PERGAMON

Vision Research 38 (1998) 2109–2118

Vision Research

Linearity across spatial frequency in object recognition

Elizabeth S. Olds [a,*], Stephen A. Engel [b]

[a] Department of Psychology, University of Waterloo, 200 University Avenue West, Waterloo, Ont. N2L 3G1, Canada
[b] Department of Psychology, University of California at Los Angeles, Box 951563, Franz Hall, Los Angeles, CA 90095-1563, USA

Received 22 January 1997; received in revised form 14 October 1997

Abstract

In three experiments, we measured recognition as a function of exposure duration for three kinds of images of common objects: component images containing mainly low-spatial-frequency information, components containing mainly high-spatial-frequency information, and compound images created by summing the components. Our data were well fit by a model with a linear first stage in which the sums of the responses to the component images equalled the responses to the compound images. Our data were less well fit by a model in which the component responses combined by probability summation. These results support linear filter accounts of complex pattern recognition.

Keywords: Object; Recognition; Spatial frequency; Linear; Animal; Human

1. Introduction

The human visual system analyzes images at many spatial scales. Classic physiological work showed that each neuron in primary visual cortex analyzes information at a relatively narrow band of spatial frequencies, and that cortical area V1 contains neurons that cover a wide range of spatial scales [6]. Psychophysical studies also found evidence for early narrowly tuned mechanisms [1,5,9,16]. Yet, somewhere in the visual system, information from multiple scales must be integrated to allow performance of complex tasks. Here we report the results of a mixture of experiments that measured how people combine information from multiple scales while recognizing common objects.

The simplest model of information combination is a linear model: linearity predicts that the response of the visual system to the sum of two component stimuli will be the sum of the responses to the component stimuli presented individually. Mixture experiments permit a straightforward test of the linear combination model. These experiments compare responses to compound stimuli and responses to individually presented component stimuli. For example, an experiment might measure people's ability to detect individual sinusoidal gratings (component stimuli) and compound gratings created by adding together two components.

For detection of compound sinusoidal stimuli, the visual system appears to use a linear combination rule when components are close in spatial frequency; when components are far apart in spatial frequency, many experimenters have found evidence for nonlinear combination [8,9,15,21].

Only a few experiments have addressed the issue of linearity in the recognition of more complicated stimuli. Legge et al. (1985) [11] and Solomon and Pelli (1994) [20] have proposed that letter identification depends on a single linear filter, similar to those mediating simple grating detection. Braje et al. (1995) [3] also modeled object recognition results using a single linear filter. None of these studies performed an explicit test of linearity, however. Loftus and colleagues [4,12,13] tested and found support for a linear model of integration over time, rather than space, in experiments where the task was identification of digit strings.

Nonlinearities are common in cognitive models of object recognition, particularly those that emphasize the role of 'top-down' processing. Data from recent studies support the idea that responses to components at different scales combines non-linearly over the time-course of object recognition [18,19]. In particular, these

* Corresponding author. Fax: +1 519 7468631; e-mail: esolds@cgl.uwaterloo.ca.

0042-6989/98/$19.00
PII: S0042-6989(97)00393-3

studies found that the value of high-spatial-frequency component stimuli depends upon the amount of low-spatial-frequency information that has already been processed. This argues for multiplicative combination of responses to different scales.

Using mixture experiments, we tested whether responses to component stimuli combine linearly in object recognition. In all the experiments described below, we used the same classes of component and compound stimuli. The compound stimuli were black and white images of common objects. These were termed the intact images. One set of component stimuli was generated by blurring the images with a Gaussian filter. This procedure removed much of the high-spatial-frequency content of the images; these component stimuli were termed the low-spatial-frequency or LSF images. A second set of component stimuli was generated by subtracting the LSF images from the intact images. These were termed the high-spatial-frequency or HSF component images. Subjects viewed the images and named the objects that they contained.

Prior work has indicated that recognition performance does not increase linearly with exposure duration [13]. Instead, percent correct recognition increases roughly exponentially, and asymptotes, as expected, at 100%. Because of this, we cannot test whether component images combine linearly in recognition by simply adding response curves. Instead, we developed a model that contains a linear first stage, followed by a static non-linearity that maps the response of the first stage into percent correct recognition performance. We fit this model to data from our LSF, HSF and intact conditions as a test of linearity.

2. A model of recognition

Recently, Loftus and his colleagues developed a model relating recognition performance to exposure duration [4,12,13]. They measured performance in a task where subjects viewed and reported digit strings from brief presentations. They used a variety of viewing conditions that included varying stimulus contrast, inserting a gap within the stimulus presentation, and using a patterned masking stimulus. Data from these experiments were well fit by a model that combined a linear-filter sensory response with an exponential information extraction process.

The first component of the model is a sensory response, $a(t)$, that behaves linearly with respect to stimulus duration and contrast (see Fig. 1C). This sensory response, $a(t)$, models the early response of the visual system as the convolution of a stimulus input function, $f(t)$, and an impulse response function, $g(t)$:

$$a(t)f(t)*g(t) \tag{1}$$

The stimulus input function $f(t)$ is physical stimulus contrast plotted as a function of time (see Fig. 1A); the impulse response function is a gamma function with two free parameters (see Fig. 1B, n and τ). The sensory response, $a(t)$, is effectively a temporally blurred version of the stimulus input function.

