WHAT KINDS OF VISUAL INFORMATION ARE USED TO MAKE ULTRA-RAPID CATEGORY DECISIONS?

by

JENNIFER LEIGH SOLBERG

(Under the Direction of James M. Brown)

ABSTRACT

The present study was designed to determine what types of stimulus input are required to make rapid category decisions. In 3 experiments, this question was addressed using ultra-rapid visual categorization (URVC) in conjunction with backward masking. The first experiment showed typical URVC accuracy with unmasked photographs of natural scenes and established masked categorization performance with these stimuli. Experiment 2 tested the hypothesis that line drawing information is sufficient to perform rapid categorizations. It was found that URVC accuracy was comparable, if not higher, with line drawings than photographs. The third experiment tested the hypothesis that a particular range of spatial frequency information is used in URVC. Results indicated participants were more accurate with low band passed than high passed scenes. Taken together, these results indicate high contrast, global form information may be necessary for URVC.

INDEX WORDS: Ultra-Rapid Visual Categorization, Natural Scenes, Line Drawings, Spatial Frequency

WHAT KINDS OF VISUAL INFORMATION ARE USED TO MAKE ULTRA-RAPID CATEGORY DECISIONS?

by

JENNIFER LEIGH SOLBERG

B.A., Davidson College, 1998

M.S., The University of Georgia, 2000

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GA

2004

WHAT KINDS OF VISUAL INFORMATION ARE USED TO MAKE ULTRA-RAPID CATEGORY DECISIONS?

by

JENNIFER LEIGH SOLBERG

Major Professor: James M. Brown

Committee: Richard Marsh
Zachary Estes
Billy R. Hammond, Jr.
Brett Clementz

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
May 2004

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

What Kinds of Visual Information Are Necessary to Make Ultra-Rapid Category Decisions?

In our daily experience, we take for granted that our visual world is readily available to us. Objects in our environment are typically still, allowing us ample time to examine them fully. Movement and change certainly occur, but not usually at speeds with which we cannot follow with our eyes. Occasionally, however, our environment requires us to make split-second decisions about things we see only briefly. What aspects of stimuli are used to make these rapid decisions? The present study was designed to address this question; in a series of experiments, the types of stimulus information used in rapid decisions was investigated using the ultra-rapid visual categorization paradigm in conjunction with backward masking.

Throughout the history of vision research, many models have been proposed to account for our ability to extract information about objects in our environment from a two-dimensional visual array. The majority of these models posit that visual processing takes place through a series of stages through which retinal information is extracted, manipulated and combined, ultimately resulting in a three-dimensional representation of our surroundings. The first stages in this process are of particular importance. They form the basis of our representation, and it may be argued they carry the most crucial information for making rapid decisions about the nature of objects (Delorme, Richard, & Fabre-Thorpe, 1999). This study investigated the types of stimulus information used in the first stages of visual processing as specified by two current models of

object recognition: Biederman's (1987) Recognition-By-Components (RBC) and a spatial frequency-based approach (e.g., DeValois & DeValois, 1988; Shyns & Oliva, 1994). This was accomplished by comparing observers' accuracy on a rapid categorization task when they are presented with photographs and model-specific versions of natural scenes.

The paradigm used in these experiments is a variant on the "ultra-rapid visual categorization" (URVC) task first described by Thorpe, Fize, and Merlot (1996). In URVC, participants are presented natural scenes very briefly and asked to make category judgments about their contents. The results of experiments using this paradigm, described in more detail below, suggest sufficient information for object categorization is available in the first milliseconds of visual processing. In these experiments, the URVC paradigm was combined with backward masking to further constrain observers' processing time. By limiting visual processing to its earliest stages, the types of stimulus information necessary to perform rapid categorizations can be studied. In these experiments, two main questions were addressed regarding the nature of early visual representations. First, do edge-based representations of natural scenes provide sufficient information to make rapid category decisions? Several current models of object recognition, such as Biederman's (1987) Recognition-By-Components model, would suggest this is the case. Secondly, which, if any, particular spatial frequencies inform the earliest visual processes? Most spatial frequency-based models of object recognition imply that low spatial frequencies, rather than higher ones, provide this information (e.g., DeValois & DeValois, 1988).

Both RBC and spatial frequency-based models make specific claims about which aspects of stimuli are used in the first milliseconds of visual processing. Biederman's highly influential model holds that object recognition begins with an edge extraction process, which generates a line drawing representation of the visual array. From this edge-based representation, primitive

object components (geons) are identified and combined to form structural representations of objects. Following from this model, a line drawing representation of an object should provide sufficient information for recognition. Furthermore, if line drawings are the foundation of our visual representation, observers should perform comparably on visual tasks when presented line drawings and more natural stimuli, such as photographs. Indeed, a comparison of performance with line drawings and photographs showed no differences between these types of stimuli in identification and name verification tasks (Biederman & Ju, 1988).

Spatial frequency-based approaches to vision also posit that certain types of visual information are processed earlier than others. These models are based on the physiological distinction between the two primary visual pathways, the magnocellular (M) and parvocellular (P) pathways. The faster M pathway is known to carry low spatial frequency information whereas the P pathway is responsible for processing high spatial frequencies (Merigan & Maunsell, 1993). Experimental evidence indicates lower spatial frequencies, which correspond to coarse luminance changes and general object form, are used early on in object recognition to develop a rough but stable representation of the visual field. Later, higher spatial frequency information is incorporated into this representation, adding object details used in tasks such as recognition and categorization (e.g., Parker, Lishman & Hughes, 1992; 1997). Recent evidence suggests low spatial frequency information is not always given priority by the visual system, however, and that task demands mediate processing of spatial scales. Oliva and Shyns (1997) found high frequency information in natural scenes may be accessed first when observers are sensitized to that particular range of spatial frequencies.

These models, as well as several others (e.g., Marr, 1982), specify the types of stimulus information used in the first stages of object recognition. The present study tests the predictions

of these models by comparing observers' rapid categorization performance with photographs of scenes, which are similar to our actual visual input, and versions that contain aspects specified by the models as primary. If, as RBC holds, object recognition is based on an initial line drawing representation, categorization accuracy was expected to be equivalent between photograph and line drawing versions of scenes. If low spatial frequency information is used in early image processing, low band passed versions of scenes would provide similar information to a full-spectrum photograph when processing time is limited. This model predicts accuracy will be higher with low band passed scenes than with high band passed scenes.

The task used in this study is a variant of a new paradigm introduced by Thorpe, Fize and Merlot (1996) to test the temporal limits of object recognition. Their "ultra-rapid visual categorization" (URVC) involves participants making speeded judgments about the contents of scenes that are presented very rapidly, often less than 30 ms.

Ultra-Rapid Visual Categorization

In the URVC paradigm, participants are presented photographs of natural scenes for a very brief amount of time, usually under 30 ms. The participants' task is a go/no go categorization task in which they are asked to determine whether the photograph contains a member of a particular target category, such as "animals". In addition to accuracy measurements, reaction times are recorded, and ERP measurements are typically recorded to determine the neurophysiological time course of the category decision.

Using this paradigm, Thorpe and his colleagues consistently report findings that challenge our understanding of the speed of object and scene recognition. Both human and monkey observers can make a category decision based on a very brief exposure with near-ceiling accuracy (80-90% correct). Furthermore, this decision seems to be made very quickly; although

the mean reaction time in these studies is usually approximately 400 ms, analyses of the fastest reaction times in which correct responses significantly outnumber incorrect ones suggest that the minimum amount of time needed to perform this task is under 250 ms (VanRullen & Thorpe, 2001b). These reaction times are much shorter than typically reported (e.g., Breitmeyer, 1975), which may be partly due to a touch- sensitive response plate used to record responses instead of a mouse or keyboard. The authors argue that these short reaction times are evidence of very rapid visual processing.

ERP analyses seem to corroborate this claim; a frontal negativity specific to no-go trials within 150 ms of stimulus onset suggests the visual processing necessary to perform this task occurs in under 150 ms. Interestingly, this divergence at 150 ms appears to be task-related and not the result of perceptual differences between the target category and distractors. In one study, ERPs were recorded for participants performing a dual categorization task in which they reported whether images contained animals or vehicles. In the animal task, half the distractor images were of vehicles and vice versa. An ERP analysis showed a category-specific divergence from zero at roughly 75 ms. The divergence appeared to be independent of the task, as animal and vehicle stimuli produced similar waveforms irrespective of their status as target or distractor. Thus, it appears that although task-related differences in waveforms appear at 150 ms, participants are perceptually distinguishing the two categories well before then (VanRullen & Thorpe, 2001a).

Contrary to most current models of object recognition that hold the visual processing of a scene involves a multi-stage, iterative process (e.g., Biederman, 1987; Marr, 1982), the results of these studies suggest the categorization of objects in naturalistic scenes can be done using a strictly feed-forward process. Ultra-rapid visual categorization can be carried out with minimal exposure to achromatic stimuli, does not require directed attention, and does not indicate

influences of cognitive factors, such as practice effects or category effects (Delorme, Richard, & Fabre-Thorpe, 1999; 2000; Thorpe, Gegenfurtner, Fabre-Thorpe, & Bülthoff, 2001; Lei, VanRullen, Koch, & Perona, 2002; Fabre-Thorpe, Delorme, Marlot, & Thorpe, 2001; VanRullen & Thorpe, 2001b).

Despite other evidence that color information affects object and scene recognition (e.g. Oliva & Schyns, 2000), URVC performance does not depend on the chromaticity of stimuli (Delorme et al., 1999; 2000). In these studies, 400 color or black and white photographs of natural scenes were presented for 32 ms. Human and monkey observers responded whether the photograph depicted a member of a particular category, either food or animals. Performance on this task was near ceiling in all conditions, indicating color does not significantly influence the ability to categorize naturalistic scenes. The authors took this result as evidence for a role of the magnocellular (M) pathway in URVC, as this pathway is generally colorblind. They argue because motion and luminance information from the M pathway reaches the visual cortex 20 ms prior to color-based information from the parvocellular (P) pathway (Nowak, Munk, Girard, & Bullier, 1995), a rapid process such as URVC would be likely to rely on the faster M pathway.

Further evidence that URVC is mediated by a primary, feed-forward process is that the rapid categorization of natural scenes can be carried out independently of focused attention. Thorpe et al. (2001) presented natural images for 28 ms at locations throughout nearly the entire visual field. While performance was near ceiling for centrally presented scenes, participants could categorize objects presented in the far periphery at levels significantly above chance. In this study, often participants described not being able to determine the content of the picture, although they were able to categorize it. This discrepancy in reported perception and performance suggests focused attention is not necessary to perform a URVC task. A more

convincing example of the independence of URVC and attention is the finding that performance on this task is not affected by a concurrent centrally demanding task (Lei et al., 2002). A characteristic of preattentive visual tasks is that they can be carried out simultaneously with other tasks with little or no cost. In Lei et al.'s (2002) study, participants were engaged in a centrally demanding letter discrimination task with one of three concurrent peripheral tasks: a natural scene categorization, a letter discrimination, or a color pattern discrimination. Performance on these peripheral tasks varied; although the URVC task was carried out without difficulty, the letter and color pattern discriminations proved difficult. This finding suggests URVC, and not other comparable tasks, can be executed with little directed attention and is likely to be a primary, bottom-up type of process.

The idea that URVC is carried out by a feed-forward process is supported by the finding that the ability to categorize natural scenes does not improve with repeated exposures to stimuli (Fabre-Thorpe et al., 2001). In this study, viewers practiced responding to 200 images 5 times a week for 3 weeks. Afterward, they were tested with 1200 new scenes and 6 repetitions of the old scenes. Although performance improved slightly, further analyses determined this effect was driven by an improvement to a few ambiguous stimuli. The authors argue the time course of URVC appears to be relatively fixed, which indicates its independence of top-down processing, which depends heavily upon experience and would therefore be susceptible to practice effects.

URVC performance does not differ based on the particular categories involved in the task (VanRullen & Thorpe, 2001b). No differences were found between natural and artifactual categories, despite a literature suggesting observers treat these types of categories differently (e.g., Forde & Humphreys, 1999; Laws & Neve, 1999). Here, participants were engaged in an URVC task with either "animals," a natural category, or "modes of transportation," an artifactual

category, as the target categories. Neither measures of accuracy nor response times differed between categories. The finding that URVC occurs similarly despite target category further points to an early, feed-forward basis of this phenomenon.

