Psychologie

To See or not to See - Action Scenes out of the Corner of the Eye

Inaugural-Dissertation
zur Erlangung des Doktorgrades
der
Philosophischen Fakultät
der
Westfälischen Wilhelms-Universität
zu
Münster (Westfalen)

vorgelegt von

Reinhild Glanemann

aus Hamm

2007

Tag der mündlichen Prüfung: 28. März 2008

Dekan: Prof. Dr. Dr. h.c. Wichard Woyke

Referent: Prof. Dr. Pienie Zwitserlood

Korreferent: Prof. Dr. Markus Lappe

Table of Contents

Introduction

CHAPTER 1

Everyday vision is fascinating. We can recognize a familiar face from millions of different ones, and our visual system can adapt to the different degrees of luminance encountered when skiing on a glacier or finding our way through near darkness. Moreover, at any point in time, we experience our visual world as being complete, continuous, highly detailed and stable, despite the fact that the images, which are projected upon the retina by a steady alternation of saccades and fixations of the eyes, are only discrete snapshots of our surroundings, with only the central two degrees of visual angle being acute.

However, this is only one extreme of the broad spectrum of human visual performance, namely the high-performance end. At the other extreme, there are the striking phenomena of change blindness and inattentional blindness, which reveal the limits of visual cognition: Substantial changes within our field of view go undetected when the change is unexpected or when we do not attend to the changing image region (for reviews see Rensink, 2002; Simons & Rensink, 2005). These phenomena demonstrate that visual cognition is not a passive and completely automatic, but an active and dynamic process, largely dependent on such factors as attention, knowledge, expectation and intention.

One topic in the research on visual cognition, which is particularly relevant to the present experiments, is the nature and detailedness of internal states that are thought to represent the external visual world, the so-called *internal visual representations*.

In this dissertation, I studied the early visual representations of complex visual scenes. More specifically, I was interested in the type of information that can be extracted from very briefly presented photographs depicting two people acting in a (meaningful or meaningless) action. These photographs were presented in a manner that prevents eye fixations on any detail of the action scene. By using stimulus exposure times of 150 ms and less, this work is devoted to the high-performance end of visual perception.

Now, what is special about visual scenes and why are action scenes particularly relevant for experimental research in cognitive psychology? I intend to answer these two questions in the remainder of this introductory section. Furthermore, I briefly introduce the two research projects reported in Chapters 2 and 3.

The ultimate goal of vision research is undoubtedly to understand the cognitive processes underlying everyday vision. One approach to understanding how we perceive our enormously complex, often moving and rapidly changing visual surroundings, is to break down the large variety of visual information into its components. Most vision research has adopted this approach.

> "[The] ultimate purpose [of visual perception] is to allow one to know what objects are present so as to behave appropriately and in accordance with one's current behavioural goals."
>
> (Yantis, 2001, p. 1)

This approach yields invaluable and detailed knowledge about the complex processes underlying vision, ranging from the so-called low-level processing of basic visual features, such as colour or orientation, to high-level processes, such as object categorization and identification.

Compared to the large body of research devoted to the perception of (static or moving) single objects, the study of more complex visual stimuli has, thus far, received much less attention. Clearly, the visual world that surrounds us consists not only of single objects. We are surrounded by inanimate and animate objects that are usually parts of scenes and events.

In the following, I provide definitions of some key terms relevant to my dissertation. By the term *environmental scene*, I refer to a "human-scaled view of a real-world environment comprising background elements and multiple discrete objects arranged in a spatially licensed manner" (Henderson, 2005, p. 849), such as beach, kitchen, party, classroom, underwater world and so forth. An *event* is even more complex than an environmental scene, in that it involves a change of state that unfolds over time, such as a thunderstorm. Thus, compared to objects and scenes, an event has an additional temporal aspect. If the event is controlled by a living entity, called *agent,* it is referred to as *action event*, such as 'A is kicking a ball'. Due to their temporal aspect, events are best depicted by dynamic stimuli such as film sequences. However, the pre-testing of our materials demonstrated that a static snapshot of an action event, which captures its characteristic properties, can satisfactorily activate the corresponding memory entry of the represented action. The stimulus material for all studies and experiments in this dissertation are photographs of action events. These photographs are referred to as *action scenes*.

So, why not transfer what is known about object perception to environmental and action scenes? After all, are scenes not just simply collections of objects? The answer is "no", and this is why scenes are important for researching human visual perception. A scene is more than just the sum of its parts. The specific spatial and semantic combination of the scene's components conveys additional meaning beyond simple co-occurrence. For example, a typical arrangement of wooden benches, long tables with plaid tablecloths and large mugs is easily recognized as a beer garden. Similarly, the specific spatial arrangement of sand, water and sky is immediately perceived as a beach. Indeed, research on scene perception suggests that the so-called *gist* of a scene, here 'beer garden' and 'beach', is processed in a different manner by the human visual system than objects. Evidence from behavioural, computational and neuroimaging studies (reviewed by Oliva & Torralba, 2006) demonstrates that global scene information, that is, the spatial layout of the scene's components, plays a significant role in apprehending the scene's gist.

> "…, just as actors cannot act without a stage, objects cannot appear except within the context of a scene. Thus, one salutary aspect of studying scene perception is that it expands our conception of what vision is for. Vision scientists have spent many years studying the actors; now it is time to direct some attention to the stage."
>
> (Epstein, 2005, p. 974)

Taken together, next to asking what type of information can be extracted from briefly presented action scenes, one can also ask whether the same or similar mechanisms underlying the perception of environmental scenes also apply for the perception of action scenes.

As described above, action scenes constitute a specific type of complex scene. The fact that action scenes comprise an additional temporal dimension renders them more complex than environmental scenes. Thus, the study of static action scenes may be a good start to investigate the perception and cognition of dynamic action events.

The action scenes used here depict two human participants involved in a joint action. All actions are of the agent-patient type, that is, one protagonist (*agent*) is acting upon another protagonist (*patient*). Imagine person A taking a photograph of person B. Similar to static environmental scenes, the specific arrangement of agent and patient (and of an optional object) in a spatially and semantically licensed manner conveys additional meaning over and above the scene's elements, namely the action. Violations with respect to spatial or semantic 'laws' render environmental scenes and action scenes incoherent and meaningless.

There is yet another motivation for studying the visual perception of action scenes, which also is relevant to the experiments reported in this dissertation. Action scenes can be used to study visual perception per se, but they can also serve as stimulus material in research on other cognitive functions, such as memory, attention, or language. In particular, they are of growing interest for the interface between language and vision.

For the study of language comprehension and production at the sentence level, eye-tracking is becoming a pet paradigm. As an online-method it reveals what is in the focus of visual attention at any time during a given task. For example, in research on sentence comprehension, eye movements can indicate the moment at which an ambiguous speech input is disambiguated by the listener (e.g., Kamide, Altmann & Haywood, 2003). In research on speech production, it was observed that people tend to look at scene components roughly one second before mentioning them (e.g. Griffin, 2004). In other words, the eyes are about one second ahead of speech. Thus, eye-tracking can be used to examine speech-planning processes in more detail.

However, to understand the temporal coupling between eye movements and speech planning, we need to know how tight the link is between eye fixations and cognitive processes. In other words, are higher cognitive processes, such as recognition and speech planning, restricted to the fixated scene region? For example, when the task is to produce a sentence that describes an action scene, and the first fixation goes to the head of the agent, it is important to know which aspects of the action were already identified before initiating this eye movement. Is the agent fixated first due to its visual salience? Or were thematic roles, or even the depicted action itself, identified by peripheral vision before the eyes started moving? The first case would imply that it is visual salience only that guides first fixations. In the latter case, the initial eye movement may indicate what the mind has already chosen as a suitable starting point for sentence production. This touches upon the issue as to whether visual scene apprehension and sentence formulation are timely distinct processes, as suggested by Griffin and Bock (2000), or whether these processes occur in parallel, possibly with mutual influence, as suggested by Gleitman, January, Nappa, and Trueswell (in press). Previous research suggests that cognitive processes, such as object identification (reviewed by Irwin, 2004), utterance planning (Bock et al., 2003), and possibly even lexical and phonological planning (Morgan & Meyer, 2005) are not restricted to visual information at the current locus of fixation.

To address such questions, the studies in *Event Conceptualization in Free View and at an Eyeblink* (Chapter 2) examined the rapid extra-foveal uptake of verb-related information from action scenes involving two actors. By registering eye-movements and using brief extra-foveal presentation, the first two studies investigated whether thematic roles of the two actors could be identified before the action scene was fixated for the first time. The third and forth study were concerned with the identification and naming of the depicted action. In addition to brief presentation, Study 4 employed blurry versions of the stimuli used in Studies 1 – 3, to simulate the reduced acuity of peripheral vision.

One of the first studies that examined the rapid perception of action scenes used coherent and incoherent actions (Dobel, Gumnior, Bölte & Zwitserlood, 2007). Coherence was manipulated by mirroring both involved actors, rendering an action scene either meaningful or meaningless. When viewers were presented with these action scenes for only 100 ms, they could correctly judge on coherence in 80 % of the cases. However, identification and naming of the components of the scene, such as agent, recipient, action and the involved object, was clearly worse. These results suggest that the decision on coherence was made on the basis of global scene properties rather than by identifying scene components first.

As a follow-up, the experiments in *Rapid Apprehension of Coherence of Action Scenes* (Chapter 3) investigated in more detail what type of scene information is most relevant to coherence judgements. With presentation times between 20 and 100 ms, two types of manipulation were employed. One manipulation altered the global spatial layout of the action, by mirroring the two actors. In contrast to the Dobel et al. (2007) study, actors were mirrored individually, resulting in four instead of two different body-orientation combinations. The second manipulation concerned the object used in the action. By using an appropriate or inappropriate object for a given action, coherence was varied as a function of the semantic consistency between action and object.

The results of the investigations reported in the following two chapters demonstrate that internal visual representations of action scenes can be built up extremely fast.

Although presentation times were too brief to focus attention on any scene region, these representations were detailed enough to extract essential semantic information from the depicted action. However, we will also see where this rapid processing reaches its limits. Chapter 4 discusses the general conclusions and implications of my results, taking also into consideration the limits and loose ends of the presented experiments.

Event Conceptualization In Free View And At An Eyeblink

CHAPTER 2

Abstract

In what detail can action events be recognized without taking a close look? In four studies, we examined the rapid peripheral uptake of visual information from naturalistic, meaningful action scenes of agent-patient-type. In Study 1, participants indicated the patient's location by pressing a button while their eye movements were monitored. These movements revealed a tendency to fixate agents first. In most cases both actors were inspected before an overt response was made. However, in Study 2, a brief peripheral presentation (150 ms) of stimuli also produced highly accurate answers (93 % correct). Study 3 showed that even actions could be named in 58 % of the cases with 150 ms exposure. Study 4 demonstrated that correct action naming depended on whether the global scene layout, in particular body posture, allowed only a few alternative actions. Apparently, visual event recognition is characterized by a rapid peripheral analysis of the scene's spatial layout that allows for role and, in many cases, action identification, followed by a period of overt attention shifts that are necessary for perceiving more detailed information.

Introduction

A large body of research deals with the question of how viewers extract meaningful information from complex visual scenes. This is of interest not only for models of vision and attention, but also for other cognitive research areas employing visual scenes as stimulus material, as is the case for the research on language processing. For example, in naming depictions of multiple objects and actions, the temporal relationship between eye fixations and speech provides insight into the underlying speech planning and execution processes (for reviews, see Griffin, 2004; Meyer & Dobel, 2004).

Research in scene perception has yielded, amongst others, two results that are important for psycholinguistic research. First, the *gist* of a scene, that is, its general semantic interpretation including some aspects of the global spatial layout, can be extracted very rapidly (for a review, see Henderson & Ferreira, 2004). Second, some semantic categorization tasks can be accomplished even when the informative parts of the scene are located outside the small area of fixation (e.g., Thorpe, Gegenfurtner, Fabre-Thorpe, & Bülthoff; 2001).

Although there is growing interest in both the rapid perception of scenes and eye movements during language production, researchers seldom look into both aspects of processing. However, the extent to which certain scene regions, and thus potential constituents of a sentence, can be processed without fixation is of significant interest for interpreting the tight linkage between eyes and mouth.

Therefore, the present study uses a specific type of action scene that has repeatedly been used in studying the interface between language and vision, namely *agent-patient actions*. These actions consist of two human actors: the agent, who performs the action, and the patient, who is acted upon (e.g. A is kicking B). Our aim is to investigate how much verb-relevant information can be extracted prior to a first fixation into a meaningful area of the scene. More specifically, we address the following questions. First, are the thematic roles of the two participants of the action (agent, patient), or even the verb corresponding to the action, accessible from

peripheral vision? Second, does the peripheral uptake of visual information guide the first fixation into the scene? Decoding a depicted action and assigning thematic roles are prerequisites for the verbalization of actions. Thus, the focus of our present research is on the interface between vision, perception of actions and conceptualization of thematic roles.

In what follows, we first summarise some basic findings on vision and attention that are relevant to our research. Next, we summarize data on rapid scene perception, before turning to language production. We review the few studies on eye movements related to speech production, and discuss how their data are in line with current models of language production.

Vision and Attention

It is important for the relationship between eye-movements and language to keep in mind that visual recognition is not limited to foveal viewing. Acuity is highest in the centre of the fovea and falls off rapidly and continuously with growing distance from this centre. However, this does not imply that no recognition is possible outside the fovea, which comprises barely 2° of visual angle (Irwin, 2004). The region around a fixation from which useful information can be extracted for a given task, the so-called functional field of view, is quite variable in size and can extend far into the visual periphery. The size of the functional field of view in scene perception depends on a variety of factors such as task, stimulus type and display complexity (for a review, see Henderson & Ferreira, 2004; Irwin, 2004).

How does the functional field of view interact with attentional processes during the rapid extraction of information from visual scenes? Distinguishing between covert and overt deployment of attention, covert or 'hidden' attention can be directed to a region in the visual scene before the eyes move, which allows for enhanced processing of this region (Hoffman, 1975; Posner, 1980). A covert shift of attention always precedes overt allocation of attention to this location, the latter by moving the eyes and, if necessary, the head (Deubel & Schneider, 1996; Hoffman

& Subramaniam, 1995; Shepherd, Findlay, & Hockey, 1986; Wolfe, 1998). Thus, eye-movements can be taken as an on-line indicator as to where covert attention was located just before a saccade was executed. In this way, we can gain insight in the attentional system, which makes studying eye-movements highly attractive in cognitive research.

Rapid Scene Perception

How well peripheral information can be processed up to the conceptual level with very short exposure was impressively shown by Thorpe et al. (2001). Naturalistic photographs of animals, vehicles, buildings, and so on, were classified at high levels of accuracy in an animal versus non-animal (vehicle vs. non-vehicle, etc.) decision task, when presented for a mere 28 ms (unmasked), and with an eccentricity of 57°. Clearly, the short presentation time prevented overt attention shifts. Li, VanRullen, Koch, and Perona (2002) went one step further and demonstrated that the same categorization could be performed at 6° eccentricity in a dual-task situation, with almost no covert attention.

The fast processing of visual scenes was also studied with tasks other than categorization, such as the object-detection paradigm (e.g., Biederman, 1972; Biederman, Mezanotte, & Rabinowitz, 1982; Davenport & Potter, 2004) and with rapid serial visual presentation (e.g., Intraub, 1981; Potter, 1976). With these methods, it was repeatedly shown that the gist of a scene can be recognized within a single glance.

In our earlier work, we demonstrated that apprehending meaningfulness of a depicted action can also be accomplished with very brief presentation times (Dobel, Gumnior, Bölte, & Zwitserlood, 2007). In this study, line-drawings of two animate actors being involved in the same action were presented randomly in one of the four quadrants of the computer screen, with a centred pre-trial fixation cross. Meaningfulness of actions was manipulated by individually mirroring both actors.

With 100 ms masked presentation, meaningfulness was correctly judged in most cases. These results point to rapid parafoveal extraction of the event structure.

Scene perception can also be studied with early eye-movements after image onset (e.g., Antes, 1974; Castelhano & Henderson, 2007; Underwood & Foulsham, 2006). In inspection and visual search tasks, the eyes were almost immediately oriented to task-specific scene regions, which supports the idea that the gist is available during (and, in principle, even before) the first fixation and is used to guide subsequent eye-movements. But what guides eye-movements (and attention) in scene perception? Two classes of factors seem to be important. First, exogenous, stimulus-based features such as colour, orientation, motion or luminance, attract attention (for a review, see e.g., Parkhurst, Law, & Niebur; 2002). Second, endogenous, cognitive factors lead to specific eye-movement patterns on a scene. Examples of such cognitive factors are task requirements (Buswell, 1935; Yarbus, 1967), knowledge about the scene and its typical layout (Torralba, Oliva, Castelhano, & Henderson, 2006), or consistency of scene gist and objects in this scene (e.g. Biederman et al., 1982; De Graef, 2005; see Henderson, Weeks, & Hollingworth, 1999, for different findings). In a visual search task, Underwood and Foulsham (2006) showed that, even in early eye movements, task demands can reduce the visual saliency weights of objects in a scene. They termed this effect "cognitive override". It is frequently argued that visual stimulus characteristics are particularly influential in guiding early eye-movements, whereas cognitive factors increasingly come into play during the course of viewing the scene (Egeth & Yantis, 1997; Henderson & Ferreira, 2004; but see Underwood & Foulsham, 2006). Thus, when investigating the interplay of top-down (cognitive) and bottom-up (stimulus-driven) factors in scene perception, time, from stimulus onset, seems to be a critical aspect (Henderson et al., 1999). Recently, Kirchner and Thorpe (2006) employed a forced-choice saccade task for animal detection in natural scenes with parafoveal presentation. They found that the initiation of an appropriate saccade can be accomplished within as little as 120 ms. The authors ruled out the potential influence of differences in low-level image

descriptors between targets and distractors, that is, it was only the semantic content of the stimulus that controlled the direction of the first saccade.

The present study investigates whether a 'cognitive override' occurs in the very first fixation into an action scene. This contrasts with most of the scene-perception studies cited above, which allowed at least one initial fixation into the stimulus before assessing eye-movements, button presses, or verbal answers as indices of what had been perceived up to that moment.

