Published in final edited form as:
J Exp Psychol Gen. 2013 August ; 142(3): 880–905. doi:10.1037/a0030045.

Getting the Gist of Events: Recognition of Two-Participant Actions from Brief Displays

Alon Hafri[a], **Anna Papafragou**[b], and **John C. Trueswell**[a]
[a]University of Pennsylvania
[b]University of Delaware

Abstract

Unlike rapid scene and object recognition from brief displays, little is known about recognition of event categories and event roles from minimal visual information. In three experiments, we displayed naturalistic photographs of a wide range of two-participant event scenes for 37 ms and 73 ms followed by a mask, and found that event categories (the event *gist*, e.g., 'kicking', 'pushing', etc.) and event roles (i.e., Agent and Patient) can be recognized rapidly, even with various actor pairs and backgrounds. Norming ratings from a subsequent experiment revealed that certain physical features (e.g., outstretched extremities) that correlate with Agent-hood could have contributed to rapid role recognition. In a final experiment, using identical twin actors, we then varied these features in two sets of stimuli, in which Patients had Agent-like features or not. Subjects recognized the roles of event participants less accurately when Patients possessed Agent-like features, with this difference being eliminated with two-second durations. Thus, given minimal visual input, typical Agent-like physical features are used in role recognition but, with sufficient input from multiple fixations, people categorically determine the relationship between event participants.

Keywords

action recognition; event cognition; event roles; scene perception; scene gist

People are quite adept at apprehending what is happening around them, even from a single glance or fixation. Most demonstrations of this ability come from experiments showing that even under brief exposures (sometimes less than 100 ms), individuals are able to categorize the type of scene depicted in an image, e.g. as a city, a park, or a mountain (e.g., Castelhano & Henderson, 2008; Fei-Fei, Iyer, Koch, & Perona, 2007; Greene & Oliva, 2009a, 2009b; Intraub, 1981; Oliva, 2005; Potter, 1975, 1976; Potter & Levy, 1969). In addition, it has been found that this ability mutually facilitates object recognition (Biederman, Mezzanotte, & Rabinowitz, 1982). This work has led to the development of theories of scene recognition, some of which hypothesize that our ability to rapidly recognize a scene is achieved by making use of both local and global image features that tend to correlate with scene categories, such as texture and spatial layout (Greene & Oliva, 2009b; Oliva & Torralba, 2001, 2007).

Yet apprehension of the world requires making additional, more complex categorizations from visual input. Most notably, individuals must be able to recognize events, and the roles that entities are playing in these events. Relative to the study of scene and object

Send correspondence to: Alon Hafri and John Trueswell, Department of Psychology, University of Pennsylvania, 3720 Walnut Street, Solomon Lab Bldg., Philadelphia, PA, 19104-6241, ahafri@sas.upenn.edu, trueswel@psych.upenn.edu.

recognition, little work has been done on rapid event recognition (see, e.g., Shipley & Zacks, 2008, and references therein). Research in this area has tended to focus on infants' development of knowledge about causation (Huttenlocher, Smiley, & Charney, 1983; Kersten & Billman, 1997; Leslie & Keeble, 1987; Muentener & Carey, 2010; Rakison, 2005) and about various dynamic events (Baillargeon, Li, Gertner, & Wu, 2011; Göksun, Hirsh-Pasek, & Golinkoff, 2010; Spelke, Phillips, & Woodward, 1995), the perception of biological motion (e.g., Blake & Shiffrar, 2007; Cutting & Kozlowski, 1976; Giese & Poggio, 2003; Grossman & Blake, 2002; Johansson, 1973; Lange, Georg, & Lappe, 2006; Lange & Lappe, 2006; Singer & Sheinberg, 2010; Troje, 2008; van Boxtel & Lu, 2011; Vangeneuden et al., 2011; Vangeneuden, Pollick, & Vogels, 2009), the segmentation of events from dynamic displays (Newtson, 1973, 1976; Zacks, Speer, Swallow, Braver, & Reynolds, 2007; Zacks, Tversky, & Iyer, 2001), and more recently the neural bases of human action representations (Kable & Chatterjee, 2006; Kable, Kan, Wilson, Thompson-Schill, & Chatterjee, 2005; Kable, Lease-Spellmeyer, & Chatterjee, 2002; Tranel, Kemmerer, Adolphs, Damasio, & Damasio, 2003).

Indeed, it is not known how much exposure to visual information is needed to recognize common events or actions, nor have researchers made explicit connections between scene recognition and event recognition, both of which involve making abstract categorizations of visual stimuli in ways that directly impact the interpretation of objects and entities that are often integral components of these categories. The work described below is intended to fill this gap by positing possible mechanisms by which people could extract information rapidly from an event scene, and by examining experimentally the speed and ability of people to recognize events and event roles.

What Is an Event?

A definition of an event must capture the many ways that humans conceive of changes over time, including fundamental conceptual categories such as 'move' but also complex categories such as 'give'. In this context, events are those conceptual categories that require reference to a location in time (Shipley, 2008) and whose temporal domain captures those qualities and conceptions of the world that predictably recur in such a way as to be relevant to the ways humans interact with their environment (e.g., Talmy, 2000). This may include inanimate entities changing or moving in time (e.g., a volcanic eruption, a tree blowing in the wind), or even animate entities performing an action independently (e.g., a girl dancing), the latter of which has been the primary focus of study on biological motion (e.g., Giese & Poggio, 2003).

Of interest to the present work is understanding how humans recognize from visual input those classes of events that involve common human interactions (e.g., 'hit', 'chase', 'push'). Recognition of such events is intertwined with object (entity) recognition in important ways, since this class of events involves participants playing particular roles that are likely to generalize across event categories (e.g., Agent, Patient, Recipient). For the purposes of this paper, we will focus on two-participant events involving just the conceptual event roles Agent and Patient. Agents are typically the originators of events, and Patients are typically the entities affected by those events. These event roles correspond to fundamental conceptual distinctions that have been assumed to be available to humans at a very young age (e.g., Goldin-Meadow, 1985; Golinkoff, 1975; Golinkoff & Kerr, 1978; Gordon, 2003). Furthermore, these roles map readily onto the verbal arguments of Agent and Patient in linguistic event descriptions (at least in the most typical cases; cf. Dowty, 1991). For example, in the description "A boy is pushing a girl", the boy is the one performing the action and the girl is the one being acted upon, so the noun phrase "a boy" is called the (linguistic) Agent whereas the noun phrase "a girl" is called the (linguistic) Patient.

Linguistic Agents and Patients, often referred to as “thematic relations”, reflect the way that the noun phrases in a linguistic description function with respect to the verb, just like conceptual Agents and Patients reflect the way non-linguistic entities participate in the corresponding event.

From Scene Gist Extraction to Event Recognition

Though it is not clear whether and how events can be rapidly understood from visual input, it is already well known that people can apprehend scenes and objects rapidly from the world. In Potter’s landmark studies of rapid scene comprehension (Potter, 1975, 1976; Potter & Levy, 1969), viewers could glean the semantic content of scenes from very briefly displayed images. In Potter (1976), subjects viewed Rapid Serial Visual Presentations (RSVPs) of scene images and had to detect a pictured or named target (e.g. “a picnic”) during the sequence or demonstrate recognition post-display. Even from displays lasting less than a typical single eye fixation (durations as short as 113 ms), subjects could detect and recognize targets, suggesting that the semantic content of the scene was extracted without attending to specific parts of the scene itself (though at least some attention is necessary to categorize scenes; see Cohen, Alvarez, & Nakayama, 2011; Intraub, 1984). In the scene recognition literature, many researchers have used the term *gist* to refer to the basic conceptual representation described above, or more specifically, to coarse information about a scene’s basic-level category, such as ‘picnic’, irrespective of the details of the scene (Oliva, 2005). Other work suggests that, just as with scenes, objects can be recognized with very little presentation time (Biederman & Ju, 1988; Biederman et al., 1982; Biederman, Rabinowitz, Glass, & Stacey, 1974), and even basic-level object names are activated without the need for fixation (Malpass & Meyer, 2010; Morgan & Meyer, 2005; Morgan, van Elswijk, & Meyer, 2008).

Given the prior work on rapid scene gist extraction and object recognition, one might expect event and role recognition to be possible even under short displays. Events, like scenes and objects, are core generalizations necessary for navigation and interaction in the world, and as such should be identified efficiently, and perhaps automatically. What is less clear is what kind of visual information would be used to achieve recognition of event categories and event roles. In the work that follows, we attempt to connect object recognition and scene gist extraction to event recognition in order to identify possible mechanisms by which event category and roles might be extracted, given limited visual information.

Event category recognition

We define the gist of an event scene as the basic-level event category, i.e., the type of action being performed by the figures in a scene, such as ‘punching’, apart from information about the particular entities that participate in the event (e.g., girl, boy, boxer, etc.). Work in the scene and object recognition literature suggests that there could exist perceptual properties of event scenes that facilitate rapid gist extraction in parallel to or before recognition of the entities in the scene. Indeed Oliva and colleagues have proposed that there are global properties of scenes (e.g., navigability, openness, depth) that allow for rapid categorization of a scene’s semantic category even before object recognition has been achieved (Greene & Oliva, 2009a, 2009b; Oliva & Torralba, 2001). They took as evidence the fact that briefly displayed images of one semantic category (e.g., forest) that happened to match the global property dimensions of another category (e.g., river) were classified incorrectly by both human observers and a computer classifier utilizing only the global properties (Greene & Oliva, 2009a). Whether or not the global properties specified by Oliva and colleagues are manifested in the human perceptual system, their work suggests that spatial and perceptual properties of scenes as a whole may correlate with the scene’s semantic category (see also Oliva & Schyns, 1997; Schyns & Oliva, 1994).

It seems plausible that, just as with natural outdoor scenes, event scenes would have coarse-grained spatial or global features that are correlated with the higher-level event category itself, and that people could use these features upon first view of an event scene to rapidly categorize it. For example, the spatial layout of the entities in an event scene may cue the observer in to the event category of the scene, perhaps apart from identifying the specific roles of the entities themselves. Of course it may be that this is less possible for some events than others, in particular if an event category's typical spatial layout is not consistent and unique ('punching', for example, may resemble many other event categories, while 'shooting' may have a more unique layout).

An alternative path to event category recognition is through role information: If an observer can assign roles rapidly to the entities in an event, such as Agent or Patient (the possibility of which is discussed in depth below), then this information may contribute to event categorization through a mutually constraining process. Both possible cues to event category, global scene features and event role information, may have varying degrees of independence and immediate availability, depending on the event scene. Even so, if either of these pathways to event categorization is available rapidly enough, then people should be able to extract an event scene's gist (i.e., the event category) with even less visual information than is available from an average fixation. We address this possibility in Exp. 1.

Event role recognition

Obviously, understanding the relationship between entities in an event scene is an essential component of knowledge about the event – indeed the interaction of event participants is part of how we operationalized the term *event* above. Yet few studies have examined the amount and depth of relational information extracted from scenes, whether spatial or causal, and how this interacts with scene recognition. In a pair of such studies, Biederman and colleagues placed objects in unexpected relational contexts (e.g., a tea kettle with a fire hydrant) and spatial contexts (e.g., a fire hydrant on top of a mailbox) and found that such violations of relations impaired scene and object recognition (Biederman, Blickle, Teitelbaum, & Klatsky, 1988; Biederman et al., 1982). Green and Hummel (2006) found that with brief displays of line drawings, objects placed in functional groupings (e.g., a water pitcher facing toward a glass) were recognized more readily than those in which there were no functional relationships (e.g., a water pitcher facing away from a glass). Additionally, in the first study to examine the extraction of relational content from rapidly displayed event scenes, Dobel and colleagues (Dobel, Gumnior, Bölte, & Zwitserlood, 2007) presented subjects with brief displays of line drawings of 'giving' and 'shooting' event scenes; afterwards, subjects had to name as many people and objects from the display as possible, as well as the event category. The authors found that people were better at identifying people and objects in coherent scenes as opposed to incoherent ones (i.e., scenes in which the shooter and the shootee were both facing away from one another), and concluded that people can apprehend event relations at durations as low as 100 ms. However, the fact that there were only two event categories and that subjects had to name scene entities without specifying roles leaves ample room for further exploration of recognition of event roles.

There is some evidence that at least the Patient role can be extracted early in scene viewing. In a study by Griffin and Bock (2000) that investigated the connection between event apprehension and the verbal description of events, different groups of subjects had to perform a variety of tasks while examining line-drawings of simple events, including description, free-viewing, and Patient identification, while their eye movements were recorded. The authors found that people in the Patient-identification task could look to the Patient as early as 300 ms after picture onset, thus showing rapid recognition of the Patient role. However, the Griffin and Bock stimuli were line drawings, which are likely to contain illustrative cues to agency and nonagency not present in photographic images.

Successful event role recognition would seem, then, to depend on computing the abstract causal relationship between entities. However, one possibility is that, in addition to the use of global perceptual features in classifying the event category, people may use any and all immediately available perceptual features that are probabilistically associated with abstract role categories to determine event roles even without having identified the action. For example, even without observing a particular event scene, if one were told that a person (e.g., a boy) is leaning towards someone (e.g., a girl) with his arms outstretched, it would seem probable that he is about to act on the girl. Knowledge of the specific event category is not needed to make this assertion confidently, as it is unlikely that that such a pose would indicate 'being acted upon'.

Some suggestions for what these perceptual features may be come from the work of Dowty (1991), who proposed a prototype theory of thematic role assignment in which the roles for a given verb have some subset of verb-general "proto-role" entailments (see also McRae, Ferretti, & Amyote [1997] for a more empirically derived proposal involving verb-specific features). For instance, prototypical Agents (or Proto-Agents) are characterized by (*a*) "volitional involvement in the event"; (*b*) "causing an event or change of state in another participant"; and (*c*) "movement relative to the position of another participant". Inversely, prototypical Patients (or Proto-Patients) are *(a)* not necessarily volitionally involved in the event; *(b)* "causally affected" by the event; and *(c)* "stationary" (Dowty, 1991, p. 572). We propose that visual event and role recognition depends upon visual features that arise from physical instantiations of the above Proto-Agent and Proto-Patient features (for our purposes, we will call the physical instantiations "event role features"). Candidate event role features would include:

1. Head orientation (toward vs. away from other event participant)
2. Body orientation (toward vs. away from other event participant)
3. Extremities (outstretched vs. contracted)
4. Body lean (toward vs. away from other event participant)

These features correspond to Dowty's entailments in the following way: How oriented the head and body are towards the other event participant (features 1–2) may correspond to the degree of volition of the event participant (entailment [a]). How outstretched the extremities are (feature 3) may correspond to the degree to which the event participant is able to causally effect change in the other participant (entailment [b]). And body lean (feature 4) may indicate degree of volition (entailment [a]) and direction and degree of movement (entailment [c]).[1] For all these event role features, the degree to which an event participant possesses them should be an indicator of Agent-hood, while a lack of the features should indicate Patient-hood.

We are not claiming that these visual features of event participants are independent in ways analogous to the independence of, e.g., color, texture, or motion in object perception. These event role features can be thought of as elements of the same thing, i.e., general body posture. Rather we call them features because they are likely to vary somewhat independently across events. In support of our proposal, there is evidence that body posture information can be rapidly integrated with other visual information as early as 115 ms (facial and bodily emotions: Meeren, van Heijnsbergen, & de Gelder, 2005), and electrophysiological recording has revealed body-selective neural responses in humans as early as 190 ms from stimulus onset (Pourtois, Peelen, Spinelli, Seeck, & Vuilleumier, 2007;

[1]Though the stimuli in our experiments are still-frame photographs, there is evidence that movement is commonly inferred from still images (Freyd, 1983; Kourtzi & Kanwisher, 2000; Urgesi, et al., 2007; Verfaillie & Daems, 2002).

Thierry et al., 2006). In addition, related features have been used successfully as the first-level components in some computer models of automated human action and interaction recognition (Park & Aggarwal, 2000, 2004). For example, Park and Aggarwal (2004) built a hierarchical Bayesian network for event recognition, in which the first step was deriving separate estimates for head, body, leg, and arm poses, followed by the integration of these with spatio-temporal domain knowledge into semantically meaningful interactions.

If the perceptual features of the entities in an event scene (features 1–4 above) are available rapidly enough to the human observer, then it is likely that people can categorize scene participants into their event roles with limited visual information. The ability to recognize the roles of participants in a briefly displayed event scene is investigated in Exp. 2, and the way in which the event role features contribute to this rapid recognition is explored in Exps. 3 and 4.

