RECOGNIZING THE SETTING BEFORE REPORTING THE ACTION: INVESTIGATING HOW VISUAL EVENTS ARE MENTALLY CONSTRUCTED FROM SCENE IMAGES

by

ADAM M. LARSON

B.S., Iowa State University, 2006
M.S., Kansas State University, 2010

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Psychology
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2012

Abstract

While watching a film, the viewer begins to construct mental representations of it, which are called events. During the opening scene of a film, the viewer is presented with two distinct pieces of information that can be used to construct the event, namely the setting and an action by the main character. But, which of these two constructs are first cognitively represented by the viewer? Experiment 1 examined the time-course of basic level action categorization with superordinate and basic level scene categorization using masking. The results indicated that categorization occurred in a course-to-fine manner, inconsistent with Rosch et al.'s (1976) basic level theory. Interestingly, basic level action categorization performance did not reach ceiling when it was processed for a 367 ms SOA, suggesting that additional scene information and processing time were required. Thus, Experiment 2 examined scene and action categorization performance over multiple fixations, and the scene information that was fixated for each categorization task. Both superordinate and basic level scene categorization required only a single fixation to reach ceiling performance, inconsistent with basic level primacy, whereas basic level action categorization took two to three fixations, and led to more object fixations than in either scene categorization task. Eye movements showed evidence of a person bias across all three categorization tasks. Additionally, the categorization task did produce differences in the scene information that was fixated (Yarbus, 1967). However, could basic level theory still be correct when subjects are given a different task? When the same scene images were named, basic level action terms were used more often than basic level scene category terms, while superordinate level action terms were used relatively less often, and superordinate level scene category terms were hardly ever used. This shows that linguistic categorization (naming) is sensitive to informative, middle-level categories, whereas early perceptual categorization makes

use of coarse high level distinctions. Additionally, the early perceptual advantage for scene categorization over basic level action categorization suggests that the scene category is the first construct that is used to represent events in scene images, and maybe even events in visual narratives like film.

RECOGNIZING THE SETTING BEFORE REPORTING THE ACTION: INVESTIGATING HOW VISUAL EVENTS ARE MENTALLY CONSTRUCTED FROM SCENE IMAGES

by

ADAM M. LARSON

B.S., Iowa State University, 2006
M.S., Kansas State University, 2010

A DISSERTATION

submitted in partial fulfillment of the requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Psychology
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2012

Approved by:

Major Professor
Lester C. Loschky

Copyright

Abstract

While watching a film, the viewer begins to construct mental representations of it, which are called events. During the opening scene of a film, the viewer is presented with two distinct pieces of information that can be used to construct the event, namely the setting and an action by the main character. But, which of these two constructs are first cognitively represented by the viewer? Experiment 1 examined the time-course of basic level action categorization with superordinate and basic level scene categorization using masking. The results indicated that categorization occurred in a course-to-fine manner, inconsistent with Rosch et al.'s (1976) basic level theory. Interestingly, basic level action categorization performance did not reach ceiling when it was processed for a 367 ms SOA, suggesting that additional scene information and processing time were required. Thus, Experiment 2 examined scene and action categorization performance over multiple fixations, and the scene information that was fixated for each categorization task. Both superordinate and basic level scene categorization required only a single fixation to reach ceiling performance, inconsistent with basic level primacy, whereas basic level action categorization took two to three fixations, and led to more object fixations than in either scene categorization task. Eye movements showed evidence of a person bias across all three categorization tasks. Additionally, the categorization task did produce differences in the scene information that was fixated (Yarbus, 1967). However, could basic level theory still be correct when subjects are given a different task? When the same scene images were named, basic level action terms were used more often than basic level scene category terms, while superordinate level action terms were used relatively less often, and superordinate level scene category terms were hardly ever used. This shows that linguistic categorization (naming) is sensitive to informative, middle-level categories, whereas early perceptual categorization makes

use of coarse high level distinctions. Additionally, the early perceptual advantage for scene categorization over basic level action categorization suggests that the scene category is the first construct that is used to represent events in scene images, and maybe even events in visual narratives like film.

Table of Contents

List of Figures

List of Tables

Acknowledgements

I would like to give a very special thanks to my advisor, Dr. Lester Loschky, for mentoring me to be the critical scientist that I am today, in addition to his continued friendship during my graduate education and beyond. I'd also like to thank Bernardo de la Garza, Tyler Freeman, Ryan Ringer, and Trey Hill for all of their discussions on this present dissertation project. I'd also like to thank the research assistants from the Visual Cognition lab, namely Lori Flippo, Conor O'Dea, Karen Akao, Caitlyn Badke, Margarita McQuade, Josh Hendry, John Zeurcher, and Kathryn Williams for helping me collect data. Many thanks go to Kevin Dean for helping me implement the study on Mechanical Turk, and a very special thanks goes to Conor O'Dea and Karen Akao for helping to code all of the Mechanical Turk data.

Finally, I'd like to thank my dissertation committee, Lester Loschky, Joseph Magliano, Richard Harris, Gary Brase, David Gustafson, and Shannon Washburn, for their critical insights on this body of work which will be heavily influencing my research for years to come.

Chapter 1 - Recognizing the Setting Before Reporting the Action: Investigating How Visual Events Are Mentally Constructed from Scene Images

The opening shot of the film, *The Dark Knight* shows a cityscape (Nolan, Thomas, Roven, & Nolan, 2008). A storyboard of the opening sequence is presented in Figure 1.1. The camera approaches one of the buildings, then suddenly one of the windows shatters. The second shot cuts to a man wearing a clown mask, holding a gun, next to the recently shattered window. The camera is positioned behind the armed man, looking over his shoulder, out the window onto the city. The armed man inserts a grappling hook into his weapon and fires it through the hole in the window. A jump cut presents the third shot. It is the back of a man, holding a clown mask and a duffle bag, standing on a street corner. The camera zooms in on the menacing clown mask. A van enters the frame, and the menacing looking man gets in the van.

The first three shots of the film begin by presenting two specific pieces of information to the viewer, namely the setting and action. These two constructs must be used by the viewer in order to comprehend this sequence of shots. From the beginning of the first shot, the viewer is introduced to the setting, the cityscape. The end of the first shot introduces an action, the shattering of the window. The second shot introduces the agent, namely a masked man holding a gun who is responsible for shattering the glass window. Therefore, in these two shots, the viewer has begun to comprehend the situation in a specific temporal order. First the viewer cognitively represents the setting followed by the action.

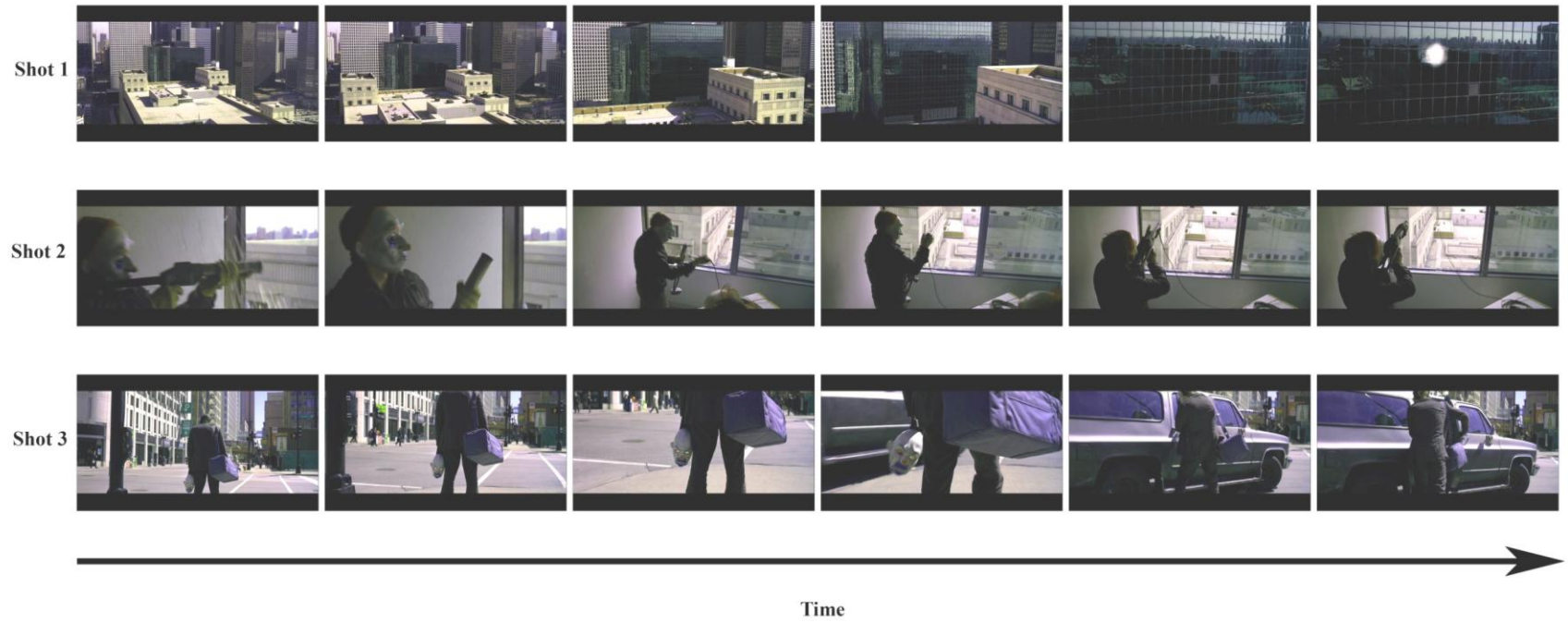

Figure 1.1 Story board of the first three shots from the film The Dark Knight (Nolan, Thomas, Roven, & Nolan, 2008).

The beginning of the third shot appears to be unrelated to the previous two shots, since a brand new character, action, and setting are introduced. Therefore, the viewer may perceive the first two shots as a separate event compared to what is currently viewed, which is presenting different character and location information (Zacks, Tversky, & Iyer, 2001). An event is defined as a goal-related activity that has a beginning and ending (Kurby & Zacks, 2007; Zacks, Speer, Swallow, Braver, & Reynolds, 2007; Zacks, & Tversky, 2001). As the viewer begins to comprehend the new event at the start of the third shot, both the setting and action information are available simultaneously. Therefore, the viewer is allowed to comprehend the setting and action for the current event without the same temporal restrictions which were placed on the viewer during the first two shots. This raises the question, how does the viewer begin to cognitively represent the event? Does the viewer begin by representing the setting first, implied by the first two shots of the film, or the character's action?

Events in Visual Narratives: Is the Scene Category or Action Represented First?

The identification of events has been investigated for many different tasks, including reading (Speer, Zacks, & Reynolds, 2007; Zacks, Speer, Reynolds, 2009), viewing pictorial narratives (Gernsbacher, 1985; Gernsbacher, Varner, & Faust, 1990), and watching short films (Zacks, Speer, Swallow, & Maley, 2010; Zacks, Tversky, Iyer, 2001). According to current theory, important psychological processes occur during these perceived events. Namely, at the start of an event, the viewer will begin to construct a working memory representation of the narrative, called an event model (Kurby & Zacks, 2007; Zacks, Speer, Swallow, Braver, & Reynolds, 2007). According to event segmentation theory (EST), perceptual information from the visual narrative is used to construct the event model (Reynolds, Zacks, & Braver, 2007;

Zacks, Kurby, Eisenberg, & Haroutunian, 2011). Once constructed, a gating mechanism prevents variability in the perceptual environment from directly affecting the event model's representation. Such a mechanism allows for the event model to remain stable for a given situation, even when perceptual information is occluded, or when perceptual encoding is inhibited due to saccadic suppression (Ross, Morrone, Goldberg, & Burr, 2001; Volkmann, 1986). Perceptual predictions are made based on the newly constructed event model. If the event model's predictions are upheld, then the same event is assumed to still be occurring since the mental representation is consistent with the incoming perceptual information in the observed narrative. Conversely, if the model's predictions are falsified, then a new event must be identified, because the current narrative is inconsistent with the mental representation of the narrative. As a result a new event model begins to be constructed. Therefore in Figure 1, the jump cut to the third shot, where a new character and a new location is introduced is likely to result in the identification of a new event and the construction of a new event model.

Many studies have examined the factors used to identify a new event. Some of these include perceptual features like changes in motion, based on research on film narratives (Smith, 2012; Zacks, 2004; Zacks, Kumar, Abrams, & Mehta, 2009). Others include breaks in coherence in conceptual/semantic features, based on research from written narratives. The event indexing model proposes a set of semantic dimensions that are likely to result in a new model for the narrative, which include changes in the character, intentionality, causality, time, and place (Magliano & Zacks, 2011; Zacks, Speer, & Reynolds, 2009; Zwaan, Langston, & Graesser, 1995; Zwaan & Radvansky, 1998). The critical factors that are important for the current study are the spatio-temporal setting and the action of the character, with the latter being related to the dimensions of character and intentionality in the event indexing model (Zwaan, Langston,

Graesser, 1995; Zwaan & Radvansky, 1998). Story grammar theories propose that narratives are organized such that the setting is introduced first (Mandler & Johnson, 1977; Thorndyke, 1977), since it establishes the place and time for subsequent actions to take place in. Based on this idea, it would be reasonable to hypothesize that in real-time perception, viewers categorize the setting prior to categorizing the action. Conversely, theories of narrative comprehension argue that narratives follow the actions and goals of a protagonist; thus attending to the actions of the character is absolutely critical to understanding the story (Mandler & Johnson, 1977; Rinck & Weber, 2003; Scott-Rich & Taylor, 2001; Thorndyke, 1977). Based on this attentional bias to the character and their actions, it would be reasonable to hypothesize that in real-time perception, viewers would categorize the action prior to the setting.

Although the event indexing model was developed specifically to account for written narratives, the same factors have been shown to be important for comprehending visual narratives as well (Zacks, Speer, & Reynolds, 2009). For instance, event-related potential (ERP) studies have shown that the brain compares the semantic information that is currently being viewed with what was previously viewed in order to detect semantic incongruities, as evidenced by N300 and N400 ERPs (described below)(McPherson & Holcomb, 1999; Stinikova, Kuperberg, & Holcomb, 2003; West & Holcomb, 2002). This detection of semantic incongruity also occurs when the action and setting changes in a picture story (West & Holcomb, 2002). In that study, a series of picture stories was presented, and the last image of the story was either semantically congruent or incongruent with the story. ERPs were compared between those who viewed a congruent or incongruent final scene, and the incongruent image ERP produced a greater negativity at 300 ms (N300) and 400 ms (N400). The N300 was suggested to indicate semantic incongruity for strictly pictorial stimuli (McPherson & Holcomb, 1999), whereas the

N400 identifies semantic incongruity for both text and pictorial information (McPherson & Holcomb, 1999; Stinikova, Kuperberg, & Holcomb, 2003; West & Holcomb, 2003). Therefore, this suggests that changes in scene category and action information create difficulties for viewers in integrating such information with the previously viewed story narrative. Assumedly, this is the sort of failure to fulfill predictions that event segmentation theory argues is used to signal the need for a new event model.

Behavioral research on scene perception has shown that both the setting and action can be identified within the first eye fixation on the scene, referred to as scene gist. Scene gist is typically operationalized as a person's ability to categorize a briefly presented image with a short description or a one-word label (Bacon-Macé, Macé, Fabre-Thorpe, & Thorpe, 2005; Larson & Loschky, 2009; Loschky & Larson, 2010; Potter, 1976). This could include a description like "dog with a frisbee" (Fei-Fei et al., 2007) or simply labeling a scene as a "beach" or "forest" (Larson & Loschky, 2009; Loschky & Larson, 2010; Rousselet, Joubert, & Fabre-Thorpe, 2005), which is more typical in scene gist research. When a scene is processed for as little as 100 ms, using a 100 ms masking stimulus onset asynchrony (SOA), people's performance for categorizing the scene is generally at ceiling (Bacon-Mace, Mace, Fabre-Thorpe, Thorpe, 2005; Biederman, Rabinowitz, Glass, & Stacy, 1974; Loschky, Sethi, Simons, Pydimarri, Ochs, & Corbeille, 2007; Loschky & Larson, 2010; Potter, 1976; Rousselet, Joubert, Fabre-Thorpe, 2005). However, scenes may be categorized at different levels of a scene taxonomy. Namely, a general category label can be applied to a scene, called a superordinate category (e.g., Indoors), or a more specific label may be applied to the same image, called the basic level (e.g., Kitchen). Classic categorization studies have shown evidence for the basic level category being represented prior to the superordinate for scenes and objects (Rosch et al., 1976; Tversky &

Hemenway, 1983), while recent work has shown the opposite (Green & Oliva, 2009; Joubert, Rousselet, Fize, & Fabre-Thorpe, 2007; Loschky & Larson, 2010; Rogers & Patterson, 2007). The time-course of superordinate and basic level scene categorization was examined in the current studies, and extended by comparing it to the time course of action categorization.

Actions are also processed during this same time frame. For instance, people and animals can be detected when a scene image is briefly presented for 20-26 ms (Mace, Joubert, Nespoulous, & Fabre-Thorpe, 2009; Thorpe, Fize, & Marlot, 1996). Indeed, animals can be detected just as easily when either one or two scene images are presented simultaneously (Rousselet, Fabre-Thorpe, & Thorpe, 2002) and when they are presented in the far periphery (Thorpe, Gegenfurtner, Fabre-Thorpe, & Bultoff, 2001). More to the point, actions can be named quite well after masking a brief peripheral presentation of the scene. Actions were named well for line drawings after a 300 ms masked presentation (Dobel, et al., 2007) and 150 ms masked presentation for real scene images (Glanemann, 2007). Likewise, eye movement research suggests that the first saccade in a scene is likely to be directed to the person (Fletcher-Watson, Findlay, Leekam, & Benson, 2008; Humphrey & Underwood, 2010; Zwickel & Võ, 2010). Thus, these studies suggest that the action may be categorized as quickly as the scene category, which raises the question of which of these constructs is recognized first?

The first construct that is recognized should also be the first construct the viewer uses to construct an event model regarding the visual narrative. For example, in the opening shot from the film, The Dark Knight, only the setting is present for the viewer to encode. The viewer is not presented with a character and action until the start of the second shot. Thus, these two sequential shots present the setting and action in a specified temporal order, such that the viewer constructs the event model from the setting first and then the actions of the character.

Conversely, the beginning of the third shot presents both a different setting and character which, presumably, is interpreted as a new event by the viewer; thus a new event model is needed. The third shot presents the setting and character's action simultaneously. In this case, which construct does the viewer recognize first in order to begin constructing the event model?

Experiments 1 and 2 examined the time course of image categorization within a single fixation and across multiple fixations, respectively. In both experiments, scene images were categorized at either the superordinate scene category level ("Indoor" vs. "Outdoor" scenes), the basic level scene category ("Park," "Kitchen," "Yard," and "Office"), or the basic level action category ("Eating" versus "Reading" in a Park). Comparing the two scene categorization levels addresses a debate within the categorization literature, specifically which categorical level is represented first. Rosch et al.'s (1976) basic level theory hypothesizes that the basic level should be categorized prior to the superordinate level. For instance, the superordinate level is defined as a general category label, whereas the basic level allows for finer category distinctions to be differentiated. Thus, in Figure 1, the third shot presents a scene, which could be labeled first as a "street" at the basic level and then "Outdoors" at the superordinate level. Conversely, recent evidence has suggested that scenes are first represented at a coarse level, which then becomes more detailed over time, with the superordinate scene category being represented before the basic level (Joubert, Rousselet, Fize, & Fabre-Thorpe, 2007; Loschky & Larson, 2010; Macé, Joubert, Nespoulous, & Fabre-Thorpe, 2009; Rogers & Patterson, 2007). These two levels of scene categorization are then compared to basic level action categorization, which may plausibly be hypothesized to be represented first due to a strong bias to immediately attend to people within scenes (Fletcher-Watson, Findlay, Leekam, & Benson, 2008; Humphrey & Underwood, 2010). Experiment 2 examined the fixation locations for these three different categorization

tasks. Previous research has shown that a viewer's task affects the information that is fixated in a scene (DeAngulus & Pelz, 2009; Yarbus, 1967). Thus, based on this research, the categorization task should produce differences in the scene information that is fixated. However, an alternative hypothesis is that eye movements are biased to fixate the person in a scene. All of the scene images contained a person performing an action; thus the eye movements in the three categorization tasks should be the same, due to fixations falling on the person across all of the tasks.

Experiment 3 then examined the descriptions that were used to label these same scene images in order to determine if image categorization differs between early perceptual categorization and later linguistic categorization (naming) processes. Previous studies have shown that early perceptual categorization of scenes occurs at the superordinate level before the basic level. Thus, it is hypothesized that scene images may be described according to the initial category that is recognized, namely the superordinate level scene category. Conversely, later linguistic categorization processes may be biased to label objects and scenes according to finer-level category distinctions. Thus, an alternative hypothesis predicts that scenes will be described according to the basic level image descriptors rather than the superordinate level terms. Furthermore, if image descriptions focus on the basic level, then are the descriptions based on the action or the scene category? Thus, a final alternative hypothesis is that basic level actions would be used more to describe scene images, based on research that shows that the first eye movement in a scene is biased to fixate people. The results from all three experiments are then discussed in terms of their contribution to research on scene perception, and extended to implications for research on event perception and comprehension of visual narratives.

Chapter 2 - Experiment 1

The classic study supporting the idea of basic level superiority by Rosch et al. (1976) showed that it was easier for people to name the basic level category of an object compared to naming its superordinate or subordinate category. For example, indicating an object was a "dog" was shown to be faster than describing it as an "animal" (superordinate level) or a "German Shepherd" (subordinate level). The basic level hypothesis has been supported by numerous studies (e.g., Gosselin & Schyns, 2001; Jolicour, Gluck, & Kosslyn, 1984; Murphy & Smith, 1982). Rosch's theory reasoned that the basic level was easier to access since the features shared at the basic level were not likely to be shared by other object categories, thus it is inclusive. However, the basic level must also be cognitively economical, since it would be impractical to have a different name for every single discrete object. Therefore, the basic level has properties of being both general and inclusive. For example, the features of dogs allow them to be recognized as a separate category than birds, which have very different features. Therefore dog is inclusive of a specific type of animal. However, the category dog is also a general term, since it includes all types of dogs, from Golden Retrievers to Chihuahuas. However, superordinate categories (e.g., animal), would be harder to recognize because objects contained in a superordinate category share few perceptual features, and thus are more difficult to recognize than basic level categories. Rosch et al. (1976) showed that recognizing objects at the basic level produced faster reaction times than at the superordinate level. Thus, according to the theory, object categorizing occurs first at the basic level, and then the object's superordinate category is inferred from its basic level category (e.g., we know that Fido is an "animal" by virtue of the fact that he is a "dog" and the fact that all "dogs" are "animals").

The basic level advantage was also found with environmental scenes (Tversky & Hemenway, 1983). Participants were able to list significantly more scene parts (perceptual features) and actions (behaviors consistent with the environmental scene) when categorizing basic level scenes (e.g., Park or Home) compared to superordinate level scenes (e.g., Outdoors and Indoors). However, there was no increase in identifying parts and actions when categorizing scenes at the subordinate level (e.g., City Park or Apartment). Additionally, when participants were asked to name different environmental scenes, basic level scene category names were more frequent than superordinate or subordinate scene categories. On the basis of this study and Rosch et al.'s theory, then, one can hypothesize that basic level scene categories should be recognized more accurately and faster than the superordinate level scene categories; however, based on Rosch et al.'s theory, it is unclear how basic level action categories and scene categories relate to each other in terms of a processing hierarchy.