The bulk of the findings from the URVC literature suggest despite the complexity of the task of categorizing natural scenes, the processes involved can be carried out using mainly data-driven mechanisms. This view is supported by evidence from certain neural network models of visual processing which perform difficult tasks using strictly feed-forward mechanisms. One such model, proposed by VanRullen and Thorpe (2002), is based on relative spike timing of populations of neurons rather than the firing rate of single neurons. This model was tested using SpikeNet, a program that simulates the asynchronous firing of groups of neurons, with some success (Thorpe, 2002). (For a review of this model, see Appendix A.)

Although feed-forward processes probably play a major role in URVC, Fabre-Thorpe et al. (2001) acknowledge top-down influences cannot be ruled out, although their involvement would necessarily be limited in order to be consistent with the reaction times found in these studies. Furthermore, these results do not challenge the existence of other feed-forward mechanisms that may not operate as quickly as URVC and extensively involve re-entrant processes (e.g., Ro, Breitmeyer, Burton, Singhal, & Lane, 2003); rather it may be the case that slower processes are simply not necessary to perform this type of task. Fabre-Thorpe et al. (2001) suggest the detection of a target object in a scene could involve a feed-forward parallel search for several features characteristic to category members, which could be specified early on by a higher-level mechanism.

Studying Early Visual Processing With URVC

According to Thorpe and his colleagues, our ability to rapidly categorize natural scenes is a function of the earliest stages of visual processing. If this is the case, the URVC paradigm would be a valuable tool in determining the fundamental aspects of objects necessary for recognition and categorization.

The URVC paradigm is particularly appealing for several reasons. One advantage of this paradigm is the use of natural scenes as stimuli. Compared to other stimuli, natural scenes more closely resemble the visual input in daily life. Also, the scenes typically used in URVC vary in the location and number of target objects within them. This prevents participants from making predictions about target objects and adds to the complexity of the task. However, this variability may result in unwanted perceptual and cognitive influences. For example, stimulus variables such as contrast, mean luminance, and wavelength may differ across images. It may be easier to detect an object in the center of a scene than in the background. Likewise, observers may find it easier to detect several target objects in a scene than just one. One means of dealing with these issues is to use many individual stimuli. Typically, hundreds of photographs are used in URVC experiments. Thus, by averaging across many different photographs, the potential differences between them are minimized. Another benefit of using the URVC paradigm is that the stimuli are presented very briefly. Fabre-Thorpe et al. (2001) argue that these short stimulus presentations add to the complexity of the task and minimize the potential for eye movements.

While the URVC paradigm features very short stimulus presentations, the stimuli are usually followed by a blank interval. Therefore, the possibility of the persistence of an iconic representation of the images cannot be ruled out. In order to limit visual processing to its earliest stages, the present study employed backward masking.

In a backward masking paradigm, the presentation of a target stimulus is followed, usually after a brief inter-stimulus interval, by a masking stimulus, which limits the availability of the target for further processing. Phenomenologically, the viewer has no perception of the target, and is unable to perform above chance on identification and detection tasks (Breitmeyer, 1984; Breitmeyer & Ogmen, 2000). Backward masking has recently been shown via single-cell recordings to greatly reduce, if not obliterate, visual processing to an ideal stimulus (Rolls, Toveé, & Panzeri, 1999; Rolls & Toveé, 1994). When face-selective cells in macaques are measured for firing to briefly presented faces (20 ms), both firing rate and response time of the cells are affected by masking. When the stimulus is masked after a 20 ms presentation, significant decreases are found in overall neural firing rate, duration of firing to the stimulus, and the peak of the firing envelope. Interestingly, it has been found that during masking, the spontaneous activity of the neuron is not affected; decreases in firing are only found in selective firing to the stimulus (Rolls, Toveé, & Panzeri, 1999; Rolls & Toveé, 1994).[1]

A backward masking paradigm has been used in conjunction with URVC in order to test whether feedback loops are involved in rapid categorization (VanRullen & Koch, 2003). In this study, category decisions about briefly presented natural scenes were compared to a condition in which scenes were masked immediately after presentation with random noise. Afterward, participants viewed the same masked images and were asked to report their subjective confidence in perceiving the target objects in the scenes. In the masked trials, when no confidence was reported in their perceptions, participants performed at 55% accuracy (statistically above chance), but when confidence was high, accuracy was at 75%. More telling for the authors was the finding that the minimum response times required for accurate performance did not differ between masked and unmasked conditions. This result was taken to

indicate that in both conditions, some responses were based on information gathered during the first 26 ms of presentation, and thus reflects a feed-forward mechanism (VanRullen & Koch, 2003).

VanRullen and Koch's (2003) study raises some important issues about introducing a visual mask into the URVC paradigm. A main issue related to this proposal has to do with the use of minimum reaction time as a dependent measure of URVC with masking. The authors found comparable minimum reaction times for both masked and unmasked stimuli. However, accuracy measurements in these two conditions were never equal; accuracy was consistently at near ceiling (90%) in the unmasked trials, while performance in the masked condition never surpassed 75%. The mask was clearly affecting performance, despite similarities in response time. Why would masking affect accuracy but not speed? This may be the result of complications arising from the use of a visual mask with the types of stimuli typically used in URVC experiments.

In a masking paradigm, the extent to which a mask is effective is dependent upon the "strength" of the mask relative to the target, which is defined in terms of duration and contrast. A mask will effectively disrupt processing of a target only if it is greater in contrast or duration (or both) than the target (Breitmeyer, 1984). VanRullen and Koch (2003) themselves point out the possibility that local contrast differences between the natural scenes could lead to differential effectiveness of their mask. If some images were higher in contrast than others, they would be less affected by the mask. This could lead to faster responses, increased accuracy, and higher confidence ratings for these particular scenes, and might account for the relationship between accuracy and confidence found in their study. Furthermore, if these stimuli were particularly visible, it would not be surprising to find that in some cases, the minimum reaction time is

comparable between masked and unmasked conditions. The difficulty with this particular issue is that short of creating a unique mask for each individual natural scene (which would be quite an endeavor given they used 1536 unique stimuli), it is nearly impossible to ensure that each scene is equally masked. An ideal alternative is to utilize a very strong mask, so that all target stimuli would be maximally interrupted. Another means of dealing with this issue would be to ensure that the scenes themselves were as perceptually similar as possible so that the mask is comparably effective for all of them. Finally, it may be necessary to recognize that despite these measures, some images will be more visible than others, and will consequently be less masked. Given the potential ineffectiveness of the mask for some scenes, the minimum reaction time measure may not be the best dependent variable to use with a masking paradigm. When masking is involved, a better measure of performance would be participants' accuracy.

To maximize the effectiveness of URVC as a tool for testing models of visual object recognition, it is important to examine the time course of performance on this task. While the results of VanRullen and Koch (2003) are informative in that they show masking has an effect on URVC, they do not examine the effectiveness of the mask over time. In a backward masking paradigm, such an investigation is possible by varying the amount of time between a target stimulus and a mask. Introducing a mask after a target stimulus prevents further processing of it; therefore increasing the time between these stimuli, the stimulus onset asynchrony (SOA), effectively lengthens the amount of processing time for the target post-stimulus. VanRullen and Koch (2003) immediately masked the target, and on average, performance was at 60%. Increasing the SOA would result in a gradual increase in accuracy until performance is similar to that in an unmasked condition, and the resulting function would provide information regarding the time course of URVC.

The Present Study

The present study employed a slightly modified version of the URVC paradigm to investigate what stimulus aspects are used in the early visual processing. Ultimately, the aim of this study is to use this technique to test predictions made by Biederman's RBC and spatial frequency-based models of object recognition about these first stages of object recognition.

The first experiment described here serves more of a methodological purpose than a theoretical one. The time course of URVC previously had yet to be examined using a backward masking paradigm, and in order to evaluate performance using different versions of scenes, it was necessary to determine observers' baseline performance on this task using photograph stimuli. This measure is particularly important due to the procedures employed in this study to minimize stimulus differences (described below). To this end, the first experiment compared participants' accuracy on a typical URVC task to performance when a visual mask is introduced. Participants viewed rapidly presented visual scenes followed either by a blank interval or a mask after a variable SOA and were asked to determine if the scene contains a member of a particular category (animals). If the mask was effective, participants were predicted to be generally more accurate in the unmasked condition. Moreover, accuracy should vary as a function of SOA in the masked condition; at shorter SOAs, performance should not differ from chance. As the time between the scene and the mask increases, participants should become more accurate and eventually, performance should match that of the unmasked condition.

The first experiment provides a masking function for URVC using photographs of natural scenes. Having established masked accuracy with natural scenes, this performance can be compared to accuracy with other model-specific versions of the scenes in order to test these models' predictions about the first stages of visual processing. The second experiment tested the

hypothesis that a line drawing representation is sufficient to perform URVC tasks, as is suggested by edge-based models of object recognition. In this experiment, URVC performance was assessed using both photographs of scenes and line drawing renditions of the same scenes.

While edge-based models hold a line drawing representation is fundamental to the earliest stages of visual processing, other theorists hold that particular ranges of spatial frequencies are necessary for object recognition and categorization. Some research indicates visual processing operates using coarse, low frequency information first, followed by detailed, high frequency information (e.g., Parker, Lishman & Hughes, 1992; 1997). In their rank order coding model of vision, VanRullen and Thorpe (2002) hold that low spatial frequency information is essential to early vision, and is the information used in processes such as URVC. However, other research (e.g., Oliva & Schyns, 1997) suggests the processing of different spatial scales is more flexible and that high frequency information may be given priority by the visual system if the task demands it. The third experiment in this study addressed this issue by contrasting URVC performance with low and high band passed versions of natural scenes. If low frequency information is essential to the early stages of object recognition, then participants would be more accurate with low passed than with high passed scenes.

Throughout these experiments, the same set of 400 natural scenes was employed and modified for each experiment. These natural scenes were similar to those used in typical URVC experiments, but whereas stimuli are usually chosen to maximize variability, the scenes used in this study were manipulated to minimize differences between them. Whereas variability between the stimuli is usually an asset to the URVC paradigm, as noted earlier, in a masking study, differences between the scenes could introduce potential confounds. If the scenes differed too much in terms of perceptual variables, the mask may be differentially effective across the

images, and thus differences between experimental conditions might actually be due to differences in masking.

In order to minimize differences in contrast, an equalization process was used. This process normalizes the range of gray values in an image so that the resulting overall contrast is approximately 50%. Color has been found to influence object recognition in certain instances in which it is diagnostic (e.g., Oliva & Shyns, 2000), so achromatic stimuli were used. This manipulation was not expected to influence URVC itself, as the process is colorblind (Delorme et al, 1999; 2000). Only one target figure was presented per scene, and each figure was centered in the scene in order to ensure that all figures were equally visible. Half the scenes will depict members of the target category, "Animals," while the other half portrayed distractor figures. These distractor scenes depicted members of several different categories and included natural objects that are not animals, modes of transportation and other artifacts. The purpose of this control was to minimize the likelihood of a viewer categorizing the distractors into a single separate category. Any legible text was edited out of the scenes, so as to minimize any automatic reading processes.

Just as the same set of natural scenes will be used throughout the proposed study, the same mask was used in throughout the experiments. When deciding how to design a mask for this study, a potential problem arose. A study by Delord (1998) suggests the most appropriate mask for a given target depends upon the specific spatial frequency range of information most relevant to perform the task. When the perceptual task involves discrimination of global stimulus features, a low-frequency noise mask appears to be the most effective. However, when naming of a specific object is involved, a pattern mask is the most efficient. The problem was that in order to choose the most appropriate mask for rapidly presented natural scenes, one must know

what type of spatial information will be used in their categorization. Obviously, as this question is the topic of the third experiment in this study, it had yet to be answered decisively. The most effective solution was to employ a very strong mask that contains a broad range of spatial frequencies. Consequently, an achromatic random noise mask was used throughout this study. A spectrum analysis of this mask indicated it contained a wide range of spatial frequencies, and it was additionally very high in local contrast. Also, pilot investigation suggested this mask was effective when used with these achromatic natural scenes.