Interdependence of event category and event role recognition

One question that arises is whether and how much the extraction of the event scene gist might facilitate role recognition, or vice-versa. In the scene and object recognition literature, a large body of research supports the idea that extraction of the scene gist occurs rapidly and facilitates recognition of or memory for objects consistent with the scene gist (Biederman et al., 1982; Boyce, Pollatsek, & Rayner, 1989; Friedman, 1979; Hollingworth, 2006; Oliva & Torralba, 2007). According to this view, objects and their settings are processed interactively (Davenport, 2007; Davenport & Potter, 2004; but see Hollingworth & Henderson, 1998, 1999). If the event scene gist can be extracted even before higher-level visual features (i.e., shapes and objects), then the semantic information automatically available upon activation of the scene's gist could have a facilitatory effect on recognizing the roles in the event (Oliva, 2005). Alternatively, event scene gist extraction may depend on the relational information from entities in a scene, i.e. the event roles, and thus information about the event gist may become available only after recognition of event roles.

If such interdependence exists, one might expect that, for events in which the event category is hard to determine, establishing the event roles would likewise be difficult, and vice-versa. Below we will examine this relationship by comparing performance between Experiment 1, which probes knowledge of event category, with performance in Exp. 2, which probes knowledge of event roles.

Current Study

To investigate the extent to which viewers can extract an event's gist and event roles from a display lasting less than a single fixation, we use a modified version of the masked display paradigm of Davenport and Potter (2004), in which a single photographic image was displayed for a brief (80 ms) duration, followed by a scrambled mask. The masking effectively blocks visual processing of images, only allowing for additional processing of a conceptual memory of the masked images beyond the display durations (Potter, 1976). We use the same general trial structure here, but instead of free responses to a pre-display probe, subjects are presented with a forced-choice response task, e.g. "Was the boy performing the action?" A distribution of true "yes" responses and false alarms, as opposed to free responses (e.g., Dobel et al., 2007), will establish a baseline and allow us to determine whether the information probed is reliably extracted despite potential response bias.

As an improvement to previous work on rapid event recognition (e.g., Dobel et al., 2007; Gleitman, January, Nappa, & Trueswell, 2007; Griffin & Bock, 2000), we use naturalistic photographs instead of line drawings, as there is evidence that naturalistic scenes are processed differently and probably more efficiently than drawn stimuli (Davenport, 2007;

Davenport & Potter, 2004; Fabre-Thorpe, Delorme, Marlot, & Thorpe, 2001; Henderson, 2005; Henderson & Ferreira, 2004). In addition, we use a larger and more diverse set of two-participant event categories than those found in previous work (24 event categories across our experiments) to increase the generalizability of our findings to event recognition.[2]

In what follows, we first establish that event recognition is indeed possible from very brief displays by probing for event category information (Exp. 1, "Did you see 'pushing'?"). In a subsequent experiment, we confirm that likewise, event roles can be recognized from brief displays (Exp. 2A, Agent Probe: "Is the girl performing the action?"; Exp. 2B, Patient Probe: "Is the boy being acted upon?"; Exp. 2C, Sentence Verification Probe: "The boy is pushing the girl." True/False?). Then in Exp. 3, we explore whether performance on the first two experiments can be predicted by the degree to which event participants in our images possess the event role features we discussed above, looking for commonalities across events that could aid in rapid role recognition. Finally, in Exp. 4 we ask whether manipulation of the relevant role features within event does indeed have expected systematic effects on event role recognition under brief displays.

Experiment 1: Recognition of Event Category

We begin by examining whether event category information (e.g., 'pushing', 'chasing') can be extracted from a very brief visual exposure to an event lasting only 37 ms or 73 ms.

Methods

Subjects—Sixteen native English-speaking individuals participated in the experiment, comprising both University of Pennsylvania undergraduate students in an intro psychology course and other adults from the Penn community. Students received course credit and non-students received five dollars for participation.

Stimuli and apparatus—Targets consisted of 32 photographic images, each of two actors (one male, one female) engaging in an event that would typically be described using a transitive sentence (e.g., "A girl is pushing a boy"), as confirmed by a norming study (described below). Sixteen different transitive events were used to create the 32 images. All items involved an Agent and a Patient, though one item ('scaring') may be more appropriately classified as a Stimulus-Experiencer verb. There were two photographs for each event, an original and a role-reversed (gender-swapped) version, which were staged identically except that the roles of the boy and the girl were reversed (i.e., the second 'pushing' photograph was of the same boy and girl in the same poses, except with the boy pushing the girl instead of the girl pushing the boy). The left and right position of the actors and the left and right position of the Agent were counterbalanced, such that half of the images had a male on the right and half the images had an Agent on the right.

We wanted to make sure that we used images for which there is general agreement in what was depicted in the image, such that, for example, most people would agree that our pushing image depicted a 'pushing' event. Thus in preparation for the study, we had eight different actor pairs pose for each event in front of their own unique background, thereby creating multiple example image pairs for each event. The actors posed in a way that we judged to be most representative of each event. We then selected a single best actor pair for each event to be used in the experiment based on the results of a separate norming study, in which a

[2]In a recent chapter, Dobel, Glanemann, Kreysa, Zwitserlood, and Eisenbeiss (2010) describe the results of several experiments that attempt to extend the results of their earlier study (Dobel et al., 2007) to naturalistic event scenes and more event categories, although they did not investigate what kinds of features contribute to event role recognition, and specific details of the studies were not included.

different group of 32 undergraduates participated for course credit. The survey included multiple examples of each target event, plus 26 filler items that we judged would typically be described using intransitive sentences (e.g., “A boy and a girl are dancing”), which were only used in Exp. 2C. Subjects typed a simple one-sentence description for each image, from which we coded verb use. Then, for each target event, we selected the actor pair that generated the highest agreement in verb use. Synonyms counted toward the total, so for example, “yelling to” and “calling after” counted toward the same total, but passives (e.g., “the girl is being pushed by the boy”) did not. In one case we used the actor pair with slightly lower verb name agreement in order to maximize the number of times each actor pair would be present in our experiment. After selecting the target images, four actor pairs appeared in the target images twice, two pairs appeared three times, and the other two pairs appeared once. Proportion of verb name agreement had the following characteristics: for each image, it was over 50% (range 30% to 100%, $M = 81\%$, $SD = 14\%$); the difference in name agreement between Male- and Female-Agent versions of each event category did not exceed 41% (range 0% to 41%, $M = 11\%$, $SD = 10\%$); and Female-Agent versions had a higher name agreement score than Male-Agent versions 50% of the time, and equal name agreement 13% of the time. Throughout the rest of the paper, we will refer to each item by its normed verb label, e.g. “punching” or “kicking”. Example images can be found in Appendix A.

Each image was a 640×480-pixel color image (dimensions 19.2 cm × 14.4 cm), and subtended 20.4 degrees visual angle horizontally and 15.4 degrees vertically at a distance of approximately 54 cm (the average distance subjects sat from the computer screen). Masks were 20×20 blocks of scrambled images formed from a set of unused images, with an equal number of masks coming from every pair of actors.

Stimuli were displayed on a 17″ Dell P793 CRT monitor (diagonal 39 cm standard viewing size) at a resolution of 1024×768 pixels with a refresh rate of 85 Hz. The experiment was run in Windows XP using E-Prime experiment design software version 1.2.1.791 (Psychology Software Tools, Pittsburgh, PA), running on a Dell Precision M4400 laptop with 3.48 GB RAM, a 2.66 GHz dual processor, and an NVIDIA Quadro FX 770M card with 512MB video memory.

Procedure—Subjects were run individually in a dimly lit room. They sat at a computer monitor and were told they would see very briefly displayed photographs of people engaged in actions. They were to use the keyboard to answer a yes/no question that appeared after each picture. If they did not know the answer they were required to guess.

The trial structure appears in Figure 1, with each trial consisting of the following: A crosshair in the center of the screen for 413 ms, a blank screen for 198 ms, the target image for either 37 or 73 ms, and the mask for 245 ms. Both the target and the mask were framed by the rest of the screen in black. Following the display, subjects saw a sentence and had to press one of two buttons on the keyboard to answer either “yes” or “no” to the probe question, depending on whether the probe was consistent or inconsistent with the image. This trial structure and timing is based on past single-picture gist extraction studies reported in the literature (e.g., Davenport & Potter, 2004). Subjects could use either hand to respond in all experiments reported in this paper.

Probe sentences asked about the event category of the image and required a yes or no response (“Did you see ‘pushing’?”). The verb was either consistent or inconsistent with the image. For example, after seeing an image of a girl pushing a boy, a consistent probe would ask “Did you see ‘pushing’?” while an inconsistent probe would ask about a different event category, e.g. “Did you see ‘scratching’?” The consistent verb was always the one most

commonly used to describe the picture in the norming study. The inconsistent verb was chosen from the list of consistent verbs used for other stimuli. Though the degree of similarity of consistent and inconsistent event category pairings inevitably varied between items, a criterion for the pairings was that the body position of the scene participants in the experimental item image would be unlikely in an event described by the inconsistent verb, as judged by the experimenters (e.g., it would be unlikely that a 'punching' event would be misconstrued as a 'filming' event). In addition, no action described by the inconsistent verb was simultaneously occurring within the test image (e.g., in the 'lifting' event, where 'looking at' was the inconsistent verb, the participants were not looking at one another).

The visual angles between the center crosshair position and the event participants in each image were computed separately for the center of the head and the center of the torso of each participant, and were as follows: Agent head (range 3.6 to 8.5 degrees, $M = 6.0$, $SD = 1.3$); Patient head (range 3.0 to 8.6 degrees, $M = 6.1$, $SD = 1.6$); Agent torso (range 2.3 to 6.8 degrees, $M = 4.6$, $SD = 1.3$); Patient torso (range 1.3 to 6.4 degrees, $M = 4.1$, $SD = 1.4$).

List design—Two practice items were shown before the test trials. Both consisted of male-female pairs who were not used in the target stimuli performing actions typically described by transitive verbs but not used as experimental items ('tripping', 'spraying'). A stimuli list consisted of two blocks of 16 items each, all target items. Within each block, half the items had a short duration (37 ms) and half had a long duration (73 ms), and for each duration type, half the items were followed by a consistent probe and half by an inconsistent probe. Thus, within a block, items were equally divided among the four conditions (Short-Consistent, Short-Inconsistent, Long-Consistent, Long-Inconsistent). Agent gender (Male/Female) and Agent position (Left/Right) were also counterbalanced across conditions.

The second block was the same as the first except for the following changes. For each item, the test image was replaced with the image version showing the actors in opposite roles (e.g., if the Agent was male in the first block, the Agent was female in the second), and likewise the verb in the probe sentence was switched. Thus, for each target item the expected response (Consistent or Inconsistent) was different between blocks. The consistency for a given item varied between blocks to keep the frequency that a subject saw a given verb in the sentences constant (e.g. they saw the verb "punching" once after the 'punching' item and once after the 'scaring' item).

Each block had a different fixed pseudo-random order, with the following criteria: the same male-female actor pair could not appear on consecutive trials; across the two blocks an item had to be separated from its role-reversed version by at least seven intervening trials; and the same verb probe had to be separated by at least two intervening trials. Three additional stimuli lists were generated by rotating the items through each of the four conditions in a consistent manner, using a Latin square design. Reverse-order versions of these four lists were also generated.

Results and Discussion

The average proportions of correct responses for Consistent and Inconsistent trials appear in Table 1, along with 95% confidence intervals. Judgments were just slightly above chance for the Short Duration (37 ms) but well above chance for the Long Duration (73 ms), where chance is .50. Figure 2 presents these data in terms of d' (d-prime, a bias-free sensitivity measure), which is derived from the hit and false-alarm rates for each condition. For all subject and item means, mean hit and false alarm rates of 0 or 1 were approximated using the procedure standard in the literature (Macmillan & Creelman, 2005, p. 8). Zeros were replaced by 1/(2N), where N equals the maximum number of observations in a group, and

ones were replaced by 1-1/(2N). A perfect score by this approximation method would yield a d' value of 3.07 for subject and item means in Exps. 1 and 2, unless stated otherwise.

Throughout the paper, separate two-tailed *t*-tests were performed on subject and item means, which are called t_1 and t_2 respectively. Here these *t*-tests reflect one-sample *t*-tests, testing whether d' was reliably different from zero. As can be seen in the figure, subjects were able to extract information about event category at both the Short ($t_1(15) = 6.73$, $p < .001$, $d = 1.68$; $t_2(15) = 4.33$, $p < .001$, $d = 1.08$) and Long Duration ($t_1(15) = 21.4$, $p < .001$, $d = 5.35$; $t_2(15) = 13.3$, $p < .001$, $d = 3.33$).

In addition, separate analyses of variance (ANOVAs) on subject and item means (F_1 and F_2 respectively) were carried out. In Exps. 1 and 2, ANOVAs had the following factors unless noted otherwise: Duration (Short or Long); Agent Gender (Male or Female), List (1–4), and either List Order (Forward or Reverse; subject ANOVAs only) or Item Group (1–4; item ANOVAs only). Any significant effects of grouping variables (List, List Order, and Item Group) and their interactions are reported in footnotes.

The ANOVAs revealed a reliable effect of Duration ($F_1(1, 12) = 41.7$, $p < .001$, $\eta_p^2 = .78$; $F_2(1, 14) = 22.8$, $p < .001$, $\eta_p^2 = .62$), with Long durations yielding higher d' values. This indicates that subjects' ability to extract role information improves at longer display durations. It should be noted that there was no effect of Agent Gender (both Fs < 1), nor was there a Duration × Agent Gender interaction (both Fs < 1). Additionally, the visual angle between the center crosshair position and the head and torso positions of the event participants underwent Pearson tests of correlation with d' scores for each image to examine any effects of distance on performance. No significant correlations were found for this experiment or any others using this set of stimuli.

The results indicate that viewers can reliably extract event category information even at very short stimulus durations. That people are able to do so shows that the gist of an event is available to the viewer without the need for fixating specific entities in a scene and motivates our next set of experiments in which we ask whether event role information is among the properties that can be extracted from a very brief exposure to an event.

Experiment 2: Recognition of Event Roles

Here we modify the procedure of Experiment 1 to ask whether viewers can identify Agent and Patient roles of events from brief displays. In Experiment 2A we ask subjects to identify Agents, in 2B Patients, and in 2C we implicitly probe recognition for both roles via a sentence verification procedure.

Methods

Subjects—Sixty-four additional individuals participated in total (16 in Exp. 2A, 16 in Exp. 2B, and 32 in Exp. 2C). They had the same background as those in Exp. 1.

Stimuli—Target stimuli for all three experiments were the same as those used in Exp. 1. Additionally, 26 filler items were included in Exp. 2C, all taken from the norming study described in Exp. 1. These filler items depicted events involving two people that, in the norming study, were most typically described using intransitive sentences (e.g., "A boy and a girl are dancing"). With filler items included, in Exp. 2C, three pairs of actors appeared seven times, two pairs appeared five times, two pairs appeared four times, and one pair appeared three times.

Procedure—The procedure and experimental design were all the same as in Exp. 1, except that the probe sentence now asked about the Agent or Patient of the event. Exps. 2A and 2B did this without labeling the event itself. Subjects in 2A were asked about the Agent, i.e., "Is the boy performing the action?" or "Is the girl performing the action?", while subjects in Exp. 2B were asked about the Patient, i.e., "Is the boy being acted upon?" or "Is the girl being acted upon?". Exp. 2C was like 2A and 2B, except the probe was no longer a question. Instead, subjects saw a sentence and had to respond whether the sentence was consistent or inconsistent with the image. Consistent target sentences conveyed the Agent and Patient roles in a way consistent with the image (e.g., "The boy is pushing the girl") whereas inconsistent sentences reversed the Agent and Patient roles (e.g., "The girl is pushing the boy"). This kind of probe would discourage subjects from focusing exclusively on one role or the other when preparing to view an image.

Probes for filler items used intransitive sentences (e.g., "The boy and the girl are dancing"), with the inconsistent version using an incorrect verb (e.g., "The boy and the girl are walking"), much like the manipulation in Exp. 1. The addition of these fillers further discouraged a strategy of focusing on event roles, as incorrect filler probes could only be rejected based on a mismatch with the event category.

