Action Observation Influences Scene Perception in 18-Month-Olds

Maja Blesić (blesic_maja@phd.ceu.edu)
Department of Cognitive Science, Central European University, Quellenstraße 51, 1100 Wien, Austria

Ágnes Melinda Kovács (kovacsag@ceu.edu)
Department of Cognitive Science, Central European University, Quellenstraße 51, 1100 Wien, Austria

Abstract

Understanding how infants perceive real-world scenes and the type of information they rely on when recognizing different kinds of scenes remains unexplored. In this study, we aimed to investigate the relationship between action and scene information in infants. In a preferential looking paradigm, 18-month-olds were exposed to several trials in which they observed a human performing a given action and a subsequent simultaneous display of two scenes. One of the scenes was congruent with the action, representing the environment where the action is more likely to occur, whereas the other was incongruent. Results revealed a significant preference for looking at the congruent scene, accompanied by a longer first visit duration of that scene. Our findings show that the relation between action and scene information, previously reported for adults, is present already in infancy, suggesting a potential role of action information in shaping the construal of scene representation.

Keywords: action; scene; development; priming

Introduction

Visual scenes are meaningful units of human cognition, easily recognizable by observers at a categorical level, such as "a city" or "a kitchen", despite the complex information they contain (Oliva, 2005; Potter, 2012). However, the process by which we acquire the ability to recognize scenes so efficiently, remains largely unexplored in cognitive science. What kind of information can infants rely on when construing a scene representation?

Man-made scenes, in particular, present a distinctive challenge as they frequently encompass multiple objects, textures, and complex layouts (Malcolm, Groen & Baker, 2016). Relying solely on perceptual similarities may not be the optimal approach to building knowledge about scene categories. Scenes that share low-level similarities might belong to different categories (e.g., a park and a playground), as well as scenes with perceptual differences could belong to the same category (e.g., a kitchen in a restaurant and a kitchen at home). An alternative possibility is that infants are already grouping scenes based on more abstract commonalities, such as the types of activities and goals encountered in different environments.

Previous studies on objects have shown that action information plays a crucial role in infants' representation of artefacts (Booth & Waxman, 2002; Futó, Téglás, Csibra, & Gergely, 2010; Oakes & Madole, 2008). This knowledge about function and how an object is used influences the formation of object categories in 11-12-month-olds. In their study, Träuble and Pauen (2007) have shown that infants can categorize novel objects based on what an experimenter did with the object rather than based on the mere perceptual similarity of the objects. Importantly, infants seem not to necessitate direct experience to build expectations of object use but can rely on their sensitivity to learn from the observation of other's actions (Hunnius & Bekkering, 2014). For instance, at 6 months of age, although still not having vast experience of certain actions themselves, infants already possess some expectations about how some artefacts are typically handled, such as that a cup goes to a mouth and a phone to the ear (Hunnius & Bekkering, 2010).

Support for the idea that information about actions is relevant for the identification of scenes, comes from some recent studies based on adult data (Ciesielski, Webb & Spotorno, 2023; Greene et al., 2016). For instance, Greene and colleagues (2016) compared human categorization patterns with predictions made by object-, feature-, and function-based models, where function refers to the types of human activities that could take place in each scene (e.g., eating, hiking, socializing). A significant similarity was discovered between function-based distances and category distances, suggesting that a scene's function contributed the most to explaining why humans perceive two images to belong to the same category, surpassing models that relied on objects or visual features.

Here, building on the established role of function and action information in infants' representations of objects, we wanted to investigate the possibility that infants can use this information to build knowledge about different kinds of scenes that they are encountering daily. While we acknowledge that this is unlikely the sole factor contributing to infants' knowledge about scenes, we posit that it plays a pivotal role in building abstract representations of scenes, similarly to how information about object use helps build categories that go beyond perceptual similarity.