Importantly, however, recent research has begun to challenge the basic level advantage hypothesis. Oliva and Torralba (2001) theorized that scene categorization occurs in a global-to-local fashion based on a computational model they proposed that first identifies global scene properties and then makes more finer scene category distinctions. Scenes would first be represented by only a very gross categorical distinction (Natural/Man-made), and over time, finer category distinctions would be made (e.g., open field vs. beach)(Greene & Oliva, 2009). The same theorized order of categorization was proposed by Rogers and Patterson's (2007) Parallel Distributed Processing theory. Rogers and Patterson (2007) showed that superordinate object representations were recognized prior to the basic level when participants were forced to make a speeded response. However, when participants were allowed to respond at their own pace, then the basic level was better recognized than the superordinate object category. Further behavioral

evidence for such coarse-to-fine categorization has been shown when participants rapidly categorized objects contained in scenes. Specifically, categorizing scenes as containing an animal versus non-animal (superordinate categorization) was faster than categorizing the animal as a bird or a dog (basic level categorization)(Macé, Joubert, Nespoulous, & Fabre-Thorpe, 2009; Poncet, Reddy, & Fabre-Thorpe, 2012). These experiments briefly presented scene images and required participants to categorize the object, similar to the procedures used in studying the speed at which scene gist is recognized. Finally, a recent study compared scene categorization at the superordinate versus basic level (Loschky & Larson, 2010). The results showed that superordinate level gist was categorized more quickly than the basic level. For example, when given a 12 ms masking SOA to process the category of a scene image at the superordinate level (Natural or Man-made), sensitivity reached .75, as measured by A', a non-parametric signal detection measure of sensitivity. However, in order to reach the same level of sensitivity when categorizing scene gist at the basic level (e.g., Street or Highway), about 24 ms SOA was required (Loschky & Larson, 2010). These studies show that rapid scene categorization begins at a coarse level, and further processing allows for finer distinctions to be made. For example, an object is recognized as an "animal" prior to being recognized as a "dog," or a scene is recognized as being "outdoors" before it is recognized as being a "street" scene. These studies would suggest the hypothesis that a scene's superordinate category should be recognized more accurately at an earlier processing time than either the basic level or action/event level.

A third competing hypothesis can be proposed based on recent eye-movement research, which shows that the eyes are biased to attend to people in scene images (Fletcher-Watson, Findlay, Leekam, & Benson, 2008; Humphrey & Underwood, 2010). For instance, Fletcher-

Watson, Findlay, Leekam, & Benson (2008) presented two images simultaneously on a computer screen, to the left and right of fixation. One of the scenes contained a person, while the other scene did not. On half of the trials, subjects were told simply to look at the pictures, while on the other half of the trials, subjects were told to identify the gender of the person in the image. When participants were given the free viewing task, their first saccade went to the image with the person 67.0% of the time. The authors also created a measure of bias that took into account both the probability of fixating a region and the relative size of that region, and showed that the first fixation on the person-present scene was more biased to target the person's face than the scene background. According to the Sequential Attention model, this indicates that prior to the initiation of the saccade, covert attention moved to the spatial location of the person, which was followed by an eye movement to the same location (Henderson, 1992). Thus, these results show that there is a strong bias to process people in an image during the very first fixation on a scene, even when the people appear in the viewer's visual periphery.

There are also studies suggesting that a simple event can be recognized within a single fixation, which is similar to the processes and time-course used to recognize scene gist. Dobel, Gumnior, Bölte, and Zwitserlood (2007) presented masked action scenes in the visual periphery for durations varying from 100-300 ms. During the experiment, the computer screen was divided into four equally sized quadrants, and a scene was presented in one of them, while the subject fixated the center of the screen—thus scenes were always presented in parafoveal or peripheral vision and the subject did not know where to expect them. When given the task of identifying the location of the person performing the action (i.e., the agent), performance reached asymptote (74% correct) when a scene was presented for 200 ms, indicating the semantic relationship between characters in the scene had been categorized. However, when given the

task of naming the action, performance was worse, varying from 19 % correct when scenes were presented for 100 ms, to 46 % correct when scenes were presented for 300 ms (Dobel, Gumnior, Bölte, & Zwitserlood, 2007). Interestingly, only two different actions were used in the study (Shooting vs. Giving), suggesting that basic level action identification in the visual periphery is quite poor. However, Glanemann (2007) briefly presented real-life action scenes in the periphery and examined whether assigning agent/patient roles in a scene was the result of identifying the action. If so, then performance on these two tasks should be similar. Scenes were constructed around twelve different actions and the scenes were presented for 150 ms. The results showed that patient was correctly detected 93% of the time, however this was not due to correctly identifying the action, since basic level action naming only reached 58% correct when the image was presented for 150 ms. Thus, both studies showed that basic level action naming does occur in briefly presented scenes in the visual periphery, though it appears that agent/patient categorization is even faster than action naming.

An interesting question is the degree to which the action-naming results of Dobel et al (2007) and Glanemann (2007), which were rather poor at short scene durations, bear on predictions of basic level action recognition. In fact, there are several reasons why action recognition performance may be better than action naming. First, recognition tasks are less cognitively taxing than recall tasks, and therefore action recognition performance should be better than action naming. Second, the scenes in Glanemann (2007) and Dobel et al., (2007) contained no background scene information, and thus cannot speak to comparisons between action or event recognition and scene gist recognition. Furthermore, previous studies have shown that person and object recognition improves when they are contained in a congruent scene background compared to an incongruent scene background (Davenport & Potter, 2004; Palmer,

1975). Therefore, the results for action recognition reported in these action naming studies may be a lower bound on performance, and action recognition in natural scene contexts should be even better than reported by Dobel et al. (2007) and Glanemann (2007). In sum, since attention is biased towards people in scenes during the first fixation, and actions are recognized within a single fixation, these results suggest the hypothesis that action categorization may occur prior to either basic level or superordinate level scene gist recognition.

The current experiment examined which level of image categorization is processed first during briefly presented scenes. Specifically, is the scene category, at the superordinate or basic level, or the action category, at the basic level, recognized first in a scene? Previous research has suggested three competing hypotheses regarding which level of categorization is processed earliest. Specifically, recent studies have found a coarse-to-fine processing order for scene categorization, which suggests that the superordinate level scene categorization will be processed earlier than either basic level scene categorization or basic level action categorization. However, numerous studies over the last 30 years have found evidence of a basic-level advantage for objects as well as scene categorization, which predicts that basic level recognition should be processed earliest. Finally, eye movement studies have found a bias to attend to people in scenes, and other studies have shown that basic level actions are processed very rapidly, which suggests that the basic level action should be recognized first.

Method

Participants

A total of 59 undergraduate students (23 Females) enrolled in General Psychology participated in the study. The mean age of the participants was 20.53 (SD = 4.88) years old and

all had at least 20/30 visual acuity. Participants signed an informed consent sheet prior to participating in the study. All participants received course credit for their participation.

Materials

Images were collected using on-line image databases (Flicker, Google Images, etc.) or by taking pictures with a digital camera. All images contained a person performing one of two basic level actions in a specific scene location that could be categorized at either the superordinate level (indoor vs. outdoor) or basic level (kitchen, office, yard, or park). Basic level actions in Kitchen scenes were either Cooking or Washing Dishes, and in Office scenes were either Phoning or Typing; basic level actions in Yard scenes were either Raking or Mowing, and in Parks were either Reading or Eating. Example scene images are presented in Figure 1.2. There were 34 images per basic level action category and each had a resolution of 1024 x 768 pixels. All of the experimental images were converted to monochrome and their mean luminance and contrast were equalized across the entire image set. Texture masks were generated from all 240 target scene images, based on the texture generation algorithm of Portilla and Simoncelli (2000). Previous studies have used texture masks and have found them to be strong masks due to sharing 2nd order and higher order image properties with scene images (Loschky, Hansen, Sethi, & Pydimarri, 2010). Scene images were presented on a Samsung SyncMaster 957 MBS monitors (17 inch) (85 Hz refresh rate) and were viewed at a distance of 53.34 cm using a chin rest, from which they subtended 37.78° x 29.03° of visual angle.

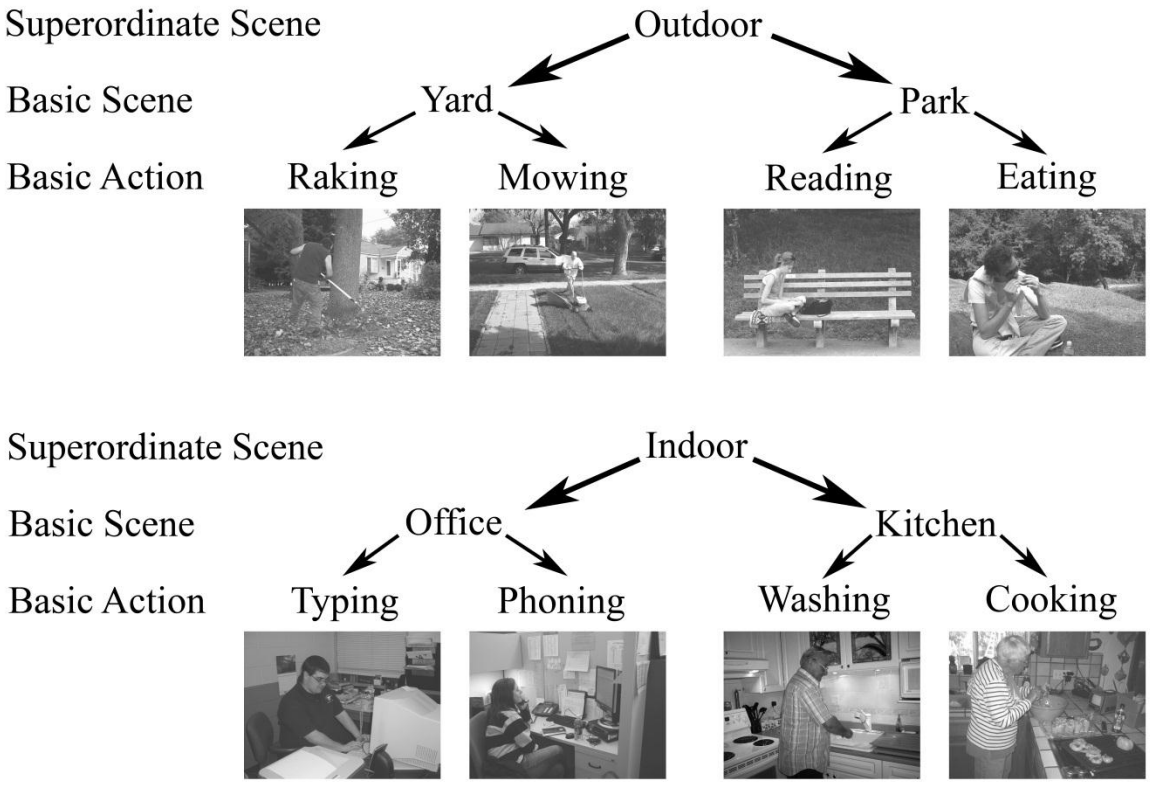

Figure 2.1 Example scene images for each basic level action category, basic level scene category, and superordinate level scene category.

Two images per basic level action category were used to familiarize participants with the image categories on which they would be tested and another two images per basic level action category were used for practice trials. These images were not used in the actual experiment.

Procedures and Design

The experiment used a 3 (Categorization level [Basic level action, Basic level scene, or Superordinate level scene]) x 5 (Stimulus Onset Asynchrony: SOA [24, 60, 106, 200, and 365 ms]) mixed design, with categorization level as a between-subjects factor, and SOA as a within-subjects' factor. Participants were randomly assigned to the between-subjects factor of the categorization level. Processing time for each scene image was equally distributed between the five masking SOAs, which is a measure of processing time for the image.

As in Loschky and Larson (2010, Exp. 2) image categories were blocked throughout the experiment, so that in each block only two image categories were tested. The pair of image categories was always from the same parent scene category. In this way, the next higher categorization level could not be used to provide information about the category being tested. For example, "Washing" and "Cooking" were always in the same block since both occurred within the same "Kitchen" basic level parent scene category. Thus, knowing that the scene was a "Kitchen" would not help in discriminating whether the person in the scene was "Washing" or "Cooking." Similarly, participants categorizing images according to their basic level scene gist categories always viewed "Kitchen" and "Office" scenes within the same block, since they came from the same "Indoor" superordinate level parent scene category. The superordinate level scene categorization task only had one block, since there were only two scene categories, "Indoor" and "Outdoor." Thus, although the same set of 240 target scene images was used for all categorization tasks, there were four separate blocks in the basic level action categorization task, two blocks in the basic level scene categorization task, and one block for the superordinate level scene categorization task. The five different SOAs were presented equally often and randomly within each block.

Before beginning the experimental trials, participants were familiarized with the scene categories by seeing sixteen images, in which the name of the image category was presented and the participant pressed the "NEXT" button to view the images. Each image was presented for two seconds. After familiarization, participants were given 32 practice trials, which shared the same design parameters as the actual experiment.

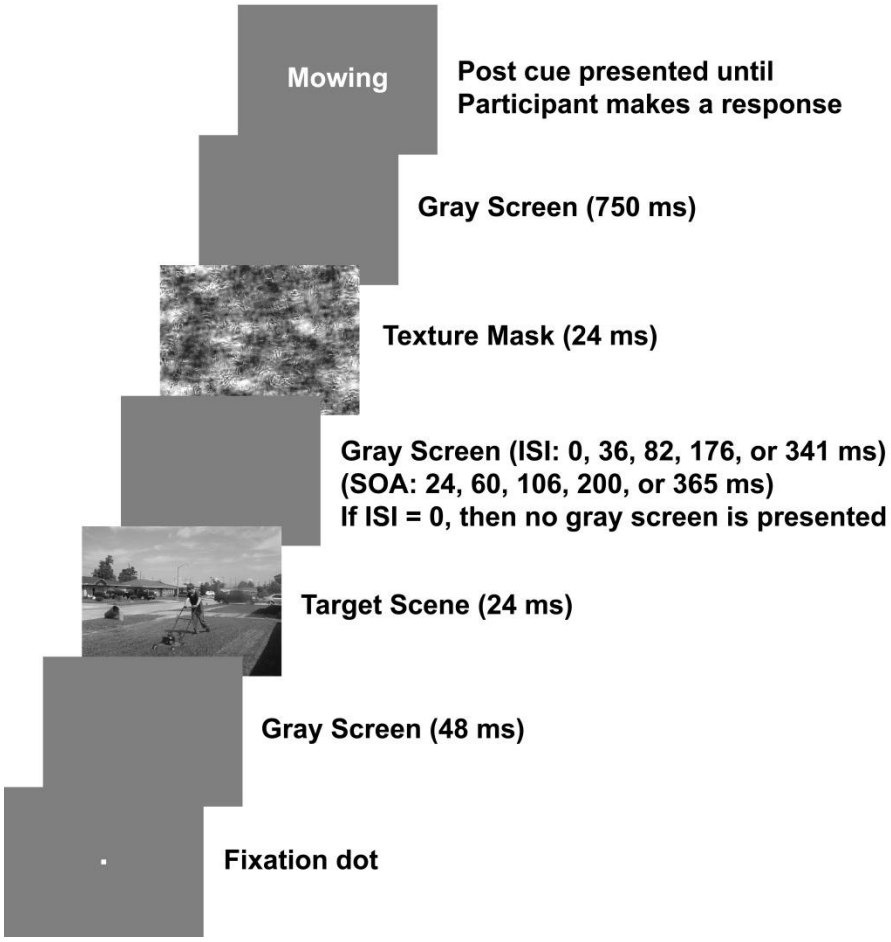

Figure 2.2 Trial schematic for the tachistoscopic presentation of scene images in Experiment one. The Inter-Stimulus Interval (ISI) is the duration between the onset of the grey screen and the onset of the mask. The Stimulus Onset Asynchrony (SOA) is the target duration plus the ISI.

In the experiment, a total of 240 trials were presented. Figure 2.2 shows the events that occurred within a single trial. An instruction screen indicated the pair of scene categories that were presented in each block. Participants pressed the "Next" button, and a white fixation dot was presented on a neutral gray background. When participants were ready for the trial to begin, they pressed the "Next" button again. A target image was presented for 24 ms. If the SOA was

greater than 24 ms, then a grey screen was presented (for ISIs of 36, 82, 176, and 341 ms). If the SOA for the trial was 24 ms, then the mask was presented immediately after the target (ISI of 0 ms). A texture mask image was randomly paired with the target scene image and was presented for 24 ms, followed by a grey screen for 750 ms. A post cue was presented on the screen until participants made a response. Validly cued trials would present a scene category that matched the post cue. If participants believed the post cue matched the scene category presented, then participants were instructed to press the "Yes" button; otherwise, if the cue word was invalid (i.e., the cue did not match the scene category viewed), they were to press the "No" button. Half of the trials within each block were randomly selected to be validly cued, while the other half was invalidly cued.

Results

Two mixed factorial ANOVAs with a 3 (Categorization task [Superordinate level scene category, basic level scene category, or basic level action category]: Between-subjects) x 5 (SOA [24, 60, 106, 200, and 365 ms]: Within-subjects) design were used to analyze sensitivity and response bias, as measured by signal detection theory's d' and c (Macmillan & Creelman, 2005)[1]. Figure 2.3 shows the sensitivity and response bias for the three categorization conditions as a function of processing time (SOA). Tables 2.1 and 2.2 present the descriptive statistics for sensitivity and bias by the categorization task for each level of processing time. As can be seen in Figure 2.3, as expected, categorization sensitivity increased monotonically with increasing processing time (F (4, 224) = 40.17, $p < .001$), and appears to reach asymptote at 200 ms SOA for all three tasks. There was no difference between 200 ms and 365 ms SOA for any of the three categorization tasks (ts (37) $\leq$ 1.39, ps $\geq$ .18). Of greater interest, the scene image

[1] Sensitivity (d') is calculated as Z(Hit) - Z(False Alarm). Bias (c) is calculated as -0.5[Z(Hit) + Z(False Alarm)].

categorization task level had a large effect on sensitivity (F (2, 56) = 45.71, p < .001). Sensitivity was greater for the superordinate Indoor/Outdoor scene category distinction than the basic level categorization task (t (37) = 4.50, p < .001), and greater for the basic level scene categorization task than the basic level action categorization task (t (37) = 10.09, p < .001). There was no interaction between the categorization tasks and processing time on categorization sensitivity (F (8, 224) = 1.22, p = .29).

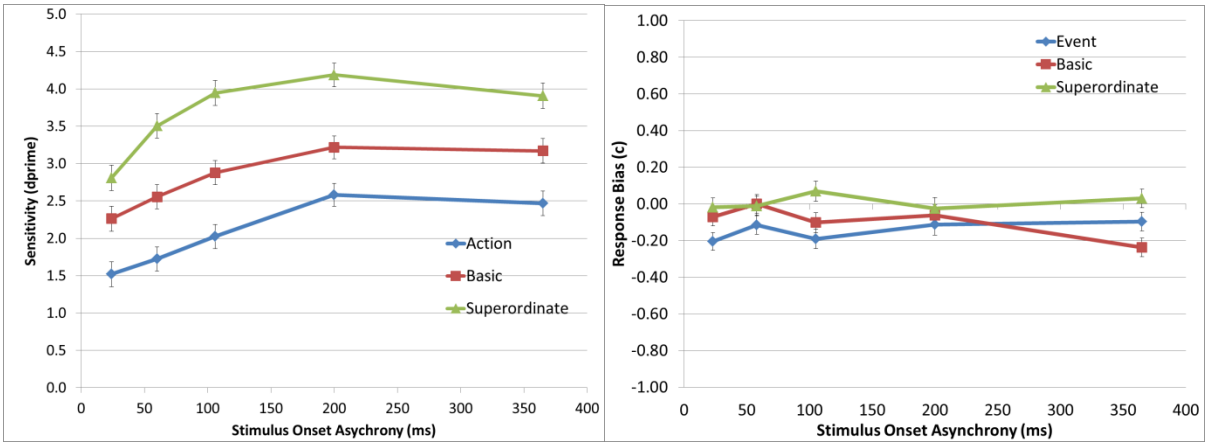

Figure 2.3 Sensitivity (d prime) and bias (c) for the three categorization tasks as a function of processing time as measured by SOA. Error bars represent the standard error.

Table 2.1 Sensitivity (d') descriptive statistics for categorizing events and scene gist, at the basic and superordinate level, at each SOA.

	Stimulus Onset Asynchrony (ms)									
	24 ms		58 ms		108 ms		200 ms		365 ms	
Categorization task	M	SD	M	SD	M	SD	M	SD	M	SD
Action	1.52	0.45	1.72	0.46	2.02	0.68	2.58	0.62	2.47	0.65
Basic	2.26	0.79	2.55	0.75	2.88	0.75	3.21	0.78	3.17	0.68
Superordinate	2.81	0.92	3.51	0.89	3.95	0.72	4.19	0.66	3.91	0.88

Table 2.2 Bias (c) descriptive statistics for categorizing events and scene gist, at the basic and superordinate level, at each SOA.

	Stimulus Onset Asynchrony (ms)									
	24 ms		58 ms		108 ms		200 ms		365 ms	
Categorization task	M	SD	M	SD	M	SD	M	SD	M	SD
Action	-0.22	0.20	-0.12	0.24	-0.23	0.29	-0.12	0.27	-0.09	0.27
Basic	-0.05	0.26	-0.01	0.25	-0.11	0.21	-0.03	0.31	-0.26	0.23
Superordinate	-0.02	0.23	0.00	0.20	0.07	0.21	0.02	0.18	0.00	0.15

There was no effect of image processing time on response bias (F (4, 224) = 1.19, p < .318). As shown in Figure 2.3, participants tended to be unbiased. However, there was a small but statistically significant liberal bias for basic level action categorization compared to both scene category recognition tasks (F (2, 56) = 9.15, p < .001). Scene processing time and categorization task did interact with respect to response bias (F (8, 224) = 2.31, p = .021), which was driven by the increase in "Yes" responses in the basic level scene categorization task at 365 ms SOA. This is shown by the fact that the categorization task x SOA interaction became non-significant (F (6, 168) = 0.78, p = .58) when the 365 ms SOA condition was removed from the analysis.

Discussion

Experiment 1 examined the time course of categorizing basic level actions and scene categories at the superordinate or basic level within a single fixation. The data shows that the superordinate scene categorization task resulted in greater sensitivity than both the basic level scene categorization task and the basic level action categorization task. Interestingly, this

advantage was seen at every processing time, including the processing time (367 ms) that exceeds the average fixation duration (generally 330 ms)(Rayner, 1998).

