Rapid apprehension of the coherence of action scenes

Reinhild Glanemann¹ · Pienie Zwitserlood² · Jens Bölte² · Christian Dobel³

Published online: 29 January 2016
© Psychonomic Society, Inc. 2016

Abstract Some information about complex naturalistic scenes, such as the scene's gist (a beach, a restaurant) or the category of an object depicted in a scene, can be extracted within a fraction of a second. The present study focused on the rapid apprehension of scene coherence in action scenes involving two actors. Coherence was manipulated by either varying global (body posture) or local (object details) action-scene properties. Scenes were presented for 20, 30, 50, or 100 ms, and subsequently masked. Viewers were able to determine scene coherence with only 30 ms presentation duration for global scene layout, but not even 100 ms sufficed for local object-action consistency. A second experiment ruled out that the difficulties in detecting local object-action consistency resulted from a strategic adaption by the participants. Overall, the results suggest that, as with scene gist, rapid extraction of coherence information from action scenes relies

more on a scene's overall Gestalt than on detailed visual or even semantic representations.

Keywords Scene perception · Action recognition · Scene gist · Visual representation · Event cognition

Humans have remarkable abilities to extract meaningful information from complex visual scenes at a single glance and without overt attention on any scene detail. People can decide whether a scene contains an object of a particular type, such as an animal (Thorpe, Fize, & Marlot, 1996) or a vehicle, or make global semantic categorizations, such as “a park,” “in a restaurant,” also known as the scene's *gist*, within an eye blink (Biederman, Mezzanotte, & Rabinowitz, 1982; Potter, 1976; Schyns & Oliva, 1994). Similarly, the coherence or meaningfulness of an action scene can be apprehended without overt attention shifts at very brief presentation durations (Dobel, Gummior, Bölte, & Zwitserlood, 2007). As with object categories and scene gist, the mechanisms that allow coherence apprehension are not well known. Thus, our current objective was to assess which type of visual information enables fast coherence apprehension. To this aim, we contrasted the manipulation of global scene layout (body orientation of actors) with the manipulation of local scene information (appropriateness of the action-relevant object). To establish how timing¹ affects the perception of each coherence type, we varied presentation duration. Before providing details of our study, we will summarize the relevant results on rapid information uptake from complex visual scenes (for an overview, see Henderson & Ferreira, 2004).

✉ Reinhild Glanemann
r.glanemann@uni-muenster.de

Pienie Zwitserlood
zwitser@uni-muenster.de

Jens Bölte
boelte@uni-muenster.de

Christian Dobel
christian.dobel@med.uni-jena.de

¹ Clinic of Phoniatrics and Pedaudiology, University Hospital Münster, Kardinal-von-Galen-Ring 10, 48149 Münster, Germany

² Institute for Psychology, Westfälische Wilhelms-Universität Münster, Fliegerstraße 21, 48149 Münster, Germany

³ Department of Otorhinolaryngology, Friedrich-Schiller University Jena, Lessingstr. 2, 07743 Jena, Germany

¹ Here and in the following, presentation, time is used to capture the timing of visual processing. We use “early,” “rapid,” and “fast” for data from brief presentations, not for the speed with which responses were made.

The first line of research focuses on rapid decisions on the presence or absence of a predefined object category member (such as “animal”). Successful categorization has been demonstrated with presentation durations shorter than 100 ms, even for stimuli displayed in far visual periphery (Kirchner & Thorpe, 2006; Thorpe, Gegenfurtner, Fabre-Thorpe, & Bülthoff, 2001). Performance is hardly affected if attention is focused on a secondary task (Li, VanRullen, Koch, & Perona, 2002), or if two images are presented simultaneously (Li, VanRullen, Koch & Perona, 2005). Two explanations have been proposed for fast categorization: (1) a rapid but crude first pass through the visual system that relies on an unbound collection of image features that are diagnostic of particular object categories (Crouzet & Serre, 2011; Evans & Treisman, 2005), and/or (2) faster and more efficient processing of familiar naturalistic stimuli, for example, due to “more intense neuronal” representations (Bacon-Macé, Macé, Fabre-Thorpe, & Thorpe, 2005; Braun, 2003; Li et al., 2005). Note that fast detection of an exemplar of a particular category (“animal”) does not necessarily imply that the object itself (“lion”) has been identified.

Another line of research examines the early recognition of a scene’s gist. Gist has been defined as “knowledge of the scene category (e.g., kitchen) and the semantic information that may be retrieved based on that category” (Henderson & Ferreira, 2004, p. 15). Given enough processing time, gist processing includes “all levels of processing, from low-level features (e.g., color, spatial frequencies), to intermediate image properties (e.g., surface, volume) and high-level information (e.g., objects, activation of semantic knowledge)” (Oliva, 2005, p. 251). Typically, gist-study stimuli involve multiple scene elements and their spatial relations. This is different from most of the categorization studies mentioned previously, in which usually one critical object has a prominent foreground position within the scene. Early gist apprehension has been shown to emerge at 30 ms to 50 ms after scene onset (cf. Henderson & Ferreira, 2004). Spatial layout information, that is, the spatial arrangement of the whole scene and/or its objects, seems to play a crucial role in gist recognition (e.g., Castelano & Henderson, 2007; Greene & Oliva, 2009; Oliva & Schyns, 1997; Oliva & Torralba, 2006; Sanocki & Epstein, 1997; Schyns & Oliva, 1994). Note that the identification of objects in a scene is not a prerequisite for gist recognition (Oliva & Schyns, 1997; Schyns & Oliva, 1994). Some recent evidence suggests that not only the internal consistency of scenes but also the observers’ task matters. Evans, Horowitz, and Wolfe (2011) showed that both scene gist and object category can be extracted from a scene within one glance (e.g., “beach” and “animal”). Importantly, the task influences whether the two types of information facilitate or inhibit each other.