In a previous study, we found that at 18-months of age but not at 12-months of age, infants are able to categorize scenes in a preferential looking paradigm, when all the scenes belonging to the same category are accompanied by a the same pseudoword in the familiarization phase (Blesic et al., 2023). However, we do not know how infants represent categories of scenes in the first place. To test our hypothesis whether infants' representation of scenes can be construed

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

around actions, we designed a preferential looking study to test 18-month-olds. In each trial, infants were primed with an action (e.g., eating) that would occur more frequently in one of the two scenes presented subsequently (e.g., kitchen and bedroom). Crucially, the actions were performed without the use of objects in front of a homogenous background, allowing us to isolate the influence of action on infant scene perception. We hypothesized that if action and scene knowledge are related early in development, infants will show longer looking at the scene that is congruent with the previously presented action.

Methods

Participants

A sample size of minimum 23 participants (d = 0.80, alpha = 0.05, power = .95) was determined using G*Power 3.1.9.7 (Faul, Erdfelder, Buchner & Lang, 2009). We collected data from 28 participants, anticipating a dropout rate of approximately 25%. The final sample comprised 22 18-month-olds (N = 22, Mean age = 551 days, SD = 13.4 days, age range: 17 months 16 days to 18 months 18 days; 9 females). Six infants had to be excluded from the analysis due to fussiness (e.g., crying, N = 2), technical failure (N = 1), and an insufficient number of valid trials (N = 3). All infants received a small toy as appreciation for their participation after the experiment.

Materials

The materials consisted of videos of actions used as primes and images of real-world scenes used as test.

Videos of actions Videos of actions were used as primes in the experiment. The actions comprised: eating, sleeping, crossing the street, shopping, jumping playfully, and handwashing. These actions were specifically chosen by the authors because they can be commonly observed in different man-made environments in everyday activities of infants and therefore, they may be recognized by infants. The videos consisted of one person performing the action without the use of any objects or other context. To remove contextual scene information the videos were filmed against a green screen that was replaced with a gray background using a video editing software. Each action was enacted twice by two different protagonists, resulting in two videos for each action and twelve videos in total. All videos were 8 seconds long.

Images of scenes In the test, we used images of scenes representing both outdoor and indoor man-made scenes. The images belonged to six categories of scenes: kitchen, bedroom, city with a street junction, supermarket, playground, and bathroom. In each selected scene, one of the actions described above could typically take place. Each scene category was represented by two different images (e.g., playground_1, playground_2), for a total of 12 unique images. The images were then separated into two sets (Set 1 and 2), both containing one image from each scene category. Furthermore, images of scenes were organized in fixed pairs that would occur together at test. The saliency of the images showed together (e.g., kitchen_1, bedroom_1) was compared using the MATLAB Saliency Toolbox (Walther & Koch, 2006) to ensure that differences in low-level saliency would not predominantly guide infant's looking behavior at test. The Saliency Toolbox computes a saliency value of each pixel in the image relative to the surroundings. The averaged saliency values of the images occurring together at test were compared with a t-test. If a significant difference in saliency was discovered, one or both images were replaced until a suitable combination was found. See Table 1 for the saliency comparisons of the final selection of test images. Two of the images of scenes were captured by the authors, while the remaining ten were obtained from Wiki Commons and Flickr using a keyword corresponding to the name of the category of the scene. All images used in the study are licensed under Creative Commons Attribution 2.0 (CC BY 2.0) and have been resized to 750 x 500 pixels in size.

Table 1: Saliency comparison of test images

Set	Test Pair	Mean diff.	P-value	T value
1	Kitchen-Bedroom	0.0050078	0.49902	0.67612
	Bathroom-Supermarket	0.069075	0.56188	1.4502
	Playground-City	0.0031333	0.56188	0.58009
2	Kitchen-Bedroom	0.0069532	0.38374	-0.38559
	Bathroom-Supermarket	0.0035163	0.14715	-0.38559
	Playground-City	0.0029637	0.67755	-0.41585

Design

The experiment comprised 12 trials, each involving the presentation of prime videos and two test images. The duration of videos was 8 seconds, followed by an interstimulus interval of 0.5 seconds, and a subsequent display of two test scenes for 5 seconds (see Figure 1).