CHAPTER 2

EXPERIMENT 1: RAPID CATEGORIZATION OF NATURAL SCENES WITH BACKWARD MASKING

The purpose of this experiment was two-fold. First, due to the perceptual differences between the images used in this study and the types of stimuli usually used in the URVC paradigm, it was important to establish that rapid categorization was possible using this set of images. To this end, a group of observers viewed the set of natural scenes in a typical URVC scenario; each image was presented very briefly and followed by a blank interval. Participants were asked to indicate, as quickly and accurately as possible, whether or not the image contained an animal. It was expected that observers would perform this task with near ceiling accuracy.

The second purpose of this experiment was to determine the baseline function of URVC accuracy over time when visual processing is limited to its earliest stages. This was accomplished by the introduction of a noise mask into the standard URVC paradigm. In this condition, a random noise mask followed the presentation of the photograph after a variable stimulus onset asynchrony (SOA). As the onset of a mask is thought to disrupt further processing of a target stimulus, this manipulation effectively limited processing time to the duration of the SOA. With a fixed target duration of 12 ms and SOAs of 12, 24, 36, 48, 60, 72, 84, and 96 ms, participants' total processing time was limited to under 100 ms throughout this experiment.

It was predicted that categorization accuracy would be near chance levels at the shortest SOAs, and would gradually increase with SOA as more time was available for visual processing to occur. It was also expected that when enough processing time was available, categorization accuracy would be comparable to that in the no-mask condition.

Method

Participants

Twenty University of Georgia students participated in this experiment. Sixteen of these were recruited from the Psychology Department's research pool and received class credit for their participation. Others participated on a voluntary basis. All participants showed normal or corrected-to-normal acuity, as tested by an Orthorater™. Participants gave informed consent prior to testing, and were naïve to the purposes of the experiment until its completion, at which time they were debriefed.

Apparatus

Stimuli were presented and responses recorded on a Dell Dimension XPS R400 computer equipped with a Pentium 3 processor and an 18" diagonal, 75 Hz ViewSonic GS790 color monitor (mean luminance = 40 cd/m^2). E-Prime experimental software (Psychological Software Tools, 2001) was used to mediate stimulus presentation and data collection. Participants responded by keypress using a Dell Quietkey keyboard.

During the experiment, participants were seated comfortably in a darkened room. A chinrest was used to minimize head movements.

Stimuli

A mixed-subjects 2 (mask/no mask) x 2 (target/distractor) x 8 (SOA) design was used with masking condition as a between-subjects variable and SOA as a nested variable within the masked condition. Examples of the stimuli used in this experiment can be found in Figure 1. Four hundred digital photographs of natural scenes were obtained from publicly available Internet databases. Using Adobe Photoshop 5.5 software (Adobe, 1999), the photographs were manipulated. Each image subtended 10º visual angle at a viewing distance of 24". Color

information was removed from the photographs and contrast was normalized using an equalization process that resulted in approximately 50% contrast across stimuli. All text information was edited out of the photographs. Likewise, all human forms were removed. Each image depicted a single figure in the center of the scene. Of these figures, half featured a member of the target category, "animals," while the other half served as distractors. The 200 distractors were composed of natural objects that are not animals, modes of transportation and other artifacts.

Procedure

Displays consisted of a white fixation cross presented for 2 sec. followed by a black interval for 500 ms. The black interval served to minimize any perceptual effects of the fixation as well as to reduce anticipatory responses. Following this interval, a natural scene was presented for one refresh of the monitor (12 ms). In the no-mask condition, the scene was followed by a blank interval of 3000 ms, during which participants performed a two-alternative forced choice categorization, indicating by keypress whether or not the scene contained an animal.

In the masked condition, the scene was followed by a blank interval of either 0, 12, 24, 36, 48, 60, 72, or 84 ms and then by a random noise mask for 300 ms. The combination of the image and blank interval durations resulted in SOAs of 12, 24, 36, 48, 60, 72, 84, and 96 ms. The mask was followed by a 3000 ms interval for participants' categorization responses, after which the fixation cross returned.

Scenes were randomly assigned to SOA for each participant, so that in the masked condition, each subject viewed 25 randomly selected images at each SOA. The order of trials was randomized for each participant. The primary dependent measure in this experiment was

accuracy of correct categorizations (either as "animal" or "not animal"), although reaction times were also recorded. After testing, participants were debriefed and credited for their participation.

Results

Accuracy

In this, and all other experiments, $\alpha = .05$, unless otherwise noted. As was expected, categorization accuracy was near ceiling in the unmasked condition ($M = 93.97$, $SE = .89$). Importantly, this result indicated typical URVC accuracy with the particular stimulus set used in this study. In the masked condition, accuracy with both animals and distractors was near chance at the shortest SOAs (12ms: $M = 57.64\%$, $SE = 3.52$) and increased with SOA until it neared ceiling at the longest SOAs (96 ms: $M = 90.76\%$, $SE = 1.28$). A univariate analysis of variance (ANOVA) conducted on percentage of correct responses in masked trials revealed a main effect of SOA, $F(7,56) = 54.60$, $p < .05$, which reflects this trend.

Inspection of the data indicated a strong participant bias toward responding "Not Animal," particularly at shorter SOAs. To obtain an unbiased measure of accuracy, a signal detection analysis was used (Green & Swets, 1988). Participants' *d'* was calculated as

$$d' = \frac{\text{Hits}}{\text{Number of Hits}} - \frac{\text{False Alarms}}{\text{Number of Distractors}}$$

with correct "Animal" responses defined as "Hits" and trials in which participants responded "Animal" to a non-animal stimulus as "False Alarms." All subsequent analyses were conducted on *d'* values.

Mean *d'* measures can be found in Table 1 and graphically represented in Figure 2. Interestingly, planned weighted comparisons between participants' *d'* measures in each of the 4 longest SOA conditions and the unmasked condition showed that accuracy in the masked

condition never reached the same level as in the unmasked condition, even at the longest SOAs (60 ms: $t(12) = 2.16$, $p < .05$; 72 ms: $t(11) = 3.66$, $p < .05$; 84 ms: $t(11) = 2.70$, $p < .05$; 96 ms: $t(14) = 2.30$, $p < .05$). Furthermore, *d'* measures in the masked condition varied only somewhat after the 60 ms SOA condition, which suggests masked accuracy would not improve had slightly longer SOAs been used. This result is not surprising, considering the presentation of a mask has been found to prevent target detection even after as long as 100 ms after target offset (Francis, 2003).

Reaction Times

Although participants' reaction times were not considered a primary dependent measure in this study, one of the hallmarks of URVC is the speed at which participants can make rapid category decisions. Thus, reaction times to unmasked stimuli will be discussed here briefly to serve as a comparison of the present results to the URVC literature in general. Generally speaking, participants' reaction times were considerably slower than those typically reported in URVC studies. In the unmasked condition, the mean reaction time to "Animal" stimuli was 573.66 ms ($SE = 17.51$) and 607.06 ms ($SE = 15.96$) to distractors. Usually, reaction times in URVC studies average approximately 400 ms (e.g., Thorpe et al., 1996). There are several potential explanations for this result, which will be discussed below.

Discussion

There are two primary conclusions to be drawn from this experiment. Firstly, typical URVC accuracy was found with the stimulus set used throughout this study in the unmasked condition. This conclusion was particularly important given the perceptual differences between these images and those typically used in the URVC paradigm. High accuracy was not a

surprising finding, however, considering that by making the scenes perceptually similar, participants should have found the categorization task easier than the standard URVC paradigm.

In contrast, participants' reaction times to these stimuli were unusually slow; typical mean response times in this paradigm are almost 200 ms faster than those found in this experiment. Several factors could contribute to this discrepancy. In this experiment, participants responded on both target and distractor trials. Other URVC studies have used a go/no go task, requiring participants to only respond when a target is perceived. Perhaps the two-alternative paradigm is more cognitively demanding, resulting in longer response times. Also, Thorpe and his colleagues collect responses using a highly sensitive touchpad, which enables them to measure response times with precision. Although the keyboard used to collect responses in this experiment is accurate within several milliseconds, timing errors could have artificially inflated reaction times. Finally, although observers were instructed repeatedly to respond as quickly and accurately as possible, it is possible that some participants were simply not motivated to do so, or felt that they could not respond quickly without sacrificing accuracy. While it is unlikely that any one of these explanations suffices to explain the long reaction times found in this experiment (and throughout this study), perhaps these and other factors worked in concert to affect these results.

The most important conclusion from this experiment is the function of participants' accuracy relative to SOA when a noise mask is introduced after the presentation of the target image. The main effect of SOA showed that accuracy improved as the time between target and mask increased, with performance near chance levels at the shortest SOAs and approaching ceiling as SOA lengthened. The resulting function has several interesting facets. First, while accuracy at the 12 ms SOA is near chance, with $d' = .14$, a one-sample t-test showed that

categorization accuracy is significantly above chance ($t(8) = 2.32$, $p < .05$). The implication of this finding is that when visual processing time is limited to only 12 ms, observers are able to make category decisions about complex natural scenes at above chance levels. This finding suggests the stimulus information necessary to make these decisions is available in the very earliest processes. Upon debriefing, many participants reported not seeing the target at all on many trials, which suggests rapid categorizations can be made with some accuracy without conscious awareness of the stimulus.

Another interesting aspect of these results lies at the other end of the masking function, where SOAs are longest. While accuracy in the masked condition approached ceiling, categorization performance never reached the level of accuracy found in the unmasked condition. This finding speaks to the effectiveness of the mask to interrupt the processing of the target. However, this is not a particularly surprising result, given previous studies which report complete masking of target stimuli at much longer SOAs (e.g., Breitmeyer, 1988; Francis, 2003).

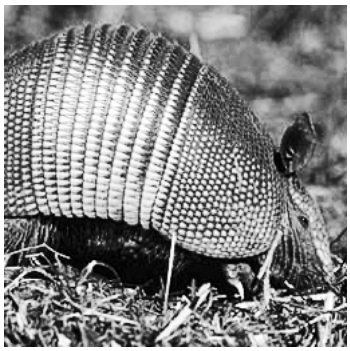
The purpose of establishing a masking function in this experiment was to determine how categorization accuracy varies with respect to SOA when participants view intact photographs of natural scenes. As various types of stimulus information are removed from these photographs in subsequent experiments, the results of this experiment serve as a standard of comparison. To be able to determine the whether these manipulations are effective, it was necessary to establish how participants responded with full images. Furthermore, these results indicate the masking paradigm used in this study is a valid means of limiting visual processing to its earliest stages.