While most studies used categorization tasks or tasks in which participants were probed with multiple-choice

questions, Fei-Fei, Iyer, Koch, and Perona (2007) used a different approach. After brief presentation of photographic scenes, participants were asked to report in a free-answer format what they had seen. The results showed that low-level sensory features (e.g., shading) and coarse shape information needed less sensory input than semantic-level information. The authors argued for a “mutual facilitation between overall scene recognition and object recognition” (p. 23).

In sum, the brief presentation of a visually complex natural scene suffices to extract its gist and to decide about the presence of objects from a specified superordinate category. Apparently, no overt attention is needed for either task (but see Cohen, Alvarez, & Nakayama, 2011). Whereas category-membership decisions are based on local features diagnostic of the object category, thus requiring high spatial frequency information, gist recognition seems to rely on global scene information, probably subserved by low spatial frequency processing. Note that neither task requires the actual identification of objects in the scene.

This is probably different when the semantic consistency of scene components – one of the manipulations in the current study – is at issue. Most studies agree that scene context influences object recognition (see Bar, 2004; Hollingworth, 2007). Objects are detected faster and more accurately when they appear in typical surroundings (e.g., a coffee machine in a kitchen) compared to atypical surroundings (e.g., a violin in a bathroom; Biederman et al., 1982; De Graef, Christiaens, & d’Ydewalle, 1990). Note that there is evidence that some of these results might be due to a response bias: When this bias is appropriately controlled for, object perception is no longer facilitated by scene context (Hollingworth & Henderson, 1998). The object-context interaction seems bidirectional because objects also seem to influence context processing (Davenport & Potter, 2004; Joubert, Fize, Rousselet, & Fabre-Thorpe, 2008). Inconsistent objects do not capture gaze from an initial glimpse of a peripherally presented scene, and it seems questionable whether objects can be identified without attention (Vo & Henderson, 2011). To our knowledge, none of the studies in this field investigated the mechanisms and time course of object and scene consistency judgments, which is our focus.

Most of the aforementioned studies used static scenes (a chair in a kitchen), but scenes of actions (a girl pushing a boy) are much more frequently encountered in human vision (and not only in action movies). The rapid processing of action scenes – also called event scenes (Hafri, Papafragou, & Trueswell, 2013) – has been investigated in few studies so far. In an eye-tracking study with line drawings of agent-patient-actions (e.g., a girl shooting at a man), the patient of the action was fixated and identified within 500 ms (Griffin & Bock, 2000). With photographs of similar types of actions, even 150 ms of masked peripheral presentation was sufficient for patient detection (Dobel, Glanemann, Kreysa,

Zwitserslood, & Eisenbeiss, 2010). Similarly, Hafri et al. (2013) observed that event roles (agent, patient) can be detected above chance with (foveal) presentation of 37 ms. Actions were correctly identified and named in nearly 60 % of trials in the study by Dobel et al. (2010), and well above chance with 73 ms foveal presentation in the Hafri et al. study.

Dobel et al. (2007) manipulated the coherence of agent-patient action scenes by varying body orientation of the actors, which led to coherent (face-to-face) and incoherent (back-to-back) actions. With 100 ms masked peripheral presentation, coherence detection was high (80 % correct), but it remained unclear whether performance was based on low spatial-layout information or on a semantic representation of the action (note that action identification was low at 100 ms). In incoherent actions, the action's contour was more ragged; body parts and the instrument used in each scene (e.g., rifle, tray) pointed toward the scene center in the coherent version but outwards in the incoherent version. Such features seem to indicate event roles (Hafri et al., 2013).

In sum, there is evidence that essential information about action scenes can be extracted in the absence of overt attention. Much of this information could be extracted from coarse, low spatial frequency visual information, but this does not hold for the detection of category members, which requires at least some featural detail. Some argue that an “unbound” collection of image features could be hardwired into the visual system, facilitating the detection of ecologically important stimuli (Crouzet & Serre, 2011). In fact, evidence from highly time-sensitive methods, such as EEG, points toward the existence of detailed, high spatial frequency information early in the visual processing stream. These data come from experiments that investigate motivationally relevant (mainly emotional) stimuli with natural scenes (e.g., Junghöfer, Bradley, Elbert, & Lang, 2001) and faces (Schupp, Öhman, Junghöfer, Weike, Stockburger, & Hamm, 2004) but also words (Keuper et al., 2014; Keuper et al., 2013) and gratings (Stolarova, Keil, & Moratti, 2006). The data show enhanced processing of emotional stimuli as early as 50 ms to 100 ms after stimulus onset. Given the detail needed to distinguish motivationally relevant and neutral stimuli, these results suggest that detailed features are available early in the processing stream and allow to distinguish between emotional and neutral stimuli, at least on a neurophysiological level. These data fit with evidence for the categorization of natural images on the basis of ERPs (Rousselet, Fabre-Thorpe, & Thorpe, 2002) but also of saccadic eye movements (Kirchner & Thorpe, 2006).

Thus, it is still unclear how detailed rapidly constructed visual representations of actions are, and which information subserves coherence judgements. From the perspective of global precedence (see Kimchi, 1992, for an overview), global aspects of a scene are processed more rapidly than local scene detail. On the other hand, local features can be (rendered) salient, either because they are

indicative of a salient category (animal, for example) and/or because of their motivational relevance. Thus, there is evidence for both global and local information, but it is not clear whether high detail is readily available when this information is not motivationally or ecologically salient. To fill this gap, we investigated which type of visual information drives fast coherence judgments: coarse, low spatial information, or more fine-grained, featural information about object detail. Unlike most other studies, we used naturalistic action scenes that were presented briefly, and subsequently masked. Coherent action scenes depicted meaningful actions. In incoherent action scenes, the visual or causal relation between the action components rendered the action meaningless. All actions involved a human agent who performed the action and a patient who was acted upon (e.g., one person serving coffee to another person). Action coherence was manipulated by reversing none, one, or both involved actors, or by using an action-appropriate or action-inappropriate object (see Fig. 1).