The combinations of action and scene categories were predetermined resulting in six unique combinations (see Table 2). Each image-test pair was presented twice: once with the action congruent with scene 1 and once with the action congruent with scene 2. Thus, a given scene of the pair functioned both as a distractor and a target, allowing us to discern the impact of action priming from potential preferences for a particular image within the pair. Each combination was replicated with a different protagonist and a distinct set of scene images, resulting in a total of 12 trials. The order in which trials were presented to each participant followed a pseudo-random sequence. The criterion for the sequence was that trials associated with the same scene-test

pair appeared at the greatest distance. Additionally, the target image appeared equally on the left and right sides. The side of target scene was not repeated more than twice consecutively on the same side.

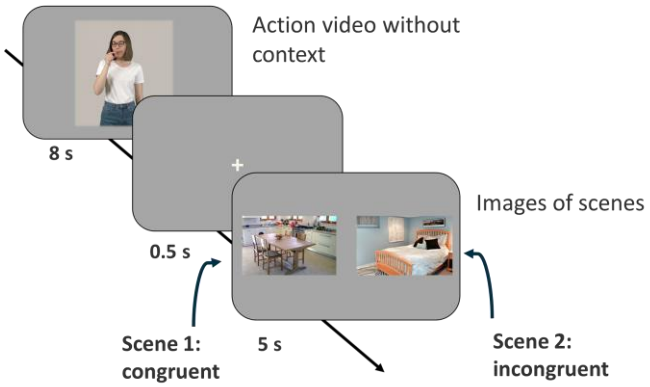

Figure 1: Trial Structure

Table 2: Combinations of Action-Type Scene Categories

Action-Type	Scene categories	
Eating Sleeping	Kitchen	Bedroom
Washing Shopping	Bathroom	Supermarket
Crossing Jumping	City	Playground

Procedure

During the experiment, infants were seated on the caregiver's lap approximately 60cm from the eye tracker's monitor. The caregiver was instructed to wear opaque glasses and refrain from interacting with the infant. The experimenter and a research assistant were in the control area throughout the experiment, which was adjacent to the testing area behind an occluding curtain. Before presenting the stimuli, a five-point calibration with sound was performed. After each trial, an inter-trial attention getter was shown to the infant and stayed as long on the screen as the infant turned back towards it.

Gaze data was recorded using a Tobii T60XL eye tracker (sampling rate: 60 Hz). The visual stimuli were displayed on a 24-inch monitor (resolution: 1920 × 1080px) on a gray homogeneous background. Sounds were played through loudspeakers located behind the monitor of the Tobii device. For the presentation of the stimuli and collection of gaze data Psychopy was used (Peirce et al., 2019). The experiment was terminated if the infant cried or became fussy. Ethical approval for the study was obtained from the local ethical review committee and the study was conducted according to the ethical rules and standards for psychological experimentation.

Data analysis

Exclusion criteria On average, infants gazed at the screen for 86% of the time (*SD* = 0.25) during the action presentation, 78% of the time during the test phase (*SD* = 0.29) and completed 11.66 trials on average (*SD* = 1.14) out of 12. Trials were excluded from analysis if the infant did not look at the screen for at least 60% of the time during the action presentation (prime) and 60 % of the time during the scene presentation in the test phase. Participants with fewer than six valid trials were excluded from the analysis. Out of 287 total number of trials in the sample, 224 trials passed the above defined criteria.

Measurements To analyze infants' preferential looking during the test, two areas of interest (AOIs) were defined to overlap in size and position with the test images (750 x 500 px). The proportion of looking at the congruent scene (propCongruent = TVDcongruent / (TVDcongruent + TVDincongruent), TVD: total visit duration within an AOI) was computed. TVD values for each AOI in each trial represented the time the average gaze position spent within the designated AOI. For each participant, the average TVD ratio of looking at the congruent scene was computed. Values higher than 0.5 indicated longer looking at the congruent scene, while values lower than 0.5 indicated longer looking at the incongruent scene. The averaged propCongruent as compared to the chance level (0.5) using one-sample t-test.