Language Production and Eye Movements

The results from studies on the fast perception of scene gist seem to indicate that the uptake of peripheral visual information can result in first fixations, which are driven more by task requirements than by visual-perceptual salience. This assumption is supported by data from Bock, Irwin, Davidson, and Levelt (2003), who examined initial gaze behaviour in a clock-time-telling task. Participants were instructed to tell the time from analogue or digital time displays. Depending on the speaker's preference and the specific instruction, the time expression varied with respect to hour and minute order, that is, either hour-first ("ten fifteen") or minute-first ("fifteen past ten"). The authors observed that within 400 ms or less after clock exposure the eyes targeted that part of the display, which provided the first-named information of the verbal expression, which unfolded only afterwards. The conclusion from these results was that the eyes were not primarily influenced by perceptual prominence which would have been the larger minute hand in the analogue displays or the leftmost digit in digital time displays. Instead, the properties of the upcoming utterance determined the direction of initial eye movements, no matter whether the expression started with the minute or the hour information. Thus, the structure of the expression in mind guided the first saccade to the informative part of the clock display. The authors took this as evidence for first fixations being task-driven, in that the eyes reveal the starting-point of the utterance under planning.

These results fit with data from Griffin & Bock (2000), who were the first to examine eye-movements related to sentence production. They observed that the speakers' gaze anticipated the first constituent of their subsequent utterance before they started speaking. Critically, the constituent selected as starting point for the utterance did not correspond to the most salient component of the depicted action (which was shown by a control condition of mere viewing). Thus, the eyes indicated which sentence structure had apparently already been selected, instead of choosing the visually most salient constituent as the sentence's subject. In contrast to Bock et al.'s (2003) clock-telling findings, initial fixations were not yet driven by the structure of the upcoming sentence. In a non-linguistic control task, Griffin & Bock examined how long it took to extract the action's causal structure. For this purpose, participants had to fixate the patient of an action as quickly as possible after stimulus onset *(patient detection task)*. This task was accomplished within about the same time frame as the selection of a starting point in the sentence production task. Thus, in both the time-telling (Bock et al., 2003) and the sentence production task (Griffin & Bock, 2000), the first few hundred milliseconds served for general comprehension of the event and the rapid extraction of the scene structure *(= apprehension phase)*, before the eyes targeted the sentence constituents in their order of mention *(= formulation phase)*.

Similar observations were made by Van der Meulen (2001) in a multiple-object description task, and by Meyer & Dobel (2004) in a sentence-production task. The eyes initially inspected parts of the display that did not correspond to the first constituent of the ensuing utterance before they then fixated objects and scene elements in order of their mentioning. By labeling this initial period *preview phase*, Van der Meulen emphasized that these fixations are carried out for planning the conceptual and syntactic structure. In contrast, the term *apprehension phase* puts more emphasis on the cognitive processes taking place before task-relevant regions of the display are fixated. Note that, unlike for the *preview phase*, eye-movements are not a mandatory component of *apprehension* (Bock et al., 2003).

In current models of sentence production, *apprehension* and *formulation* could correspond to non-linguistic and linguistic planning levels, respectively (cf. Bock & Levelt, 1994; Bock et al., 2003; Garrett, 1976; Levelt, 1989). Apprehending an action scene means understanding the causal structure of the action and assigning thematic roles, such as action, agent, patient or recipient, to the scene's entities. During speech planning, we attend more closely to language-specific aspects than to other features of the (visual) surroundings (Jackendoff, 1997; Slobin, 1996), a phenomenon termed *thinking for speaking* (Slobin, 1996). Analogously, one can ask whether *seeing for speaking* occurs. Bock et al.'s (2003) results seem to point in this direction. Remember that the eyes were rapidly sent to the scene region containing the relevant information for the first sentence constituent. However, the authors also showed that eye-movements were not mandatory for telling the correct time. The same clock-telling task was accomplished to near perfection with stimulus duration as short as 100 ms, which is too short for eye-movements into the display. Although these results seem intriguing, it should be kept in mind that time-telling is a highly overlearned task, and a simple one, compared to describing an action scene with a sentence.

Evidence for a parafoveal preview benefit was also found for object naming in studies employing the 'boundary technique' (Morgan & Meyer, 2005; Pollatsek, Rayner, & Collins, 1984; Rayner, 1975). The boundary technique allows for changes to occur in target pictures when the eyes cross an (invisible) boundary on the way from one display element to the next. This method revealed that the word form of an object was already activated before a first fixation landed on this object (Morgan & Meyer, 2005; Pollatsek et al., 1984; but see Griffin & Spieler, 2006, for an alternative interpretation). For the idea of *seeing for speaking*, it can be assumed that *seeing* comprises both foveal and extra-foveal visual perception.

In sum, several studies have demonstrated that the uptake of language-relevant visual information can be extremely fast. It thus seems crucial to assess the level up to which (a specific part of) the stimulus is already processed *before* it is fixated to be able to interpret the close time-locking of eye-movements and speech production

(and comprehension). Whereas extra-foveal pre-processing was already shown for line drawings of single objects (Morgan & Meyer, 2005), for clock displays (Bock et al., 2003), and for identifying actors in line-drawings of actions (Dobel et al., 2007), none of these studies employed naturalistic stimulus material. The use of naturalistic action scenes in the present study is expected to facilitate fast visual recognition (cf. Braun, 2003).

The question we ask is, whether thematic roles and possibly even the action itself in naturalistic two-participant action scenes can be apprehended within the first 100 - 200 ms after stimulus onset. If so, this would support models of scene perception that propose a fast and holistic uptake of scene information at image onset before local features of the scene are identified subsequently (e.g., Hochstein & Ahissar, 2002; Torralba et al., 2006). In this case, we expect, in parallel to Bock et al.'s (2003) findings for less complex displays, that peripheral *apprehension* is sufficient to direct the first gaze to the search target, which is the patient in our study. If apprehension can not be achieved peripherally, we expect initial fixations to be equally distributed over agent and patient. This question is also of methodological interest. If an action and its causal structure could only be identified when action-relevant regions of the scene are fixated, the eye movement record would indicate when action information became available to the speaker. Consequently, we could infer the earliest point in time when the speaker accesses speech-relevant information. If, on the other hand, identification is already possible prior to fixation, the interpretation of eye-movements related to action scene perception and language production is much more sophisticated.

Overview of Studies

In the course of four studies, we manipulated presentation time, task, and blurring of naturalistic action scenes. We investigated whether verb-relevant information can be extracted during an initial, peripheral viewing phase. In Study 1, we assessed whether role assignment is possible during such a preview phase, which should lead

to task-specific initial gazes. If the task is to detect the patient and the preview benefit is sufficient to determine thematic roles, gazes should be predominantly guided directly to the patient. Study 2 tested whether viewers are able to report which of the two actors is the patient even on the basis of 150 ms peripheral presentation, which is too brief to carry out a saccade. Studies 3 and 4 are of a more explanatory nature. Given the overwhelming evidence from Study 2 that thematic roles can indeed be identified, we assessed, in a post-hoc manner, whether the actions themselves can be perceived and reported after brief, peripheral scene presentation. Study 4 assessed how thematic role assignment and action identification in Study 2 and 3 could have been accomplished. For this purpose, we simulated the reduction of acuity in the periphery in our stimuli, by presenting them in blurred, black-and-white form.

Study 1: Patient Detection with Unlimited Exposure and First Gazes

Study 1 investigated the peripheral preview benefit in an event apprehension task. The question was whether initial eye movements into an action scene are guided by peripherally processed scene information. In particular, are such initial gazes influenced by a task that involves the identification of scene elements which play a particular role in non-linguistic and linguistic planning for speaking? We used a variant of the patient-detection task employed by Griffin and Bock (2000) for this purpose. Participants had to decide which of the two actors involved in an action was the patient by pressing a button corresponding to the patient's location in the scene (left or right). The dependent measure is different from the one used by Griffin and Bock, who required their participants to fixate on the patient as quickly as possible. The latter instruction might have lead to controlled eye-movement behaviour. Therefore, we used button presses to keep eye movements as natural as possible and strictly task-driven. If thematic roles can be apprehended within the first 150 ms of extra-foveal viewing, we expected the majority of first gazes to fall on the patient region.

Method

Participants. The participants were 16 students at the University of Münster, native speakers of German, and aged between 20 and 31. They received either course credit or were paid for their participation. All reported normal or corrected-to-normal visual acuity.

Stimuli & Design. Digital colour photographs of 20 action scenes were used, each depicting two actors involved in a meaningful action in front of the same neutral background (see Figure 1, and Appendix for a list of all actions). As each action was performed by two different pairs of actors, this resulted in 40 pictures. The material was shown to 40 students in a pre-test whose task was to name each depicted action. Each participant saw only one version of an action. Eight actions were not unmistakably identified. They either had less than 80% naming agreement (synonyms allowed), or the chosen verb did not unambiguously differentiate the thematic roles (agent and patient) in at least 90% of trials. These actions were excluded from the experimental set. The remaining 24 photographs (12 actions x 2 actor pairs) were also mirrored to counteract effects of preferred (left-to-right) scanning direction (cf. Buswell, 1935). Normal and mirrored pictures were distributed over two lists, with two randomized versions of each list. Each participant saw each action twice, but with different actors, once with the patient on the left and once with the patient on the right. The size of the images on the screen subtended a 22.2° horizontal and a 16.7° vertical visual angle. The mean distance between the pre-trial fixation point and the heads of the actors was 6.8° (patients: 7.6°, sd: 1.5; agents: 6.0°, sd: 1.6), and 5.6° (sd: 2.0) between the fixation point and the object, which was present in 10 of the 12 actions. Together with six practice trials, this resulted in 30 experimental stimuli.

Apparatus. The pictures were displayed on a 21-inch Samsung Syncmaster 1100p monitor with a screen resolution of 1024 x 768 pixels and a refresh rate of 100 Hz. The software 'SR Research Experiment Builder' was used to run the study. Eye-movements were recorded with the head-mounted 'SR Research Ltd. EyeLink II' system, operating at a sampling rate of 250 Hz and with a spatial accuracy of better

than 0.5° visual angle. Recording was controlled by an IBM-compatible PC, and an EyeLink button box was used for manual responding.

Procedure. The viewing distance between participants and the screen of the monitor was approximately 80 cm. The task was to decide as quickly as possible and to indicate by button press (left or right) on which side of the picture the patient of the action was located (left or right). After the participants read the instruction with example pictures, the room was shaded for the duration of the study. A nine-point-calibration and validation procedure was performed first. At the beginning of each trial, participants had to fixate a fixation cross centred at the top of the rectangle in which the picture would appear. This served as drift correction for the software. We randomly varied the presentation time of the fixation cross (2000 ms +/- 1000 ms) to minimize premature saccade onsets due to expectation. At the disappearance of the fixation cross the picture was displayed until the participants made their decision by button press. Because the eyes were at the fixation cross when the picture appeared, an initial phase of peripheral viewing was guaranteed, which was essential for our purposes. The peripheral preview phase is the phase in which the eyes are on the fixation cross and have not yet fixated a picture region. Eye-movements were recorded for both eyes. Only the eye with the best validation values for each participant was analysed. The button press ended the image presentation.

Eye-movement analysis. The Eye Link Software identifies those saccades as eye-movements which have a minimum velocity of 30°/s, a minimum motion of 0.2° and a minimum acceleration of 8000°/s². For the eye-movement analysis, regions of interest, with a surrounding margin of approximately 2.5°, were drawn around the patient and the agent. The agent region comprised the agent and (if existing) the object held by the agent. For the analysis of eye-movements, gazes were analyzed rather than single fixations. We define gaze as the time interval between the onset of the first fixation and the offset of the last of consecutive fixations falling into the same interest area.

Figure 1. Examples of pictures *(to shoot, to take a picture)* used in Study 1-3.
Note: the original photographs were fully coloured

Results & Discussion

Accuracy. An answer was scored as correct when the participant pressed the button on the side corresponding to the location of the patient. Accuracy averaged 98% (sd: 3), with non-significant differences between actions ($\chi^2 = 11.96$, df = 11, p = .367). Trials with wrong answers were excluded from the eye-movement and response-time analysis.

Response time. Response time was defined as time elapsed between image onset and button press. The mean response time was 1002 ms (sd: 375). Our participants responded faster than those in Griffin and Bock's study (2000), who report mean patient-detection latencies of 1690 ms. Note that our participants were free to inspect all scene regions and, in 74% of the cases, inspected both agent and patient before pressing the button. In contrast, Griffin and Bock's participants had to locate the patient by fixating as soon as possible, and to subsequently press a button. Therefore, their long reaction times could well be due to monitoring and checking processes performed by peripheral viewing.

Eye movements. Trials with technical errors (7.2 %: failed drift correction, blinks at image onset) were excluded from the analysis. The onset latency of the first saccade away from a fixation point depends on stimulus type and several other factors, such

as the processing load on the current fixation, or whether the fixation cross is temporarily overlapping with the stimulus, or is removed before stimulus onset (e.g. De Graef, 2005; Fischer, Gezeck, & Hartnegg, 1997; Kirchner & Thorpe, 2006; Rayner, 1998). Here, we excluded trials with anticipatory saccade onsets < 80 ms (3 %; cf. Fischer et al., 1997). One participant was replaced because of more than 30 % data exclusion. For the remaining data, the mean onset of initial saccades and fixations was 211 ms (sd: 48) and 251 ms (sd: 88), respectively. The distribution of first gazes (Table 1) shows that 10 % of the initial fixations did not target either of the two actors but went to some intermediate position. This is known as the 'centre of gravity effect' for targets consisting of two elements (cf. Rayner, 1998). Regarding the two actors, the agent region (56.5 %) captured initial fixations more often than the patient (33.5 %). The difference of 23 % was significant (Binomial Test: $z = 4.19$, $p < .001$). As mentioned in the *Stimuli & Design* section, agents were on average located slightly closer to the pre-trial fixation cross than patients. However, excluding the four actions where this was the case resulted in a comparable distribution of first gazes (agent: 53.9 %, patient: 35.1 %, centre: 11.0 %). Consequently, eccentricity effects were not responsible for the observed agent - patient distribution. Similarly, excluding three actions for which the object was within the parafoveal area (at 3° and 3.5°) did not substantially alter the distribution (patient: 33.3 %, agent: 55.6 %, centre: 11.1 %). In addition to the agent preference, actors placed on the left side of the screen were favoured over right-placed actors with 68.6 % versus 31.4 % (Binomial Test: $\chi = 6.99$, $p < .001$). In agent-left actions, the agent preference was even more pronounced (82.4 % versus 17.6 %; Binomial Test: $\chi = 12.09$, $p < .001$, one-tailed, test proportion = .373). It was found before that in Western cultures agents are conceptualized predominantly on the left side in scenes (Chatterjee, Southwood, & Basilico, 1999; Dobel, Diesendruck & Bölte, in press). Gazes on patients were 107 ms longer than gazes on agents. This difference was significant by subjects and items ($t_1 = -4.05$, $df = 15$, $p = .001$; $t_2 = -4.04$, $df = 11$, $p = .002$). Furthermore, gazes on patients started 33 ms later than gazes on

agents, which was significant by subjects (t_1 = -2.60, df = 15, p = .02) but not by items (t_2 = -1.30, df = 11, p = .217).

Table 1. Results of Study 1: Distribution of first gazes and their onset latencies and durations

Region of interest	*Proportion*	*Mean onset (sd)*	*Mean duration (sd)*
Patient	33.5 %	274 ms (78)	345 ms (81)
Agent	56.5 %	241 ms (39)	238 ms (84)
Centre	10.0 %	262 ms (77)	143 ms (79)

Although the observation that first gazes to patients were longer than first gazes to agents can be related to the task, the 'agent effect' in the distribution of first gazes is the reverse of what we had expected for this study. However, an overall 'agent advantage effect' was also found in other studies (Dobel et al., 2007; Kreysa, Zwitserlood, Bölte, Glanemann, & Dobel, subm.; Segalowitz, 1982). Most importantly, the fact that there was a preference for sending the first gaze to the agent region reflects that even initial eye-movements were not made randomly. Some kind of visual and/or semantic representation must have been built up already during the peripheral viewing phase. What could be the reason for the attractiveness of the agent region? We can exclude proximity to fixation location as main reason. One explanation is that the preference for the agent reflects a combined effect of higher visual saliency and a strategy of observers. As mentioned earlier, the agent region comprises the agent as well as the action. This action region in most cases includes the object, which enhances the semantic and probably also the visual saliency of the agent region. Although the object can presumably not be identified during the peripheral preview, due to its small size and the distance from the fixation point (Henderson & Ferreira, 2004; Nelson & Loftus, 1980), the expectancy

of its mere presence might attract attention. Looking at the object provides valuable information about thematic roles and thus for solving the task.

There are alternative explanations for the eyes favouring the agent over the patient region, based on the representation built during the peripheral preview. First, it is possible that thematic roles had already been identified, but that the agent region attracted the eyes by higher visual salience and/or higher semantic salience for task-related monitoring processes via exclusion. Alternatively, if thematic roles could not be assigned within the short peripheral preview, the 'agent effect' would only be due to greater visual salience. Deciding between such explanations was not the focus of our study and has to be postponed to future research.

In the following three studies we further pursue our main question as to how much verb-relevant information can be apprehended peripherally. Presentation conditions of the picture stimuli were designed to mimic the peripheral preview phase.

Study 2: Patient Detection with Brief Peripheral Presentation

The purpose of Study 2 was to establish whether thematic roles can still be identified when foveal inspection of the action scene is ruled out by drastic reduction of the presentation time. Mean saccade onset time in Study 1 reduced by 1 SD (211 ms - 48 ms) is even higher than the duration of 150 ms which is often used as brief presentation time in scene perception (e.g., Calvo & Lang, 2005; De Graef, 2005). Moreover, event-related potential experiments show that the visual processing necessary for a categorization task can be achieved within 150 ms after stimulus onset (Thorpe, Fize, & Marlot, 1996). We therefore reduced the presentation time of our images to 150 ms and subsequently masked them, to simulate the peripheral preview phase of Study 1. We expected that viewers would still be able to make correct decisions about the location of the patient. If so, this indicated that thematic roles of the actors can be identified with brief peripheral presentation and that it is unnecessary to focus the actors.