The first manipulation influences the action scene's global spatial layout, and body orientation can serve as a reliable cue for action coherence, even at brief presentation durations (cf. Dobel et al., 2007). In contrast, the semantic consistency of an action and its object cannot be extracted from the global spatial properties of the image. Instead, such judgments require parsing the scene into action and object and most probably the identification of both components. This necessitates a more detailed scene representation, involving processing of high spatial frequency information. We chose coherence judgments because they can be made without (correct) perception of all scene elements.

Given that global spatial layout information is available early during visual processing, we predicted better performance for body orientation than for action-object stimuli with short presentation durations. If only ecologically or motivationally relevant stimuli can be perceived on the basis of detailed featural information during early vision, we predicted that with our stimuli, coherence cannot be made on the basis of action-object consistency with short presentation durations.

Experiment 1

Method

Participants

Sixty-four students from Münster University (38 female, ages 19–28 years) with normal or corrected-to-normal vision participated in this study for course credits or payment (3.00 €). Sixteen participants were tested per presentation duration.

Fig. 1 A–D Body orientation (“to handcuff someone”); E–F: action-object consistency (“to shoot someone”). Coherent actions: A, B, and E; incoherent actions: C, D, and F

Materials

Color photographs of two-participant action scenes served as stimuli, depicting a coherent or an incoherent agent-patient action with two human actors in front of a neutral background. Coherence was manipulated by varying body orientation (10 actions, Set 1) or by exchanging the object of the action (10 actions, Set 2). In Set 1, body orientation of agent and patient was varied independently, resulting in face-to-face, agent-facing-patient's-back, patient-facing-agent's-back, and back-to-back orientations (see Fig. 1, A–D). The first two orientations resulted in coherent actions (C+), and the latter two in incoherent ones (C-). In Set 2, coherence was varied by replacing the action-appropriate object with an inappropriate one (see Fig. 1, E–F). These actions were shown slightly enlarged relative to the body-orientation actions to ensure proper identification at long presentation durations. All stimuli were pretested with 20 participants, who did not participate in the main study. They were asked to decide whether the depicted action was coherent or incoherent. Stimuli judged correctly by at least 75 % of the participants were included in the main experiment (see the [Appendix](#) for materials).

The complete set consisted of the 10 actions from Set 1, each in all four orientations, and each with four different actor pairs (160 stimuli). Each of the 10 actions from Set 2 had congruent or incongruent objects, again with four different actor pairs (80 stimuli). Half of these stimuli were presented mirrored, to balance agent and patient position. These 240 stimuli were distributed over four lists, with the actor pairs assigned to the four lists by means of a Latin-square design.

Each participant saw two of the four lists. Each list contained all 20 actions in all conditions, but with different actors (40 items from Set 1, 20 from Set 2). Each experimental list included 20 filler items, half with coherent back-to-back

actions, and eight warm-up actions that were not part of the experimental set. In all, each participant saw 168 action scenes. Agent position (left/right), agent and patient sex, and coherence were balanced within lists.

Apparatus

The stimuli subtended $20^\circ \times 28^\circ$ visual angle and were presented on a 21" CRT-monitor (Samsung Syncmaster 1100p, resolution: 1024×768 pixels; refresh rate: 100 Hz). Stimulus presentation and online response collection was controlled by SR Research Experiment Builder software run on an IBM-compatible computer. A game pad served as device for the manual response.

Procedure

Participants were tested individually, seated approximately 90 cm from the display monitor. In the written instruction with verbal examples, coherent actions were described as scenes with two actors involved in the same action, and incoherent actions scenes with an inappropriate object for the action or with one of the actors uninvolved in the action. Participants were asked to decide as quickly and accurately as possible whether they saw a coherent or incoherent action by pressing the right or the left button of a game pad. Each trial started with the display of a fixation cross in the center of the screen. After a random delay of 1,000, 1,500, 2,000, or 2,500 ms, an action scene was presented centered for 20, 30, 50, or 100 ms (between subjects). Action scenes were immediately followed by a perceptual mask, presented for 250 ms. It consisted of 80 uninformative small squares cut out of the filler items. Time-out was 2,000 ms. The experiment lasted about 20 minutes.

Results

The data were analyzed with mixed ANOVAs with Duration (20, 30, 50, 100 ms) as between-subjects factor, and Body Orientation (face-to-face [C+], agent-facing-patient's-back [C+], patient-facing-agent's-back [C-], and back-to-back [C-]) or action-object consistency (two levels) as a within-subjects factor. Percentage of correct responses were arcsine-transformed with replacing 100 percentage values by $(1-1/2n)$ and 0 percentage by $(1/2n)$. For post hoc comparisons, independent and paired t tests were computed; error levels were adjusted according to the Bonferroni correction using presentation durations as test families.

Body orientation

The mixed ANOVA on the arcsine-transformed proportions with the factors body orientation and duration showed a main effect of Body Orientation, $F(3, 180) = 21.65$, $p < .001$, $\eta_g^2 = .22$, and of Duration, $F(3, 60) = 51.40$, $p < .001$, $\eta_g^2 = .31$. Moreover, there was a significant interaction, $F(9, 180) = 5.53$, $p < .001$, $\eta_g^2 = .18$ (see Table 1 for descriptive data). As expected, coherence judgements were most accurate for face-to-face (C+) and back-to-back (C-) actions. Accuracy increased from chance level to clearly above-chance level with longer presentation durations. Even in the patient-facing-agent's-back (C-) condition, accuracy increased from 53 % to 95 % with longer presentation durations. The agent-facing-patient's-back (C+) condition, for which the correct response is “coherent,” showed the smallest number of correct responses (see Table 1). There seems to be a bias toward responding “incoherent” when the two scene participants have the same orientation (both either looking to the right or to the left). We therefore calculated d' as a bias-free sensitivity measure based on the difference between hits and false alarms, replacing proportions of 0 and 1 as suggested by Macmillan and Creelman (2005, p. 8).