The nonlinear portion of the model is based on a feature sampling process that takes as input, the sensory response. One property of this stage is a threshold, θ, below which information extraction does not occur (see Fig. 1C). Formally, the thresholded sensory response, $a_\theta(t)$, is defined as:

$$a_\theta(t) = \begin{cases} a(t) - \theta & \text{if } a(t) \geq \theta \\ 0 & \text{if } a(t) < \theta \end{cases} \tag{2}$$

Information extraction is assumed to occur at a rate that is proportional to the cumulative sum of the above-threshold sensory response, $A_\theta(t)$, where

$$A_\theta(t) = \int a_\theta(t) \tag{3}$$

Given this rate of extraction, and assuming that performance is proportional to the total number of features acquired from a stimulus, Loftus and Ruthruff (1994) [13] showed that percent correct, p, varies with exposure duration, d, such that:

$$p = 1 - e^{-A_\theta(d)\,\phi} \tag{4}$$

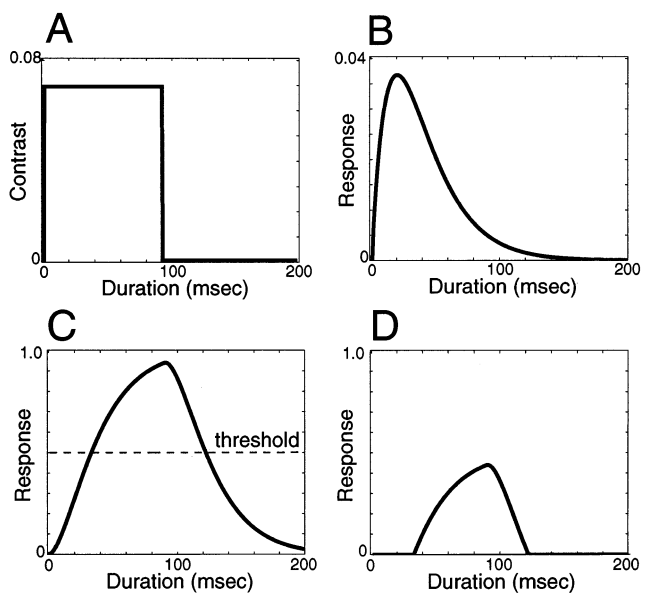

Fig. 1. Model of recognition. (A) An example stimulus input function, $f(t)$. (B) An example impulse response function, $g(t)$. (C) The sensory response, $a(t)$, which is the convolution of the stimulus input function with the impulse response function $f(t)*g(t)$. The dotted line indicates the threshold, below which feature sampling does not occur. (D) The thresholded sensory response, $a_\theta(t)$. Feature sampling occurs at a rate proportional to the area under the curve. Percent correct recognition is proportional to the total number of features sampled.

where ϕ is the constant of proportionality that relates the information-extraction rate to the cumulative sum of the sensory response. Loftus and his colleagues found that this model provided good fits to data collected from a range of experimental conditions. While this model was developed using data from a character reading task, it was inspired by models used for explaining detection data. Busey and Loftus (1994) [4] discuss the close similarity between this model and Watson's (1979) [22] model of detection data. In addition, several linear filter models have been proposed recently for recognizing letters [11,20] and objects [3].

3. Extending the model

We tested for linearity by evaluating the hypothesis that the LSF and HSF component images generate sensory responses that are linearly combined prior to feature sampling. This requires a slight extension to the model: we allowed the sensory stage to be differentially sensitive to LSF and HSF information. These sensitivities, s_L, and s_H, scale the stimulus contrast yielding two sensory responses:

$$a_L(t) = s_L f(t)*g(t) \qquad a_H(t) = s_H f(t)*g(t) \tag{5}$$

When viewed alone, the sensory responses generated from the component (LSF, HSF) stimuli are processed according to the original model yielding percentages correct, p_L, and p_H, respectively:

$$p_L = 1 - e^{-A_{L\theta}(d)\phi} \qquad p_H = 1 - e^{-A_{H\theta}(d)\phi} \tag{6}$$

If linearity holds at the sensory response stage, then the compound (intact) stimulus will generate a sensory response that is the sum of the two sensory responses generated by the component stimuli. Our model makes this assumption of linearity, and computes the sensory response as $a_I(t)$:

$$a_I(t) = a_L(t) + a_H(t) \tag{7}$$

It should be noted that this is subthreshold summation. The rest of the model is unchanged, and so after thresholding percent correct for the compound stimulus, presented for duration d, is:

$$p_I = 1 - e^{-A_{I\theta}(d)\phi} \tag{8}$$

where $A_{I\theta}$ is the thresholded integral of the intact sensory response, $a_{I(t)}$.

In experiments 1–3, we tested for linearity by evaluating the fit of the model to performance curves for LSF, HSF, and intact images. The model has six parameters: s_L and s_H, the relative sensitivities to low and high spatial frequencies; θ, the threshold; ϕ, the constant that relates information extraction rate to sensory response; and n and τ, the parameters of the sensory impulse response function. In the model fits reported here, for all three experiments, n was fixed at 9 and τ at 3; these values were fixed based on fits to pilot data. Only the other parameters (s_L, s_H, θ, and ϕ) varied, leaving four free parameters. Model fits were maximized and evaluated using a maximum likelihood method [22].