Method

Participants. There were 16 participants (19-28 years of age) from the same pool as in Study 1. None had participated in Study 1. Compensation was the same as before.

Procedure. The method of Study 2 was the same as in Study 1 with one exception. 150 ms after image onset, the presentation was terminated by a visual mask of the same size as the image, consisting of 16 randomly arranged scrambled squares of the original images. The mask was shown until participants pressed the decision button. Masking served to terminate low-level visual processing. Eye-tracking served only as a control here, assessing whether the eyes did indeed not reach the stimulus.

Results & Discussion

Eye Movements and Response Time. In less than 1% of all trials, a fixation into one of the interest areas occurred. These trials were excluded from the analysis. Mean response time was 664 ms (sd: 50), which is some 350 ms faster than in Study 1.

Accuracy. Answers were scored as in Study 1. Participants were again extremely good at detecting the patient (93% correct answers, sd: 5). This performance did not differ significantly from the performance in Study 1 ($\chi^2 = 1.36$, df = 1, p = .301). As in Study 1, there was no difference in accuracy between actions ($\chi^2 = 15.20$, df = 11, p = .173). The result was the same (93 %) when excluding the actions with differing eccentricities between the two actors and the fixation cross. That is, roles were not inferred by choosing the actor closest to the fixation cross as agent.

Obviously, peripheral viewing of action scenes sufficed for role identification. There are two possibilities how this could be achieved. First, participants could have identified the peripherally presented action, which then allowed them to infer the roles of the two actors. The second possibility is that the global structural layout, that is, the arrangement (body posture and orientation) of the two actors and the space in between, which is available as low-frequency spatial information in the visual periphery, was used to identify the more active person of both. A bias to use low-frequency spatial information versus high-frequency spatial information at early

stages of scene identification was found by Schyns and Oliva (1994; but see Morrison & Schyns, 2001). The first explanation necessitates knowledge about the scene's action, the second does not. To decide between these explanations, in Study 3 we investigated whether actions can be recognised with 150 ms peripheral presentation.

Study 3: Action Naming with Brief Peripheral Presentation

Study 3 tested whether not only thematic roles but also actions could be identified correctly after 150 ms of peripheral viewing. If the high patient detection performance in Study 2 (93%, with low variability) was achieved through action recognition, action naming performance under the same conditions should reach a comparable level, allowing for some deviation due to the binary response required for patient detection. Note that a forced-choice task (patient detection) necessarily produces more correct answers than a naming task (action recognition) with many potential answers. However, if role assignment was inferred from the global layout of the picture, that is, from the body posture of the actors and their relation to each other, actions need not be encoded. We hypothesized that the latter was responsible for the results of Study 2, assuming that encoding the action in our type of scene needs foveal vision into action-relevant regions, as also found in the study by Dobel et al. (2007). In contrast to our stimuli, black-and-white line drawings of actions were used in Dobel et al. with smaller size and unpredictable location on screen. These actions were identified correctly in 19 % and 34 % with 100 ms and 200 ms presentation time, respectively. Despite the advantage of naturalness and larger size in our stimuli, we expected action recognition to be clearly worse than patient detection in Study 2, with accuracy presumably less than one third correct.

At this point we want to stress that we did not plan to directly compare the performance in a two-alternative-forced-choice task with the performance in a open-ended naming task. The rationale behind choosing different tasks was that

they represent best the different demands on observers in Study 2. In the case that action identification was a prerequisite for identifying the location of the patient, the action had to be chosen from an open set but not from two alternatives.

Method

Participants. There were 16 new participants (20-31 years of age) from the same pool as in Study 1. None had participated in Study 1 or 2. Compensation was the same as before.

Procedure. The only difference between Study 2 and Study 3 was the task. Instead of pressing a button to indicate patient location, participants were instructed to name the action they had perceived peripherally. The experimenter wrote an answer down and then started the next trial.

Results & Discussion

Eye movements. Again, trials with fixations on the stimulus were excluded from the analysis (1.6 %).

Accuracy. All correct verbs and their synonyms (as produced in the pretest of the material) were accepted as correct answers. The mean correct performance was quite impressive and higher than expected: 58 % (sd: 32). But the variance in task performance for individual actions was also impressively large, as indicated by the high standard deviation. Some actions were named correctly in 90 - 100 % of trials, for example, *to kick sb.*, but other actions, such as *to feed sb.* (9 %), could hardly be identified. This indicates that peripheral action identification was achievable under specific circumstances, but not in general. Comparing the photographs corresponding to actions with high (e.g., *to kick sb., to throw a ball at sb., to shoot sb.)* and low response performance (e.g., *to feed sb., to help sb., to give sb. a present)* we observed differences in the body postures of the two actors. Highly dynamic actions, such as *to kick at sb.,* show typical relative positions of arms and legs of both actors. Furthermore, there seems to be hardly any alternative interpretation for the

overall spatial layout of these body postures. Such actions were usually named correctly. In contrast, actions with low naming rates were those which require typical facial expressions and typical objects for correct identification. Body postures alone are ambiguous for such pictures, and are congruent with more than one action, such as *to talk to sb., to approach sb., to give sb. sth.* as alternatives for the depicted action *to help sb.*. Note that these actions were identified correctly in the pre-test. We addressed the question as to whether the number of possible alternatives for action-specific body postures played a role in action naming performance in Study 4.

In addition to differences concerning the global spatial layout, there were three other differences between our pictures that seemed worth examining. First, the image region that contained action-relevant details, i.e. involved body parts and objects, was of varying size (mean: 25,072 px^2, sd: 6,552), and second, this region was located at varying distance from the pre-trial fixation point (mean: 6.5°, sd: 1.5). The larger and nearer to this point, the larger the peripheral preview advantage (e.g., Dobel et al., 2007; Nelson & Loftus, 1980). Third, most of the actions had an action-typical object (e.g. a camera for *to photograph sb.*, see Appendix), which would most likely facilitate action recognition, given that such so-called 'diagnostic objects' help rapid identification of scene gist (Friedman, 1979). Although findings concerning the eccentricity at which objects can be identified are inhomogeneous (for a review, Henderson & Ferreira, 2004), for our type of stimuli and task it seems reasonable to expect that objects can only be identified within about 4 - 5° of visual angle. Three objects on our images fell within this radius (3° and 3.5°). We calculated the correlations between the action-naming performance and the size of the action region (in pixels), the distance between the centre of the action region and the fixation cross (in ° visual angle), and the distance between the object and the fixation cross (in ° visual angle). Success in action naming did not correlate significantly with size or eccentricity of the action region (Pearson correlations: size: $r = .11$; distance: $r = .43$) but it correlated negatively with the distance from the action's object ($r = .64$, $p = .024$, one-tailed). The correlation disappeared, however,

when the three actions with the object lying within the parafovea were excluded ($r = -.59$). As found earlier, a diagnostic object within the parafovea facilitates action identification, however, beyond the foveal region it does not.

Note that most of these differences between stimuli are a natural consequence of using realistic stimuli. For example, if the pre-trial fixation point was displayed at a location with equal distance to agent, patient and action, this would result in different distances between stimuli. Obviously, this variable could be controlled for better in line drawings. However, naturalistic stimuli undoubtedly represent the natural setting of perception and language better than line drawings.

From the results of Study 3, we conclude that action identification needs fixations on action-relevant regions of the scene, unless the global spatial layout of the action scene is unambiguous with respect to the depicted action. To test this hypothesis, we carried out the final study.

Study 4: Action Naming with Blurred Pictures

Study 4 was conducted to clarify whether the degree of ambiguity of the images' overall layout, as perceivable with short peripheral presentation, could be held responsible for the large variance in action naming performance of Study 3. Visual resolution declines rapidly with growing eccentricity from the current point of fixation. Whereas identification of facial expressions and small objects needs acute vision, body postures in form of low-frequency spatial information may still be identifiable when out of focus. When peripherally perceived body postures and the spatial relationship between both actors can unambiguously be interpreted with respect to the underlying action, the correct verb can be inferred. However, when the layout is uninformative in that it lacks typical arm, leg, and body posture, and when faces and objects are blurred, the chance of inferring the depicted action is low and action naming performance should be worse. Thus, the variance between items observed in Study 3 could be due to the variance in typicality of body

postures. To test this hypothesis, we altered the appearance of the images. The stimuli were presented this time in foveal vision, but in a way that simulates peripheral viewing. This was done by reducing acuity and eliminating colour from the pictures. If the correct naming of almost two thirds of the actions with short peripheral presentation was mediated by the relative uniqueness of their body-posture layout, we expected a negative correlation between the number of alternative verbs considered suitable for a picture in this study and action-naming performance in Study 3.

Method

Participants. There were new 40 participants (18-29 years of age), none of whom had participated in the previous studies. Compensation was the same as before.

Figure 2. Examples of blurred pictures (to kick sb., to help sb.) used in Study 4.

Stimuli & Design. We used a Gaussian filter (rad. 10px) to eliminate high-frequency spatial information from the photographs used in Studies 1-3. In addition, all photographs were transformed into grey-scaled pictures (see Figure 2). Each action had only one realisation per presentation (agent location again counterbalanced), so the material was reduced to 12 experimental items and 6 warm-ups, distributed over

two lists. The participants saw each stimulus (ca. 26.9° x 21.5°) for 250 ms. This duration was slightly prolonged compared to Studies 2 and 3, to accommodate for the modified experimental setting.

Procedure. In contrast to Studies 1-3, Study 4 was carried out with more than one participant at the same time, without eye-tracking. The experimental lists were shown as a slide presentation in a seminar room, each to a group of 20 students. Participants were given a protocol sheet and asked to write down the action they found most suitable for a specific stimulus. They had about 6 seconds time for their answer to each picture.

Data analysis. All verbs as produced for a specific scene by the participants were submitted in the analysis. A verb and its synonyms were summarized as one item. The sum of all alternative names for each action was correlated with the percentage of correct answers for this action in Study 3.

Results & Discussion

The number of alternatives for the 12 actions varied from 7 verbs for *to kick at sb.* to 23 verbs for *to feed sb.*. As expected, there was a negative correlation between the number of alternatives and performance in the action naming task of Study 3, with short peripheral presentation ($r = .71$, $n = 12$, $p = .005$, one-tailed). The more alternatives for an action were listed, the less correct the answers in Study 3. Conversely, the rate of correct answers (Study 3) was high in the case of fewer alternatives. However, the correlation was not very high. Furthermore, even for actions that were named correctly after short peripheral presentation in Study 3, at least 7 alternatives were given with the blurred images. Together, these results demonstrate that the uniqueness of the actions' general spatial layout was one factor in recognizing them by peripheral vision, but clearly, it was not the only factor. Moreover, the finding that actions were inferred on basis of the spatial layout rather than directly identified, corroborates our hypothesis that patient detection in Study 2 was hardly mediated by identification of the depicted action.

General Discussion

The present studies addressed whether and how thematic roles of an action and the action itself can be apprehended on the basis of a brief peripheral preview phase. In Study 1 we found that, when the task was to detect the patient, the corresponding region received – counter to our expectations – only one third of first gazes, whereas the agent region, which also included the action, received more than half of first gazes. However, it was not clear whether this 'agent advantage' was due to visual or semantic salience, or a combination of both. Studies 2 and 3 revealed that short presentation times of 150 ms, simulating the peripheral preview phase of Study 1, were sufficient to enable the respondent to decide correctly about thematic roles but not to correctly report the verb describing the depicted action. With blurred photographs, the final study yielded that high naming performance in action naming after short peripheral presentation (data Study 3) correlated with relatively low ambiguity of the spatial layout with respect to the corresponding action (data Study 4). In sum, our results show that within a peripheral glance, sufficient information was extracted from the action scene to identify thematic roles by means of the global spatial layout. Furthermore, although action identification was not possible per se, the correct action could be inferred under certain circumstances. One factor influencing successful action naming when parafoveal information, such as an object, was not available, was the number of 'competitor actions' for a given global layout of body postures.

In the following we discuss the implications of our results with respect to (1) comparing the perception of action scenes to the perception of scene gist, (2) the time course of role and action identification, (3) the size of the functional field in action scenes, and (4) the application of eye-movement analysis to research on language production processes.

Action Events and Scene Gist

Dobel et al. (2007) showed that coherence of an agent-recipient action event, manipulated by mirroring both actors (either face-to-face or back-to-back), could be judged correctly with similar short peripheral presentations as used here. Thus, the event's structure, embodied by thematic roles or by coherence, was perceived very rapidly. In contrast, action identification was only possible when an action was unambiguously inferable from the global spatial layout. Otherwise, foveal vision on faces, objects, or body parts seems mandatory. Therefore, the perception of actions constitutes a middle ground between the perception of scene gist (e.g., bathroom, in the street) and of object arrays. Whereas scene gist can be identified within a single glance, via a holistic low-level representation of the scene (Biederman et al., 1982), arrays of (unrelated) objects can only be identified one by one (e.g. Henderson et al., 1999). In the present study, thematic roles were identified far more rapidly than would have been possible if both actors and the optional object of the action had to be fully identified first. Together with the results by Dobel et al. (2007) our findings suggest that brief visual presentation of agent-patient action scenes affords inferring essential semantic information about the depicted action on the basis of the scene's global spatial layout.

The present findings fit well with the account put forward by Oliva and Torralba (2001, 2006). According to their *Spatial Envelope* theory, the visual system forms a spatial representation from global image features within a single glance into a scene, via feed-forward and parallel processing. This representation allows for scene categorization, activation of related semantic information, and possibly for identifying a few objects that are large and near to the fixation point. It is not rich enough, however, to identify peripheral objects or exact spatial relationships. A similar but more interactive model is the *Reverse Hierarchy Theory* (Hochstein & Ahissar, 2002). In this theory, implicit feedforward processing of low-level features results in high-level object and category representations. Explicit feedback to the basic level occurs only later in order to process feature details ("vision with scrutiny"). Both theories account well for our finding that thematic roles can be

assigned even when the action was not identified, given that the temporal restriction of 150 ms ruled out vision with scrutiny.

The Time Course of Role and Action Identification

Although the tasks in Studies 2 and 3 were accomplished with stimuli presented for 150 ms, it is unknown when exactly these answers became available to the participants at the conscious level. They responded on the basis of what they had perceived during 150 ms presentation. Even if further low-level visual processing was ruled out by masking, the cognitive processing that resulted in the identification of thematic roles and some of the actions was not halted by masking. Therefore, the identification of thematic roles and actions could have been reached at any point in time between stimulus onset and response time. For example, if patient detection in Study 2 was possible only after the visual-cognitive processing had proceeded for some time after stimulus offset, this could be an alternative explanation for the fact that, in Study 1, first gazes did not directly target the patient region.

The Functional Field of View in Action Scenes

The present results add to our knowledge concerning the task-specificity of the functional field of view, which is defined as the area from which useful (task-relevant) information can be extracted (Henderson & Ferreira, 2004). Our results show that, with this type of naturalistic action scenes, the functional field of view is larger for identifying thematic roles than for identifying actions. Otherwise, actions should have been reported more accurately. This is in line with the notion that the size of the functional field of view is different for different types of scene information and that gross spatial scene layout can be processed from a larger area surrounding the current fixation than it is possible for detailed semantic information (Henderson & Ferreira, 2004). There are other characteristics of our stimuli and design, apart from the task, that had an influence on the size of the functional field of view. First and as mentioned above, all photographs had the same global scene

arrangement, with the two humans on the right and left of the image facing each other in front of an identical, neutral background. It is likely that these characteristics were beneficial for recognising the global layout. Second, all our actions were agent-patient actions, with both participants being human and, apart from their roles, the same individuals shown repeatedly. Actions with two or more actors carrying out the same action (e.g. playing ball), or only one of the participants being animate (e.g. feeding a cow), might lead to easier recognition of the event's structure. Third, there is evidence that natural visual stimuli are processed more efficiently than artificial stimuli (Braun, 2003; Rousselet, Joubert, & Fabre-Thorpe, 2005). On the other hand, the more complex and cluttered natural scenes are, the higher the demands they pose on visual recognition due to lateral masking effects and, hence, the smaller the functional field of view (De Graef, Christiaens, & d'Ydewalle, 1990; Henderson et al., 1999). As our photos were all taken in front of the same neutral background, the facilitative influence of natural scenes was probably predominant.

A further issue relevant to the size of the functional field is the role of attention. In our studies, task-knowledge directed covert attention, prior to stimulus onset, from the fixation point downwards into the region where the picture would appear. As covert attention can shift faster than overt attention (e.g., Wolfe, 1998), covert attention probably targeted one of the meaningful regions in a more 'spot-light' manner soon after image onset. The fact that the global layout of the stimuli was always the same, with actors located on the right and left side of the image, may have facilitated this process of covert orienting. In Study 1, this covert shift became evident in first fixations on the agent. However, it remained invisible when the display disappeared too early, as in Studies 2 and 3.

Eye Movements and Language Production

Having found that, after a peripheral information uptake of 150 ms, a good deal of verb-related information is already available, we can now ask what this implies for

the use of eye-movements in language production research. Not at all surprisingly, we see that linguistic processing is not restricted to information from foveal vision. Central components of the preverbal message, namely roles (and in many cases also actions), could be assigned (roles) or verbally reported (actions) from scenes that were never fixated. This loosens the presumed tight coupling between looking and speaking, as eye movements do not indicate, at which point in time particular and speech-relevant scene information, such as roles or actions, becomes available to the speaker. Furthermore, not fixating a region is not identical with not recognising it, and consequently, gaze durations serve as a relative rather than as an absolute measure (Irwin, 2004; Morgan & Meyer, 2005).

The interpretation of potential peripheral preview effects becomes even more sophisticated when interpreting eye-movements that are not at the initial but in an advanced position within a given gaze path. Clearly, a gaze path on an action scene is not the sum of looks at its constituent parts. Instead, the duration of gaze *n* on a scene region is influenced by what aspects of this region could be perceived during gaze *n-1*. Also, there is evidence that the processing load of the current fixation can limit the amount of peripheral preview benefit (Calvo & Lang, 2005, Lavie & Fox, 2000) and the onset of the next saccade (Findlay, 2004). Whereas in our study, the simple task of staying with the eyes on the fixation point demanded little cognitive effort, the situation is different as soon as language-related gazes on the informational scene parts have begun. Therefore, the present findings obtained with an 'isolated' preview phase should not be over-interpreted with respect to the preview benefit within a gaze path under more natural viewing conditions (cf. Findlay, 2004).