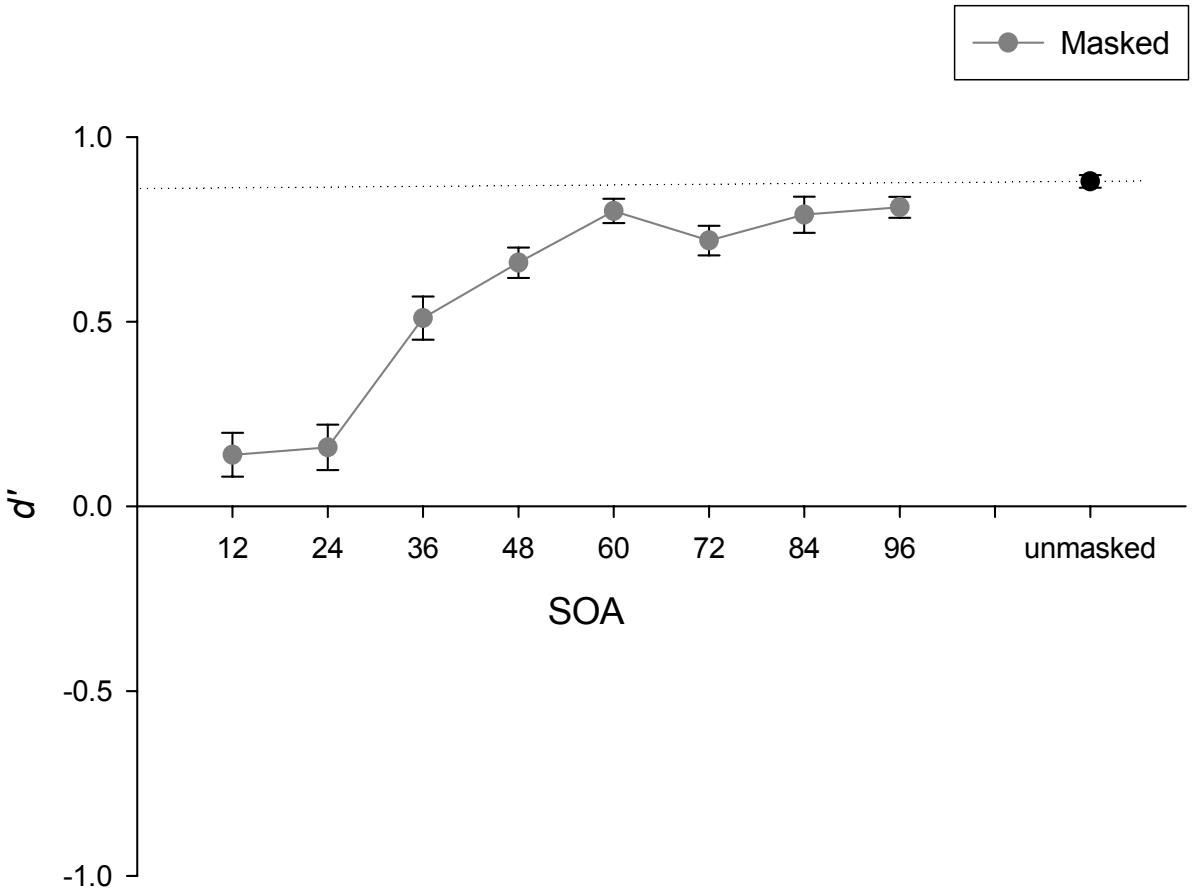
Table 1

Masked URVC Accuracy for Photographs of Natural Scenes as a Function of SOA

SOA (ms)	*d′*	*SE*
12	0.14	0.05
24	0.16	0.06
36	0.51	0.06
48	0.66	0.04
60	0.80	0.03
72	0.72	0.04
84	0.76	0.04
96	0.80	0.03

Examples of Stimuli Used in Experiment 1




Masked
1.0
0.5
0.0
-0.5
-1.0
d'
12
24
36
48
60
72
84
96
unmasked
SOA

CHAPTER 3

EXPERIMENT 2: A COMPARISON OF URVC WITH PHOTOGRAPHS AND LINE DRAWINGS

Having established masked URVC accuracy for photographs of natural scenes in Experiment 1, this paradigm can be used to determine the types of stimulus information used to make rapid categorizations. Assuming URVC relies on the first information to be processed by the visual system, the information used to perform this task is likely to reflect the first stages of object recognition. The purpose of this experiment was to determine if edge-based information is sufficient to make rapid categorizations. If participants are able to perform a URVC task with only line drawing representations of scenes, it would suggest edge-based information is used in the earliest stages of object recognition.

Many current models of object recognition hold that representations of objects are derived from two-dimensional, edge-based extractions of the visual field (e.g., Marr, 1982; Hummel & Biederman, 1992). Biederman's (1987) RBC model has arguably been the most influential. According to this model, the first stage of object recognition is an edge extraction stage, in which a line drawing representation of an object is generated. The line drawing is then parsed into volumetric primitive components (geons), which combine to form a structural description of the object. This representation is then identified as an object through comparisons with other object representations in memory.

Since its introduction into the literature, RBC has been a dominant theory in object recognition. Consequently, it has been assumed that line drawings are sufficient stimuli for the investigation of object recognition, and much research of these processes has involved line

drawings. Although this assumption is critical to the validity of these experiments, the question of whether line drawing stimuli are actually equivalent to more realistic stimuli, such as photographs, has yet to be sufficiently addressed.

Tests of recognition for information contained in line drawing versus photographic stimuli have produced mixed results. Nelson, Metzler, and Reed (1974) found no difference in either short-term (7 minutes) or long-term (7 weeks) recognition for line drawings and photographs. However, using the same stimuli, Loftus and Bell (1975) found an advantage for recognition of photographs. Methodological differences are likely sufficient to explain the discrepancies in these results; while Nelson et al. (1974) presented stimuli for 10 sec., Loftus and Bell (1975) varied the presentation time of the stimuli from 60-500 ms. Considering performance in the short-term recognition task was at ceiling in the former study, it may be the case that a 10 sec. exposure is long enough for observers to extract enough information about the content of any type of stimulus to distinguish it from others in a recognition task.

When compared to caricatures, photographs of famous faces are recognized and recalled more accurately and matched to the target individuals' names more quickly. Furthermore, photographs of famous faces are rated as more characteristic of the target (Tversky & Baratz, 1985). These results suggest that photographic stimuli are more similar to memory representations for famous faces than caricatures, even though caricatures highlight distinctive facial features, which should be more salient in memory. However, when asked to rate the typicality of objects in line drawing or photographic scenes, no differences were found, which may suggest that familiarity does not depend upon the level of detail in a stimulus (Beltran & Duque, 1993).

In a series of experiments, Biederman and Ju (1988) investigated whether line drawings are identified as quickly as color photographs. In most cases, no differences were found between line drawings and photographs, and the differences that were found favored line drawings. Further analyses indicated color was not a diagnostic factor in naming the photographic stimuli, suggesting edge-based information is of primary importance in object recognition

In this experiment, URVC performance was compared between photographs and line drawing versions of natural scenes. Participants viewed 200 photographs and 200 line drawings of scenes. Each scene was presented very briefly, and followed by a noise mask after a variable SOA. As in Experiment 1, the participants' task was a two-alternative forced choice categorization; they were asked to report whether or not the picture depicted an animal. It was predicted that if edge-based information is used in the first stages of object recognition, URVC accuracy would be comparable between these two conditions.

Unfortunately, equalization of stimuli is a major difficulty in comparing line drawings to photographs. Black and white line drawings are inherently high in local contrast, while photographs typically feature lesser changes in contrast. This issue is particularly relevant in this study, as masking is highly contrast-dependent (Breitmeyer, 1984). If the line drawings and photographs are naturally different in contrast, the mask would be differentially effective, and differences in threshold may simply be attributable to stimulus differences. This problem is exacerbated by the photograph stimuli themselves, which, although equalized, vary in local contrast changes. In other words, although the average contrast of the pictures is 50%, the contrast at any given area within the picture could take on any possible value. To control for this factor, two measures will be taken. Firstly, the line drawing stimuli were reduced to 50% contrast. Thus, the line drawings matched the photographs in average contrast while maintaining

the visibility of the drawings' contours. Additionally, the noise mask used throughout this study was designed to be high in local contrast in order to maximize its effectiveness as a mask for both types of stimuli.

An important aspect of the line drawing stimuli in this experiment is that the entire scene, not just the figure, was represented in the line drawing. By maintaining the entirety of the image, the task remained comparable in both line drawing and photograph conditions; in order to identify the figure, the participants necessarily segregated it from the background. As a result, the task was more similar to how objects are identified in actual visual space, unlike previous studies in which line drawings were presented in isolation. (although, see Sanoki, Bowyer, Heath, & Sarkar, 1998).

Method

Participants

Thirty undergraduates were recruited from the University of Georgia psychology research pool. Participants showed normal or corrected-to-normal visual acuity as tested by an Orthorater™. All participants gave informed consent prior to testing, and were naïve to the purposes of the experiment until its completion, at which time they were debriefed. The students received class credit for their participation.

Apparatus

The apparatus used in this experiment was the same as in Experiment 1.

Stimuli

Examples of the stimuli used in this experiment can be found in Figure 3. A 2 (line drawing/photograph) x 8 (SOA) within-subjects design was used. The stimuli used in this experiment consisted of the 400 photographs used in Experiment 1 and subtended 10° visual

angle at a distance of 24". The photographs were selected and converted into line drawings by first tracing the scenes onto paper, scanning them into .bmp files, and using Adobe Photoshop 5.5™ software to remove stray marks and adjust contrast levels. The photographs were selected so that the relative proportions of 50% animals and 50% distractors remained consistent between line drawing and photograph conditions. The line drawing stimuli were reduced to 50% contrast to match the contrast of the photographs. The random noise mask described in Experiment 1 was used in this experiment.

Procedure

Displays involved the same masking paradigm described in Experiment 1. A white fixation cross was presented for 2 sec followed by a blank interval for 500 ms. Following the blank interval, a scene was presented for 12 ms (one scan of the monitor), followed by the noise mask after a variable SOA of either 12, 24, 36, 48, 60, 72, 84, or 96 ms. Mask duration was 300 ms, and was be followed by another blank interval for 3000 ms so that participants responses could be recorded. The participants' task was the same two-alternative, forced choice categorization task used in Experiment 1. The primary dependent measure was observer's accuracy.

Each participant viewed 200 randomly selected line drawings as well as 200 randomly selected photographs for a total of 400 trials. Scenes were randomly assigned to SOAs in order to minimize effects of differences between the images. After testing, the participant was debriefed and credited for his or her participation.

Results

Accuracy

In both line drawing and photograph conditions, participants' percentage of correct responses increased with SOA; in the 12 ms SOA condition, accuracy was near chance levels (photos: $M = 60.17\%$; $SE = 2.67$; line drawings: $M = 55.15\%$; $SE = 1.97$) and near ceiling at the 96 ms SOA (photos: $M = 89.20\%$; $SE = 1.41$; line drawings: $M = 90.99\%$; $SE = 1.53$).

Mean *d'* scores for line drawing and photograph stimuli at each SOA can be found in Table 2 and graphically represented in Figure 4. Participants' *d'* values were subjected to a 2 (Stimulus Type) x 8 (SOA) within subjects analysis of variance (ANOVA). Due to a violation of Mauchly's test of sphericity, all F values were adjusted using the Greenhouse-Geisser correction. A main effect of Stimulus Type indicated participants were generally more accurate with line drawing stimuli than photographs ($F(1,29) = 72.320$; $p < .05$), a result that supports the predictions made by edge-based models of object recognition. A main effect of SOA was also found, $F(7,203) = 150.20$; $p < .05$, with accuracy improving as SOA increased as expected. These effects were mediated by a Stimulus Type x SOA interaction, however ($F(7,147) = 4.41$, $p < .05$. Planned comparisons indicated accuracy between photographs and line drawings did not differ at the shortest SOAs, but was greater with line drawing stimuli at SOAs of 48-84 ms (48 ms: $t(56) = 3.84$, $p < .05$; 60 ms: $t(56) = 2.56$, $p < .05$; 72 ms: $t(56) = 4.48$, $p < .05$; 84 ms: $t(56) = 2.50$, $p < .05$). Thus, participants' accuracy was always at least as good, if not better, with line drawings than photographs. It is likely ceiling and floor effects can account for the lack of differences between stimulus types; at the shortest SOAs, accuracy may be comparable because the categorization task is particularly difficult given those viewing conditions. Similarly, as

performance approaches ceiling at the longer SOAs, differences between stimulus types may be minimized due to the relative ease of the task in both conditions.

Discussion

Several important conclusions can be drawn from this experiment. Importantly, the results of the photograph condition replicated those in Experiment 1 (see Figure 2). However, the most important finding in this experiment was that participants' categorization accuracy was similar, if not better, with line drawing versions of the images than photographs at every SOA. It is clear from these results that edge-based information, as is found in line drawings, is sufficient to perform rapid categorizations. This finding lends support to edge-based models of object recognition, such as Biederman's (1987) Recognition-By-Components, as well as to others (e.g., Marr, 1982; Lowe, 1985).

Finding that line drawing stimuli were more easily categorized than photographs may seem counterintuitive when one considers that photographs more closely resemble our actual visual input. However, when Biederman and Ju (1988) compared performance with line drawings and photographs on identification and name verification tasks, occasional differences were found favoring line drawings. Also, there are other factors that may have affected the results of this particular experiment. As was discussed earlier, line drawings are inherently high in contrast compared to photographs. Although the contrast of the line drawings was reduced so that the average contrast of these images matched that of the photographs, the line drawings remained relatively high in local contrast. Unfortunately, it was not feasible to match the line drawings and photographs in terms of local contrast. Furthermore, any attempts at doing so would necessarily result in a line-drawing representation that would not resemble the types of information specified by edge-based models of object recognition. Masking is dependent upon

the relative contrast between target and mask; consequently, the mask may have been less effective with the higher-contrast line drawings, rendering them more visible to observers. This might explain the more accurate performance with line drawings at intermediate SOAs.

