PERGAMON

Vision Research 41 (2001) 1931–1950

Vision Research

www.elsevier.com/locate/visres

Size tuning in the absence of spatial frequency tuning in object recognition

József Fiser [a,*], Suresh Subramaniam [b], Irving Biederman [b]

[a] *Department of Brain and Cognitive Sciences, Center for Visual Science, University of Rochester, Rochester, NY 14627-0268, USA*
[b] *Department of Psychology, University of Southern California, Los Angeles, CA 90089, USA*

Received 29 January 2000; received in revised form 16 August 2000

Abstract

How do we attend to objects at a variety of sizes as we view our visual world? Because of an advantage in identification of lowpass over highpass filtered patterns, as well as large over small images, a number of theorists have assumed that size-independent recognition is achieved by spatial frequency (SF) based coarse-to-fine tuning. We found that the advantage of large sizes or low SFs was lost when participants attempted to identify a target object (specified verbally) somewhere in the middle of a sequence of 40 images of objects, each shown for only 72 ms, as long as the target and distractors were the same size or spatial frequency (unfiltered or low or high bandpassed). When targets were of a different size or scale than the distractors, a marked advantage (pop out) was observed for large (unfiltered) and low SF targets against small (unfiltered) and high SF distractors, respectively, and a marked decrement for the complementary conditions. Importantly, this pattern of results for large and small images was unaffected by holding absolute or relative SF content constant over the different sizes and it could not be explained by simple luminance- or contrast-based pattern masking. These results suggest that size/scale tuning in object recognition was accomplished over the first several images (< 576 ms) in the sequence and that the size tuning was implemented by a mechanism sensitive to spatial extent rather than to variations in spatial frequency.

Keywords: Size invariance; Spatial frequency; Human object recognition; Coarse-to-fine tuning; Rapid serial visual presentation

1. Introduction

Humans and animals can identify objects appearing at a variety of sizes in their visual field without much apparent cost. This apparent invariance over size changes poses a challenge to computational theories of visual recognition because the early cortical representation of features appearing at different sizes will be very different. For example, a slightly rounded L-shaped corner of an object, when the object is shown at a small size, will activate feature detectors sensitive to sharp curves at a given scale. The image of the same object at a large size might not activate curve detectors at that scale at all.

The major account of how the visual system handles images of different sizes, which we will refer to as the 'Scaling Hypothesis', is based on the multiscale spatial frequency representation characteristic of V1 (Sergent, 1986; Schyns & Oliva, 1994; Hughes, Nozawa, & Kitterle, 1996). Specifically, whenever a new object appears in the scene the visual system performs a coarse-to-fine tuning, in that additional time is required for information at higher spatial frequencies to become available and information presented at low spatial frequencies guide the processing of information in higher frequencies. Olshausen, Anderson, and Van Essen (1993; 1995) proposed a specific routing mechanism, a 'shifter circuit', that implements 'coarse-to-fine' tuning in the visual system. This shifter circuit selects the most salient information on the lowest scale and adjusts the size and sampling of the input window to higher processing centers based on this scale to achieve a size-normalized representation of a given object in the scene. Such a

* Corresponding author. Tel.: +1-716-2757259; fax: +1-716-4429216.
E-mail addresses: fiser@bcs.rochester.edu (J. Fiser), bieder@usc.edu (I. Biederman).

0042-6989/01/$ - see front matter
PII: S0042-6989(01)00062-1

tuning mechanism can accommodate voluntary attentional as well as involuntary mechanisms for size- and position-invariant recognition. It is the 'Scaling Hypothesis' assumption of SF-based size tuning that the present study investigates. Specifically, we address the question as to whether efficient responding to images of different sizes requires information to be represented at different scales. Would the pattern of results obtained in tasks that required processing images of various sizes be different if the SF content of the images was held constant over the different sizes?

An alternative to the Scaling Hypothesis for how we code images of various sizes could be tuning that was based directly on spatial extent, in retinotopic maps, rather than indirectly, by scale. The tuning for different sizes would thus be directly in response to the size variations, insofar as different size images occupy different regions of space. Such a proposal would circumvent the apparent difficulty that the Scaling Hypothesis would have in handling how real-time tuning to large images, e.g. one of 20°, could be made, when such images greatly exceed the receptive fields of the scale-sensitive cells in the earlier stages, i.e. V1, V2 and V4 of the ventral pathway.

The basic assumption behind size tuning, i.e. that a subrange of sizes is selected by a tuning mechanism during visual recognition, implies that not all sizes are equally usable at a given moment. This implication received indirect experimental support from attentional tasks, where the participants had to perform matching between objects of different sizes (Sekuler & Nash, 1972; Bundesen & Larsen, 1975; Larsen & Bundesen, 1978; Ward, 1982; Larsen, 1985; Cave & Kosslyn, 1989) or localization and identifying a target of various sizes (Farell & Pelli, 1993). The general conclusion of these studies (with the exception of Farell & Pelli, 1993) was that participants were unable to attend to two targets of different sizes at the same time and their performance in matching objects monotonically declined as the difference in the sizes of the targets increased. Using several experimental methods, Farell and Pelli (1993) found that the ability to attend to different sizes at the same time depended on the nature of the task. Participants were able to attend to both small and large sizes in identification tasks, but they failed to do so in localization tasks.

A similar capacity of the visual system to voluntarily select a limited spatial frequency range was proposed by Julesz (1975). Indeed, several studies found improved performance in detecting sinusoid gratings when the participants knew the spatial frequency at which the grating would appear (Graham, Robson, & Nachmias, 1978; Davis & Graham, 1981). Similarly, Shulman and Wilson (1987) found that when participants focused on the global structure of a compound, their detection of low SF gratings at low contrast was facilitated, whereas focusing on the local structure of the compound stimulus facilitated detection of high spatial frequency gratings. These findings set the stage for the Hughes et al. (1996) hypothesis that linked spatial frequency and size: SF-based size tuning might govern involuntary object recognition processes.

Interestingly, none of the above studies interpreted as evidence for the Scaling Hypothesis provided an explicit test of SF-based tuning for size in object recognition. Either they failed to control independently the size and SF of the stimuli or their task was not that of object recognition but rather one in which participants detected or matched a very limited set of simple stimuli.

We present here the results of two single-presentation and six rapid serial visual presentation (RSVP) object-identification experiments, in which the size and scale of the object images were manipulated independently. In the RSVP experiments, on each trial a sequence of 40 gray level images of different everyday objects was presented, with each image displayed for 72 ms. The participants were to detect whether a target, specified verbally before each sequence, was present in the sequence. The RSVP paradigm has several advantages. First, it requires the identification of unanticipated complex object images at a speed that approximates the uncertainty and rate of real-time object recognition. Typical voluntary fixations are made at a rate of three to four times per second, but there is considerable evidence that recognition is generally achieved within the first 100 ms (Thorpe, Fize, & Marlot, 1996), with the additional time for the fixation required for memory consolidation (Subramaniam, Biederman & Madigan, 2000). More importantly, the high rate allows performance differences to be manifested and lessens the likelihood of concluding that two perceptual inputs are processed simultaneously, when in reality they are processed sequentially.

The second advantage of using the RSVP paradigm was that it allowed us to test entry-level object recognition rather than a specialized visual search or matching task. Two types of sequences were used in all of our RSVP experiments: 'homogeneous' sequences in which all the images were of the same size and/or SF and 'heterogeneous' sequences, which were identical to homogeneous sequences except that one, and only one, image (half of the time the target) was of a different size or SF than all the other images in the sequence. The rationale for using an RSVP task was that under such conditions for speeded recognition, participants would have no recourse, without any instruction, but to employ their natural strategy for optimal recognition, whether it involved tuning for a particular size/SF or not.

The third advantage of the RSVP paradigm was that independent size and scale manipulation could be performed naturally within the paradigm by resizing and/

or filtering the original object images without changing the task. This allowed an explicit test of whether the spatial frequency content of the image was what governed the perception of objects at different sizes.

The present investigation assessed whether: (a) tuning to images of different sizes during the RSVP sequences could be achieved so efficiently that it eliminated the advantage in the recognition of large over small images evident in the single-presentation tasks; and (b) this tuning to small sizes could be accomplished in the absence of SF information correlated with the size changes.

To address these issues we first established, in Experiments 1 and 2, whether there was any difference in recognition performance, as manifested by RTs and error rates, between suprathreshold object images of different sizes and spatial frequency content in single-presentation (i.e. non-RSVP) naming or verification tasks. With the size and SF values that we used, we found that large images were more accurately and quickly named and LowSF better verified than small or HighSF images, respectively. Next, in two RSVP experiments (Experiments 3 and 4) we tested whether performance in verifying a target when it was embedded in a stream of similar images showed the same size/SF differences that were apparent in the single presentations. We found that in homogeneous sequences, the Large/LowSF advantage was lost, but that in the heterogeneous sequences, the accuracy of recognizing a Large/LowSF target among Small/HighSF distractors increased compared to the homogeneous sequences, whereas Small/HighSF targets among Large/LowSF distractors resulted in large reductions in accuracy. To investigate whether these interdependencies between target and distractors could be explained by low level masking, such as the greater change in luminance and/or the larger spatial cover of the large images over the small images, we ran two RSVP tests with interleaved masks between images (Experiments 5 and 6). The result of these studies ruled out the possibility of explaining the results of Experiment 3 by low-level luminance or contrast masking. Finally, in the last two RSVP tests (Experiments 7 and 8), we explicitly tested the notion of SF-based size tuning by changing image sizes but keeping the absolute (cpd) SF content or the relative (cycle per object) SF content of the images unchanged during the switch. In both experiments, we found evidence against the notion that tuning to appropriate size is accomplished by SF-based mechanisms. Specifically, holding the band of the absolute or relative spatial frequency content constant did not affect the pattern of costs and gains evident when switching to a different size.

2. General method

Fig. 1 summarizes the different types of stimuli used in the experiments. The same small and large unfiltered images were used in the single-trial naming experiment (Experiment 1), in the Pure Size RSVP task (Experiment 3) and in the two masked RSVP experiments (Experiments 5 and 6). The same high and low bandpass filtered images were used in the single-trial verification experiment (Experiment 2) and in the pure scale RSVP task (Experiment 4). In Experiment 7 (absolute scale RSVP), both small and large images were filtered at the same high center frequency, so the absolute spatial frequency content was held identical. In Experiment 8 (relative scale RSVP), where the large images were filtered at low whereas the small images were filtered at high SF, the *relative* spatial frequency contents of the small and large images used (defined by cycles per object, see Harmon & Julesz, 1973; Parish & Sperling, 1991; Solomon & Pelli, 1994) were identical.

3. Experiment 1: naming singly presented objects at different sizes

The goal of both Experiments 1 and 2 was to establish a baseline condition for the relative difficulty of large- and small-sized images and low- and high-SF passed images for the RSVP experiments.

3.1. Stimuli and method

Twenty-four images of everyday objects, each in two sizes, were used in this experiment.[1] The largest extent of the object images was scaled to 1.2° of visual angle for the small images and to 6° for the large images (Fig. 1). The gray scale images were obtained by video frame capturing of real objects in natural lighting and then replacing the background of the images with a white homogeneous background. Because of the large size difference and white backgrounds the mean luminance and the RMS, contrast of the images were assessed by considering only pixels belonging to the object and discarding the white background pixels. By this measure, the mean luminance and RMS contrast of large and small images was comparable (Fig. 2).[2] The

[1] Images of the following objects (24 targets and eight buffers) were used in all the experiment: airplane, battery, bottle, brush, bus, calculator, camera, cap, car, comb, flashlight, glass, goggles, gun, hanger, headphones, helicopter, key, knife, light bulb, lock, paperclip, pot, saw, scissors, screw, shoe, spatula, stapler, sword, tractor, truck. These images can be obtained from the authors.