4. Experiment 1

In Experiment 1, we measured performance as a function of stimulus duration, for intact, LSF, and HSF images of objects. If information is combined in a predominantly linear fashion, then intact performance should be predictable from the performances on the component stimuli according to the model.

4.1. Methods

Subjects. The two authors and two naive subjects participated. All subjects had performed at least 1000 trials with these stimuli prior to beginning this experiment. Subjects' vision was normal or corrected-to-normal.

Stimuli. Black and white photographs of 32 common objects were digitized for presentation on a Macintosh computer. (See Fig. 2, Appendix A). In all the image operations reported here, images were represented in terms of gamma-corrected linear intensity units. The intact versions of these images were scaled by a factor of 0.333 and added to a constant image of 0.333. We did this to insure that the subtraction operation described below did not result in images with negative intensities. To avoid ceiling levels of performance, the contrast of the images was further reduced by an additional 50%. These images were used as the intact images in Experiment 1. Images of objects subtended a visual angle of $\sim$5.7 by 5.7°.

LSF versions of the stimuli were created by convolving each intact image with a 2-D Gaussian filter. The Gaussian filter had a support of ten pixels square and a standard deviation of two pixels in each dimension. This filtering procedure removed much of the high-spatial-frequency content from the images. In the Fourier domain, the filter had a standard deviation of $\sim$20 cycles per object. This corresponded to a standard deviation of $\sim$3.5 cycles per degree.

We created HSF images by subtracting each LSF image from its corresponding intact image, and adding this difference to the mean of the intact image. Thus, when measured in contrast, each pair of LSF and HSF subimages summed pixelwise to an intact image[1]. When

[1] There are several possible definitions of contrast for complex images. In this paper, we define each pixel's contrast as its difference from the image mean, measured as a percentage of the image mean.

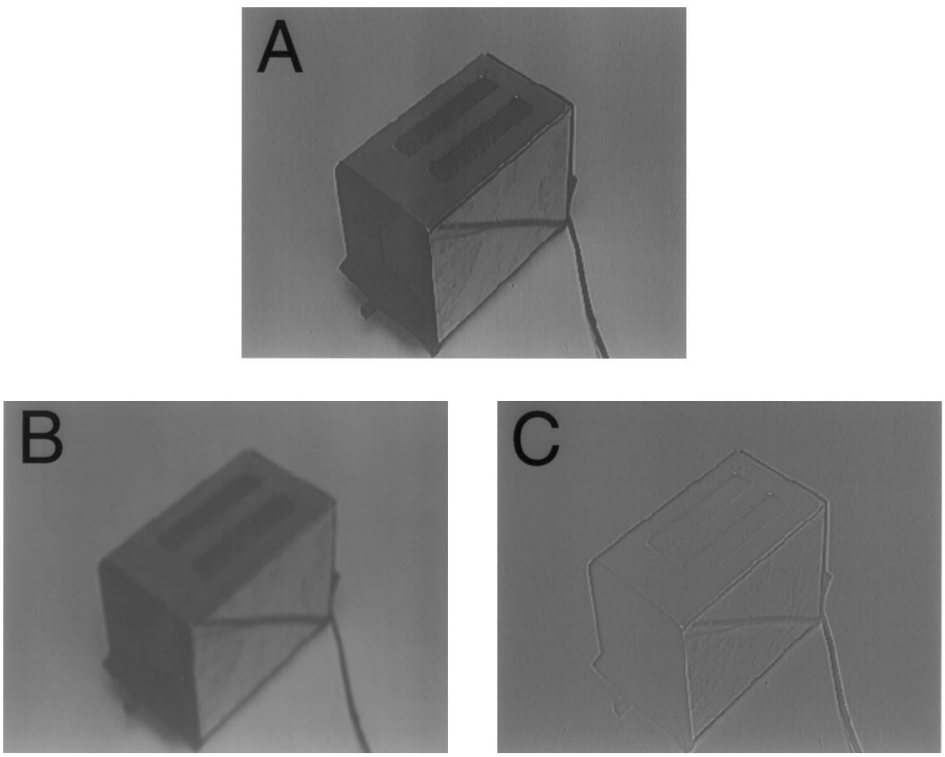

Fig. 2. Sample stimuli. (A) An intact image of a toaster. (B) An LSF image of a toaster, created by blurring with a Gaussian filter. (C) An HSF image created by subtracting the LSF image from the intact image, and adding back the intact image mean. The contrast of these images has been raised to aid in reproduction.

measured in intensity, each intact image was the pixel-wise mean of a pair of LSF and HSF subimages. We also created a mask image, consisting of a jumble of lines on a grey background, which was presented at full contrast. The mean luminance of each image was 54 cd/m^2.

Design. Each of the 32 objects was presented in LSF, HSF, and intact images at each of six exposure durations, yielding 576 trials in 18 conditions. The exposure durations used were 17, 33, 50, 67, 83 and 100 ms. Subjects completed two independently randomized sets of 576 trials each. Each set was completed in a session that lasted $\sim$1 h.

4.2. Procedure

The experiment was run on a Macintosh computer using the Psychophysics toolbox [2] for MATLAB (Mathworks). Subjects used a chin-rest to maintain a fixed viewing distance.

Subjects viewed a fixation point in the center of the computer screen. Following a warning tone, a stimulus image was presented at the center of the screen for a given exposure duration. The image was then replaced by the mask image, which was displayed for 1500 ms in all conditions. Subjects responded by typing the name of the object depicted in the image, and the computer then displayed the correct object name.