Except for the 20 ms presentation duration, all d' values were significantly different from zero (see Fig. 2) as determined by t tests: for “orientation different” at 20 ms: $t(15) = 3.45$, $p = .021$,

95 % CI [0.16, inf], at 30 ms: $t(15) = 4.50$, $p = .003$, 95 % CI [0.71, inf], at 50 ms: $t(15) = 14.60$, $p < .001$, 95 % CI [2.29, inf], at 100 ms: $t(15) = 18.10$, $p < .001$; 95 % CI [2.35, inf]; for “orientation same” at 20 ms: $t(15) = -.88$, $p > .999$, 95 % CI [-0.44, inf], at 30 ms: $t(15) = 4.68$, $p = .002$, 95 % CI [0.26, inf], at 50 ms: $t(15) = 4.12$, $p = .005$, 95 % CI [0.44, inf], at 100 ms: $t(15) = 12.35$, $p < .001$, 95 % CI [1.64, inf].

In an ANOVA with Duration as between-subjects factor and Orientation (Same, Different) as within-subjects factor, there was a main effect of Duration, $F(3, 60) = 52.46$, $p < .001$, $\eta_g^2 = .63$, with longer durations yielding higher d' values. This shows that the ability to judge coherence improves with longer presentation times (note that this is a between-subjects factor). There was also a main effect of Orientation $F(1, 60) = 90.76$, $p < .001$, $\eta_g^2 = .35$: The Orientation-Different conditions (face-to-face [C+]; back-to-back [C-]) yielded higher d' values than the Orientation-Same conditions (agent-facing-patient's-back [C+], patient-facing-agent's-back [C-]). This indicates that the correct assessment of action coherence is much easier when the overall scene layout supports the decision. There was a significant interaction, $F(3, 60) = 9.76$, $p < .001$, $\eta_g^2 = .15$.

Semantic consistency between action and object

The repeated measures ANOVA on the arcsine-transformed proportions with the factors action-object consistency (Consistent, Inconsistent) and Duration (20, 30, 50, 100 ms) showed a large effect of Consistency, $F(1, 60) = 111.41$, $p < .001$, $\eta_g^2 = .61$, as well as a significant interaction, $F(3, 60) = 18.59$, $p < .001$, $\eta_g^2 = .44$. The effect Duration was also significant, $F(3, 60) = 3.00$, $p = .037$, $\eta_g^2 = .02$ (see Table 1 for descriptive values). Whereas the number of correct responses increased across durations in the action-object consistent condition, the reverse happened in the inconsistent condition. In fact, the action-object inconsistent scenes were judged as coherent, and increasingly so with increasing presentation durations (50 %, 72 %, 83 %, and 84 % for durations 20, 30, 50, and 100 ms).

The analysis of d' data showed that there might be some differentiation at duration 100 ms, but not at shorter presentation

Table 1 Percentage correct and 95 % CI in brackets as a function of body-orientation, action-object consistency and presentation duration

Body-Orientation	Coherence	Presentation duration			
		20 ms	30 ms	50 ms	100 ms
face-to-face	C+	53 [48, 59]	76 [71, 80]	84 [80, 88]	80 [75, 84]
agent-facing-patient's- back	C+	43 [38, 49]	60 [55, 65]	47 [41, 52]	60 [54, 65]
patient-facing-agent's-back	C-	53 [47, 58]	55 [50, 61]	78 [73, 82]	95 [93, 97]
back-to-back	C-	60 [54, 65]	60 [54, 65]	92 [89, 95]	96 [93, 98]
Action-Object Consistency					
consistent	C+	45 [40, 51]	70 [65, 75]	84 [80, 88]	91 [87, 94]
inconsistent	C-	50 [44, 55]	28 [23, 33]	17 [13, 21]	16 [12, 20]

Note. C+ = coherent action, C- = incoherent action

Fig. 2 The d' results based on hits and false alarms, as a function of Body-Orientation and Action-Object Consistency. Orientation same = agent-facing-patient's-back (C+), patient-facing-agent's-back (C-); orientation different = face-to-face (C+), back-to-back (C-); action-object consistency = consistent (C+), inconsistent (C-). Error bars indicate 95 % confidence intervals

durations: at 20 ms: $t(15) = -1.23$, $p > .999$, 95 % CI [-0.36, inf], at 30 ms: $t(15) = -0.57$, $p > .999$, 95 % CI [-0.31, inf], at 50 ms: $t(15) = 0.13$, $p > .999$, 95 % CI [-0.20, inf], at 100 ms: $t(15) = 2.88$, $p = .069$, 95 % CI [0.13, inf]. Note that in the object-action inconsistent condition, 84 % responses are still incorrect at 100 ms duration. What these data show is that the face-to-face orientation that is common to all scenes, with consistent and with inconsistent objects, seems to signal coherence. The small object details cannot be used to judge coherence correctly, even at the longest presentation duration.

There are two aspects to Experiment 1 that might have induced overall biases in responding. The first concerns the fact that body-orientation stimuli were more frequent than action-object stimuli – in fact, including the fillers, there were three times as many body-orientation than action-object trials. Second, the face-to-face orientation apparently induced a bias toward “coherent” responses. To investigate whether these two factors had influenced the results, we ran Experiment 2, with the same number of body-orientation and action-object stimuli (40 each). Next, we omitted the face-to-face stimuli of the body-orientation condition, keeping only the coherent and incoherent stimuli with the same body orientation. As a result, the only face-to-face stimuli were those from the action-object condition. We selected the longest presentation duration (100 ms), given that Experiment 1 showed an indication for a differentiation between coherent and incoherent action-object scenes at this presentation duration.