[2] Calculations based on the entire image yielded similar relative results, in particular, mean luminance of 66.3 vs. 66.2 cd/m^2 and RMS contrast of 32.65 vs. 30.8 cd/m^2 for the large and small images, respectively.

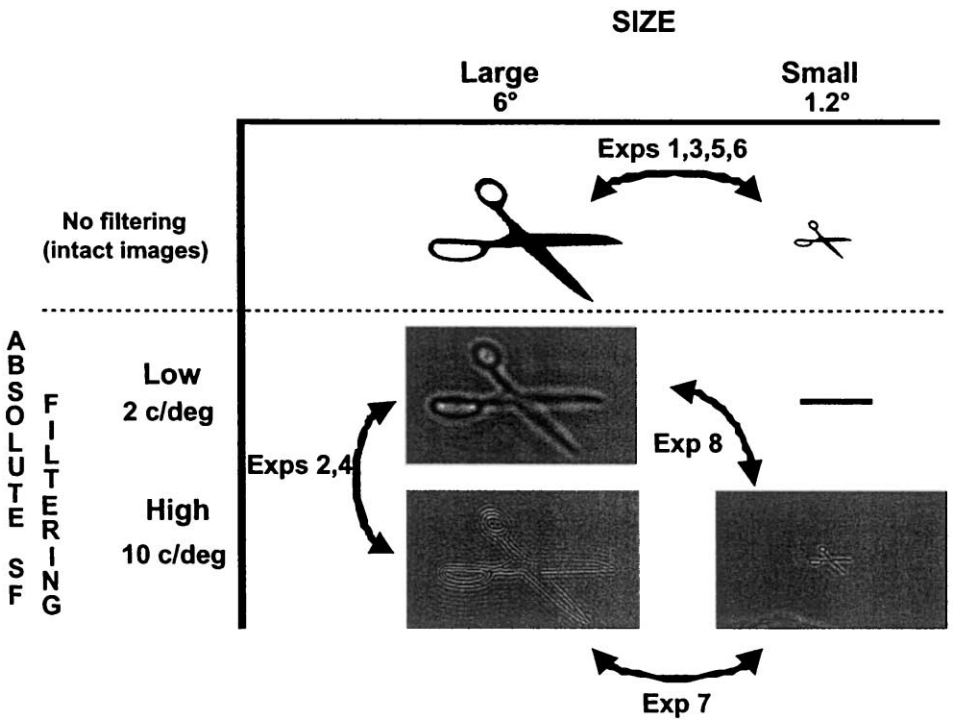

Fig. 1. The stimuli used in the different experiments. Size change (in terms of maximum extent of the object) is represented on the horizontal axis; changes in spatial frequency filtering (in center frequency) on the vertical axis. The filtering at 2 and 10 cpd was performed with 1.5 octave bandwidths. The first single-presentation naming experiment, the Pure Size RSVP and the two masking RSVP experiments (Experiments 1, 3, 5 and 6) used unfiltered large and small images. The second single presentation verification experiment and the Pure Scale RSVP Experiments (Experiments 2 and 4) used large size images filtered at the two center frequencies. The Absolute RSVP Experiment (Experiment 7) used the large and small images filtered ≈ 10 cpd, whereas the Relative RSVP Experiment (Experiment 8) used different center frequencies in proportion with the size changes between large and small images, that is, the large, 2 cpd and small, 10 cpd images. The only untested combination of variables, small images filtered ≈ 2 cpd, could not be tested as such images would be unidentifiable blobs.

average mean luminance of the large and small masks used in the experiment were 34.34 and 36.35 cd/m^2 and the average RMS contrast were 21.09 and 21.53 cd/m^2, respectively.

Sixteen naïve subjects participated in the experiment. Each subject saw 24 different images, 12 large and 12 small ones, in a randomly mixed sequence of 24 trials. The trial sequences were balanced across participants by having different images in the large and small conditions. The trials were self-paced, initiated by a mouse press. On each trial, a small fixation point appeared on the screen for 500 ms followed by the object image for 72 ms and then by a mask of large size for 500 ms (Fig. 3). Four gray scale masks, created by superimposing small segments of different images and random patterns blurred to different degrees, were used in random order. The spatial frequency spectrum of the masks closely matched that of the target images. The use of varying masks was based on the Intraub (1981) finding that masking effectiveness is diminished by repetition of a mask.

The participants were instructed to name the object after its appearance as quickly as possible using common basic-level category names. The participants were not familiarized with the categories prior to the experi-

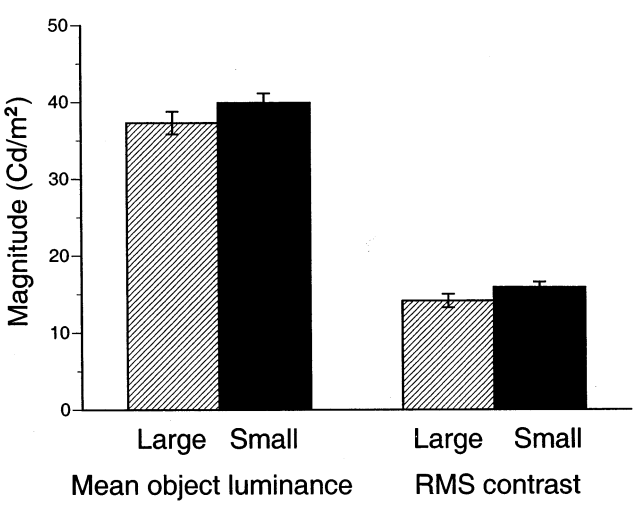

Fig. 2. Mean luminance and RMS contrast of the large and small unfiltered images used in Experiment 1. The same images were used in Experiments 3, 5 and 6. (Error bars show S.E.).

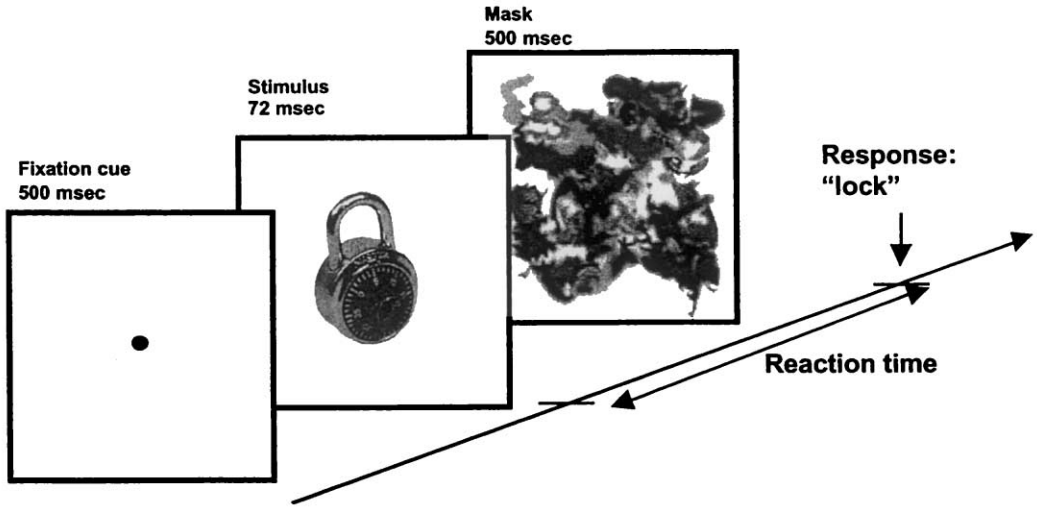

Fig. 3. Schematic representation of one trial in Experiment 1. Targets were everyday objects against a white background. Masks were composed of small fragments of images. Squares represent time frames, i.e. there were no black frames around the cue, the target or the mask.

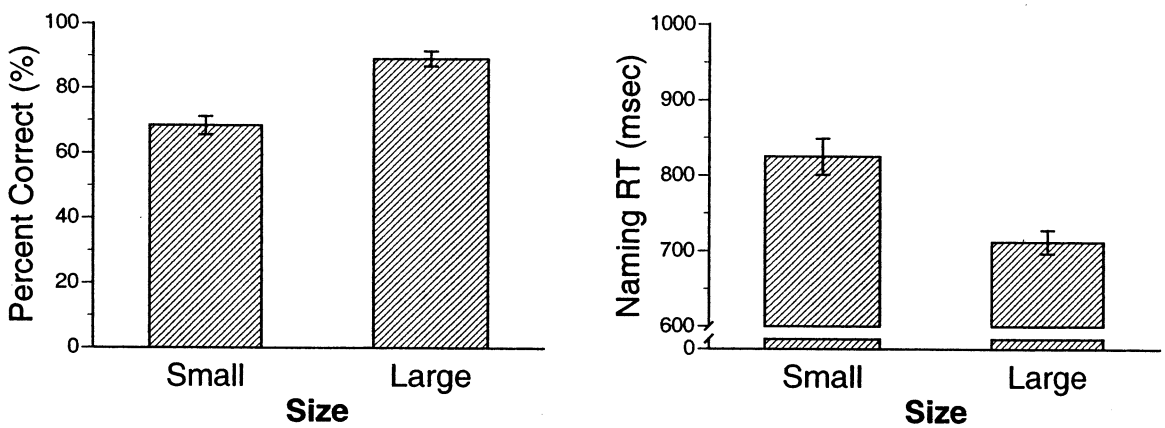

Fig. 4. Results of the single-presentation naming experiment with small and large images (Experiment 1). Both percent correct responses and RTs showed a significant advantage in the naming of large-size images over small-size images. (Note the break in the y-axis on the right panel. Error bars show S.E.).

ment.[3] A delay in naming longer than 2.5 s was counted as an error. Errors were recorded manually and reaction times of the correct responses were measured by means of a voice key (Lafayette Inst. Co. Model: 18010). There were ten practice trials with images not used in the main experiment, to familiarize the participants with the experimental procedure.

3.2. Results and discussion

Fig. 4 presents the results of Experiment 1. Naming accuracy was significantly higher for large- than small-sized stimuli, $t(15) = 5.97$, $P < 0.001$. In addition, the correct naming reaction times (RTs) of the small images were significantly longer than those of the large images, $t(15) = 6.86$, $P < 0.001$. These findings agree with the Biederman and Cooper (1992) results that the naming of 100-ms object images followed a U-shaped curve with an optimum at $\approx 5°$ maximal extent of the image. According to this curve, images with a size of $\approx 1.2°$ would be expected to be harder to identify than those with a maximal extent of 6°. Notice that these results do not suggest that a smaller image is *always* harder to identify than a large one. In fact, the inverted U-shape reported by Biederman suggests that it is possible to select a smaller image which is easier to name. Our results do not identify the cause of the difference in performance either. This difference could come from differences in stimulus information, specifics of the experimental setup or selection of the particular images. Our results simply establish a performance baseline under our experimental conditions and confirm that the recognition performance with our images follows the commonly observed pattern of better identification of large than small images.

[3] A number of studies showed that reading the names of objects a few minutes before naming pictures or judging whether they are real objects or not has no effect on RTs or error rates (Kroll & Potter, 1984; Biederman, 1987).

4. Experiment 2: verifying singly presented low and high bandpass filtered images

Our second baseline study, Experiment 2, focused on human recognition performance with bandpass filtered images. In general, recognition of filtered images depends strongly on the actual image content, the task, the center frequency and the bandwidth of the filtering operation, the presentation time and other parameters of the experiment. The purpose of this experiment was to assess whether there would be an advantage of low-passed images in object recognition tasks at stimulus durations of 72 ms. Such an outcome would be compatible with coarse-to-fine tuning given the task parameters: Under the assumption that low frequency information is available earlier in time than high frequency information, greater accuracy can be achieved under brief presentations with low-passed images.