The results of the present experiment support the findings of Biederman and Ju (1988), who found similar performance with photographs and line drawings. Furthermore, these findings expand on the results of this widely-cited study. Whereas the stimuli used by Biederman and Ju were manipulated so that they resembled a geon-based representation of an object, such measures were not taken with the images used in this experiment. On the contrary, many of the scenes depicted natural objects, such as rocks, water, and plants, for which no specific geon structure is defined.

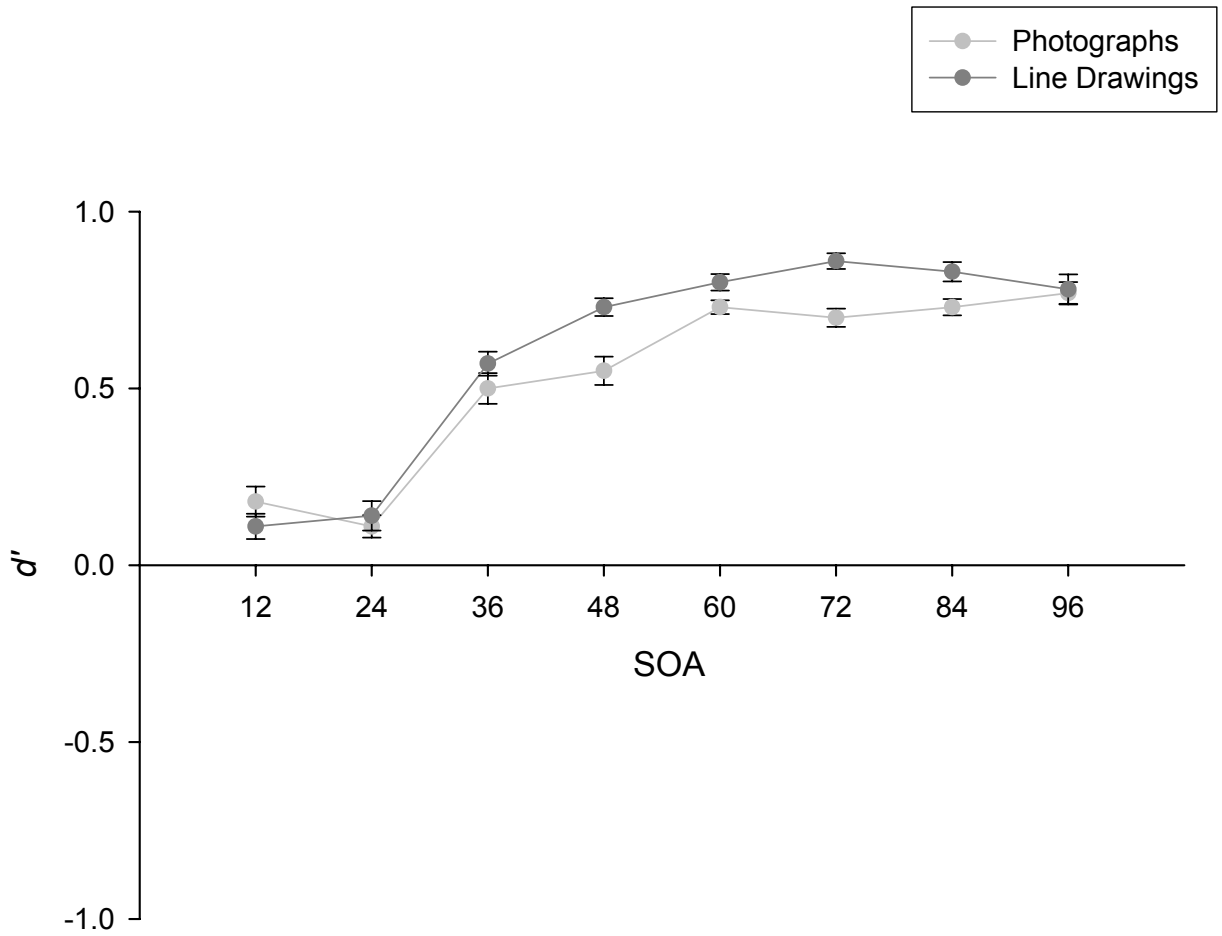
It is important to note that this experiment was not designed to compare different edge-based models of object recognition; rather, its results lend support to all models that suggest an edge-based representation is used in the first stages of visual processing. More importantly, the results of this study show that edge-based information is sufficient to perform rapid categorization tasks.

Table 2

Categorization Accuracy for Photograph and Line Drawing Images as a Function of SOA

	Photographs		Line Drawings	
SOA (ms)	*d′*	*SE*	*d′*	*SE*
12	0.18	0.04	0.11	0.04
24	0.11	0.03	0.14	0.04
36	0.50	0.04	0.57	0.03
48	0.55	0.04	0.73	0.03
60	0.73	0.02	0.80	0.02
72	0.70	0.03	0.86	0.02
84	0.73	0.02	0.83	0.03
96	0.77	0.03	0.78	0.04

Examples of Stimuli Used in Experiment 2


Photographs
Line Drawings
1.0
0.5
0.0
-0.5
-1.0
d'
12
24
36
48
60
72
84
96
SOA

CHAPTER 4

EXPERIMENT 3: INVESTIGATING THE ROLE OF COARSE AND FINE SCALE PROCESSING IN URVC

The results of Experiment 2 suggest edge-based information can be used to make rapid categorizations. However, this finding does not rule out the possibility that other stimulus aspects are available in early visual processing as well. Experiment 3 was designed to test the hypothesis that spatial frequency-based information can be used to perform URVC as well. This was accomplished by comparing categorization accuracy when observers viewed low and high band passed versions of natural scenes in the masked URVC paradigm.

It has been long known that visual information can be represented in terms of its Fourier components (see DeValois & DeValois, 1988, for a review), with different ranges of spatial frequencies carrying various aspects of the visual scene. Low spatial frequency information corresponds to coarse luminance changes in stimuli that delineate general form aspects of objects, whereas high spatial frequencies represent the finer details in a scene. In psychophysics, low and high spatial frequencies are distinguished in terms of cycles per degree of visual angle, a measurement that takes into consideration both the size of the image and the viewing distance of the observer. In the image processing literature, however, spatial frequency is discussed in terms of cycles per image, as this measurement is a more accurate reflection of the amount of information in a given image (Morrison & Schyns, 2001). Using this terminology, information found in 8 cycles/image and under is considered low spatial frequency information (roughly 2 cycles/degree) whereas high spatial frequencies (roughly 6 cycles/degree) comprise the information found in 24 cycles/image or above. It is also well known that independent,

specialized spatial frequency channels in the visual system are selectively sensitive to ranges of spatial information (Blakemore & Campbell, 1969).

Given the physiology of early vision, it seems intuitive that higher visual processes might use the different types of information provided by low and high spatial frequencies for various aspects of object perception. Experimental findings suggest coarse scale, or low frequency, information is used to first develop a quick, stable representation of the visual field, whereas high frequency information is integrated later to flesh out object details needed for tasks such as discrimination and categorization. Some of the first evidence for coarse-to-fine scale scene processing was found by Parker, et al. (1992), who presented observers with three-image sequences depicting a single scene at different spatial scales. In these sequences, images were presented for 40 ms each and were composed of a low-pass, a high-pass, and a full scale version of the image, presented in various orders. Participants rated low-to-high sequences higher in terms of quality than high-to-low sequences. Using a similar paradigm, Parker et al. (1997) found that participants were unable to distinguish between low-to-high sequences and the full-scale version of both scenes and faces. These results support the idea that scene processing occurs in a coarse-to-fine scale fashion, with low spatial frequencies processed first, followed by higher frequencies.

Schyns and Oliva (1994) found evidence of coarse-to fine processing using hybrid stimuli in which the low spatial frequency components of one scene are superimposed upon the high spatial frequency components of another. In one experiment, a sample hybrid scene was presented for either 30 or 150 ms, followed by a mask and then a target full-scale image. The task was to indicate whether the hybrid depicted the same scene as the target. When presented for 30 ms, hybrids were matched to the target scenes that shared their low spatial components,

whereas hybrids and targets were matched according to their high spatial components at the 150 ms duration. A second experiment involved a recognition task. In this experiment, hybrid animations were created depicting two hybrid scenes in sequence. Over the course of the animation, one scene within the hybrid was presented shifting from low to high frequency components, while the other scene was depicted shifting from high to low. The participants' task was to name the sequence they perceived. Participants overwhelmingly recognized the low-to-high scene as the scene contained in the hybrid, although both low-to-high and high-to-low were presented simultaneously. The findings from both these experiments suggest that both matching and recognition processes involve coarse-to-fine processing of visual information.

On the basis of this evidence, it appears higher-level vision mimics lower-level vision in the temporal processing of spatial information. However, Morrison and Schyns (2001) suggest task demands may mediate the processing of spatial scales. They argue coarse-to-fine processing may be the most efficient means of extracting information needed for tasks such as form recognition, but if a task demands the use of high frequency information, the temporal aspects of scale-dependent processing may be more flexible. For example, if an observer is required to determine whether a particular image depicts a face, coarse-scale information is sufficient. However, if the task involves determining the age of the individual portrayed in the image, fine-scale information would be necessary. Perhaps the visual system is capable of prioritizing, in terms of processing speed, particular scales of information based on the requirements of the task at hand.

Oliva and Schyns (1997) showed priming for full spectrum scenes using both high and low band passed versions of the scene, suggesting both coarse and fine scale information are available at the early stages of processing. In a subsequent experiment, scenes were presented for

135 ms in an identification task. At first, the scenes were hybrids containing either high or low spatial frequency information and noise. During the course of the experiment, the hybrids changed to include both types of spatial information. Hybrids were identified according to the scale to which participants had been sensitized; if the early images contained high frequencies and noise, subsequent hybrids were identified according to their high frequency components. These findings support the idea that processing does not necessarily progress from coarse-to-fine, rather, the priority of spatial scale information is determined by task.

Other evidence that either low or high spatial frequency information is sufficient for object recognition was found by Peyrin, Chauvin, Chokron, and Marendaz (2003), who tested categorization accuracy with low and high passed natural scenes in a masking paradigm. In this experiment, participants viewed band passed scenes for 100 ms followed by a mask, and made a go/no go category decision about the scene. The authors found accurate performance with both low and high passed scenes, indicating both types of spatial frequency information can be used to make category judgments about natural scenes.

The present experiment was designed to determine which, if any, range of spatial frequencies was sufficient to perform rapid categorizations. If the visual system operates using coarse-to-fine processing, low frequency information should be sufficient to categorize figures in natural scenes, and accuracy with low band passed scenes would be higher than with high band passed scenes. Delorme et al.'s (1999; 2000) suggestion of the role of magnocellular pathway activity in URVC would be consistent with this hypothesis, as this pathway is responsible for the processing of low spatial frequency information. On the other hand, if fine details of scenes are necessary for this type of task, high frequency information would be necessary. If this were the

case, it would be expected that URVC performance with high band passed versions of scenes would be more accurate than with low passed versions.

Method

Participants

Twenty-two undergraduates were recruited from the University of Georgia psychology research pool. Participants were required to show normal or corrected-to-normal visual acuity as tested by an Orthorater™. All participants gave informed consent prior to testing, and were naïve to the purposes of the experiment until its completion, at which time they were debriefed. The students received class credit for their participation.

Apparatus

Stimuli were presented using the same apparatus described in Experiments 1 and 2.

Stimuli

Examples of the stimuli used in this experiment can be found in Figure 5. A 2 (Spatial Scale) x 8 (SOA) within-subjects design was used. In this experiment, new versions of the 400 natural scenes were generated. Low passed versions were filtered so that no spatial frequencies over 2 c/deg were present. High passed versions were also generated, removing any frequencies below 6 c/deg. These spatial frequency ranges reflect those used by researchers investigating the role of spatial scales in face and scene categorization (e.g., Schyns & Oliva, 1999). The same mask was used in this experiment as in Experiments 1 and 2.

Procedure

The same masking procedure was used as in previous experiments. Accuracy was the primary dependent measure. Participants viewed 200 randomly selected low passed images and

200 randomly selected high passed images. As in Experiment 2, images were randomly paired with SOAs.