Previous studies examining eye-movements in scenes found that if there is a person in a scene, then the first saccade in an image is strongly biased to go to it (Fletcher-Watson, Findlay, Leekam, & Benson, 2008; Humphrey & Underwood, 2010; Zwickel & Võ, 2010). This shows that during the first fixation on a scene, attention is biased to select information about people prior to the execution of a saccade to that spatial location (Henderson, 1992). Thus, if we are biased to attend to people in scenes, we might expect that basic level action categorization should be processed earlier than either the basic level or superordinate level scene categorization tasks. However, the current data are inconsistent with this person bias hypothesis. Likewise, the basic level advantage, first proposed by Rosch et al., (1976), specifies that the basic level should be categorized first, for objects as well as for scene categories (Tversky & Hemenway, 1983). The basic level theory predicts that the superordinate category is inferred from the basic level category after that is determined. The data are also inconsistent with the predictions based on the basic level theory.

Recent research has begun to question the assumption of basic level superiority, by showing that early visual processes are better able to distinguish between superordinate level categories before distinguishing between basic level categories (Large, Kiss, & McMullen, 2004; Loschky & Larson, 2010; Macé, Joubert, Nespoulous, & Fabre-Thorpe, 2009; Rogers & Patterson, 2007). For instance, contrary to the basic level theory, a picture would be categorized as an "animal" prior to being categorized as a "dog," or a picture of a scene would be categorized as "Man-made" before the same image was categorized as a "Street." The data from the current experiment supports the coarse-to-fine hypothesis that scene images are categorized first at the

superordinate level scene category prior to the basic level scene category, which is also categorized prior to the basic level action. As the viewer is given more processing time with a scene image, they are able to make finer scene category distinctions (Large, Kiss, & McMullen, 2004; Loschky & Larson, 2010; Macé, Joubert, Nespoulous, & Fabre-Thorpe, 2009; Rogers & Patterson, 2007). According to the present data, the scenes were categorized at the superordinate level quite well with an SOA of 23 ms (d' = 2.5). For the basic level scene categorization task, that same level of sensitivity was not reached until 80 ms of processing, and it was not reached in the basic level action categorization task until around 200 ms of processing. Therefore, this shows that coarse category distinctions require less processing time than distinguishing finer scene categories. It is expected that this superordinate scene category advantage would also be present for other superordinate scene categories, such as distinguishing between "Natural" versus "Man-made" scene images. It seems likely that the superordinate advantage, found for scene categorization, would also be found for action categorization, though that is an empirical question.

As noted earlier, other recent studies have also found an advantage for processing the superordinate level prior to the basic level (Rogers & Patterson, 2007; Large, Kiss, & McMullen, 2004; Macé, Joubert, Nespoulous, & Fabre-Thorpe, 2009; more). Rogers and Patterson (2007) showed better recognition of superordinate object categories than the basic level, when participants were forced to make a speeded response. However, when participants were not forced to make a speeded response, the basic level category was superior to superordinate object categorization. Rogers and Patterson (2007) explained their results in terms of a parallel distributed processing network in which conceptual representations are searched for in a

semantic space and the search enters the superordinate level prior to the basic level.[2] A similar coarse-to-fine pattern of categorization has been theorized for scene gist recognition (Oliva & Torralba, 2001) and supported by empirical research on human scene perception (Greene & Oliva, 2009; Mace, Joubert, Nespoulous, & Fabre-Thorpe, 2009; Loschky & Larson, 2010). Large, Kiss, and McMullen (2004) also found evidence supporting a superordinate advantage for objects. Specifically, ERPs distinguished the superordinate categories earlier than the basic level categories, and basic level categories earlier than subordinate level categories.

These findings conflict with the basic level superiority hypothesis (Gosselin & Schyns, 2001; Jolicour, Gluck, & Kosslyn, 1984, Rosch et al., 1976; Tversky & Hemenway, 1983). As shown by previous research (Gosselin & Schyns, 2001; Jolicour, Gluck, & Kosslyn, 1984, Rosch et al., 1976; Tversky & Hemenway, 1983), the basic level is more informative than categorizing an object or scene at the superordinate level. For example, Rosch et al., (1976) presented category names and asked participants to list attributes that were representative of that category. For instance, features representative of "bird" would include wings, beaks, and feathers. They showed a significant increase in the number of attributes listed for basic level category names than for superordinate names or subordinate names. Thus, the basic level category names elicit a greater number of features for an object than a superordinate category name, making the basic level more informative than the respective levels. Additionally, when asked to categorize objects, reaction times were faster for basic level categorization than the superordinate level (Rosch, 1976, Experiment 7). This was taken as evidence that the basic level was represented prior to the superordinate level. However, the objects were presented until a response was made.

[2] The speeded categorization task limits the amount of processing time in which competing object concepts (i.e., basic level and superordinate level concepts) can become activated. Therefore, in this task, categorizing an object according to its superordinate representation is more accurate because the proportion of activation for the superordinate concept is greater than the proportion of activation for the basic level concept.

Therefore, as predicted by the parallel distributed processing model, proposed by Rogers and Patterson (2007), when not forced to respond quickly, the basic level is recognized first; however if required to make a speeded response, a superordinate advantage is observed. By manipulating the amount of image processing time in study 1, via masking SOA, our task may be conceptually similar to requiring a speeded response in Rogers and Patterson's (2007) study. This appears to be the case, as shown by better superordinate categorization than basic level categorization at earlier image processing times.

Nevertheless, Rogers and Patterson (2007) would predict that all three category representations should be activated when a sufficient amount of processing time is allowed. Our results show that each representation is not equally available within the processing time of a single fixation, especially in the basic level action categorization task. This suggests that a single fixation is not a sufficient amount of processing time to categorize the basic level action. In addition, the visual information extracted regarding the person, and the object that is being manipulated, may not be sufficiently detailed to categorize the basic level action. Therefore, the person shown in the scene may need to be fixated in order to extract a sufficient amount of detailed visual information, and allow more processing time, in order to categorize the basic level action. Study 2 examined the number of fixations required in order to reach ceiling performance for basic level action categorization compared to both scene categorization tasks at the superordinate and basic levels, and the visual information needed to categorize basic level actions accurately, as shown by what viewers fixated.

Chapter 3 - Experiment 2

The human visual field can be divided into two distinct regions—central vision and peripheral vision. A small area extending out to 5° eccentricity around the point of fixation is

known as central vision, and includes both foveal vision (out to 1° eccentricity) and parafoveal vision (between 1° and 5° eccentricity). Information encoded in this region contains high-resolution visual information, like sharp edges. However, as the eccentricity in degrees of visual angle from fixation increases, visual resolution decreases (Peli & Geri, 2001; Séré, Marendaz, & Hérault, 2000; Strasburger, Rentschler, & Juttner, 2011). This large region beyond central vision is referred to as peripheral vision. Although the resolution of visual information varies over the visual field, useful information can still be extracted within a single eye fixation. For instance, the brief masked presentation of a scene for ≤ 100 ms is more than enough information to allow for accurate scene gist recognition (Bacon-Mace, 2005; Potter, 1976; Loschky et al., 2007; Beiderman, Rabinowitz, Glass, & Stacy, 1974). In the case of scene gist recognition, high resolution scene information from the central 5° of a scene can be completely removed, such that scene information is only viewed using peripheral vision, with no decrease in scene gist recognition compared to a fully intact scene image (Larson & Loschky, 2009).

Study 1 was able to show that scene gist categorization at the basic and superordinate level was good with as little as 200 ms of processing. However, even when given 365 ms of processing, which is slightly longer than the average duration for an eye fixation, the basic level action categorization task was unable to reach the same level of performance as either scene categorization task. This suggests that the scene information needed to increase basic level action categorization performance is either a) insufficient when presented in peripheral vision (assuming that the important information for basic level actions are presented in the visual periphery, specifically the person and the object they are manipulating), or b) requires more time to process, or c) a combination of both a) and b), namely, critical scene information needs to be

foveated (i.e., looked at with central vision), in addition to needing more time processing that information.

If additional foveal scene information is required to recognize a basic level action, then the question becomes what information is needed to improve it? Glanemann (2007) suggested that the body posture of a person had a large effect on participant's ability to name the action. Their analysis of action naming showed that the more stereotypical body postures (e.g., Kicking) were more easily identified than ambiguous body postures (e.g., helping). An additional source of information for identifying an action would likely be the object that is being manipulated by a person in the scene. An important question then becomes whether or not useful object information contained in the visual periphery can be encoded.

Several eye-movement studies have examined how well objects can be recognized in our visual periphery. Henderson, McClure, Pierce, and Schrock (1997) examined whether foveal vision was necessary to categorize an object. The study used a gaze-contingent Scotoma paradigm. During the experiment, participant's gaze was tracked and, in the Scotoma-centered condition, visual information that was within 1.2° eccentricity from fixation was removed from view. Thus, only objects outside the visual scotoma (i.e., in parafoveal vision or beyond) were presented. A second condition presented the Scotoma to the right of a fixation, allowing foveal information to be encoded. A control condition allowed for normal object viewing with no Scotoma. The results showed that object recognition in the Scotoma-centered condition was no different from the control condition or the off-centered condition. This study showed that foveal vision is not needed to identify objects, however, in the Scotoma centered condition, objects were only 2.4° eccentricity from fixation, well within parafoveal vision.

Studies examining priming have been able to show that objects presented in the visual periphery decreased the naming latency to the target (Henderson, Pollatsek, & Rayner, 1987; Pollatsek, Rayner, & Collins, 1984). Pollatsek, Rayner, and Collins (1984) presented objects at either 5° or 10° eccentricity, and participants were asked to make a saccade as quickly as possible to the object. While the eye was performing the saccade, the visual system is suppressing visual encoding (Chekaluk, & Llewellyn, 1990; Ross, Morrone, Goldberg, & Burr, 2001), and during that saccade, the object was either changed to a different object or was presented unchanged. The results show that when the object was changed, naming latency for the object increased by around 100 ms compared to when the object was unchanged, suggesting that the visual periphery was processing object information prior to the execution of a saccade. When the object was changed during the saccade, additional processing time, as indicated by an increase in object naming latency, was required to process the new object, compared to when the object remained unchanged. A similar procedure was used by Henderson, Pollatsek, and Rayner (1987), which also required participants to make a saccade to an object in the periphery. The study manipulated the presence of a peripheral object as well as its relationship with a foveal object. Specifically the foveal object could be related (for example a hand and a foot) or unrelated (a hand and a rodent) to the peripheral object. The results show that fixating a related object decreased the naming latency of the peripheral object. However, there was no difference in the amount of priming when the peripheral target was present or not. This suggests that priming occurs between two subsequent fixations, but priming is not the result of the presence of a related peripheral object. Thus, it seems that identifying an object during the current fixation can help to speed the identification of a related object on the next fixation. However, similar to the Henderson et al. (1997) study, a limitation of these studies is that only the line drawn objects

were presented, and they were not part of real-world scenes. Thus, since there was a lack of foveal information present to encode, attention may have been able to quickly deploy to the spatial location of the peripheral object resulting in a large priming effect. However, if the peripherally presented object were part of a real-world scene image, it is unclear if any peripheral object information would be encoded due to the presence of scene information at the fovea.

The extreme limit of scene perception in peripheral vision was tested in another study using an animal detection task (Thorpe, Gegenfurtner, Fabre-Thorpe, & Bultoff, 2001). Participants were asked to detect animals in real-world scenes presented at varying eccentricities from the center of fixation ranging from 0°-60° eccentricity. The results showed that animals could still be detected at above chance levels in the far visual periphery (60° eccentricity). These studies show that objects presented in the periphery can be encoded to some degree (Thorpe, Gegenfurtner, Fabre-Thorpe, & Bultoff, 2001).

The results from study 1 may indicate that some object information was likely encoded from parafoveal and peripheral vision during the brief presentation of the scene image, but was insufficient to recognize the basic level action. Thus, it may be necessary to allow viewers to make eye-movements within the scene. A fixation on the object that the person is manipulating would allow for high-resolution details to be encoded in order to differentiate between two different events occurring in the same basic level scene category (e.g., cooking or washing the dishes in a kitchen). Due to the relatively worse performance in the basic level action categorization task in Experiment 1, it is hypothesized that the basic level action categorization requires a fixation on the manipulated object in the scene. If so, then as one allows viewers to make from one to several (e.g., 2, 4, or 8) fixations on a scene, their performance on basic level action recognition should increase monotonically with increasing numbers of fixations.

It may be expected that different areas of a scene will be fixated when given a superordinate or basic level scene categorization task versus a basic level action categorization task. Evidence for such differences have been found in previous eye-movement studies that have shown different parts of an image were fixated when the participant's task differed (DeAngulus & Pelz, 2009; Yarbus, 1967). In the classic study by Yarbus (1967), the same scene image was presented to a participant while different questions were asked about the scene. For instance, if asked to "estimate the material circumstances of the family shown in the picture," then participants fixated the clothing of the women and the furniture in the scene. Conversely, if asked to "estimate the ages of the people shown in the picture," then participants fixated people's faces (pp. 192). A limitation of the work by Yarbus (1967) is that the conclusions were based on one participant and no quantitative data was reported. However, a recent study by DeAngelus and Pelz (2009) replicated Yarbus' (1967) study with 17 participants and reported quantitative eye-movement results. Overall, DeAngelus and Pelz (2009) confirmed most of the conclusions drawn by Yarbus. Specifically, different participants fixated similar locations for six of the seven questions asked. For instance, when asked to "estimate the financial state of the family," then faces were fixated most, followed by the furniture and clothing, which is exactly what Yarbus (1967) observed. These findings from DeAngelus and Pelz (2009) confirm the top-down task effects described by Yarbus (1967), as well as showing, quantitatively, that these eye-movement patterns are similar between participants. Therefore, due to the task requirements, eye-movements may differ between the event and scene categorization tasks. Specifically, when categorizing the scene gist, saccades may be more likely to fixate background scene information than the person or the object being manipulated by that person, whereas when recognizing the basic level action, saccades to the person and object may be more likely.

An alternative hypothesis is that eye-movements made when categorizing scene gist at the basic and superordinate levels will be very similar to those made for event categorization due to the person bias when viewers look at real-world scenes (Fletcher-Watson, Findlay, Leekam, & Benson, 2008; Humphrey & Underwood, 2010; Zwickel & Võ, 2010). Namely, people may spend most of their fixations looking at the person regardless of the task. This person bias is also observed in the results described above by DeAngelus and Pelz (2009), who state that "the faces of the people in the image were invariably fixated, regardless of the task" (p. 804). This could be especially true since scene categorization, whether at the superordinate or the basic level, can be acquired in the first fixation. All of the remaining fixations on the scene may be influenced by the presence of the person in the scene, which would result in participants fixating similar regions of an image regardless of the task that they are given.

Method

Participants

A total of 70 General Psychology students (33 female) participated in the study. The mean age of the participants was 19.93 (SD = 3.52) years old and all had normal or corrected-to-normal vision (20/30 visual acuity). Two participants were removed from the analysis, one was removed due to low sensitivity and the other was removed due to a computer error which failed to record all of their trials during the experiment. Participants signed an informed consent sheet prior to participating in the study. All participants received course credit for their participation.

Materials

Six images per event category were added to the image set used in study 1 for a total of 288 scene images used. Target images were used to create texture masks. All scene targets and masks were converted to grey scale and the mean luminance (lum = 126.84) and contrast was

equalized (RMS contrast = 0.38). Images were presented on a 17" ViewSonic Graphics Series CRT monitor (Model G90fb) with a refresh rate of 85 Hz. Images were viewed at a distance of 60.96 cm, resulting in the scene image subtending 33.67° x 25.50° of visual angle. An EyeLink 2000 (SR research) was used to track eye-movements throughout the experiment. Fixations and saccades were detected by using algorithms developed by SR research. Specifically, saccades were detected by the velocity of a saccade exceeding 30° per second and/or an acceleration of greater than 8,000° per second^2, while fixations are periods of time that are not classified as a saccade or an eye blink.

Procedures and Design

The experiment used a 3 (Categorization level [Basic level Action categorization, Basic level scene categorization, or Superordinate scene categorization]: Between-subjects) x 6 (scene processing durations [24 ms, 200 ms, or 1, 2, 4, or 8 fixations]: Within-subjects) mixed factorial design. Participants were randomly assigned to a categorization level. All participants were presented with all of the 288 scene images, which were divided equally among all scene processing durations, and counter-balanced across subjects. Scene image categories were blocked as in Experiment 1. Sixteen scene images were presented to participants in order to familiarize them with the basic level action or scene category names. Afterword, participants performed 32 practice trials prior to starting the experiment. None of the images presented during the familiarization and practice phase were included in the actual experiment. Participants were calibrated to the eye-tracker with a nine-point calibration screen. Calibration was considered acceptable if the mean spatial error was less than .5° and the maximum spatial error was less than 1.0° of visual angle.

A total of 288 trials were presented. Figure 3.1 presents a trial schematic for the events that occurred in Experiment 2. For trials involving tachistiscopic image presentations, the trials and procedures were the same as in Experiment 1.

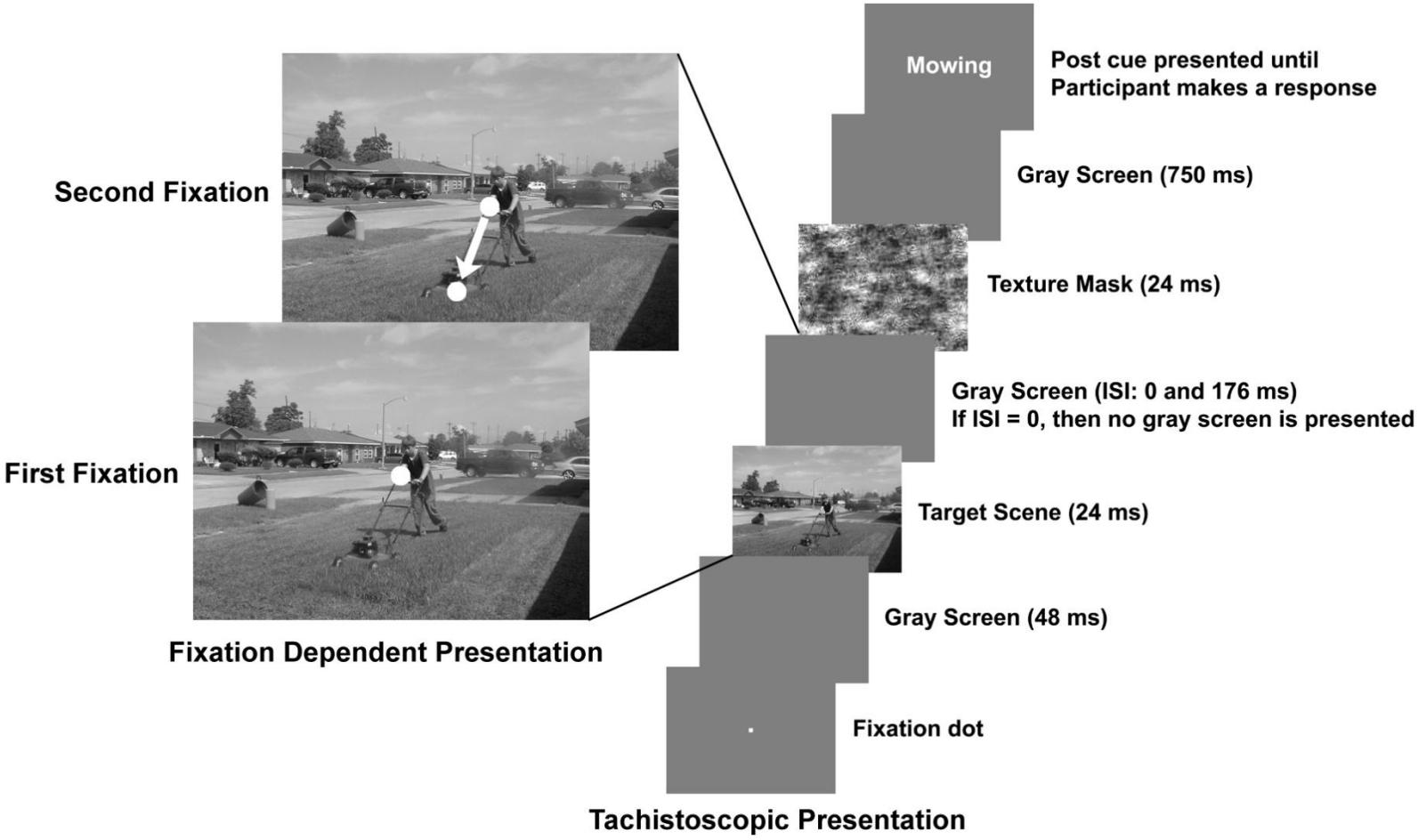

Figure 3.1 Experiment 2 trial schematic. The left side represents the events occurring during a two fixation-dependent image presentation, whereas the right side represents the events occurring during a tachistoscopic image (24 ms) presentation. Fixations are indicated by white circles and the arrow indicates a saccade.

When scenes were presented for either a single or multiple fixations, then scene images were presented until the onset of the second, third, fifth, or ninth saccade was detected (resulting in one, two, four, or eight fixations on the scene, respectively), and then followed by a 24 ms texture mask, a 750 ms grey screen, and a cue word which was presented until the participant made a response. However, due to a programming error in the fixation condition, scene images were presented for 1, 3, and 7 full fixations. The programming error was identified and corrected for a future replication study and follow-up studies.

Results

The Effect of Processing Time on Scene Image Categorization

Two Mixed Factorial ANOVAs with a 3 (Categorization task [Superordinate level scene category, basic level scene category, or basic level action]: Between-subjects) x 5 (processing [24, 200 ms SOA, 1, 3, and 7 fixations]: Within-subjects) designs were used to analyze sensitivity and bias, as measured by signal detection theory (McMillen & Crealman, 2005). Figure 3.2 shows sensitivity (d') and bias (c) for each categorization task as a function of processing time for the image. Tables 3.1 and 3.2 present the descriptive statistics for sensitivity and bias at each processing time for the three categorization tasks. The data shows a main effect for processing time (F (4, 256) = 110.91, $p < .001$), where sensitivity significantly increased from the 23 ms SOA to a single fixation (ts (66) $\geq$ 5.60, ps < .001). There was no difference in sensitivity when the image was viewed for one, three, or seven fixations (ts (66) $\leq$ 0.07, ps $\geq$ .95). There was also a main effect for the categorization task on sensitivity (F (2, 64) = 15.25, p < .001), where sensitivity was not different between the Superordinate and Basic level scene categorization tasks (t (42) = 1.89, p = .065), although both tasks produced greater performance than the basic level action categorization task (ts (43) $\geq$ 3.20, ps $\leq$ .003). There was a significant

interaction between the categorization task and processing time (F (8, 256) = 7.88, p < .001). Bonferroni-corrected t-test were used to examine the significant differences between the image categorization tasks at each processing level (critical p-value = .003). The Superordinate and Basic level scene categorization tasks were significantly better than the basic level action categorization task for the sub-fixation processing times (24 and 200 ms SOAs)(ts (43) ≥ 4.82, ps < .001). When the image was processed for a fixation, superordinate level scene categorization was better than basic level action categorization (t (43) = 4.13, p < .001), but neither was different from the basic level scene categorization task (ts (42) ≤ 1.78, ps ≥ .082). When the scene image was processed for three fixations or more, basic level action categorization sensitivity converged with the superordinate and basic level scene categorization tasks (ts (42) ≤ 1.20, ps ≥ .238).