An advantage of low spatial frequencies, as measured by reaction time differences, is known to exist in detecting the onset of suprathreshold gratings (Breitmeyer, 1975). Several experiments have found, however, that in more complex visual tasks, such as letter or face recognition, low-passed images did not always have an advantage over high-passed images. Rather, identification depended on the wavelengths of the band relative to the size of the stimulus and this relationship was characterized by an inverted U-shape with an optimal band (Parish & Sperling, 1991; Solomon & Pelli, 1994; Costen, Parker & Craw, 1996). The cause of this inverted U-shape with an optimal band could be either the more efficient use of available information or simply more information in the optimal band (Gold, Bennett, & Sekuler, 1999), but the present test does not address this issue. Rather it seeks to establish whether recognition of the low and high band-pass filtered versions of our images show a pattern similar to that of the unfiltered large and small images under the present experimental conditions.

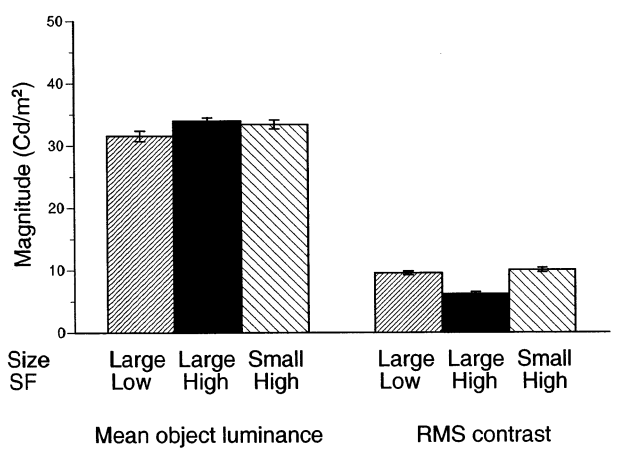

Fig. 5. Mean luminance and RMS contrast of the large and small filtered images used in the experiments. In Experiments 2 and 4, the large, low-pass filtered and the large, high-pass filtered images were used (the first two bars in the two groups). In Experiment 7, large and small high pass filtered images were used, whereas in Experiment 8 large low-pass and small-high pass images were used. (Error bars show S.E.).

4.1. Stimuli and method

Only the large size images of Experiment 1 were used in this experiment. Two versions of each image were generated by bandpass filtering the image around a center frequency of 2 and 10 cpd, with a 1.5 octave bandwidth. This filtering generated an octave wide gap in the spatial frequency spectrum between the two types of images ensuring that different spatial frequency channels would carry the relevant information for the different scaled images.

Thirty-two naïve participants participated in the experiment. A verification task was used instead of the naming task used in Experiment 1, because at 72 ms, identification of filtered images is very difficult: A pilot experiment found naming accuracy to be lower than 35%, rendering the RTs for naming unreliable. As in Experiment 1, each subject viewed 24 images, 12 low-passed and 12 high-passed, in a randomly mixed sequence of 24 trials, with the trial sequences balanced across participants. On each trial, first the name of an everyday object appeared for 1 s followed by a fixation cue for 500 ms, then by the object image for 72 ms and finally by a mask for 500 ms. The object name matched the object on half the trials. The cue was a bandpassed version of the fixation spot of Experiment 1 with a bandwidth spanning the gap between the low- and highpass images. The masks were the same as in Experiment 1, carrying information in all spatial frequency channels. Mean luminance and RMS contrast of the images was assessed the same way as in Experiment 1 (Fig. 5).

The participants were instructed to press one of two keys on a button box depending on whether they thought the name and the object image matched or not. There were ten practice trials to familiarize the participants with the experimental procedure.

4.2. Results and discussion

Fig. 6 shows the results of the verification experiment. Analyses of variance (ANOVAs) were computed separately for percent correct and RTs, each with two fixed factors, SF filtering (high versus low) and match (same versus different match of name and object). For percent correct there was a significant advantage of low spatial filtering over high, $F(1,31) = 8.09$, $P < 0.008$, no effect of matching (same versus different trials), $F(1,31) < 1.00$, *ns* and no interaction between filtering and matching, $F(1,31) < 1.00$, *ns*. For correct RTs,

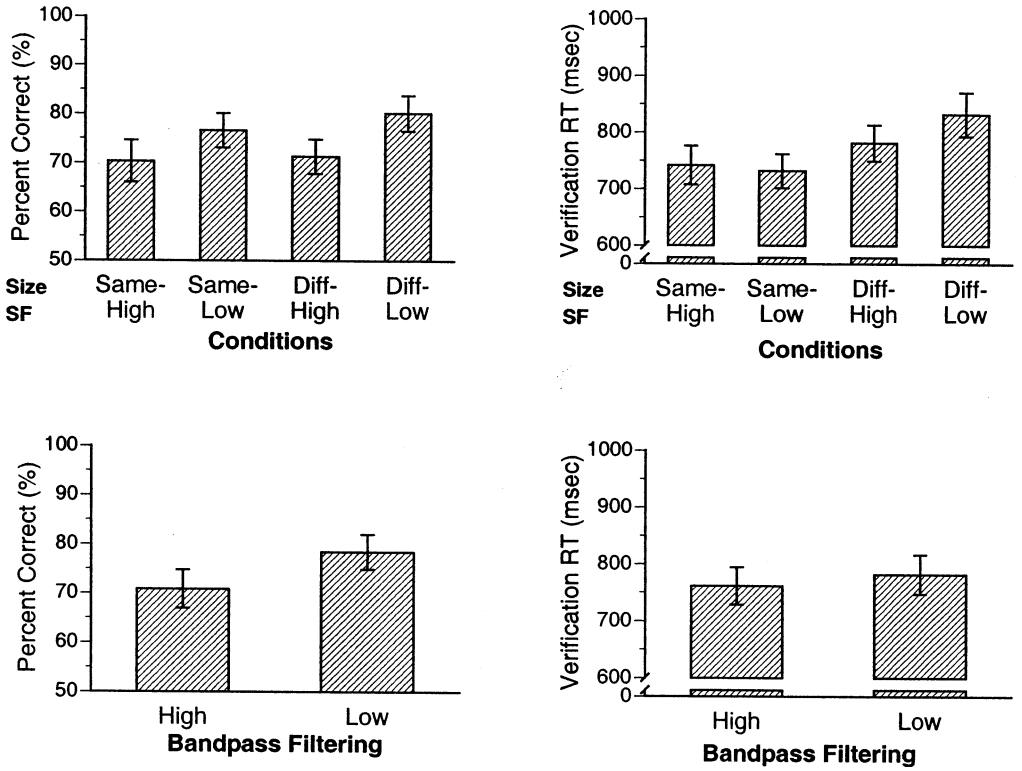

Fig. 6. The results of the single-presentation verification task with bandpass filtered images (Experiment 2). Upper panel: error rates and RTs from Experiment 2. Lower panel: the same results collapsed over same-different conditions. For percent correct, low bandpassed images have a significant advantage, but RTs show no clear advantage. (Error bars show S.E.).

there was a significant advantage of same trials over different ones, $F(1,31) = 11.2$, $P < 0.003$, but no main effect of filtering, $F(1,31) < 1.00$, *ns* and a close to significant interaction between filtering and match, $F(1,31) = 3.47$, $P < 0.07$.

These results establish a strong advantage in the accuracy of identification of lowpass images with our stimuli and allow a test of coarse-to-fine tuning. The nonsignificant difference in RTs could be the result of stimulus selection artifact, since it includes only the correctly verified images. As in the previous experiment with unfiltered images, the results of this experiment do not isolate the cause of the difference in performance between high and low bandpassed images and do not imply that low-pass images are always easier to identify than high-pass images. They merely establish that under our presentation conditions, the common finding of an advantage of low-pass images is obtained and thus can provide a baseline for testing the SF-based size-tuning hypothesis in the subsequent experiments.

5. Experiment 3: identifying objects at different sizes in RSVP sequences

The goal of this experiment (Pure Size RSVP) was to investigate how human recognition performance is affected by the size of the preceding and following distractors when a target is embedded in a sequence of other images and each image has to be identified. Would the advantage of large over small sized images found in Experiment 1 also hold for homogeneous RSVP sequences? Or would adjustments to small images, even though the images were changing, eliminate the advantage enjoyed by large images? What would be the effect on target detection of variation in the relative sizes of target and distractors? There is an asymmetry in the adjustments of contrast sensitivity to variations in suprathreshold contrast such that a transient *increase* in contrast sensitivity is observed in switching from low to high contrast stimuli compared to steady state sensitivity to high contrast, whereas a transient *decrease* in sensitivity compared to steady low contrast sensitivity is observed in switching from high to low contrast (Victor, Conte, & Purpura, 1997). Would a parallel phenomenon occur with size such that performance improvement or decrement would occur depending on the relative sizes of targets and distractors? Or would there be no cost or a uniform cost of target-distractor size disparity?

If there is no effect of the size of the target relative to the size of the surrounding distractors on target recognition performance, the differences between large and small images found in the single trial presentations of

Experiment 1 should be apparent for the targets in this RSVP experiment. If, however, the size of the preceding and following images influence recognition of the target, then the equivalence of target identification in large and small homogeneous sequences might be a reflection of size tuning, whereas differences between homogeneous and heterogeneous sequences could demonstrate the difference between recognition where there is versus is not an opportunity for size tuning.

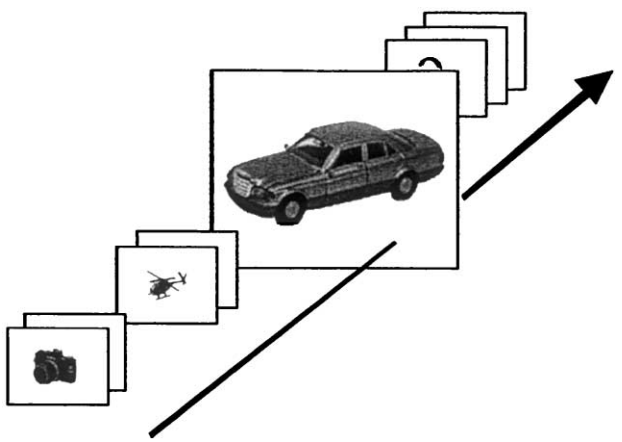

Fig. 7. A heterogeneous trial example from the Pure Size RSVP Experiment (Experiment 3). In a sequence of small images there is one large image in a random position, which might or might not be the target. In the other type of heterogeneous sequence, one image would be small and all the others large. The arrow represents the time axis (there were no visible frames around the images).

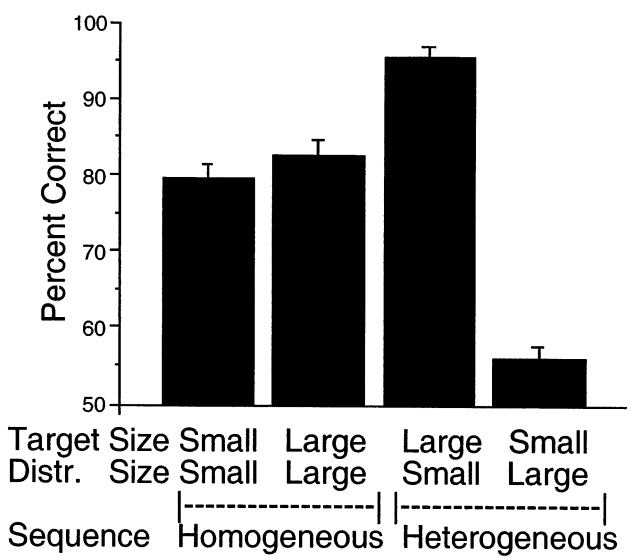

Fig. 8. The results of the Pure Size RSVP Experiment (Experiment 3). Overall target recognition performance in the two homogeneous conditions were not significantly different from each other in contrast with the results in the single-target naming conditions in Experiment 1. Performance in the large target hetergeneous condition was significantly better and performance in the small target heterogeneous condition significantly worse than their corresponding homogenous conditions. Distr. = Distractors. (Error bars show S.E.).