Results

Accuracy

Participants' accuracy with both low and high passed images was considerably worse than with full-spectrum pictures. At the 12 ms SOA, percentage of correct responses in both stimulus conditions was at chance levels (Low pass: $M = 51.92\%$; $SE = 4.66$; High Pass: $M = 49.17\%$; $SE = 3.28$). Accuracy did not rise above chance in the high passed condition, even at the longest SOA ($M = 51.53$; $SE = 4.56$). With low passed images, accuracy increased slightly with SOA, (96 ms SOA: $M = 64.55$; $SE = 5.21$).

Mean *d'* scores for high and low band passed images as a function of SOA can be found in Table 3 and graphically represented in Figure 6. A 2 (Stimulus Type) x 8 (SOA) within subjects analysis of variance (ANOVA) was conducted on participants' *d'* values. A main effect of Stimulus Type was found, $F(1,21) = 24.20$, $p < .05$, which indicated generally higher accuracy with low passed than high passed images. Performance improved with SOA, evidenced by a main effect of SOA, $F(7,147) = 4.10$, $p < .05$. A Stimulus Type x SOA interaction mediated these effects, $F(7,147) = 2.56$, $p < .05$, which reflects the improvement in performance over SOA in the low passed condition, but not the high passed condition.

Discussion

The main result of this experiment was that URVC performance was significantly more accurate when participants viewed low band passed versions of scenes relative to high passed versions. This finding suggests low spatial frequency information can be used to make rapid category decisions, whereas high spatial frequencies do not provide sufficient information to perform this task. That said, performance with both low and high band passed scenes was

extremely poor; accuracy with the high passed scenes never surpassed chance levels, and although performance with the low passed scenes improved with SOA, accuracy in this condition was still considerably worse than with either photographs or line drawings. There are several potential explanations for this performance.

One explanation for low accuracy in both band passed conditions has to do with the stimulus energy of both types of images relative to the mask. The process of band passing the scenes resulted in images with lower contrast and less spatial frequency content than not only the original scenes, but the mask as well. Consequently, while the noise mask used throughout this study was effective with photographs, it would be even more effective at interrupting the processing of these weaker versions of the stimuli.

It may also be the case that more time between the target and mask is needed to process the band passed images. Accuracy with the low passed scenes tended to increase with SOA; perhaps with longer SOAs, performance could reach near ceiling levels with these stimuli. It is unlikely that longer SOAs would improve accuracy with the high passed pictures, however, considering URVC performance failed to improve even at the 96 ms SOA.

The particularly poor performance with the high passed images raised questions about stimulus visibility. These stimuli were very low in contrast, although they truthfully reflected the high spatial frequency components of the original photographs. Perhaps accuracy was low simply because participants could not see the figures in the stimuli at all. To ensure this was not the case, a small group of participants viewed the high passed images in a typical URVC paradigm. Presentation time was very brief – 12 ms – but visual processing was uninterrupted by masking. In this scenario, mean accuracy was 60.5% (SE = 2.45). A one-tailed t test indicated this result was above chance ($t(7) = 4.30$, $p < .05$), although it was still low relative to the ceiling

accuracy found with unmasked photographs in Experiment 1. Another group viewed the high passed images with unlimited viewing time; the scenes remained onscreen until the participant made a category decision. Here, accuracy was 89.5% (*SE* = 3.34). (Participants' accuracy in these conditions compared to the 96 ms SOA condition is depicted in Figure 7.) While still not equivalent to unmasked photograph accuracy, performance in this case was considerably better, indicating the high passed images did contain sufficient information to make category judgments about them. Given this finding, it seems safe to conclude that high spatial frequency information, while providing information used in normal viewing circumstances to identify objects, is not used in rapid categorizations. Low spatial frequency information, on the other hand, appears to be involved at least to some extent in making URVC judgments. This finding is consistent with coarse-to-fine models of spatial frequency based object recognition.

Table 3

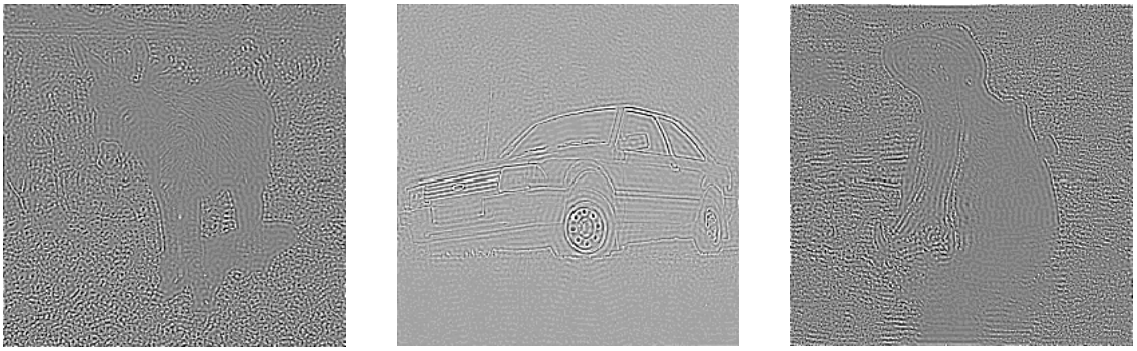
Categorization Accuracy for High and Low Band Passed Images as a Function of SOA

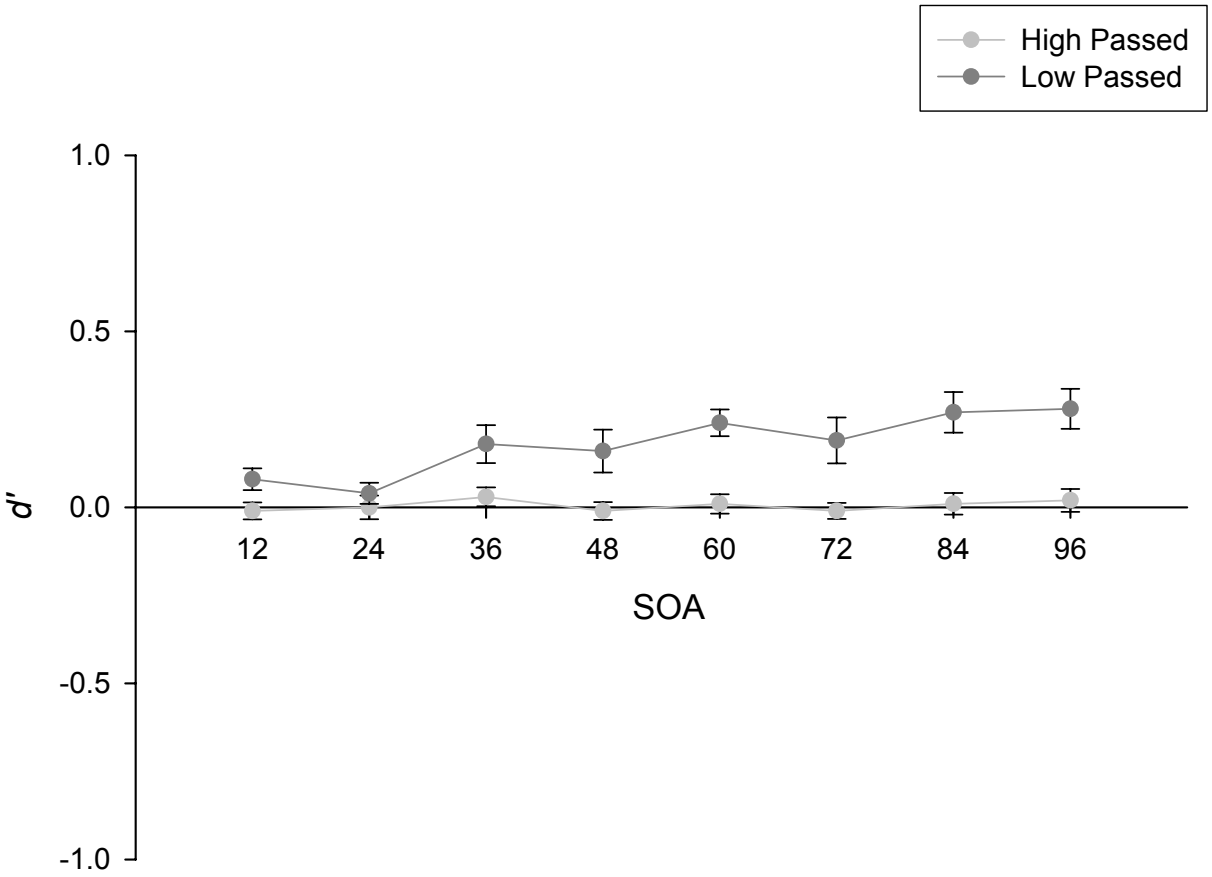
	High Passed		Low Passed	
SOA (ms)	*d′*	*SE*	*d′*	*SE*
12	-0.01	0.02	0.08	0.03
24	0.00	0.03	0.04	0.03
36	0.03	0.03	0.18	0.05
48	-0.01	0.03	0.16	0.06
60	0.01	0.03	0.24	0.04
72	-0.01	0.02	0.19	0.06
84	0.01	0.03	0.27	0.06
96	0.02	0.03	0.28	0.06

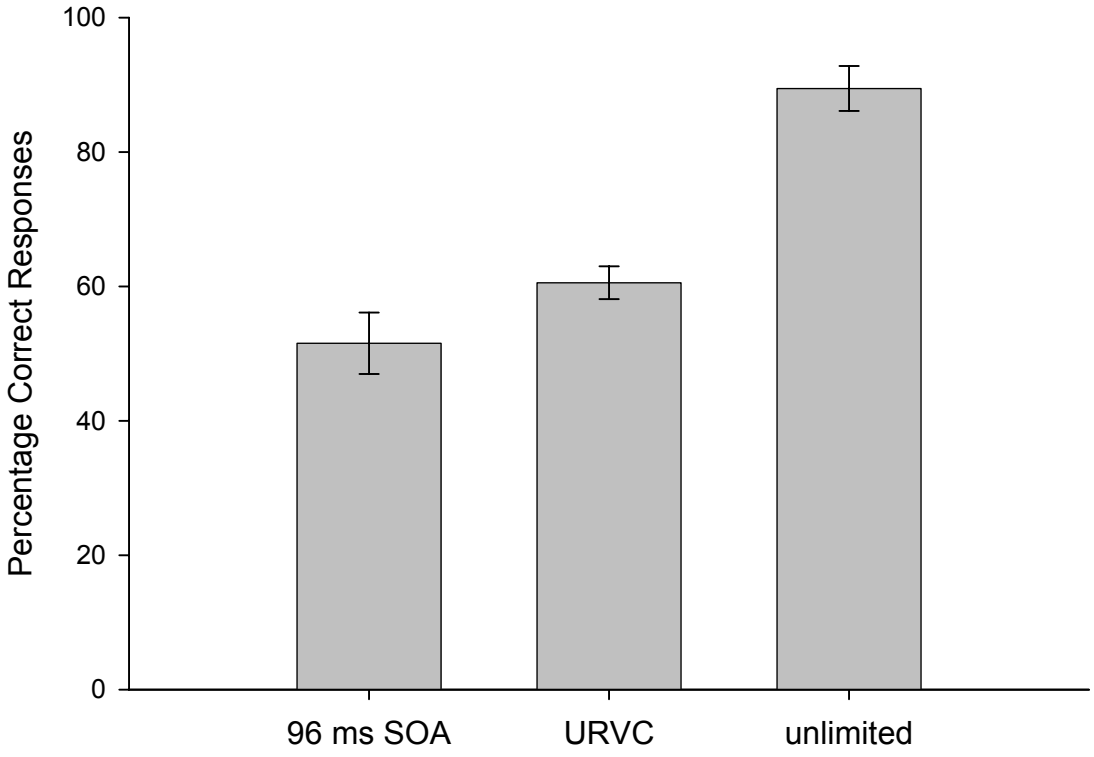
Examples of Stimuli Used in Experiment 3

Low Band Passed

High Band Passed




High Passed
Low Passed
1.0
0.5
0.0
-0.5
-1.0
d'
12
24
36
48
60
72
84
96
SOA


Percentage Correct Responses
100
80
60
40
20
0
96 ms SOA
URVC
unlimited

CHAPTER 5

GENERAL DISCUSSION

The purpose of this study was to determine the aspects of visual scenes used in rapid categorizations, when processing time is limited. In the first experiment, masked URVC accuracy was examined with photographs of natural scenes in order to establish participants' ability to make rapid categorizations with intact stimuli. Experiments 2 and 3 tested the predictions of two models of object recognition about what types of information might be used in making this type of judgment. The results of Experiment 2 clearly showed that edge-based information is sufficient to make rapid categorizations. This finding supports models of object recognition, such as Biederman's (1987) RBC, which hold the first stages of visual processing involve an edge-based extraction of the visual array. Experiment 3 showed that low spatial frequencies can be used to perform URVC tasks, although less effectively, whereas high spatial frequencies provide insufficient information to make rapid category decisions. The results of this experiment support coarse-to-fine models of object recognition, which hold low spatial frequencies are processed early on to provide a rough sketch of the visual array, while high spatial frequency information is integrated later on. Given these results, what conclusions can be drawn about the stimulus aspects used in very early visual processing?