To conclude, verb-relevant semantic information can be extracted from agent-patient action photographs on the basis of a short peripheral presentation. Our data suggest that this is possible for those aspects that can be inferred from the spatial layout of the whole scene: thematic roles and distinct action layouts. However, beyond this holistic exploitation, foveal vision seems necessary to extract the more

specific and detailed features needed for verb and sentence production. Recent work in our lab confirms this hypothesis by demonstrating that rapid understanding of the meaningfulness of action scenes is driven more by body orientation and gaze direction of the participants than by the object used in this action (Glanemann, Dobel, Bölte, & Zwitserlood, 2007). The results of the present studies suggest that the degree to which eye-movements and speech-planning processes are time-locked early after stimulus onset is influenced by a whole variety of stimulus and task factors, all of which need to be considered when interpreting the underlying relationship between vision and speaking.

Appendix

List of actions and objects for all pictures

Action	Translation	Object
jdn. bedienen	to serve sb.	tray with cup
jdn. beschenken	to give sb. a present	parcel
jdn. bewerfen	to throw a ball at sb.	ball
jdn. erschießen	to shoot sb.	pistol
jdn. erschrecken	to scare sb.	n. a.
jdn. fotografieren	to photograph sb.	camera
jdn. füttern	to feed sb.	bowl, spoon
jdm. helfen	to help sb.	first-aid kit
jdn. interviewen	to interview sb.	recorder, microphone
jdn. schlagen	to hit sb	stick
jdn. stoppen	to stop sb.	traffic signal
jd. treten	to kick sb.	n. a.

Rapid Apprehension of Coherence of Action Scenes

CHAPTER 3

Abstract

Some information about complex naturalistic scenes, such as the scene's gist and an object's category, can be extracted within a fraction of a second. The present study focussed on scene coherence, for action scenes that involve two actors. Scenes were presented for 100, 50, 30, or 20 ms. Coherence was manipulated either through varying global scene layout, by mirroring one or both actors, or by varying the semantic consistency between the action and the action-involved object. Viewers were able to extract scene coherence with very brief presentation durations (from 30 ms onwards for global scene layout, but at least 100 ms for semantic consistency). Thus, the recognition of holistic scene properties, such as the spatial relation between the scene's components and the contour of the depicted action, was faster than the identification of the action and the object. The results suggest that the rapid extraction of semantic information from action scenes relies more on the scene's overall *Gestalt* than on a detailed visual and semantic representation.

Introduction

Studies on visual scene perception have revealed that the extraction of semantic information from a complex scene can be achieved without overt attention on any scene detail. With presentation for 100 ms or less, viewers can decide whether a scene contains a member of a predefined object category (e.g., Thorpe, Fize & Marlot, 1996). Under similar conditions, the category of an environmental scene can be recognized, such as 'in a restaurant' or 'at the beach', also known as the scene's *gist* (e.g., Biederman, Mezzanotte, Rabinowitz, 1982; Potter, 1975, 1976; Schyns & Oliva, 1994). This seems at odds with classic accounts of visual perception (e.g., Neisser, 1967; Treisman & Gelade, 1980), which propose that high-level visual representations can only emerge from the attention-demanding binding of basic image features, such as orientation and colour. However, other accounts propose that fast and parallel processing of image feature sets at a "relatively low" (Oliva & Torralba, 2001; 2006) or "intermediate" level (Evans & Treisman, 2005) can also account for these intriguing findings.

The present study is concerned with the rapid apprehension of a particular type of complex scene, namely with action scenes. Action scenes are particularly interesting as they are depictions of dynamic events. This temporal aspect renders them more complex than static scenes. In our earlier work, we found that people can apprehend the coherence (= meaningfulness) of action scenes based on presentation durations as short as 100 ms (Dobel, Gumnior, Bölte & Zwitserlood, 2007). As with object categories and gist, the means and mechanisms underlying this remarkable ability are not yet well known. Thus, our current objective is to assess on the basis of what type of visual information such fast coherence decisions are made. To this aim, we contrast the manipulation of the global scene layout (mirroring of actors) with the manipulation of local scene information (appropriateness of action-relevant object). We also varied presentation duration, to establish how timing affects the perception of each coherence type.

In the following, we first review the main findings in rapid scene recognition. We then discuss the data currently available on the influence of object-scene consistency on object perception. Next, we summarize the few studies that dealt with action scenes to date. The introduction concludes with an overview of our experiments and hypotheses.

Rapid Categorization of Objects and Scenes

During the last four decades, two information types have been investigated with respect to their rapid uptake from natural scenes (for an overview see, e.g., Henderson & Ferreira, 2004). One line of research uses forced-choice categorization tasks, in which a member of a predefined object category (such as 'animal') is either present in the scene or not (e.g., Evans & Treisman, 2005; Thorpe et al., 1996; VanRullen & Thorpe, 2001). The exposure times employed in these studies were well below the minimal time needed for a single fixation, thus excluding overt attention on scene details. Successful categorization was possible, even for stimuli displayed in the far visual periphery (Thorpe, Gegenfurtner, Fabre-Thorpe & Bülthoff, 2001). Moreover, performance with peripheral stimuli was hardly affected by the binding of attention to a simultaneous central task (Li, VanRullen, Koch & Perona, 2002), or by presenting two images simultaneously (Li, VanRullen, Koch & Perona, 2005). There are two explanations for this fast categorization. One is based on a rapid, but crude first pass through the visual system relying on local-feature sets diagnostic of particular object categories (Evans & Treisman, 2005). The other proposes that familiar and naturalistic stimuli are processed with particular efficiency and speed, due to, for example, pre-existing or more intense neuronal representations (Bacon-Macé, Macé, Fabre-Thorpe & Thorpe, 2005; Braun, 2003; Li et al., 2005). Note that the fact that viewers can decide whether an object belongs to a superordinate object category after ultra-rapid presentation does not necessarily imply that these objects have been identified.

Another approach in rapid scene perception examines the fast detection 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), "including all levels of processing, from low-level features (e.g., colour, 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). That is, the gist of a scene is more than the sum of its components; it also concerns the shared meaningful content, as, for example, kitchen or farmyard, as conveyed by the specific semantic and spatial relationships between the scene's components. Typically, stimuli in gist studies involve multiple scene elements and their spatial relations. This is different from most categorization studies mentioned above, where one critical object has a prominent foreground position within the scene. Gist apprehension has been shown to emerge at 30 - 50 ms after scene onset (cf. Henderson & Ferreira, 2004). It has been proposed that spatial layout information, that is, the spatial arrangement of the whole scene or its objects, plays a crucial role in gist identification (Biederman, 1988, 1995; Castelhano & Henderson, 2007; Mannan, Rudock, & Wooding, 1995; Oliva & Schyns, 1997; Sanocki & Epstein, 1997; Schyns & Oliva, 1994; Oliva & Torralba, 2001, 2006). These studies also suggest that the identification of the objects in a scene is not a prerequisite for gist recognition. For example, Schyns & Oliva (1994; Oliva & Schyns, 1997) used blurred stimuli to show that the layout of principal contours in a given scene stimulates the recognition of the scene's category before the identity of the scene's objects is available. With only 30 ms masked presentation, observers exploited low spatial-frequency information for scene recognition (Schyns & Oliva, 1994). According to Oliva & Torralba (2001, 2006), specific configurations of local scene features, such as colour and orientation, constitute so-called "global features" that in turn form a "Spatial Envelope representation" and capture the diagnostic structure of the image. This scene-wide processing is proposed as being complementary, parallel, and supportive to object-centred mechanisms in scene recognition (Oliva & Torralba, 2006).

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. No overt attention is needed for either task. Whereas object-category decisions are based on local scene information, decisions on a scene's gist seem to rely on global scene information. Which of the described mechanisms underlie the rapid apprehension of the coherence of action scenes, has not been examined yet.

Rapid Apprehension of Object-Scene Consistency

Whether the semantic consistency of scene components has an influence on early scene perception has been studied, for example, with the object detection paradigm (e.g., Biederman et al., 1982; Boyce, Pollatsek & Rayner, 1989; Hollingworth & Henderson, 1998, 1999; for reviews see Bar, 2004; Henderson & Hollingworth, 1999), and with eye movement paradigms (e.g., DeGraef, Christiaens & d'Ydewall, 1990; Friedman, 1979). In the object detection paradigm, for example, viewers have to detect a target object in a briefly presented scene (usually 100 - 200 ms), and accuracy rates are interpreted as a measure of object identification. Most studies found an effect of semantic consistency, that is, objects can be detected more accurately if they appear in their typical surroundings (e.g., a coffee machine in a kitchen) compared to an atypical surrounding (e.g., a violin in a bathroom). These results were challenged by Hollingworth and Henderson (1998) who attributed the consistency effect to methodological problems of the studies. Controlling for higher-level response biases, they failed to replicate the effect and concluded that the consistency effect does not arise at early levels of perceptual analysis, but rather results from later, post-identification processing stages (Hollingworth & Henderson, 1998, 1999). Recently however, the consistency effect of objects and their background was observed with brief presentation (80 ms), using naturalistic colour photographs with large and foregrounded objects (Davenport & Potter, 2004). Moreover, the effect was also found for background information, that is, backgrounds were reported more accurately when the object was consistent with the scene (e.g., a jogger in a park) than when it was not (e.g., a ballerina in a church). The authors argued that in the Hollingworth and Henderson experiments, the task might have been not sensitive enough, and the design had created an asymmetrical

advantage for inconsistent objects. From their own results, they concluded that scenes and embedded objects are processed interactively, at the early perceptual level.

Taken together, and important for our present purposes, there is good evidence for the rapid understanding of whether scene components, or a scene component and its background, are semantically consistent.

Rapid Apprehension of Action Scenes

Action scenes are a special type of complex scenes. They are static depictions of events, which inherently stretch out in time. Our knowledge about the processing of action scenes is limited because only few studies have been conducted so far. Fast apprehension of thematic roles from action scenes was first demonstrated by Griffin and Bock (2000) in an eye-tracking study with line drawings of agent-patient-actions (e.g., a girl shooting at a man). Viewers could identify and fixate the patient of the action within 500 ms after picture onset, which implied that they had already extracted the roles of both participants. With photographs and similar types of actions, even 150 ms of masked peripheral presentation were sufficient to apprehend thematic roles (Glanemann et al., under revision). Strikingly, actions were correctly identified and named in nearly 60% of trials. Data from blurred versions of the same stimuli demonstrated that the global layout of the action, conveyed by action-typical body postures of agent and patient, was mainly responsible for this result.

In the study by Dobel et al. (2007) mentioned above, the coherence of action scenes was manipulated by mirroring both actors at the same time, which led to coherent (face-to-face) and incoherent (back-to-back) actions. With 100 ms masked peripheral presentation, coherence detection was 80 %. However, it remained unclear whether this extremely rapid identification resulted from spatial layout information, or from a high-level semantic representation of the action. In incoherent versions, the whole action's *Gestalt* was more ragged. Body parts and the

instrument used in each scene (e.g. rifle, tray) pointed towards the scene centre in the coherent version, and outwards in the incoherent version. In addition, the orientation of the actor's faces, and thus the direction of their gaze, may have played an important role. Gaze direction attracts the attention of observers (Weith, Castelhano & Henderson, 2003), and recognizing gaze directions of acting people helps to understand their intention (Baron-Cohen, 1995). Thus, both factors might have led to the interpretation of back-to-back actions as being incoherent and face-to-face actions as being coherent. However, it is also possible that the visual representations were detailed enough to understand coherence at a conceptual level.

In sum, there is evidence that essential information about action scenes can be extracted in the absence of overt attention. It is still unclear how detailed the rapidly construed visual and semantic representations of the actions are. As a consequence, it is not well understood exactly which information enables coherence extraction.

Overview of Experiments

The aim of this study was to investigate what kind of visual information drives the judgment on the coherence of an action scene with very brief stimulus exposure. We refer to an action scene as coherent, if it depicts a meaningful action, and as incoherent, if the manipulation of visual or causal relation between the action's components renders the action meaningless. To our knowledge, no research with stimulus exposures below 100 ms has been conducted including manipulations of properties of actions. We examined two types of information potentially useful for coherence: spatial layout aspects and detailed semantic information about the action-relevance of objects. We also varied presentation duration, because we expected the two types of information to be differentially affected. We hypothesized that spatial-layout cues to coherence should be less susceptible to presentation duration than cues relying on scene details.

Action scenes were presented for 100, 50, 30, or 20 ms, and immediately followed by a perceptual mask. All actions involved two humans: the agent, performing the

action, and the patient, who was acted upon (e.g., one person serving coffee to the other person). Action coherence was manipulated by either mirroring none, one, or both involved actors (resulting in four different combinations of body orientation, see Figure 1a - d), or by the use of an action-appropriate or action-inappropriate object (Figure 1e - f). The first kind of variation had an influence on the action scene's general spatial layout, whereas the second kind produced consistent and inconsistent action-object relations within the scenes.

Body orientation can be an easily identifiable cue for action coherence, even with brief presentation times (cf. Dobel et al., 2007). Face-to-face actions, as in Figure 1a, are typically coherent, and back-to-back actions (Figure 1d) are typically incoherent. However, coherence detection becomes more demanding if one actor faces the other's back, and requires a more detailed visual representation of the scene. Actions of this type could be coherent or incoherent. In our material, an agent facing the back of the patient, as in Figure 1b, constitutes a coherent action, whereas a patient facing the agent's back constitutes an incoherent action, see Figure 1c. Due to the very nature of coherent and meaningful actions, agents are usually oriented towards the person they act upon. This leads to a confound of the orientation of the agent, in contrast to orientation of the patient, with coherence. We therefore included some atypical actions that were coherent, although both actors stood backwards to each other (see Figure 1g - h). These items served as probes to assess whether coherence was exclusively judged by the body orientations of the two actors, or whether viewers were able to extract more detailed information from the scene. In the first case, data for back-to-back probe actions (c +)[1] should be quite similar to those for the back-to-back actions (c -) of the Body Orientation set. In the second case, the probe items (c +) should be judged as coherent more often than their (c -) counterparts.

[1] for the reason of easier reading, hereafter (c +) is used to mark coherent action types and (c -) is used to mark incoherent action types

In contrast to body orientation, the semantic consistency of an action and its object cannot be judged by relying on global spatial properties of the image. Instead, judgments of semantic consistency between object and action require parsing the scene into action and object, as well as the identification of both these components. Thus, we predicted that a correct judgment with this stimulus type necessitates a more detailed representation of the action scene. Given that all scenes for which semantic consistency was varied had face-to-face actors, we predicted the following.

Correct coherence decisions for the semantic consistency pictures would demonstrate that visual representations were detailed enough to identify both the action and the object. On the other hand, if presentation times are too short for the identification of both the action and the object, actions should be judged as coherent on the basis of the actors' orientation, regardless of the appropriateness of the object used for the action. The same result is expected if only the action or only the object but not both can be identified.

Based on what is known about rapid (action) scene perception, we hypothesized that, in the present experiment, global spatial layout information would be easier to perceive than the identity of both action and object. Therefore, we expected viewers to perform better on the Body Orientation scenes than on the Action-Object Consistency scenes.

Method

Participants. A total of 64 students from Münster University (26 m, 38 f), between the ages of 19 and 28 years, with normal or corrected-to-normal vision, participated in this study. They received course credits or 3 €.

Materials. The stimuli were colour photographs of two-participant action scenes, created for the purpose of this experiment. They depicted either 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,

e.g., to scare sb., to brush sb.'s hair) or exchanging the object of the action (10 actions, e.g., to serve sb. a coffee, to give sb. money). Body orientation of agent and patient were varied individually, resulting in four different body orientation variations: face-to-face, agent-facing-patient's-back, patient-facing-agent's-back, and back-to-back (see Figure 1a - d). The first two of these variations constituted coherent actions, the other two were incoherent. Note again that, whereas the patient could be facing the agent or standing backwards to him in both coherent and incoherent actions, the orientation of the agent was confounded with coherence. However, this was unproblematic for the statistical analysis, because combinations of body orientations rather than the individual orientation of the agent and patient were compared (see results section).

In the second action set, coherence was varied by exchanging the original object (e.g. scissors for cutting hair) with an inappropriate object (e.g. a wooden spoon for cutting hair), resulting in two variations per action. These 10 actions were depicted in slightly enlarged size to guarantee that all objects could be identified (see Figure 1e - f).

All 20 actions were photographed with four different actor pairs from a pool of 10 actors (5 f, 5 m). This resulted in four different experimental sets, each of which comprised the 40 Body Orientation targets (10 actions * 4 variations) and the 20 Action-Object Consistency targets (10 actions * 2 variations). The actor pairs were distributed over the four sets in a Latin square design. In addition, each list contained 10 probe items for the Body Orientation condition. These items depicted actions that were coherent although the actors stood back-to-back (e.g., to spray sb. backwards with water, see Figure 1g - h). Another 10 incoherent filler items counterbalanced these probe items with respect to coherence. To increase the power, always two of the above described sets were combined so that all actions were repeated once but with different actors. This resulted in 160 items per experimental list. Another eight items served as warm-ups. Actions used as filler and warm-up items differed from the actions of the experimental set. Agent position (left/right) and coherence were balanced within lists. All actions used were pre-

tested. The task of the 20 participants (none of which took part in the main experiment) was to decide whether a specific stimulus depicted a meaningful (= coherent) or a meaningless (= incoherent) action. Only actions that were at least 75% judged correctly were used in the experiment (see the Appendix for a list of all actions).

