Research Note

An Upper Limit to the Binocular Combination of Stimuli

LAWRENCE K. CORMACK,* SCOTT B. STEVENSON,† CLIFTON M. SCHOR†

Received 3 November 1993; in revised form 2 February 1994

It is important to know the spatial extent over which the binocular visual system searches for "matches" or image correspondence. Most models of stereopsis define fixed neighbourhoods in one monocular image in which a search is conducted for a match to some element in the other image. We were unable to experimentally determine fixed values for the extent of these neighbourhoods. We were, however, able to derive a simple rule that predicts performance on a binocular matching task over a large spatio-temporal region, and from which we can calculate the efficiency of our observers. We conclude that the human visual system does not use neighbourhoods of a single, fixed extent to perform binocular matching in a single region of the visual field, but uses more flexible scheme that allows it to function well under a broad range of stimulus conditions.

Introduction

An understanding of the matching process (or the act of solving the "correspondence problem") is of fundamental import in the study of binocular vision and stereopsis. The aim of this study was to determine the spatial and temporal bounds over which this matching process operates; in other words, this study sought to determine the spatial and temporal integration areas for binocular matching.

Typically, studies have used a disparity based criterion by which to gauge the success or failure of the matching process. For example, the detection of a form in depth or a depth (disparity) defined edge is used as an indication of successful matching (e.g. Fender & Julesz, 1967). Yet most, if not all, computational studies or models of stereopsis divide stereoscopic processing into three stages: preprocessing, matching and depth recovery. These stages seem inescapable, as they are distinguishable in models of all kinds, whether feature-based or area-based. Thus, if one wishes to measure aspects of the matching stage, it would seem best to avoid tasks which rely heavily on depth recovery for successful judgments (this is not meant to suggest that useful information does not come from tasks that rely on depth recovery).

This study employs a paradigm in which thresholds for the detection of interocular correlation (IOC) are measured (Tyler & Julesz, 1976; Cormack, Stevenson & Schor, 1991; Stevenson, Cormack, Schor & Tyler, 1992). In this paradigm, subjects are not required to make any judgements requiring depth recovery, they are only judging whether the images of the two eyes are statistically correlated with one another. The logic behind this approach is simply that if one wants to measure how well the visual system matches the images of the two eyes, the appropriate stimulus parameter to vary is the degree to which the two eye's images match. One can quantify how well two images match by computing the IOC of the two images. This is particularly intuitive and convenient in the case of random-element stereograms (Aschenbrenner, 1954; Julesz, 1971) in which the IOC is linearly related to the proportion of matching elements. Figure 1 shows a stereogram in which the IOC ramps from 1.0 in the uppermost and lowermost raster lines to 0 in the central raster line.

The binocular combination of contrast information must occur in piecemeal fashion across the visual field. For example, a model of binocular combination based on a global cross-correlation of the two eyes' inputs would fail to discriminate between a random-dot stimulus composed of two surfaces alternating in depth across...
the visual field (a cyclopean depth grating) and a stimulus composed of two overlaid surfaces (cyclopean transparency). In fact, if the alterations in depth between two random-dot surfaces are sufficiently fine in space (Parker & Yang, 1989) or in time (Stevenson, Cormack & Schor, 1989; Cormack, Stevenson & Schor, 1993) transparency (or disparity averaging, depending on the disparity's magnitude) will be perceived. Weinshall (1991) has created displays that present a challenge to most matching algorithms but are solved by a spatially restricted cross-correlation. These findings hint at the finite spatial and temporal extent over which the combination of binocular information of the two eyes takes place, that is the size of a cyclopean visual direction. Other evidence for the localized nature of binocular combination comes from the observation that, while stereopsis and rivalry can coexist (Kaufman, 1974; Julesz & Miller, 1975), they cease to do so when the targets are made relatively small (Blake, Yang & Wilson, 1991).