5.1. Stimuli and method

Twenty-four target and eight buffer images scaled to two sizes (6 and 1.2°) were used in this experiment. The target images were the same images used in Experiment 1 with the single presentation conditions. The general procedure of presenting the sequences was identical in all of the RSVP experiments in this study. Each subject viewed 192 sequences. Each sequence consisted of 40 images including eight 'buffer' images at the beginning and eight at the end of the sequence where the target, if present, never appeared. The presentation time for each image was 72 ms with no gap between successive images. There were four types of sequences run in randomized order, so the participants could not know what type of sequence would be presented on any particular trial. Two types of sequences were homogeneous in that all the stimuli in a sequence were of the same size, one with large-sized images and the other with small-sized images. The other two types of sequences were heterogeneous, in which a single image (half of the time the target) differed in size from that of the other images (the distractors) in the sequences (Fig. 7). Eight naïve participants participated in the experiment for credit in the 'Introduction to Psychology' course.

At the beginning of each trial, the participants heard a name of an everyday object. They were instructed to watch the rapid sequence of images and say whether the image of the named object was among the presented stimuli. In half of the trials, the target image was included in the sequence, in the other half of the trials, the image of the verbally specified object never appeared during the experiment. The position of the target image was randomized, but, as noted previously, the target never appeared in the first or last eight positions in the sequence. Avoiding the first eight positions ensured that participants became aware of the size of the images and allowed size tuning (if any) before the target would occur in the sequence. Avoiding the last eight positions excluded the possibility of a benefit from memory recency contributing to the results (e.g. Murdock, 1962). The homogeneous sequences were designed to assess performance when the size of the images preceding the target was a good predictor of the size of the target. The heterogeneous sequences tested recognition performance when the subject needed to react to a sudden switch in size. A given image was selected to serve as a target only twice in each condition. Only error rates were recorded.

5.2. Results and discussion

Fig. 8 shows the results of the Pure Size RSVP Experiment. Two aspects of these results are noteworthy. First, the significant advantage in identification of

large images over small images apparent in the single trial control experiments was eliminated in the homogeneous sequences $t(7) = 0.89$, $P > 0.4$, *ns*. In other words, small images were identified as accurately as large ones in the RSVP tasks. Second, in the heterogeneous conditions there was a strong asymmetry. Large images were significantly more accurately identified among small images than among large images $t(7) = 4.85$, $P < 0.002$, whereas the identification of small images among large images was almost at chance, significantly lower than performance with the same targets among small distractors $t(7) = 9.85$, $P < 0.001$.

Why did the difference found with the single presentation naming experiment (Experiment 1) disappear during the homogeneous RSVP trials? The participants had the same amount of time, 72 ms, to identify each image in the sequences as they had in Experiment 1, since the subsequent image in the sequence can be viewed as an effective mask. Effects of becoming familiar with the images can be excluded for two reasons. First, performance was not near ceiling, as evidenced by the substantially higher accuracy in the heterogeneous sequences with large targets. Second, under RSVP conditions, priming of distractor objects is negligible. Subramaniam et al. (2000) showed that even after extensive repetition (up to 31 times) of nontargets in RSVP sequences, participants showed no benefit in verification accuracy from these prior exposures when the nontargets finally became a target. Their general result of extremely poor memory for nontargets in RSVP sequences is consistent with a large number of studies suggesting little or no memory consolidation when the subject fails to attend to an image in the few hundred milliseconds *following* that which is sufficient for its identification as a nontarget (Subramaniam et al., 2000). Therefore, the most important difference between the single and the rapid serial presentations was that in the latter there were images surrounding the target that had to be processed and that were of the same or different size as the target. The results thus suggest that it was the size of the preceding (or following) images that crucially influenced recognition of the target images and eliminated the difference between small and large targets in the homogeneous trials.

The heterogeneous results suggest that whatever the nature of the effect of embedding a target in a RSVP sequence with a size change, the influence of the distracting images cannot be based purely on the magnitude of the size change per se functioning as a cue, as size changes produced facilitation of large targets and high costs for small targets relative to their verification in the homogeneous sequences. In addition, these results cannot be described by differences in the available information in the stimulus as assessed, for example, by an ideal observer, because the only way available information can be changed by context compared to the homogeneous case is by reducing positional uncertainty within the sequence. This, however, can lead to zero or positive information gain only with respect to the homogeneous results and thus cannot explain the drop of performance to a level slightly above chance in the case of the small heterogeneous condition. A more plausible explanation of the asymmetry could be based on luminance and contrast masking, a hypothesis tested (and rejected as an account of a substantial portion of the effects) in Experiments 5 and 6. An explanation based on size tuning is given in the general discussion.

6. Experiment 4: identifying objects with different spatial frequency content in RSVP sequences

The homogeneous results of Experiment 3 (Pure Size RSVP) could be interpreted as partial evidence for size tuning in processing a sequence of images. However, if a size-tuning mechanism based on SF does exist and affects size tuning in object recognition, it should be evident when images of constant size vary in their spatial frequency content. The goal of Experiment 4 (Pure Scale RSVP) was to address similar questions to those of Experiment 3, but in the spatial frequency rather than in the size domain. That is, would the advantage of LowSF images established in the single-trial experiment (Experiment 2), hold for homogeneous RSVP sequences? In heterogeneous sequences, would the detection of LowSF targets be facilitated and HighSF targets depressed?

According to the coarse-to-fine hypothesis, the change in size of images in the RSVP sequence in Experiment 3 produced a change in spatial frequency content that, in turn, engaged the SF-based size tuning mechanism. Following this logic, holding or changing spatial frequency content of the images should mimic the effect of size changes. Thus, if the spatial frequency content of the surrounding images influence recognition of the target qualitatively in the same way that the size of those images did in Experiment 3, it could provide additional support for the notion of size tuning based on spatial frequency.

6.1. Stimuli and method

The 24 large target images and the eight buffer images of Experiment 3 (Pure Size RSVP) were used in this experiment. All images were bandpass filtered according to the method in Experiment 2. The general procedure of presenting the sequences was identical to that of Experiment 3. Homogeneous sequences had all high or all low bandpass filtered images. In the heterogeneous sequences, one image was filtered differently than all the others (Fig. 9).

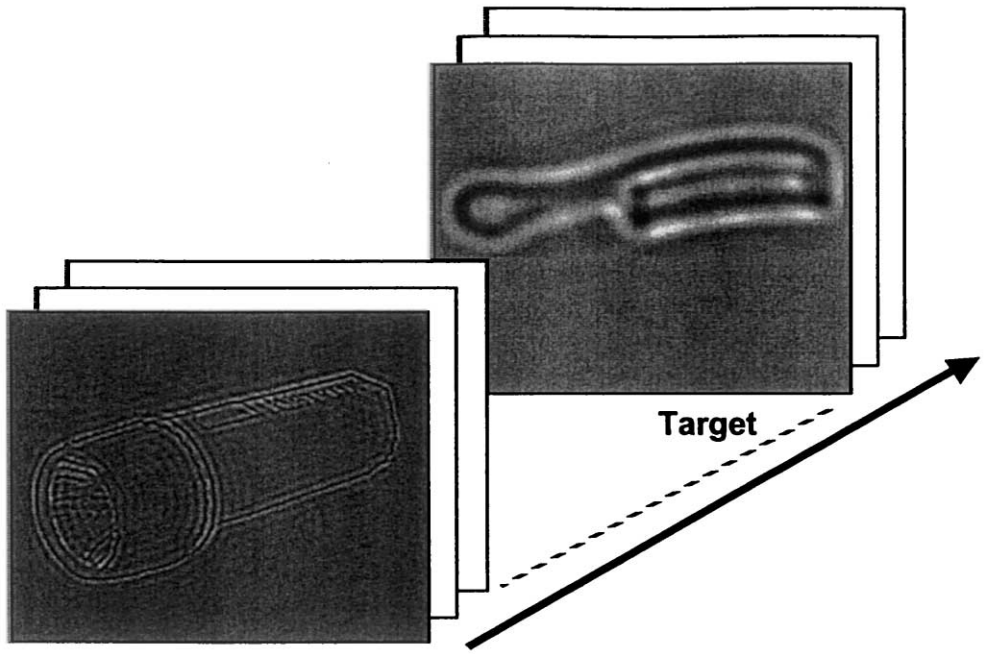

Fig. 9. An example of a trial from the heterogeneous sequences, LowSF target, HighSF distractors, in the pure scale RSVP experiment (Experiment 4). In the heterogeneous sequences, if a target was present in the sequence (i.e. a match trial), which it was on half the trials, it was always the image that differed in SF from the distractors. On the other half of the trials (Nonmatch trials), the image that differed in SF from the distractors did not match the target. All images in all conditions had the same size. (The arrow represents time).

6.2. Results and discussion

Fig. 10 shows the results of the Pure Scale RSVP experiment. The overall pattern of the results was very similar to that of the Pure Size Experiment, except that the magnitude of the costs and gains were reduced in the heterogeneous conditions. Parallel to the absence of an effect of size in the homogeneous sequences of Experiment 3, the difference between low and high bandpass images disappeared in the homogeneous sequences of the present experiment $t(15) = 0.23$, *ns*. In the heterogeneous conditions, low bandpassed images were significantly easier to identify among high bandpassed images than in homogeneous conditions, $t(15) = 3.01$, $P < 0.01$. Although the accuracy of identifying high bandpassed images among low bandpassed images was significantly above chance, it was also significantly below performance in the homogeneous conditions, $t(15) = 2.74$, $P < 0.016$.

The similar pattern of results in Experiments 3 and 4 is consistent with a size-tuning mechanism can presumably adjust its processing to the appropriate range of sizes based on the dominant spatial frequency content of the preceding images. Before turning to an explicit test of this hypothesis, however, in the next two experiments we assessed (and rejected as a complete account) the possible role of luminance and contrast masking in the RSVP experiments.

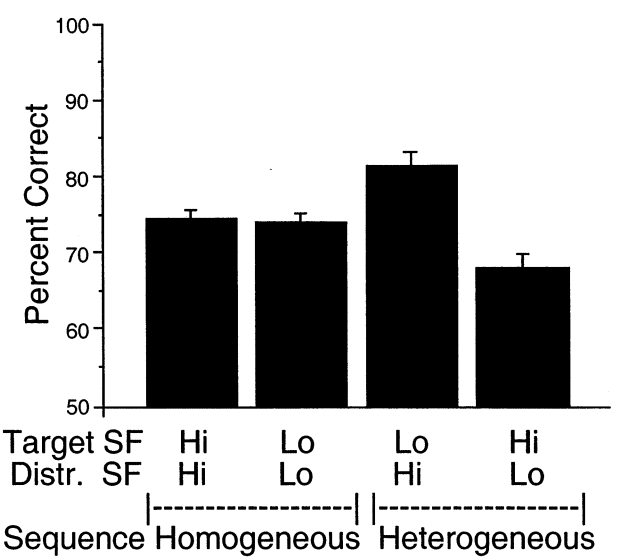

Fig. 10. The results of the Pure Scale RSVP Experiment (Experiment 4). Similar to the Pure Size Experiment (Experiment 3), accuracy on the two homogeneous sequences did not differ from each other. In the heterogeneous conditions, accuracy on the low-bandpassed targets was significantly higher and the high-bandpassed images significantly lower than that of the two homogeneous conditions. The effects of switching in the heterogeneous conditions were smaller than the effects revealed in the heterogeneous conditions of Experiment 3, the Pure Size RSVP Experiment (Error bars show S.E.).