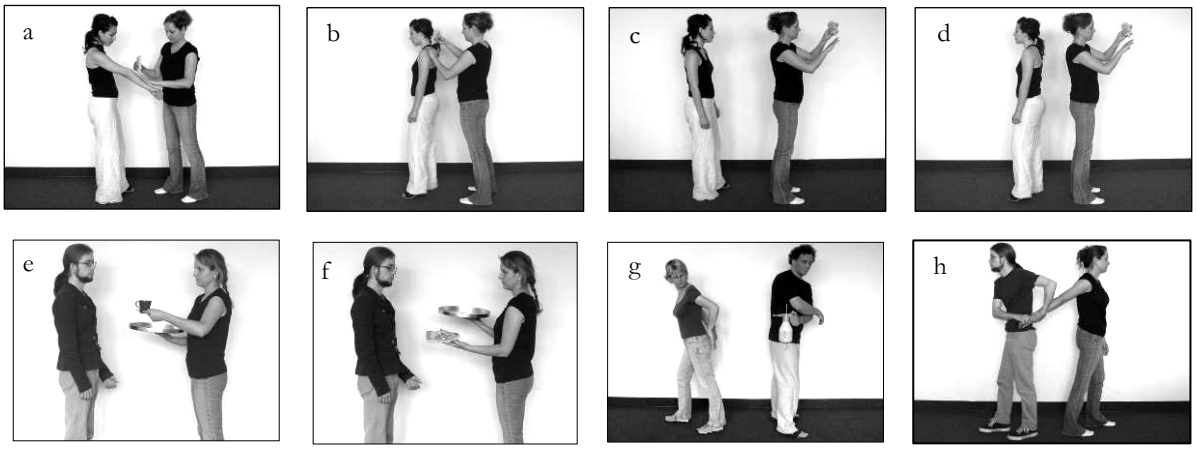

Figure 1. Examples of the stimuli used in this study. 1a - d: example action for the Body Orientation condition with its four variations (to put lotion on sb.), 1e - f: example action for the Action-Object Consistency condition with its two variations (to serve sb. a coffee/shoe), 1g - h: two example actions of the probe items (to spray sb. backwards with water, to pull sb.). Images of the type 1a, b, e, g, and h depict coherent actions, stimuli of the type 1c, d, and f depict incoherent actions.
Note. The original images were fully coloured.

Apparatus. The stimuli had a size of 20° x 28° visual angle and were presented on a 21" computer screen (Samsung Syncmaster 1100p), with a resolution of 1024 x 768 pixels. The refresh rate of the monitor was 100 Hz. The presentation of the stimuli and the on-line collection of the manual responses were controlled by the software SR Research Experiment Builder®, on an IBM-compatible computer. A Microsoft Sidewinder Freestyle Pro Game Controller served as device for the manual answer.

Procedure. Participants were tested individually, seated at a distance of approximately 90 cm from the display monitor. With a written instruction, a meaningful action was

explained as depicting two actors being involved in the same action. A meaningless action was explained as depicting an inappropriate object for the action (e.g., one person interviews the other person with a banana instead of a microphone) or as one of the actors being uninvolved in the action (e.g., one person is packing a parcel, the other person is just an observer). For both types of coherence variation, written example sentences were given for meaningful and meaningless actions. Participants were asked to respond as quickly and accurately as possible by pressing one key for a meaningful action and another key for a meaningless action. Each trial started with the display of a fixation cross in the centre of the screen where the images were to appear. After a randomly varied delay between 1000 and 2500 ms, an action scene was presented for 100 ms, 50 ms, 30 ms, or 20 ms. The offset of the stimulus was followed by a perceptual mask that consisted of 80 uninformative small squared parts cut out of the filler-items. The mask disappeared after another 250 ms and the participant had 2000 ms to answer by key press before the next central fixation cross was displayed. The total duration of an experimental session averaged 20 minutes. Sixteen participants were tested with each presentation duration.

Results

Data Analysis

An answer was scored as correct if a coherent action was judged as meaningful and if an incoherent action was judged as meaningless. Percentages of correct answers are provided in Table 1. One-sample *t*-tests were used to analyse whether proportions of correct answers differed significantly from chance level (test value = 0.5, two-tailed).

For the comparison of task performance between the two scene types, arcsine-transformed percentages of correct answers (see Winer, Brown & Michels, 1991) were submitted to a 2 * 4 ANOVA with the within-subjects factor

INFORMATION TYPE (Body Orientation, Action-Object Consistency) and the between-subjects factor PRESENTATION TIME (100, 50, 30, 20 ms). In the analysis by items, INFORMATION TYPE was a between-items factor, and PRESENTATION TIME was a within-items factor.

Further ANOVAS were conducted on the data within each information type. For the Body Orientation stimuli, the arcsine-transformed percentages were submitted to a 4 * 4 ANOVA with the within-subjects factor ORIENTATION (face-to-face, agent-facing-patient's-back, back-to-back, patient-facing-agent's-back) and the between-subjects factor PRESENTATION TIME (100, 50, 30, 20 ms). In the analysis by items, both were within-items factors. Similarly, for the Action-Object Consistency stimuli, the arcsine-transformed percentages were submitted to a 2 * 4 ANOVA with the within-subjects factor CONSISTENCY (consistent, inconsistent) and the between-subjects factor PRESENTATION TIME (100, 50, 30, 20 ms). In the analysis by items, both were within-items factors. Degrees of freedom were adjusted with the conservative lower bound procedure in all ANOVAs. For post-hoc comparisons, independent and paired *t*-tests were computed.

For reasons of brevity, we report only *t*-tests that yielded significance. The probe items for the Body Orientation condition underwent only a descriptive analysis due to the limited number of items.

Comparison of the two Information Types

As expected, the Body Orientation stimuli generally yielded much better performance rates than the Action-Object Consistency stimuli. Averaged over all four presentation times, coherence was judged correctly in 68 % when manipulated by body orientation, and only in 50 % (i.e., chance level) when manipulated by semantic consistency of action and object (main effect of INFORMATION TYPE, $F_1(1, 60) = 244.69$, MSE = 0.02, $p < 0.001$, and $F_2(1, 18) = 136.87$, MSE = 0.02, $p < 0.001$). There was also a significant main effect for PRESENTATION TIME

($F_1(3, 60) = 32.90$, MSE = 0.03, $p < .001$, and $F_2(1, 18) = 52.31$, MSE = 0.03, $p < 0.001$).

The two main effects were qualified by a significant interaction ($F_1(3, 60) = 25.43$, MSE = 0.02, $p < 0.001$, and $F_2(1, 18) = 27.05$, MSE = 0.03, $p < 0.001$). Whereas correct judgements declined steadily with shorter presentation time for Body Orientation scenes, this was not so for Action-Object Consistency scenes. For the former scenes, the performance of 83 % with 100 ms exposure declined in roughly 10 % - steps with decreasing presentation times to 52 % with 20 ms. All differences between consecutive presentation durations were significant (all $t_{1,2}$ s > 2.85, all ps ≤ 0.004, one-tailed). In contrast, Action-Object Consistency was hardly ever identified correctly. The performances at presentation times of 50 ms and shorter corresponded to chance level. Only the 3 % - difference between 100 ms and 50 ms reached significance ($t_1 = 1.78$, df = 30, $p = 0.043$, and $t_2 = 2.18$, df = 9, $p = 0.029$, one-tailed). Consequently, the linear trend found for PRESENTATION TIME, $F_2(1, 18) = 129.61$, MSE = 0.01, $p < 0.001$, was caused mainly by the Body Orientation scenes.

Table 1. Percentages of correct answers for all action types in the Body Orientation and Action-Object Consistency condition, for all presentation times.

		100 ms	50 ms	30 ms	20 ms	*mean*
Body Orientation						
face-to-face	c +	80 ***	84 ***	76 ***	53	*73 ****
ag facing pat	c +	60	47	60 *, 2	43 *, 2	*52*
pat facing ag	c -	95 ***	78 ***	55 *, 2	53 *	*70 ****
back-to-back	c -	96 ***	92 ***	60 *, 2	60	*77 ****
mean		*83 ****	*75 ****	*63****	*52 *, 2*	*68 ****
Act-Obj Consistency						
consistent	c +	91 ***	84 ***	70 **	45	*73 ****
inconsistent	c -	16 ***	17 ***	28 **	50	*28 ****
mean		*53 **, 1*	*50*	*49*	*47*	*50*

Note.

One-sample t-tests were significant at least at the p < *0.05 / **0.01 / ***0.001 level

One-sample t-tests were significant only in the analysis by [1]subjects / [2]items

c + = coherent action, c - = incoherent action

Body Orientation

One-sample *t*-tests yielded that performances with all image types and all presentation times differed from chance level with two exceptions. Images of the type agent-facing-patient's-back produced inconclusive coherence judgments at close to chance level irrespective of presentation time. The same applied for all image types at 20 ms presentation.

The four different body orientation combinations yielded different results across presentation times (main effect of ORIENTATION, $F_1(1, 60) = 22.22$, MSE = 0.59, $p < .001$, and $F_2(1, 9) = 37.39$, MSE = 0.21, $p < 0.001$). As expected, coherence judgements were very similar for face-to-face (M = 73 % correct) and back-to-back actions (77 %). However, contrary to our expectations, patient-facing-agent's-back actions (70 %) were judged equally well. Only the 7 %-difference between the latter two was significant ($t_1 = 4.05$, df = 63, $p < 0.001$, and $t_2 = 4.33$, df = 9, $p = 0.001$). In contrast, performance for the agent-facing-patient's-back actions (M = 52 %) differed significantly from the other three action types (all $t_{1,2}$ s < -4.11 or > 4.05, all ps < 0.001) and was at or near chance level for all presentation times.

A main effect for PRESENTATION TIME, $F_1(3, 60) = 50.37$, MSE = 0.13, $p < 0.001$, and $F_2(1, 9) = 106.07$, MSE = 0.12, $p < 0.001$; and a linear trend ($F_2(1, 9) = 205.95$, MSE = 0.01, $p < 0.001$ was observed. Both effects were already described above.

Table 2 shows that judgements on the three coherent back-to-back actions that served as probes (e.g., to spray sb. backwards with water, Figure 1g) differed from judgements made for the incoherent actions with the same body orientation (e.g., Figure 1d). Note that contrary to our reports so far, these percentages represent frequencies of the answer "meaningful", instead of frequencies of correct answers. Thus, the percentages in the first and second row represent incorrect and correct answers, respectively. The data illustrate that viewers judged these two action types differently, especially with 100 ms and 50 ms presentation time. The coherence of back-to-back probe actions was correctly judged in about 20 % more often than the incoherence of back-to-back actions from the Body Orientation set. With shorter presentation times, this difference disappeared. As mentioned above, a statistical comparison between these two action types was not carried out due to the limited number of probe actions.

Table 2. Percentages of the answer "meaningful" for coherent (probe items) and incoherent (Body Orientation condition) back-to-back actions, for all presentation times.

		100 ms	50 ms	30 ms	20 ms	*mean*
Body Orientation	c -	4 %	8 %	40 %	40 %	*23 %*
probe items	c +	28 %	28 %	43 %	49 %	*37 %*

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

Semantic Consistency between Action and Object

There was a marginally significant difference between the four presentation times (main effect of PRESENTATION TIME *(F_1*(3, 60) = 2.69, MSE = 0.05, p = 0.05, and F_2(3, 27) = 2.67, MSE = 0.03, p = 0.07). As can be seen in Table 1, performance clearly differed between coherent and incoherent actions (main effect of CONSISTENCY, F_1(1, 60) = 115.83, MSE = 0.30, p < 0.001, and F_2(1, 9) = 149.50, MSE = 0.14, p < 0.001). Averaged over all presentation times, actions with appropriate objects were judged correctly in 73 % of the cases, whereas actions with inappropriate objects were judged correctly in only 28 % of the cases. That is, both image types were judged as coherent in more than 70 %. The shorter the presentation time, the less strong was this tendency (significant interaction between PRESENTATION and CONSISTENCY *(F_1*(3, 60) = 19.12, MSE = 0.30, p < 0.001, and F_2(3, 27) = 53.56, MSE = 0.07, p < 0.001). However, there was no general bias for the answer "meaningful" as demonstrated by a comparison to the results of the Body Orientation condition.

Again, looking at the data from a different perspective, as percentages of "meaningful" answers, is very informative (see Table 3; and see Table 1 for a comparison with percentages of correct answers in this condition). Averaged over consistent and inconsistent actions, the data closely resemble those for the face-to-face actions (c +) of the Body Orientation condition. Hereby, they clearly show that

Action-Object consistency did not play a role in judging the coherence of face-to-face actions.

Table 3. Percentages of the answer "meaningful" for the face-to-face actions in the Action-Object Consistency and Body Orientation condition, for all presentation times.

		100 ms	50 ms	30 ms	20 ms	*mean*
Act-Obj Consistency	consistent	91 %	84 %	70 %	45 %	*73 %*
	inconsistent	84 %	83 %	72 %	50 %	*72 %*
Body Orientation	face-to-face	80 %	84 %	76 %	53 %	*73 %*

Discussion

This study investigated the rapid build-up of internal visual representations of briefly presented action scenes. We tested the hypothesis that, for judging coherence of the depicted action, coarse spatial layout information is processed more efficiently than information on the semantic consistency between object and action. Further, we determined the lower limit of stimulus exposure for correctly judging scene coherence with these information types.

We first discuss the results for the spatial-layout set and for the semantic-consistency set. We evaluate next how our results on action scenes fit with existing findings and theories of early visual scene perception. Finally, we consider the benefits of rapid action apprehension.

Spatial Layout

The minimal presentation time needed for judging action coherence correctly was 30-50 ms, when coherence was conveyed by the global spatial layout of the action

scene and the mirroring of one or both involved actors. This corresponds to the time frame suggested for extracting first information about the gist of an environmental scene, such as a kitchen or a farmyard (Henderson & Ferreira, 2004). For both types of scenes, the uptake of task-relevant information within this ultra-brief time span is a remarkable feat. However, it seems even more remarkable for action scenes, given that a photograph is only a snapshot of an action and does not reflect the whole temporal course of the event it stands for. Presumably, the memory entry of such dynamic events is more or less automatically activated very early in action scene perception.

The performance rate of 83 % correct answers with 100 ms presentation time study resembles what we found earlier for coherence ratings of action scenes (Dobel et al., 2007), although compared to the task in the study by Dobel and co-authors, our images were larger, consisted of coloured photographs, and were presented in central view instead in the visual periphery. On the other hand, the fact that the stimulus set comprised not only 'face-to-face' and 'back-to-back' combinations but also actions with only one mirrored actor rendered the present task more difficult.

As expected, the two action types face-to-face (c +) and back-to-back (c -) were judged best. Contrary to our expectations, however, only one of the other two combinations was judged clearly worse than the hypothesized easy combinations. Whereas patient-facing-agent's-back actions (c -) yielded equally good performance rates as face-to-face (c +) and back-to-back actions (c -), performance for actions of the type agent-facing-patient's-back (c +) hardly ever differed from chance level. This large difference between the two action types demonstrates that factors other than body orientation played a role in the viewers' decision on coherence. Looking at the overall Gestalt of the actions, particular layout features might explain this difference. Although we tried to realise the actions without expansive forward-gestures of the agent, this - due to the nature of an agent acting upon another person - certainly could not be avoided. That is, the arms of an agent often reached out to some extent. Consequently, a backwards oriented agent resulted in a non-continuous edge of the whole action's contour, as well as in an empty interspace

between the two actors. Together, this made the Gestalt more open and apparently gave reason to judge the patient-facing-agent's-back actions (c -) correctly as incoherent. In contrast, agent-facing-patient's-back actions (c +) had a filled interspace combined with an even contour. These two layout criteria for coherence apparently carried less weight than an empty interspace combined with a ragged contour as criteria for incoherence[2].

Looking at the back-to-back probe actions (c +), we find support for the idea that viewers did not rely on body orientation alone to decide on coherence. With 100 ms and 50 ms presentation time, this action type was judged as coherent more often than its incoherent counterpart (Table 2). That is, details such as sidewards turned heads (Figure 1g), or a filled interspace between the actors (Figure 1h), were likely identified and responsible for judging the actions correctly as coherent in nearly one third of cases. However, the fact that in two thirds of cases, these actions were incorrectly judged as meaningless, underlines that body orientation and contour still were important cues for viewers' decisions, and thus – with atypical stimuli like the probe actions - misleading.

Together, for judgments on action coherence, global layout aspects can be extracted from the scene even with presentation times as brief as 30-50 ms. We assume that observers made their coherence decision on the basis of a set of layout features. However, the internal representation of the action is rather coarse and would need focussed attention to provide enough detail to verify or correct the decision.

Action-Object Consistency

As expected, performance rates for scenes with variation of local scene information were worse than for those from the global scene layout set. Detecting whether the action and the object were semantically consistent was hardly possible with all

[2] We also considered analyzing our data by using *agent orientation* and *patient orientation* as two levels of the factor Body Orientation instead of the four orientation combinations as levels. However, this type of analysis does not illustrate the global layout effects adequately.

exposure times. Only with 100 ms, viewers sometimes judged coherence correctly, which indicates that they managed to identify both action and object. However, most of the action scenes were judged as meaningful, with percentages corresponding roughly to those of the 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 wasn't possible to identify both action and object. Recall that both scene types were presented in random order to the subjects. Due to the very similar configuration of all images, our observers were probably not aware of being confronted with two coherence conditions, and therefore applied the same strategy for all items.

As the participants were not asked to name actions and objects, it is unclear whether any of these scene components were identified. In the Dobel et al. study (2007), actions and objects could be hardly identified with 100 ms presentation. The fact that, in the present study, the objects were naturalistic (see Braun, 2003), coloured, larger, and positioned centrally in the images, enhances the likelihood of their identification. Whether an action can be identified after such brief presentation depends on the uniqueness of its spatial layout (Glanemann et al., under revision). Given that in the present study, the identification of the object or the action alone did not suffice for a correct decision, we do not go into any detail here.

Underlying Mechanisms of Early Action Scene Processing

Our results show that, after 30 – 100 ms masked presentation of an action scene, internal visual representations can be detailed enough to infer a correct coherence judgement from the global spatial layout formed by the two actors. However, the visual representations are not detailed enough to yield correct semantic representations of both the depicted object and the action. These findings can be easily explained by mechanisms previously suggested for the rapid perception of complex scenes.

We propose that the coherent/incoherent categorization of the actions was mainly based upon feed-forward and parallel processing of a set of image features related to the whole action's spatial layout. The potentially relevant features were an action's contour (symmetric or ragged), body and face orientations of the two actors (towards or backwards the other person) and the interspace (filled or empty) between them. Note that the term "feature" has been used in the literature with regard to different processing levels, namely to visual information at a low (e.g., Kaping, Tzvetanov & Treue, 2007; Oliva & Torralba, 2006: local features), intermediate (Evans & Treisman, 2005; Oliva & Torralba, 2006: global features; Ullman, Vidal-Naquet & Sali; 2002) or high processing level (e.g., Li et al., 2005; Rousselet et al., 2005). The features related to the spatial layout reported here fit best to the intermediate category.