Experiment 2

Method

Participants

Twenty-two students from Münster University (18 female, ages 18–27 years) with normal or corrected-to-normal vision participated in this study for course credits or payment (3.00 €).

Materials

We used a subset of the material from Experiment 1, with two versions including 80 pictures each. Each version comprised 20 scenes with coherent body orientation (agent-facing-patient's-back; see Fig. 1B), 20 scenes with incoherent body orientation (patient-facing-agent's-back; see Fig. 1C), 20 stimuli depicted coherent action-object scenes and 20 incoherent action-object scenes. Five pictures were used as warm-up trials.

Procedure

The procedure and instruction were the same as in Experiment 1. Scene presentation duration was 100 ms. Trial order was randomized for each participant.

Results

Proportions were arcsine-transformed for inferential statistics as before. An ANOVA with two within-subject factors (Coherence: coherent vs. incoherent; Variation: body orientation vs. action-object consistency) was computed. Both main effects, Coherence: $F(1, 21) = 8.76$, $p < .01$, $\eta_g^2 = .13$; Variation: $F(1, 21) = 39.35$, $p < .001$, $\eta_g^2 = .19$, and the interaction were significant, $F(1, 21) = 67.29$, $p < .001$, $\eta_g^2 = .46$. Considering the correct responses, the difference between coherent and incoherent conditions is smaller for the body-orientation scenes (21 %) than for the action-object stimuli (55 %) (Table 2). In fact, there is a crossover interaction, with worst performance for incoherent action-object stimuli and best performance for incoherent body-orientation stimuli – a very similar interaction as the one observed for the data from the 100 ms presentation of Experiment 1.

T tests confirmed that arcsine-transformed percentage correct was lower for coherent than for incoherent body-orientations, $t(21) = 3.62$, $p = .001$; 95 % CI [0.11 .40], and

Table 2 Percentage correct and 95 % CI as a function of coherence and global-local change

Body orientation	Coherence	
consistent	C+	62 [51, 73]
inconsistent	C-	83 [76, 91]
Action-object consistency		
consistent	C+	80 [71, 89]
inconsistent	C-	25 [17, 34]

that this was the other way around for action-object relations, $t(21) = -6.37, p < .001$; 95 % CI [-0.83 -0.43].

Using d' as dependent variable, a within t test showed that global changes ($M = 0.99, SD = 0.63$) were detected better than local changes ($M = 0.06, SD = 0.20$); $t(21) = 7.25, p < .001$; 95 % CI [0.65 1.18]. T tests against zero showed that participants were sensitive to global changes, $t(21) = 7.31, p < .001$; 95 % CI [0.71, 1.26], while they lacked such sensitivity for local changes, $t(21) = 1.51, p = .147$; 95 % CI [-0.02, 0.15]. These data replicate what was observed for this presentation duration in Experiment 1, where the d' for local changes was also not significantly different from zero. Thus, despite numerical differences, the overall pattern of results was very similar for the two experiments. This shows that the results from Experiment 1 were not due to biases induced by the preponderance of body-orientation scenes, nor by the specific stimuli used in that condition.

Discussion

Our experiments demonstrate that for judging the coherence of agent-patient action scenes, coarse spatial layout information is processed far more efficiently than information on the semantic consistency between action and object. When coherence was conveyed by the global spatial layout of the action scene, the minimal presentation duration needed for correctly judging action coherence was 30 ms to 50 ms. This corresponds to the time frame suggested for extracting first information about the gist of an environmental scene (Henderson, 2005; Henderson & Ferreira, 2004). However, the fast information uptake is even more remarkable for action scenes, given that a photograph is only a snapshot of an action and does not reflect the temporal course of the event. Our data fit well with results from Hafri et al. (2013), who reported above-chance level action and role verification with 37 ms masked scene presentation.

As expected, face-to-face (C+) and back-to-back (C-) actions were judged best. As the d' data clearly showed, coherence decisions were more difficult when agent and patient faced the same direction – although performance, reflected by d' , was reliably above chance from 50 ms durations onward. The percent-correct data in Table 1 show better performance for the patient-facing-agent's-back condition than for the agent-facing-patient's-back condition. This demonstrates that factors other

than body orientation determine coherence decisions. Looking at the overall contour of the action scenes, particular layout features might explain this difference. Although we tried to realize actions without expansive forward gestures by the agent, this could not be avoided due to the nature of an agent acting upon another person. That is, the arms of an agent often reached out to some extent. Consequently, a backward-oriented agent resulted in a noncontinuous edge of the action contour, as well as in an empty interspace between the two actors (see Hafri et al., 2013, for similar features). This made the contour more open and apparently gave reason for judging the patient-facing-agent's-back actions (C-) correctly as incoherent. In contrast, agent-facing-patient's-back actions (C+) had filled interspaces combined with even contours. These two layout criteria for coherence apparently carried less weight than an empty interspace combined with a ragged contour as criteria for incoherence.

As such, this suggests that actions represent a specific type of scene, where the meaningfulness of an action can be, at least for some actions, derived from the global layout. Based on data from a multitude of indoor/outdoor scenes, Fei-Fei et al. (2007) argued that shape information precedes semantic-level information. Our manipulated action scenes show a different result, because the global layout already determines semantic aspects of the scene (i.e., meaningfulness). Apparently, this information is used early on, when the task is to judge semantic coherence.

As expected, performance was worse for scenes with variation of local scene information than for global-layout scenes. Detecting whether the action and the object were semantically consistent was hardly possible. The percent-correct data for the coherent scenes should not be taken as evidence for correct decisions on the basis of local action-object coherence. In fact, as the d' data show, there is hardly any sensitivity to local scene aspects, even more clearly so in Experiment 2. Interestingly, action scenes were judged as coherent independent of local object-action consistency. In Experiment 1, “correct” responses in consistent and inconsistent roughly corresponded to those of the (correct) face-to-face scenes in the body-orientation condition. This seems to reflect a bias of the observers to rely on the action's spatial layout when it is not possible to identify both action and object. This bias is not induced by the presence of face-to-face stimuli in the body-orientation condition, given that it was only slightly reduced in Experiment 2.