The findings from Experiments 2 and 3 suggest that a line drawing representation is sufficient to perform URVC tasks, and low spatial frequencies are preferred over high frequencies in this type of scenario. Perhaps line drawings and low spatial frequencies provide similar information to the visual system. Both types of stimuli are high in local contrast changes. This similarity is important because changes in contrast typically delineate object boundaries,

providing information about the global form of objects in the visual field. It may be the case that global information is necessary to perform rapid categorizations. Some evidence supports this idea. Observers are able to make quick, accurate judgments about whether objects are real based on their silhouettes, which also convey global form aspects (Dell'Acqua, Job, & Grainger, 2001). It has also been well established that when observers are presented compound stimuli consisting of both global and local types of information (e.g., the form of a square made of smaller circles), global information takes precedence (Navon, 1977; Miller & Navon, 2002).

Furthermore, neither line drawings nor low spatial frequencies contain information about object detail. Although this hypothesis has yet to be empirically tested, it may be that object details are not only unnecessary for rapid categorization, they may actually hinder the ability to make quick judgments based on brief exposures to stimuli. Perhaps the combination of global form information in conjunction with little detail information is ideal for URVC performance, regardless of whether this information is conveyed in an edge-based representation, a spatial-frequency specific representation, or even perhaps a silhouette. Obviously, further research is necessary to test this hypothesis.

While an examination of the types of stimulus information used in early visual processing was the ultimate goal of this study, several other exciting conclusions can be drawn from these experiments. The most important of these was the finding across experiments that when presented sufficient information within an image, participants are able to perform significantly above chance when asked to determine category membership of a figure in a scene when visual processing is limited to as little as 12 ms! Previously, when masking had been used in conjunction with URVC, above chance performance was found at an SOA of roughly 30 ms (VanRullen & Koch, 2003). Although the difference between 12 and 30 ms is slight, it should be

noted that an SOA of 12 ms is less than half as long. This result supports and furthers VanRullen and Koch's (2003) already amazing finding.

To put this finding in realistic terms: Most desktop computer monitors refresh at a rate of 60-85 Hz, or once every 12-17 ms. This constant updating is imperceptible to the casual observer (largely because the desktop display rarely changes significantly every 12 ms). However, the findings of this study imply that if one were watching the desktop display on a computer monitor, and a complex natural scene appeared for only as long as it takes a computer monitor to refresh once, followed by the previous desktop display, one would be able to make a correct category judgment about a figure in that scene roughly 60% of the time. Furthermore, these judgments seem to be made with little, if any, conscious perception of the scene.

Additionally, performance on this task nears ceiling levels at an SOA of 60 ms, which suggests all the information necessary to accurately perform rapid categorizations is available to the visual system within the first 60 ms of processing. This finding is congruent with ERP results in the URVC literature, which suggest waveform differences between target and distractor trials as early as 75 ms (VanRullen & Thorpe, 2001a). Thorpe and his colleagues argue the ability to make such rapid category decisions calls into question the assumptions of current models of object recognition, which hold categorization is the result of lengthy, iterative processing. The results of the present study certainly seem to support this claim.

The results of this study, while impressive, leave many questions unanswered. The findings from Experiments 2 and 3 suggest global form information and not object detail, is necessary to perform URVC tasks. However, this hypothesis has yet to be directly tested. If this is not the case, what aspects of a line drawing representation provide sufficient information to perform rapid categorizations? A recent study by Levin, Takarae, Miner, and Keil (2001) in

which participants performed a category-specific visual search task suggests categories may be distinguishable by between-category differences in contour shape. In this experiment, participants searched for a member of a target category among members of a distractor category. It was found that not only was search highly efficient, but speed of search was predicted by target-distractor differences in rectilinearity and typicality. Perhaps factors such as these enable rapid decision-making in a URVC task as well.

The results of Experiment 3 leave important questions to be answered. How much exposure time is needed to accurately categorize high passed versions of natural images? The finding that participants can use high spatial frequency information to make category decisions given sufficient processing time begs the question of exactly how much time is sufficient. Another question raised by this experiment has to do with which, if any, one range of spatial frequencies is sufficient to perform rapid categorizations. Due to differences in stimulus energy, it is unlikely that masked URVC performance with band passed images would be equivalent to that with full spectrum images, but perhaps another range of spatial frequencies would provide more relevant information than the ranges used in this study. Finally, it may be the case that low spatial frequencies do provide sufficient information to perform masked URVC tasks at ceiling levels, given longer SOAs. Testing performance with the low band passed images used in this study at longer SOAs would be informative to that end.

REFERENCES

Beltran, F. S. & Duque, Y. (1993). Processing typical objects in scenes: Effects of photographs versus line drawings. *Perceptual and Motor Skills, 76,* 307-312.

Beiderman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review, 94,* 115-147.

Biederman, I. & Ju, G. (1988). Surface versus edge-based determinants of visual recognition. *Cognitive Psychology, 20,* 38-64.

Blakemore, C. & Campbell, F. W. (1969). On the existence of neurons in the human visual system selectively sensitive to orientation and size of retinal images. *Journal of Physiology, 203,* 237-260.

Breitmeyer, B. G. (1975). Simple reaction times as a measure of the temporal response properties of transient and sustained channels. *Vision Research, 15,* 1141-1142.

Breitmeyer, B. G. (1984). *Visual Masking: An Integrative Approach.* New York: Oxford University Press.

Breitmeyer, B. G., & Ganz, L. (1976). Implications of sustained and transient channels for theories of visual pattern masking, saccadic suppression, and information processing. *Psychological Review, 83,* 1-36.

Breitmeyer, B. G. & Ogmen, H. (2000). Recent models and findings in visual backward masking: A comparison, review and update. *Perception and Psychophysics, 62,* 1572-1595.

Davidoff, J. B. & Ostergaard, A. L. (1988). The role of colour in categorical judgments. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 40,* 533-544.

Dell'Acqua, R., Job, R., & Grainger, J. (2001). Is global shape sufficient for automatic object identification? *Visual Cognition, 8,* 801-821.

Delord, S. (1998). Which mask is most efficient: A pattern or noise? It depends on the task. *Visual Cognition, 5,* 313-338.

Delorme, A., Richard, G., & Fabre-Thorpe, M. (1999). Rapid processing of complex natural scenes: A role for the magnocellular visual pathways? *Neurocomputing, 26-27,* 663-670.

Delorme, A., Richard, G. & Fabre-Thorpe, M. (2000). Ultra-rapid visual categorization of natural scenes does not rely on colour cues: A study in monkeys and humans. *Vision Research, 40,* 2187-2200.

DeValois, R. L. & DeValois, K. K. (1988). *Spatial Vision.* New York: Oxford University Press.

Enns, J. T., & Di Lollo, V. (1997). Object substitution: A new form of masking in unattended visual locations. *Psychological Science, 8,* 135-139.

Enns, J. T., & Di Lollo, V. (2000). What's new in visual masking. *Trends in Cognitive Sciences, 4,* 345-352.

Fabre-Thorpe, M., Delorme, A., Marlot, C., & Thorpe, S. (2001). A limit to the speed of processing in ultra-rapid visual categorization of novel natural scenes. *Journal of Cognitive Neuroscience, 13,* 171-180.

Forde, E. M. E., & Humphreys, G. W. (1999). Category-specific recognition impairments: a review of important studies and influential theories. *Aphasiology, 13,* 169-193.

Francis, G. (2003). Developing a new quantitative account of backward masking. *Cognitive Psychology, 46,* 198-226.

Green, D. M., & Swets, J. A. (1966). *Signal Detection Theory and Psychophysics*. New York: Wiley.

Humle, M. R., & Merikle, P. M. (1976). Processing time and memory for pictures. *Canadian Journal of Psychology, 30,* 31-38.

Hummel, J. E., & Biederman, I. (1992). Dynamic binding in a neural network for object recognition. *Psychological Review, 99*, 480-517.

Larochelle, S., Richard, S., & Soulieres, I. (2000. What some effects might not be: The time to verify membership in "well-defined" categories. *Quarterly Journal of Experimental Psychology, 53A,* 929-961.

Laws, K. R., & Neve, C. (1999). A 'normal' category-specific advantage for naming living things. *Neuropsychologia, 37,* 1263-1269.

Lei, F. F., VanRullen, R., Koch, C. & Perona, P. (2002). Rapid natural scene categorization in the near absence of attention. *Proceedings of the National Academy of Sciences, 99,* 9596-9601.

Levin, D. T., Takarae, Y., Miner, A. G., & Keil, F. (2001). Efficient visual search by category: Specifying the features that mark the difference between artifacts and animals in preattentive vision. *Perception and Psychophysics, 63,* 676-697.

Lichtenstein, M. (1961). Phenomenal simultaneity with irregular timing of components of the visual stimulus. *Perceptual and Motor Skills, 12,* 47-60.

Loftus, G. R. & Bell, S. M. (1975). Two types of information in picture memory. *Journal of Experimental Psychology: Human Learning and Memory, 104,* 103-113.

Marr, D. (1982). *Vision.* San Francisco: Freeman

Merigan, W.H. and Maunsell, J.H.R. (1993). How parallel are the primate visual pathways?. *Annual Review of Neuroscience, 16*, 369–402.

Miller, J. & Navon, D. (2002). Global precedence and response activation: Evidence from LRPs. *The Quarterly Journal of Experimental Psychology, 55A,* 289-310.

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

Navon, D. (1977). Forest before the trees: The precedence of global features in visual perception. *Cognitive Psychology, 9,* 353-383.

Nelson, T. O., Metzler, J. & Reed, D. A. (1974). Role of details in the long-term recognition of pictures and verbal descriptions. *Journal of Experimental Psychology, 102,* 184-186.

Nowak, L. G., Munk, M. H., Girard, P., & Bullier, J. (1995). Visual latencies in areas V1 and V2 of the macaque monkey. *Visual Neuroscience, 12,* 371-384.

Oliva, A. & Schyns, P. G. (1997). Coarse blobs or fine edges? Evidence that information diagnosticity changes the perception of complex visual stimuli. *Cognitive Psychology, 34,* 72-107.

Oliva, A. & Schyns, P. G. (2000). Diagnostic colors mediate scene recognition. *Cognitive Psychology, 41, 176-210.*

Ostergaard, A. L. & Davidoff, J. B. (1985). Some effects of color on naming and recognition of objects. *Journal of Experimental Psychology: Learning, Memory and Cognition, 11,* 579-587.

Parker, D. M., Lishman, J. R. & Hughes, J. (1992). Temporal integration of spatially filtered images. *Journal of Experimental Psychology: Human Perception and Performance, 22,* 147-160.

Parker, D. M., Lishman, J. R. & Hughes, J. (1997). Evidence for the view that temporospatial integration in vision is temporally anisotropic. *Perception, 26,* 1169-1180.

Peyrin, C., Chauvin, A., Chokron, S., & Marendaz, C. (2003). Hemispheric specialization for spatial frequency processing in the analysis of natural scenes. *Brain and Cognition, 53,* 278-282.

Reinagel, P. & Reid, R. C. (2000). Temporal coding of visual information in the thalamus. *Journal of Neuroscience, 20,* 5392-5400.

Ro, T., Breitmeyer, B. G., Burton, P. Singhal, N. S. & Lane, D. (2003). Feedback contributions to visual awareness in human occipital cortex. *Current Biology, 11,* 1038-1041.