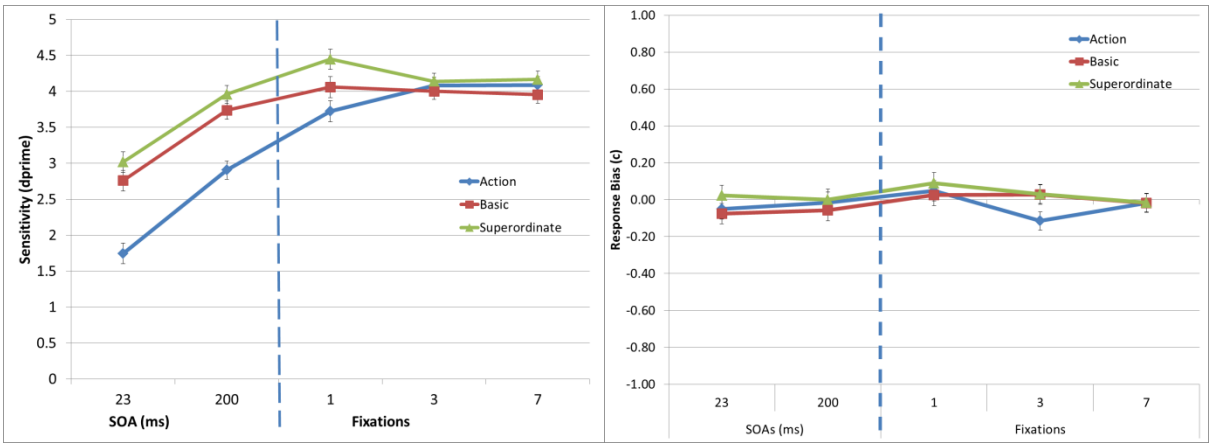

Figure 3.2 Sensitivity (d') and bias (c) performance for the three categorization tasks as a function of the processing time for the image (as measured in SOA and the number of fixations). Error bars represent standard errors.

Table 3.1 Sensitivity (d') descriptive statistics for categorizing the basic level action and scene category, at the basic and superordinate level, at each SOA.

	Image Processing (SOA and fixations)									
	24 ms		200 ms		1 fixation		3 fixations		7 fixations	
Categorization task	M	SD	M	SD	M	SD	M	SD	M	SD
Action	1.74	0.44	2.91	0.49	3.73	0.60	4.09	0.41	4.09	0.49
Basic	2.76	0.80	3.74	0.65	4.06	0.84	4.00	0.69	3.96	0.65
Superordinate	3.02	0.76	3.96	0.63	4.45	0.57	4.14	0.50	4.17	0.52

Overall, participants' mean response bias was near zero. There was no difference in response bias between the three categorization tasks (F (2, 64) = 1.24, p = .30) or between processing times (F (4, 256) = 1.35, p = .25). There was also no interaction between the categorization task and processing time on response bias (F (8, 256) = 0.64, p = .74).

Table 3.2 Bias (c) descriptive statistics for categorizing the basic level action and scene gist, at the basic and superordinate level, at each SOA.

	Image Processing (SOA and fixations)									
	24 ms		200 ms		1 fixation		3 fixations		7 fixations	
Categorization task	M	SD	M	SD	M	SD	M	SD	M	SD
Action	-0.05	0.16	-0.02	0.35	0.05	0.29	-0.11	0.25	-0.02	0.25
Basic	-0.07	0.32	-0.06	0.19	0.03	0.25	0.03	0.21	-0.02	0.27
Superordinate	0.02	0.28	0.00	0.25	0.09	0.25	0.03	0.27	-0.02	0.20

Precursors to the Eye Movement Analyses

Eye movements were analyzed to determine if top-down, task effects would influence the information that was fixated in scene images. Five interest areas were created for each scene image. These interest areas included the body, head, hands, the object that the person was manipulating, and the scene background. To determine the size and eccentricity of each interest

area, the smallest possible rectangular box needed to completely encompass a particular feature was drawn around it, thus constituting an *interest area* for that feature. The size of each interest area was determined by using the height and width of the rectangular interest area to compute the length of the hypotenuse of a right triangle for that respective interest area (i.e., the distance from the bottom-left corner to the top-right corner of each interest area). Interest area eccentricity was computed by determining the distance from the center of the image to the center of each respective interest area. This distance, in pixels, was then calculated into degrees of visual angle, and the average eccentricity of each interest area was computed. Table 2.3 presents descriptive statistics for the eccentricity and size of each of the interest areas. The grand average distance for all interest areas was 9.31°, with the smallest average distance being 8.35° (for bodies), thus all interest areas were, on average, in the visual periphery (> 5° eccentricity). Likewise, the body was the largest interest area, on average, followed by the head, which was similar in size to the object being manipulated. The smallest interest area was the person's hands. The rectangular interest areas were used only to estimate the size and eccentricity of each semantic feature but they were not used in the following eye movement analysis, which instead used more precise free-hand drawn interest areas.

Table 3.3 Descriptive statistics for the average size and eccentricity of each interest area in the scene images.

	Interest Areas							
	Body		Object		Head		Hands	
Degrees of Visual Angle	M	SD	M	SD	M	SD	M	SD
Size	19.19	6.31	7.42	4.67	7.74	3.63	3.34	1.79
Eccentricity	8.35	5.06	8.63	5.44	11.34	6.26	8.90	7.32

Another set of interest areas were created to analyze eye movements. Namely, free-hand tracings were made for each respective semantic feature. These free-hand interest areas completely encompassed only that respective semantic feature. Importantly, free-hand interest areas were mutually exclusive—an area encapsulated by one interest area did not overlap with any other interest area. The following eye movement analyses are based on the free-hand interest areas.

Attentional Selection and Encoding Processes in Eye Movements

Attention plays a large role in selecting a subsequent fixation location in a scene image (Carmi & Itti, 2006; Henderson, 1992; Rizzolatti, Riggio, Dascola, & Umilta, 1987; Torralba, Oliva, Castelhano, & Henderson, 2006). Thus, attentional selection processes in scene images can be investigated by examining the differences between the categorization tasks on 1) the percentage of fixations on different semantic scene features and 2) the amount of time that was required to first fixate that semantic information, called the *fixation latency*. If the categorization task affects the attentional selection processes for eye movements, then different semantic scene information may be more useful for certain categorization tasks (DeAngelus & Pelz, 2009; Yarbus, 1967). As a result, the percentage of fixations on each interest area should differ. Likewise, the fixation latency to these interest areas should also differ, with faster latencies to interest areas important for a given categorization task.

Additionally, encoding processes, as measured by the average fixation durations and the total amount of time spent fixating an interest area (i.e., dwell time), may be affected by the informativeness of different scene information. Namely, informative scene information may result in longer fixation durations than non-informative scene information. Thus, if the categorization task influences attentional selection processes, it may also affect the perceptual

encoding of scene information, which should produce differences in fixation durations or dwell times for different semantic scene features. These analyses are presented later in the results section.

In the present study, when participants were allowed to make more than a single fixation on the scene, we assume that attentional selection processes were used to select the location of the subsequent fixation. Attentional selection processes affected by gist processing should be strongest for the first eye movement in the scene. Therefore, the impact of the categorization task on the attentional selection processes in eye movements were assessed by examining the percentage of fixations on the five semantic features first fixated (i.e., the body, head, hands, object, or scene background) and the first fixation latency on each feature. The first eye movement was analyzed from the trials that allowed participants 3 and 7 fixations to view the scene image. The first fixation condition was eliminated from the analysis since participants were fixated at the center of the image and did not have an opportunity to execute a saccade and fixate any other area of the scene image. Later on in the results section, encoding processes were analyzed by including all of the eye movements in the data analysis except of the first fixation duration on the scene. If encoding processes differed between the three categorization tasks, the average fixation duration, or dwell time, on each interest area should differ.

Effects of the Categorization Task on the First Eye Movement

The percentage of fixations per interest area for the first eye movement on the scene was analyzed using a 3 (categorization task [basic level Action, Basic level scene, and Superordinate scene categorization task]) x 5 (Interest Area [Head, Body, Hands, Object, and Background areas]) Mixed Factorial ANOVA. Figure 32.3 and Table 32.4 present the descriptive statistics for the percentage of fixations on each interest area. There was a significant difference in the

percentage of fixations falling in each interest area (F (4, 256) = 339.94, $p < .001$). Bonferroni-corrected t-tests were used to determine the significant differences in the percentage of fixation on each interest area (critical p-value = .005). The scene background had the greatest percentage of fixations compared to all other interest areas (ts (62) $\geq$ 16.81, ps < .001). This result is unsurprising, since the scene background comprised, on average, 82.7% of the image. The body received the second greatest percentage of fixations and was significantly greater than the percentage of fixations on the head, object, and hands (ts (66) $\geq$ 4.41, ps < .001). There was no difference in the percentage of fixations between the head and object (t (66) = 1.71, p = .093), however both were greater than the percentage of fixations on the hands (ts (62) $\geq$ 6.74, ps < .001).

Of critical interest, the interest area first fixated differed depending on the categorization task being performed (F (8, 240) = 4.14, $p < .001$), thus supporting the task-based hypothesis. This is consistent with the results of Yarbus (1967) and the more recent replication by Angelus and Pelz (2009) showing that there are top-down task-based effects on where people look in scenes. Bonferroni-corrected t-tests were used to determine the significant differences between each categorization for each interest area (critical p-value = .003). The basic level action categorization condition had a greater percentage of fixations on the object than the basic level and superordinate scene categorization tasks (ts (40) $\geq$ 4.18, ps < .001). Additionally, the basic level action categorization task had a greater percentage of fixations for the hands than the basic level scene categorization task (t (40) = 3.53, p = .001). The superordinate scene categorization task had a greater percentage of fixations on the body than the basic level action categorization task (t (42) = 3.95, $p < .001$). There were no other significant comparisons.

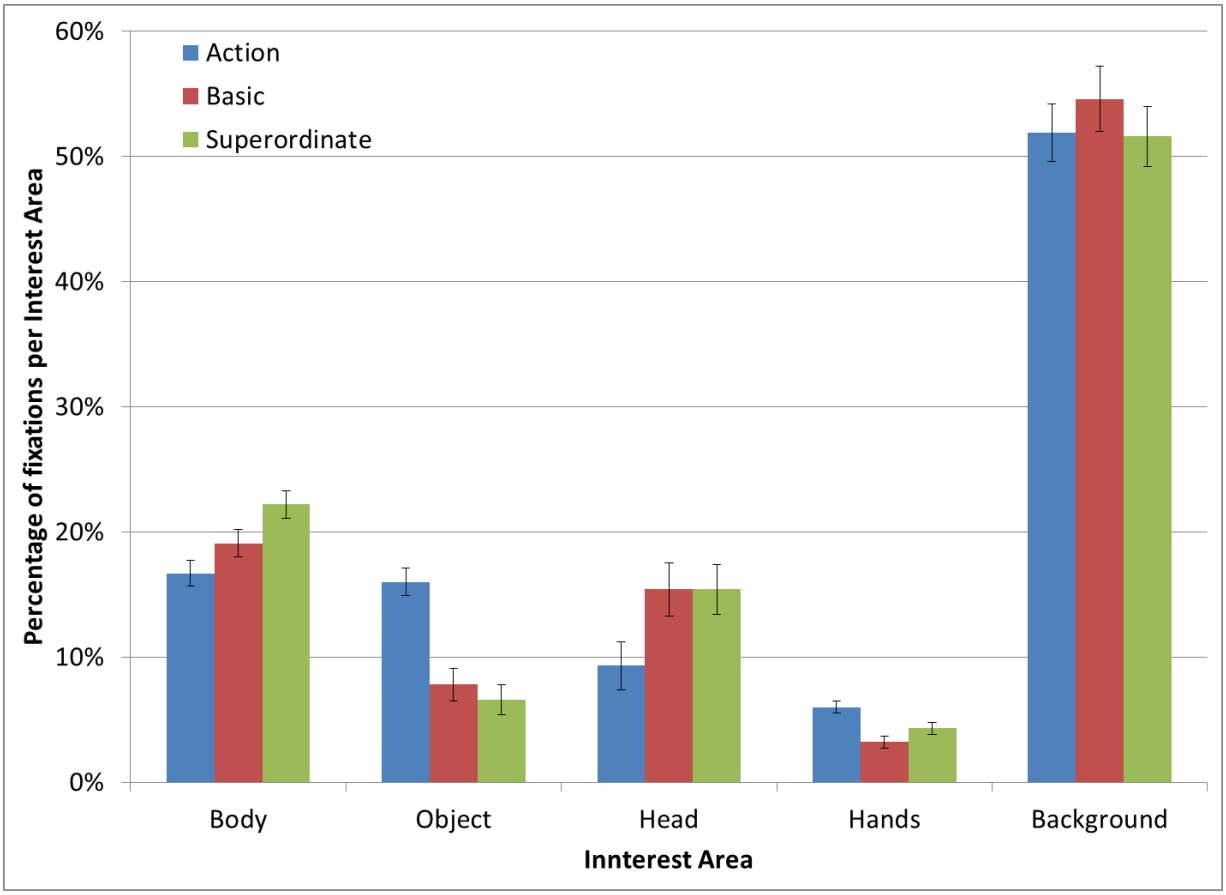

Figure 3.3 The mean percentage of fixations on each interest area between the three categorization conditions. Data represents only the first eye movement in the scene. Error bars represent the standard error.

Table 3.4 Descriptive statistics for the percentage of first fixation on each respective interest area for the Basic level Action, Basic, and Superordinate scene categorization tasks.

	Interest Areas									
	Body		Object		Head		Hands		Background	
Categorization task	M	SD	M	SD	M	SD	M	SD	M	SD
Action	16.73	4.62	16.00	7.95	9.33	6.07	6.02	3.00	51.92	9.21
Basic	19.06	5.46	7.81	3.40	15.36	9.12	3.21	1.91	54.56	12.02
Superordinate	22.17	4.49	6.55	3.36	15.42	11.52	4.28	1.91	51.58	12.31

As shown in Figure 3.3, the scene background was the most likely of the five interest areas to be fixated. However, since the scene background composed a greater percentage (*M* = 82.7%) of scene image compared to the other areas, one could predict that the background would receive a greater percentage of fixations, if only by chance. Following the same logic, the body, object, head, and hands should be less likely to be fixated by chance since these areas compose a smaller total percentage (*M* = 17.3%) of the scene image. Thus, based solely on chance, the percentage of fixations to land on each respective interest area is clearly expected to be dependent on the size of that interest area. Therefore, a follow-up analysis was conducted based on the procedures described by Fletcher-Watson et al. (2008), in order to control for the effect of relative image area on the percentage of fixations on each interest area. To do this, a new metric was calculated by dividing the percentage of fixations within each interest area by the percentage of the total image area that each interest area encompassed in the scene image. This ratio will be referred to as the domain-relative ratio. If the domain-relative ratio produced a valued of one, then the interest area was fixated as much as would be expected by chance for the size of that area. For example, if the background composed 50% of the scene, then it would be expected that, by chance, the background would be fixated 50% of the time. In this case, a domain-relative ratio of 1.0 would be expected by chance. However, a value greater than 1.0 indicates that there must be a bias to process that semantic information at a rate greater than chance.

The Fletcher-Watson analysis of domain-relative fixation ratio for the first eye movement on the scene image was analyzed using a 3 (categorization task [Basic level Action, Basic scene category, and Superordinate scene category]) x 5 (Interest Area [Head, Body, Hands, Object, and Background areas]) Mixed Factorial ANOVA. Figure 3.4 shows the domain-relative ratios per

interest area for each image categorization task. Descriptive statistics are presented in Table 3.5. There was no significant difference in the domain-relative ratios between the three categorization conditions (F (2, 60) = 1.63, p = .21). However, there was a difference in the domain-relative ratios between the five interest areas (F (4, 240) = 39.18, p < .001). Bonferroni-corrected t-tests were used to compare the domain-relative ratios between interest areas (critical p-value = .005). The domain-relative ratios for the object, hands, and head were not significantly different from each other (ts (62) ≤ 1.55, ps ≥ .13). However, these three interest areas had a greater domain-relative ratio than the body and scene background (ts (62) ≥ 5.26, ps < .001). Additionally, the domain-relative ratios for each interest area depended on the categorization task (F (8, 240) = 3.38, p = .001). Bonferroni-corrected t-tests were used to compare domain-relative ratios for interest areas within each categorization task (critical p-value = .001). For the basic level action categorization task, the domain-relative ratios were not different between the object and hands (t (22) = 0.09, p = .93). The domain-relative ratio for the object was greater for the head, body, and scene background (ts (22) ≥ 3.78, ps ≤ .001), whereas the domain-relative ratio for the hands was only greater than for the body and scene background (ts (22) ≥ 4.60, ps < .001). Finally, the domain-relative ratio for the head was greater than for the body and background (ts (22) ≥ 3.95, ps ≤ .001), and the body was greater than for the background (t (22) = 6.41, p < .001). For both the basic level and superordinate level scene categorization tasks, the domain-relative ratios for the head, object and hands are not significantly different from each other (basic: ts (18) ≤ 2.22, ps ≥ .04, superordinate: ts (20) ≤ 1.28, ps ≥ .21). The object and hands had a greater domain-relative ratio than the body and background (basic: ts (18) ≥ 3.12, ps ≤ .006, superordinate: ts (20) ≥ 3.92, ps ≤ .001). Lastly, the body had a greater domain-relative ratio than the scene background (basic: t (18) = 8.30, p < .001, superordinate: t (20) = 10.22, p < .001). The only

difference between the two scene categorization tasks was that the domain-relative ratio for the head was not different from the body in the superordinate task (t (20) = 2.41, p = .026), however the domain-relative ratio for the head was significantly larger than for the body in the basic level scene categorization task (t (18) = 4.38, p < .001). Importantly, this shows that for the first eye movement in the scene, attention selects different semantic features depending on the categorization task. Namely, for the two scene categorization tasks, the object, head, and hands received a similar percentage of fixations relative to the area they subtend. On the other hand, for basic level action categorization, attention primarily selects the object and the hands to be fixated with the first eye movement on the scene. This supports similar findings of the task affecting the selection of semantic scene features (DeAngulus & Pelz, 2009; Yarbus, 1967).

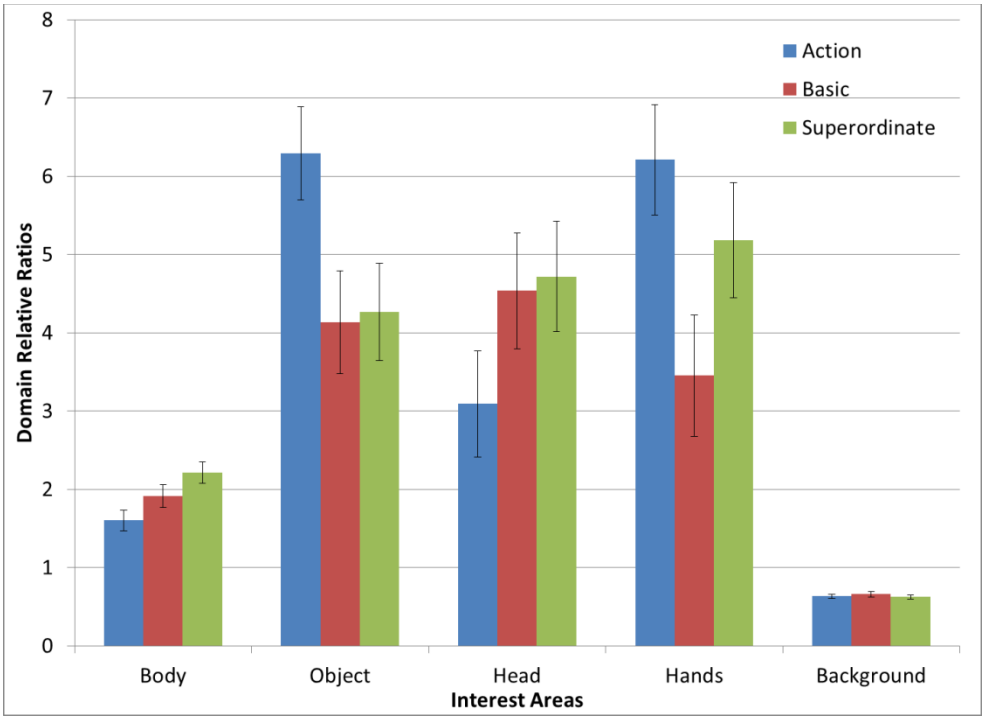

Figure 3.4 The mean domain-relative ratio for the first eye movement on each interest area between the three categorization tasks. Error bars represent the standard error.

Table 3.5 Descriptive statistics for the domain relative ratios for first fixation on each respective interest area for the Basic level Action, Basic level scene, and Superordinate scene categorization tasks.

	Interest Areas									
	Body		Object		Head		Hands		Background	
Categorization task	M	SD	M	SD	M	SD	M	SD	M	SD
Action	1.60	0.68	6.29	3.25	3.09	1.73	6.21	4.73	0.63	0.11
Basic	1.92	0.55	4.13	2.44	4.54	2.69	3.45	2.23	0.66	0.15
Superordinate	2.21	0.63	4.26	2.74	4.72	4.65	5.18	2.25	0.62	0.15

It is clear that the categorization task affected the semantic information that was first fixated on the scene. However, does the categorization task also effect the time it takes to fixate that information, namely the latency to fixate? The average latency to fixate on an interest area was analyzed using a 3 (categorization task [Basic level Action, Basic level scene, and Superordinate scene categorization task]) x 5 (Interest Area [Head, Body, Hands, Object, and Background areas]) Mixed Factorial ANOVA. Figure 3.5 presents the data and Table 3.6 presents the descriptive statistics. There was no difference in the average latency to fixate an interest area between the three categorization tasks (F (2, 64) = 0.66, p = .52). However, the average latency to fixate did differ between the five interest areas (F (4, 256) = 14.69, $p < .001$). Bonferroni-corrected t-tests were used to compare mean fixation latencies between each interest area (critical p-value = .005). The scene background was fixated prior to the head, body, object, and hands (ts (66) $\geq$ 3.43, ps $\leq$ .001). This is predictable due to the large visual area that the scene background encompassed relative to the other interest areas. Overall, the latency to fixate the head was not different from the body (t (66) = 2.22, p = .030), but it was significantly faster

than the object and hands (*t*s (66) ≥ 3.53, *p*s ≤ .001). There was no difference between the latency to fixate the object and hands (*t* (66) = 0.67, *p* = .51).

Critically, there was evidence that the average latency to fixate a particular interest area significantly depended on the categorization task that was performed (*F* (8, 256) = 2.21, *p* = .027). To explore this interaction, Bonferroni-corrected t-tests were used to compare the average latency to fixate between categorization tasks for each interest area (critical *p*-value = .001). For the basic level action categorization task, the scene background had a marginally faster latency than the body (*t* (22) = 3.51, *p* = .002) and head (*t* (22) =3.41, *p* = .003). All other comparisons between interest area latencies were not significantly different (*t*s (22) ≤ 2.66, *p*s ≥ .014). Conversely, for the basic level categorization task, the latency to fixate the scene background was significantly faster than the latency for body, object, and hands (*t*s (21) ≥ 4.18, *p*s < .001). Additionally, the latency to fixate the head was marginally faster than the latency to fixate the object (*t* (21) = 3.49, *p* = .002). This is consistent with previous research regarding the person bias (Fletcher-Watson, Findlay, Leekam, & Benson, 2008). For the superordinate categorization task, the scene background latency was significantly faster than the latency to the hands and object (*t*s (21) ≥ 3.97, *p*s ≤ .001). Additionally, the latency to fixate the object was significantly slower than to the body (*t* (21) = 4.49, *p* < .001) as well as marginally slower than to the head (*t* (21) = 3.53, *p* = .002).