In the first experiment, we set out to determine the spatial and temporal bounds over which the binocular matching process operates, as determined by measuring IOC thresholds for dynamic random-element stimuli of various sizes and durations. In the initial experiment stimulus size and duration were varied as the independent variables and, since element size and the size of an element in time were held constant, the total number of elements present in the stimulus varied as well. Commensurate with earlier experiments addressing spatial and temporal integration (e.g. Davila & Geisler, 1991; Gorea & Tyler, 1986, respectively) we anticipated that threshold would decrease linearly with increasing area (or duration) until the integration region was surpassed, after which threshold would remain constant as area (or duration) was increased further. In other words, a plot of threshold vs area (or duration) on log-log coordinates should show a slope of -1 until the integration region is surpassed, and a plateau (or much shallower slope) afterwards.

To anticipate, we found that no consistent estimate of a spatial or temporal integration could be gleaned from the data. Rather the absolute number of stereogram elements presented to the subject (the product of the number of elements per video frame and the total number of video frames) proved to be the best predictor of subjects performance. This dependence on the total number of stereogram elements was tested and confirmed in a second experiment, in which the number of stereogram elements was varied as the independent variable, but the total area and duration of the stimuli were held constant.

EXPERIMENT 1

Methods

The stimuli were one-bit dynamic random-element stereograms of a flat surface at zero disparity (no "form-in-depth" was ever present). Stimuli (random-bit streams) were generated in hardware and fed to a pair of matched video monitors viewed through a haploscope. In this experiment, dynamic noise was continuously present. During one of the two intervals, the noise changed from zero IOC to some positive IOC. Stimulus duration was controlled by varying the number of video frames which composed an interval, and the area was controlled via iris diaphragms in front of the monitors. Element size was constant at 2.3 min arc vertical x 5 min arc horizontal. More exhaustive details of the stimulus generation and presentation have been given elsewhere (Cormack et al., 1991).

Vergence was controlled with a fixation mark (a high contrast "+" the arms of which were 12 min arc long x 5 min arc wide) flanked vertically by nonius lines (high contrast lines 48 min arc tall x 5 min arc wide). While viewing dynamic random element displays, our subjects can detect vernier offsets in the nonius lines corresponding to <0.5 min arc of vergence fluctuation. Moreover, they can maintain vergence posture with a standard deviation of below 5 min arc under these conditions as assessed by objective eye movement recording with an SRI DPI Generation V.eye tracker (Stevenson, Cormack, Schor & Tyler, 1994).

A temporal two-alternative forced-choice (2AFC) paradigm was used in which the subject's task was always to indicate (via a key press) the interval which contained the higher interocular correlation. The two intervals, which were delimited by tones, were separated by 500 msec.

Five levels of interocular correlation were used, chosen to bracket the subjects threshold for a particular condition such that the lowest stimulus level yielded chance performance and the highest yielded close to perfect performance. A block comprised each of the five levels of correlation and 30 such blocks composed a single run. The presentation order of stimuli within a block was determined by the method of constant stimuli. Each run, then, produced a psychometric function composed of five data points representing 30 trials each. Threshold (75% correct) was estimated for each run from the best-fit Weibull function. The subjects each did three runs per condition for this experiment.

To reiterate, the independent variables in this experiment were the area and duration of the stimulus presentation. The area was varied between 0.14 and 7.00 deg² and the duration was varied between 49.2 and 800 msec.

Two of the three authors (LKC and SBS) served as subjects. Both have normal (LKC) or corrected-to-normal (SBS) acuity, normal contrast sensitivity, and excellent (<5 sec arc) stereoacuity.

Results

The threshold vs area data (Fig. 2) from Expt 1 generally indicate a corner at roughly 1 deg², but this is more apparent in the data from longer than shorter durations.

The threshold vs duration data (Fig. 3) generally indicate a corner at roughly 200 msec, but this is much more apparent in the data from larger areas than in the data from smaller areas.
However, it must be stressed that in neither case (threshold vs area or threshold vs duration) was there a consistent, obvious corner in the data. This could be due to a number of reasons. First, we realized that the transition from perfect integration (in space or time) to probability summation between independent mechanisms adjacent in space or time might obviate the presence of sharp corners in the data.