Mechanisms underlying early action scene processing

We now turn to the question as to what type of information drives the categorization of briefly presented, complex action scenes as coherent or not coherent: global spatial layout information, provided by low spatial frequencies, or detailed featural information, provided by high spatial frequencies? We propose that categorization of actions as coherent or incoherent is not based on feed-forward processing from features to objects to

scenes (Mishkin & Ungerleider, 1982), but on the immediate and parallel processing of image features relating to the action's overall spatial layout. Potentially relevant features are (a) an action's perceptual contour (symmetric or ragged), (b) body and face orientations of the actors (toward or backwards to the other person), and (c) the interspace (filled or empty) between them. We have reported earlier that action scenes presented briefly in the visual periphery can only be named correctly when their action-typical spatial layout, or Gestalt, is unambiguous, such as in “to kick someone” (Dobel et al., 2010). Similarly, in our present categorization task, a layout feature that is atypical for coherent action scenes – a patient oriented backward to the agent – predominantly led to incorrect responses.

The features just mentioned may be mainly conveyed by the low spatial frequencies specifying the global spatial layout of a scene, but it seems that such relatively coarse features do more than convey coherence: they allow the verification of event roles (agent, patient) in briefly presented scenes (see Hafri et al., 2013). In addition, we assume a strong top-down influence in that already initial visual representations are task specific (cf. Hochstein & Ahissar, 2002; Oliva & Torralba, 2006; Torralba, Oliva, Castelhana & Henderson, 2006; Treisman, 2006). In our study, the instruction to judge action coherence produced strong constraints on the extraction of information from the image. Whereas color or background texture were noncritical features for the coherence decision, the configuration of the bodies' contours, as coded by low-level spatial frequency information, was a key diagnostic cue.

The limits of this initial stage in visual processing without focused attention become apparent when coherence decisions need high spatial frequency information to detect features, as well as their binding. Whereas some such features can be detected, when they indicate the presence of ecologically relevant or emotional stimuli, their binding, allowing for object recognition, was not possible with brief and masked scene representation. Our data show that even 100 ms do not suffice to convey enough visual detail to differentiate coherent and incoherent object-action scenes. This fits with observations by Evans and Treisman (2005), who demonstrated that performance in animal/nonanimal categorization decreases, the more the feature sets of the two categories overlap. Moreover, tasks that necessitated more detailed representations than categorization, such as target localization or identification of the target's subcategory, also reduced performance.

These findings demonstrate that visual representations built on the fly are incomplete. These representations are of great value for a first impression of what is in our field of vision, and constrain the analysis of local features (Oliva & Torralba, 2006), prime semantic categories in the recognition network (Treisman, 2006), and guide the eyes to informative scene regions (Torralba et al., 2006). However, this first “best guess” (Kaping, Tzvetanov, & True, 2007) needs focused attention for generating a more detailed representation that allows verification of the first

impression. If subsequent “vision with scrutiny” (Hochstein & Ahissar, 2002) by means of eye fixations is prevented, the incomplete visual representation is prone to erroneous inference, in particular for atypical and ambiguous visual scenes.

To conclude, we have shown that extracting visual information that is useful for judging the coherence of agent-patient action scenes can be as fast as perceiving the gist of an environmental scene. Our results support the hypothesis that the representations in early visual perception, while sufficient for perceiving the spatial layout of an action scene, are insufficient for correctly understanding the relation between an action and the object used in this action.

We can only speculate to what extent the results obtained with action scenes reported here and elsewhere can be transferred to the perception of dynamic action events. Moreover, the present results go beyond visual perception per se. Action scenes serve as stimuli in the research on the interface between vision and language, specifically for sentence comprehension (e.g., Knoeferle & Crocker, 2006) and sentence production (e.g., Gleitman, January, Nappa, & Trueswell, 2007). Knowledge of the depth and abstractness of information from nonfixated scene regions will be crucial for deciding whether visual apprehension and linguistic formulation are serial processes or proceed in a temporally overlapping fashion.

Acknowledgments Thanks go to our actors, Malte Viebahn and Kerstin Funnemann, for running the experiment. We thank Dirk Vorberg for valuable comments on the manuscript. Research was supported by the Deutsche Forschungsgemeinschaft, grant DO-711. A brief report on this research appeared in the Proceedings of the European Cognitive Science Conference, Greece, 2007.

Appendix

Table 3 List of actions used in the experiment

A1. Body-orientation set		
to shove someone, to kick someone, to blindfold someone, to push someone, to scare someone, to brush someone's hair, to hit someone, to put lotion on someone, to handcuff someone, to stab someone		
A2. Probe items for the body-orientation set		
to pull someone, to kick backwards at someone, to spray someone		
B. Action-object consistency set		
	<i>Appropriate object</i>	<i>Inappropriate object</i>
to shoot someone	pistol	hand brush
to portray someone	paintbrush	croissant
to light a cigarette	cigarette	toothbrush
to serve someone a coffee	cup	shoe
to varnish someone's nails	applicator	cake fork
to help someone into a coat	coat	bin liner
to bandage someone's arm	bandage	wire mesh
to feed someone	spoon	eyeglasses
to give someone money	banknote	compact disc
to cut someone's hair	scissors	wooden spoon