List design—The lists were the same as in Exp. 1, except with seven of the 32 target trials swapped with one another within-block for Exps. 2A and 2B, and 11 of the 32 target trials swapped with one another within-block for Exp. 2C. The lists still met the same pseudo-random order criteria as in Exp. 1 (no consecutive actor pairs and at least seven intervening trials between images of the same event item across blocks). Experiments 2A and 2B used the same practice trials, whereas Exp. 2C used just one practice trial, which was an extra filler image not used in test trials.

Like in Exp. 1, lists consisted of two blocks, such that the recurrence of a test image in Block 2 was replaced with the image version showing the actors in opposite roles. Unlike Exp. 1, the consistency of the probe remained the same for an item across blocks. In particular, in Exps. 2A and 2B, the probe asked about the opposite gender character. For example, a boy-punch-girl image with a probe "Is the girl performing the action?" (Inconsistent) in Block 1 became a girl-punch-boy image with a probe "Is the boy performing the action?" (Inconsistent) in Block 2. Thus unlike in Exp. 1, the expected response for each item was the same (Consistent or Inconsistent) between blocks but both the image and the probe had reversed roles across blocks. Likewise in Exp. 2C, the probe reversed roles for an item across blocks, such that the probe would have been, e.g., "The girl is pushing the boy" (Inconsistent) in Block 1 and "The boy is pushing the girl" (Inconsistent) in Block 2. Finally, in Exp. 2C, filler trials were randomly intermixed with targets in different orders across the blocks.

Results and Discussion

Experiment 2A: Agent probe—Accuracy for the Agent role probe was above chance at both the Short (37 ms) and Long (73 ms) Duration (see Table 1), suggesting that subjects could extract Agent information from very brief displays.

Figure 2 presents these data in terms of d'. The d' value was reliably above zero at both the Short ($t_1(15) = 8.81$, $p < .001$, $d = 2.20$; $t_2(15) = 5.07$, $p < .001$, $d = 1.27$) and Long display durations ($t_1(15) = 15.8$, $p < .001$, $d = 3.96$; $t_2(15) = 7.22$, $p < .001$, $d = 1.80$). In addition, ANOVAs on subject and item means revealed a reliable effect of Duration ($F_1(1, 12) = 50.9$, $p < .001$, $\eta_p^2 = .81$; $F_2(1, 14) = 7.43$, $p < .05$, $\eta_p^2 = .35$) with Long durations yielding higher *d'* values. The ANOVAs showed no effect of Agent Gender ($F_1(1, 12) = 1.91$; $F_2(1, 14) = 2.31$) nor a Duration × Agent Gender interaction ($F_1(1, 12) = 1.41$; $F_2(1, 14) < 1$).

Experiment 2B: Patient probe—Like the results for Agent probes, accuracy for the Patient role probe was also above chance at both the Short (37 ms) and Long (73 ms) Duration (see Table 1), suggesting that subjects could extract Patient information from very brief displays.

Figure 2 presents these data in terms of d'. Subjects were able to extract information about the Patient role at both the Short ($t_1(15) = 4.22$, $p < .001$, $d = 1.06$; $t_2(15) = 4.18$, $p < .001$, $d = 1.04$) and Long Duration ($t_1(15) = 15.4$, $p < .001$, $d = 3.86$; $t_2(15) = 10.2$, $p < .001$, $d = 2.55$). In addition, ANOVAs on subject and item means revealed a reliable effect of Duration ($F_1(1, 12) = 23.3$, $p < .001$, $\eta_p^2 = .66$; $F_2(1, 14) = 23.6$, $p < .001$, $\eta_p^2 = .63$) with Long durations yielding higher d' values.[3]

Somewhat surprisingly, the subject ANOVA revealed an effect of Agent Gender on performance ($F_1(1, 12) = 5.69$, $p = .03$, $\eta_p^2 = .32$), as well as a Duration × Agent Gender interaction ($F_1(1, 12) = 8.28$, $p = .01$, $\eta_p^2 = .41$), with Male-Agent items yielding higher d' values than Female-Agent items at the Short Duration (paired t-test, adjusting for multiple tests by Holm's method; $t_1(15) = 3.23$, $p = .006$, $d = .81$).[4] Similar effects were not observed in the analysis of item means. Such a pattern could suggest that subjects are influenced by stereotypical gender knowledge about events (e.g., males are more likely to be Agents in two-person interactions), and is the first hint that perceptual features, in this case gender, may be used to rapidly identify roles. However, we suspect that this effect is being driven by relative size of event participants, such that larger participants are more likely to be acting upon smaller participants. Indeed males tended to be larger than females in our stimuli. We return to this issue in Exp. 3, where we explore in more detail how certain features might contribute to role recognition.

Experiment 2C: Sentence verification probe—Like the results for Agent and Patient probes, accuracy for the sentence probe was also above chance at both the Short (37 ms) and Long (73 ms) Duration (see Table 1), suggesting that subjects could extract role information from very brief displays even when they were not explicitly asked about event roles at all.[5]

Figure 2 presents these data in terms of d'. For item means, a perfect d' score by the standard approximation method would be 3.73 instead of 3.07, due to there being a greater number of subjects in this experiment compared to Exps. 1, 2A, and 2B. As can be seen in the figure, d' performance was reliably above zero at both the Short ($t_1(31) = 4.55$, $p < .001$, $d = 0.80$; $t_2(15) = 4.30$, $p < .001$, $d = 1.07$) and Long display durations ($t_1(31) = 11.2$, $p < .001$, $d = 1.98$; $t_2(15) = 11.1$, $p < .001$, $d = 2.77$). In addition, ANOVAs on subject and item means revealed a reliable effect of Duration ($F_1(1, 28) = 63.1$, $p < .001$, $\eta_p^2 = .69$; $F_2(1, 14) = 69.6$, $p < .001$, $\eta_p^2 = .83$) with the Long Duration yielding higher d' values. The ANOVAs revealed no effect of Agent Gender (both Fs < 1) nor a Duration × Agent Gender interaction (both Fs < 1).

Although the results show that subjects could extract role information even with the subtle manipulation of this experiment, d' values were lower than those in Exps. 2A and 2B. This

[3]A Duration × List Order × Agent Gender interaction was found on subject means ($F_1(1, 12) = 5.87$, $p = .03$, $\eta_p^2 = .33$).

[4]Cohen's d for paired t-tests was calculated using the mean of the differences divided by the standard deviation (*SD*, or s) of the differences. For all independent sample comparisons, we used the difference of the means divided by a pooled *SD*, equivalent to Hedges' g and given by the following formula: $s_{pooled} = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$.

[5]For Exp. 2C, due to a syncing error between the experimental setup and monitor refresh, 15% of Short Duration and 34% of Long Duration trials lasted for 25 ms and 60 ms, respectively. Since the display times were shorter as opposed to longer than the intended duration, we did not deem it necessary to rerun the experiment.

is perhaps because subjects in Exp. 2C did not have the advantage of knowing in advance which aspect of the event would be probed by the test sentence (though veridical event category information for target items was included in the probe itself, which could in principle have aided conceptual processing of the stimulus post-mask). Indeed, separate ANOVAs on subject means comparing Exp. 2C to Exps. 2A and 2B (with Experiment as a factor) confirmed that subjects performed worse in 2C (vs. Exp. 2A (Agent probe): $F_1(1, 40) = 5.94$, $p = .02$, $\eta_p^2 = .13$; vs. Exp. 2B (Patient probe): $F_1(1, 40) = 3.96$, $p = .05$, $\eta_p^2 = .09$).[6]

Correlations between event category (Exp. 1) and event role (Exps. 2A – 2C) probes—As discussed in the Introduction, it is possible that event role and event category recognition interact with each other in a mutually constraining fashion. If this were so, we would expect performance on Exp. 1 (event category probe) to be highly related to performance on Exps. 2A–2C (event role probes). If event category and role information can be extracted independently, we would expect performance on Exp. 1 to be unrelated to performance in Exps. 2A–2C. Finally, if subjects were extracting role information in a similar way across the role probe experiments (2A–2C), we would expect performance to be related across these experiments.

To investigate this, we calculated Pearson's correlation coefficient between the mean d' values for items at the Short Duration for each experiment (we did not perform these tests on the Long-Duration d' values, as most were at or near ceiling). Subject performance on items in all of the experiments probing about role correlated positively with each other, significantly so or close to it (Exps. 2A and 2B: $r = .58$, $p = .02$, $n = 16$; Exps. 2A and 2C: $r = .60$, $p = .01$, $n = 16$; Exps. 2B and 2C: $r = .37$, $p = .16$, $n = 16$). This suggests that subjects used similar information to perform these tasks. However, the results also indicate that event category and event role identification are at least partially independent. Performance on items in the experiment probing about event category (Exp. 1) did not correlate with performance in the experiments probing about role information (all ps > .26). We will discuss the implications of this in the General Discussion.

Experiment 3: Event Role Features and Role Recognition

In Exps. 1 and 2, subjects were able to extract both event category and event role information even from very short display durations. The lack of correlations in performance between a task asking about event category information (e.g., "Did you see 'pushing'?") and event role information (e.g., "Is the girl performing the action?") suggests that subjects used different information for recognition of event categories and event roles and/or that the processes are partially independent. Correlations among the event role experiments themselves suggest that subjects used similar information to accomplish these tasks.

Here we offer a preliminary exploration of the visual cues subjects used to identify event roles, by asking whether variation in performance in Exps. 1 and 2 can be accounted for by the presence or absence of the event role features we proposed in the Introduction as being indicators of Agent- or Patient-hood, namely head direction, body direction, outstretched extremities, and leaning in. Indeed, there was significant variation in performance on extracting event role and category information, especially at the Short (37 ms) presentation duration, with subjects performing extremely well on some events (e.g., 'brushing', 'scratching') and quite poorly on others (e.g., 'biting', 'pulling') (see Appendix B, which lists item d' performance at the Short Duration separately for each experiment). In addition, the fact that item d' performance only correlated across the event role experiments (Exps.

[6]ANOVAs on item means were not carried out since the d' approximation for item means in this experiment was different from the others due to the difference in number of subjects in each experiment, 32 vs. 16.

2A–2C) and not the event category experiment (Exp. 1) suggests that it may be possible to extract event roles without the event category, and vice-versa. As such, we predict that the presence or absence of event role features will explain variation in performance only in the role probe experiments, and not in the verb probe experiment.

To test these predictions, we asked additional subjects to rate the Agents and Patients in our stimuli for the degree of presence of these event role features, and then we used these feature scores to model variation in the item d' scores from Exps. 1 to 2. With these post-hoc correlations in hand, we will then directly manipulate these features in Exp. 4.

Methods

Subjects—Eight additional undergraduates with the same background as those from Exps. 1 and 2 participated for course credit.

Stimuli—The images were the same as the target photographs from Exps. 1 and 2.

Procedure—Subjects completed an online survey. For each trial they viewed a photograph and were asked to rate the Agents and Patients along the following physical features (the text used for survey questions are enclosed quotation marks, where the first instance of "person" was replaced by either "person performing the action", i.e. the Agent, or "person being acted upon", i.e. the Patient):

1. Head-Facing: "The head of the person is facing towards the other person"
2. Body-Facing: "The body of the person is facing towards the other person"
3. Extremities: "The arms or legs of the person are outstretched towards the other person"
4. Leaning: "The person is leaning in towards the other person" [7]

Ratings were on a Likert scale of 1 (strongly disagree that the event participant possesses the feature) to 7 (strongly agree that the event participant possesses the feature). Each subject rated either the Male- or Female-Agent version of each event category, and viewed the images in a random order, one per survey page. On each image page, either the Agent feature questions came first or the Patient feature questions did, such that in one half of an image page, subjects rated the Agent's features, and in the other half, the Patient's. This order (Agent or Patient features first) was the same for all images within subject, and was counterbalanced across subject. Mirror-image versions of each image were also included in the surveys, but their results are not reported here.

Results and Discussion

Ratings were normalized within subject such that the ratings from 1 to 7 were transformed to *z*-scores centered at zero with a standard deviation of one. Median ratings were calculated for each feature and appear in Table 2. Scores higher than zero indicate that a particular feature rating was higher than average, and scores below zero indicate a lower than average rating. Feature distributions may be visualized in Figure 3, which shows a histogram of the normalized Agent and Patient ratings for each of the four features overlaid on top of one another.

[7]Participants were also asked whether the person was displaying negative emotion and about the proximity of the two participants to one another, but we will not discuss these features further (Agents and Patients did not significantly differ with respect to expressed emotion).

As stated in the Introduction, we propose that possession of these features is an indication of Agent-hood, whereas absence of these features is an indication of Patient-hood. Indeed, median feature scores for Agents were all positive while median feature scores for Patients were all negative; Agents tended to be facing and leaning toward the other participant, with extremities outstretched, whereas Patients tended to be facing and leaning away from the other participant, with extremities not outstretched. Nevertheless, there was a reasonable amount of variation in feature ratings across items (especially for Patient scores), as can be seen in the range and *SD* values, suggesting that some items may have better feature support for extraction of event information than others. A table of the feature difference scores for each item and the corresponding d' performance at the Short Duration for each experiment can be found in Appendix B.

Based on these feature scores, we wanted to derive a single score for each image that would reflect the degree to which the features present in an image would allow an observer to distinguish the Agent from the Patient. To do this we summed the Agent and Patient feature scores separately and then took the difference between these sums. A large positive difference score would indicate that the Agent possessed more event role features than the Patient, a score of zero would indicate little difference between Agent and Patient, and a negative score would actually indicate that the Patient rather than the Agent possessed the event role features that are consistent with Agent-hood.

Table 3 presents the results of a series of multilevel models that relate the Sum Feature Difference Score to item d' scores separately for each experiment, as well as all role experiments combined. As predicted, we found evidence that this feature difference score partially predicts variation in performance on the role recognition experiments (Exps. 2A–2C) but not the event category recognition experiment (Exp. 1).

Despite the significantly (or marginally) better fits that most of these models demonstrate over the null model (i.e., a model with no fixed effects) in the role experiments, a few caveats of our exploration should be noted. First, the rating scores of some features were highly correlated (specifically Head-Facing and Body-Facing: $r = .67$, $p < .001$, $n = 32$; and Head-Facing and Extremities: $r = .51$, $p = .002$, $n = 32$). This means that such features are artificially magnified in the Sum Feature Difference Score. Also, summing the scores assumes that each accounts for an equal amount of d' score variation, but some features are likely to have greater explanatory power than others.

To partially address these issues, we also related d' performance to each individual feature difference score. Table 4 presents the chi-square values and significance patterns for each feature model as compared separately to the null model. The significance patterns are similar to the Sum Feature Difference Score results, such that only feature score models of the data for the role experiments show reliable or marginal improvement over the null model, while none of the feature models shows improvement over the null model for the event category experiment.

We must note that these results do not imply that event role features are not used in some way to recognize event categories. Indeed it is likely that there are characteristic combinations of event role features that are at least partial indicators of various event categories. However, it does suggest that the simple degree of presence of these features (singly or in combination) is unlikely to aid in recognition of event categories.

Size and Agent Gender—It is possible that gender or relative size of event participants could be contributing factors to rapid role extraction. Recall that in Experiment 2B (Patient probe), an effect of Agent Gender was found: people extracted the event role better at the

Short duration in the Male- vs. Female-Agent images. In addition, in our stimuli, males tend to be larger than females. We wanted to see what improvement, if any, Size or Agent Gender could offer over and above the event role features in predicting item d' performance. To do this, we measured the size difference (in pixels, head to foot) between Agents and Patients in each image, and we compared multilevel models that included the Sum Feature Difference Score to models that had either Size or Agent Gender as an additional fixed effect. The results show that there is only scattered improvement in adding either Size or Agent Gender to the models. Marginal or significant improvement for Size over the Sum Score is seen only in Exp. 2A (Agent probe: $\chi^2(1) = 4.91$, $p = .03$) and in all role experiment data combined ($\chi^2(1) = 4.26$, $p = .04$), and for Agent Gender only when all role experiment data are combined ($\chi^2(1) = 3.61$, $p = .06$). Such a finding suggests that gender and relative size of event participants were not major contributors to subject performance.

Finally, Size and Agent Gender appear to covary, as male actors tended to be larger than female actors (eight of 16 times in Male-Agent images, and 13 of 16 times in Female-Agent images). In fact, apart from Size, there is little difference between the Male- and Female-Agent images, which are essentially the same in terms of the feature difference scores (paired *t*-tests: all *p*s > .11). Such a lack of difference is to be expected, as we intended the differences between Male- and Female-Agent depictions to be minimal when posing the photographs.