Further, we assume a quite strong top-down influence in that even initial visual representations are already task-specific (cf. Treisman, 2006; Oliva & Torralba, 2006, Torralba, Oliva, Castelhano & Henderson, 2006; Hochstein & Ahissar, 2002). In the present experiment, the instruction to judge an action's coherence produced strong constraints on the extraction of available image information. Whereas, for example, colour or background texture were non-critical features for the decision on coherence, the configuration of the bodies' contours as coded by low-level spatial frequency information was one key diagnostic cue.

The limits of this initial stage in visual processing without focussed attention become apparent when the task becomes more difficult. For example, Evans and Treisman (2005) demonstrated decreasing performance in animal/non-animal categorization when there is overlap between the feature sets of the target category and the category the targets have to be discriminated from, such as animals versus humans. Moreover, in that study, tasks that necessitated more detailed representations than categorization, such as target localization or identification of the target's subcategory, also resulted in worse performance. With respect to action scenes, we found earlier that actions 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 sb.* (Glanemann et al., under revision). Similarly, in the present categorization task, layout features that were atypical for coherent action scenes predominantly led to incorrect responses.

These findings demonstrate that visual representations built on the fly are incomplete. They are of great value for a first – and most often correct - impression of what is in our field of vision. Moreover, they 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 et al., 2007) via a "rapid pass through the hierarchy of visual processing" (Evans & Treisman, 2005) needs focussed attention to generate a more detailed visual representation, which then allows for 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.

In sum, we believe that mechanisms that generate a coarse spatial scene representation are the basis of coherence decisions on action scenes.

The Value of Rapid Action Scene Processing

As shown here and in our earlier work, the extraction of substantial information from an action scene, such as coherence, thematic roles or even the action itself, does not necessarily need selective attention on the action-relevant image regions (Dobel et al., 2007; Glanemann et al., under revision). Why is our brain capable of such a remarkable achievement? Research on gist apprehension of natural scene categories repeatedly suggested that fast understanding of briefly viewed scenes may have survival value (Bacon-Macé et al., 2005; Li et al., 2002). In the same line, it was suggested that familiar and meaningful visual stimuli may have a processing advantage (Li et al., 2005). These arguments can be applied to action scenes, too. Actions are essentials of our everyday life. We perform them ourselves for own or

other-person directed purposes. Furthermore, we observe and interpret the actions of others, and sometimes, we have to *re-act* quickly on other persons' actions, even if these happen in the visual periphery, or, if our attention is bound to something else, or even both. Imagine, for example, a mother talking to her friend while her two younger children are playing in the background. She would certainly want to interfere in a potentially aggressive action, and even more if a sharp instrument is involved. Following the argumentation of Li et al. (2002, 2005) and Bacon-Macé et al. (2005), a top-down guided 'pre-tuning' for familiar visual stimuli, in our case for everyday actions, that supports rapid perception and interpretation, particularly with respect to meaningfulness, danger, etc. would be extremely useful. To err on the side of caution, the mother would probably intervene even if the instrument of threat turns out to be a rubber spoon. In this context, it would be interesting, whether actions with high emotional value, such as threatening or kissing, are faster and easier to identify than actions with low emotional value, such as giving or talking (see, e.g., Calvo & Lang, 2005).

To conclude, we have shown that the extraction of visual information that is useful for judging coherence of an agent-patient action scene can be as fast as perceiving the gist of an environmental scene. The results of the present study support our hypothesis that the representations of early visual perception are sufficient to perceive the spatial layout of an action scene but hardly to identify the semantic relation between an action and the object used in this action.

It will be interesting to see 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 results are not only of interest for the study of visual perception per se. Action scenes also 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, in press). As claimed by the latter authors, knowledge about the depth and abstractness of information from non-fixated scene regions is crucial for

the question as to how serial or parallel visual apprehension and linguistic formulation are.

Appendix

List of actions used in the experiments

A1. Body Orientation set

to shove sb., to kick sb., to blindfold sb., to push sb., to scare sb., to brush sb.'s hair, to hit sb., to put lotion on sb., to handcuff sb., to stab sb.

A2. Probe items for the Body Orientation set

to pull sb., to kick backwards at sb., to spray sb.

B. Action-Object Consistency set

	appropriate object	*inappropriate object*
to shoot sb.	pistol	hand brush
to portray sb.	paintbrush	croissant
to light a cigarette	cigarette	toothbrush
to serve sb. a coffee	cup	shoe
to varnish sb.'s nails	applicator	cake fork
to help sb. into a coat	coat	bin liner
to bandage sb.'s arm	bandage	wire mesh
to feed sb.	spoon	eyeglasses
to give sb. money	banknote	compact disc
to cut sb.'s hair	scissors	wooden spoon

Summary and Conclusions

CHAPTER 4

Human vision is an active process (e.g. Hayhoe, 2000). What we see at any point in time is not an automatically created, complete and detailed internal representation of the external world. Rather, it is a subjective and transient representation that is based on individual knowledge, expectation, and the momentary state of attention, the latter two being strongly influenced by the current task (e.g., Triesch, Ballard, Hayhoe & Sullivan, 2003; Ullman, 1984). This is why human vision exhibits the fascinating wide spectrum ranging from 'not seeing what we are looking at' (change/inattentional blindness), to 'seeing what we are not looking at' (perception of stimuli presented briefly or in the visual periphery). In other words, looking at something (by eye fixations) and seeing it (i.e., consciously perceiving) can be dissociated under certain circumstances.

The experiments in this dissertation dealt with the rapid generation of visual representations of naturalistic action scenes. My overall interest was whether these representations, which are built within a fraction of a second and in the absence of focussed attention, are detailed enough to conceive essential semantic aspects of the depicted action. By registering eye movements and using brief and/or peripheral visual stimulus presentation, I studied what can be 'seen without looking at' action scenes.

More specifically, there were two general research questions that I sought to find answers for. The first was related to the *what*, the second to the *how* in rapid action-scene analysis. With respect to *what*, I investigated what type of abstract semantic information can be extracted from a novel action scene with 150 ms (and less) exposure time. Is it possible to assign thematic roles to the two actors?

Is it even possible to recognize the depicted action? Can viewers decide whether an action makes sense or not, when 'sense' is varied by the individual mirroring of the two involved actors? Can they decide whether the object used in an action is the appropriate one for that action? Concerning *how,* the main question was whether the visual information most efficient for solving the experimental tasks is related to the action's 'whole' or 'its parts'. The 'whole' refers to the global spatial layout of the scene as formed by the body postures of the two actors, whereas the 'parts' refer to the scene's components, such as the identity of the object used in this action.

In what follows, I first briefly summarize the experiments reported in Chapters 2 and 3. Secondly, I discuss which information could be extracted from briefly presented action scenes, what the studies reveal about the time course and about the underlying mechanisms of rapid action scene perception. Next, I compare the action scenes of the type used here with other types of action scenes and with environmental scenes. This is followed by suggesting which other methods and research areas seem promising for contributing to my research questions. Finally, I will discuss the implications of my work with respect to interpreting eye-movements in language tasks and generally in scene perception.

Summary of Chapter 2

The studies in *Event Conceptualization at free View and at an Eyeblink* investigated the rapid perception of thematic roles with briefly and peripherally presented photographs of agent-patient-actions. The first study employed eye-tracking as on-line measure of patient detection. Although the viewers' eyes went preferentially to the agent it was unclear whether – prior to this initial fixation - roles had been identified (and first gazes reflected, for example, monitoring processes), or whether the agent was merely targeted due to higher visual saliency. That is, the results were indecisive as to whether the patient was identified during the initial peripheral preview phase. The second study revealed that, indeed, 150 ms masked presentation were sufficient to identify the patient of the action. Astonishingly, the same

presentation time allowed for the correct naming of the depicted action in nearly 60 % (Study 3). Although generally, performance in a naming task cannot be compared to performance in a two-alternative forced choice task, it can be concluded nevertheless that action identification was not a necessary prerequisite for patient detection in Study 2. The fourth study strengthened the hypothesis that the global spatial layout of an action scene, as perceived peripherally, facilitated action identification in Study 3.

In sum, 150 ms masked peripheral presentation of action scenes allowed for the assignment of thematic roles to the two involved persons, and in nearly 60 % of cases for the correct naming of the action. It is most likely that for both tasks, the gross spatial layout as formed by the scene's components played the main role for this excellent performance.

At this point, I should emphasize that this hypothesis has yet to be proved. The initial aim of this work was to find out whether thematic roles and actions can be identified after short and peripheral presentation. The global layout-hypothesis developed more or less over the course of the investigations. All I suggest here is that, with the type of action scenes used here, the findings are consistent with the idea that the gross spatial layout of an action scene can be perceived extremely rapidly and used to decide on thematic roles and actions.

Summary of Chapter 3

In contrast to the studies summarized above, the key interest in *Rapid Apprehension of Coherence of Action Scenes* was to contrast two types of coherence manipulation, in order to see which kind of visual information is processed more efficiently in early action scene perception. Previous work already demonstrated that action coherence manipulated by body orientation of the two involved actors can be judged correctly on the basis of 100 ms-presentation (Dobel et al., 2007). The experiments reported here investigated in more detail how coherence perception comes about. Variations of the global layout were contrasted with variations of the semantic consistency of

action and object. Additionally, I examined whether stepwise reduction of the presentation time from 100 to 20 ms would differentially affect the viewers' performance with the two types of information.

Action-object consistency could hardly be recognized, even with 100 ms. In contrast, with three of the four body orientation combinations, masked presentation of 50 ms duration was sufficient to judge coherence correctly in 80 – 90 % of cases. And with 30 ms, overall performance in this condition was still significantly better than chance. The results confirmed the initial hypothesis that holistic layout properties of the scene can be extracted earlier than visual information needed to judge on action-object consistency.

Which information about the action can be extracted based on brief presentation?

This question is closely related to the issue of how detailed early visual scene representations are (e.g., Dobel et al., 2007; Simons & Rensink, 2005; O'Regan & Noe, 2001). As to the detailedness of representations, the answer is not straight-forward. The terminology mostly used to characterize the quality of internal visual representations is inconsistent: "One investigator's 'relatively detailed' representation may be another investigator's 'relatively sparse' representation" (Peterson & Rhodes, 2003, p. 16). Therefore, it seems more adequate to describe which tasks can be performed with a particular stimulus type, and which other tasks cannot.

From the results in Chapters 2 and 3, we learned that the rapidly built visual representations were detailed enough to assign thematic roles, to decide whether an action was meaningful or not (when meaningfulness was manipulated by spatial layout) and, in many cases, to name the depicted action. In most cases, they afforded a correct "first best guess" (Kaping, Tzvetanov & Treue, 2007) on the task. However, these representations were incomplete, as demonstrated by the viewer's worse performance with (1) atypical (e.g. coherent back-to-back actions in coherence judgements) and (2) ambiguous stimuli (e.g. ambiguous body postures in

action naming), and (3) with tasks that cannot be fulfilled on basis of the spatial layout alone (e.g. meaningfulness manipulated by action-object consistency).

Despite this qualification, viewers extracted a great deal of semantic information from an action scene when the scene was presented for just a fraction of a second. Nevertheless, before drawing premature conclusions about how fast the understanding of action scenes might be, it is worth looking closer at the time course of scene perception and response giving in the present experiments.

The Time Course of Action Scene Perception

Visual representations of agent-patient actions can be built extremely rapidly. However, there is one caveat related to the interpretation of the presentation times I used in the experiments. Whereas brief and masked presentation was sufficient to extract complex visual information from an action scene, we still do not know at which point in time – after stimulus onset – this information was consciously available. Hence, the presentation times used in the experiments here should not be misinterpreted as the absolute time needed for *understanding* thematic roles or action coherence. Rather, the presentation time that was minimally needed for a particular correct answer displays only the lower time-limit of apprehension. That is to say, although the backward mask terminated the extraction of image information, visual-cognitive processing continued until viewers gave their answer (and theoretically even beyond). The upper time limit for decision-relevant information processing was provided by the response (button press or naming, if recorded), minus the time needed for monitoring and response execution. As a consequence, it may well be that making correct responses was only possible after a certain amount of continued cognitive processing after stimulus offset. This issue is of particular interest for studying the interface between vision and language. The earlier visual information is processed beyond a basic perceptual level, the more likely it has immediate influence on the ongoing visual-cognitive and linguistic processes (for a more detailed discussion on this issue, see the paragraph on language-vision interface below).

A means that has been repeatedly used to investigate the time course of higher-level decision making in categorization tasks are event-related potentials (ERPs; e.g., Thorpe, Fize & Marlot, 1996; Rousselet, Fabre-Thorpe & Thorpe, 2002). ERPs reveal signs of neural decision-related processing well before any motor output. For tasks, such as, whether the patient is located on the left or right side of the scene, or, whether an action is meaningful or meaningless, ERPs might illuminate the exact time-course of decision making by distinguishing between category-specific brain activity related to low-level image analysis and activity related to high-level decision making. In addition to the temporal information, ERPs measured by electroencephalography or magnetencephalography reveal spatial information, which allows to relate a specific ERP to a particular region of the brain. This in turn aids to identify the cognitive process underlying the specific ERP. For example, if there is differential brain activity found for yes- versus no-answers with respect to the meaningfulness of an action as soon as 50 ms after stimulus onset, and further, this differential activity can be assigned to (pre-) frontal brain areas, this finding would suggest that a great deal of visual processing has been completed by this time (see e.g. Thorpe et al., 1996). On the other hand, if differential frontal brain activity arises only some hundred milliseconds after stimulus onset, it is reasonable to assume that the visual-cognitive processes that continue after short stimulus presentation are mandatory for the correct judgments on meaningfulness.

Mechanisms underlying early Action Scene Perception

With respect to the mechanisms of early action scene perception, the results of both research projects reported here assign a key role to the gross spatial layout or Gestalt of the scene. This layout is formed by body postures and orientations of agent and patient, the spatial relationship between them and the overall contour of the scene. The spatial layout of the actions used here vary, for example, with respect to whether the two actors have physical contact (e.g., *to brush sb.'s hair* as opposed to *to scare sb.* and to the back-to-back actions in the coherence experiment), or whether

arms and legs keep close to an actor's body or reach out (e.g., *to photograph sb.* as opposed to *to kick sb.*).

The results indicate that the viewers utilized these layout differences between action scenes for task completion rather than having access to a detailed semantic representation of an action. In the Patient Detection task, viewers could assign thematic roles correctly by detecting the less active actor in the scene. Naming the depicted action was achieved by mapping an action's Gestalt onto an existing action representation (with the consequence of errors in case it could be mapped onto more than one representation).

In the experiment *Rapid Apprehension of Coherence of Action Scenes*, the differential results for the four combination types suggest that viewers made their judgements on the basis of so-called sets of features. The features are related to properties of the spatial layout, such as the orientation of an actor (towards or backwards to the other), whether the overall contour is ragged due to the out-reaching arms of the actor (back-to-back and patient-facing-agent's-back actions), or whether the interspace between the two actors is filled or empty. The term set is used to demonstrate that more than one feature can be perceived in parallel. A given set, that is, a specific combination of features, is assumed to be diagnostic for a task-specific judgement. For example, the co-occurrence of a ragged contour, both actors being oriented backwards to each other, and an empty interspace between them is diagnostic of an incoherent action. On the other hand, a symmetric contour, towards orientated actors and a filled interspace point to a coherent action. As described in the discussion of Chapter 3, the individual features of a particular set can carry different weights, and thus be more or less important for the coherence judgments than other features within the set. Although the results of the coherence experiment can be plausibly explained by task-specific, i.e., top-down-guided feature sets, this post-hoc hypothesis needs verification in future research.

Together, I assume that early action scene perception is characterized by parallel and feed-forward processing of task-specific (sets of) features related to the scene's

gross spatial layout. Next, I discuss to what extent the present findings apply for other types of actions.

Generalisation of the Findings to other Action Scene Types

Action scenes are a specific type of scene, and the agent-patient actions used here represent a specific type of action. Therefore, our results cannot readily be transferred to other types of action scenes, such as two actors occupying the same thematic role (e.g. shaking hands), many actors performing the same action (e.g., playing football), or two actors independently performing different actions (e.g., dancing and fencing), and many more. Certainly, other types of actions imply other research questions. However, research on diverse action types might be very insightful as to what the mechanisms are, which generally underlie rapid action scene perception irrespective of action diversity and apart from using the overall spatial layout. For example, if the repeatedly proposed hypothesis that highly familiar stimuli have a processing advantage due to pre-existing neuronal representations (Li, VanRullen, Koch & Perona, 2005) and a stronger top-down presetting (Rousselet, Joubert & Fabre-Thorpe, 2005) is true, we should find better performance with high-frequent than with low-frequent actions. Equally, if action scenes are found to be processed faster when their emotional (positive or negative) value is high than when it is low, this would corroborate the hypothesis that visual stimuli, which have survival value, have a processing advantage (e.g., Bacon-Mace et al., 2005; Li, VanRullen, Koch & Perona, 2002).

The following section compares action and environmental scenes with respect to their characteristics and underlying mechanisms of the early perceptual analysis. From the comparison between the two scene types we can gain knowledge about which mechanisms are specific for all visual scenes and which play a role only in certain types of scenes. This in turn is insightful with respect to our understanding of the build-up and nature of internal visual representations.

Commonalities and Differences between Action Scenes and Environmental Scenes

The present experiments demonstrate that task-specific internal visual representations of agent-patient action scenes can be built extremely rapidly, and within an approximately similar time frame as suggested for environmental scenes (Henderson & Ferreira, 2004). Both scene types have in common that they contain semantic information that exceeds the identity of the scene's components, such as the gist in environmental scenes and the action and its structure (coherence, thematic roles) in action scenes. So, given what is known about the rapid extraction of the gist of environmental scenes, my findings may seem unsurprising at first glance. For example, spatial layout information was also found to be diagnostic for the categorization of briefly presented environmental outdoor scenes (Schyns & Oliva, 1994; Oliva & Schyns, 1997). However and as mentioned earlier, action scenes differ from environmental scenes in some critical aspects.