References

- Bacon-Macé, N., Macé, M. J.-M., Fabre-Thorpe, M., & Thorpe, S. J. (2005). The time course of visual processing: Backward masking and natural scene categorisation. *Vision Research*, 45(11), 1459–1469. doi:10.1016/j.visres.2005.01.004
- Bar, M. (2004). Visual objects in context. *Nature Neuroscience Reviews*, 5, 617–629. doi:10.1038/nrn1476
- Biederman, I., Mezzanotte, R. J., & Rabinowitz, J. C. (1982). Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology*, 14, 143–177. doi:10.1016/0010-0285(82)90007-0
- Braun, J. (2003). Natural Scenes upset the visual appletart. *Trends in Cognitive Sciences*, 7(1), 7–9. doi:10.1016/S1364-6613(02)00008-6
- Castelhano, M. S., & Henderson, J. M. (2007). Initial scene representations facilitate eye movement guidance in visual search. *Journal of Experimental Psychology: Human Perception and Performance*, 33, 753–763. doi:10.1037/0096-1523.33.4.753
- Cohen, M. A., Alvarez, G. A., & Nakayama, K. (2011). Natural-scene perception requires attention. *Psychological Science*, 22(9), 1165–1172. doi:10.1177/0956797611419168
- Crouzet, S. M., & Serre, T. (2011). What are the visual features underlying rapid object recognition? *Frontiers in Psychology*, 2, 326. doi:10.3389/fpsyg.2011.00326
- Davenport, J. L., & Potter, M. C. (2004). Scene consistency in object and background perception. *Psychological Science*, 15, 559–564. doi:10.1111/j.0956-7976.2004.00719.x
- De Graef, P., Christiaens, D., & d'Ydewalle, G. (1990). Perceptual effects of scene context on object identification. *Psychological Research*, 52, 317–329. doi:10.1007/BF00868064
- Dobel, C., Glanemann, R., Kreysa, H., Zwitserlood, P., & Eisenbeiss, S. (2010). Visual encoding of meaningful and meaningless scenes. In E. Pedersen & J. Bohnemeyer (Eds.), *Event representation in language: Encoding events at the language-cognition interface* (pp. 189–215). Cambridge: Cambridge University Press.
- Dobel, C., Gummior, H., Bölte, J., & Zwitserlood, P. (2007). Describing scenes hardly seen. *Acta Psychologica*, 125(2), 129–143. doi:10.1016/j.actpsy.2006.07.004
- Evans, K. K., Horowitz, T. S., & Wolfe, J. M. (2011). When categories collide: Accumulation of information about multiple categories in rapid scene perception. *Psychological Science*, 22(6), 739–746. doi:10.1177/0956797611407930
- Evans, K. K., & Treisman, A. (2005). Perception of objects in natural scenes: Is it really attention free? *Journal of Experimental Psychology: Human Perception and Performance*, 31(6), 1476–1492. doi:10.1037/0096-1523.31.6.1476
- Fei-Fei, L., Iyer, A., Koch, C., & Perona, P. (2007). What do we perceive in a glance of a real-world scene? *Journal of Vision*, 7(1), 10. doi:10.1167/7.1.10. 1–29.
- Gleitman, L. R., January, D., Nappa, R., & Trueswell, J. (2007). On the give and take between event apprehension and utterance formulation. *Journal of Memory and Language*, 57(4), 544–569. doi:10.1016/j.jml.2007.01.007
- Greene, M. R., & Oliva, A. (2009). Recognition of natural scenes from global properties: Seeing the forest without representing the trees. *Cognitive Psychology*, 58(2), 137–176. doi:10.1016/j.cogpsych.2008.06.001
- Griffin, Z., & Bock, K. (2000). What the eyes say about speaking. *Psychological Science*, 11(4), 274–279. doi:10.1111/1467-9280.00255
- Hafri, A., Papafragou, A., & Trueswell, J. C. (2013). Getting the gist of events: Recognition of two-participant actions from brief displays. *Journal of Experimental Psychology: General*, 142(3), 880–905. doi:10.1037/a0030045
- Henderson, J. M. (Ed.). (2005). Real-world scene perception. *Visual Cognition*, 12(6). doi:10.1080/1350628044000544
- Henderson, J. M., & Ferreira, F. (2004). Scene perception for psycholinguists. In J. M. Henderson & F. Ferreira (Eds.), *The interface of language, vision, and action: Eye movements and the visual world* (pp. 1–58). New York, NY: Psychology Press.
- Hochstein, S., & Ahissar, M. (2002). View from the top: Hierarchies and reverse hierarchies in the visual system. *Neuron*, 36, 791–804. doi:10.1016/S0896-6273(02)01091-7
- Hollingworth, A. (2007). Object-position binding in visual memory for natural scenes and object arrays. *Journal of Experimental Psychology: Human Perception and Performance*, 33, 31–47. doi:10.1037/0096-1523.33.1.31
- Hollingworth, A., & Henderson, J. M. (1998). Does consistent scene context facilitate object perception? *Journal of Experimental Psychology: General*, 127(4), 398–415. doi:10.1037/0096-3445.127.4.398
- Joubert, O. R., Fize, D., Rousselet, G. A., & Fabre-Thorpe, M. (2008). Early interference of context consistence on object processing in rapid visual categorization of natural scenes. *Journal of Vision*, 8(13), 1–18. doi:10.1167/8.13.11
- Junghöfer, M., Bradley, M. M., Elbert, T. R., & Lang, P. J. (2001). Fleeting images: A new look at early emotion discrimination. *Psychophysiology*, 38(2), 175–178. doi:10.1111/1469-8986.3820175
- Kaping, D., Tzvetanov, T., & Treue, T. (2007). Adaptation to statistical properties of visual scenes biases rapid categorization. *Visual Cognition*, 15, 12–19. doi:10.1080/13506280600856660
- Keuper, K., Zwanger, P., Nordt, M., Eden, A., Laeger, I., Zwitserlood, P., . . . Dobel, C. (2014). How 'love' and 'hate' differ from 'sleep': Using combined electro/magnetoencephalographic data to reveal the sources of early cortical responses to emotional words. *Human Brain Mapping*, 35(3), 875–888. doi:10.1002/hbm.22220
- Keuper, K., Zwitserlood, P., Rehbein, M. A., Eden, A. S., Laeger, I., Junghöfer, M., . . . Dobel, C. (2013). Early prefrontal brain responses to the hedonic quality of emotional words: A simultaneous EEG and MEG study. *PLoS One*, 8(8). doi:10.1371/journal.pone.0070788
- Kimchi, R. (1992). Primacy of wholistic processing and global/local paradigm: A critical review. *Psychological Bulletin*, 112(1), 24. doi:10.1037/0033-2909.112.1.24
- Kirchner, H., & Thorpe, S. J. (2006). Ultra-rapid object detection with saccadic eye movements: Visual processing speed revisited. *Vision Research*, 46(11), 1762–1776. doi:10.1016/j.visres.2005.10.002
- Knoeferle, P., & Crocker, M. W. (2006). The coordinated interplay of scene, utterance, and world knowledge: Evidence from eye tracking. *Cognitive Science*, 30(3), 481–529. doi:10.1207/s15516709cog0000_65
- Li, F. F., VanRullen, R., Koch, C., & Perona, P. (2002). Rapid natural scene categorization in the near absence of attention. *Proceedings of the National Academy of Sciences*, 99(14), 9595–9601. doi:10.1073/pnas.092277599
- Li, F. F., VanRullen, R., Koch, C., & Perona, P. (2005). Why does natural scene categorization require little attention? Exploring attentional requirements for natural and synthetic stimuli. *Visual Cognition*, 12(6), 893–924. doi:10.1080/1350628044000571
- Macmillan, N. A., & Creelman, C. (2005). *Detection theory: A user's guide* (2nd ed.). Mahwah: Erlbaum.
- Mishkin, M., & Ungerleider, L. G. (1982). Contribution of striate inputs to the visuospatial functions of parieto-preoccipital cortex in monkeys. *Behavioural Brain Research*, 6(1), 57–77.
- Oliva, A. (2005). Gist of the scene. In L. Itti, G. Rees, & J. K. Tsotsos (Eds.), *The encyclopedia of neurobiology of attention* (pp. 251–256). San Diego: Elsevier.
- Oliva, A., & Schyns, P. G. (1997). Coarse blobs or fine edges? Evidence that information diagnosticity changes the perception of complex