Experiment 4: Manipulation of Event Role Features

Here we conduct one final gist extraction experiment that directly manipulates the event role features of participants in order to test potential effects on event role recognition. Although the modeling in Exp. 3 supports such a link, a serious concern is that the analysis was post-hoc and only accounts for variation between different events (e.g., 'hitting' vs. 'pulling'), leaving open the possibility that we are describing inherent differences between events with regard to role recognition (e.g., it may be easier to tell a 'hitter' from a 'hittee' than a 'puller' from a 'pullee'). Under our account, however, one would expect that direct manipulation of event role features within a single event would yield the same result. We are also interested in generating better-controlled stimuli in which the Agent and Patient are otherwise physically identical, except for the manipulated event role features. To accomplish this, rather than using a man and a woman as event participants, who naturally differ, for instance, in relative size and gender, we use here identical twins only distinguishable by shirt color. The predictions are that we should be able to replicate the results of Exps. 2A and 2B (using additional events and completely new images) and, via feature manipulation, observe a causal link between event role recognition and event participants' possession of event role features.

Since we constructed the photographic stimuli for Exps. 1 and 2 using our best judgment rather than by sampling random instances of events in the world, an important question concerns knowing to what extent Agents and Patients actually possess the event role features in prototypical instances of two-participant causative events. To this end, we conducted a brief "mental imagery" experiment in which subjects ($N = 14$) told us, for prototypical instances of a larger set (33) of two-participant event categories, to what extent the Agent and Patient possess the event role features. The distributions of Agent and Patient features as imagined by these subjects are very similar to those of the Agents and Patients in our photographs (Exp. 3), as can be seen in Figure 4, suggesting that our posing of the actors was consistent with the way in which people would imagine prototypical instances of those events taking place.

However, this mental imagery norming experiment also shows that Agents are quite uniform in their possession of event roles features across events, whereas the degree to which Patients possess the features varies considerably across (and even within) events. As such, in the current experiment, we keep the Agent consistent within event while manipulating the extent to which Patients possess "Agent-like" features. Thus, in this experiment we compare two event types for each event: one we label "Prototypical", and the other "Bad-Patient", in which the Patient possesses Agent-like features.

Methods

Subjects—Eighty native English-speaking individuals participated in the experiment, comprising both University of Pennsylvania and University of Delaware undergraduate students in introductory psychology courses. Subjects received course credit for participation. Sixty-four participated in the primary brief-display experiment and 16 served as controls.

Stimuli and apparatus—Targets consisted of 48 photographic images. Images featured two male identical twin actors (age 29) engaging in an event that would typically be described using a transitive sentence. The actors had identical haircuts and facial hair styles, and wore the same clothing (similar jeans and sneakers, same shirt style), except that the color of one's shirt was blue, and the other's orange-red. The colors were distinct, even to color-blind individuals. Thus the two actors were essentially indistinguishable other than by shirt color, which was the manipulated factor in the post-stimulus probe (as gender was for Exp. 2). In all images in which an instrument was part of the event (e.g., 'stabbing'), both Agent and Patient held duplicates of those objects (e.g., both actors held identical knives). The actors were photographed in front of a plain light-blue background and dark floor, and the image backgrounds were manually postprocessed to a uniform brightness level.

Image size was the same as in Exps. 1 and 2, and masks were again composed of 20×20 blocks of scrambled images formed from a subset of unused images from the same photo set. Stimuli were presented on the same display as in the earlier experiments. The experiment was run in Windows XP using E-Prime experiment design software version 2.0.8.90, running on a Dell Latitude E6410 laptop with 3.24 GB RAM, a 2.67 GHz processor, and an NVIDIA NVS 3100M card with 512MB video memory.

Creation and selection of photographic stimuli: The 48 target images consisted of 24 distinct two-participant events, each with two versions (two event types). In the "Prototypical" version, the actors posed in a way that was prototypical for the event in question, whereas in the "Bad-Patient" version, the Agent posed identically to the first version, but the Patient now posed such that he possessed Agent-like features.

We selected these images from a larger set of photographs specifically generated for this experiment. In particular, we staged and photographed 33 events, corresponding to the ones found in the mental imagery norming experiment described above. For each Prototypical event depiction, we used the Agent and Patient feature ratings from the mental imagery norming results as a reference for the pose (e.g., the Agent in the prototypical 'punching' event was rated as having his or her extremities outstretched, so he posed as such; the Patient was not, so he posed accordingly). For the Bad-Patient event depiction, the Agent was staged identically to the Agent in the Prototypical version, but the Patient was manipulated such that he would possess all four of the Agent-like features (barring five single exceptions out of the 96 combinations of event category and feature). For most event categories, we staged several versions of both event types that differed in what features the Agent and Patient had.

As with the stimuli in Exps. 1 and 2, we wanted to ensure that we indeed depicted the intended events in our photographs. However, this experiment necessitated a more particular vetting of the stimuli, as it was crucial that there be no difference in agreement on the depictions between the Prototypical vs. the Bad-Patient event types. Therefore we conducted an image description norming survey ($N = 24$), in which differently staged versions for each of the 33 event categories were included to increase the chance that we would find both Prototypical and Bad-Patient versions that were described equally frequently as the intended event category. Subjects provided both a single verb and a full sentence to describe each event scene. As in the previous norming survey, synonyms were counted, so for example, "yelling to" and "calling after" contributed toward the same event category total. Each subject only saw one Agent Color/Location combination for each staged version of an event category. Name agreement scores were collapsed across Agent Color and Agent Location.

Example staged versions of event categories were only considered for selection in the current experiment if name agreement was higher than 50% for both Prototypical and Bad-Patient versions, and if the difference between name agreement for the Prototypical and Bad-Patient versions was 25% or less. After exclusion, 24 event categories remained. For each event category, the staged version with the highest name agreement score was chosen.

Next we had to select the Agent Color/Location combination for each event category such that name agreement scores would be balanced across all conditions. To this end, the 24 event categories were pseudo-randomly assigned to eight item groups of three event categories each, and then within item group, three different Agent Color/Location combinations were assigned, in order that the four combinations of Agent Color and Location appeared equally among all event categories. This process was performed such that name agreement scores were fundamentally the same between Prototypical and Bad-Patient versions for all factors manipulated in this experiment (discussed below). Indeed the maximum name agreement difference between the two event types (Prototypical or Bad-Patient) along any factor was just 5.6%. Example images can be found in Appendix C.

Event role features of new stimuli: Our intended depiction of the features in the two event types was confirmed by eight additional undergraduates, who rated the two event participants in each image along the event role features on a Likert scale of 1 to 7, as in Exp. 3. Agents for both event types resembled the typical Agent from the mental imagery experiment, in that they were facing and leaning in towards the Patient, with extremities outstretched. Patients in the Prototypical stimuli also followed the patterns of the typical Patient from the mental imagery experiment, in that they were not facing or leaning in towards the Agent, nor did they generally have extremities outstretched.

However, Patient features in the Bad-Patient stimuli were almost indistinguishable from the Agent features. Indeed the feature difference scores (i.e., Agent feature ratings minus Patient feature ratings) in the Bad-Patient event type were at or close to zero for every role feature, with only the Extremities difference score being significantly different from zero ($t(23) = 2.99$, $p = .007$), most likely because the Patient's extremities were not quite as outstretched as the Agent's. Additionally, every Patient feature was significantly different between the Prototypical and Bad-Patient event types (ps $< .001$). Thus we can be confident in three aspects of our stimuli: 1) Agents are effectively the same across the Prototypical and Bad-Patient event types; 2) in the Bad-Patient event type, Patients are similar to the Agents; and 3) the Patient's features differ significantly between event types.

Procedure—The procedure and trial structure were the same as in Exps. 2A and 2B (Agent and Patient probes), except that instructions included an additional sentence indicating that the events all involved two boys, one in a blue shirt and one in a red shirt.

Probe Type (Agent or Patient) was varied between subjects, such that each subject was asked either about the Agent ("Is the blue boy performing the action?") or the Patient ("Is the blue boy being acted upon?"). The probe was either consistent or inconsistent with the image (i.e., after a Blue-Agent item, a Consistent probe would ask "Is the blue boy performing the action?" while an Inconsistent probe would ask "Is the red boy performing the action?").

The visual angles between the center crosshair position and the event participants in each image were computed separately for the center of the head and the center of the torso of each participant, and were as follows, listed separately for Prototypical and Bad-Patient event types: Agent head (Prototypical: range 1.8 to 9.6 degrees, $M = 5.8$, $SD = 1.7$; Bad-Patient: range 2.1 to 9.0 degrees, $M = 5.7$, $SD = 1.5$); Patient head (Prototypical: range 4.5 to 9.8 degrees, $M = 6.4$, $SD = 1.3$; Bad-Patient: range 4.2 to 9.1 degrees, $M = 5.9$, $SD = 1.2$); Agent torso (Prototypical: range 2.4 to 8.2 degrees, $M = 4.3$, $SD = 1.8$; Bad-Patient: range 2.4 to 7.9 degrees, $M = 4.2$, $SD = 1.5$); Patient torso (Prototypical: range 2.7 to 8.5 degrees, $M = 4.3$, $SD = 1.6$; Bad-Patient: range 2.3 to 7.9 degrees, $M = 4.1$, $SD = 1.4$).

List design—Four practice items were shown before the test trials, which were all Prototypical versions of event categories that were excluded due to lack of name agreement in the norming survey ('chasing', 'facepainting', 'shoving', and 'tapping'). Agent Color and Location were counterbalanced among these practice items. Each stimuli list consisted of two blocks of 24 items each. Within each block, Duration (Short/Long), Consistency (Consistent/Inconsistent), and Event Type (Prototypical/Bad-Patient) were counterbalanced among the items, such that items were equally divided among the eight conditions. Agent Color (Blue/Red) and Agent Location (Left/Right) were also counterbalanced across conditions.

The second block was the same as the first except that for each item, the alternate Event Type appeared (e.g., if an item was Prototypical in the first block, it would be Bad-Patient in the second, and vice-versa). Thus across blocks, the expected response (Consistent or Inconsistent) for an item was the same. Additionally, for a given item, the Event Type that appeared in the first block was counterbalanced between subjects.

Each block had a different fixed pseudo-random order, with the only criterion that across the two blocks an item had to be separated from its role-reversed version by at least seven intervening trials. Seven additional stimuli lists were generated by rotating the items through each of the eight conditions in a consistent manner, using a Latin square design. Reverse-order versions of these eight lists were also generated. Since Probe Type was varied between subjects, a total of 32 different lists were used.

Control condition—If we find that people perform worse in the Bad-Patient condition than the Prototypical condition, it would suggest that given very limited information, people make decisions about event roles in part by using the event role features described above. However, a concern would be that even given ample time to examine the event scene, despite our sentence norming results, people may still judge that the Patient in a Bad-Patient image is in fact the Agent, and thus perform lower on Bad-Patient images than on Prototypical ones.

To rule out this possibility, 16 undergraduates participated in a control condition in which test trials were identical to those above, except that the images were displayed for two full seconds instead of only briefly. Since Duration was no longer a factor, we had eight lists (four reverse) for each Probe Type (Agent or Patient). If people use categorical definitions of Agent and Patient to assign event roles given information from multiple fixations, any

difference in performance between the two Event Types that we see at very brief durations should be eliminated.

Results and Discussion

Results of brief display conditions—Table 5 presents the average proportion of correct responses for Consistent and Inconsistent trials, along with 95% confidence intervals, split by Duration, Event Type, and Probe Type. Figure 5 reports these values in terms of d'. Perfect d' scores by the standard approximation method would be 2.77 and 3.07 for subject and item means, respectively.

As can be seen in the table and the figure, the core findings of Exp. 2 were replicated here with new images and additional events. In particular, subjects were able to recognize the Agent and the Patient even at the Short display duration. Moreover, as predicted, images containing a Patient who possessed Agent-like features (i.e., the Bad-Patient event type) resulted in poorer performance than images in which the Patient did not. This was true at both the Short and Long display durations.

These conclusions found support in tests of significance. In particular, d' performance was reliably above zero at all levels of Duration, Event Type, and Probe Type based on one-sample t-tests (all ps < .005). In addition, separate ANOVAs on subject and item means were carried out, with the following factors: Duration (Short or Long), Event Type (Prototypical or Bad-Patient), Probe Type (Agent or Patient), List (1–8), and either List Order (Forward or Reverse; subject ANOVA only) or Item Group (1–8; item ANOVA only). Both ANOVAs revealed a reliable main effect of Duration ($F_1(1, 32) = 59.6$, $p < .001$, $\eta_p^2 = .65$; $F_2(1, 16) = 46.0$, $p < .001$, $\eta_p^2 = .74$), with the Long Duration (73 ms) yielding higher d' values, as well as a reliable main effect of Event Type ($F_1(1, 32) = 83.4$, $p < .001$, $\eta_p^2 = .72$; $F_2(1, 16) = 18.7$, $p < .001$, $\eta_p^2 = .54$), with the Prototypical Event Type yielding higher d' values than the Bad-Patient Event Type. The ANOVAs revealed no effect of Probe Type (both Fs < 1) nor any reliable two-or three-way interactions between Duration, Event Type, or Probe Type (both Fs < 1, except for Duration × Probe Type interaction: $F_1(1, 32) = 1.76$, $p = .19$; $F_2(1, 16) = 2.25$, $p = .15$). Thus, subjects were able to extract role information at both durations and for both event types, but were able to do so better at the longer duration, as well as when the event type was Prototypical rather than Bad-Patient. In addition, these effects were similar whether subjects were asked about the Agent or Patient.

As noted in the Procedure section above, there were some small differences between the Prototypical and Bad-Patient stimuli in terms of the distance between the crosshair and the Agent and Patient. Indeed item d' scores at the Short Duration were found to correlate with at least two of these distances (Agent head and torso). To be sure that that the effect of Event Type was not due to these small differences in distances, we also ran equivalent item ANOVAs on d' values for the Short Duration that were first residualized for distance (four ANOVAs, one for each distance measure: Agent head, Agent torso, Patient head, Patient torso). In other words, ANOVAs were run on values for which variance due to distance had been first partialed out. In all four cases, a strong reliable effect of Event Type still held, indicating that distance between the event participants and the crosshair could not explain the effect of Event Type.

Additionally, it is unlikely that in the current experiment, we simply induced a strategy (the implicit or explicit use of event role features to determine event role assignment) that would not normally be utilized outside of the laboratory. Indeed, if anything, the presence of our Bad-Patient stimuli would discourage use of such a strategy, since in those stimuli, the event role features are not a reliable cue to event roles. Furthermore, one might expect that if the strategy developed over the course of the experiment, performance on Prototypical event

types would start out more similar to performance on Bad-Patient items and improve with exposure to our stimuli; that is, one might expect to see an interaction between Event Type (Prototypical or Bad-Patient) and Portion of Experiment (First or Last), possibly interacting with Duration (Short or Long). However, ANOVAs that included Portion of Experiment as an additional factor revealed no such interactions, whether Portion was split by halves (both Fs < 1.64) or thirds (both Fs < 1.70). This analysis does not rule out the possibility that subjects developed such a strategy immediately, after exposure to the first few items, but does cast doubt on the possibility that subjects developed such a strategy over the course of the experiment, nonetheless.

Results of control condition—When images were displayed for two seconds before being masked, subject performance was essentially at ceiling: Accuracy scores for all conditions were at 93% or above for both Consistent and Inconsistent probes, as presented in Table 5. The resulting subject mean d' scores were thus very similar (for Agent Probe: Prototypical, d' = 3.15, and Bad-Patient, d' = 3.06; for Patient Probe: Prototypical, d' = 3.01, and Bad-Patient, d' = 2.83). ANOVAs conducted separately on subject and item means of d' scores revealed no effects of Probe Type (both Fs < 1.70) or Event Type (both Fs < 1), nor their interaction (both Fs < 1), indicating essentially identical performance on Bad-Patient and Prototypical event types across Agent and Patient probes.[8]

Thus it appears that given enough visual information, people have no trouble recognizing the Agent and Patient in these events, despite the fact that in the Bad-Patient versions, the Patient may at first glance appear more Agent-like. In fact, only one of the 16 control subjects, when asked post-experiment about what changed from one instance of an event to the next, mentioned the physical poses of the actors, which suggests that subjects were not consciously aware of any feature manipulation, even given a substantially greater time to view the images than subjects in any of the previous experiments. These results, along with the similarity in name agreement between the Prototypical and Bad-Patient event types from the sentence norms, argue against an alternative interpretation of our results, namely that the Bad-Patient event types are simply not depicting the same event roles as the Prototypical event types.