First, remember that an action scene is only a discrete snapshot of an event that unfolds over time. Hence, one of the first processes in action scene perception is the activation of the dynamic mental event representation that a given scene stands for. With environmental scenes, however, this mental representation is of static nature. It seems reasonable to assume that activating a dynamic representation needs more processing capacity than activating a static representation.

Second, the most elaborated model of the fast perception of scene gist to date (Oliva & Torralba, 2006) implements a scene-wide analysis of global features, which are in turn specific configurations of low-level features. For the type of action scenes used here, however, this analysis would presumably not result in differential responses to the experimental conditions as they hardly differ with respect to basic local scene information, such as colour or texture. This becomes easily apparent if one compares the action stimuli between the conditions in my experiments (i.e., patient right - patient left, action A - action B, coherent action - incoherent action), to the variety of environmental scenes between which are differentiated in the account of Oliva & Torralba (2006; e.g., beach - snowy mountains - urban scene). Obviously, the environmental scenes vary much more with respect to low-level

image features than the action scenes used here. In the latter, most low-level information is very similar between all items and thus non-differential with respect to the tasks. Differences between the action scenes become apparent only at a higher processing level, when spatial relations come into play (e.g., in form of global image features; Oliva & Torralba, 2006).

Third, the main components of action scenes are human beings whereas environmental scenes may contain humans or animals, such as in a street or farmyard scene, but they are not mandatory components and, critically, they are not placed in the foreground of the scene. On the one hand, this should not influence global layout processing. On the other hand, there are accounts that propose a general processing advantage for living entities (e.g., Thorpe, Fize & Marlot, 1996; Li et al., 2005).

A further difference lies in the composition of the two scene types. Typically, the components of environmental scenes are distributed over different depth levels, that is, they can be found in the fore- or background of the image, or somewhere in between. In contrast, the actors (and optional objects) in our scenes are positioned at the same depth level. Whether depth differences within an action scene are easier or more difficult to process compared to one-depth-level scenes can be a topic of future research.

Taken together, although both environmental and action scenes share many characteristics they also differ in some critical aspects. I suggest that the mechanisms underlying the rapid information extraction from the two scene types are very similar in that they rely on spatial layout aspects, but they are not quite the same. The follwing section is concerned with other research methods that might also be insightful for understanding early action scene perception and potential differences to environmental scene processing.

Potential Contribution from other research methods and areas

Cognitive Neuroscience contributes significantly to investigating the underlying mechanisms of scene processing. As mentioned above, ERP-measurements may shed light on the exact time course. Moreover, neuroimaging researchers found, for example, that a region in the medial occipito-temporal cortex, the parahippocampal place area (PPA), is a critical component of scene processing (Epstein, 1998). Importantly, the PPA does not respond to objects. More recently, differentially located brain activity was observed for photographs of familiar versus unfamiliar locations (Epstein, Higgins, Jablonski & Feiler, 2007), for indoor versus outdoor scenes (Henderson, Larson & Zhu, 2007) and for rapid categorization versus conscious report for photographs of real-world scenes (Marois, Yi & Chun, 2004). The latter study employed the attentional blink paradigm and functional magnet resonance imaging. Even though in- and outdoor scenes remained frequently undetected by the viewers the PPA area showed activation. With conscious perception this activity increased.

Although only environmental scenes have been investigated so far, it will be interesting to see what the neurophysiological and neuroanatomical correlates of action scene processing are. Do they activate the same regions as environmental scenes or is there more overlap with object or face processing? Can we even find action-scene specific activity, potentially (and highly speculatively) in motor-related brain areas in the sense of 'mirror activity'? Do emotional value and familiarity of an action influence the intensity or the temporal course of brain activity? Especially the loci of rapid categorization versus conscious report, as in the study of Marois et al. (2004), would provide insight into the mechanisms underlying the perception and understanding of action scenes.

A further discipline that has advanced the understanding of human visual processing, which might be insightful for studying action scene perception, is Cognitive Neuropsychology. Studying people with acquired neuropsychological deficits has broadened our knowledge about functional and structural brain organisation. For example, there is the phenomenon of visual agnosia, the inability

to recognize certain types of visual stimuli. As in the research on normal vision, the main focus in the research of visual agnosia is on object cognition, not on scene cognition. However, subtypes of visual agnosia have been described with specific difficulties in recognizing spatial representations between multiple objects (*visual simultanagnosia*; see, e.g., Behrmann, 2005). It seems reasonable to assume that these patients also exhibit difficulties in perceiving scene-related image information that relies – among other factors – on spatial relations, such as the gist of an environmental scene and (the coherence of) a depicted action.

A subtype of visual agnosia in which people have a specific problem with recognizing faces, might also advance our understanding of rapid (action) scene perception. The main problem in *prosopagnosia* is the failure to process configurational information (e.g., Righart & de Gelder, 2007; for a review, see Behrmann & Avidan, 2005). Critically, this deficit is much more expressed in face than in object perception. "A configural impairment may affect other visual stimuli too if, as is true for faces, the spatial relations between the components need to be represented to differentiate perceptually similar exemplars" (Behrmann & Avidan, 2005, p. 183). Recently, this account was corroborated by the finding that people with congenital prosopagnosia also displayed a severe impairment in processing biological motion (Lange, de Lussanet, Kuhlmann, Zimmermann, Lappe, Zwitserlood & Dobel, in prep.).

As a third approach, our understanding of rapid scene perception may benefit from looking at parallel processes in vision and hearing. Although the main focus in auditory scene research is on *separating* the sounds and sound streams that are often coevally present in the environment (reviewed by Goldstein, 2007, p. 217ff.), we also find a stimulus type, namely the musical chord, which exhibits the key characteristic of a scene: In the same way as a scene is more than the sum of its components, a musical chord is more than just the sum of its tones. The additionally conveyed information that evolves from particular relations between the single tones is whether the chord is disharmonious or harmonious, and in the latter case, whether it is a 'major' and 'minor' chord. Although in musical chords, there are no

analogue categories to the environmental and action scenes in vision, they can be examined with respect to holistic processing. Investigating the psychology and neural base of rapid chord perception and comparing it to rapid scene perception in vision can tell us whether holistic processing mechanisms are modality-specific or if they have a more general function across domains of cognition as was suggested, for example, for the "unification" role of Broca's area for syntactic, semantic, and phonological unification (Hagoort, 2005).

Together, in addition to behavioural experiments on the rapid visual scene perception of healthy people, studying the neuroanatomical correlates of action scene perception and people with deficits in holistic visual processing enables us to examine the neuronal and psychological mechanisms underlying rapid action scene processing and their relationship to non-action-scene processing. Moreover, comparing the early perception of holistic stimulus characteristics in the visual and auditory sense might reveal whether the mechanisms found for rapid visual scene perception are specific to the modality of vision.

The last two sections of this chapter discuss the implications of my investigations in more general, firstly, with respect to the application of eye-movement recording in language tasks and, secondly, with respect to what can (not) be concluded from my findings for non-foveal visual processing.

The Language-Vision Interface

What do the present results imply for the use of action scenes in the study of the interface between vision and language? The interpretation of fixating a scene region for the first time depends largely on how far this region had already been processed prior to this fixation. The results reported here confirm the findings of other work (Bock et al., 2003; Dobel et al, 2007; Gleitman et al., in press; Morgan & Meyer, 2005) in that the processing of not fixated scene regions goes already beyond a low-level perceptual analysis and can deliver speech-relevant scene information. Thus, a saccade towards a given scene region could serve the apprehension of the scene, but

it could also be the *result* of pre-attentive (i.e. in the absence of overt attention) apprehension, and, most likely, "visual apprehension and linguistic formulation in scene description are tightly coupled, rapid and not dissociable in a stage-like fashion" (Gleitman et al., in press). In other words, the present work would be consistent with the notion that apprehension and formulation are not two distinct processes, but timely and functionally intertwined. Importantly, the quick extraction of information is likely to influence the eye movement behaviour not only at stimulus onset but throughout the whole scan path.

The consequence for using eye-tracking in the research on speech comprehension and production is that the possibility of parafoveal high-level processing should always be considered in result interpretation.

Does Peripheral Preview supersede Eye Fixations in Action Scenes?

With our type of stimuli, quite a lot of essential semantic information about the depicted action can be accessed without fixating agent, patient or object. Does this imply that, in tasks that require role identification, eye movements on agent or patient do not serve the identification of thematic roles? Certainly not. Remember that the rapidly extracted visual representation is a 'first best guess' that needs subsequent verification. Eye movements are made at very low cost. In sentence production, for example, they can fulfil many functions, such as monitoring, memory support, interference avoidance or production support (Griffin, 2004). That is to say, the present findings should not be over-interpreted with respect to the achievements of parafoveal and peripheral visual processing. It is important to consider that the presentation mode we chose is excellent for investigating high-performance vision. At the same time, it is a quite artificial situation to view scenes for such short time and miles away from normal viewing conditions. The same argument was made earlier by Findlay (2004), with regard to the finding that covert attention and eye fixation can be dissociated as demonstrated in the seminal work of Posner (1980). "The assumption is often implicit that covert attention can substitute

for eye movements […]. I suggest that this conclusion is misleading. Covert attention generally acts to supplement eye movements rather than to substitute for them" (Findlay, 2004, p. 156). Similarly, I suggest that parafoveal vision supports eye movements rather than to substitute for them.

The research in this dissertation belongs to the very few work so far that deals with the rapid build-up of visual representations of action scenes. Altogether, the speed of information extraction from very briefly and peripherally presented action scenes is intriguing. Some of the present results were the contrary of what I had expected before conducting the experiments. Future research will investigate the issues in more detail and will be able to answer the questions that remained open.

References

Antes, J. R. (1974). The time course of picture viewing. *Journal of Experimental Psychology, 103,* 62-70.

Bacon-Mace N., Mace, 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.

Bar, M. (2004). Visual objects in context. *Nature Neuroscience Reviews, 5,* 617–629.

Baron-Cohen, S. (1995). *Mindblindness. An Essay on Autism and Theory of Mind.* Cambridge, MA: MIT Press.

Behrmann, M. (2005). Neuropsychological Approaches to Perceptual Organization – Evidence from Visual Agnosia. In M. A. Peterson & G. Rhodes (Eds.) *Perception of Faces, Objects, and Scenes: Analytic and Holistic Processes.* OUP, USA.

Behrmann, M., & Avidan, G. (2005). Congenital prosopagnosia: Face-blind from birth. *Trends in Cognitive* Sciences, 9(4), 180-187.

Biederman, I. (1972). Perceiving real-world scenes. *Science, 177(4043),* 77-80.

Biederman, I., Mezzanotte, R. J., & Rabinowitz, J. C. (1982). Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology, 14,* 143-177.

Biederman, I. (1988). Aspects and extensions of a theory of human image understanding. In Z. W. Pylyshyn (Ed.): *Computational Processes in Human Vision: An Interdisciplinary Perspective* (pp. 370-428). Norwood, NJ: Ablex.

Biederman, I. (1995). Visual object recognition. In S. F. Kosslyn & D. N. Osherson (Eds.): *An Invitation to Cognitive Science: Visual Cognition* (pp. 121-165). Cambridge, MA: MIT Press.

Bock, K., Irwin, D. E., Davidson, D. J., & Levelt, W. J. M. (2003). Minding the clock. *Journal of Memory and Language, 48*, 653-685.

Bock K., & Levelt, W. (1994). Language production. Grammatical encoding. In: M. A. Gernsbacher (Ed.): *Handbook of psycholinguistics*, Academic Press.

Boyce, S. J., Pollatsek, A., & Rayner, K. (1989). Effect of background information on object identification. *Journal of Experimental Psychology: Human Perception and Performance, 15,* 556-566.

Braun, J. (2003). Natural Scenes upset the visual applecart. *Trends in Cognitive Sciences, 7(1)*, 7-9.

Buswell, G. T. (1935). *How People Look at Pictures.* Chicago: University of Chicago Press.

Calvo, M. G., & Lang, P. J. (2005). Parafoveal semantic processing of emotional visual scenes. *Journal of Experimental Psychology: Human Perception and Performance, 31(3),* 502-519.

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.

Chatterjee, A., Southwood, M. H., & Basilico, D. (1999). Verbs, events and spatial representations. *Neuropsychologia 37,* 395-402.

Davenport, J. L., & Potter, M. C. (2004). Scene Consistency in Object and Background Perception. *Psychological Science, 15,* 559-564.

De Graef, P. (2005). Semantic effects on object selection in real-world scene perception. In G. Underwood (Ed.): *Cognitive processes in eye guidance* (pp. 189-211). Oxford: University Press.

De Graef, P., Christiaens D., & d'Ydewalle G. (1990). Perceptual effects of scene context on object identification. *Psychological Research, 52*, 317-329.

Deubel, H., & Schneider, W. X. (1996). Saccade target selection and object recognition: Evidence for a common attentional mechanism. *Vision Research, 36,* 1827-1837.

Dobel, C., Diesendruck, G., & Bölte, J. (in press). How writing system and age influence spatial representations of actions - a developmental, crosslinguistic study. *Psychological Science.*

Dobel, C., Glanemann, R., Kreysa, H., Zwitserlood, P., & Eisenbeiss, S. (in press). Visual encoding of meaningful and meaningless scenes. In: E. Pedersen & J. Bohnemeyer (Eds.): *Event Representation in Language: Encoding Events at the Language-Cognition Interface.* Cambridge University Press.

Dobel, C., Gumnior, H., Bölte, J. & Zwitserlood, P. (2007). Describing scenes hardly seen. *Acta Psychologica, 125 (2),* 129-143.

Egeth, H. E., & Yantis, S. (1997). Visual attention: control representation and time course. *Annual Review of Psychology, 48*, 269-297.

Epstein, R. A. (2005). The cortical basis of visual scene processing. *Visual Cognition: Special Issue on Real-World Scene Perception, 12,* 954-978.

Epstein, R. A., Higgins, J. S., Jablonski, K., & Feiler, A. M. (2007). Visual scene processing in familiar and unfamiliar environments. *Journal of Neurophysiology, 97(5*), 3670-3683.

Epstein, R., & Kanwisher, N. (1998). A cortical representation the local visual environment. *Nature*, *392 (6676),* 598-601.

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.

Findlay, J. M. (2004). Eye scanning and visual search. In J. M. Henderson, and F. Ferreira (Eds.): *The interface of Language, Vision, and Action: Eye movements and the Visual World* (pp. 135-159). New York: Psychology Press.

Fischer, B., Gezeck, S., & Hartnegg, K. (1997). The analysis of saccadic eye movements from gap and overlap paradigms. *Brain Research Protocols 2 (1),* 47-52.

Friedman, A. (1979). Framing pictures: The role of knowledge in automatized encoding and memory for gist. *Journal of Experimental Psychology: General, 108*, 316-355.

Garrett, M. F. (1976). Syntactic processes in sentence production. In R. J. Wales & E. Walker (Eds.): *New Approaches to Language Mechanisms* (pp. 231-255), Amsterdam: North Holland.

Glanemann, R., Dobel, C., Bölte, J., & Zwitserlood, P. (2007). Rapid Apprehension of Gist in Action Scenes. *Proceedings of the European Cognitive Science Conference,* Greece, 904.

Glanemann, R., Zwitserlood, P., Bölte, J., Kreysa, H., & Dobel, C. (under revision) *Event conceptualization in free view and at an eyeblink.*

Gleitman, L. R., January, D., Nappa, R., & Trueswell, J. (in press). On the give and take between event apprehension and utterance formulation. *Journal of Memory and Language,* doi:10.1016/j.jml.2007.01.007.

Goldstein, B. E. (2007). *Sensation & Perception* (7th ed.). Belmont, CA: Wadsworth.

Griffin, Z. (2004). Why look? Reasons for eye movements related to language production. In J. Henderson & F. Ferreira (Eds.): *The Interface of Language, Vision, and Action: Eye Movements and the Visual World* (pp. 213-247). New York: Psychology Press.

Griffin, Z., & Bock, K. (2000). What the eyes say about speaking. *Psychological Science, 11(4*), 274-279.

Griffin, Z., & Spieler, D. H. (2006). The influence of age on the time course of word preparation in multiword utterances. *Language and Cognitive Processes, 21(1-3),* 291-321.

Hagoort, P. (2005). On broca, brain, and binding: A new framework. *Trends in Cognitive Sciences, 9(9),* 416-423.

Hayhoe, M. (2000). Vision using routines: A functional account of vision. *Visual Cognition, 7(1-3),* 43-64.

Henderson, J. M. (2005). Introduction to real-world scene perception. *Visual Cognition: Special Issue on Real-World Scene Perception, 12,* 849-851.

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: Psychology Press.

Henderson, J. M., & Hollingworth, A. (1999). High-level scene perception. *Annual Review of Psychology, 50,* 243-271.

Henderson, J. M., Weeks, P. A. Jr., & Hollingworth, A. (1999). Effects of semantic consistency on eye movements during scene viewing. *Journal of Experimental Psychology: Human Perception and Performance, 25*, 210-228.

Hochstein, S., & Ahissar, M. (2002). View from the top: hierarchies and reverse hierarchies in the visual system. *Neuron, 36,* 791–804.

Hoffman, J. E. (1975). Hierarchical stages in the processing of visual information. *Perception and Psychophysics, 18 (5),* 348-354.

Hoffman, J. E., & Subramaniam, B. (1995). The role of visual attention in saccadic eye-movements. *Perception & Psychophysics, 57(6),* 787-795.

Hollingworth, A., & Henderson, J. M. (1998). Does consistent scene context facilitate object perception? *Journal of Experimental Psychology: General, 127*, 398–415.

Hollingworth, A., & Henderson, J. M. (1999). Object identification is isolated from scene semantic constraint: evidence from object type and token discrimination. *Acta Psychologica, 102,* 319–343.

Intraub, H. (1981). Rapid conceptual identification of sequentially presented pictures. *Journal of Experimental Psychology: Human Perception and Performance, 7,* 604-610.

Irwin, D. E. (2004). Fixation location and fixation duration as indices of cognitive processing. In J.M. Henderson & F. Ferreira (Eds.): *The Interface of Language, Vision, and Action: Eye Movements and the Visual World* (pp. 105-133).New York: Psychology Press.