4.3. Results and discussion

The individual subjects' data and model fits are presented in Fig. 3. Performance on the LSF images was worse than that on the intact images, and performance was lowest on the HSF images.

The difference between LSF and HSF performance is most likely due to the size of the filter used to create the LSF and HSF images. A larger filter would probably have resulted in more equal performance between the LSF and HSF conditions (Experiment 3 verified this intuition).

The model fit the data well, supporting the idea that sensory responses to different spatial frequencies are combined linearly in the object recognition process. The model explained 99.5, 99.6, 95.8 and 99.3% of the variance in the data from the four subjects, respectively. These results are consistent with the findings of Solomon and Pelli (1994) [20] and Braje et al. (1995) [3]. They do not support interactive models of object recognition.

The model fit least well for subject KS. This may be a genuine individual difference, or it may simply be due

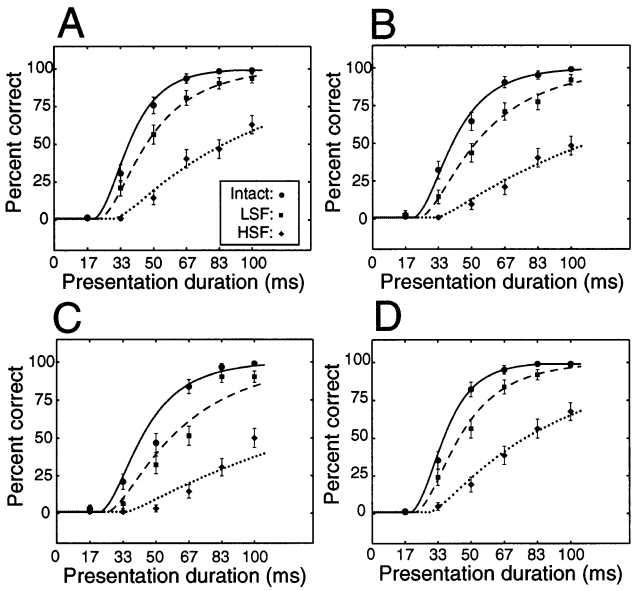

Fig. 3. Results from Experiment 1. Performance is plotted for (A) Subject EO; (B) Subject SE; (C) Subject KS; and (D) Subject PB. Performance (percent correct) for intact (circles), LSF (diamonds), and HSF (squares) images. The solid line plots the model fit for intact performance. The dashed line plots the model fit for LSF performance. The dotted line plots the model fit for HSF performance. Error bars show two S.E.M.

to noise in the data. See Table 1 for the model parameters that produced the optimal model fits shown in Fig. 3. As stated before, $\tau = 3$ and $n = 9$.

5. Experiment 2

In the previous study, observers were well practiced at recognizing our stimuli, having viewed over 1000 presentations prior to beginning the experiment. It is possible that our findings of linearity only pertain to situations where stimuli are overlearned in this manner. So, to test whether unpracticed object recognition also shows linear combination of information across spatial frequencies, we conducted an experiment in which each subject only viewed each stimulus once. Because of our limited number of stimuli, we used a between-subjects design.

Again we measured performance as a function of stimulus duration. If responses to LSF and HSF stimuli combine in a linear fashion, then intact performance should be related to the performances on the component stimuli according to the model.

5.1. Method

Subjects. A total of 48 Stanford University students participated for credit in an introductory psychology course. All had normal or corrected-to-normal vision.

Stimuli. The stimuli were created with the same filters as those described in Experiment 1, but they were presented at twice the contrast. The display used, and the mask used, were the same as those used in Experiment 1.

Design. Spatial frequency was manipulated between subjects; one third of our subjects viewed one stimulus type (intact, LSF and HSF) only. The exposure durations were 50, 83, 117 and 150 ms.

The 32 images were divided into four groups of eight. Each subject was tested on one group of images at each exposure duration. To avoid confounding the effects of individual objects and exposure duration, object groups and exposure durations were rotated through a between-subjects Latin-square design.

5.2. Procedure

The procedure was the same as that used in Experiment 1, except that each subject only performed 32 trials. Each subject performed two practice trials before performing the experimental trials. The objects used in the practice trials were different from those used in the rest of the experiment.

5.3. Results and discussion

Percent correct scores for each subject, exposure duration, and stimulus condition (intact, LSF, HSF) were calculated from the data. These were averaged to produce a mean and standard deviation for each exposure duration and stimulus condition. The best-fitting model parameters are included in Table 1. The model explained 97.4% of the variance in the data.