General Discussion

In our first two experiments, we found that subjects were able to make use of the limited visual information available in briefly displayed event scenes to form abstract conceptual representations of event categories (Exp. 1) and event roles (Exps. 2A-2C). That they could do so from less than the visual input from one fixation demonstrates that event information can be extracted without actively fixating on specific components of the event scene, such as event participants or event regions. Our most significant finding is that gist extraction is not limited to general scene content (Potter, 1976) or the spatial relationship between objects (Biederman et al., 1982, 1988), but also includes abstract relational information between event participants in natural scenes.

We explored what makes this rapid recognition possible (Exp. 3), and found that the degree to which Agents and Patients were distinct in their event role features predicted performance on the previous experiments, but only for extracting the event roles, not the event category. In fact, the features may be properly called “Agent-like”: people imagining these events taking place agree that Agents but not Patients almost always possess these features. Finally,

[8]In the subject ANOVA, the interaction of List Type with the other factors was not included because there would not have been enough degrees of freedom to estimate the error term. A Probe Type × Item Group interaction was found on item means ($F_2(7, 16)$ = 3.18, p = .03, η_p^2 = .58).

when we systematically manipulated the event role features within event by instilling Patients with Agent-like features (the Bad-Patient event type; Exp. 4), role recognition was worsened; however, given enough time to observe the same Bad-Patient events (Exp. 4, control condition), subjects performed essentially perfect role recognition.

Implications for Event Category Recognition

Certain aspects of our results lead us to believe that people can use global perceptual properties of event scenes to rapidly recognize them independent of event roles. Performance on the first set of stimuli in the gist experiments probing about role information (Exps. 2A-2C) correlated with one another, but none correlated with the experiment probing about event category information (Exp. 1). Moreover, the difference in Agent and Patient event role features in our first set of stimuli predicted performance on event role experiments, but not the event category experiment. These patterns suggest that one does not need to classify the entities in an event scene to recognize the event category or vice-versa. There is evidence that the spatial layout and other global properties of non-event scenes contribute to rapidly extracting their gist (Greene & Oliva, 2009a, 2009b; Oliva, 2005; Oliva & Schyns, 1997; Oliva & Torralba, 2001; Schyns & Oliva, 1994), and we suspect that people may use global features of event scenes that are correlated with certain event categories – including information about the spatial layout of entities in the scene – to rapidly extract event gist.

It would follow that event scenes with spatial layouts less common among other event scenes (i.e., more “unique”) should be easier to categorize. In contrast, event scenes with spatial layouts that share many properties with other event categories would be harder to distinguish. Spatial layout is, however, unlikely to be the sole means for performing event category recognition. An alternative path to event recognition is role discrimination, which certainly occurred reliably enough in our experiments that it could in principle contribute to event category recognition. The lack of significant correlations between the performance on our event category and role experiments suggests otherwise, but we recognize that the analyses were post-hoc. Future work manipulating the uniqueness of spatial layout within and across events could illuminate the information people use to extract event gist. Dobel and colleagues (2010) describe an experiment using blurred event images that partially addresses this possibility: they found that low ambiguity of event scene spatial layouts predicted high performance in naming briefly displayed events. Nevertheless, further exploration of event scene spatial layout and its effect on gist extraction is warranted.

Implications for Event Role Recognition

There are several aspects of the current findings that inform theories of event gist extraction and event role recognition more specifically. Our proposal in the Introduction that people use rapidly emerging perceptual features to aid in role recognition was confirmed. Nevertheless, it remains the case that the categories Agent and Patient are largely abstract and depend on the relationship between entities. These roles’ more conceptual nature must be reconciled with the apparent use of features that only probabilistically predict (but do not define) the categories themselves.

In fact, the use of event role features comports well with a dual theory of concepts as laid out by Armstrong, Gleitman, and Gleitman (1983), in which one maintains both a categorical definition for a concept to determine membership in a category, and an identification function that makes use of prototype representations, especially more readily available perceptual and functional properties, to quickly categorize things in the world. In their experiments, Armstrong et al. took rating and categorization tasks that were previously used by Rosch (1973) as evidence for the prototype view and extended them to well-defined

categories (odd number, female, plane geometry figure). Surprisingly, these well-defined categories elicited graded responses and varied categorization times, even after subjects explicitly confirmed the categorical nature of the above categories beforehand. The authors took this as evidence for the presence of both definitional and identificational (prototype-driven) roles for concepts.[9] In our experiments, that performance in role recognition was modulated by manipulation of certain physical features (e.g., extremities outstretched) suggests the utilization of an identification function for rapid categorizations of things in the world. In other words, Agent is defined as *the person performing the action*, but to identify the Agent as such when fixating on aspects of the event scene is not possible, people may first use prototypical features of Agents, such as *facing the other person with arms outstretched*, to make the less perceptually obvious event role categorization.

Further support for the dual theory of concepts, as applied to event roles, lies in the control condition of Exp. 4: When subjects could observe the scene for two seconds and thus make multiple fixations within the scene itself, they had no trouble assigning role categories to the event participants, even when such participants were atypical posed. Such performance indicates that people default to the more abstract, categorical definitions for role concepts, given enough time and visual input. The additional fixations allowed subjects to analyze aspects of the scene (event participants or event regions) in detail to form a coherent representation of the event as a whole, as a means of confirming or adjusting their initial hypotheses about the event roles, so for example, to ensure that the putative Agent is indeed the Agent (and not an Agent-like Patient).

The perceptual features available early on to classify event participants as Agents or Patients would presumably extend far beyond just physical stance or relative position of extremities to e.g., facial expression, movement, and instruments. Though we noted no difference in facial expression between our Agents and Patients, rapid integration of emotion implicit in faces and body postures is possible, at least from a first-person viewpoint (Meeren et al., 2005). In addition, the onset of motion is among the most salient and immediately available types of visual information in a scene (Abrams & Christ, 2003; Hillstrom & Yantis, 1994; Sekuler, Watamaniuk, & Blake, 2002), motion is more probabilistically associated with Agents than with Patients (Cicchino, Aslin, & Rakison, 2011; Rakison, 2005), and movement features contribute to event segmentation (Zacks, 2004).

Event role features are probably not limited to event-general properties, either. Certain early visual information (for example, whether a participant is holding an instrument) is likely to contribute to event-specific role recognition. In principle one could compile a list of event-specific features that may contribute to early role (and event) categorization (see McRae et al. [1997] for a similar project in sentence processing). The purpose of our own feature exploration was to show that visual features probabilistically associated with event roles are used for early role recognition, so this leaves open the possibility for further exploration into what other features may contribute to this process. Interestingly, the linguistic norming results of McRae and colleagues (1997) reveal a more highly variable set of features for the Patient than for the Agent role, similar to what we observed for event scenes (see the feature distributions for Agents and Patients in both Exp. 3 and in the mental imagery experiment described in Exp. 4, found in Figures 3 and 4, respectively). McRae and colleagues found

[9]Armstrong et al. (1983) emphasize that even a dual theory of concepts does not help address all of the difficulties that arise when features are included as a part of any theory of concepts. The main issue is that describing and categorizing the features in a way that adequately fits a prototype (or core) remains very elusive. Likewise the combinatory power of features is not clear. However, we do not claim that either core concepts or identification functions can be defined solely by a set of features. Shying away from Armstrong et al.'s (1983) admonition of the lack of feature theory's explanatory power, we have used the term *feature* to describe body part position and other cues as possible perceptual identification heuristics. The event role features may or may not have any bearing on the conceptual core itself, and they are by no means features independent of the "role" concepts.

that event-specific features listed for Patients were less consistent across subjects than those listed for Agent roles, and furthermore, subjects' ratings allowed more "Agent-like" event participants (e.g., a policeman) to be associated with the Patient role than "Patient-like" event participants (e.g., a baby) with the Agent role. Of course contributing to such a pattern may be the fact that we (and McRae et al.) were investigating events involving two animate participants rather than those involving an inanimate participant.

Neural Processes Involved in Event Recognition

Our finding that visual features related to body posture can inform rapid event role recognition dovetails nicely with recent neural models of event recognition. Such models have focused both on what areas of the brain are involved in recognition of biological motion and actions, and on how information from brain regions involved in different aspects of action recognition might be integrated. Giese and Poggio (2003) proposed a neurally motivated computational model in which form and motion information are processed by two different streams and then integrated in a feed-forward manner for recognition of biological movements. Specifically the form pathway analyzes "snapshots" of body shapes and activates neurons that encode subsequent snapshot neurons in the sequence. Variants or extensions of this model all follow the same general principle, namely that static body form information is integrated into sequences (Downing, Peelen, Wigget, & Tew, 2006; Lange, George, & Lappe, 2006; Lange & Lappe, 2006; Singer & Sheinberg, 2010), and indeed the models have support in electrophysiological recordings in monkeys (Singer & Sheinberg, 2010; Vangeneuden et al., 2011; Vangeneuden, Pollick, & Vogels, 2009). Areas of the brain that have been implicated in the process of event recognition are the superior temporal sulcus (STS) for both form and movement recognition (Grossman, Jardine, & Pyles, 2010; Oram & Perrett, 1996; Vangeneuden et al., 2009); area MT for motion signals (Giese & Poggio, 2003); extrastriate and fusiform body areas (EBA/FBA) for body part and full body form, shape, and posture (Downing, Jiang, Shuman, & Kanwisher, 2001; Downing & Peelen, 2011); and ventral premotor cortex (vPMc) for action discrimination (Moro et al., 2008). Human imaging studies have found that one or more of these areas can represent higher level action categories independent of viewpoint (Grossman et al., 2010) or the actor involved (Kable & Chatterjee, 2006; Wiggett & Downing, 2011).

Our results lend some support to the static pose model: Subjects in our study viewed just one brief "snapshot" of an event and could already extract information about the entities and even the event category itself, most likely using perceptual features about body pose to do so. This is in line with evidence that pose information is extracted rapidly (Meeren et al., 2005) and that poses are integrated into action representations over about 120 ms (Singer & Sheinberg, 2010). Pose information may be made available due to the EBA/FBA and other areas responsible for body form (for a review of the functions of EBA and FBA, see Downing & Peelen, 2011). Importantly, studies of the neural correlates of action representation almost exclusively focus on one-participant actions, often with point-light displays lacking full form information. The fact that our studies demonstrate that humans can integrate rapidly displayed information about multiple actors into coherent action representations should provide impetus to researchers to extend their work to the neural substrates of action representations with multiple actors.

Automated Computer Event Recognition

For insight into human event recognition, we may look to the successes of work on automated event recognition. Here too our findings connect well to the existing research. Perhaps most relevant is the work on computer classification of actions from still-frame images (Delaitre, Laptev, & Sivic, 2010; Ikizler, Cinbis, Pehlivan, & Duygulu, 2008; Ikizler-Cinbis, Cinbis, & Sclaroff, 2009; Lan, Wang, Yang, & Mori, 2010; Wang, Jiang,

Drew, Li, & Mori, 2006; Yang, Wang, & Mori, 2010). The methods used involve discriminating the poses of the human participants from their background, and either matching the poses to training sets (Delaitre et al., 2010; Ikizler et al., 2008) or creating classes of actions in an unsupervised manner (Wang et al., 2006). This work relates to the work discussed above on global scene properties, as the systems generally function best when the spatial layouts of the participants are most unique. For example, in Ikilzer-Cinbis et al. (2009), 'dancing' was often misclassified as 'running', since many dance poses resemble those of 'running'. It may even be enough for only certain *parts* of the human form to be unique in a spatial layout: Yang, Wang, and Mori's (2010) model actually learns which features of poses are more important for classification of specific actions (e.g., for 'sitting', triangle or A-shaped legs). However, so far the work on still-frame images has focused on single-participant actions like 'walking', 'running', or 'throwing' (though a notable exception is Lan et al., 2010, who modeled group activities and individual actions interactively in still images).

Recent work on the recognition of human actions and interactions from video rather than still images applies more directly to our work on two-participant events (see Aggarwal & Ryoo, 2011, for a review). Some successful models follow a hierarchical approach, starting with basic body part movements (gestures), atomic actions (simple Agent-motion-target specifications), and eventually recognition of events by previously specified domain knowledge of relative poses and event causality as it unfolds over time. These systems can achieve greater than 90% accuracy even with complex interactions (e.g. 'punching', 'pointing', etc.; Park & Aggarwal, 2004; Ryoo & Aggarwal, 2006). Errors arise from failures to successfully identify lower-level components (i.e., gestures and body part movements). There is evidence that human event recognition is similarly hierarchical (see Giese & Poggio's [2003] computational model of human event recognition), so it is not surprising that both automated event recognition models and our own human observers suffer from poorer performance when a participant's body poses are manipulated slightly from the norm. Indeed lack of success to identify events due to failed pose recognition also corresponds to difficulties found in automated event recognition from one-participant still-images.

How information is either integrated or discarded in certain automated event recognition models may have direct applications to how humans might process events and their components. One model (Vahdat, Gao, Ranjbar, & Mori, 2011) learns what "key poses" of an action sequence are useful for discriminating interactions, discarding information in the action sequence not relevant to recognition of the event, and these key poses are incorporated into the model, along with spatiotemporal ordering of the poses between actors. Other models include a dependency between event and object recognition such that the possible actions and concomitant objects are mutually constraining (e.g., Gupta & Davis, 2007; Gupta, Kembhavi, & Davis, 2009; Yao & Fei-Fei, 2010). Improvements shown with these models indicate that any robust event recognition system, human or automated, must exploit useful and reliable cues (e.g., object identity) and disregard less relevant cues that may contribute to noise (e.g., transition motions or poses that occur between the crucial ones).

Conclusions

We extended work in rapid object and scene gist extraction to include event category and role information from event scenes. In some ways the ease and speed of event category and role recognition are striking, as event category and roles are relatively abstract categories, yet event scenes in reality change quite quickly. It may even be that people's representations of unfolding event scenes are not all that different from the kinds of static visual input like in

our stimuli, especially across fixations to different parts of an event scene as the event is unfolding (Verfaillie & Daems, 2002).

In addition, we found that the features we hypothesized to be relevant for event role recognition do indeed contribute to the role recognition process. In doing so, we offered a preliminary theory of event category and role recognition that draws from work in both scene gist extraction and object recognition.

Our work should be seen as a starting point for those who wish to investigate the first moments of event recognition. Further work will address the level of abstraction of event representations in the first moments of observing an event, as well as the kinds of information that can be extracted from types of events other than two-participant causal interactions (e.g., single- and multi-participant events, or human-object interactions). Further work will also address the relationship between event recognition and language production (see Gleitman et al., 2007; Griffin & Bock, 2000; Papafragou, Hulbert, & Trueswell, 2008).

Acknowledgments

This work was partially funded by Grant 5R01HD055498 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development to A.P. and J.T. We especially want to thank the 24 actors who were photographed for our event scenes.

References

Abrams RA, Christ SE. Motion onset captures attention. Psychological Science. 2003; 14(5):427–432.10.1111/1467-9280.01458 [PubMed: 12930472]

Aggarwal JK, Ryoo MS. Human activity analysis: A review. ACM Computer Surveys. 2011 Apr. 43(3):Article 16.10.1145/1922649.1922653

Armstrong SL, Gleitman LR, Gleitman H. What some concepts might not be. Cognition. 1983; 13:263–308.10.1016/0010-0277(83)90012-4 [PubMed: 6683139]

Baillargeon, R.; Li, J.; Gertner, Y.; Wu, D. How do infants reason about physical events?. In: Goswami, U., editor. The Wiley-Blackwell handbook of childhood cognitive development. 2. Oxford: Blackwell; 2011. p. 11-48.