An additional problem is the unknown manner in which channels interact in the binocular system. Assume for the moment that spatial frequency tuned channels operate independently for stereoscopic matching (cf. Julesz & Miller, 1975; Mayhew & Frisby, 1978; Yang & Blake, 1991) and that the size of the integration area, or the local neighbourhood over which matching is performed, is inversely proportional to the peak spatial frequency of the mechanism (cf. Marr & Poggio, 1979; Ohzawa, DeAngelis & Freeman, 1990). Since our stimuli contain power across a broad frequency spectrum, they will stimulate a large number of the putative spatial channels. As the size of our stimuli increase, they will progressively, rather than simultaneously, surpass the

![Disparity](image)

**FIGURE 1.** Upper panel: a stereogram in which the interocular correlation varies with vertical position. The interocular correlation is 1.0 at both the top and bottom of the stereogram and ramps to 0 at the centre of the figure. The reader's correlation threshold can be estimated by noting the vertical position at which the percept of a flat surface, apparent at both the top and bottom, completely gives way to one of an incoherent volume. Lower panel: cross-correlation functions resulting from cross-correlating horizontal slices through stereograms like those in the upper panel. These functions illustrate the nature of the task confronting the observers: a signal (the flat surface at zero disparity) must be detected against background noise. The background noise decreases (the cross-correlation functions become smoother) as the number of elements in the stimulus increases.
integration areas for these mechanisms. This would clearly prevent a clean transition from integration within a localized mechanism to probability summation between independent mechanisms sensitive to various spatial frequencies. Parallel arguments apply to the relationship between stimulus duration and temporal frequency channeling in the visual system.

Still, we hoped that the families of curves in Figs 2 and 3 would furnish a clearer estimate of the spatial and temporal bounds over which the mechanisms responsible for binocular combination operate. While the lack of an obvious corner was somewhat disconcerting, it was certainly not at odds with what we know about early visual processing.

Another factor needs to be considered however. Even though an experimenter specifies a particular level of correlation (essentially guaranteeing that a certain minimum number of dots will match) the actual correlation on any given trial will vary (because a great number of the dots are left to match, or not by chance). The variance of the actual correlation is proportional to the number of elements which compose the stimulus, which of course, varies with both the area and the duration of the stimulus.

Subjects in our 2AFC experiment were thus faced with the task of discriminating whether the random-dot sample presented on each trial came from one
distribution of correlation or another. Even an ideal stimulus comparator would have been limited by the statistics of the stimulus in this situation.

At this point, it is useful to consider a description of optimal performance in this task given the statistical nature of these stimuli. For this purpose, each eye's stimulus can be envisaged as a one-dimensional binary array (where $p = q = 0.5$. The $i$th element of the first array is set to 1 or 0 depending on whether a value drawn from uniformly distributed random variable is below $p$. The $i$th element of the second array is then made identical to that of the first array if a second random value is below $C$, the correlation between the arrays, otherwise it too is selected at random. A "binocular" array can then be constructed such that the $i$th value is 1 if the $i$th values of the two monocular arrays match, and 0 otherwise (this is equivalent to inverting a bitwise exclusive OR of the two arrays).

For this binocular array, with $p$ representing the probability of a match,

$$p = \frac{(C + 1)}{2}; \quad q = \frac{(1 - C)}{2}. \quad (1)$$

Thus given stimuli comprising $N$ elements, the expected number of interocular matches, $\mu$, is

$$\mu = \frac{N(C + 1)}{2} \quad (2)$$

with variance, $\sigma^2$, given by

$$\sigma^2 = \frac{N(1 - C^2)}{4}. \quad (3)$$

![Correlation threshold as a function of stimulus duration for two subjects. While, for some areas, a corner is present at between 200 and 400 msec, the data do not clearly define a period of temporal integration.](image)
Correlation threshold:
Human Observers

- $\mu_1 - \mu_2 / (\sigma_1^2 + \sigma_2^2)^{1/2}$

so, using equations (2) and (3):

$$\Psi \propto d' = \sqrt{2N(C_1 - C_2)} / \sqrt{2 - (C_1 + C_2)}.$$ (5)

Where $C_1$ and $C_2$ are the interocular correlations of the two signals to be discriminated. In these experiments, $C_2$ was always 0 and $C_1 \approx 2$, so equation (5) reduces to:

$$\Psi \propto C \sqrt{N}.$$ (6)

If observers were limited by the noise inherent in the stimulus, a plot of log threshold against log $N$ should follow a slope of $-0.5$. Moreover, there must come a point at which the visual system ceases to gain from additional information. We know this from informal observations in the lab and by comparison of the thresholds using the largest areas in the present study with thresholds from previous studies (e.g. Cormack et al., 1991) using much larger areas and comparable durations.