7. Experiment 5 and 6: testing the effect of luminance and contrast masking in RSVP sequences

The results of RSVP Experiments 3 (size variation) and 4 (SF variation) could be explained in terms of perceptual masking (Breitmeyer, 1984) if the following assumption is made: in the homogeneous condition in the Pure Size Experiment, small images surrounding a target provide less masking than large images, so the easier large images (as established in the single-trial experiment) will be masked more by the preceding and following large images than small images by the surrounding small images. Thus, the stronger masking effects in the homogeneous conditions could have eliminated the differences between performances with small and large images found with single presentations. In the heterogeneous case, the large images would be masked even less by the small ones, so performance with large targets would be much better than that in the homogeneous case. Conversely, small images would be more effectively masked by the large images, leading to depressed performance. Similar arguments can be made for the bandpassed images under the assumption that low SF images or components in masks are more effective masks than high SF images or components in masks. Experiments 5 and 6 tested whether such a masking effect could explain the results of Experiment 3, by inserting pattern masks (rather than real objects) of different sizes after each object image in the RSVP sequence.

If masking was a function of luminance or contrast energy, then the mask images should be as effective as the object images in terms of perceptual masking since, in general, there was no difference between the object images and the mask images in terms of luminance and contrast values. Thus, variations in the size of the masks relative to the images should influence participants' performance in the same manner as expected from variation in object image sizes, except for a component of 'cognitive masking' (Intraub, 1981). For example, the large masks should reduce the identifiability of the small images in the homogeneous sequences relative to the large images, so that the previously obtained equivalence between large and small images in homogeneous sequences is lost. On the other hand, the masks differed from real object images in that they did not have definite contours and pronounced shapes that could be analyzed by a higher-level pattern mechanism. Therefore, if the masking effect was due to higher-level shape processes, such as size tuning, inserting the masks should not alter the general pattern of results obtained in the Pure Size Experiment (Experiment 3).

7.1. Stimuli and method

Both Experiments 5 and 6 were similar to the Pure Size RSVP Experiment (Experiment 3) with one important difference. After each image in the RSVP sequence, a mask appeared for 72 ms, the same amount of time as would an image in Experiment 3. Thus, in both experiments the sequences consisted of 80 images alternating between object image and mask (Fig. 11). Experiment 5 used the four large size masks used in Experiment 1, whereas Experiment 6 used both the large and the small versions of those masks. In Experiment 5 each object image was followed by one of the four large size masks, whereas in Experiment 6, each image was followed by a randomly selected mask that had the same size as the preceding object image. In a separate pilot study, the masks were tested and they demonstrated effective masking with the object images.

Eight participants, all undergraduates, at the University of Southern California, participated in Experiment 5 and another eight in Experiment 6 for course credit. The experimental procedure was identical to that of Experiment 3 with a single difference. Participants were told that they would see not only everyday objects, but random nonsense pictures as well and that they should ignore such pictures. As in all of the RSVP experiments, 192 sequences of images were run in both Experiments 5 and 6.

7.2. Results and discussion

Fig. 11(b) shows the results of the two experiments. In both cases, the pattern of the overall results was very similar to that of the Pure Size Experiment with no significant difference between the two homogeneous types of sequences, $t(7) = 0.23$, *ns* and $t(7) = 1.66$, $P > 0.1$, for Experiments 5 and 6, respectively. The heterogeneous sequences yielded significant facilitation in the identification of large targets compared to the homogeneous case, $t(7) = 2.43$, $P < 0.05$ and $t(7) = 3.9$, $P < 0.006$ for Experiments 5 and 6, respectively, and significant decrements in the recognition of small targets, $t(7) = 5.86$, $P < 0.001$ and $t(7) = 10.79$, $P < 0.0001$ for Experiments 5 and 6, respectively. In order to see how these results relate to the findings in Experiment 3 with respect to luminance or contrast-based perceptual masking, Fig. 12 shows in more detail the types of sequences used in the masking experiments and the recognition accuracy for those sequences.

According to the explanation based on luminance or contrast masking, the large differences in the heterogeneous sequences in Experiment 3 were caused by the difference in areas covered by the neighboring images, with large images causing more masking on small targets than vice versa. However, the equivalence of performance levels with the homogeneous sequences of small images, 81.4% and large images 80.9% (Experiment 5, first two rows), indicates that larger masks were insufficient to produce greater masking of the small objects than of the large objects. It is not the case that

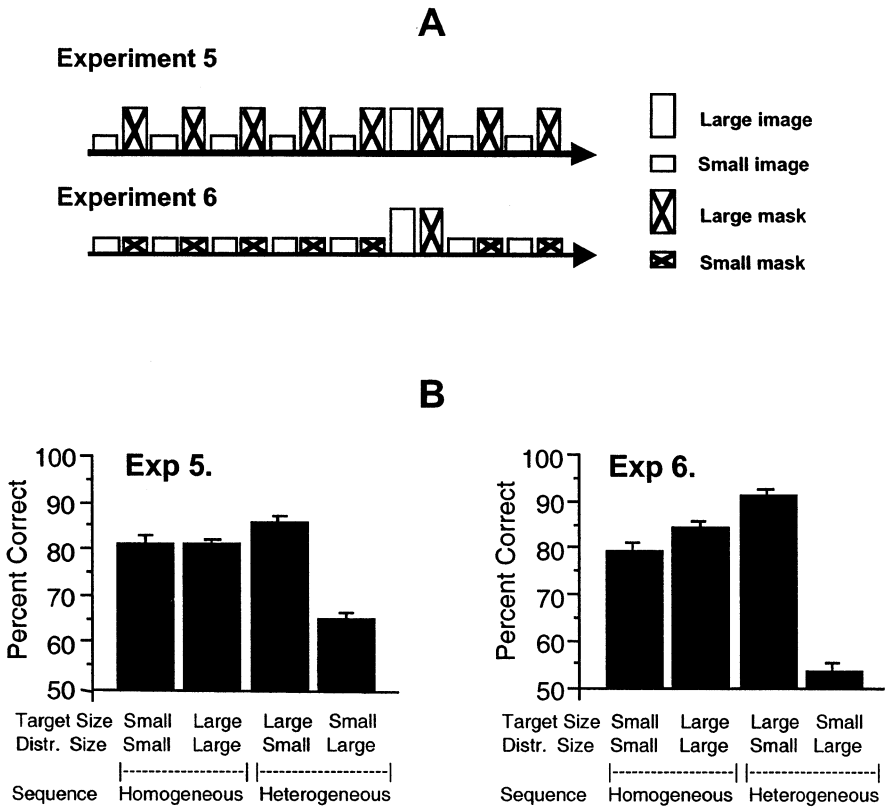

Fig. 11. The structure of the sequences and results of the two control RSVP experiments. (A) Schematic representation of a large-target heterogeneous sequence in the two masking experiments. In Experiment 5, a large size mask followed each image, whereas in Experiment 6, the size of the mask was always identical to the size of the preceding object image. There was no gap between subsequent images. (B) Results in both experiments had the same general pattern as that for the Pure Size RSVP Experiment. (Error bars show S.E.).

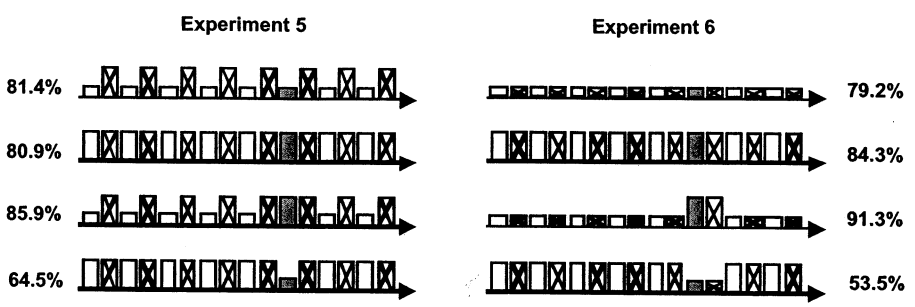

Fig. 12. Schematic representation of the sequences and results in the two masking RSVP experiments. The left block shows the sequences of Experiment 5, the right block of Experiment 6. The crossed rectangles represent masks, open rectangles represent object images. The size of the rectangles refers to the size of the images. The target in all sequences is depicted by the shaded open rectangle. From the top to bottom the rows depict small homogeneous, large homogeneous, large heterogeneous, small heterogeneous sequences, respectively. The percentages next to each row are the mean accuracy scores for identifying targets in sequences.

the 81.4 and 80.9% accuracy levels were an anomaly: the performance level for small images followed by small objects (rather than masks) in the homogeneous sequence of Experiment 3 had an almost identical accuracy level of 79.6%, whereas the performance level for large images followed by large objects was 82.6%. This equivalence also suggests that if conceptual masking (other than size tuning) was present in Experiment 3, its effect was, at best, quite modest.

In Experiment 5, large targets were always preceded and followed by large masks covering the entire area of the object image, yet the large targets in the heterogeneous sequences manifested a significant 5% increase in accuracy compared to those in the homogeneous se-

quences, 85.9 vs. 80.9% (Fig. 10). Similarly, the size of the preceding and following masks were the same in the homogeneous and heterogeneous sequences with small targets in Experiment 5, yet the latter targets suffered a 16.9% reduction in accuracy compared to the former (81.4 vs. 64.5%). Pure luminance or contrast masking cannot account for these results.

Further support for our contention that perceptual masking based on luminance or contrast played only a minor role in these experiments derives from the near equivalence in accuracy of verifying small targets in the homogeneous condition of Experiments 5 and 6, $t(7) < 1.00$, *ns*, where the masks were large in Experiment 5 but small in Experiment 6 (top row, Fig. 12). This equivalence is not likely to be a consequence of more able participants in Experiment 5 overcoming the effect of the large masks: The homogeneous sequences with large targets were exactly the same in the two experiment (Fig. 12, second row) and participants in Experiment 5 performed *worse* on that condition than participants in Experiment 6.

A similar case can be made when comparing the two heterogeneous sequences with small targets across the two experiments. The sequences were identical except that in Experiment 5 a large mask followed the small target, whereas in Experiment 6 the target was followed by a small mask (Fig. 12, fourth row). If the magnitude of masking was related to mask size, accuracy in this condition in Experiment 5 should have been lower than in Experiment 6, but the opposite was the case: a significant advantage of performance of Experiment 5 over Experiment 6, $t(7) = 3.15$, $P < 0.02$.

Insofar as masking cannot completely explain the pattern of results in Experiment 3, size tuning assumes greater plausibility as a factor responsible for a considerable portion of the effects. However, this does not mean that luminance and contrast masking were not present or such masking would not contribute to the participants' performance. Luminance and contrast based perceptual masking are well-documented phenomena and such masking probably had a role in Experiments 5 and 6. The magnitudes of the facilitation and cost in the heterogeneous conditions of Experiment 3 — an increase of 12.8% for large targets and a decrease of 23.6% for small targets — were greater than in the corresponding conditions of Experiments 5 and 6, indicating the effect of the masks in Experiments 5 and 6. However, the results combined from the three experiments suggest the presence of another major factor that we believe to be size tuning.