The excellent model fits (Fig. 4) agree with the results of Experiment 1. Linear combination across spatial

Table 1
Model parameters for Experiments 1–3

Subject	Expt. 1				Expt. 2	Expt. 3	
	EO	SE	KS	PB	48 Ss	SE	CF
s_L	0.12	0.12	0.23	0.09	0.07	0.03	0.11
s_H	0.06	0.06	0.13	0.05	0.04	0.02	0.11
ϕ	0.49	0.55	0.20	0.50	0.19	9.60	1.29
θ	0.04	0.03	0.09	0.03	0.01	0.001	0.006

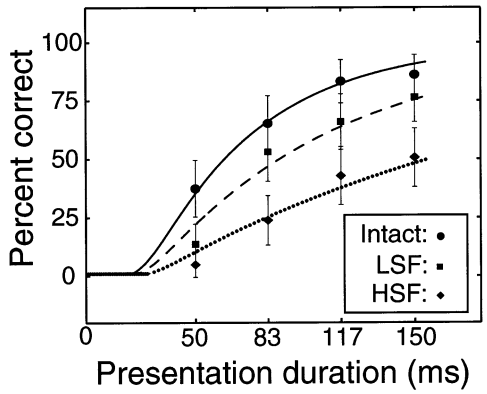

Fig. 4. Results from Experiment 2. Data and fits are plotted as in Fig. 3.

frequencies holds for the recognition of objects in images that have not been practiced.

6. Experiment 3

The previous experiments provided preliminary evidence for linear combination of responses in object recognition. In Experiment 3 we provided a much stronger test of linearity by fitting the model to a much larger data set. We created a second set of LSF and HSF component images by linearly scaling the first set, and we measured performance on both sets of component images and the compound images formed by summing them in various combinations. Again, we tested linearity by fitting our model, in which the sensory responses to the component stimuli sum to equal the sensory response to the compound stimuli.

This experiment also tests a second property of linear models: that scaling an input causes an equal scaling in response. This is an assumption of our model; hence if our model fits the data, then we have evidence that sensory responses scale with contrast in object recognition. In the model, a doubling of stimulus contrast would correspond to multiplying the right hand side of Eq. (5) by a factor of 2.

In this experiment, percent correct recognition was measured for LSF, HSF, and compound images. LSF and HSF images were presented at two different levels of contrast, and compound images were created by summing all possible combinations of the component images. This yielded four conditions for the component images presented alone, and four compound conditions, for a total of eight conditions.

In addition to the greater number of stimulus conditions, this experiment differed from the previous experiment in several ways. First, the sizes of the images were normalized such that the largest dimension of each object was a constant size, and the mean of each image was set to the same level. This was intended to reduce noise in the data by making each image more similar. In addition, a larger Gaussian filter was used to produce the LSF and HSF images (Fig. 5). This was intended to make performance in the LSF and HSF conditions more equal. Finally, to increase efficiency in data collection, exposure durations were controlled using a staircase procedure.

6.1. Method

Subjects. Author SE and naive subject CF participated. Both subjects were well practiced in the task, having performed at least 1000 trials prior to beginning the experiment.

Stimuli. The stimuli used in Experiment 1 were modified for Experiment 3. First, the size of the objects in the images was adjusted so that the largest dimension of each object subtended $\sim 4°$ of visual angle. The mean luminance of each image was again 54 cd/m^2.

Additionally, a larger Gaussian filter was used to create the LSF images from the intact images. The filter had a standard deviation of ~ 5 cycles per object (on the longest dimension), or 1.2 cycles per degree of visual angle. As in Experiment 1, HSF images were created by subtracting the LSF images from the intact images, and adding back the image mean.

Eight sets of images were created as follows. First, four sets of component images were created by scaling the contrast of the LSF and HSF filtered images by 0.07 and 0.14. Second, four sets of compound stimuli were created by taking the pixelwise sum of each LSF component with each HSF component and subtracting the image mean from each pixel. As in Experiment 1, the compound images are sums of the component images in terms of contrast and means in terms of intensity.

The mask in this experiment was different from the one used in Experiment 1. We created this mask by tiling 9-pixel-square patches drawn from random locations in the original images.

Design. Performance was measured for each of the eight image sets at eight exposure durations (13, 27, 53, 80, 107, 160, 213 and 426 ms). Exposure duration was controlled using a staircase procedure. Trials were performed in blocks of 64 trials, where each block used one image set and consisted of two repetitions of two interleaved 16-trial staircases. Image sets were assigned to blocks in random order. Subjects SE and CF participated in six blocks of trials for each image set.

6.2. Procedure

The procedure on each trial was identical to that used in Experiment 1, except that the correct object name was only presented after trials on which the subject responded incorrectly.