Thus, for each of the size-duration combinations used in this experiment, we computed the total number of elements presented to the subject by multiplying the number of elements contained in the stimulus area by the stimulus duration in video frames. The data from Figs 2 and 3 were replotted as a graph of threshold vs total number of elements, which is shown in Fig. 4. When plotted in this manner, the data collapsed onto a single curve. This curve follows the theoretical limit of discriminability (log-log slope of $-0.5$) up to a point, and then shows a well defined plateau over which performance is constant. Thus, over the range of stimulus conditions used, no fixed temporal or spatial integration region limited the performance of our observers. Rather, they continued to utilize the information available until some maximum number of stimulus elements was reached, regardless of the spatio-temporal distribution of those elements.
As this was a rather surprising result, a second, more direct, experiment was undertaken. Experiment 2 was designed to explicitly test the notion that the total number of dots is indeed the controlling variable in determining subjects' thresholds for the detection of IOC. In this experiment, stimulus area and duration were held constant while the total number of stimulus elements was varied. An ideal observer was also run in this experiment to demonstrate the theoretical dependence on performance on the total number of elements.

EXPERIMENT 2

Methods

In the second experiment, we measured IOC thresholds for stimuli comprising a variable number of total elements while holding total stimulus duration and area constant. We achieved this by allowing the element size to co-vary with the total number of elements [the proportion of white (or black) elements was constant at 0.5, i.e. stimulus density was fixed at 50%].

This experiment was run on a Macintosh microcomputer. Stimuli (pseudo random one-bit arrays) were generated in software trial-by-trial and displayed on an Apple 13 in. color monitor. The output of the system supplied random number generator was improved by the Bayes-Durham algorithm (Press, Flannery, Teukolsky & Vetterling, 1988). Stimulus area was held constant at just over 4 deg² and stimulus duration was 12 video frames (200 msec). Dynamic noise was only present during the two intervals, thus the stimulus "region" was sharply delimited in time as well as in space.

The fixation target in this experiment consisted of a small box (20 x 20 min arc) flanked vertically by nonius lines (20 x 2 min arc). The fixation target disappeared immediately prior to each stimulus presentation and reappeared on the subsequent video frame.

The psychophysics and data analysis were identical to that of the first experiment except that the subjects did five, rather than three, runs per condition.

An "ideal observer" was run in this experiment also. It ran in parallel with the human observers, making judgments on the same stimuli. These runs, however, did not produce useful psychometric functions due to the high sensitivity of the ideal observer relative to the stimulus levels used. Hence, the ideal observer was run on much lower stimulus levels, but using identical psychophysical procedures (five runs per condition, etc.).

Results

The data from Exp 2 are shown in Fig. 5. The range of the data from Fig. 4, omitting one datum per subject, is shown by the gray region. The omitted datum, identical for both subjects and lying almost directly above the point where the two line segments meet, is from the shortest duration/largest area combination and indicates that, for very brief durations, subjects failed to gain from an increase in stimulus area from 2.2 to 7 deg² (cf. Fig. 3).

Both sets of data are in good agreement and are in accord with the theoretical limits on performance imposed by statistics of the stimulus.

The open diamonds in Fig. 5 plot the performance of the ideal observer. The solid line through the half-solid diamonds indicates the theoretical ideal observer performance as determined by equation (6). The close agreement between these two indicates that the assumptions behind the ideal formulation are valid.

By comparing the human and ideal data, we can also say that the efficiencies of the human observers are rather high, between 3% and 10%. Moreover, it can be seen that the variability of threshold, as indicated by the vertical error bars (±1 SD) are roughly equal for the human and ideal observers. This indicates that much of the human subjects' response variability, and hence the slope of the psychometric function, is determined by the stimulus.

