hafri et al.'s (2013) results with real-world news images.

References

Green, C., & Hummel, J. D. (2006). *Familiar interacting object pairs are perceptually grouped. 32*(5), 1107–1119. https://doi.org/10.1037/0096-1523.32.5.1107

Hafri, A., Papafragou, A., & Trueswell, J. C. (2013). Getting the gist of events: Recognition of two-participant actions from brief displays. *Journal of Experimental Psychology: General*, *142*(3), 880–905. https://doi.org/10.1037/a0030045

Papeo, L., Stein, T., & Soto-Faraco, S. (2017). The Two-Body Inversion Effect. *Psychological Science*, *28*(3), 369–379. https://doi.org/10.1177/0956797616685769

Papeo, L., Vettori, S., Serraille, E., Odin, C., Rostami, F., & Hochmann, J.-R. (2024). Abstract thematic roles in infants' representation of social events. *Current Biology*, *34*(18), 4294-4300.e4. https://doi.org/10.1016/j.cub.2024.07.081