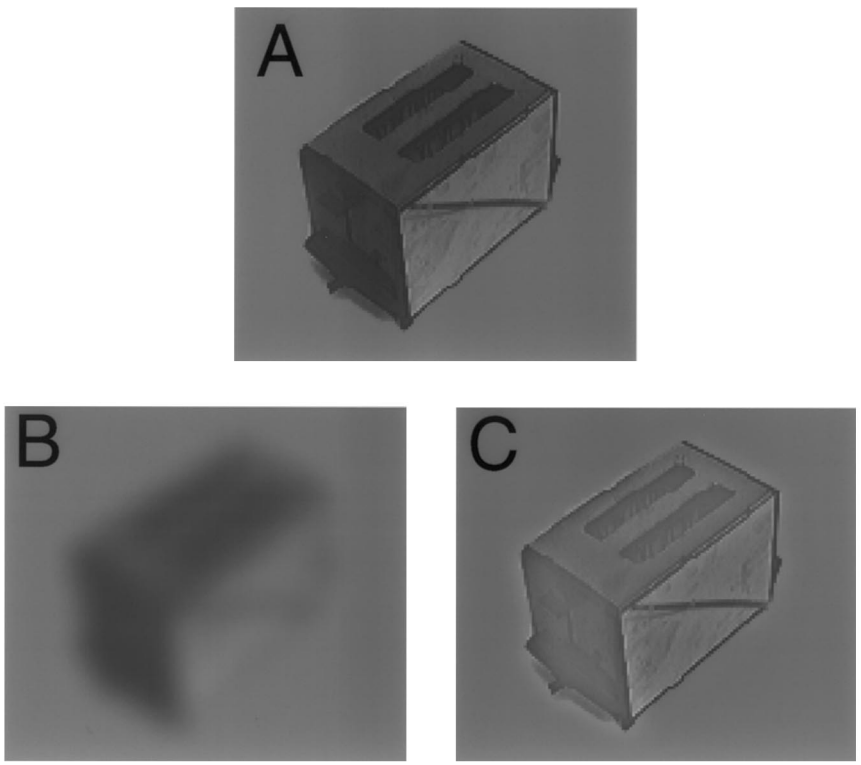

Fig. 5. Sample stimuli used in Experiment 3. (A) An intact image of a toaster. (B) An LSF image of a toaster, created by blurring with a Gaussian filter, which was larger for these stimuli than for the stimuli used in Experiments 1 and 2. (C) An HSF image created by subtracting the LSF image from the intact image, and adding back the intact image mean. The contrast of these images has been raised to aid in reproduction.

6.3. Results and discussion

The subjects' performance in this experiment is plotted in Figs. 6 and 7. Performance increased with image component contrast. In addition, performances on the LSF and HSF images were more equal in Experiment 3 than in Experiment 1. This is presumably because creating the component images with the larger Gaussian filter increased the amount of information in the HSF component stimuli.

The results of fitting the model simultaneously to the data for all eight stimulus sets are shown in Figs. 6 and 7. The model fits are good; see Table 1 for the model parameters. The model explained 97 and 93% of the variance in the data for the two subjects. The data from the images with the lowest contrast do not seem as

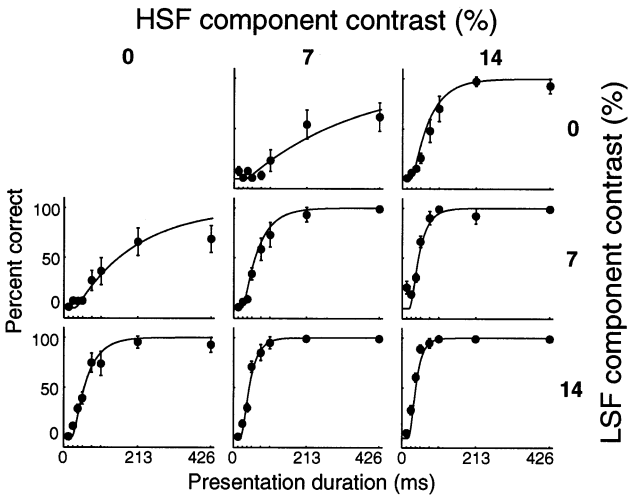

Fig. 6. Results of Experiment 3 for subject SE. The symbols plot percent correct recognition as a function of exposure duration, for eight stimulus conditions. The position of the plot indicates the contrast of the two component stimuli that were summed to create the compound stimuli. Zero contrast indicates that only one of the components was presented. The solid lines show the fit of the model.

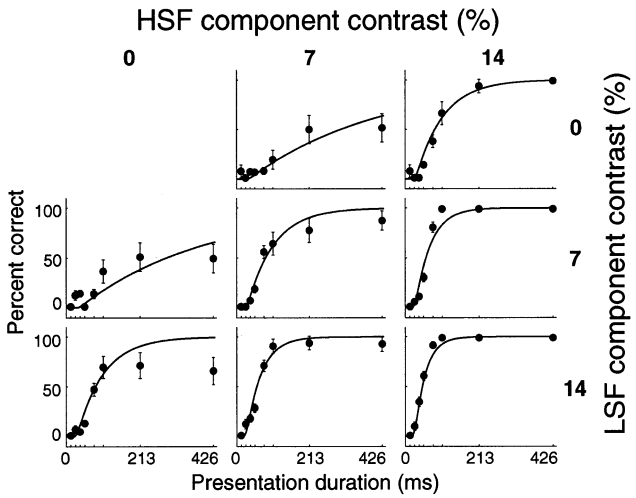

Fig. 7. Results of Experiment 3 for subject CF.

reliable and are not fit as well by the model as the other data.

Note that this experiment contained two kinds of stimulus variations. The model accounted for effects due to adding stimuli together, and also for effects due to scaling stimuli. Hence, in object recognition, responses appear to scale with stimulus contrast. Responses also appear to combine linearly, confirming the results of Experiments 1 and 2.

7. General discussion

The results of all three experiments provide evidence for an early linear stage in object recognition. This was demonstrated by our ability to account for the data with a model whose first stage sums responses to low and high spatial frequencies. The generality of our findings was tested using two sets of component stimuli (Experiments 1 and 2 vs. Experiment 3) that differed in the relative amounts of low and high spatial frequency information that they contained. We further tested the generality of our results by testing subjects who had no previous exposure to our stimuli (Experiment 2) and by varying the contrasts of the component stimuli (Experiment 3).

In general, our results support models of object recognition that possess a linear first stage, including two recent linear filter models proposed for recognizing letters [11,20] and objects [3]. While both these studies varied the spatial frequency content available for recognition, neither conducted an explicit test of the linear combination rule underlying their models. Hence, our results provide an important confirmation of linear filter models of recognition.