To further characterize the priming effect, we also analyzed the time infants spent looking when the first visited scene at test was congruent versus incongruent with the action. The first-visit duration (FVD) was defined as the interval between the time when the gaze first landed within one of the two AOIs corresponding to the test images and until it first landed outside of that AOI. For each participant, the average FVD for congruent and incongruent scenes was computed. The averaged FVD of congruent and incongruent scenes was compared with a paired sample t-test. To ensure that infants directed their initial look comparably towards both the congruent and incongruent scenes across trials, we conducted a binomial test. All statistical analysis was carried out in R (R Core Team, 2013).

Results

During the test phase, infants spent overall more time looking at the scene that was congruent with the action, compared to chance (M = 0.53, SD = 0.06), $t(21) = 2.56$, p = .018, 95% CI = [0.51, 0.57], Cohen's d: 0.55, BF10: 3.07. See the proportion of looking to congruent scene in Figure 2. Furthermore, the comparison of first visit duration confirmed our prediction, as the average FVD to the congruent scene (M = 1820.61 ms, SD = 725.14) was longer than to the incongruent scene (M = 1505.72 ms, SD = 551.07), $t(21) = 2.80$, p = 0.01, Cohen's d: 0.49, BF10 = 4.68. The mean difference in FVD between congruent and incongruent

scenes was 314.89 ms (95% CI [80.87, 548.92]). See the average first visit duration for the congruent and incongruent scene in Figure 3. Finally, infants' first looks were not more frequent towards the congruent scene at test (117 out of 224 first looks were directed to the congruent scene), $p = 0.5477$.

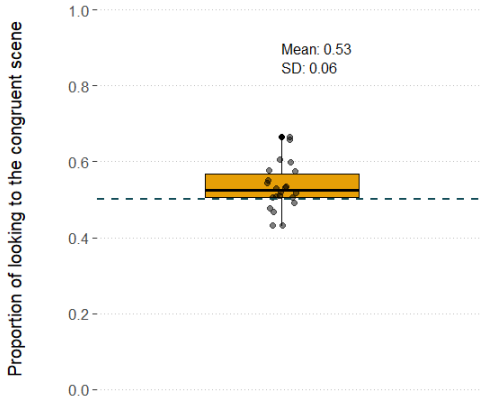

Figure 2: Preferential looking at the congruent scene during the 5 second test period

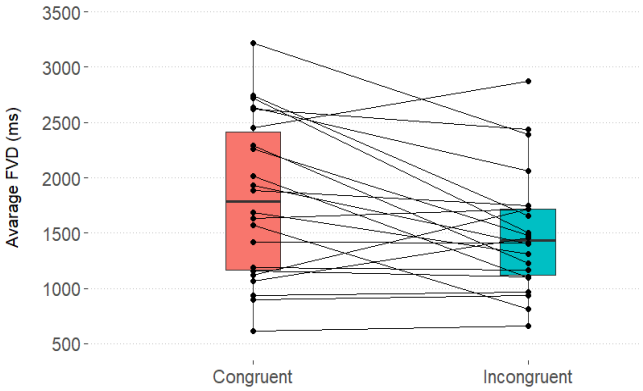

Figure 3: Average first visit duration for the congruent and incongruent scene at test

Discussion

In the current experiment, we tested whether observing agents performing actions affects infant looking behavior towards images of scenes in which the observed action is more likely to occur. We expected that if infants' representation of scenes is construed around the kinds of actions that the scene allows for, then we should observe a priming effect towards action-congruent scene compared to a scene that is unrelated to the action. Both the overall preferential looking at test, and the duration of the first visit to the scene, confirmed our hypothesis: infants' looking was longer for the congruent scene at test.