8. Experiment 7: testing the effect of absolute spatial frequency content of images in RSVP sequences

As described in Section 1, the Scaling Hypothesis (Hughes et al., 1996) assumes that with each new object appearing in the scene, the visual system performs a 'coarse-to-fine' tuning in the spatial frequency domain. This tuning provides input to a normalization process that ensures size-invariant representation of features by selecting the most salient size range and provides a reason why small images would be identified with less efficiency in single presentation tasks. However, the coarse-to-fine hypothesis with object-by-object tuning is inconsistent with the equivalence in performance in identifying small and large targets in the homogeneous RSVP sequences in Experiment 3. Instead, the equivalence suggests that tuning over the first eight images in the sequence were maintained over the subsequent images. If this tuning is based on the spatial frequency content of the images, then holding the absolute spatial frequency content of the images constant should reduce or eliminate the effect of tuning. In particular, in the heterogeneous sequences, the advantage of large targets and the disadvantage of small targets should be eliminated. Experiment 7 tested this prediction.

8.1. Stimuli and method

The same 24 target and eight buffer images were used as in the Pure Size RSVP Experiment (Experiment 3), but all images were filtered in a one and half octave bandwidth ≈ 10 cpd. Thus, the large size images were identical to the high band-passed images used in Experiments 2 and 4, whereas the small filtered images were not used in any of the previous experiments (Fig. 1). The experimental method was identical to that in all of the previous RSVP experiments with 192 trial sequences (Fig. 13). Eight naïve participants participated in the experiment.

8.2. Results and discussion

Fig. 14 shows the results of the Absolute Scale RSVP experiment. As evident from the graph, filtering the images in the same spatial frequency band did not significantly alter the heterogeneous results compared to the results of the Pure Size Experiment. Large images remained significantly easier to identify among small ones than in homogeneous conditions, $t(7) = 8.81$, $P < 0.0001$. Identification of small images among large ones stayed barely above chance and significantly worse than identification among small distractors, $t(7) = 7.89$, $P < 0.0001$. On the other hand, there was a small but significant difference between identifying small and large images in the homogeneous condition favoring the small sequences, $t(7) = 2.63$, $P < 0.05$.

If the reason for equal performance in the homogeneous case was an absolute spatial frequency based tuning mechanism, one would expect the significant

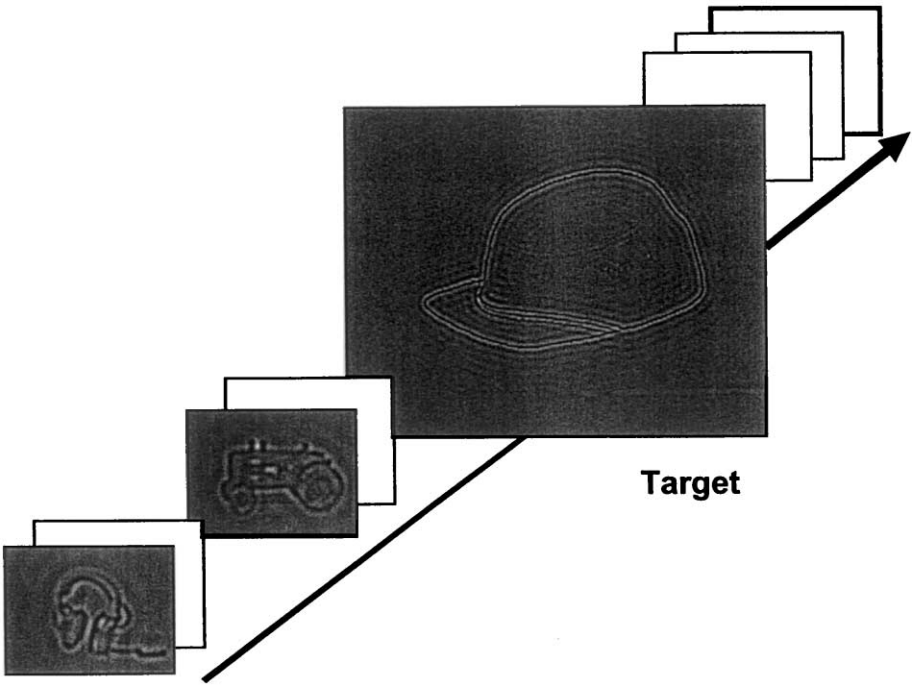

Fig. 13. A heterogeneous trial example from Experiment 7. In a sequence of small images there was one large image which might or might not be the target. All images were filtered $\approx$10 cpd in a 1.5 octave bandwidth.

advantage of large images to reappear with the bandpass filtered images in the homogeneous cases and the differences in performance to disappear in the heterogeneous cases. Neither occurred suggesting that if there is size tuning during the viewing of RSVP sequences it is not based exclusively on the absolute spatial frequency content of the images.

In fact, comparing the magnitudes of the gains for large targets and costs for small targets in the heterogeneous conditions across experiments suggests only a moderate role played by a size tuning mechanism based on spatial frequency. In Experiment 3, where both size and scale changed, accuracy of verification of large targets increased by 12.8% and the accuracy of small targets declined by 23.6% in the heterogeneous sequences compared to the homogeneous sequences. In the present Experiment 7, where only size changed but scale did not, the comparable values were 13.7 and 20.8%. (Some caution must be exercised in considering this comparison because the overall mean accuracy level of Experiment 2 with unfiltered stimuli, 81.1%, was higher than the 73.4% mean accuracy level for Experiment 7, with filtered stimuli.). In Experiment 4 where *only* scale varied but size did not, the magnitude of heterogeneous costs and gains were markedly reduced relative to Experiments 2 and 7, with a 7.4% gain in accuracy for low passed images and a 6.3% reduction in accuracy for high passed images. Thus 'size-change-only' produced much more similar results to those of 'size-and-scale-change' than 'scale-change-only' did.

However, filtering the images to the same absolute frequency band did alter the homogeneous results in an unexpected way: the same small bandpass filtered images that were identified at near chance levels in the

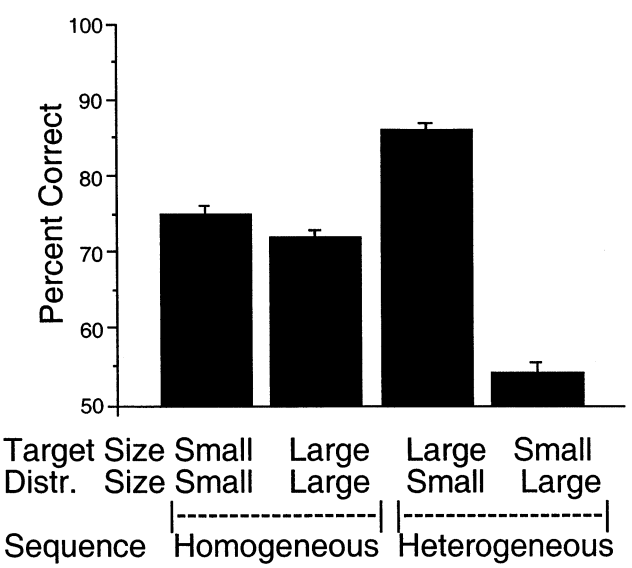

Fig. 14. The results of Experiment 7 (absolute SF). In contrast to Experiment 3 (Pure Size) where there was no difference in accuracy in the two homogeneous conditions, here the small bandpass filtered target images were slightly (but reliably) easier to identify than the large bandpassed target images. Accuracy in the large target heterogeneous condition was significantly higher and the small target heterogeneous accuracy significantly lower than in the two homogeneous conditions. (Error bars show S.E.).

heterogeneous sequences had a small but significant advantage over the large filtered images on the homogeneous conditions. This result raises the possibility that spatial frequency content of the images is taken into account according to its value relative to the object size rather than according to its absolute value.

9. Experiment 8: testing the effect of relative spatial frequency content of images in RSVP sequences

As described in Section 1, a tuning mechanism based on absolute spatial frequency would utilize information in the low-pass filtered version of the image to establish the most salient spatial frequency band and normalize the size of the input window according to this spatial frequency. Since the amplitude spectrum of natural images declines roughly with $1/f$ where f is the spatial frequency of the image summed over all orientations (Field, 1987), an octave-wide bandpass filtering has to be used for this type of tuning.

However, simple normalization according to the $1/f$ rule would not necessarily suffice. While bandpass filtering the object images according to their absolute spatial frequency content ensures that each image carries information only in a given range as measured in cycles per degree, it does not treat small and large images equally in terms of their diagnostic spatial frequencies for recognition. Several studies reported that relative frequency (measured by cycles per object, cpo) rather than absolute frequency (cycles per degree, cpd) determines object recognition performance (Harmon & Julesz, 1973; Rolls & Baylis, 1986; Parish & Sperling, 1991; Solomon & Pelli, 1994).[4] Could a tuning mechanism based on relative spatial frequencies explain our results? A possible tuning mechanism based on relative spatial frequency content of the image could work, for example, by individual simple feature detectors defined in one spatial frequency band voting for the most salient spatial frequency band based on their activity *and* the activity of the detectors in the other spatial frequency bands. Thus, the normalization process would not only depend on the absolute spatial frequency content of a band, but also on the content of other bands. Importantly, such a scheme would still implement the tuning by selecting a particular spatial frequency band and determining the proper size through this selection.

If the size tuning mechanism is based on the relative rather than the absolute spatial frequency content of the image, Experiment 7 would not provide an appropriate control for testing the Scaling Hypothesis. Filtering around 10 cpd means that small and large images were filtered in different relative frequency ranges: high bandpass filtering in terms of cpo for large sized images, but low bandpass filtering in terms of cycles per object for the small images. As recognition performance is an inverted U function of the relative SF with an optimal relative SF content, the small images might be filtered closer to their optimal relative frequencies in Experiment 7. This could also explain why performance with homogeneous sequences using small images was significantly better than performance with large images. In order to test tuning based on relative spatial frequencies, Experiment 8 repeated the test in Experiment 7 but kept the relative rather than the absolute spatial frequency content constant between the small and large images.

9.1. Stimuli and method

The same 24 target and eight buffer images were used as in the Pure Size RSVP Experiment (Experiment 3), but only the small images were filtered in a one and half octave bandwidth ≈ 10 cpd, whereas the large ones were filtered ≈ 2 cpd (Fig. 15). Since the linear size ratio between the large and small images as well as between the two center frequencies was 5:1, the resulting small and large images had the same relative frequency content. The experimental method was identical to that in all of the previous RSVP experiments.

9.2. Results and discussion

Fig. 16 shows the results of the relative scale RSVP experiment. Once again, the general pattern of the results was similar to the Pure Size RSVP experiment (Experiment 3). There was no significant difference between identifying small and large images in the homogeneous sequences, $t(7) < 1.0$, *ns*. In the heterogeneous conditions, large images remained significantly easier to identify among small ones than in homogeneous conditions, $t(7) = 3.74$, $P < 0.01$. Identifying small images among large ones still stayed significantly worse than in the homogeneous case, $t(7) = 2.5$, $P < 0.05$, although identification accuracy was significantly higher than in the Absolute Scale RSVP experiment $t(14) = 2.28$, $P < 0.05$.

The rationale of Experiment 8 was the same as that of Experiment 7. If tuning depends on the relative spatial frequency content of the image and this is why there is a strong switching effect in the heterogeneous conditions, keeping relative spatial frequency constant while changing the size of the image should eliminate the switching effect. Although the results of the two homogeneous conditions were equal, the heterogeneous results still reveal a strong effect of changing size suggesting that changing relative spatial frequency content

[4] There is some evidence that spatial frequency discrimination also acts on representation of distal (cyl/cm) rather than on retinal (cpd) size (Burbeck, 1987; Bennett & Cortese, 1996).