The half-solid diamonds in Fig. 5 plot the performance of the simulated ideal observer when it was restricted to counting edge matches only. This restriction has the effect of reducing the average number of matching elements, since luminance edges do not occur at every dot (e.g. a string of five black dots has only two edges). The decrease in ideal performance is a factor of 1.5 and the solid line through the half-solid diamonds indicates the theoretical performance under the edge sampling hypothesis. Thus, if we assume that human observers utilize only the edges of groups of black (or white) elements, we have a commensurate gain of efficiency.

DISCUSSION

These experiments demonstrate that the total number of elements is in fact the controlling variable in the determination of IOC thresholds over a substantial spatio-temporal region. This result is in accordance with both theoretical computations based on the statistics of the stimulus, and with results produced by an ideal observer run in the second experiment.

Most models of stereopsis define neighbourhoods of fixed extent (integration areas) over which to solve "the correspondence problem", that is, identifying which features in the left image correspond to features in the right image (e.g. Pollard, Mayhew & Frisby, 1985; Prazdny, 1985; Weinshall, 1991). We measured observers' ability to solve the correspondence problem directly. We presented dynamic, spatial white noise to each eye, and measured thresholds for detecting partial interocular correlations imposed on the noise. Thresholds were measured as a function of both area and duration. Performance improved with increasing sample size in accordance with the statistics of the stimuli until a fixed sample size was reached, regardless of the spatial or temporal extent of the stimuli. Thus, the amount of information observers gathered from the stimuli was not set by spatio-temporal integration limits per se, but by a fixed "buffer size", somewhat analogous to the 7 ± 2 limit for short term memory (Miller, 1956).
Harris and Parker (1992) have studied the effect of the number of stimulus elements in a stereogram on performance in a stereoscopic discrimination task. They also find that efficiency remained constant to a point, and then began to fall off with the increasing number of elements in the stimulus. The actual number of elements at which observers cease to gain additional information from additional dots was very different in the two studies; Harris and Parker (1992) found that observers failed to gain additional benefit from more than roughly 20 elements; this limit was some 2–3 log units higher in our study. Over the region where constant efficiency was observed, they also report efficiencies roughly an octave higher than those reported here. However, there are several important differences between the studies that must be considered. First, the task was different. Our observers were performing a correlation detection rather than a stereoscopic discrimination. The difference between the two tasks is analogous to the difference between detecting a spot of light (luminance detection) and judging its position relative to a second spot (vernier acuity). This task difference has been shown to manifest itself in the form of differential stimulus dependencies, even when the observers are the same and the stimulus is as similar as possible (Cormack et al., 1991).

Second, Harris and Parker (1992) used static stereograms that were displayed until the observer responded. Our displays were dynamic and lasted only the designed duration. It is likely that this difference resulted in an undersampling by our observers relative to those of the other study.

Finally, the dot densities used by Harris and Parker (1992) were very low compared to ours. Experiments currently underway in the laboratory of one of the authors indicate that high dot densities result in undersampling by observers. Naturally, this would result in a reduction of the calculated efficiencies of observers viewing high density displays. It is not known whether the two factors that resulted in a relative undersampling by

![Diagram of Correlation threshold: Human and Ideal Observers](image)

**FIGURE 5.** The data from Expt 2 are represented by the open symbols; squares show data from SBS and circles from LKC. In this experiment, the total number of elements was varied while the stimulus area and duration were held constant. The extent of the data from Expt 1 (with the exception of a single datum from each subject) is shown by the gray region. Note the close concordance between these data and those from Expt 1. This indicates that the total number of elements present, regardless of the spatio/temporal configuration limits performance in these experiments. The data from an ideal decision rule attending to individual elements (solid diamonds) or luminance edges only (half-solid diamonds) are also shown. Dashed and solid lines show the analytical formulations for these observers as given in the text. The efficiency of interocular correlation detection is roughly 5% for observer LKC (shown by the double headed arrow). This represents the fraction of elements required to equate human and ideal performance. This difference could be due to sampling inefficiency, in which case the human observer is actually sampling only 5% of the stimulus, but is able to use that 5% perfectly. Alternatively, it be due to computational inefficiency, in which case the human observer is actually sampling all of the available elements, but is doing a noisy calculation with them.
our observers can quantitatively account for the differences between the two studies, but both factors do serve to decrease, not increase, the discrepancy.