The overall goodness of fit of our extension of Loftus's model of visual processing replicates the success that Loftus's model has had in fitting other data [4,12,13]. However, we do not believe that ours is the only model that can explain our results. Other models possessing a linear first stage followed by a static non-linearity may also fit these data well [22].

Prior studies, varying the spatial frequency content for letter [17,20] and object recognition [3], have primarily been concerned with estimating the relative importance of different frequency bands for these tasks. In general, these studies found that mid-range frequencies of roughly 6–10 cycles per object are most important for recognition. Our data are in general agreement with these studies. Experiments 1 and 2 showed that fairly high spatial frequencies are of reduced importance for object recognition. However, because our component images were created by Gaussian filters, which are relatively broad in the frequency domain, our data do not permit a more fine-grained analysis of the relative importance of information from different frequency bands.

Our model proposes that the low and high spatial frequency information in object recognition are processed by a single linear filter or channel. In other tasks, such as grating detection, results suggest that low and high spatial frequency information are processed independently, and are combined relatively late in processing using probability summation. Such results are especially common when component stimuli are widely separated in spatial frequency. These multiple-channel models have successfully explained detection of many stimuli [23,24], in particular when the components differ widely in spatial frequency (for a review see Graham, 1989 [9]).

To evaluate whether our data could be explained in this manner, we developed a two-channel version of our model. Formally, the two-channel model does not differ from our original model in how it predicts performance for the component stimuli. These predictions are still described by Eq. (6). For the two-channel model, however, performance on the intact stimuli is no longer predicted by Eq. (8). Instead, intact performance is computed directly from performance on the component stimuli using probability summation:

$$p_I(d) = p_L(d) + (1 - p_L(d))\, p_H(d) \quad (9)$$

The two-channel model did not account for the data as well as our original single-channel model had. The clearest violation of the two-channel model predictions can be seen in the plots of the results of Experiments 1 and 2. Both these experiments show effects of subthreshold summation. At the shortest exposure durations, performance on one of the component stimuli is not significantly different from zero, while performance on the compound stimuli is significantly higher than performance on both components. This pattern is called subthreshold summation because the response to one of the two components is below a level where it affects behavior; such a level is typically called a threshold. Yet this subthreshold response produces a boost in performance when the two components are presented simultaneously. Probability summation, however, predicts that when performance on one component is zero, performance on the other component should be equal to performance on the compound stimulus.

We tested the two-channel model more formally by fitting it to our data. Fig. 8 shows a scatter plot of the errors for the two models plotted for the data points from all four subjects in Experiment 1. If the two models fit each data point equally well, all the plotted errors would fall along the diagonal. Since twice as many of the errors fall below the diagonal than fall above it, we conclude that the original model fit the data better than the two-channel model did.

Finally, we also evaluated the relative goodness of fit of the two models by computing the total variance in the data explained by each model (r^2). In Experiments

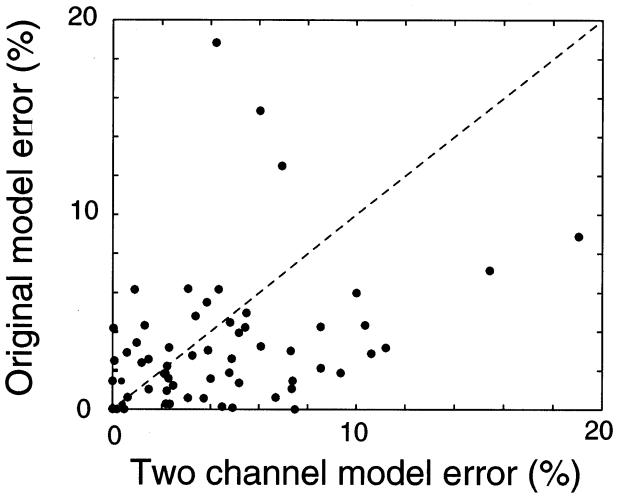

Fig. 8. Comparison of original model with two-channel model. Each symbol is one data point from Experiment 1. In general, errors are larger for the two-channel model, causing the majority of the points to fall beneath the diagonal.

1 and 2, the two models explained virtually indistinguishable amounts of variance. But larger differences were found in Experiment 3, where the original model explained 97 and 93% of the variance in the data for the two subjects. The two-channel model explained only 90 and 89% of the variance in the data. In summary, we find three sources of support for our linear combination model over a probability summation model: subthreshold summation, the scatterplot of model errors, and the overall amount of variance explained.

Probability summation models are a member of a class of multiplicative accounts of response combination in object recognition. Other accounts of object recognition have emphasized the importance of global information in providing a framework within which local information is interpreted [14,18]. These 'global-to-local' models propose that the value of local responses changes depending upon the amount of global information in the stimulus. Our results do not support such models; instead, LSF responses simply sum with whatever HSF response was generated by the stimulus. It is important to note, however, that we have tested linearity only under one particular (Fourier) decomposition of the stimulus. It is possible that other methods of decomposing images of objects into 'global' and 'local' information may reveal multiplicative interactions.