- visual stimuli. *Cognitive Psychology*, 34, 72–107. doi:[10.1006/cogp.1997.0667](https://doi.org/10.1006/cogp.1997.0667)
- Oliva, A., & Torralba, A. (2006). Building the gist of a scene: the role of global image features in recognition. *Progress in Brain Research*, 155(B), 23–36. doi:[10.1016/S0079-6123\(06\)55002-2](https://doi.org/10.1016/S0079-6123(06)55002-2)
- Potter, M. C. (1976). Short-term conceptual memory for pictures. *Journal of Experimental Psychology: Human Learning and Memory*, 2, 509–522. doi:[10.1037/0278-7393.2.5.509](https://doi.org/10.1037/0278-7393.2.5.509)
- Rousselet, G. A., Fabre-Thorpe, M., & Thorpe, S. J. (2002). Parallel processing in high-level categorization of natural images. *Nature Neuroscience*, 5(7), 629–630. doi:[10.1038/nn866](https://doi.org/10.1038/nn866)
- Sanocki, T., & Epstein, W. (1997). Priming spatial layout of scenes. *Psychological Science*, 8, 374–378. doi:[10.1111/j.1467-9280.1997.tb00428.x](https://doi.org/10.1111/j.1467-9280.1997.tb00428.x)
- Schupp, H. T., Öhman, A., Junghöfer, M., Weike, A. I., Stockburger, J., & Hamm, A. O. (2004). The facilitated processing of threatening faces: An ERP analysis. *Emotion*, 4, 189–200. doi:[10.1037/1528-3542.4.2.189](https://doi.org/10.1037/1528-3542.4.2.189)
- Schyns, P., & Oliva, A. (1994). From blobs to boundary edges: Evidence for time-scale-dependent and spatial-scale-dependent scene recognition. *Psychological Science*, 5(4), 195–200. doi:[10.1111/j.1467-9280.1994.tb00500.x](https://doi.org/10.1111/j.1467-9280.1994.tb00500.x)
- Stolarova, M., Keil, A., & Moratti, S. (2006). Modulation of the C1 visual event-related component by conditioned stimuli: Evidence for sensory plasticity in early affective perception. *Cerebral Cortex*, 16, 876–878. doi:[10.1093/cercor/bhj031](https://doi.org/10.1093/cercor/bhj031)
- Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature*, 381(6582), 520–522. doi:[10.1038/381520a0](https://doi.org/10.1038/381520a0)
- Thorpe, S., Gegenfurtner, K., Fabre-Thorpe, M., & Bülthoff, H. (2001). Detection of animals in natural images using far peripheral vision. *European Journal of Neuroscience*, 14(5), 869–876. doi:[10.1046/j.0953-816x.2001.01717.x](https://doi.org/10.1046/j.0953-816x.2001.01717.x)
- Torralba, A., Oliva, A., Castelano, M. S., & Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review*, 113(4), 766–786. doi:[10.1037/0033-295X.113.4.766](https://doi.org/10.1037/0033-295X.113.4.766)
- Treisman, A. (2006). How the deployment of attention determines what we see. *Visual Cognition*, 14(4/8), 411–443. doi:[10.1080/13506280500195250](https://doi.org/10.1080/13506280500195250)
- Vo, M. L.-H., & Henderson, J. M. (2011). Object-scene inconsistencies do not capture gaze: Evidence from the flash-preview moving-window paradigm. *Attention, Perception, and Psychophysics*, 73(6), 1742–1753. doi:[10.3758/s13414-011-0150-6](https://doi.org/10.3758/s13414-011-0150-6)