Biederman I, Ju G. Surface versus edge-based determinants of visual recognition. Cognitive Psychology. 1988; 20:38–64.10.1016/0010-0285(88)90024-2 [PubMed: 3338267]

Biederman I, Blickle TW, Teitelbaum RC, Klatsky GJ. Object search in nonscene displays. Journal of Experimental Psychology: Learning, Memory, and Cognition. 1988; 14(3):456–467.10.1037/0278-7393.14.3.456

Biederman I, Mezzanotte RJ, Rabinowitz JC. Scene perception: Detecting and judging objects undergoing relational violations. Cognitive Psychology. 1982; 14:143–177.10.1016/0010-0285(82)90007-X [PubMed: 7083801]

Biederman I, Rabinowitz JC, Glass AL, Stacy EW. On the information extracted from a glance at a scene. Journal of Experimental Psychology. 1974; 103(3):597–600.10.1037/h0037158 [PubMed: 4448962]

Blake R, Shiffrar M. Perception of human motion. Annual Review of Psychology. 2007; 58:47–73.10.1146/annurev.psych.57.102904.190152

Boyce SJ, Pollatsek A, Rayner K. Effect of background information on object identification. Journal of Experimental Psychology: Human Perception and Performance. 1989; 15(3):556–566.10.1037/0096-1523.15.3.556 [PubMed: 2527962]

Castelhano MS, Henderson JM. The influence of color on the perception of scene gist. Journal of Experimental Psychology: Human Perception and Performance. 2008; 34:660–675.10.1037/0096-1523.34.3.660 [PubMed: 18505330]

Cicchino JB, Aslin RN, Rakison DH. Correspondences between what infants see and know about causal and self-propelled motion. Cognition. 2011; 118:171–192.10.1016/j.cognition.2010.11.005 [PubMed: 21122832]

Cohen MA, Alvarez GA, Nakayama K. Natural-scene perception requires attention. Psychological Science. 2011 Advance online publication. 10.1177/0956797611419168

Cutting JE, Kozlowski LT. Recognizing friends by their walk: Gait perception without familiarity cues. Bulletin of the Psychonomic Society. 1977; 9(5):353–356.

Davenport JL. Consistency effects between objects in scenes. Memory & Cognition. 2007; 35(3):393–401.10.3758/BF03193280 [PubMed: 17691140]

Davenport JL, Potter MC. Scene consistency in object and background perception. Psychological Science. 2004; 15:559–564.10.1111/j.0956-7976.2004.00719.x [PubMed: 15271002]

Delaitre, V.; Laptev, I.; Sivic, J. Recognizing human actions in still images: A study of bag-of-features and part-based representations. Proceedings of the 21st British Machine Vision Conference; Aberystwyth. 2010.

Dobel, C.; Glanemann, R.; Kreysa, H.; Zwitserlood, P.; Eisenbeiss, S. Visual encoding of coherent and non-coherent scenes. In: Pedersen, E.; Bohnemeyer, J., editors. Event Representation in language: Encoding events at the language cognition interface. Cambridge, UK: Cambridge University Press; 2010. p. 189-215.

Dobel C, Gumnior H, Bölte J, Zwitserlood P. Describing scenes hardly seen. Acta Psychologica. 2007; 125:129–143.10.1016/j.actpsy.2006.07.004 [PubMed: 16934737]

Downing PE, Peelen MV. The role of occipitotemporal body-selective regions in person perception. Cognitive Neuroscience. 2011; 2(3–4):186–226.10.1080/17588928.2011.582945

Downing PE, Jiang Y, Shuman M, Kanwisher N. A cortical area selective for visual processing of the human body. Science. 2001; 293(5539):2470–2473.10.1126/science.1063414 [PubMed: 11577239]

Downing PE, Peelen MV, Wiggett AJ, Tew BD. The role of the extrastriate body area in action perception. Social Neuroscience. 2006; 1(1):52–62.10.1080/17470910600668854 [PubMed: 18633775]

Dowty D. Thematic proto-roles and argument selection. Language. 1991; 67:547–619.10.2307/415037

Fabre-Thorpe M, Delorme A, Marlot C, Thorpe S. A limit to the speed of processing in ultra-rapid visual categorization of novel natural scenes. Journal of Cognitive Neuroscience. 2001; 13(2):171–180.10.1162/089892901564234 [PubMed: 11244543]

Fei-Fei L, Iyer A, Koch C, Perona P. What do we perceive in a glance of a real-world scene? Journal of Vision. 2007; 7(1):10, 1–29.10.1167/7.1.10 [PubMed: 17461678]

Freyd JJ. The mental representation of movement when static stimuli are viewed. Perception & Psychophysics. 1983; 33(6):575–581.10.3758/BF03202940 [PubMed: 6622194]

Friedman A. Framing pictures: The role of knowledge in automated encoding and memory for gist. Journal of Experimental Psychology: General. 1979; 108(3):316–355.10.1037/0096-3445.108.3.316 [PubMed: 528908]

Giese MA, Poggio T. Neural mechanisms for the recognition of biological movements. Nature Reviews Neuroscience. 2003; 4(3):179–192.10.1038/nrn1057

Gleitman LR, January D, Nappa R, Trueswell JC. On the give and take between event apprehension and utterance formulation. Journal of Memory and Language. 2007; 57:544–569.10.1016/j.jml.2007.01.007 [PubMed: 18978929]

Göksun T, Hirsh-Pasek K, Golinkoff RM. Trading spaces: Carving up events for learning language. Perspectives on Psychological Science. 2010; 5:33–42.10.1177/1745691609356783

Goldin-Meadow, S. Language development under atypical learning conditions: Replication and implications of a study of deaf children of hearing parents. In: Nelson, K., editor. Children's Language. Vol. 5. Hillsdale, NJ: Lawrence Erlbaum; 1985. p. 197-245.

Golinkoff RM. Semantic development in infants: The concepts of agent and recipient. Merrill Palmer Quarterly. 1975; 21:181–193.

Golinkoff RM, Kerr JL. Infants' perception of semantically defined action role changes in filmed events. Merrill-Palmer Quarterly. 1978; 24:53–61.

Gordon, P. The origin of argument structure in infant event representations. Proceedings of the 26th Boston University Conference on Language Development; Somerville, MA: Cascadilla Press; 2003.

Green C, Hummel JE. Familiar interacting object pairs are perceptually grouped. Journal of Experimental Psychology. Human Perception and Performance. 2006; 32(5):1107–1119.10.1037/0096-1523.32.5.1107 [PubMed: 17002525]

Greene MR, Oliva A. Recognition of natural scenes from global properties: Seeing the forest without representing the trees. Cognitive Psychology. 2009a; 58:137–176.10.1016/j.cogpsych.2008.06.001 [PubMed: 18762289]

Greene MR, Oliva A. The briefest of glances: The time course of natural scene understanding. Psychological Science. 2009b; 20:464–472.10.1111/j.1467-9280.2009.02316.x [PubMed: 19399976]

Griffin ZM, Bock K. What the eyes say about speaking. Psychological Science. 2000; 11:274–279.10.1111/1467-9280.00255 [PubMed: 11273384]

Grossman ED, Blake R. Brain areas active during visual perception of biological motion. Neuron. 2002; 35:1167–1175.10.1016/S0896-6273(02)00897-8 [PubMed: 12354405]

Grossman ED, Jardine NL, Pyles JA. fMRI-adaptation reveals invariant coding of biological motion on the human STS. Frontiers in Human Neuroscience. 2010; 4:1–15.10.3389/neuro.09.015.2010 [PubMed: 20204154]

Gupta, A.; Davis, LS. Objects in action: An approach for combining action understanding and object perception. Proceedings of the 2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition; 2007. p. 1-8.

Gupta A, Kembhavi A, Davis LS. Observing human-object interactions: Using spatial and functional compatibility for recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2009; 31(10):1775–1789.10.1109/TPAMI.2009.83 [PubMed: 19696449]

Henderson, JM.; Ferreira, F. Scene perception for psycholinguists. In: Henderson, JM.; Ferreira, F., editors. The interface of language, vision, and action: Eye movements and the visual world. New York, NY: Psychology Press; 2004. p. 1-58.

Henderson JM. Introduction to real-world scene perception. Visual Cognition. 2005; 12:849–851.10.1080/13506280444000544

Hillstrom AP, Yantis S. Perception & Psychophysics. 1994; 55(4):399–411.10.3758/BF03205298 [PubMed: 8036120]

Hollingworth A. Scene and position specificity in visual memory for objects. Journal of Experimental Psychology: Learning, Memory, and Cognition. 2006; 32(1):58–69.10.1037/0278-7393.32.1.58

Hollingworth A, Henderson JM. Does consistent scene context facilitate object perception? Journal of Experimental Psychology: General. 1998; 127(4):398–415.10.1037/0096-3445.127.4.398 [PubMed: 9857494]

Hollingworth A, Henderson JM. Object identification is isolated from scene semantic constraint: Evidence from object type and token discrimination. Acta Psychologica. 1999; 102(2–3):319–343.10.1016/S0001-6918(98)00053-5 [PubMed: 10504886]

Huttenlocher J, Smiley P, Charney R. Emergence of action categories in the child: Evidence from verb meanings. Psychological Review. 1983; 90(1):72–93.10.1037/0033-295X.90.1.72

Ikizler, N.; Cinbis, RG.; Pehlivan, S.; Duygulu, P. Recognizing actions from still images. Proceedings of the 19th International Conference on Pattern Recognition; 2008. p. 1-4.

Ikizler-Cinbis, N.; Cinbis, RG.; Sclaroff, S. Learning actions from the Web. 12th International Conference on Computer Vision; 2009. p. 995-1002.

Intraub H. Rapid conceptual identification of sequentially presented pictures. Journal of Experimental Psychology: Human Perception and Performance. 1981; 7(3):604–610.10.1037/0096-1523.7.3.604

Intraub H. Conceptual masking: The effects of subsequent visual events on memory for pictures. Journal of Experimental Psychology: Learning, Memory, and Cognition. 1984; 10(1):115–125.10.1037/0278-7393.10.1.115

Johansson G. Visual perception of biological motion and a model for its analysis. Perception and Psychophysics. 1973; 14:201–211.10.3758/BF03212378

Kable JW, Chatterjee A. Specificity of action representations in the lateral occipitotemporal cortex. Journal of Cognitive Neuroscience. 2006; 18(9):1498–1517.10.1162/jocn.2006.18.9.1498 [PubMed: 16989551]

Kable JW, Kan IP, Wilson A, Thompson-Schill SL, Chatterjee A. Conceptual representations of action in the lateral temporal cortex. Journal of Cognitive Neuroscience. 2005; 17(12):1855–1870.10.1162/089892905775008625 [PubMed: 16356324]

Kable JW, Lease-Spellmeyer J, Chatterjee A. Neural substrates of action event knowledge. Journal of Cognitive Neuroscience. 2002; 14(5):795–805.10.1162/08989290260138681 [PubMed: 12167263]

Kersten AW, Billman D. Event category learning. Journal of Experimental Psychology: Learning, Memory, and Cognition. 1997; 23(3):638–658.10.1037/0278-7393.23.3.638

Kourtzi Z, Kanwisher N. Activation in human MT/MST by static images with implied motion. Journal of Cognitive Neuroscience. 2000; 12(1):48–55.10.1162/08989290051137594 [PubMed: 10769305]

Lan, T.; Wang, Y.; Yang, W.; Mori, G. Beyond actions: Discriminative models for contextual group activities. In: Lafferty, J.; Williams, CKI.; Shawe-Taylor, J.; Zemel, RS.; Culotta, A., editors. Advances in Neural Information Processing Systems. 2010. p. 23

Lange J, Lappe M. A model of biological motion perception from configural form cues. The Journal of Neuroscience. 2006; 26(11):2894–2906.10.1523/JNEUROSCI.4915-05.2006 [PubMed: 16540566]

Lange J, Georg K, Lappe M. Visual perception of biological motion by form: A template-matching analysis. Journal of vision. 2006; 6(8):836–849.10.1167/6.8.6 [PubMed: 16895462]

Leslie AM, Keeble S. Do six-month-old infants perceive causality? Cognition. 1987; 25:265–288.10.1016/S0010-0277(87)80006-9 [PubMed: 3581732]

Macmillan, NA.; Creelman, CD. Detection Theory: A User's Guide. 2. Mahwah, N.J: Lawrence Erlbaum; 2005.

Malpass D, Meyer AS. The time course of name retrieval during multiple-object naming: Evidence from extrafoveal-on-foveal effects. Journal of Experimental Psychology: Learning, Memory, and Cognition. 2010; 36(2):523–537.10.1037/a0018522

McRae K, Ferretti T, Amyote L. Thematic roles as verb-specific concepts. Language and Cognitive Processes. 1997; 12(2):137–176.10.1080/016909697386835

Meeren HKM, van Heijnsbergen CCRJ, de Gelder B. Rapid perceptual integration of facial expression and emotional body language. Proceedings of the National Academy of Sciences of the United States of America. 2005; 102(45):16518–16523.10.1073/pnas.0507650102 [PubMed: 16260734]

Morgan JL, Meyer AS. Processing of extrafoveal objects during multiple-object naming. Journal of Experimental Psychology: Learning, Memory, and Cognition. 2005; 31(3):428–442.10.1037/0278-7393.31.3.428

Morgan JL, van Elswijk G, Meyer AS. Extrafoveal processing of objects in a naming task: Evidence from word probe experiments. Psychonomic Bulletin & Review. 2008; 15(3):561–565.10.3758/PBR.15.3.561 [PubMed: 18567255]

Moro V, Urgesi C, Pernigo S, Lanteri P, Pazzaglia M, Aglioti SM. The neural basis of body form and body action agnosia. Neuron. 2008; 60(2):235–246.10.1016/j.neuron.2008.09.022 [PubMed: 18957216]

Muentener P, Carey S. Infants' causal representations of state change events. Cognitive Psychology. 2010; 61(2):63–86.10.1016/j.cogpsych.2010.02.001 [PubMed: 20553762]

Newtson D. Attribution and the unit of perception of ongoing behavior. Journal of Personality and Social Psychology. 1973; 28(1):28–38.10.1037/h0035584

Newtson D. The perceptual organization of ongoing behavior. Journal of Experimental Social Psychology. 1976; 12(5):436–450.10.1016/0022-1031(76)90076-7

Oliva, A. Gist of the Scene. In: Itti, L.; Rees, G.; Tsotsos, JK., editors. Neurobiology of Attention. San Diego: Elsevier Academic Press; 2005. p. 251-257.

Oliva A, Schyns PG. Coarse blobs or fine edges? Evidence that information diagnosticity changes the perception of complex visual stimuli. Cognitive Psychology. 1997; 34(1):72–107.10.1006/cogp.1997.0667 [PubMed: 9325010]

Oliva A, Torralba A. Modeling the shape of the scene: a holistic representation of the spatial envelope. International Journal in Computer Vision. 2001; 42:145–175.10.1023/A:1011139631724

Oliva A, Torralba A. The role of context in object recognition. Trends in Cognitive Sciences. 2007; 11(12):520–527.10.1016/j.tics.2007.09.009 [PubMed: 18024143]

Oram MW, Perrett DI. Integration of form and motion in the anterior superior temporal polysensory area (STPa) of the macaque monkey. Journal of Neurophysiology. 1996; 76:109–129. [PubMed: 8836213]

Papafragou A, Hulbert J, Trueswell JC. Does language guide event perception? Evidence from eye movements. Cognition. 2008; 108(1):155–184.10.1016/j.cognition.2008.02.007 [PubMed: 18395705]

Park, S.; Aggarwal, JK. Recognition of human interaction using multiple features in grayscale images. Proceedings of the 15th International Conference on Pattern Recognition; 2000. p. 51-54.

Park S, Aggarwal JK. A hierarchical Bayesian network for event recognition of human actions and interactions. Multimedia Systems. 2004; 10:164–179.10.1007/s00530-004-0148-1

Potter MC. Meaning in visual search. Science. 1975; 187:965–966.10.1126/science.1145183 [PubMed: 1145183]

Potter MC. Short-term conceptual memory for pictures. Journal of Experimental Psychology: Human Learning and Memory. 1976; 2:509–522.10.1037/0278-7393.2.5.509 [PubMed: 1003124]

Potter MC, Levy EI. Recognition memory for a rapid sequence of pictures. Journal of Experimental Psychology. 1969; 81:10–15.10.1037/h0027470 [PubMed: 5812164]

Pourtois G, Peelen MV, Spinelli L, Seeck M, Vuilleumier P. Direct intracranial recording of body-selective responses in human extrastriate visual cortex. Neuropsychologia. 2007; 45(11):2621–2625.10.1016/j.neuropsychologia.2007.04.005 [PubMed: 17499819]

Rakison DH. A secret agent? How infants learn about the identity of objects in a causal scene. Journal of Experimental Child Psychology. 2005; 91:271–296.10.1016/j.jecp.2005.03.005 [PubMed: 15869760]

Rosch, E. On the internal structure of perceptual and semantic categories. In: Moore, TE., editor. Cognitive Development and the Acquisition of Language. New York: Academic Press; 1973.