Finally, a number of studies have demonstrated that responses to LSF stimuli are faster than responses to HSF stimuli [10,19]. This is a separate issue from how the responses are combined; the differently lagged HSF and LSF responses can still be combined using any possible rule. The model that we fit to our data contains identical timecourse parameters for LSF and HSF responses. This is a more parsimonious explanation of our data than proposing different timecourses for the two responses. However, our data do not contain a strong test of this aspect of the model, since the stimulus timecourses for the LSF and HSF components were always the same when the two were presented together.

Psychophysical channels often have been tentatively hypothesized to correspond to relatively small populations of V1 simple cells [9,16]. Simple cells, like channels, are linear devices, summing activity generated by stimuli within a relatively narrow range of spatial frequencies [7]. Our data agree with others in providing evidence that complex pattern recognition may in some cases be supported by a single early linear channel. Hence, one interpretation of our findings is that the neural pathways that perform object recognition receive input from a relatively small population of V1 simple cells. The linearity we find in recognition may reflect the linearity of these V1 neurons. Many further tests will be required to confirm this hypothesis.

In conclusion, object recognition in humans appears to be mediated by an early linear stage that sums responses across spatial frequency. Future work will provide additional tests of the linearity, as well as examine the later, nonlinear stages of processing.

Acknowledgements

This work was supported by a Defense Department A.F.O.S.R. grant to EO. These experiments comprised part of EO's Ph.D. dissertation, at Stanford University. We are grateful for helpful comments and assistance from Geoffrey Boynton, Christopher Furmanski, David Heeger, Geoffrey Loftus, David Rumelhart, Philip Servos, David Tolhurst, Brian Wandell, and Xuemei Zhang.

Appendix A. List of objects used in Experiments 1–3

Apple	Book
Boot	Bottle
Bowl	Bucket
Can	Clipboard
Cup	Dog
Fan	Firetruck
Flashlight	Football
Frog	Glasses
Hammer	Headphones
Iron	Kettle
Lemon	Lightbulb
Lighter	Pitcher
Record	Scissors
Screwdriver	Stapler
Teapot	Television
Toaster	Umbrella

References

[1] Blakemore C, Campbell FW. On the existence of neurones in the human visual system selectively sensitive to the orientation and size of retinal images. J Physiol 1969;203:237–60.

[2] Brainard DH. The psychophysics toolbox. Spatial Vision 1997;10:433–6.

[3] Braje WL, Tjan BS, Legge GE. Human efficiency for recognizing and detecting low-pass filtered objects. Vision Res 1995;35:2955–66.

[4] Busey TA, Loftus GR. Sensory and cognitive components of visual information acquisition. Psychol Rev 1994;101:446–69.

[5] Campbell F, Robson J. Application of fourier analysis to the visibility of gratings. J Physiol Lond 1968;197:551–66.

[6] De Valois RL, De Valois KK. Spatial vision. Annu Rev Psychol 1980;31:309–41.

[7] De Valois RL, De Valois KK. Spatial Vision. New York: Oxford University Press, 1988.

[8] Graham N, Nachmias J. Detection of grating patterns containing two spatial frequencies: A comparison of single-channel and multiple-channels models. Vision Res 1971;11:251–9.

[9] Graham NVS. Visual Pattern Analyzers. New York: Oxford University Press, 1989.

[10] Hughes HC, Fendrich R, Reuter-Lorenz PA. Global versus local processing in the absence of low spatial frequencies. J Cogn Neurosci 1990;2:272–82.

[11] Legge GE, Pelli DG, Rubin GS, Schleske MM. Psychophysics of reading—i. Normal vision. Vision Res 1985;25:239–52.

[12] Loftus GR, Busey TA, Senders JW. Providing a sensory basis for models of visual information acquisition. Percept Psychophys 1993;54:535–54.

[13] Loftus GR, Ruthruff E. A theory of visual information acquisition and visual memory with special application to intensity-duration trade-offs. J Exp Psychol: Human Percept Perform 1994;20:33–49.

[14] Navon D. Forest before trees: The precedence of global features in visual perception. Cognitive Psychol 1977;9:353–83.

[15] Olzak LA. Widely separated spatial frequencies: Mechanism interactions. Vision Res 1986;26:1143–53.

[16] Olzak LA, Thomas JP. Seeing spatial patterns. In: Boff KR, Kaufman L, Thomas JP, editors. Handbook of Perception and Human Performance, Vol 1: Sensory Processes and Perception, New York: Wiley, 1986:7:1–7:56.

[17] Parish DH, Sperling G. Object spatial frequencies, retinal spatial frequencies, noise, and the efficiency of letter discrimination. Vision Res 1991;31:1399–415.

[18] Sanocki T. Time course of object identification: Evidence for a global-to-local contingency. J Exp Psychol: Hum Percept Perform 1993;19:878–98.

[19] Schyns PG, Oliva A. From blobs to boundary edges: Evidence for time- and spatial-scale-dependent scene recognition. Psychol Sci 1994;5:195–200.

[20] Solomon JA, Pelli DG. The visual filter mediating letter identification. Nature 1994;369:395–7.

[21] Thomas J, Olzak L. Cue summation in spatial discrimination. Vision Res 1990;30:1865–75.

[22] Watson AB. Probability summation over time. Vision Res 1979;19:515–22.

[23] Watson AB. Summation of grating patches indicates many types of detector at one retinal location. Vision Res 1982;22:17–25.

[24] Wilson HR, Bergen JR. A four mechanism model for threshold spatial vision. Vision Res 1979;19:19–32.