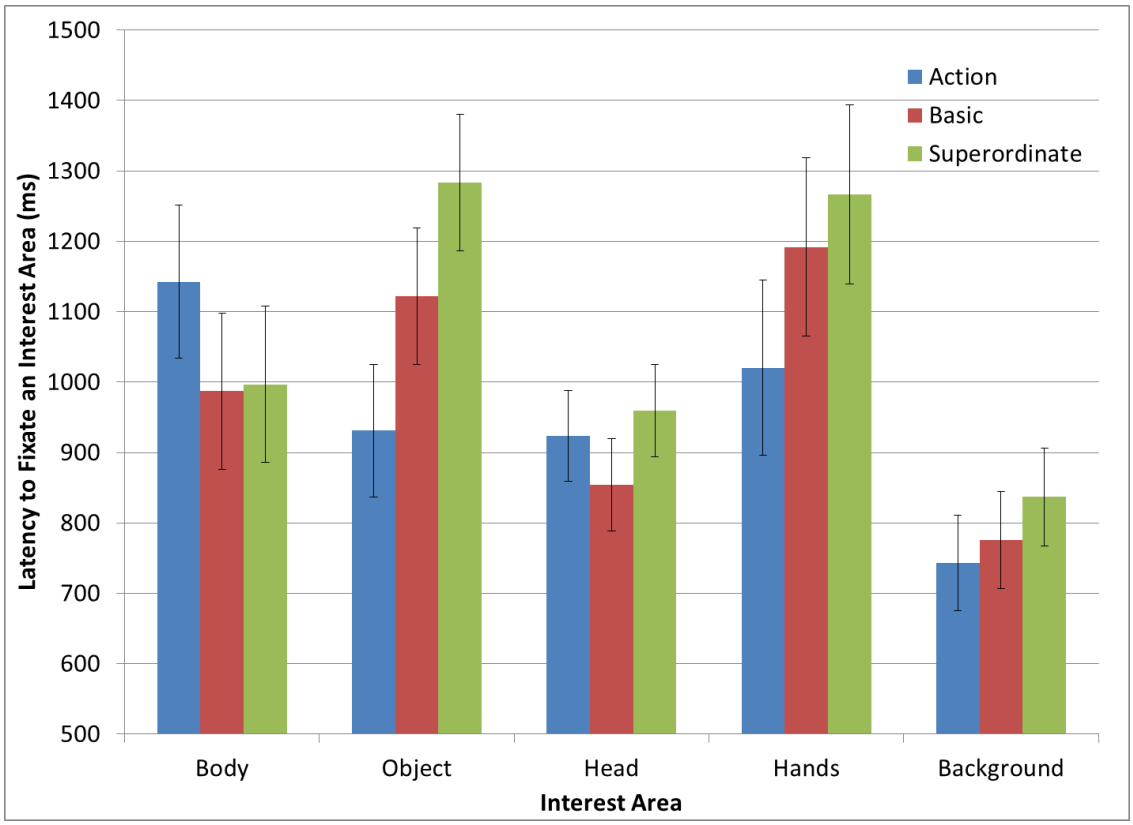

Figure 3.5 The mean latency to fixate each interest area between the three categorization tasks. Error bars represent the standard error.

Importantly, the latency to fixate an interest area varies depending on its importance to the categorization task. For example, in the superordinate task, the latency to the object is much greater than the latency to the person and head, which indicates the object is not a critical feature to attend to when performing the superordinate scene categorization task. However, when categorizing the basic level action, the latency to the object is as fast compared to the person, head, and hands. This shows that the object is an important semantic feature to be fixated in order to categorize the basic level action.

Table 3.6 Descriptive statistics for the latency (in ms) to first fixate an interest area for each image categorization task.

	Interest Areas									
	Body		Object		Head		Hands		Background	
Categorization task	M	SD	M	SD	M	SD	M	SD	M	SD
Action	1142.38	717.63	930.94	526.97	923.62	186.48	1020.34	450.01	743.35	276.70
Basic	987.12	345.91	1122.00	367.40	854.39	339.74	1191.89	642.81	775.45	357.58
Superordinate	996.85	409.96	1283.26	446.86	959.54	368.13	1266.41	674.18	836.74	337.03

Comparison of Eye Movements between Categorization Tasks

The eye movement analyses based on the first saccade in the scene show an effect of *what* was fixated and *when* it was fixated. Thus, the categorization task has a strong effect on attentional selection of different semantic features in the scene image. However, will these effects change, or become stronger, when all of the eye movement data is included in the analysis? All eye movements were analyzed from the trials that allowed participants 3 and 7 fixations to view the scene image. The first fixation condition was eliminated from the analysis since participants were not allowed to make an eye movement. The percentage of fixations falling on each of the five respective interest areas was analyzed with a 3 (categorization task [Basic level Action, Basic level scene, and Superordinate level scene category]) x 5 (Interest Area [Head, Body, Hands, Object, and Background areas]) Mixed Factorial ANOVA. Figure 3.6 and Table 3.7 present the descriptive statistics regarding the percentage of fixations on each interest area for each categorization task.

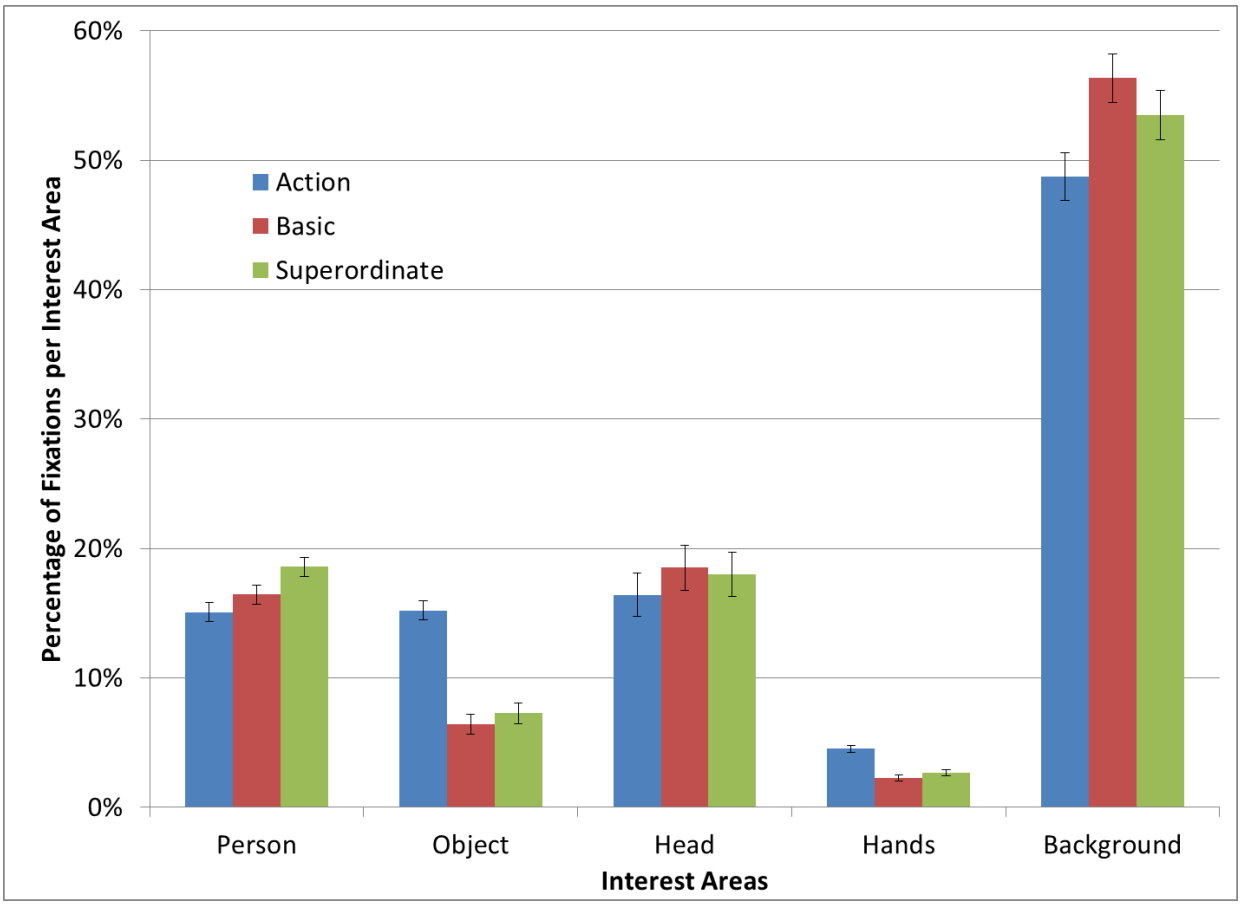

Figure 3.6 The mean percentage of fixations on each interest area between the three categorization tasks. Error bars represent standard errors.

Table 3.7 Descriptive statistics for the percentage of fixations on each interest area for each categorization task.

	Interest Areas									
	Body		Object		Head		Hands		Background	
Categorization task	M	SD	M	SD	M	SD	M	SD	M	SD
Action	15.08	3.59	15.22	5.08	16.42	6.43	4.53	1.37	48.75	6.29
Basic	16.44	3.12	6.45	2.74	18.52	9.31	2.26	1.21	56.33	10.69
Superordinate	18.59	3.55	7.26	2.41	17.99	8.26	2.68	0.94	53.48	9.07

As intended, there were no significant differences in the number of fixations between each categorization task (F (2, 64) = 0.58, p = .56). This was as intended because the number of fixations allowed on the scene image was manipulated as a key independent variable in the study. However, there was a main effect for the interest area that was fixated (F (4, 256) = 587.25, p < .001). Bonferroni-corrected t-tests were used to determine the differences between the percentage of fixations falling in each interest area (critical p-value = .005). Similar to the previous analysis examining the percentage of fixations on each interest area for the first eye movement, a greater percentage of fixations were on the scene background compared to the head, hands, body, and object (ts (66) ≥ 17.85, ps < .001). The fixation percentage on the body and head were not significantly different from each other (t (66) = 0.82, p = .42), however both had a greater fixation percentage than that for the object (ts (66) ≥ 6.61, ps < .001). The smallest fixation percentage was on the hands compared to the other four interest areas (ts (66) ≥ 11.96, ps < .001).

Critically, the interest areas that were fixated depended on the categorization task that was being performed (F (8, 256) = 5.86, p < .001), which was very similar to the results regarding the fixation location for the first eye movement in the scene. Bonferroni-corrected t-tests were used to compare the percentage of fixations on each interest area between the three categorization tasks (critical p-value = .003). Basic level action categorization resulted in a greater percentage of fixations on the object and hands compared to the basic and superordinate scene categorization task (ts (43) ≥ 5.25, ps < .001). This indicates that the key object and hands were critical semantic features for basic level action categorization, whereas basic and superordinate scene categorization may not emphasize or require these semantic features. One explanation for why fixations were greater for the hands in the basic level action categorization

task could simply be due to the proximity of the hands to the object. However, an alternative hypothesis provided by research on embodied cognition would suggest that showing the hands manipulating the object in a scene context allows the viewer to understand the intentions of the person (Iacoboni et al., 2005). However, the current data cannot distinguish these two alternative hypotheses. Additionally, superordinate scene categorization resulted in a greater percentage of fixations on the body compared to basic level action categorization (t (43) = 3.30, p = .002).

To account for the fact that the percentage of fixations on each type of interest area depended on the size of that area, domain-relative ratios were once again calculated and analyzed using a 3 (categorization task [Basic level Action, Basic level scene category, and Superordinate level scene category task]) x 5 (Interest Area [Head, Body, Hands, Object, and Background areas]) Mixed Factorial ANOVA. Figure 3.7 and Table 3.8 show the descriptive statistics for the domain-relative ratios for each categorization task and interest area. Domain-relative ratios were different between the three categorization conditions (F (2, 64) = 13.88, p < .001). Namely, the basic level action categorization task had greater domain-relative ratios than the basic and superordinate tasks (ts (43) ≥ 4.00, ps ≤ .001); however there was no difference between the basic and superordinate domain-relative ratios (t (42) = 1.11, p = .28). This indicates that those in the basic level action categorization task processed the scene image differently than those performing the scene categorization tasks. There was also a significant main effect for the interest area (F (4, 256) = 91.03, p < .001). Bonferroni corrected t-tests were used to determine the significant differences between interest areas on domain-relative ratios (critical p-value = .005). The head had overall greater domain-relative ratios than any other interest area (ts (66) ≥ 3.16, ps ≤ .002). There was no overall difference in the domain-relative ratios between the object and hands (t (68) = 2.81, p = .007), however both interest areas had greater domain-

relative ratios than the body and background (*t*s (66) ≥ 7.56, *p*s < .001). The body also had a greater domain-relative ratio than the background (*t* (66) = 16.99, *p* < .001). These data are consistent with the person bias reported by Fletcher-Watson et al. (2008). Specifically, there is a greater tendency for fixations to fall on a person compared to the other information present in the scene image. Interestingly, this person bias was more pronounced in the present analysis than the results from the first eye movement. Limiting the analysis to the first eye movement produced a weaker person bias, since the domain-relative ratio for the head was not different from the hands or object. This indicates that allowing more fixations on the scene image may produce a stronger person bias. However, certain areas of the person were more likely to be attended then other areas, particularly the face.

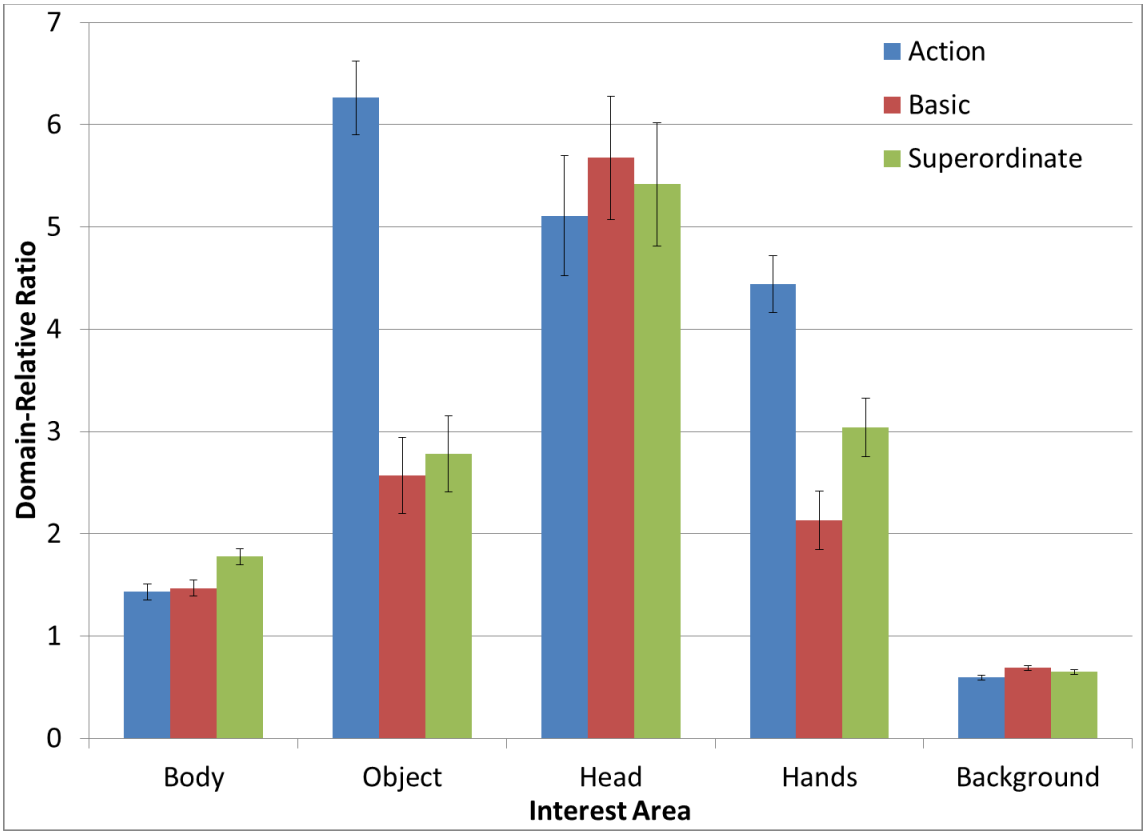

Figure 3.7 The mean domain relative ratio for each interest area between the three categorization tasks. Error bars represent the standard error.

Of particular importance, the amount of processing devoted to each interest area significantly varied depending on the categorization task that was being performed (F (8, 256) = 9.15, $p < .001$). Bonferroni-corrected t-tests were used to assess significant differences for each interest area between each categorization task (critical p-value = .003). Basic level action categorization had greater domain-relative ratios for the object and hands than the basic (ts (43) $\geq$ 5.35, ps < .001) and superordinate scene categorization tasks (ts (43) $\geq$ 3.37, ps $\leq$ .002). Indeed, the basic level action categorization task produced a domain-relative fixation ratio for the object that was six times greater than would be predicted by chance, and that was more than twice that of either of the two scene categorization tasks. There were no differences in the domain-relative ratios between the basic and superordinate categorization tasks (ts (42) $\leq$ 2.92, ps $\geq$ .006). This shows that the categorization task clearly affected the deployment of eye movements, and thus attention, in the scene. Specifically, basic level action categorization resulted in a greater percentage of fixations on the object and hands relative to the basic and superordinate conditions, particularly when the size of each interest area was controlled. These findings amplify and clarify the previous analyses of the first eye movement in the scene.

Table 3.8 Domain relative ratio descriptive statistics for each categorizing condition at each interest area.

	Interest Areas									
	Body		Object		Head		Hands		Background	
Categorization task	M	SD	M	SD	M	SD	M	SD	M	SD
Action	1.43	0.40	6.26	2.57	5.11	2.25	4.44	1.66	0.59	0.08
Basic	1.47	0.28	2.57	0.85	5.68	3.20	2.13	1.19	0.69	0.13
Superordinate	1.78	0.41	2.78	1.26	5.42	2.99	3.04	1.05	0.65	0.11

Overall, the image categorization task had a clear effect on what viewers fixated in the scene image. However, did the categorization task affect the perceptual encoding of these different types of semantic features? To answer this question, the average fixation duration in each interest area, and the total amount of time spent fixating each interest area, referred to as the dwell time, were examined.

The average fixation duration on each interest area for each categorization task was computed and analyzed using a 3 (categorization task [basic level Action, Basic level scene category, and Superordinate level scene category]) x 5 (Interest Area [Head, Body, Hands, Object, and Background areas]) Mixed Factorial ANOVA. Figure 3.8 and Table 3.9 present the average fixation duration per interest area for the three categorization tasks. There were no differences in the average fixation duration between the three categorization tasks (F (2, 64) = 2.50, p = .090). However, there was a difference in the fixation durations between the five interest areas (F (4, 256) = 4.36, p = .002). Bonferroni-correct t-tests were used to compare fixation durations between interest areas (critical p-value = .005). The average fixation durations for the object and background areas were greater than the average fixation duration for the body (ts (66) $\geq$ 3.16, ps $\leq$ .002). No other comparisons approached significance. This effect for fixation duration did not interact with the categorization task being performed (F (8, 256) = 0.172, p = .994). Therefore, the categorization task did not have an effect on the average fixation duration, and it did not interact with the interest area which was fixated. However, the dwell time, defined as the total amount of time spent fixating a given interest area across multiple fixations, may vary depending on the categorization task.

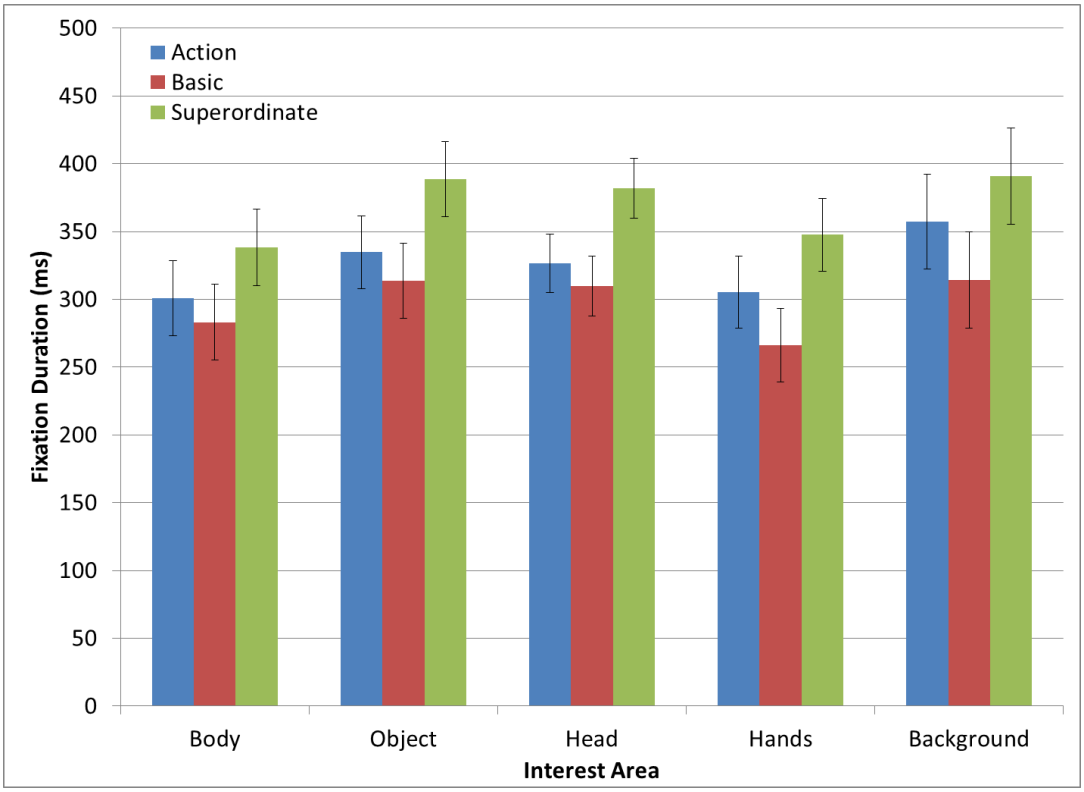

Figure 3.8 The mean fixation duration on each interest area between the three categorization tasks. Error bars represent the standard error.

Table 3.9 Descriptive statistics for the average fixation duration (ms) on each respective interest area for the Action, Basic, and Superordinate scene categorization tasks.