Ryoo MS, Aggarwal JK. Recognition of composite human activities through context-free grammar based representation. Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2006; 2:1709–1718.10.1109/CVPR.2006.242

Schyns PG, Oliva A. From blobs to boundary edges: Evidence for time- and spatial-scale-dependent scene recognition. Psychological Science. 1994; 5(4):195–200.10.1111/j.1467-9280.1994.tb00500.x

Sekuler, R.; Watamaniuk, SNJ.; Blake, R. Visual motion perception. In: Pashler, H.; Yantis, S., editors. Stevens' Handbook of Experimental Psychology: Vol. 1: Sensation and perception. 3. New York: Wiley; 2002. p. 121-176.

Shipley, TF. An invitation to an event. In: Shipley, TF.; Zacks, JM., editors. Understanding events: From perception to action. New York: Oxford University Press; 2008. p. 3-30.

Shipley, TF.; Zacks, JM., editors. Understanding events: From perception to action. New York: Oxford University Press; 2008.

Singer JM, Sheinberg DL. Temporal cortex neurons encode articulated actions as slow sequences of integrated poses. The Journal of Neuroscience. 2010; 30(8):3133–3145.10.1523/JNEUROSCI.3211-09.2010 [PubMed: 20181610]

Spelke, ES.; Phillips, AT.; Woodward, AL. Infants' knowledge of object motion and human action. In: Sperber, D.; Premack, D.; Premack, A., editors. Causal cognition: A multidisciplinary debate. New York: Oxford University Press; 1995.

Talmy, L. Typology and process in concept structuring. Vol. II. Cambridge, MA: MIT Press; 2000. Toward a cognitive semantics.

Thierry G, Pegna AJ, Dodds C, Roberts M, Basan S, Downing P. An event-related potential component sensitive to images of the human body. NeuroImage. 2006; 32(2):871–879.10.1016/j.neuroimage.2006.03.060 [PubMed: 16750639]

Tranel D, Kemmerer D, Adolphs R, Damasio H, Damasio AR. Neural correlates of conceptual knowledge for actions. Cognitive Neuropsychology. 2003; 20(3):409–432.10.1080/02643290244000248 [PubMed: 20957578]

Troje, NF. Retrieving information from human movement patterns. In: Shipley, TF.; Zacks, JM., editors. Understanding events: From perception to action. New York: Oxford University Press; 2008. p. 308-334.

Urgesi C, Candidi M, Ionta S, Aglioti SM. Representation of body identity and body actions in extrastriate body area and ventral premotor cortex. Nature Neuroscience. 2007; 10(1):30–31.10.1038/nn1815

Vahdat, A.; Gao, B.; Ranjbar, M.; Mori, G. A discriminative key pose sequence model for recognizing human interactions. interactions. Proceedings of the 2011 IEEE International Conference on Computer Vision Workshops; 2011. p. 1729-1736.

van Boxtel JJA, Lu H. Visual search by action category. Journal of Vision. 2011; 11(7):19, 1–14.10.1167/11.7.19 [PubMed: 21709212]

Vangeneugden J, De Mazière PA, Van Hulle MM, Jaeggli T, Van Gool L, Vogels R. Distinct mechanisms for coding of visual actions in macaque temporal cortex. The Journal of Neuroscience. 2011; 31(2):385–401.10.1523/JNEUROSCI.2703-10.2011 [PubMed: 21228150]

Vangeneugden J, Pollick F, Vogels R. Functional differentiation of macaque visual temporal cortical neurons using a parametric action space. Cerebral Cortex. 2009; 19(3):593–611.10.1093/cercor/bhn109 [PubMed: 18632741]

Verfaillie K, Daems A. Representing and anticipating human actions in vision. Visual Cognition. 2002; 9(1):217–232.10.1080/13506280143000403

Wang Y, Jiang H, Drew MS, Li Z-N, Mori G. Unsupervised discovery of action classes. Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2006; 2:1654–1661.10.1109/CVPR.2006.321

Wiggett AJ, Downing PE. Representation of action in occipito-temporal cortex. Journal of Cognitive Neuroscience. 2011; 23(7):1765–1780.10.1162/jocn.2010.21552 [PubMed: 20807060]

Yang, W.; Wang, Y.; Mori, G. Recognizing human actions from still images with latent poses. Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition; 2010. p. 2030-2037.

Yao, B.; Fei-Fei, L. Modeling mutual context of object and human pose in human-object interaction activities. Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition; 2010. p. 17-24.

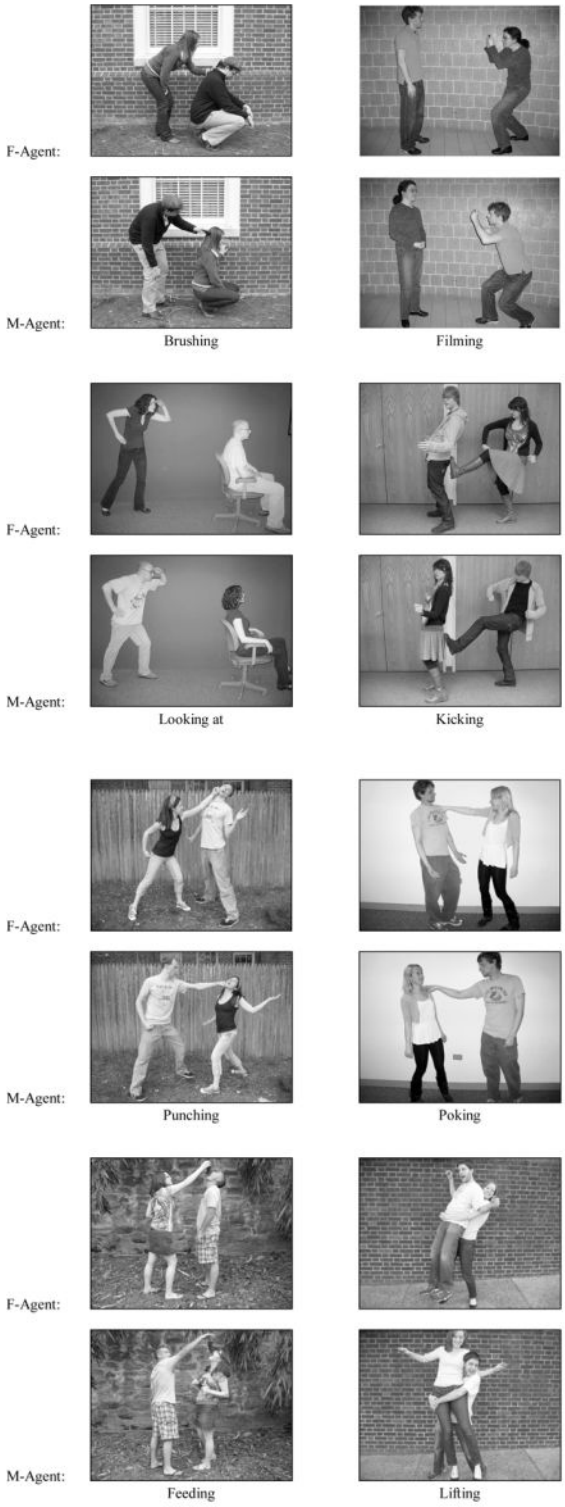
Zacks J. Using movement and intentions to understand simple events. Cognitive Science. 2004; 28(6):979–1008.10.1016/j.cogsci.2004.06.003

Zacks JM, Speer NK, Swallow KM, Braver TS, Reynolds JR. Event perception: a mind-brain perspective. Psychological Bulletin. 2007; 133(2):273–293.10.1037/0033-2909.133.2.273 [PubMed: 17338600]

Zacks JM, Tversky B, Iyer G. Perceiving, remembering, and communicating structure in events. Journal of Experimental Psychology: General. 2001; 130(1):29–58.10.1037/0096-3445.130.1.29 [PubMed: 11293458]

Appendix A

Example target images used in Exps. 1 and 2. Female Agent (F-Agent) versions appear above Male Agent (M-Agent) versions. Images were displayed in color in the experiments.



Appendix B

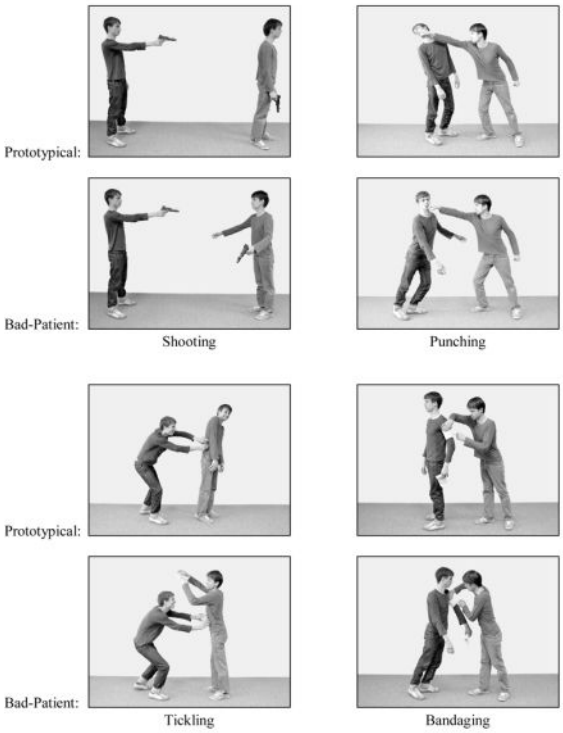
Mean item d' performance at the Short Duration for Exps. 1 and 2 (Probe Type in parentheses), with the corresponding median feature difference scores from Exp. 3.

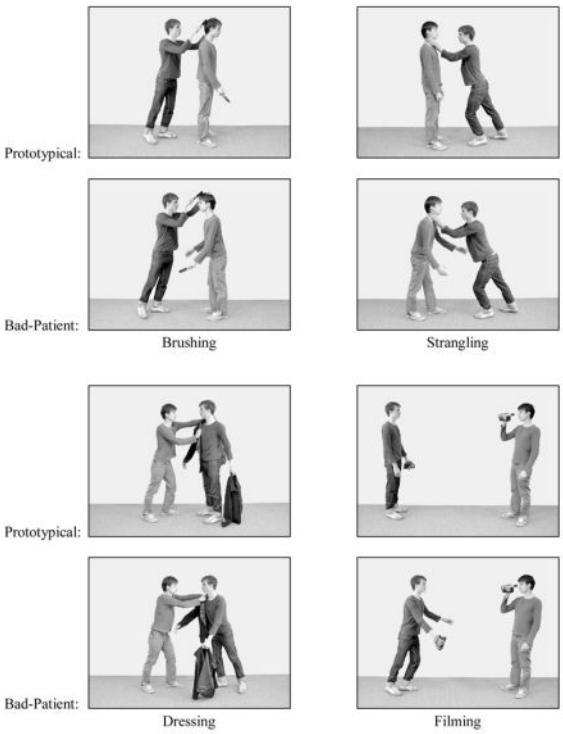
		d' scores at 37 ms (Short) Duration				Feature Difference Scores (Exp. 3)				
Event Category	**Agent Gender**	**Exp. 1 (Verb)**	**Exp. 2A (Agent)**	**Exp. 2B (Patient)**	**Exp. 2C (Sentence)**	**Head**	**Body**	**Extremities**	**Leaning**	**Sum Score**
Bite	F	0.48	0.67	0.67	0.38	−0.35	−0.35	−0.51	1.96	0.76
	M	0.48	0.67	0.48	0.00	−0.01	1.07	0.17	1.82	3.05
Brush	F	1.35	2.30	1.82	1.35	2.18	2.18	2.12	2.15	8.63
	M	2.30	2.30	2.30	1.35	2.23	2.23	2.20	2.23	8.89
Call after	F	0.00	0.00	0.67	0.36	2.18	2.18	1.29	2.12	7.76
	M	0.48	1.15	0.00	0.00	2.23	2.23	1.13	1.43	7.02
Chase	F	1.82	−0.67	0.00	1.35	0.07	2.18	0.70	0.95	3.89
	M	1.82	1.35	0.00	1.15	0.37	2.23	0.05	1.66	4.31
Feed	F	0.00	1.15	0.67	0.48	1.08	0.00	2.18	1.42	4.68
	M	−0.67	1.35	1.82	0.99	0.60	0.70	1.78	1.40	4.48
Film	F	0.00	1.82	0.67	1.15	−0.03	0.00	0.70	1.07	1.74
	M	0.00	1.82	2.30	0.67	0.00	0.00	1.98	2.09	4.07
Kick	F	0.67	1.82	−1.35	0.00	2.23	2.20	2.20	1.26	7.89
	M	1.35	2.30	2.30	0.67	2.18	2.18	2.18	0.54	7.08
Lift	F	1.35	0.00	1.82	−0.32	1.32	2.12	1.89	−1.10	4.23
	M	1.82	0.48	−1.15	0.99	0.17	1.39	1.15	1.09	3.81
Look at	F	0.00	1.35	−0.67	1.47	2.23	2.23	0.60	1.64	6.70
	M	−1.35	1.82	1.82	0.00	2.18	2.18	1.48	1.97	7.81
Poke	F	1.82	−1.35	−0.48	−0.36	0.00	−0.07	0.70	0.58	1.22
	M	1.35	1.35	1.15	−0.67	0.00	−0.56	2.18	1.51	3.13
Pull	F	0.67	0.67	−1.35	0.00	0.00	0.58	0.29	0.00	0.86
	M	0.67	−0.67	−0.67	0.48	−0.25	0.51	0.17	0.10	0.53
Punch	F	−0.67	1.82	1.82	1.15	1.98	0.12	2.02	2.09	6.21
	M	1.82	1.82	0.00	0.00	1.49	−0.15	2.15	1.78	5.27
Push	F	2.30	1.15	1.15	1.47	2.23	2.23	2.23	2.20	8.89
	M	1.15	2.30	1.35	0.99	2.18	2.18	2.12	2.15	8.63
Scare	F	1.35	0.67	1.35	−0.32	0.00	0.48	0.68	2.01	3.16
	M	−0.67	−0.48	1.82	0.00	0.00	1.34	1.25	2.10	4.69
Scratch	F	1.82	1.82	1.35	0.99	−0.47	0.34	2.21	0.87	2.95

		d' scores at 37 ms (Short) Duration				Feature Difference Scores (Exp. 3)				
Event Category	**Agent Gender**	**Exp. 1 (Verb)**	**Exp. 2A (Agent)**	**Exp. 2B (Patient)**	**Exp. 2C (Sentence)**	**Head**	**Body**	**Extremities**	**Leaning**	**Sum Score**
	M	0.48	0.67	1.82	1.47	0.10	1.43	2.29	2.12	5.94
Tap	F	1.82	0.00	−0.67	1.15	2.27	2.38	2.29	1.27	8.21
	M	2.30	2.30	1.82	1.47	2.07	2.27	2.15	1.28	7.77

Appendix C

Example images used in Exp. 4. Prototypical event type versions appear above Bad-Patient versions. Images were displayed in color in the experiment. The boy in darker clothes is the blue-shirted actor, and the other is the red-shirted actor.