Rolls, E. T. & Tovée, M. J. (1994). Processing speed in the cerebral cortex and neurophysiology of visual masking. *Proceedings of the Royal Society of London B, 257,* 9-15.

Rolls, E. T., Tovée, M. J., & Panzeri, S. (1999). The neurophysiology of backward masking: Information analysis. *Journal of Cognitive Neuroscience, 11,* 300-311.

Sanocki, T., Bowyer, K., Heath, M. & Sarkar, S. (1998). Are edges sufficient for object recognition? *Journal of Experimental Psychology: Human Perception and Performance, 24,* 340-349.

Schyns, P. G., Bonnar, L., & Gosselin, F. (2002). Show me the features! Understanding recognition from the use of visual information. *Psychological Science, 13,* 402-409.

Schyns, P. G. & Oliva, A. (1994). From blobs to boundary edges: Evidence for time- and spatial-scale-dependent scene recognition. *Psychological Science, 5,* 195-200.

Schyns, P. G. & Oliva, A. (1999). Dr. Angry and Mr. Smile: When categorization flexibly modifies the perception of faces in rapid visual presentations. *Cognition, 69,* 243-265.

Sperling, G. (1960). The information available in brief visual presentation. *Psychological Monographs, 74,* 29.

Thorpe, S. (2002). Ultra-Rapid Scene Categorisation with a Wave of Spikes. In H.H. Bulthoff et al (eds), *Biologically Motivated Computer Vision, Lecture Notes in Computer Science, 2525,* pp1-15, Springer-Verlag, Berlin.

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

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

Tversky, B. & Baratz, D. (1985). Memory for faces: Are caricatures better than photographs? *Memory and Cognition, 13,* 45-49.

VanRullen, R. Delorme, A., & Thorpe, S. J. (2001). Feed-forward contour integration in primary visual cortex based on asynchronous spike integration. *NeuroComputing, 26-27,* 911-918.

VanRullen, R., Gautrais, J., Delorme, A., & Thorpe, S. J. (1998). Face processing using one spike per neuron. *Biosystems, 48,* 229-239.

VanRullen, R. & Koch, C. (2003). Visual selective behavior can be triggered by a feed-forward process *Journal of Cognitive Neuroscience, 15,* 209-217.

VanRullen, R., & Thorpe, S. J. (2001a). The time course of visual processing: From early perception to decision-making. *Journal of Cognitive Neuroscience, 13,* 454-461.

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

VanRullen, R., & Thorpe, S. J. (2002). Surfing a spike wave down the ventral stream. *Vision Research, 42,* 2593-2615.

Wichmann, F. A., Sharpe, L. T., & Gegenfurtner, K. R. (2002). The contributions of color to recognition memory for natural scenes. *Journal of Experimental Psychology: Learning, Memory and Cognition, 28,* 509-520.

Wurm, L. H., Legge, G. E., Isenberg, L. M. & Luebker, A. (1993). Color improves object recognition in normal and low vision. *Journal of Experimental Psychology: Human Perception and Performance, 19,* 899-911.

APPENDIX A

A SPIKE-BASED ACCOUNT FOR ULTRA-RAPID VISUAL CATEGORIZATION

Thorpe and his colleagues have argued that URVC is the result of predominantly feed-forward visual processing. However, the speed at which this complex process occurs is inconsistent with current models of vision. In order to account for the findings in the URVC literature, VanRullen and Thorpe (2002) propose a new, potentially revolutionary model.

VanRullen and Thorpe's model is of particular interest because its fundamental premise is contrary to popular thought regarding how information is coded in neural processing. In most current models neurons, or neuron-like processing units, are bound together by a series of weighted connections, and the activation level of each individual unit is represented by the sum of the inputs from these connections. More importantly, this activation level is actually a representation of the unit's firing rate. The idea that a neuron's firing rate can be used to measure its activation has long been held in neuroscience. However, the findings from the URVC literature are inconsistent with this idea; the processing involved in performing this sort of task could not possibly occur at the speed it does if it were dependent upon multiple firings of a single neuron. ERP recordings taken during URVC tasks indicate the processing necessary to make rapid category decisions is completed in 150 ms (Thorpe et al., 1996). VanRullen and Thorpe argue that due to the complexity of the task, in order to complete the series of neural firings necessary to perform it in 150 ms, any given neuron would have enough time for only one firing. If this is the case, the firing rate of a single neuron is not a viable means of communicating information.

Rate coding is not the only option. VanRullen and Thorpe suggest one alternative, based not on the frequency with which a neuron fires, but on its latency to fire. The time required for a neuron to fire is dependent upon the strength of its stimulation; the more strongly a neuron is stimulated, the more quickly it will reach its threshold and fire. It may be the case that information is carried by the pattern of firing of many neurons, rather than the rate of firing of single neurons. In this scenario, the pattern of neural firing could be operationalized as the order in which certain neurons fire relative to each other. This idea is the basis of Rank Order Coding, their account for the rapid processing of natural scenes as well as vision in general.

Rank Order Coding has several advantages over traditional firing rate models, in addition to its effectiveness as an account for our ability to rapidly categorize scenes. If information is coded relative to the timing of the firing of a large population, the number of possible patterns increases with the factorial of the number of neurons involved, which would allow for a single population of neurons to transmit vast amounts of information (VanRullen & Thorpe, 2002). Although any instantiation of Rank Order Coding would necessarily include units not normally incorporated in theories of vision, such as units responsible for interpreting the delay of neurons relative to each other, VanRullen and Thorpe argue the existence of these units is not unreasonable and can be instantiated using neurons that only fire when its inputs fire in a particular order.

Using the idea that neural information can be communicated via the relative timing of many inputs, VanRullen and Thorpe propose a new theory of vision consisting of several stages. At each stage, a wave of neural input is propagated and refined. Beginning in the retina, receptive fields are stimulated according to the contrast and spatial frequency of a particular area of the visual field. The amount of stimulation in each receptive field determines the order in

which the neurons will fire. Across the retina, a wave of neural activity is formed, with the most heavily stimulated areas firing first, followed by less stimulated areas. Stimulation is determined by contrast and spatial frequency, so as a result, the neurons associated with higher changes in local luminance and lower spatial frequencies will be given priority. As the wave passes to the next level of processing, neurons that are sensitive to the order in which input arrives will respond selectively to a specific waveform. As they are stimulated, they too will fire, with the neurons receiving their ideal stimulus firing first, and ones that match less firing progressively later. Thus the wave, in a form very similar to how it arrived, is propagated to the next level of processing. Importantly, throughout the entire process of spike wave formation and propagation, each neuron only fires once, thus allowing maximal information to be processed very rapidly.

It is known that processes such as filling-in and contour integration involve lateral communication between neurons. Although it may be the case that reiterative processes may be responsible for these interactions, VanRullen, Delorme, and Thorpe (2001) developed a model of lateral influences that operates in an entirely feed-forward fashion. The crux of the model is that the earliest firing cells, which carry the most important information, influence their neighbors while they are still awaiting input from later incoming neurons. Thus, information about contour alignment can be integrated into the next resulting wave via a totally data-driven, temporal mechanism.

While wave-propagation takes place, neural responses are modulated with respect to higher-level, task related goals by an attentional mechanism. This mechanism alters the shape the propagating wave by temporarily altering receptor thresholds. Specifically, if attention is directed at one region of the visual field, the temporal thresholds of the receptors particular to that area will be lowered, enabling them to fire more quickly, despite receiving slower

information from the incoming wave. In the resulting wave, this information will be carried first. Given that in a spike wave the most important information arrives first, this attentional mechanism is able to functionally determine which information is the most important. Although attention in this model is still a top-down process, it does not constitute a feedback loop; that is, the attentional mechanism is not itself influenced by the incoming wave.

The model thus far can account for a single incoming wave, but the visual array is perpetually in a state of flux, and somehow the visual system's ability to distinguish between the tail end of one wave and the beginning of another needs to be accounted for. To do this, VanRullen and Thorpe propose a "perceptual frame" (a la Lichtenstein, 1961), which is the minimum interval between two temporally separate events for which they are perceived as one. This amount of time is 40 ms or less (Reinagel & Reid, 2000), which is just enough time for the propagation of a single spike wave. If each wave is contained within a single perceptual frame, the visual system could easily discriminate between many waves.

This model of visual processing based on Rank Order Coding has yet to be tested empirically. However, aspects of it have been tested through a computer model, SpikeNet. SpikeNet is essentially a model for a secondary relay center, which receives predetermined input in the form of a spike wave. After slightly modifying the wave, it is sent on to another relay center, which roughly approximates V1. Using lateral interactions, this layer shows selectivity to edges and orientations. These outputs, in turn, drive a series of feature detector maps that respond selectively to particular temporal waveforms. When trained on specific images, SpikeNet successfully identifies faces (VanRullen, Gautrais, Delorme, & Thorpe, 1998) and more recently has learned to localize and identify various types of stimuli in a montage of 51

natural images. Furthermore, SpikeNet is size and position invariant in its responding. Most impressively, all this is done with one firing per cell on a simple desktop computer.

One of SpikeNet's shortcomings is that in order to successively model this process, 1.5 billion synaptic connections between 3.5 million neurons (roughly one per pixel) are necessary. Thorpe (2002) himself notes this model is not particularly realistic, as humans are able of recognizing well over 51 images and the corresponding cost in neurons would be astronomical. However, he points out the purpose of this model is not necessarily biological realism, but rather a demonstration of a functional instantiation of Rank Order Coding in a functional artificial agent. Furthermore, SpikeNet serves as a call for researchers to explore the possibility of Rank Order Coding as an alternative to the traditional rate coding assumed by neuroscience today.

VanRullen and Thorpe's (2001) model raise a valid point regarding the assumption that speed of a neuron's firing reflects its level of stimulation. One of the difficulties with this model is the lack of experimental evidence in vivo to support it, by the same token, there is no evidence falsifying it. In fact, on the retinal level, the properties of the initial spike wave are consistent with evidence that luminance and low frequency-based information enjoy temporal priority (e.g., Breitmeyer, 1975). Furthermore, the idea that the most important visual information for object recognition could vary based on attentional mechanism is consistent with the finding that different spatial scales are involved in face recognition based on the task at hand (Shyns & Oliva, 1999; Shyns, Bonnar, & Gosselin, 2002).

Experiment 3 in this study speaks directly to this model. If URVC depends upon the first information through the system, then low spatial frequency information should be sufficient to allow rapid categorization. In Experiment 3, URVC using low-passed images is compared to performance using high-passed images. If accuracy is higher with low-passed images, the results

would be consistent with this model of vision, and would indicate that category judgments regarding complex images is dependent upon high contrast, low spatial frequency information.

FOOTNOTE

1. There is some debate as to the exact means by which backward masking interrupts visual processing. Traditionally, a mask is thought to obliterate a target by disrupting the ongoing generation of a representation of it (Breitmeyer, 1984; Rolls & Tovée, 1994). Therefore, according to this model, backward masking affects primarily bottom-up processing of incoming information. Although some of the processes involved in object recognition have had time to take place, the introduction of the mask prevents the completion of this process, and the observer is therefore unable to perceive the target. Recently, some authors have argued for the role of re-entrant processing in the formation of object representations (Enns & Di Lollo, 2000). According to these authors, backward masking is effective because it disrupts top-down processes that search for a match between a high-level object representation and low-level ongoing processes. When such a match is found, the representation is "locked in." This argument has been used to explain the four-dot masking effects reported by Enns and Di Lollo (1997) as well as the attentional blink and some forms of change blindness (Enns & Di Lollo, 2000).

This recent account of backward masking seems incompatible with Thorpe and his colleagues' model of object recognition. By this model, the majority of object recognition processing takes place in a single, feed-forward wave (see Appendix A). If this were the case, and if backward masking affects feedback mechanisms, one would expect to find no masking whatsoever in this study.

Although backward masking is used in this study, its purpose is not to clarify this issue. Indeed, the experiments described here are not designed to investigate the mechanisms underlying backward masking, and will not be informative to that end. Likewise, these

experiments do not provide a direct test of Thorpe and his colleagues' proposal that object recognition can take place using only feed-forward mechanisms. Backward masking is employed here as a means of disrupting the generation of a full representation of visual scenes, however it may occur.