These findings suggest that action information is related to 18-month-olds knowledge about scenes, extending and contributing to the findings of infants' use of function-related information in building abstract representations. Indeed, previous studies (Booth & Waxman, 2002; Träuble & Pauen, 2007) have only addressed the importance of functional knowledge for the categorization of manipulable human artefacts. Our results are the first to suggest that infants can also use action information to learn about their broader surrounding visual units: scenes.

In a recent priming study with adults, it has been observed that adults' identification of scenes is enhanced by action words relative to object words (Ciesielski, Webb & Spotorno, 2023). Our findings with infants are in line with these results and suggest that the link between actions and scenes is present already in development. We hypothesize that this might play an important role in acquisition of knowledge about scenes.

However, it is crucial to acknowledge some limitations of the current study. While our paradigm provides valuable insights by revealing this action-scene relation for the first time in infants, it does not address its underlying nature. It could be argued that actions are not organizing infants' knowledge about scenes on a more abstract level, but that actions and scenes are merely associated due to their high statistical co-occurrence. Given that infants were seeing the decontextualized action videos for the first time as well as the specific scene images, it is more likely, however, that the more abstract relation between knowledge of scenes and actions was guiding their looking behavior. Whether any kind of action is a good candidate for construing an abstract representation of a scene and whether there is a difference between actions that are instrumentally involving more elements of the scenes or are related to a higher-level goal in the scene, is an open question we plan to address in our future research.

Conclusion

The present findings show for the first time that infants' knowledge about actions and scenes is related and that actions can prime infant's scene recognition, opening new avenues for future research about the nature of this interaction.

References

Booth, A. E., & Waxman, S. (2002). Object names and object functions serve as cues to categories for infants. *Developmental psychology*, *38*(6), 948.

Blesic, M., Hamilton, M., Blaser, E., Kaldy, Z., & Kovacs, A. (2023). Can infants categorize scenes?. In *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Ciesielski, K., Webb, A., & Spotorno, S. (2023). Mainly the actions: Functional knowledge has a primary role in understanding real-world scenes portrayed by either fine or coarse visual information. *Journal of Vision*, *23*(9), 5689-5689.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior research methods*, *41*(4), 1149-1160.

Futó, J., Téglás, E., Csibra, G., & Gergely, G. (2010). Communicative function demonstration induces kind-based artifact representation in preverbal infants. *Cognition*, *117*(1), 1-8.

Greene, M. R., Baldassano, C., Esteva, A., Beck, D. M., & Fei-Fei, L. (2016). Visual scenes are categorized by function. *Journal of experimental psychology. General*, *145*(1), 82–94.

Hunnius, S., & Bekkering, H. (2010). The early development of object knowledge: A study of infants' visual anticipations during action observation. *Developmental psychology*, *46*(2), 446.

Hunnius, S., & Bekkering, H. (2014). What are you doing? How active and observational experience shape infants' action understanding. Philosophical Transactions of the *Royal Society B: Biological Sciences*, *369*(1644), 20130490.

Malcolm, G. L., Groen, I. I., & Baker, C. I. (2016). Making sense of real-world scenes. *Trends in cognitive sciences*, *20*(11), 843-856.

Oakes, L. M., & Madole, K. L. (2008). Function revisited: how infants construe functional features in their representation of objects. *Advances in child development and behavior*, *36*, 135–185.

Oliva, A. (2005). Gist of the scene. *Neurobiology of attention*. Academic press.

Potter, M. C. (2012). Recognition and memory for briefly presented scenes. *Frontiers in Psychology*, *3*, 32.

R Core Team, R. (2013). R: A language and environment for statistical computing.

Träuble, B., & Pauen, S. (2007). The role of functional information for infant categorization. *Cognition*, *105*(2), 362-379.

Walther, D., & Koch, C. (2006). Modeling attention to salient proto-objects. *Neural networks*, *19*(9), 1395-1407.