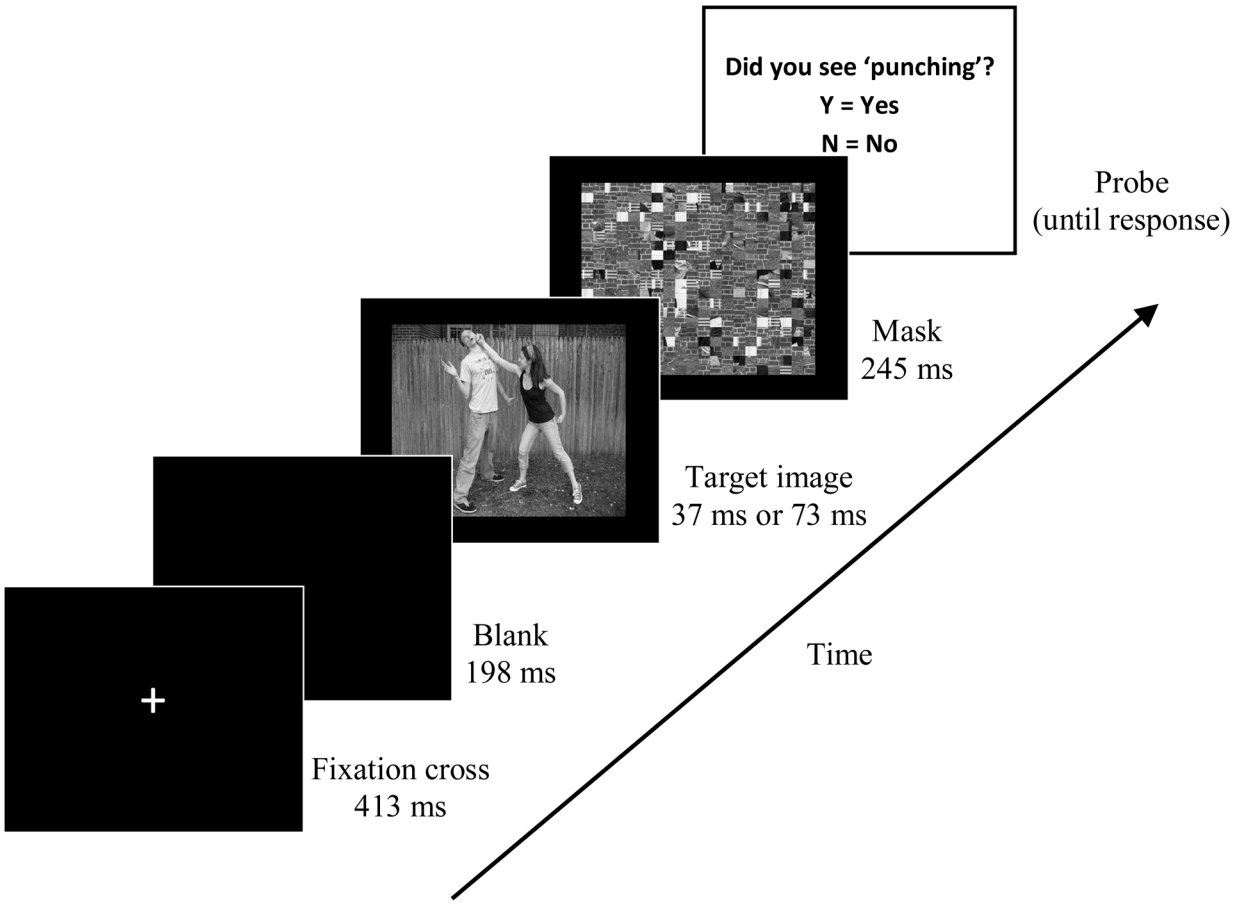


Figure 1.
Trial structure for Exps. 1, 2, and 4.

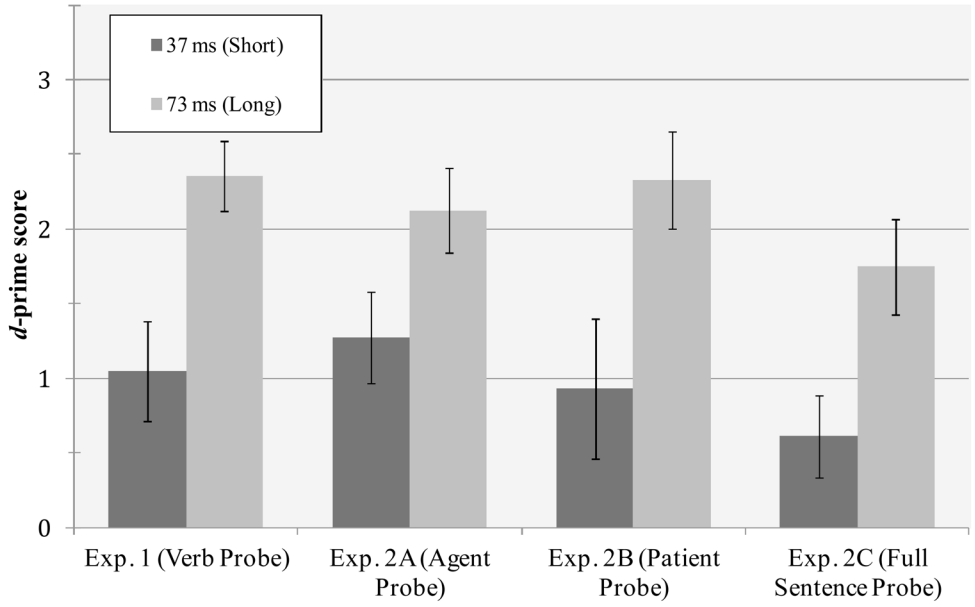


Figure 2.

Results of Exps. 1 and 2. Calculated on subject means. Error bars indicate 95% confidence interval. *d*-prime (d') is a bias-free measure of sensitivity to information, based on hit and false-alarm rates. Zero indicates no sensitivity, and at the approximation rates we used, 3.07 indicates full sensitivity (i.e., no misses or false alarms).

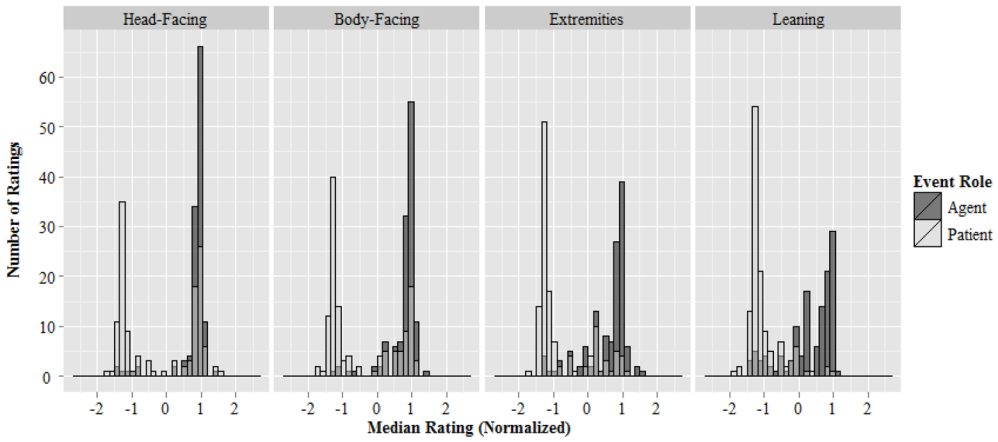


Figure 3.

Histogram of all normalized feature ratings (for each of 16 subjects and 32 images), with Event Roles (Agent or Patient) overlaid on top of one another, separately for each event role feature (gray bars indicate overlap between Agent and Patient bars). Bins are 0.15 *z*-ratings wide. Experiment 3.

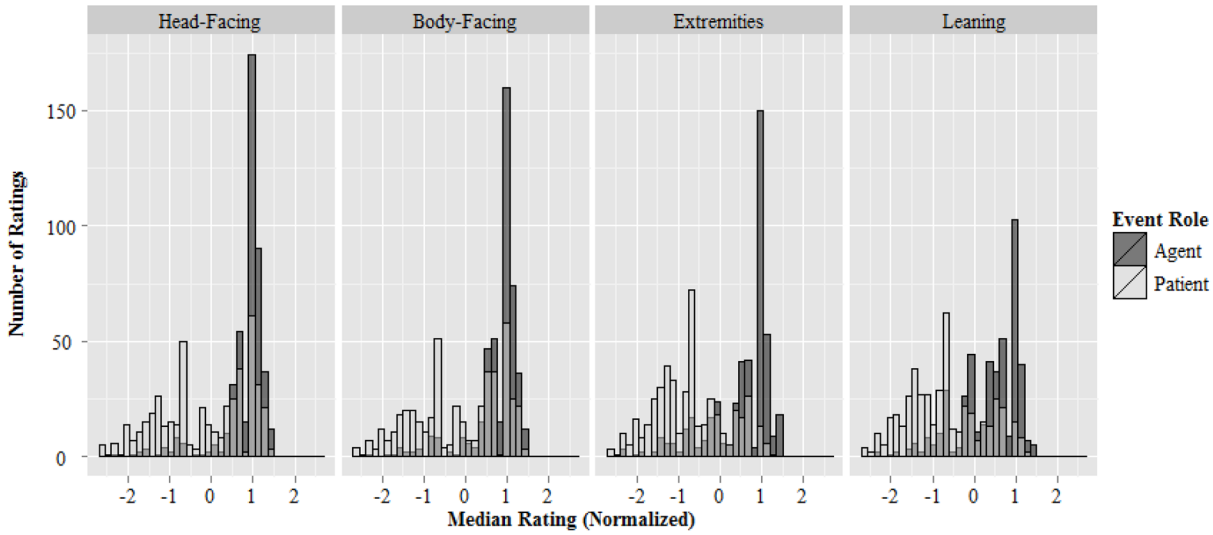


Figure 4.

Histogram of all normalized feature ratings (for each of 14 subjects and 33 event categories), with Event Roles (Agent or Patient) overlaid on top of one another, separately for each event role feature (gray bars indicate overlap between Agent and Patient bars). Bins are 0.15 *z*-ratings wide. Experiment 4, “mental imagery” experiment.

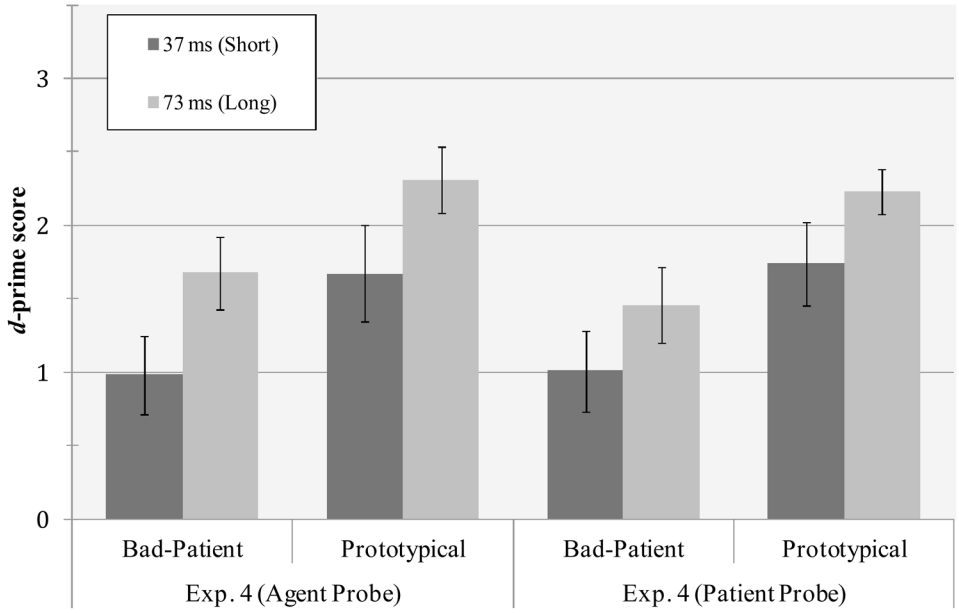


Figure 5.
Results of Exp. 4, brief display conditions. Calculated on subject means. Error bars indicate 95% confidence interval. *d*-prime (d') is a bias-free measure of sensitivity to information, based on hit and false-alarm rates. Zero indicates no sensitivity, and at the approximation rates we used, 2.77 indicates full sensitivity (i.e., no misses or false alarms).

Table 1

Mean proportion of correct responses on Consistent and Inconsistent trials, as calculated on subject means. The 95% confidence interval for each mean is given in parentheses. Experiments 1 and 2.

		Consistency	
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)
Exp. 2B (Patient)	37 ms (Short)	.63 (.09)	.70 (.09)
	73 ms (Long)	.89 (.05)	.88 (.07)
Exp. 2C (Full Sentence)	37 ms (Short)	.59 (.08)	.62 (.09)
	73 ms (Long)	.89 (.04)	.68 (.10)

Table 2

Medians (*Med*), Standard Deviations (*SD*), and Ranges of the normalized feature ratings of the target images of Exps. 1 and 2. The median feature ratings for each image were used to calculate the above values. Also included is the Sum Score, which is the sum of the median values of all four feature scores, for each item. Normalized ratings of the Agent and Patient are given separately, as well as the difference of the Agent and Patient ratings. Experiment 3.

	Agent			Patient			Difference		
Feature	***Med***	***SD***	**Range**	***Med***	***SD***	**Range**	***Med***	***SD***	**Range**
Head-Facing	0.94	0.42	(−1.16, 0.98)	−0.31	1.07	(−1.40, 1.32)	0.84	1.06	(−0.47, 2.27)
Body-Facing	0.94	0.25	(−0.10, 0.98)	−1.04	0.94	(−1.49, 0.98)	1.41	1.02	(−0.56, 2.38)
Extremities	0.88	0.57	(−1.20, 1.32)	−1.22	0.69	(−1.40, 0.85)	1.84	0.84	(−0.51, 2.30)
Leaning	0.50	0.65	(−1.23, 0.98)	−1.22	0.33	(−1.57, 0.35)	1.58	0.77	(−1.10, 2.23)
Sum Score	2.74	0.96	(0.09, 3.88)	−2.59	2.28	(−5.40, 1.19)	4.69	2.61	(0.53, 8.89)

Table 3

Estimates of the fixed effect of the Sum Feature Difference Score in multilevel models predicting item d' performance at the 37 ms (Short) Duration, run separately for each experiment. The models were run in R using the lmer function from the lme4 package, with the following formula: d.prime ~ Sum.Score + (1 + Sum.Score | Event.Category). Each model was compared separately to the null model, i.e., a model with no fixed effects (Baayen, Davidson, & Bates, 2008), and the resulting *p*-value is given, based on a chi-square test of the change in −2 restricted log likelihood (Steiger, Shapiro, & Browne, 1985). The null model included only the random intercept of Event Category and random slope of Sum.Score by Event.Category, and no fixed effect of Sum.Score. Experiment 3.

Experiment (Probe Type)	Estimate	S.E.	*t*-value	*p*-value
Exp.1 (Verb)	−0.06	0.19	−0.30	.77
Exp.2A (Agent)	0.48	0.15	3.17	.01 *
Exp.2B (Patient)	0.34	0.19	1.83	.13
Exp. 2C (Sentence)	0.24	0.12	2.07	.06
Exps. 2A–2C (All Roles)	0.31	0.12	2.53	.03 *

Table 4

Multilevel models of item d' performance at the 37 ms (Short) Duration as a function of the feature difference scores, run separately for each Experiment (Probe Type) and each feature difference score. The models were run in R using the lmer function from the lme4 package, with the following formula: d.prime ~ Feature.Score + (1 + Feature.Score | Event.Category). Each model was compared separately to the null model, i.e., a model with no fixed effects (Baayen, Davidson, & Bates, 2008), and the resulting chi-square and *p*-values (in parentheses, uncorrected) are given, based on a chi-square test of the change in −2 restricted log likelihood (Steiger, Shapiro, & Browne, 1985). The null model included only the random intercept of Event Category and random slope of Feature.Score by Event.Category, and no fixed effect of Feature.Score. Experiment 3.

Feature	Exp. 1 (Verb)	Exp. 2A (Agent)	Exp. 2B (Patient)	Exp. 2C (Sentence)	Exps. 2A–2C (All Roles)
Head-Facing	0.00 (.97)	6.61 (.01) *	0.55 (.46)	1.11 (.29)	3.62 (.06)
Body-Facing	0.66 (.42)	0.00 (> .99)	0.01 (.90)	3.03 (.08)	0.66 (.42)
Extremities	0.07 (.79)	6.93 (.008) **	7.85 (.005) **	0.05 (.83)	7.92 (.005) **
Leaning	1.89 (.17)	0.75 (.39)	3.99 (.05) *	2.63 (.11)	2.92 (.09)

Table 5

Mean proportion of correct responses on Consistent and Inconsistent trials, as calculated on subject means. The 95% confidence interval for each mean is given in parentheses. Experiment 4.

			Consistency	
Probe Type	**Duration**	**Event Type**	**Consistent**	**Inconsistent**
Agent	37 ms (Short)	Bad-Patient	.66 (.05)	.68 (.07)
		Prototypical	.76 (.07)	.83 (.04)
	73 ms (Long)	Bad-Patient	.82 (.06)	.77 (.05)
		Prototypical	.90 (.04)	.92 (.05)
	2 s (Control)	Bad-Patient	.95 (.02)	.94 (.01)
		Prototypical	.97 (.01)	.94 (.02)
Patient	37 ms (Short)	Bad-Patient	.69 (.06)	.68 (.06)
		Prototypical	.83 (.05)	.80 (.05)
	73 ms (Long)	Bad-Patient	.71 (.05)	.82 (.07)
		Prototypical	.88 (.05)	.93 (.04)
	2 s (Control)	Bad-Patient	.94 (.02)	.94 (.02)
		Prototypical	.96 (.02)	.96 (.03)