	Interest Areas									
	Body		Object		Head		Hands		Background	
Categorization task	M	SD	M	SD	M	SD	M	SD	M	SD
Action	300.29	131.74	334.75	102.01	326.70	94.68	305.16	113.16	357.38	216.42
Basic	283.13	103.96	313.66	104.48	309.82	75.95	266.24	65.40	314.21	100.33
Superordinate	338.09	155.61	388.63	170.49	381.86	130.84	347.52	176.32	390.90	161.86

The average dwell time spent in each interest area for each categorization task was analyzed using a 3 (categorization task [Basic level Action, Basic level scene, and Superordinate

level scene task]) x 5 (Interest Area [Head, Body, Hands, Object, and Background areas]) Mixed Factorial ANOVA. Figure 3.9 and Table 3.10 present the descriptive statistics for the average dwell time on each interest area for the three categorization tasks. There was no main effect for the total time viewing interest areas between the three categorization tasks (F (2, 64) = 1.87, p = .16)[3]. Again, this is unsurprising since the number of fixations allowed to view the scene image was manipulated as an independent variable and held constant across the three categorization tasks. However, there was a main effect for interest area on dwell time (F (4, 256) = 46.19, p < .001). Bonferroni-corrected t-tests were used to determine the significant differences in dwell time between the five interest areas (critical p-value = .005). Dwell time for the background was significantly greater than the dwell times for the body, head, hands, and object (ts (66) ≥ 5.89, ps < .001). There was no difference in dwell times between the body, head, and object (ts (66) ≤ 1.35, ps ≥ .18), however they were significantly greater than the dwell time for the hands (ts (66) ≥ 4.98, ps < .001). There was no interaction for dwell time between the interest areas and the categorization tasks (F (8, 256) = 0.93, p = .49). Therefore, similar to the fixation duration analysis, the categorization task had no effect on the total amount of time spent in each interest area, although the background did have a greater dwell time than the other interest areas. However, this is most likely a result of the scene background composing, on average, 82% of the entire scene image. Therefore, a domain-relative ratio was computed for dwell time to determine if it was affected by the categorization task, after controlling for the size of each respective interest area.

[3] Participants in each categorization task viewed scene images for an equal number of fixations. Thus, Figure 3.9 suggests that dwell times were slightly longer for the superordinate scene categorization compared to the other two categorization tasks. However, The dwell times for the superordinate scene categorization task were not significantly greater than the other two categorization tasks.

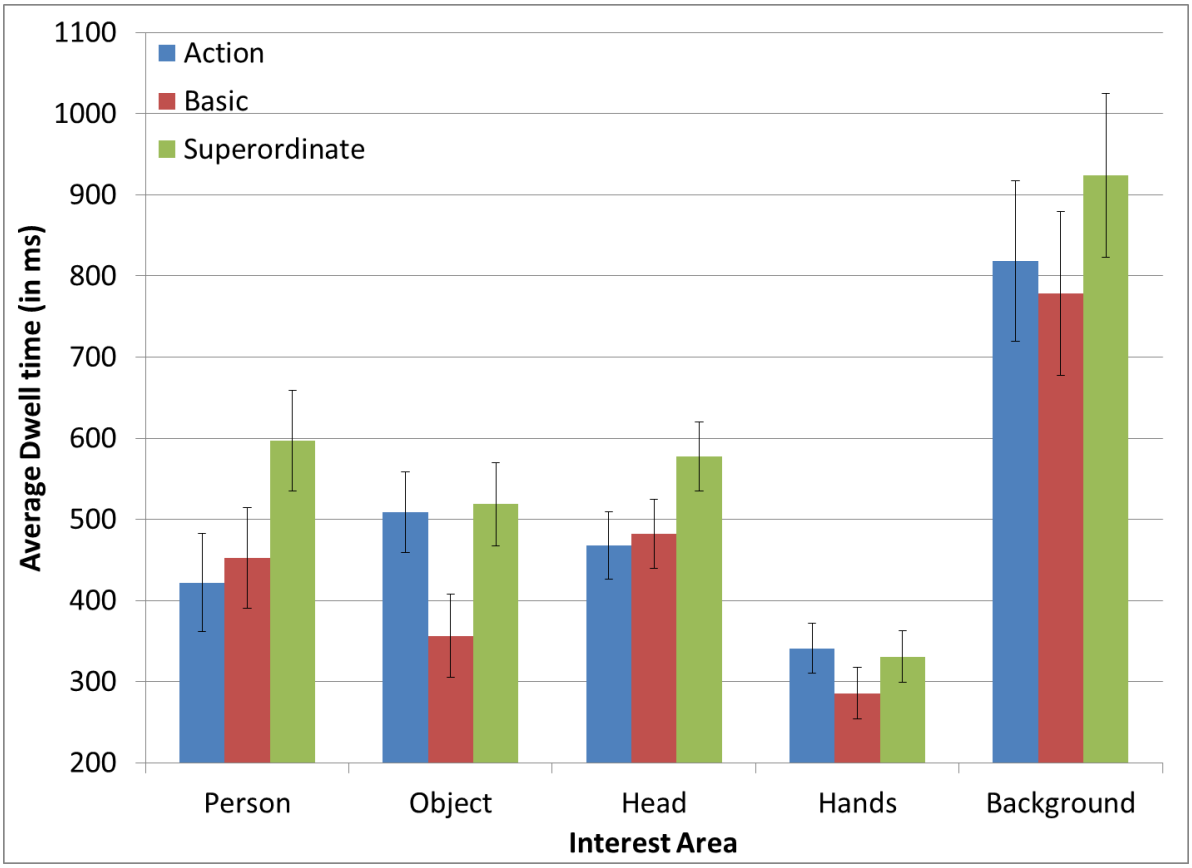

Figure 3.9 The mean dwell time per interest area for each of the categorization tasks. Dwell time is measured as the total time spent fixating an interest area. Data is from conditions with > 1 fixation. Error bars represent the standard error.

Table 3.10 Descriptive statistics for the average dwell time (ms) on each respective interest area for the Action, Basic, and Superordinate scene categorization tasks. Dwell time measures the total amount of time spent in each interest area.

	Interest Areas									
	Body		Object		Head		Hands		Background	
Categorization task	M	SD	M	SD	M	SD	M	SD	M	SD
Action	422.01	214.41	508.73	219.04	467.64	152.44	341.11	171.18	818.51	532.75
Basic	452.68	214.87	356.61	125.84	482.08	171.02	285.78	99.76	778.85	348.00
Superordinate	596.67	403.84	518.66	329.02	577.94	259.52	330.80	163.21	924.03	514.18

The domain-relative ratio was computed by dividing the percentage of time each interest area was fixated by the percentage of area that the interest area encompassed within the image. The average dwell time domain-relative ratio was analyzed using a 3 (categorization task [Basic level Action, Basic level scene, and Superordinate level scene task]) x 5 (Interest Area [Head, Body, Hands, Object, and Background areas]) Mixed Factorial ANOVA. Figure 3.10 and table 3.11 present the descriptive statistics for the dwell time domain-relative ratio. A significant main effect for the categorization task (F (2, 64) = 7.29, p = .001) resulted from greater dwell time domain-relative ratios for the basic level action task than both the basic level and superordinate scene categorization task (ts (43) $\geq$ 2.78, ps $\leq$ .008). Additionally, there was a significant main effect for the interest areas (F (4, 256) = 53.21, p < .001). Bonferroni corrected t-tests were used to determine the different dwell time domain-relative ratios between the interest areas (critical p-value = .005). There was no difference in the dwell time domain-relative ratios for the object and head (t (66) = 0.14, p = .89), however both regions were greater than the hands, body, and background (ts (66) $\geq$ 3.87, ps < .001). Likewise, the dwell time domain-relative ratio for the hands was greater than the person and background (ts (66) $\geq$ 4.98, ps < .001). Finally, the dwell time domain-relative ratio for the person was greater than the scene background (t (66) = 16.42, p < .001). Critically, there was a significant interaction for dwell time domain-relative ratios between the categorization tasks and interest areas (F (8, 256) = 6.80, p < .001). Again, bonferroni corrected t-test were used to determine the differences between dwell time domain relative ratios between each categorization task for a given interest area (critical p-value = .003). The dwell time domain-relative ratio for the object was greater in the basic level action

categorization task than the two scene categorization tasks (*t*s (43) ≥ 4.43, *p*s < .001). No other comparisons were significant.

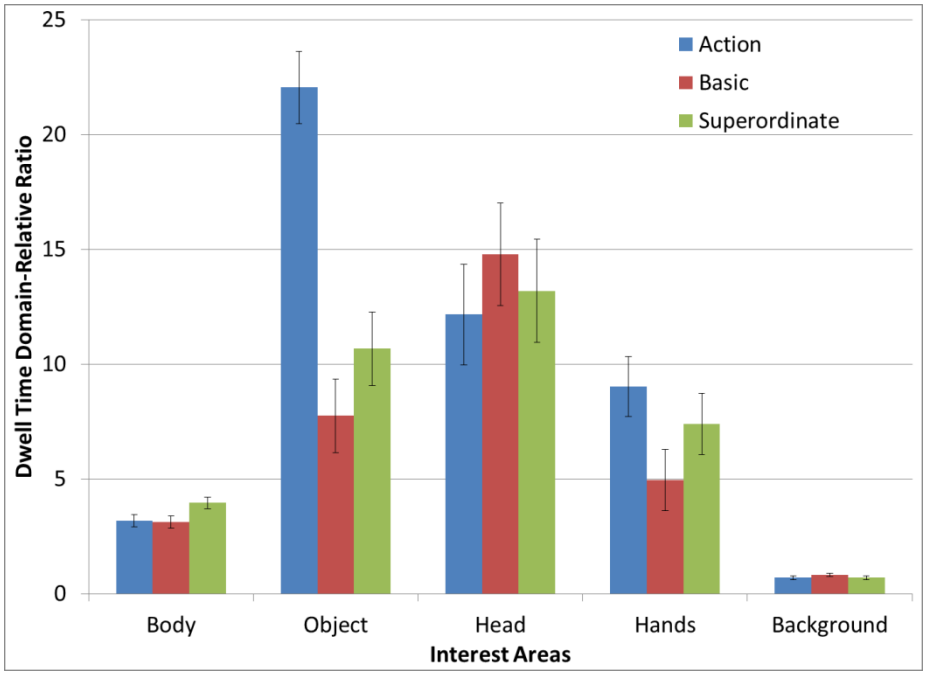

Figure 3.10 The mean dwell time domain-relative ratios for each interest area and categorization task. Data is from conditions with > 1 fixation. Error bars represent the standard error.

Table 3.11 Descriptive statistics for the average dwell time domain-relative ratios for each interest area and categorization task.

	Interest Areas									
	Body		Object		Head		Hands		Background	
Categorization task	M	SD	M	SD	M	SD	M	SD	M	SD
Action	3.17	1.11	22.05	10.23	12.16	8.22	9.02	4.71	0.69	0.37
Basic	3.13	1.10	7.74	4.61	14.79	12.27	4.94	5.76	0.81	0.44
Superordinate	3.94	1.49	10.66	6.55	13.19	10.65	7.38	7.92	0.69	0.24

The domain-relative ratio for dwell time shows that the relative amount of time spent fixating specific scene information, when considered in proportion to its area within the scene, depended on the categorization task being performed. This is consistent not only with the idea that viewing task affects what viewers look at in scenes (DeAngulus & Pelz, 2009; Yarbus, 1967) but also the time spent processing that information. Additionally, the effect of the task on relative viewing time was limited to the object being manipulated, with the effect on the relative time spent viewing the hands being much weaker. This is interesting since the previous analysis examining the domain-relative ratio for the percentage of fixations on each interest area did indicate an effect for the hands. These two analyses seem to suggest that the hands may be fixated more frequently when categorizing the action, however spending time fixating the hands is not as critical as fixating the object being manipulated. Evidence for processing the person is also evident in the relative dwell time domain-relative ratios. Namely, regardless of the categorization task being performed, more processing time was devoted to the head compared to the other interest areas, unless categorizing the action. This person processing bias is also consistent with the previous literature (Fletcher-Watson, Findlay, Leekam, & Benson, 2008; Humphrey & Underwood, 2010; Zwickel & Võ, 2010). Overall, the eye movement results show that the categorization task affected attentional selection processes in eye movements, and encoding processes, after controlling for the size of the interest areas.

Discussion

Experiment 2 examined the scene image information that was fixated and the number of fixations that were required for each image categorization task, therefore extending the results from Experiment 1, which was limited to scene image processing times within a single fixation. Specifically, consistent with previous research on scene gist recognition a single fixation was

sufficient processing time for the basic level and superordinate level scene categorization tasks to reach ceiling performance (Biederman 1972; Potter 1976; Eckstein, Drescher, & Shimozaki 2006; Greene and Oliva 2009; Mace, Joubert et al. 2009; Loschky and Larson 2010). These results are inconsistent with the research supporting basic level theory (Gosselin & Schyns, 2001; Jolicour, Gluck, & Kosslyn, 1984, Rosch et al., 1976; Tversky & Hemenway, 1983), which would predict that the basic level categorization task would be better than the superordinate scene categorization task. In contrast to this prediction, however, performance was similar between the two scene categorization tasks at each processing time. The results are also inconsistent with our person-bias hypothesis that the attentional bias to people in scenes would result in better categorization for the basic level action relative to the scene categorization tasks. In fact, basic level action categorization required two to three fixations in order for ceiling performance to be achieved. However, these results are most consistent with the hypothesis predicting coarse-to-fine processing for the scene image. Although superordinate level scene categorization was not better than basic level scene categorization, the data was in the predicted direction. Likewise, consistent with Experiment 1, both scene categorization tasks were better than the basic level action categorization task, which required more than a single fixation to reach ceiling performance. The current results are consistent with previous studies that have found a course-to-fine processing strategy for both objects and scene images (Large, Kiss, & McMullen, 2004; Loschky & Larson, 2010; Macé, Joubert, Nespoulous, & Fabre-Thorpe, 2009; Poncet, Reddy, & Fabre-Thorpe, 2012; Rogers & Patterson, 2007).

Since basic level action category distinctions required more than a single fixation to reach ceiling performance, it could be expected that the information fixated in the basic level action categorization task would differ from the information fixated in the two scene categorization

tasks. This is consistent with previous research showing that the task plays a role in determining where eye movements go in scenes (DeAngulus & Pelz, 2009; Yarbus, 1967). However, an alternative hypothesis predicted that eye movements between the three categorization tasks would be similar, based on the fact that scene categorization at the basic and superordinate level requires only a single fixation, and research showing the person bias in scene viewing (Fletcher-Watson, Findlay, Leekam, & Benson, 2008; Humphrey & Underwood, 2010; Zwickel & Võ, 2010). Thus, after acquiring the scene category in the first fixation, viewers would be free to fixate the person in all of the categorization tasks, resulting in similar fixation patterns. Although there was very clear evidence both for the person bias in all three task conditions, and that scene categorization took only a single fixation to reach asymptote, there was nevertheless strong evidence of a task effect on attentional selection processes and encoding processes in eye movements, particularly in the basic level action categorization task, consistent with previous research (DeAngelus & Pelz, 2009; Yarbus, 1967).

The top-down effect of the categorization task on eye movements was evident in the first eye movement in the scene. Namely, for the basic level action categorization task, a greater percentage of fixations fell on the object and hands than in both scene categorization tasks. This effect was even stronger after controlling for the size of each respective interest area in the scene. Likewise, the categorization task affected the latency to fixate different semantic regions of the scene. In the basic level action categorization task, the latency to fixate the object and hands were no different from any of the other interest areas. Conversely, in the basic level scene categorization task, the latency to fixate the object was slower than the time to fixate the head. Furthermore, for the superordinate task, the latency for the object was slower than the head and body. These results provide strong evidence for the effect of the task on what and when

information was fixated in the scene image. These two dependent variables provide strong evidence that the categorization task affected attentional selection processes for eye movements. Namely, prior to initiating a saccade, covert attention selects a new spatial location in the image, which is then followed by a saccade to that location (Henderson, 1992; Rizzolatti, Riggio, Dascola, & Umilta, 1987). Thus, the current study shows that attentional selection of semantic scene information differs between action and scene categorization. Specifically, basic level action categorization places more importance on object information than the two scene categorization conditions.

These effects seen when analyzing the first fixation became stronger when all of the eye movements were included in the analysis. These analyses showed that, besides the scene background, the body and head were most likely to be fixated. This shows direct evidence of a person bias when viewing a scene image (Fletcher-Watson, Findlay, Leekam, & Benson, 2008; Humphrey & Underwood, 2010; Zwickel & Võ, 2010). When the relative physical image area of the respective interest areas were controlled for, fixations in all three categorization conditions were most biased to process the person's head relative to its size, again consistent with previous research (Fletcher-Watson, Findlay, Leekam, & Benson, 2008). However, as seen with the eye movement analysis based on the first fixation, the part of the person that was fixated differed depending on the categorization task being performed. Namely, the object and hands received a greater percentage of fixations, relative to their size, when categorizing the basic level action compared to the two scene categorization tasks. Furthermore, the categorization task did not have an effect on the average fixation duration and dwell time in each interest area. However, when computing the domain-relative ratio for the dwell times for each respective interest area, then the categorization task had an effect on the encoding processes for different semantic

feature in the scene. Namely, when categorizing the basic level action, more time was spent encoding the object being manipulated compared to the two scene categorization tasks. Thus, encoding the details of the object was critical when categorizing the basic level action. Thus, the current study shows that the top-down influences what, when, and for how long semantic scene information was processed. Namely, the task influenced both the attentional selection and perceptual encoding processes involved in eye movements. When categorizing the basic level action, one could make an argument regarding the importance of the object information versus the hands. One simple explanation for the increased fixations for the hands is due to their proximity to the object. Since the hands are involved in manipulating the object, which is critical for categorizing the basic level action, then fixations may have a greater opportunity to land on the hands, possibly due to inaccuracy in saccade execution (or measurement). Conversely, it has been proposed that the hands are useful cues for perceiving events in everyday actions in dynamic film (Smith, 2012). Likewise, research has shown that the hand and object being manipulated in a specific context allows the observer to interpret the intentions of the individual (Iacoboni, Molnar-Szakacs, Gallese, Buccio, Mazziotta, & Rizzolatti, 2005). According to these studies, fixations on the object and hands would be critical to the identification of the basic level action in the scene image.

The lack of a sensitivity advantage for the superordinate level task over the basic level task at the very earliest stage of processing (24 ms) in Experiment 2 is inconsistent with the results of Experiment 1 and previous findings regarding the superordinate advantage at early processing times (Joubert, Rousselet, Fize, & Fabre-Thorpe, 2007; Loschky & Larson, 2010; Macé, Joubert, Nespoulous, & Fabre-Thorpe, 2009). This failure to replicate our previous effect may be related to the range of processing times examined between experiments. Experiment 1

used a range of processing times within a single fixation, whereas Experiment 2 had a much larger range for processing times spanning up to 7 fixations. This difference in the range of processing times may have produced different levels of motivation between the two studies. Namely, participants in Experiment 1 were presented with short processing times for the images, which made the task more difficult. As participants progressed through Experiment 1, they may have become aware of the difficulty of the task, and therefore been less motivated to give their best effort. Conversely, the range of processing times selected for Experiment 2 was much larger, and this experiment may have been perceived as being easier. The greater ease of Experiment 2 may have resulted in participants trying harder to do well in the task, especially for the 24 ms SOA condition. This motivation hypothesis would explain the differences seen in the Experiment 1 for the superordinate and basic level scene categorization performance, and the lack of the same effect in Experiment 2. If this hypothesis is correct, then Experiment 1 would show performance at early processing times when viewers are giving less effort, and would indicate that the easier scene categorization task is indeed at the superordinate level.

Chapter 4 - Experiment Three

When reading a story or viewing a film, the story follows the actions and goals of one or more characters. The importance of the character's goals and actions can been seen in theories of story grammars and in the event indexing model (Mandler & Johnson, 1977; Rinck & Weber, 2003; Scott-Rich & Taylor, 2001; Thorndyke, 1977), recognition memory for a sequence of actions (Brewer & Dupree, 1983; Lichtenstein & Brewer, 1980), and the perception of events within stories (Zacks & Tversky, 2001; Magliano & Zacks, 2011). Specifically, story grammar theories decompose a story based on the events that involve a character, and the goals and actions taken by the character (Mandler & Johnson, 1977; Thorndyke, 1977). Studies show that

when a character is introduced in a written narrative, ratings of coherence and cohesion decreased and reading time increased, relative to changes in the time and location of story events. This indicates that a change in the narrative character disrupts integration of that new information with previous story information, thus more time is required to integrate the new information with the previous event representation (Rinck & Weber, 2003; Scott-Rich & Taylor, 2001). Similarly, recall memory is better for goal-directed activities than activities that were not related to the goal (Brewer & Dupree, 1983; Lichtenstein & Brewer, 1980). Likewise, events tend to be goal-directed activities, like "Washing a car" or "Making lunch," which suggests that the basic level actions that are observed, and the inferred goals that guide them, are central to understanding the event and being able to identify a new event (Zacks & Tversky, 2001; Magliano & Zacks, 2011). Thus it would have been reasonable to expect that the first semantic construct recognized in a scene would be the basic level action. However, the results of Experiments 1 and 2 shows that image categorization starts with global scene representation at the superordinate level, then more specific basic level scene category distinctions, and finally the basic level action. Thus, image categorization occurs in a global-to-local fashion.

This global-to-local processing order replicates work by other recent research examining object and scene recognition categorization (Large, Kiss, & McMullen, 2004; Loschky & Larson, 2010; Macé, Joubert, Nespoulous, & Fabre-Thorpe, 2009; Rogers & Patterson, 2007). However, this processing order is at odds with previous research showing that the basic level is categorized prior to the superordinate level (Gosselin & Schyns, 2001; Jolicour, Gluck, & Kosslyn, 1984, Rosch et al., 1976; Tversky & Hemenway, 1983). So how can these two apparently contradictory results be reconciled? One explanation is in terms of the cognitive processes used to categorize the scene image in a given experimental task. In rapid scene

categorization tasks in the scene gist recognition literature, the processes are fast, early, perceptual categorization mechanisms that deal with visual stimuli in a course-to-fine order. In the types of tasks used more commonly in the categorization literature, the processes involve a somewhat slower, later, linguistic categorization system that is sensitive to the communicative informativeness of categories, which favors the basic level.

The classic studies by Rosch et al. (1976), for object categorization, and Tversky and Hemanway (1983), for scene categorization, show that more attributes are listed for basic level categories than for superordinate categories. Thus, the basic level allows for a substantial increase in the informativeness of communication about the object or scene being viewed. Interestingly, Rosch et al. (1976, Exp. 7) also found the basic level advantage in a perceptual categorization task, in which basic level object categorization was quicker than superordinate and subordinate level categorization. This apparently contradicts the above proposal that the basic level advantage is absent for fast, early, visual categorization processes. However, a closer look at the methods used in Rosch et al. (1976, Exp. 7) call this into question. Specifically, at the beginning of a trial, a cue word was presented 500 ms prior to the onset of the object, which then remained in view until the participant made a "Yes" or "No" response regarding whether the object matched the pre-cue. As reported by Rosch et al. (1976, Exp. 7), the average reaction time for "Yes" responses to the basic level category was 535 ms and 578 ms for "No" responses. Thus, participants were allowed to process the stimulus for more than half a second, which is slightly over 22 to 24 times longer than the earliest processing time tested in Experiment 1.

The lack of a basic level advantage when viewers are given very short processing times has been further elucidated by Rogers and Patterson (2007). Initially, Rogers and Patterson (2007) replicated the basic level advantage for an object categorization task. However, in a

follow-up experiment, when the same object categorization task required a speeded response from participants, the effect changed and an advantage was found for the superordinate category. Thus, it seems that at the early stages of visual categorization, a perceptual advantage is found for superordinate categorization, but as more time is given to categorize the object, a later (presumably linguistic) advantage is found for the basic level. An MEG study by Löw et al., (2003) has proposed a similar distinction between perceptual and linguistic categorization mechanisms. Between 170 and 210 ms after object onset, brain activity was found to be correlated with superordinate object categories in the right occipital-temporal lobe, while between 200 and 450 ms, these same distinctions were made in the left temporal hemisphere. Löw et al., (2003) claim that these spatially distinct brain areas are representing a quick perceptual categorization in the right temporal cortex, followed by a later, linguistic/semantic categorization in the left temporal cortex.