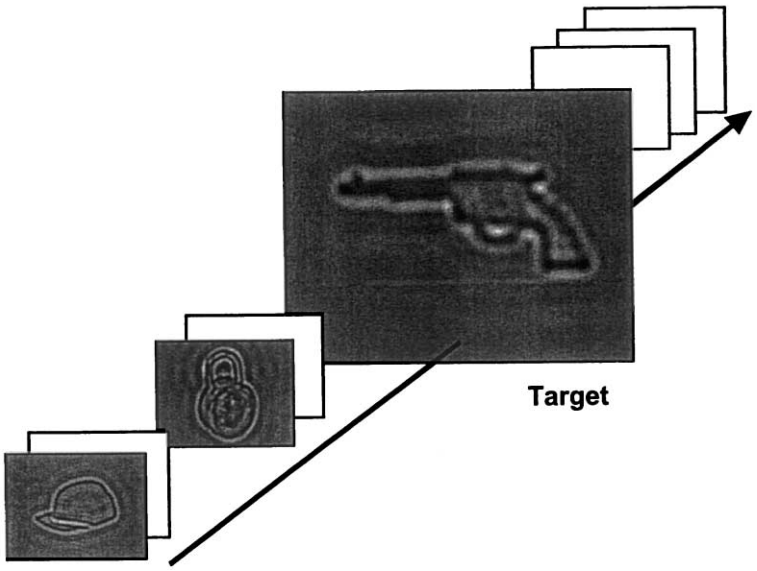

Fig. 15. An example of a heterogeneous trial with a large/2 cpd target among small/10 cpd distractors from Experiment 8. In a sequence of small images (distractors) there was one large image which was the target on half the trials. On the other half of the trials, the large image did not match the target. In the heterogeneous sequences with large distractors, a small image would be the possible target. Small images were filtered around a center frequency of 10 cpd; large images ≈ 2 cpd. Because the large images were $5\times$ larger than the small images, all images thus had the same relative frequency content.

by itself is not the sole reason for the switching effect in performance.

Comparing the magnitudes of the gains for large targets in the heterogeneous conditions across the experiments shows that accuracy of verification of large targets increased by approximately the same amount that it did in the Pure Size and the Absolute RSVP Experiments, 11.8, 12.8 and 13.7%, respectively. A one-way, 3-level ANOVA on switching amplitudes showed no main effect of Experiment, $F(2,21) < 1.00$, *ns*. However, the cost for small targets was significantly smaller than in the other two experiments, 9.2 vs. 23.6 and 20.8% in Experiments 3 and 7, with a main effect of Experiment $F(2,21) = 6.80$, $P < 0.01$. This suggest that although relative spatial frequency by itself is not the control signal for size tuning, relative spatial frequency filtering affects attributes of the image which are important in tuning to smaller sizes.

10. Discussion

The two primary issues in the present investigation were whether there is evidence for efficient size tuning in recognition tasks and whether this tuning can be achieved in the absence of spatial frequency tuning. The answer is yes to both questions, regardless of whether SF tuning is defined absolutely, in cycles per degree, or relatively, in cycles per object. The tuning can be accomplished by 576 ms (over the first eight buffer images) and maintained over a changing stream of images.

10.1. Scale-invariant size tuning

To recapitulate our results, we tested the hypothesis that the capacity to process objects at different sizes is achieved by a size-normalization process where spatial frequency is used as the guiding signal to accomplish

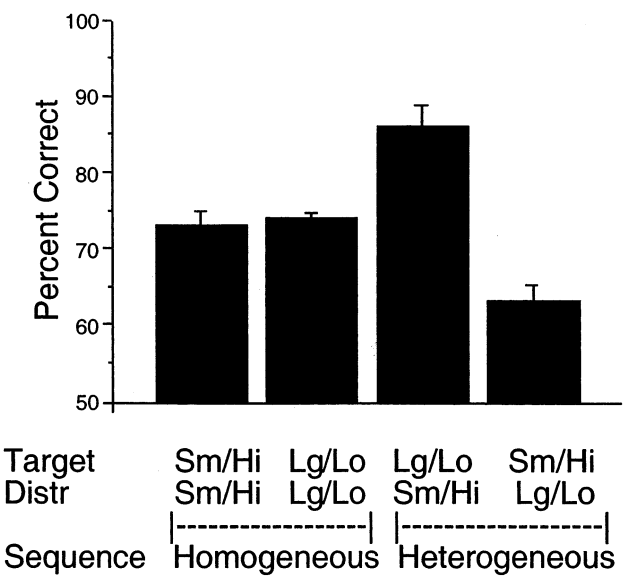

Fig. 16. The results of Experiment 8. As in the Pure Size Experiment (Experiment 3), accuracy in the two homogeneous conditions was almost identical, despite the relative spatial frequency bandpass filtering operation. In the heterogeneous conditions, accuracy with large targets was significantly higher and that for small targets significantly lower than that for the homogeneous conditions. (Error bars show S.E.).

the normalization. In the single-trial presentation conditions of Experiments 1 and 2 we showed that large and low bandpassed images had an advantage over small and high bandpassed images in single-trial recognition tasks. These results established that the size and SF differences in our stimuli were sufficient to produce differences in recognition performance. Experiments 3 and 4 showed that the large/LowSF advantage could be eliminated when images were presented in an RSVP stream, following images that matched the size or scale of the target. Experiments 7 and 8 established that this equivalence could be obtained with absolute and relative spatial frequency held constant. These results are contrary to what would be expected from the proposal of coarse-to-fine tuning (Sergent, 1986; Schyns & Oliva, 1994; Hughes et al., 1996).

If coarse-to-fine tuning was based on each image, the accuracy in verifying the high bandpassed images in the RSVP presentations should have been lower than that for the low bandpassed images (mirroring the single trial presentation results), due to a missing tuning signal of the low frequency component. One possible remedy is to assume that tuning across multiple, rapidly presented images is possible, although at slower speed. However, the heterogeneous results of the Absolute and the Relative Scale RSVP experiments suggest that the spatial frequency content of the images does not affect performance when switching from one size to another. Thus, size tuning is not controlled exclusively or even linked closely to the spatial frequency content of the image. When the magnitude of size changes are small, a common representation of local spatial filters might be useful in recognizing the same object at different sizes. However, for size changes of large magnitude, as in the present experiments, the tuning mechanism may utilize another attribute. We propose that the other attribute is spatial extent rather than spatial frequency.

The idea of separating the effect of spatial frequency and spatial extent when there is a large variation of an attribute of an image has been proposed in several earlier studies (Stuart, Bossomaier, & Johnson, 1993; Wilcox & Hess, 1995; Morgan, Perry, & Fahle, 1997). For example, Wilcox and Hess (1995) found that although stereoacuity depends on both the spatial frequency of a Gabor stimulus and the size of the Gabor patches, the upper limit of stereopsis, D_{max}, primarily depends on the size of the patch and it is relatively independent of the spatial frequency of the Gabor patch.

10.2. Asymmetric switching costs: a parallel between size and contrast tuning

In Experiments 3 (Pure Size) and 4 (pure SF), there were asymmetric switching costs such that accuracy of verification of Large/HighSF images increased compared to their homogeneous levels and declined for Small/LowSF images. In Experiments 5 and 6, we found that the interactions could not be explained in terms of low-level luminance or contrast-based masking. As with the homogeneous sequences, Experiments 7 and 8 ruled out absolute and relative spatial frequency content as the tuning signal for asymmetric size adjustments.

Earlier studies exploring the effect of size differences in pattern recognition were not designed so that they could separate the effect of upward and downward switches (Bundesen & Larsen, 1975; Larsen & Bundesen, 1978; Larsen, 1985; Cave & Kosslyn, 1989). Similarly, earlier studies of attention investigating the effect of precueing on the size of the 'attentional window' could test only downward switches since they used small targets and varying cue sizes that were always at least as large as the target (Egeth, 1977; Eriksen & St. James, 1986; LaBerge & Brown, 1989; Greenwood & Parasuraman, 1999).

Computationally, the asymmetry of costs would be unexpected from a size-tuning hypothesis if the cost were associated simply with 'novelty' or other measures that consider only the magnitude of the size change between the target and the preceding images. Within a biologically more plausible framework, however, asymmetrical costs in an adaptive system is the norm rather than the exception. DeWees and Zador (1998) showed formally that in a large class of stochastic systems optimal adaptation to the variance of the input leads to asymmetrical dynamics with faster adaptation to abrupt increases than to abrupt decreases. In addition, the upward switch causes a brief overshoot in the estimate of variance before asymptoting to its steady value. It is well known that light intensity adaptation in sensory cells such as photoreceptors and ganglion cells follows this pattern (Barlow & Mollon, 1982). Similarly, the contrast adaptation of ganglion cells in salamanders and rabbits occurs more rapidly to higher than to lower levels of contrast (Smirnakis, Berry, Warland, Bialek, & Meister, 1997). These results suggest that if tuning to a given size is accomplished by a mechanism which is similar to other adaptation mechanisms in the visual system, asymmetrical behavior could be expected.

The case of contrast can offer an analogy as to how asymmetric adaptation could lead to differential performance in the heterogeneous conditions, relative to the homogeneous conditions. Fiser and Fine (1998) recently reported an RSVP experiment which was an exact replica of the present Experiment 3 (Pure Size) with two differences. First the varying attribute of the images was their contrast (low = 15%, high = 62%) rather than their size. Second, the presentation time of each image was 50 ms. The results were very similar to those of the Pure Size RSVP Experiment: there was no

significant difference in performance between the two homogeneous conditions (72 vs. 74%), a strong improvement with high contrast targets among low contrast ones (86%) and a decrement with low contrast targets among high contrast ones (67%).

The results in the contrast domain can be explained intuitively by contrast gain control. When the system adapts its contrast gain control to the stream of high contrast images, a quickly presented low contrast target image is buried in the noise. When the system is adapted to low contrast a sudden appearance of a high contrast image gives a quick saturated signal that can be sufficient for basic level recognition. Indeed, measuring visual evoked potentials in human participants with contrast-reversal checkerboards undergoing transitions in mean contrast level, Victor et al. (1997) found evidence for both contrast gain control and a 2-fold sensitivity change in response to a sudden shift in mean contrast level.

By analogy, if there is a similar adaptation mechanism for size tuning, the asymmetrical performance in our experiments could be explained in a similar fashion. We thus propose that large switches (>4-fold) from one size to another does require tuning. In object recognition tasks, this tuning can manifest itself in either a decrement or an improvement in performance depending on the particular task and the direction of the switch. We note that this proposal is speculative since at the moment there is no clear evidence that size is encoded in the visual system in a fashion similar to the encoding of contrast.

Behavioral evidence from monkey psychophysics also support the existence of a tuning mechanism based on size, spatial frequency, contrast, color and other attributes of the stimulus, and an asymmetric effect of size differences on tuning. Schiller et al. investigated the effect of ablating a section of V4 of the macaque on odd-man-out detection of targets that were either in the visual field of the lesioned tissue or outside that sector (Schiller & Lee, 1991; Schiller, 1993). They found that the lesion had little effect on the detection of a briefly presented target in the portion of the visual field coded by the lesioned sector if it was a more salient stimulus, e.g. large/low SF item among small/high SF distractors. In contrast, the monkey's performance was severely disrupted when the target presented in the lesioned sector was the less salient element among the distractors, e.g., small/HighSF target among large/lowSF distractors. In addition, the response latencies with larger targets in the lesioned sector were only 20% longer than in the non-lesioned sector, whereas with smaller target the latencies were twice as long. When the less salient targets were presented alone in the lesioned sector the monkeys had no difficulty in detecting them, suggesting that the lesion interfered with perception only when fast tuning from the large distractors to a small target was necessary. This strongly asymmetric performance and the latency differences are in agreement with our results. The fact that Schiller's results hold across different attributes, which include stimulus size and contrast, supports our analogy of size and contrast tuning in human perception.