In modelling the combination of information from the two eyes, one successful approach is the use of a local (i.e. spatially restricted) cross-correlation of the retinal images, or some derivative of them (Weinshall, 1991; Nishihara, 1988; Stevenson et al., 1989).* Our experiment's initial aim was to determine the spatial and temporal bounds of this correlation process in human observers. We measured thresholds for the detection of interocular correlation as a function of the stimulus duration and area. We predicted that, for a fixed duration, performance would improve until the spatial integration area was exceeded, at which point performance would plateau. An analogous prediction was made for the effect of stimulus duration with stimuli of a fixed area. We failed, however, to find a particular area and/or time at which performance asymptoted. Instead, estimated integration area depended on stimulus duration and, conversely, estimated integration time depended on stimulus area. However, when the data were replotted as a function of the number of elements in the stimulus, collapsed across time and area, a consistent picture emerged. Observers' thresholds continued to improve until a fixed number of stimulus elements (i.e. a fixed sample size) was reached, regardless of the spatio-temporal configuration of the stimulus. Observers' performance improved as sample size increased in accordance with the theoretical improvement due to the increasing signal-to-noise ratio in the stimulus. This is conceptually identical with the improvement of contrast detection as a function of mean luminance, that is, photon sample size (Rose, 1942) once the absolute lower limit, set by the thermal breakdown of rhodopsin (Barlow, 1988), is exceeded. Eventually, however, a limit of about 10^4 dots was reached, beyond which observers failed to benefit from additional samples. This indicates that a storage, or information processing, limit was reached that was independent of the spatio-temporal bounds of the stimulus.

There are at least two means by which such behaviour could be realized in a simple system. First, the stimulus integration could be performed over neighborhoods whose size (in both space and time) could be adjusted to best suit current stimulus conditions. Such a scheme has been implemented in an artificial visual system by Okutomi and Kanade (1992). The notion of a malleable integration area is also not at odds with recent physiological evidence indicating that measured receptive field properties in the visual cortex are dependent upon the stimulus context in which these properties are measured (Gilbert & Wiesel, 1990).

A second means by which the observed behaviour could be realized is through the use of overlaid (concentric) receptive fields in a single region of visual space, as undoubtedly occurs in the central visual field (and which is the central tenet of models of visual processing based on parallel spatial frequency coding). The fact that receptive fields of different sizes in the same region of visual space are associated with different peak sensitivities in the spatial frequency domain doesn't effect this argument due to the broad band nature of the stimuli; the stimuli would uniformly excite neurons tuned to roughly 5 c/deg and below (including any "DC" of luminance-only cells). Moreover, this scheme would be expected to produce the gradual transition from complete spatial (or temporal) integration to complete size (or time) independence. Simple modelling of concentric, overlaid receptive fields (of "pillbox" profile) integrating an arbitrary spatial signal confirms the presence of a gradual transition as is seen in the data of Figs 1 and 2. Mechanisms and models aside, these observations demonstrate that under our conditions the performance of the visual system is quite good (on an objective level, in comparison with an ideal decision rule), and that it's performance is governed largely by the noise in the stimulus itself.

Perhaps the most remarkable aspect of these observations is the flexibility they suggest. The visual system must collect the maximum amount of information it can to perform a given task, but should not integrate beyond the spatial and temporal bounds of the stimuli, as such over integration would decrease signal-to-noise ratio (this is particularly apparent in our Expt 1, in which the stimulus was surrounded in time by uncorrelated dynamic random dots; had the visual system overintegrated, performance would have deteriorated at a much faster rate than that of the ideal observer). Thus it appears that if the binocular visual system is required to cross-correlate the information in the two eyes, it uses effective limits of spatio-temporal integration that are optimal to collect the maximum amount of information that it is capable of processing.

REFERENCES


*Cross-correlation models are by no means new in computer vision. They were out of favour for a time, particularly among vision researchers, due to their failure under general stimulus conditions. The failure of cross-correlation based models, however, can be largely blamed on improper image coding (representation) rather than on the method of image combination itself. While it is true that cross-correlation of raw intensity values is doomed to fail as a model of binocular vision, the point is moot because an image description based on raw intensity values is simply incorrect, even at the bipolar cell layer of the retina.