Therefore, the above distinctions between the time-course of linguistic and perceptual categorization suggests a way of reconciling the findings from Experiments 1 and 2 and those used to support Rosch et al.'s (1976) basic level theory. The findings presented in Experiments 1 and 2 are consistent with a fast, early, perceptual categorization mechanism that begins by making superordinate level scene category distinctions. Conversely, a slower, later, linguistic categorization system may be more concerned with communicating informative image constructs, captured by basic level scene and basic level action representations, and least concerned with coarse-level scene representations. Evidence for this slower linguistic categorization mechanism can be found for scene and object naming tasks that show basic level terms are used more often than superordinate terms, since the basic level is more informative (Rosch et al., 1976; Tversky & Hemenway, 1983). In fact, when participants were asked to

name scene images, Tversky and Hemenway (1983) found that participants never used superordinate category names like “Indoor” or “Outdoor” and subordinate names were used infrequently (pp. 137). Likewise, the basic level advantage is also present for events (Morris & Murphy, 1990; Rifkin, 1985), with events being more likely to be named at the basic level than the superordinate level (Morris & Murphy, 1990). Thus, if linguistic descriptions of scenes are sensitive to more informative scene constructs, then participant descriptions should report more basic level image descriptors than superordinate descriptors.

Additionally, a separate argument can be made for the importance and informativeness of basic level actions, based on research on the person bias shown by eye movements. These studies show that the first saccade is likely to move toward the person (Fletcher-Watson, Findlay, Leekam, & Benson, 2008; Humphrey & Underwood, 2010). This could be taken as evidence that the most important construct in the scene is the action being performed. This is even more evident in the research examining how eye movements are linked to verbal descriptions of images. Namely, scene information was verbally reported within a second of that information being fixated (Griffin & Bock, 2000; Meyer, Sleiderink, & Levelt, 1998). If so, then it could be hypothesized that the basic level action category would be reported more often than the basic level and superordinate level scene categories. Furthermore, this hypothesis can be specified to predict that the basic level action would be reported more often than the superordinate level action, which would be consistent with research showing a basic level advantage for events (Morris & Murphy, 1990; Rifkin, 1985).

Conversely, an alternative hypothesis would predict that the first image category activated is the first term used to describe the image. Experiments 1 and 2 show that the scene category is activated before the basic level action. Likewise, the superordinate level scene

category is recognized prior to the basic level, which is consistent with previous studies on scene categorization and object categorization (Large, Kiss, & McMullen, 2004; Loschky & Larson, 2010; Macé, Joubert, Nespoulous, & Fabre-Thorpe, 2009; Rogers & Patterson, 2007). This superordinate advantage was also found when participants were asked to name scene images (Fei-Fei, Iyer, Koch, & Perona, 2007). Namely, when scene image processing was limited by scene masking, then superordinate terms were used more often than finer level scene terms. This was especially evident for Outdoor scenes, however Indoor scenes were just as likely to be described as "Indoor" versus an finer-level indoor category term. Likewise, if fixations are tightly liked to verbal descriptions for a scene (Griffin & Bock, 2000; Meyer, Sleiderink, & Levelt, 1998), then the scene category should be reported first since it is represented during the first fixation in the scene image.

Experiment 3 asked participants to describe the scenes used in Experiments 1 and 2. Participant's responses were coded to determine the proportion of superordinate level and basic level scene and action terms. If the linguistic categorization system is biased to report the most informative construct present in the scene, then basic level terms should be use more often than superordinate terms. However, if the informativeness of the image is dependent on the action, then the basic level action should be used more than the superordinate action and basic level scene terms. Alternatively, if the terms that are used to describe the scene are dependent on the first image category activated, then the exact opposite pattern should be observed. Namely, superordinate scene category terms should be used more than basic level scene category terms, and both superordinate and basic level action categorization terms.

Method

Participants

549 Participants (247 female) completed an on-line survey posted on Mechanical turk.com (https://www.mturk.com/mturk/welcome)[4]. All participants were self-reported native English speakers, with a mean age of 28.8 (SD = 10.4). Participants' age ranged between 16 and 71. All participants received 16 cents for completing the on-line survey.

Materials

The same 288 images used in Experiments one and two were used for the naming study. Participant responses were collected through an on-line survey on Mechanical Turk. Each participant was limited to viewing one image from each of the 8 basic level action categories. If more than one image per basic level action category had been presented to each participant, then it is likely that the response for one basic level action category image would influence the response for a second image from the same basic level action category. In order to avoid such a bias, only one scene image per basic level action category was presented to each participant, for a total of 8 scene image presentations per participant. Furthermore, each participant was only allowed to participate in the study once.

Procedure

Restrictions were placed on the sample that could participate in the study. Namely, respondents were selected to be from North America. Native English speakers were specified in the title of the survey on Mechanical Turk. Prior to beginning the survey, participants were

[4] 15 descriptions for each of the 288 images were sought. This would produce 4, 320 total descriptions. However, since participants were presented with eight images, a total 540 participants were required.

asked to report demographic information, including their age, gender, and native language. If the participant reported that their native language was not English, they were not permitted to participate in the study. Afterwards, participants were instructed to name each image with "a very simple and common one word label or short phrase that best describes the scene image. We do not want elaborate or creative answers. Instead, we want the most simple, obvious, direct sort of label or description that ordinary people would give for each picture." These instructions were modified from Tversky and Hemenway's (1983) instructions. After an image had been labeled, the next image was presented until a label was given for the scene image. This process continued until all eight scene images were labeled.

Results

Precursors

Participant responses were only analyzed if they reported English as their native language. All other responses were removed from the analysis and data for their dataset was re-collected.

The 4,392 image descriptions were coded regarding the relative frequency of each of the three categorical levels examined in Experiments 1 and 2 and an additional categorical level, superordinate level actions. Superordinate actions were general terms like "Working," "Chores," "Cleaning," and "Busy." As defined by Rosch et al. (1976), members of a superordinate category share few attributes due to their inclusiveness, whereas members of a basic level category have the most attributes in common. Therefore, these superordinate action descriptors are more inclusive of numerous actions and thus, share few attributes, whereas basic level action descriptors share many perceptual features. This exact result was found by Rifkin (1985), where events at the basic level have a greater number of shared attributes than superordinate events.

Responses were coded for the use of superordinate or basic level scene category terms, and superordinate or basic level action category terms. Additionally, coders identified if the description identified an object or person in the scene. All scene category terms were limited to nouns specifying a location, whereas action terms were limited to responses containing a verb or a gerund which described an action occurring in the scene. Coding of participants' responses was not mutually exclusive. Specifically, if a response contained an object or person, a basic level action term, a superordinate action term, a basic level scene category term, and a superordinate scene category term, then each response was coded regarding its categorization level. For example, a response such as "A person outdoors doing yard work mowing their lawn," would be coded as containing a response from each of the five categorization levels. Finally, responses that were deemed incorrect/inappropriate were also calculated. For instance, responses like "fun time" and "Gazing" were counted as incorrect.

Participant responses were coded by two coders naïve to the hypotheses of the study. Their inter-rater reliability was measured using Cohen's Kappa. The inter-rater reliability was acceptable (Cohen's Kappa $\geq .764$) and all remaining disagreements were resolved through discussion between the two coders and the author.

Analysis

The frequency of image descriptors used to label the scene images were analyzed using a 3 (Action descriptor [Absent, Basic level, superordinate level]) x 3 (Scene Category Descriptor [Absent, Basic Level, Superordinate level]) chi-square. Figure 4.1 and Table 4.1 present the frequencies and percentages for the frequency of action descriptors x scene descriptors. Responses that contained neither a scene nor action term (the absent/absent column in Fig 3.1 or cell in Table 3.1) were composed of incorrect responses (479 out of 1288), a reference to an

object or person (499 out of 1288), or neither an incorrect response nor a reference to an object/person (310 out of 1288). In Figure and Table 3.1, these categories were collapsed into a single variable (absent/absent). These responses appear to be quite frequent, however they are simply the sum of both correct and incorrect image descriptions that were neither an action nor a scene category term, for example object or person descriptions. Thus, analyses presented later examined the conditional probability of using an object/person term when either an action or scene category term was provided. The frequency of the action terms differed depending on the level of the scene term used (χ^2 (4) = 44.09, $p < .001$). A series of chi square test of independent frequencies were used to make specific comparisons to test our proposed competing hypotheses.

The frequency of basic level and superordinate level terms was compared in order to determine the level of detail that was used to label the scene images. If viewers provided more informative descriptions of the image, then this hypothesis predicts that the basic level descriptors would be used more often than superordinate level descriptors. Alternatively, the perceptually earliest categorization level recognized could be used to describe the scene image. This predicts that the superordinate descriptors would be used more than basic level descriptors. A chi square test of independent frequencies showed that a greater frequency of basic level descriptors (2,569[5] out of 3,180) were used to label the image than superordinate level descriptors (611[6] out of 3,180)(χ^2 (1) = 1205.59, $p < .001$) This supports the basic level theory showing that later, linguistic image categorization uses more informative image terms to describe

[5] The total frequency of basic level terms used was calculated by summing the total number of basic level scene terms (241 + 201 + 77 = 519) and total number of basic level action terms (2,028 + 22 = 2,050) for a total of 2,569 basic level terms.

[6] The total frequency of superordinate level terms used was calculated by summing the total number of superordinate level scene terms (12 + 22 + 1 = 35) and total number of superordinate level action terms (499 + 77 = 576) for a total of 611 superordinate level terms.

scene images. The frequency of action and scene descriptors were compared and the results showed that a greater frequency of action descriptors (2828 out of 3382) were used to describe images compared to scene descriptors (554 out of 3382)(χ^2 (1) = 1376.99, $p < .001$).

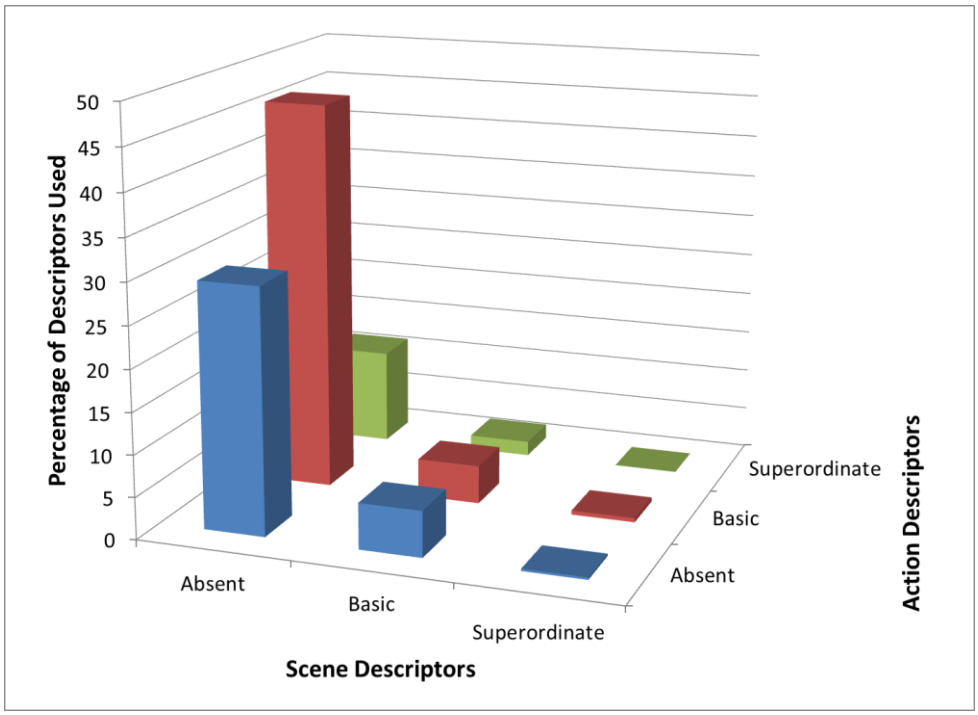

Figure 4.1 The percentage of scene and action descriptors used to label scene images

Table 4.1 Frequency and percentage of image descriptors used by basic level, superordinate level and both taxonomies used (Percentages in parentheses)

	Scene descriptor					
Action Descriptor	None		Basic		Superordinate	
None	1288	(29.35)	241	(5.52)	12	(0.27)
Basic	2028	(46.42)	201	(4.60)	22	(0.50)
Superordinate	499	(11.42)	77	(1.76)	1	(0.02)

Note. 4,369 total responses are presented in the table. An additional 23 response contained both a basic and superordinate action term, like "Busy talking on the phone."

Furthermore, two additional comparisons were made to determine if a greater frequency of basic level terms were used for both action and scene descriptions. Two chi square tests were performed comparing the frequency of basic level descriptors versus superordinate level descriptors for both action and scene category terms. The results show that action terms were more frequently reported at the basic level (2,028 out of 2,527) than at the superordinate level (554 out of 2527)(χ^2 (1) $\geq$ 925.14, $p < .001$). The same results is shown when comparing basic level scene terms (241 out of 253) with superordinate scene terms (12 out of 253)(χ^2 (1) $\geq$ 207.28, $p < .001$). This shows that the basic level advantage was present for both scene and action categories. Furthermore, a greater frequency of basic level action terms (2251 out of 2770) were used to describe the scene images compared to basic level scene category terms (519 out of 2770)(χ^2 (1) $\geq$ 207.28, $p < .001$). This indicates that the linguistic description of our scene images was primarily focused on the person and the action that was being performed in the scene, which supports the findings regarding the person bias in eye movements (Fletcher-Watson, Findlay, Leekam, & Benson, 2008; Humphrey & Underwood, 2010).

Additionally, consistent with Tversky and Hemanway's (1983) findings, superordinate level scene descriptors, like "Indoors" and "Outdoors," were almost never used to label scene images (35 out of 4,369 responses, $M = 0.80\%$), however superordinate action terms were used fairly regularly (577 out of 4,369 responses, $M = 13.21\%$).

Further analyses were computed to examine the detail in which participants described scenes. First, since each image provided a scene category and an action, a comparison was made between the percentage of scene and action terms given that the participant used an action or scene term, respectively. The results show that the conditional probability for actions given a scene term ($M = .56$, $SD = 0.45$) was greater than the conditional probability for scenes given an

action term (M = .17, SD = 0.17)(t (311) = 18.27, p < .001). This shows that when a response included an action descriptor, it was less likely to contain a superordinate or basic level scene term. However, if the image label contained a scene category, then it was likely to also contain an action term. This is consistent with the overall frequency in which action terms were used over scene category terms.

After reviewing participant's responses, it was found that 33% of the responses contained a reference to an object in the image. The object information may imply the action that is depicted in the image. If true, then viewers might be less likely to provide an action category term when an object term is used, because the object term could take the place of the action. Therefore, it was hypothesized that object terms would be used more often with a scene category term than an action category term. A t-test was used to compare the conditional probability of using an object term given that an action term was used compared to when a scene term was used. Contrary to the hypothesis, the conditional probability for reporting an object was greater when given an action term (M = .35, SD = .34) compared to a scene category term (M = .19, SD = .37)(t (311) = 9.51, p < .001). Specifically, action descriptions tended to specify the object being manipulated, whereas scene descriptions were less likely to identify an object in the image.

Discussion

Experiments 1 and 2 found evidence that was inconsistent with Rosch et al.'s (1976) basic level theory, showing that the superordinate scene category was recognized before the basic level. However, based on brain imaging studies (Löw et al., 2003), it was hypothesized that the previous studies showing a basic level advantage were based on a linguistic categorization system, which is biased to communicating informative scene information. Therefore, it was hypothesized that if participants were asked to describe an image, their use of

their linguistic categorization system would lead them to report more detailed, informative scene constructs. Specifically, the basic level descriptors would be used more than superordinate descriptors (Morris & Murphy, 1990; Rosch et al., 1976; Tversky & Hemenway, 1983). Conversely, an alternative hypothesis was that the first category activated during perception of a scene would be the primary category verbally reported. This suggested that superordinate level categorization would occur more often than basic level categorization for both scenes and objects (Large, Kiss, & McMullen, 2004; Loschky & Larson, 2010; Macé, Joubert, Nespoulous, & Fabre-Thorpe, 2009; Rogers & Patterson, 2007) and presumably also for actions.

An additional hypothesis was that action terms would be reported more often than scene category terms, due to the person bias. This hypothesis was independent of the categorization level hypothesis since it would be possible to have any of four possible outcomes (more basic level scene terms, more superordinate level scene terms, more basic level action terms, or more superordinate level action terms).

Experiment 3 clearly showed an advantage for basic level action descriptions over other image descriptors. Overall, these results show the exact opposite pattern than was found in Experiments 1 and 2. Namely, scene image descriptions using the linguistic system produced categories that were more finely detailed than coarse. Furthermore, basic level action terms were more likely to be used than basic level scene terms, indicating that the detailed descriptions were more focused on the people in the scene images and the tasks they performed. Likewise, basic level action and scene category terms were more frequently used than superordinate action and scene category terms. Furthermore superordinate scene category terms were virtually absent, consistent with Tversky and Hemenway (1983). This shows that the linguistic categorization system is more concerned with the pragmatics of providing informative terms regarding the

scene, as predicted by the basic level theory for objects, scenes, and actions (Gosselin & Schyns, 2001; Jolicour, Gluck, & Kosslyn, 1984, Morris & Murphy, 1990; Rifkin, 1985; Rosch et al., 1976; Tversky & Hemenway, 1983).

Chapter 5 - General Discussion

To summarize, Experiment 1 showed that superordinate scene categorization occurs prior to basic level scene categorization, which also occurs prior to basic level action categorization. Although the sensitivity results from Experiment 2 were not as strong as those presented in Experiment 1, the data still converged with the conclusions from Experiment 1. This supports the previous findings that rapid image categorization occurs in a global-to-local fashion (Loschky & Larson, 2010; Macé, Joubert, Nespoulous, & Fabre-Thorpe, 2009; Poncet, Reddy, & Fabre-Thorpe, 2012; Rogers & Patterson, 2007). Specifically, a scene image is recognized first according to its membership in a superordinate category. When given more time, a finer basic level distinction can be made, while discriminating the basic level action requires even more time. This effect directly contradicts the research showing that the basic level is categorized prior to the superordinate (Rosch et al., 1976; Jolicour, Gluck, & Kosslyn, 1984; Murphy & Smith, 1982). However, when participants were asked to describe the images used in Experiments 1 and 2, basic level descriptors were used more often than superordinate descriptors, which supported basic level theory. Additionally, basic level descriptions were based on the action more often than the scene category, which supports the research based on the person bias.

The effects shown in Experiments 1 and 2 appear to oppose the effect shown in Experiment 3. However, these contradictory effects may be representative of two distinctive sets of processes. Namely, the superordinate advantage shown in Experiments 1 and 2 is

representative of early perceptual processes, whereas the basic level advantage is representative of a later, linguistic categorization processes. Thus, these two effects are not opposing, but identify important psychological processes that occur at different stages of scene categorization. This line or reasoning is analogous to the discussion within the text comprehension literature regarding the importance of the task of the reader, reader's goals, and other variables used to examine apparent contradictory findings between on-line versus off-line comprehension (Rapp & Mensink, 2011). Regarding the current studies, the apparently contradictory results are, in fact, highly informative of the categorization processes which occur over time. Namely, Experiments 1 and 2 represent early perceptual scene categorization processes used to comprehend the scene image when given a specific categorization task. Conversely, Experiment 3 represents the later linguistic categorizations which are the product of comprehending and communicating informative parts or dimensions of the scene image.

This distinction between early perceptual categorization and later linguistic categorization may be explained by different brain mechanisms. Specifically, an MEG study proposed that early activation in the occipital-temporal junction was involved in perceptual categorization, whereas later activation in the left temporal lobe was specific to linguistic categorization (Löw, et al, 2003). This distinction is consistent with behavioral results indicating that early perceptual categorization first makes distinctions between superordinate categories, whereas linguistic categorization focuses on somewhat finer distinctions. Namely, superordinate categorization of objects and scenes is best at the earliest stages of processing (Loschky & Larson, 2010; Macé, Joubert, Nespoulous, & Fabre-Thorpe, 2009; Rogers & Patterson, 2007), but the basic level is best later in processing when naming or describing objects and scenes (Rosch et al., 1976; Tversky & Hemenway, 1983).

The results also showed that basic level action categorization required 2-3 fixations on the scene image, whereas scene categorization required only one fixation at both the superordinate and basic levels. This shows that distinguishing between basic level actions occurring in the same scene context requires that the viewer foveate additional information in the scene, specifically the object involved in the action, and the pictured person's hands manipulating the object. These eye movement results were consistent with the categorization task affecting the scene information that was fixated. Namely, basic level action categorization resulted in a greater percentage of fixations to the hands and object compared to both scene categorization tasks. Likewise, the eye movement results are also somewhat consistent with the person bias (Fletcher-Watson, Findlay, Leekam, & Benson, 2008; Humphrey & Underwood, 2010; Zwickel & Võ, 2010), which suggested that the eyes should fixated similar locations in the scene. Evidence of this is shown by the similar percentage of fixations and domain-relative ratios on the head and person across all three categorization tasks. In addition, the results showed that top-down information, specifically the categorization task, biased the scene information that was fixated, consistent with previous research (DeAngelus & Pelz, 2009; Malcolm, Nuthmann, & Schyns, 2011; Yarbus, 1967).

Theoretical Contributions and Implications

These experiments begin to lay the ground work for integrating what is known about cognitive processing occurring within a single fixation, namely scene gist processes, with the higher-level processes occurring over multiple fixations during event perception. Most importantly, Experiment 1 showed that the scene category, at both the superordinate and basic level, is recognized prior to the basic level action. This suggests that the scene category could be the first semantic information that is integrated into a new event model. Event Segmentation

Theory (EST) has described the circumstances in which an event model is updated, as well as proposing a neurological circuit that allows this updating process to occur (Zacks et al., 2007). In addition, the Event Indexing Model (EIM) identifies a set of dimensions within text and visual narratives that are attended to and lead to an increase in the likelihood of an event model being updated (Magliano, Miller, & Zwaan, 2001; Magliano & Zacks, 2011; Zacks, Speer, & Reynolds, 2009; Zwaan, Langston, & Graesser, 1995; Zwaan & Radvansky, 1998). While the EIM does not describe how an event model is formed, EST hypothesizes that that low-level visual and auditory information are combined, in working memory, to form an event model (Kurby & Zacks, 2007; Zacks, Speer, Swallow, Braver, & Reynolds, 2007). EST also predicts that top-down information, including previous experiences in similar situations (Avrahami & Kareev, 1994), will also influence the construction of the model (Kurby & Zacks, 2007; Zacks, Speer, Swallow, Braver, & Reynolds, 2007). However, the current study on rapid scene and action categorization delineates key semantic information acquired during the first fixation and its time course of acquisition. It also shows how immediately following fixations gather information to build upon that foundation for the event model. Thus, the current study lays the theoretical and empirical ground work for how an event model is constructed while looking at a scene. The scene category, or the spatiotemporal setting of a visual narrative, is likely recognized first, after which, the protagonist's action is recognized. This should allow the viewer to infer the character's goal(s), based on their prior knowledge, and, according to EST, make predictions about what should occur next in the visual event.