This analogy can be strengthened by viewing size tuning in humans as an involuntary attentional mechanism. Consistent with this account, when attention in human visual search tasks testing the same attributes as used by Schiller and Lee (1991) is redirected away from the target, the results were qualitatively similar to those obtained by the lesioned monkeys (Braun, 1994). Braun (1994) found that when participants' attention was directed to a concurrent letter discrimination task, they could perform a visual search for a target that was larger, lower SF, or brighter with only a small effect of the number of distractors, whereas detecting smaller, higher SF, or dimmer targets required their scrutiny, and produced a much larger cost of the distractors. Since the perceptual difficulty of the targets were equalized in the two cases and with attention there was no difference in performance between more or less salient targets, Braun suggested that removing visual attention mimicked the effect of a lesion in area V4.

10.3. Relation to other studies

Our findings suggesting that the human visual system operates on an adjustable size and spatial frequency scale during recognition rather than in a size-independent fashion is consistent with a number of earlier results. Solomon and Pelli (1994) suggested that recognition of letters and gratings at one size is mediated and constrained by a single visual filter or 'channel'. Costen et al., (1996) reported that face identification is supported by a limited band of spatial frequencies of ≈ 8–16 cycles per face (cpo). Testing identification performance with bandpass filtered letters and faces, Gold et al. (1999) found that human performance could not be predicted based on a single filter with fixed bandwidth, and they suggested that humans might adjust the center frequency of the utilized 'channel' and use spatial sampling which is optimal to a given channel.

Our findings, that spatial frequency content of the images are not relevant to size tuning is in contrast with the proposal by Hughes et al. (1996) who suggested that the visual system determines the optimal size of operation by the absolute frequency content of the image. Recently, Schyns and Oliva (1994) revised their earlier proposal about coarse-to-fine tuning suggesting that different recognition tasks might utilize information in different, 'diagnostic' scales and thus a strictly coarse-to-fine progression is not necessary during object or scene recognition (Schyns & Oliva, 1997a,b). This

explanation, in abandoning the idea of 'coarse-to-fine tuning', is consistent with our interpretation of the present results.

11. Conclusion

In conclusion, our results suggest that during object recognition the visual system can tune in to an appropriate size range in < 576 ms. This tuning can be so effective that the default biases that favor large or low SF stimuli in single trial presentations lose their advantage so recognition performance becomes independent of the actual size/SF of the stimuli. The tuning process is not dependent on constancy of an individual stimulus in an RSVP sequence, in that changing the actual image every 72 ms does not eliminate the tuning as far as the relevant attribute — size or spatial frequency content — remains constant. Finally, size tuning can be carried out independently of the absolute and relative spatial frequency content of the image, suggesting that if global-to-local dominance exists in object recognition, it is not controlled by coarse-to-fine spatial frequency tuning.

Acknowledgements

This work was supported by ARO DAAH04-94-G-0065, ARO MURI DAAG55-98-1-0293JSMF-96-44 (MURI), DOD NMA202-98-K-1089 and the Human Frontier Science Program RG0035/2000-B 102 to Irving Biederman and JSMF 96-32 to József Fiser. We thank Dianne Martinez, Trang Hong and Nancy Wang for their assistance in running the participants.

References

Barlow, H., Mollon, J. (Eds.). The senses, (2nd edn.) Cambridge: Cambridge University Press, 1982.

Bennett, P. J., & Cortese, F. (1996). Masking of spatial frequency in visual memory depends on distal, not retinal, frequency. *Vision Research*, *36*, 233–238.

Biederman, I. (1987). Recognition-by components: A theory of human image understanding. *Psychological Review*, *94*(2), 115–147.

Biederman, I., & Cooper, E. E. (1992). Size invariance in visual object priming. *Journal of Experimental Psychology-Human Perception and Performance*, *18*(1), 121–133.

Braun, J. (1994). Visual search among items of different salience: Removal of visual attention mimics a lesion in extrastriate area V4. *Journal of Neuroscience*, *14*(2), 554–567.

Breitmeyer, B. G. (1975). Simple reaction time as a measure of the temporal response properties of transient and sustained channels. *Vision Research*, *15*, 1411–1412.

Breitmeyer, B. G. (1984). *Visual masking*: *an integrative approach*. New York: Oxford University Press.

Bundesen, C., & Larsen, A. (1975). Visual transformation of size. *Journal of Experimental Psychology: Human Performance and Perception*, *1*, 214–220.

Burbeck, C. A. (1987). Locus of spatial-frequency discrimination. *Journal of the Optical Society of America (A)*, *4*, 1807–1813.

Cave, K. R., & Kosslyn, S. M. (1989). Varieties of size-specific visual selection. *Journal of Experimental Psychology: General*, *118*, 148–164.

Costen, N. P., Parker, D. M., & Craw, I. (1996). Effects of high-pass and low-pass spatial filtering on face identification. *Perception and Psychophysics*, *58*(4), 602–612.

Davis, E. T., & Graham, N. (1981). Spatial frequency uncertainty effects in the detection of sinusoidal gratings. *Vision Research*, *21*, 705–712.

DeWees, M., & Zador, A. (1998). Asymmetric dynamics in optimal variance adaptation. *Neural Computation*, *10*(5), 1179–1202.

Egeth, H. (1977). Attention and preattention. In G. H. Bower, *The psychology of learning and motivation* (pp. 277–320). New York: Academic Press.

Eriksen, C. W., & St. James, J. D. (1986). Visual attention within and around the field of focal attention: A zoom lens model. *Perception and Psychophysics*, *40*(4), 225–240.

Farell, B., & Pelli, D. G. (1993). Can we attend to large and small at the same time? *Vision Research*, *33*(18), 2757–2772.

Field, D. (1987). Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America (A) Optics and Image Science*, *4*, 2379–2394.

Fiser, J., & Fine, I. (1998). Effects of contrast adaptation on high level object recognition tasks. *Investigative Ophthalmology and Visual Science*, *39*(4), S853–S853.

Gold, J., Bennett, P. J., & Sekuler, A. B. (1999). Identification of band-pass filtered letters and faces by human and ideal observers. *Vision Research*, *39*, 3537–3560.

Graham, N., Robson, J., & Nachmias, J. (1978). Grating summation in fovea and periphery. *Vision Research*, *18*, 815–826.

Greenwood, P. M., & Parasuraman, R. (1999). Scale of attentional focus in visual search. *Perception and Psychophysics*, *61*, 837–859.

Harmon, L. D., & Julesz, B. (1973). Masking in visual recognition: Effect of two-dimensional filtered noise. *Science*, *180*, 1194–1197.

Hughes, H. C., Nozawa, G., & Kitterle, F. (1996). Global precedence, spatial-frequency channels, and the statistics of natural images. *Journal of Cognitive Neuroscience*, *8*(3), 197–230.

Intraub, H. (1981). Identification and naming of briefly glimpsed visual scenes. In D. F. Fisher, R. A. Monty, & J. W. Senders, *Eye movements*: *cognition and visual perception* (pp. 181–190). Hillsdale, NJ: Erlbaum.

Julesz, B. (1975). Two-dimensional spatial-frequency-tuned channels in visual perception. In G. F. Inbar, *Signal analysis and pattern recognition in biomedical engineering* (pp. 177–197). New York, NY: Wiley.

Kroll, J. F., & Potter, M. C. (1984). Recognizing words. pictures, and concepts: A comparison of lexical, object, and reality decisions. *Journal of Verbal Learning and Verbal Behavior*, *23*, 39–66.

LaBerge, D., & Brown, V. (1989). Theory of attentional operations in shape identification. *Psychological Review*, *96*, 101–124.

Larsen, A. (1985). Pattern matching: Effect of size ratio, angular difference in orientation, and familiarity. *Perception and Psychophysics*, *38*, 63–68.

Larsen, A., & Bundesen, C. (1978). Size scaling in visual pattern recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *4*, 1–20.

Morgan, M. J., Perry, R., & Fahle, M. (1997). The spatial limit for motion detection in noise depend on element size, not on spatial frequency. *Vision Research*, *37*, 729–736.

Murdock, B. B. Jr (1962). The serial position effect in free recall. *Journal of Experimental Psychology*, *64*, 482–488.

Olshausen, B. A., Anderson, C. H., & Van Essen, D. C. (1993). A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *Journal of Neuroscience*, *13*(11), 4700–4719.

Olshausen, B. A., Anderson, C. H., & Van Essen, D. C. (1995). A multiscale dynamic routing circuit for forming size- and position-invariant object representation. *Journal of Computational Neuroscience*, *2*, 45–62.

Parish, D. H., & Sperling, G. (1991). Object spatial frequency, retinal spatial frequencies, noise, and the efficiency of letter discrimination. *Vision Research*, *31*(7/8), 1399–1415.

Rolls, E. T., & Baylis, G. C. (1986). Size and contrast have only small effects on the responses to faces of neurons in the cortex of the superior temporal sulcus of the monkey. *Experimental Brain Research*, *65*, 38–48.

Schiller, P. H. (1993). The effect of V4 and middle temporal (MT) area lesions on visual performance in the Rhesus monkey. *Visual Neuroscience*, *10*, 717–746.

Schiller, P. H., & Lee, K. (1991). The role of the primate extrastriate area V4 in vision. *Science*, *251*, 1251–1253.

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. (1997a). Coarse blobs or fine edges? Evidence that information diagnosticity changes the perception of complex visual stimuli. *Cognitive Psychology*, *34*(1), 72–107.

Schyns, P. G., & Oliva, A. (1997b). Flexible, diagnosticity-driven, rather than fixed, perceptually determined scale selection in scene and face recognition. *Perception*, *26*, 1027–1038.

Sekuler, R., & Nash, D. (1972). Speed of size scaling in human vision. *Psychonomic Science*, *27*(2), 93–94.

Sergent, J. (1986). Microgenesis in face perception. In H. D. Ellis, M. A. Jeeves, F. Newcombe, & A. Young, *Aspects of face processing*. Dordrecht, The Netherlands: Martinus Nijhoff.

Shulman, G. L., & Wilson, J. (1987). Spatial frequency and selective attention to local and global information. *Perception*, *16*, 89–101.

Smirnakis, S., Berry, M., Warland, D., Bialek, W., & Meister, M. (1997). Retinal processing adapts to image contrast and spatial scale. *Nature*, *386*, 69–73.

Solomon, J. A., & Pelli, D. G. (1994). The visual filter mediating letter identification. *Nature*, *369*, 395–397.

Stuart, G. W., Bossomaier, T. R. J., & Johnson, S. (1993). Preattentive processing of object size-implications for theories of size perception. *Perception*, *22*, 1175–1193.

Subramaniam, S., Biederman, I., & Madigan, S. A. (2000). Accurate identification but no priming and chance recognition memory for pictures in RSVP sequences. *Visual Cognition*, *7*, 511–535.

Thorpe, S. (1988). Identification of rapidly presented images by the human visual system. Perception, 17(A77).

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

Victor, J. D., Conte, M. M., & Purpura, K. P. (1997). Dynamic shifts of the contrast-response function. *Visual Neuroscience*, *14*, 577–587.

Ward, L. M. (1982). Determinants of attention to local and global features of visual forms. *Journal of Experimental Psychology: Human Perception and Performance*, *8*(4), 562–581.

Wilcox, L. M., & Hess, R. F. (1995). Dmax for stereopsis depends on size, not spatial frequency content. *Vision Research*, *35*, 1061–1069.