Jackendoff, R. (1997). *The Architecture of the Language Faculty.* Cambridge, Mass.: MIT Press.

Kamide, Y., Altmann, G., & Haywood, S. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language, 49(1),* 133-156.

Kaping, D., Tzvetanov, T., & Treue, T. (2007). Adaptation to statistical properties of visual scenes biases rapid categorization. *Visual Cognition, 15,* 12-19.

Kirchner, H., & Thorpe, S. J. (2006). Ultra-rapid object detection with saccadic eye movements: Visual processing speed revisited. *Vision Research, 46(11),* 1762-1776.

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.

Kreysa, H., Zwitserlood. P., Boelte, J., Glanemann, R., &. Dobel C. (submitted). *Where is the action? An Eyetracking study on the Description of naturalistic events.*

Lange, J., de Lussanet, M., Kuhlmann, S., Zimmermann, A., Lappe, M., Zwitserlood, P., & Dobel, C. (in prep.). *Biological motion perception in prosopagnosia.*

Lavie, N., & Fox, E. (2000). The role of perceptual load in negative priming. *Journal of Experimental Psychology: Human Perception and Performance, 26(3),* 1038-1052.

Levelt, W. J. M. (1989). *Speaking: From intention to articulation.* Cambridge (MA), London: MIT Press.

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.

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.

Mannan, S., Ruddock, K. H., & Wooding, D. S. (1995). Automatic control of saccadic eye movements made in visual inspection of briefly presented 2-D images. *Spatial Vision, 9,* 363-386.

Marois, R., Yi, D.-J., & Chun, M. M. (2004). The Neural Fate of Consciously Perceived and Missed Events in the Attentional Blink. *Neuron*, *41(3),* 465-472.

Meyer, A. S., & Dobel, C. (2004). Application of eye tracking in speech production research. In: J. Hyöna, J.R. Radach & H. Deubel (Eds.): *The Mind's Eye: Cognitive and Applied Aspects of Eye Movement Research* (pp. 253-272). Oxford: Elsevier Science.

Morgan, J., & Meyer, A. (2005). Processing of extrafoveal objects during multiple-object naming. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31(3),* 428-442.

Morrison, D. J., & Schyns, P. G. (2001). Usage of spatial scales for the categorization of faces, objects, and scenes. *Psychonomic Bulletin and Review, 8*(3), 454-469.

Neisser, U. (1967). *Cognitive psychology*. NY: Appleton-Century-Crofts.

Nelson, W. W., & Loftus, G. R. (1980). The functional visual field during picture viewing. *Journal of Experimental Psychology: Human Learning and Memory, 6,* 391-399.

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, CA: 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.

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision, 42(3),* 145-175.

Oliva, A., & Torralba, A. (2006). Chapter 2 Building the gist of a scene: the role of global image features in recognition. *Progress in Brain Research, 155 B*, 23-36.

O'Regan, J. K., & Noe, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences, 24(5),* 939-973.

Parkhurst, D., Law, K., & Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision Research, 42,* 107-123.

Peterson, M. A., & Rhodes, G. (2003). Analytic and Holistic Processing—The View Through Different Lenses. In M. A. Peterson & G. Rhodes (Eds.) *Perception of Faces, Objects, and Scenes: Analytic and Holistic Processes.* OUP, USA.

Pollatsek, A., Rayner, K., & Collins, W. E. (1984). Integrating pictorial information across eye movements. *Journal of Experimental Psychology: General, 113,* 426-442.

Posner, M. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology, 32(2)*, 3-25.

Potter, M. C. (1975). Meaning in visual search. *Science 187*, 965-966.

Potter, M. C. (1976). Short-term conceptual memory for pictures, *Journal of Experimental Psychology: Human Learning and Memory, 2,* 509-522.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124(3),* 372-422.

Rensink, R. (2002). Change detection. *Annual Review of Psychology, 53,* 245-277.

Righart, R., & de Gelder, B. (2007). Impaired face and body perception in developmental prosopagnosia. Retrieved December 10, 2007 from http://www.pnas.org/cgi/reprint/0707753104v1.pdf

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.

Rousselet, G. A., Joubert, O. R., & Fabre-Thorpe, M. (2005). How long to get to the "gist" of real-world natural scenes? *Visual Cognition, 12(6),* 852-877.

Sanocki, T. & Epstein, W. (1997). Priming spatial layout of scenes. *Psychological Science, 8,* 374-378.

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.

Segalowitz, N. S. (1982). The perception of semantic relations in pictures. *Memory & Cognition, 10,* 381-388.

Shepherd, M., Findlay, J., & Hockey, R. (1986). The relationship between eye-movements and spatial attention. *Quarterly Journal of Experimental Psychology Human Experimental Psychology, 38(3),* 475-491.

Simons, D. J. & Rensink, R. A. (2005). Change blindness: past, present, and future. *Trends in Cognitive Sciences, 9,* 16–20.

Slobin, D. I. (1996). From "thought and language" to "thinking for speaking". In J. Gumperz & S. C. Levinson (Eds.): Rethinking linguistic relativity. *Studies in the social and cultural foundations of language* (pp. 70-96). New York: Cambridge University Press.

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature, 381(6582),* 520-522.

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.

Torralba, A., Oliva, A., Castelhano, 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.

Treisman, A. (2006). How the deployment of attention determines what we see. *Visual Cognition, 14(4-8),* 411-443.

Treisman, A., & Gelade, G. (1980). A feature integration theory of attention. *Cognitive Psychology, 12,* 97-136.

Triesch, J., Ballard, D. H., Hayhoe, M. M., & Sullivan, B. T. (2003). What you see is what you need. *Journal of Vision, 3(1)*, 86-94.

Ullman, S. (1984). Visual routines. *Cognition, 18(1-3),* 97-159.

Ullman, S., Vidal-Naquet, M., & Sali, E. (2002). Visual features of intermediate complexity and their use in classification. *Nature Neuroscience, 5,* 682–687.

Underwood, G., & Foulsham, T. (2006). Visual saliency and semantic incongruency influence eye movements when inspecting pictures. *Quarterly Journal of Experimental Psychology, 59(11),* 1931-1949.

Van der Meulen, F. F. (2001). *Moving eyes and naming objects.* Nijmegen, NL: MPI Series in Psycholinguistics 17.

VanRullen R., & Thorpe, S. (2001). Is it a bird? Is it a plane? Ultra-rapid visual categorisation of natural and artifactual objects. *Perception, 30,* 655-668.

Weith, M., Castelhano, M. S., & Henderson, J. M. (2003). I see what you see: Gaze perception during scene viewing. *Poster presented at the Annual Meeting of the Vision Sciences Society,* Sarasota, Florida.

Winer, B. J., Brown, D. R., & Michels, K. M. (1991). *Statistical Principles in Experimental Design.* McGraw-Hill, New York..

Wolfe, J. (1998). Visual Search. In H. Pashler (Ed.): *Attention* (pp. 13-73). East Sussex, GB: Psychology Press.

Yantis, S. (2001). *Visual Perception: Essential Readings.* Philadelphia: Psychology Press.

Yarbus, A. L. (1967). *Eye movements and vision.* New York: Plenum Press.

Zusammenfassung

Die visuelle Wahrnehmung des Menschen ist ein faszinierendes Feld. Wir können ein bekanntes Gesicht unter Millionen von anderen Gesichtern wiedererkennen. Unser visuelles System ist in der Lage, sich so unterschiedlichen Lichtverhältnissen anzupassen, wie sie uns begegnen, wenn wir auf einem Gletscher skilaufen oder unseren Weg durch die Dunkelheit suchen. Subjektiv sehen wir ein kontinuierliches, vollständiges und detailliertes Bild unserer Umgebung, obwohl durch den ständigen Wechsel von Fixationen und schnellen, ruckartigen Augenbewegungen (Sakkaden) tatsächlich nur Einzelbilder auf die Netzhaut projiziert werden, von denen zudem jeweils nur eine sehr kleine zentrale Region scharf gesehen wird.

Allerdings ist dies nur das ‚Hochleistungsende' vom weiten Spektrum der visuellen Wahrnehmung. Am anderen Ende finden wir so interessante Phänomene wie *change/inattentional blindness*: wir sind für markante Veränderungen in unserem Blickfeld ‚blind', wenn wir unsere Aufmerksamkeit nicht auf diese Region richten, oder wenn eine Veränderung völlig unerwartet eintritt. Diese Phänomene weisen darauf hin, dass es sich bei der visuellen Wahrnehmung nicht um einen passiven, sondern um einen aktiven und dynamischen Prozess handelt, der in bedeutendem Maße von Faktoren wie momentaner Aufmerksamkeit, Wissen und Erwartung des Betrachters abhängt.

Ein Gegenstand der visuellen Wahrnehmungsforschung, der speziell für die Experimente der vorliegenden Dissertation relevant ist, sind die inneren Abbilder, von denen man annimmt, dass sie die externe visuelle Welt repräsentieren, die sogenannten *internen visuellen Repräsentationen*.

Die hier geschilderten Experimente beschäftigen sich mit der Frage, wie schnell unser visuell-kognitives System interne Repräsentationen von Handlungsszenen aufbauen kann. Während eine Reihe von Studien zeigt, dass Präsentationszeiten von wenigen Zehntel Sekunden ausreichen, um Objekte zu kategorisieren (z. B. Tier/kein Tier) oder die Kategorie einer Umgebungsszene (z. B. Küche, Biergarten, Berglandschaft) zu bestimmen, ist die schnelle Extraktion von Bildinformation aus Handlungsszenen noch kaum erforscht.

Bei den in dieser Arbeit verwendeten Handlungsszenen handelt es sich um Fotos von jeweils zwei Akteuren, die in eine gemeinsame (sinnvolle oder sinnlose) Handlung involviert sind. Dabei ist jeweils ein Akteur der *Agent*, der die Handlung ausführt, der andere ist der *Patient*, mit dem etwas gemacht wird (z. B. A fotografiert P).

Die Untersuchungen im Kapitel *Event Conceptualization in Free View and at an Eyeblink* gingen der Frage nach, ob die Darbietung einer Handlungsszene für 150 ms im extra-fovealen Sehfeld ausreicht, um die thematischen Rollen (Agent/Patient) der beiden Akteure sowie die dargestellte Handlung zu identifizieren. Die Ergebnisse zeigen, dass die thematischen Rollen nahezu immer richtig zugeordnet (93 %), und selbst die Handlung in fast 60 % aller Fälle richtig benannt werden konnte. Die abschließende Untersuchung weist darauf hin, dass das globale räumliche Szenenlayout, das durch die Körperhaltungen der beiden Akteure und ihre Stellung zueinander entsteht, für die Identifikation der Handlung eine wesentliche Rolle spielte. Daraus lässt sich schließen, dass frühe visuelle Repräsentationen von Agent-Patient Handlungsszenen, die sich außerhalb des fovealen Blickfeldes befinden, anhand des groben räumlichen Layouts aufgebaut werden. Für eine detaillierte Bildanalyse sind allerdings offene Aufmerksamkeitswechsel in Form von Fixationen in handlungsrelevante Bildregionen notwendig.

Das Experiment *Rapid Apprehension of Coherence in Action Scenes* untersuchte, welche von zwei verschiedenen Bildinformations-Typen bei der Kategorisierung einer Handlung bezüglich ihrer Kohärenz (sinnvoll/sinnlos) schneller verarbeitet werden kann. Die Darbietungszeiten betrugen zwischen 20 und 100 ms, und Kohärenz

wurde auf zwei verschiedene Arten variiert. Die eine Variation veränderte das globale räumliche Layout der Szene, indem die beiden Akteure (Agent und Patient) einzeln gespiegelt wurden, wodurch sich vier verschiedene (zwei sinnvolle, zwei sinnlose) Kombinationen der Körperorientierungen ergaben. Die zweite Variation veränderte lokale Bildinformation, indem das handlungsrelevante Objekt durch ein nicht zur Handlung passendes anderes Objekt ersetzt wurde (z.B. *A serviert B eine Tasse Kaffee* vs. *A serviert B einen Schuh*). Auf diese Weise wurde die semantische Kohärenz von Handlung und Objekt variiert.

Es stellte sich heraus, dass die maskierte Darbietung von nur 30 – 50 ms ausreichend war, um die Szenen bezüglich ihrer Sinnhaftigkeit zu kategorisieren. Allerdings galt dies nur für die Manipulation des globalen räumlichen Layouts. Erwartungsgemäß konnte Agent-Objekt-Kohärenz nur mit Darbietungszeiten von mindestens 100 ms korrekt beurteilt werden. Wie in den oben genannten Untersuchungen zur Benennung einer sehr kurz präsentierten Handlung, erwies sich auch bei der Kategorisierung in sinnvolle und nicht sinnvolle Handlungen das globale räumliche Layout als wichtigste Bildinformation.

Die vorliegenden Untersuchungen sowie die Arbeit von Dobel et al. (2007) zeigen, dass unser visuelles System in der Lage ist, innerhalb weniger Zehntel bzw. Hundertstel Sekunden interne Repräsentationen von Agent-Patient-Handlungsszenen aufzubauen. Diese Repräsentationen ermöglichen dem Betrachter, essentielle semantische Szeneninhalte (Sinnhaftigkeit, thematische Rollen, Handlung) zu identifizieren. Fokussierte offene Aufmerksamkeit in Form von Fixationen auf Details der Szene ist dafür nicht notwendig.

Des Weiteren lassen die Ergebnisse auf die Hypothese schließen, dass das globale räumliche Layout der Handlungsszene, ähnlich wie bei Umgebungsszenen, bei dem schnellen Aufbau interner visueller Repräsentationen eine wesentliche Rolle spielt.

Diese Ergebnisse sind nicht nur für die Erforschung der visuellen Wahrnehmung per se interessant. Handlungsszenen werden auch als Stimulusmaterial in der Untersuchung anderer kognitiver Funktionen verwendet, im Speziellen für die

Schnittstelle zwischen dem visuellen und dem sprachlichen System. Sowohl für die Interpretation von Augenbewegungen, die zeitlich parallel zur Sprachproduktion oder Sprachwahrnehmung ausgeführt werden, als auch für die Frage, wie seriell oder parallel visuelle Wahrnehmung und linguistische Formulierung ablaufen, ist es von großer Bedeutung zu wissen, welche Information einer Szene bereits entnommen werden kann, bevor Bilddetails mit den Augen fixiert werden.

Curriculum Vitae

Reinhild Glanemann
born 1970 in Hamm/Westfalen

since 2005	PhD student (major subject: Psychology, minor subject: Neurology), University of Münster
since 2004	Research Assistant in the DFG-project 'Cognitive and linguistic event representation in language comprehension and production', Group of Prof. Dr. P. Zwitserlood, Department of Psychology, University of Münster
2004 -2005	Speech and Language Therapist (part-time), Practice for SLT Spenthof, Münster
2002 - 2003	Master of Science in 'Human Communication Sciences', University of Newcastle upon Tyne, England, with a Scholarship of the Carl-Duisberg Gesellschaft e.V.
1994 - 2002	Speech and Language Therapist, Practice for ENT, Phoniatrics & Audiology Lübben, Koch, Doleschal & Bremken, Münster
1990 - 1993	Study of Speech and Language Therapy, University of Tübingen
1989 - 1990	Voluntary Social Year, Rehabilitation Centre (Neurology), Bad Heilbrunn, Oberbayern
1980 - 1989	Anne-Frank-Gymnasium Werne a. d. Lippe

Danksagung

Mein allergrößtes ‚Danke' geht an Pienie Zwitserlood und Christian Dobel, die Ihr mich während der Promotionszeit einzigartig betreut habt. Irgendwann las ich ein einer Zeitungskolumne über zwei verschiedene Führungsstile: *Führen mit Druck* und *Führen mit Sog*. Beide habe ich schon kennen gelernt, und letzterer ist mir mit Abstand der liebste... Ihr beiden beherrscht ihn perfekt: Bei Eurer eigenen Arbeit enthusiastisch, kreativ und mit Freude dabei, immer ein offenes Ohr für Fragen und Diskussionen, so habt Ihr mich mitgerissen. Darüber hinaus habt Ihr uns Promis in die Forscherwelt im In- und Ausland so eingeführt, wie man es sich als wissenschaftlicher Nachwuchs nur träumen lassen kann. Doch hier hört der Dank gar nicht auf. Zu Eurem Sog gehören unter anderem auch die klasse Partys, die wir bei Euch gefeiert haben: zu Weihnachten, zu Halloween, zum Semesteranfang, zum Semesterabschluss, zur Verabschiedung von .... ich glaube wir haben keine Gelegenheit ausgelassen. Was hab ich mit Euch beiden für ein Glück gehabt!

Genauso ein Glück hatte ich mit meiner einzigartigen Arbeitsgruppe, neben Pienie und Christian bestehend aus Jens Bölte, Gerrit Hirschfeld, Heidrun Bien und meinen drei Mit-Promis Andrea Böhl, Annett Jorschick und Heidi Gumnior. Anfangs war mir neu, dass man bei und neben der Arbeit soviel Spaß miteinander haben kann, aber ich habe mich sehr schnell daran gewöhnt!

Letzteren, meinen Mit-Promis Andrea, Annett und Heidi gilt ein besonderer Dank für die schöne Promi-Zeit, die wir uns mit leckeren Kochabenden, Sing-Star-Wettbewerben, Power-Hüpfen, einem Meck-Pomm-Ausflug und viel Humor angenehm gemacht haben. Pienie, dank je wel dafür, dass Du damals in Maastricht nicht auf Manolo's Tauschangebot eingegangen bist, und wir weiter als die ‚The

Silliest Group of Science' in Münster bleiben durften (wer will schon auf Teneriffa leben?).

Vielen Dank an Nadine Kloth, Stefanie Enriquez-Geppert, Malte Viehbahn und Kerstin Funnemann für Euren unermüdlichen Einsatz bei der Durchführung und Diskussion meiner Experimente.

Danke auch an Uschi Husemann und Jutta Linke für das supernette und reibungslose ‚Back-Stage-Management'.

Liebe Helene Kreysa, Dir danke ich für die nette und fruchtbare Zusammenarbeit und Diskussionen nicht nur über Eye-Tracking.

Ein ganz spezieller Dank an meine Eltern und Geschwister!