Future studies will need to examine both the bottom-up effects of how scene gist is incorporated in an event model, and the top-down effects of the event model on scene gist. To our knowledge, this is the first study to investigate how both the scene category and an

action/event taking place in the scene is recognized over the course of both a single and multiple fixations. A related question is how consistent scene constructs may contribute to action categorization. For instance, is an action easier to categorize in an appropriate scene context or does the scene context have no effect on action categorization? Conversely, knowledge of the processes occurring during event perception can also be incorporated into our understanding of how scene gist is recognized. For example, EST describes the variation in cognitive processing of perceptual information over the course of an event. The bottom-up and top-down processes involved in scene gist recognition can be compared by presenting scene image stimuli either in the form of a coherent picture story or in randomized order. The randomized order condition would present the same perceptual information as the picture story condition. However, by randomizing the image order, any higher level structure regarding the observed actions could not be used to predict the action in the next scene image. Conversely, in the coherent story condition, being able to accurately predict the action or scene category that is about to be presented should prime that information, resulting in faster recognition of the to-be-presented scene.

By presenting scene image information in the form of a visual narrative, then it is hypothesized that the narrative should begin to be parsed into separate events, similar to the segmentation processes reported for reading (Speer, Zacks, & Reynolds, 2007; Zacks, Speer, Reynolds, 2009), pictorial narratives (Gernsbacher, 1985; Gernsbacher, Varner, & Faust, 1990), and film (Magliano & Zacks, 2011; Zacks, Speer, & Reynolds, 2009). Theories of event segment on written narratives have identified some of the factors that increase the probability of perceiving a new event. These factors are also used when viewing visual narratives. These perceptual factors that identify the beginning of a new event may contain information regarding

the action being performed than compared to actions that typically occur during the middle or the end of an event. Previous research comparing memory for the perceptual information at the event boundary versus the middle of an event has found that recognition memory is better for the boundary (Newston & Engquist, 1976; Swallow et al., 2009; 2010). Information presented at the boundaries also contains critical information for summarizing the visual narrative (Schwan & Garsoffky, 2004). Summaries produced from a visual narrative only containing the event boundaries did not differ from summaries produced from viewing the entire film (Schwan & Garsoffky, 2004). These studies suggest that the perceptual information viewed during the event boundary is more diagnostic of an action than perceptual information during the middle of an event. This diagnostic information may include the body posture, as noted by Glanemann (2007). One can imagine a "Kicking" action where one raises their foot behind themself, in preparation to kick, as seen on the left side of Figure 4.1. This may be a better example of kicking compared to the same action where the foot is slightly in front of the body, as shown on the right side of Figure 4.1.

Figure 5.1 Diagnostic and non-diagnostic body posture information for a "Kicking" action.

Similarly, there may be differences in scene and action categorization for frames that are from an event boundary versus frames from a non-boundary. Future research could compare

rapid scene and action categorization using scene images from event boundaries versus non-event boundaries. One hypothesis would predict that scene categorization would be better during the event boundary since the perceptual information encoded during that time is being used to construct a predictive event model. Conversely, it could be argued that having a predictive event model would allow a viewer to have an expectation for the scene gist that will be presented next. In this case, it would be predicted that scene category would be recognized faster during the non-boundary, which has a fully constructed event model that can prime the gist, compared to the event boundary, which is in the process of constructing the event model, when the gist may have to be acquired from scratch based only on bottom-up input.

The present study showed how the categorization task affected sensitivity within a single fixation and multiple fixations. Namely, these categorization tasks resulted in vastly different time-courses for scene category and action category information. Likewise, top-down effects were evident in the fixation locations for the three categorization tasks. For action categorization, a greater percentage of fixations were on the object and hands than for the two scene categorization tasks, indicating the importance of these features for categorizing the action. These two features were, on average, contained in the visual periphery, thus an eye movement was required to foveate this information. Therefore, a ceiling level of categorization sensitivity required 2-3 fixations. However, one possible limitation to this study is that the visual eccentricity of the objects and hands were not controlled. Given that object and hand information appeared to be required for fully accurate action categorization, the time-course for categorizing the action may vary as a function of the visual eccentricity at which that information is presented. Thus, if the objects and hands were presented at the fovea at the onset of the trial, then no saccade would have to be planned and executed, presumably resulting in ceiling

performance occurring on the first fixation on the scene image. However, the current results can be argued to be representative of real-world viewing, since it is unlikely that the spatial location of an object and the hands would always fall on the fovea at the onset of viewing a scene. It is more likely that an object will fall outside of the fovea and central visual field, simply based on the percentage of viewable area that these two regions subtend relative to the visual periphery.

Further research could examine the relative importance of scene context and object consistency on the time-course of action categorization. Previous research has found that the scene category interacts with object perception and vice versa (Davenport & Potter, 2004; Palmer, 1974). For example, a loaf of bread in a kitchen is easier to categorize than a mailbox in the same kitchen, even though the loaf of bread and mailbox share perceptual features (i.e. shape). Thus, object recognition is affected by the scene context in which it is presented. However, no research to our knowledge has been done to investigate the relative importance of the scene background on action categorization. The same set of action scene images could be presented in context versus with no context. The scene images in the no context condition could be PhotoShopped such that the scene background would be removed from the image, leaving only the person and the object that was being manipulated. Conversely, the context condition would present the entire scene image with the scene background, person, and object. It could be hypothesized that the scene context plays a role in categorizing the action at early processing times, based on the current study's findings showing that the scene category is categorized prior to the action. Research by Barker (1968) indicates that actions are typically performed in certain settings. For example, cooking typically occurs more often in the Kitchen than in a Bedroom. Thus, knowledge of the location may make categorizing an action easier if it is consistent with the setting. Relating to the current action and scene categories, mowing and raking are limited to

yards, thus the setting may have helped to identify those actions. Conversely, actions which are less typical for a given setting show have less of an impact of the scene category on action categorization. For example, it may be more difficult to categorize the actions of eating and reading in a park since both of these actions can occur in multiple scene categories. Thus, the scene category may be having an impact on the viewer's ability to categorize actions. For instance, potential action categories could be eliminated from consideration based on an early superordinate scene category distinction. Knowing that a scene image is an "Indoor" scene may help to eliminate "outdoor" action categories such as raking. Therefore, lacking the scene context would produce worse action categorization at early stages of scene processing than if the action was presented in context. Conversely, an alternative hypothesis would predict that action categorization is better with no scene context. By presenting action scenes in no context, there is very little perceptual information to interfere with the encoding and processing of the person and object via crowding (Chung, Levi, & Legge, 2001; Pelli, Palomares, & Majaj, 2004), relative to the context condition. Additionally, the person and object would not have to be parsed from the scene background, therefore all attentional resources could be focused on perceptual processing of the body and object producing better performance. Evidence consistent with the latter prediction was shown by Davenport and Potter (2004) where object naming was best when presented without a scene context. Lastly, there may be no difference between action categorization in the appropriate scene context versus in no-context. This hypothesis is based on the findings of Hollingworth and Henderson (1998; 1999) suggesting that processing of objects is functionally isolated from scene processing (Hollingworth & Henderson 1998; 1999). This was based on findings showing that object recognition sensitivity was no different when

presented in an appropriate scene context versus in an inappropriate scene context. Therefore, it is an open question as to how scene context influences action categorization.

Another critical visual cue that was missing from the current experiments is motion. In the real world, an action rarely occurs without motion, and research has shown that motion onset captures attention (Abrams & Christ, 2003; Mital, Smith, Hill, & Henderson, 2011) and motion helps to identify the onset of a new event (Speer, Swallow, & Zacks, 2003; Smith 2012; Zacks et al. 2006). Therefore, will motion information in short video clips have an impact on the time-course of action and scene categorization? The onset of motion would likely capture attention, resulting in the first eye movement being made toward the moving stimulus. This would result in quicker latencies to fixate the moving features (object and/or person) relative to a static scene image. Likewise, if this information is critical for action categorization, then it would be expected that action and scene categorization performance would converge at the second fixation. Alternatively, attentional capture to motion may bias the perceptual processing of the image to only the motion/action information. It may be possible that this attentional capture could interfere with the processing of the scene category, with useful scene information contained in the periphery (Larson & Loschky, 2009), which would result in action categorization occurring prior to scene categorization.

The perceptual similarity of action categories is a potential factor that could affect the time course of action categorization. If action categories are more perceptually similar, then more detailed perceptual information should be needed to discriminate between them. Conversely, if action categories are more perceptually distinct, then diagnostic features may be able to quickly distinguish one category from another. For example, in the present study distinguishing between "Typing" and "Phoning" in an office may have only been discriminable

by fixating the keyboard or phone. However, actions like "Typing" and "Greeting a visitor" in the office may have been more discriminable by the body position. Typing usually requires one to be sitting in a chair, whereas greeting a visitor usually requires the person to be standing. Using this diagnostic feature to distinguish between actions may have increased the percentage of fixations made to the body. Conversely, if that diagnostic feature is sufficiently large in the visual field, then no fixation would be required to encode and categorize the scene, but instead it could be recognized using peripheral vision (Larson & Loschky, 2009).

The working hypothesis in the current studies is that scene gist, gathered in the first fixation on a scene, is the first piece of semantic information that is used to construct an event model. The current research has provided evidence consistent with this hypothesis based on the time course of rapid scene and action categorization, but these studies have not provided evidence showing how scene gist is utilized in constructing an event model. Future research will be needed to investigate that question. If the hypothesis that scene gist is used to construct an event model is supported, then the "bottom-up" processes involved in scene gist can be incorporated into the emerging theories and research on event model construction and event perception. Likewise, the "top-down" knowledge contained within event models can then help to guide theories of scene gist recognition and scene perception.

References

Avrahami, J., & Kareev, Y. (1994). The emergence of events. *Cognition, 53,* 239 – 261.

Bacon-Macé, N., Macé, M. J., Fabre-Thorpe, M., & Thorpe, S. J. (2005). The time course of visual processing: Backward masking and natural scene categorization. *Vision Research, 45,* 1459 – 1469.

Barker, R. G. (1968). *Ecological Psychology.* Stanford, CA: Standford University Press.

Carmi, R., & Itti, L. (2006). Visual causes versus correlates of attentional selecting in dynamic scenes. *Vision Research, 46,* 4333 – 4345.

Chung, S. T. L., Levi, D. M., & Legge, G. E. (2001). Spatial frequencies and contrast properties of crowding. *Vision Research, 41*(14), 1833 – 1850.

Daventport, J. L., & Potter, M. C. (2004). Scene consistency in object and background perception. *Psychological Science, 15*(8), 559 – 564.

DeAngelus, M., & Pelz, J. P. (2009). Top-down control of eye-movements: Yarbus revisited. *Visual Cognition, 17*(6), 790 – 811.

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

Eckstein, M. P., Drescher, B. A., & Shimozaki, S. S. (2006). Attentional cues in real scenes, saccadic targeting, and Bayesian priors. *Psychological Science, 17*(11), 973 – 980.

Fei-Fei, L., Iyer, A., Koch, C., & Perona, P. (2007). What do we perceive in a glance of a real-world scene? *Journal of Vision, 7*(1):10, 1 – 29.

Fletcher-Watson, S., Findlay, J. M., Leekam, S. R., & Benson, V. (2008). Rapid detection of person information in a naturalistic scene. *Perception, 37,* 571 – 583.

Gernsbacher, M. A. (1985). Surface information loss in comprehension. *Cognitive Psychology, 17*(3), 324 – 363.

Gernsbacher, M. A., Varner, K. R., & Faust, M. E. (1990). Investigation differences in general comprehension skill. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*(3), 430 – 445.

Glanemann, R. (2007). Too see or not to see – Action scenes out of the corner of the eye. Unpublished doctoral dissertation, Westfälischen Wilhelms-Universität, Münster.

Gosselin, F., & Schyns, P. G. (2001). Why do we SLIP to the basic level? Computational constraints and their implementation. *Psychological Review, 108*(4), 735 – 758.

Green, M. R., & Oliva, A. (2009). The briefest of glances: The time course of natural scene understanding. *Psychological Science, 20*(4), 464 – 472.

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

Henderson, J. M. (1992). Visual attention and eye movement control during reading and picture viewing. In K. Rayner (Ed.), Eye movements and visual cognition: Scene perception and reading (pp. 260 – 283). New York: Springer-Verlag.

Henderson, J. M., McClure, K. K., Pierce, S., & Schrock, G. (1997). Object identification without foveal vision: Evidence from an artificial scotoma paradigm. *Perception and Psychophysics, 59*(3), 323 – 346.

Henderson, J. M., Pollatsek, A., & Rayner, K. (1987). Effects of foveal priming and extrafoveal preview on object identification. *Journal of Experimental Psychology: Human Perception and Performance, 13*(3), 449 – 463.

Hollingworth, A., & Henderson, J. (1998). Does consistent scene context facilitate object perception? *Journal of Experimental Psychology: General, 127*(4), 398 – 415.

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

Humphery, K., & Underwood, G. (2010). The potency of people in pictures: Evidence from sequences of eye fixations. *Journal of Vision, 10*(10):19, 1 – 10.

Iacoboni, M., Molnar-Szakacs, I., Gallese, V., Buccino, G., Mazziotta, J. C., & Rizzolatti, G. (2005). Grasping the intentions of others with one's own mirror neurons. *PLoS Biology, 3*(3), 529 – 535.

Joubert, O. R., Rousselet, G. A., Fize, D., & Fabre-Thorpe, M. (2007). Processing scene context: Fast categorization and object interference. *Vision Research, 47*(26), 3286 – 3297.

Jolicoeur, P., Gluck, M. A., Kosslyn, S. M. (1984). Pictures and names: Making the connection. *Cognitive Psychology, 16*, 243 – 275.

Kurby, C. A., & Zacks, J. M. (2007). Segmentation in the perception and memory of events. *Trends in Cognitive Science, 12*(2), 72 – 79.

Large, M-E., Kiss, I., McMullen, P. A. (2004). Electrophysiological correlates of object categorization: Back to basics. *Cognitive Brain Research, 20*, 415 – 426.

Larson, A. M., & Loschky, L. C. (2009). The contributions of central versus peripheral vision to scene gist recognition. *Journal of Vision, 9*(10):6, 1 – 16.

Lichtenstein, E. H., & Brewer, W. F. (1980). Memory for goal-directed events. *Cognitive Psychology, 12*(3), 412 – 445.

Loschky, L. C., Hansen, B. C., Sethi, A. & Pydimarri, T. N. (2010). The role of higher order image statistics in scene gist recognition. *Attention, Perception, & Psychophysics, 72*(2), 427 – 444.

Loschky, L. C., Sethi, A., Simons, D. J., Pydimarri, T. N., Ochs, D., & Corbeille, J. L. (2007). The importance of information localization in scene gist recognition. *Journal of Experimental Psychology: Human Perception and Performance, 33*(6), 1431 – 1450.

Loschky, L. C., & Larson, A. M. (2010). The natural/man-made distinction is made prior to basic-level distinctions in scene gist processing. *Visual Cognition, 18*(4), 513 – 536.

Macé, M. J-M., Joubert, O.R., Nespoulous, J-L., & Fabre-Thorpe, M. (2009). Time-course of visual categorizations: You spot the animal faster than the bird. *PloSONE, 4*(6), e5927.

Macmillan, N. A., & Creelman, D. C. (2005). Detection theory: A user's guide (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Magliano, J. P., Miller, J., & Zwaan, R. A. (2001). Indexing space and time in film understanding. *Applied Cognitive Psychology, 15*, 533 – 545.

Magliano, J. P., & Zacks, J. M. (2011). The impact of continuity editing in narrative film on event segmentation. *Cognitive Science, 35*, 1 – 29.

Malcolm, G. L., Nuthmann, A. & Schyns, P. G. (2011). Ordinate and subordinate level categorizations of real-world scenes: An eye movement study. Poster presented at the 11th annual meeting of the Vision Sciences Society, Naples, FL.

Mandler, J. M., & Johnson, N. S. (1977). Remembrance of things parsed: Story structure and recall. *Cognitive Psychology, 9*, 111 – 151.

McPherson, W. B., & Holcomb, P. J. (1999). An electrophysiological investigation of semantic priming with real world pictures. *Psychophysiology, 36*, 53 – 65.

Meyer, A. S., Sleiderink, A. M., & Levelt, W. J. M. (1998). Veiwing and naming objects: Eye movements during noun phrase production. *Cognition, 66,* B25 – B33.

Morris, M. W., & Murphy, G. L. (1990). Converging operations on a basic level in event taxonomies. *Memory & Cognition, 18*(4), 407 – 418.

Murphy, G. L., & Smith, E. E. (1982). Basic-level superiority in picture categorization. *Journal of Verbal Learning & Verbal Behavior, 21*(1), 1 – 20.

Nolan, C., Thomas, E., & Roven, C. (Producers), & Nolan, S. (Director). (2008). The Dark Knight [Motion picture]. U.S.: Legendary pictures.

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

Palmer, S. E. (1975). The effects of contextual scenes on the identification of objects. *Memory & Cognition, 3*(5), 519 – 526.

Peli, E., & Geri, G. A. (2001). Discrimination of wide-field images as a test of a peripheral-vision model. *Journal of Optical Society of America, A, Optics, Image Sciences & Vision, 18*(2), 294 – 301.

Pelli, D. G., Palomares, M., & Majaj, N. J. (2004). Crowding is unlike ordinary masking: Distinguishing feature detection and integration. *Journal of Vision, 4*(12), 1136 – 1169.

Pollatsek, A., Rayner, K., Collins, W. E. (1984). Integrating pictorial information across eye movements. *Journal of Experimental Psychology: General, 113*(3), 426 – 442.

Poncet, M., Reddy, L., & Fabre-Thorpe, M. (2012). Presentation time does not affect the superordinate advantage in ultra-rapid categorization. Paper presented at the 12th annual meeting of the Vision Sciences Society. Naples, FL.

Portilla, J., & Simoncelli, E. P. (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision, 40,* 49 -71.

Potter, M. C. (1976). Short-term conceptual memory for pictures. *Journal of Experimental Psychology: Human Learning and Memory, 2*(5), 509 – 522.

Rapp, D. N. & Mensick, M. C. (2011). Focusing effects from online and offline reading tasks. In M. T. McCrudden, J. P. Magliano, & G. Schraw (Eds.), *Text Relevance and Learning from Text.* Greenwich, CT: Information Age Publishing.

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

Rifkin, A. (1985). Evidence for a basic level in event taxonomies. *Memory & Cognition, 13*(6), 538 – 556.

Rizzolatti, G., Riggio, L., Dascola, D., & Umilta, C. (1987). Reorienting attention across the horizontal and vertical meridians: Evidence in favor of a premotor theory of attention. *Neuropsychologia, 1-A,* 31-40.

Rogers, T. T., & Patterson, K. (2007). Object Categorization: Reversals and explanations of the basic-level advantage. *Journal of Experimental Psychology: General, 136*(3), 451 – 469.

Rousselet, G. A., Fabre-Thorpe, M., & Thorpe, S. J. (2002). Parallel processing in high-level categorization of natural images. *Nature neuroscience, 5,* 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.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology, 8*(3), 382 – 439.

Ross, J., Morrone, M. C., Goldberg, M. E., & Burr, D. C. (2001). Changes in visual perception at the time of saccades. *Trends in Neurosciences, 24*(2), 113 – 121.

Strasburger, H., Rentschler, I., & Juttner, M. (2011). Peripheral vision and pattern recognition: A review. *Journal of Vision, 11*(5):13, 1 – 82.

Séré, B., Marendaz, C., & Hérault, J. (2000). Nonhomogeneous resolution of images of natural scenes. *Perception, 29,* 1403 – 1412.

Smith, T. J. (2012, May). The relationship between overt attention and event perception during dynamic social scenes. Paper presented at the 12th annual meeting of the Vision Sciences Society. Naples, FL.

Speer, N. K., Zacks, J. M., & Reynolds, J. R.(2007). Human brain activity time-locked to narrative event boundaries. *Psychological Science, 18*, 449 – 455.

Stinikova, T., Kuperberg, G., & Holcomb, P. J. (2003). Semantic integration in videos of real-world events: An electrophysiological investigation. *Psychophysiology, 40*, 160 – 164.

Swallow, K. M., Zacks, J. M., & Abrams, R. A. (2009). Event boundaries in perception affect memory encoding and updating. *Journal of Experimental Psychology: General, 138*(2), 236 – 257.

Thorndyke, P. W. (1977). Cognitive structures in comprehension and memory of narrative discourse. *Cognitive Psychology, 9*, 77 – 110.

Thorpe, S. J., Gegenfurtner, K. R., Fabre-Thorpe, M., & Bulthoff, H. H. (2001). Detection of animals in natural images using far peripheral vision. *European Journal of Neuroscience, 14*(5), 869 – 876.

Tversky, B., & Hemenway, K. (1983). Categories of environmental scenes. *Cognitive Psychology, 15*, 121 – 149.

Volkmann, F. C. (1986). Human visual suppression. *Vision Research, 26*(9), 1401 – 1416.

West, W. C., & Holcomb, P. J. (2002). Event-related potentials during discourse-level semantic integration of complex pictures. *Cognitive Brain Research, 13*, 363 – 375.

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

Zacks, J. M. (2004). Using movement and intentions to understand simple events. *Cognitive Science, 28*, 979 – 1008.

Zacks, J. M., Kumar, S., Abrams, R. A., & Mehta, R. (2009). Using movement and intentions to understand human activity. *Cognition, 112*, 201 – 216.

Zacks, J. M., Kurby, C. A., Eisenberg, M. L., & Haroutunian, N. (2011). Prediction error associated with the perceptual segmentation of naturalistic events. *Journal of Cognitive Neuroscience, 23*(12), 4057 – 4066.

Zacks, J. M., & Tversky, B. (2001). Event structure in perception and conception. *Psychological Bulletin, 127*, 3 – 21.

Zacks, J. M., Tversky, B., & Iyer, G. (2001). Perceiveing, remembering, and communicating structure in events. *Journal of Experimental Pscyhology: General, 130*, 29 – 58.

Zacks, J. M, Speer, N., & Reynolds, J. R. (2009). Segmentation in reading and film comprehension. *Journal of Experimental Psychology: General, 138*, 307 – 327.

Zacks, J. M., Speer, N, K., Swallow, K. M., Braver, T. S., & Reynolds, J. R. (2007). Event perception: A mind/brain perspective. *Psychological Bulletin, 133*, 273 – 293.

Zacks, J. M., Speer, N. K., Swallow, K. M., & Maley, C. J. (2010). The brain's cutting-room floor: Segmentation of narrative cinema. *Frontiers in Human Neuroscience, 4*:168

Zwaan, R. A., Langston, M. C., & Graesser, A. C. (1995). The construction of situation models in narrative comprehension: An event-indexing model. *Psychological Science, 6*(5), 292 – 297.

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin, 123*(2), 162 – 185.

Zwickel, J., & Võ, M. (2010). How the presence of persons biases eye movements. *Psychonomic Bulletin & Review, 17*(2), 257 – 262.