Gepshtein S & Kubovy M (2000). Proceedings of the National Academy of Sciences, USA 97 (14), 8186-8191.

The emergence of visual objects in space–time

Sergei Gepshtein* and Michael Kubovy*

Department of Psychology, P.O. Box 400400, Charlottesville, VA 22904–4400

Communicated by Julian Hochberg, Columbia University, New York, NY, April 20, 2000 (received for review September 7, 1999)

It is natural to think that in perceiving dynamic scenes, vision takes a series of snapshots. Motion perception can ensue when the snapshots are different. The snapshot metaphor suggests two questions: (*i*) How does the visual system put together elements within each snapshot to form objects? This is the spatial grouping problem. (*ii*) When the snapshots are different, how does the visual system know which element in one snapshot corresponds to which element in the next? This is the temporal grouping problem. The snapshot metaphor is a caricature of the dominant model in the field—the sequential model—according to which spatial and temporal grouping are independent. The model we propose here is an interactive model, according to which the two grouping mechanisms are not separable. Currently, the experiments that support the interactive model are not conclusive because they use stimuli that are excessively specialized. To overcome this weakness, we created a new type of stimulus—spatiotemporal dot lattices—which allow us to independently manipulate the strength of spatial and temporal groupings. For these stimuli, sequential models make one fundamental assumption: if the spatial configuration of the stimulus remains constant, the perception of spatial grouping cannot be affected by manipulations of the temporal configuration of the stimulus. Our data are inconsistent with this assumption.

Vision uses small receptors to sample optical information. Spatial grouping is the process by which samples are linked across space to form more complex visual entities, such as objects and surfaces. Temporal grouping is the process by which visual entities are linked over time. Spatial grouping and temporal grouping either are sequential or they interact. If the perception of dynamic scenes is the result of the successive application of these two kinds of grouping, we have a sequential model of motion perception. If the perception of dynamic scenes is the result of spatial and temporal grouping operating in parallel, we have an interactive model.

The Sequential Model. Let us consider two versions of the sequential model: one in which spatial grouping comes first, the other in which temporal grouping comes first. The perception of most dynamic stimuli can be explained by describing them as a succession of snapshots (1). For example, according to Ullman (2), vision first does grouping within each snapshot and then finds a mapping between these groupings across the snapshots. These groupings are often called matching units. In this view, spatial grouping alone determines the matching units that will undergo temporal grouping.

Some stimuli, however, are designed to prevent us from applying spatial grouping first. These are, for example, random-dot cinematograms (3). These are dynamic displays in which each frame contains a different random texture. If the frames are not correlated, one sees random dynamic "snow." If dots in a patch of the display are correlated across frames, the patch will segregate, and its shape will be visible. A sequential model can account for the perception of such displays by assuming that temporal grouping extracts coherently moving elements [a process known as grouping by common fate (4)], which then undergo spatial organization.

Some data that appear to imply an interactive model can actually be explained by a sequential model. Consider, for example, the Ternus display (5), which consists of two rapidly alternating frames—f_1 and f_2—in which dots can occupy four equally spaced collinear positions *pqrs* (Fig. 1). The dots in f_1 are at *pqr*; the dots in f_2 are at *qrs*. This display can give rise to two percepts: (*i*) Element motion (*e*-motion) is seen when the two dots in positions *q* and *r* appear immobile, while one dot appears to move between the positions *p* and *s*. (*ii*) Group motion (*g*-motion) is seen when three dots appear to move, as a group, back and forth between *pqr* and *qrs*. The longer the interstimulus interval (ISI; interframe interval in this context), the higher the likelihood of *g*-motion (6). This is called the ISI effect. It is tempting to view this phenomenon as evidence for an interactive model (7). To see why, assume that the shorter the ISI, the stronger the temporal grouping. When the ISI is sufficiently short, temporal grouping could be stronger than spatial grouping. Thus, rather than grouping with the dot at *r* in f_1, the dot at *q* in f_1 would group with the dot at *q* in f_2. The result is *e*-motion. As ISI grows, the strength of temporal grouping drops, and concurrent dots group within frames, resulting in *g*-motion. However, a sequential model can also account for the ISI effect. Suppose that longer ISIs have two effects: (*i*) they weaken temporal grouping, and (*ii*) they give spatial grouping more time to consolidate the organization of concurrent dots. According to (*ii*), the ISI effect is caused by spatial rather than temporal grouping and is consistent with a sequential model (8).

The Interactive Model. Only an interactive model can account for the perception of motion when neither spatial grouping alone nor successive spatial and temporal grouping operations could derive matching units. Unfortunately, the only persuasive evidence in favor of the interactive model comes from displays in which objects and surfaces (transparent or opaque) overlap, and their spatial relation changes dynamically. For example, in certain kinetic occlusion (9–11) displays, we see a hitherto visible part of the scene become occluded by an opaque surface. In such displays, there is no simple correspondence between successive frames, because one frame contains a different number of elements than the next.†

We are concerned that the evidence from such displays is not general enough to refute sequential models as a class, because such displays may trigger a specialized mechanism that processes kinetic occlusion in which spatial and temporal grouping interact. For example, kinetic occlusion offers two characteristic clues: (*i*) the accretion or deletion of texture, as a textured object emerges from or disappears behind an occluder (9), and (*ii*) the presence of "T-junctions" between the contours of an occluder and the contours of an object it occludes (14, 15). To address these concerns, we designed displays that are not likely to trigger specialized motion-perception mechanisms.

Abbreviations: ISI, interstimulus interval; element motion, *e*-motion; group motion, *g*-motion; log-odds, logarithm of odds.

*To whom reprint requests should be addressed. E-mail: sergei@virginia.edu or kubovy@virginia.edu.

†A similar problem occurs if the moving object or surface is transparent (12, 13): finding correspondence between successive frames is hampered because the appearance of the region that is seen through a transparent surface changes as it becomes uncovered.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

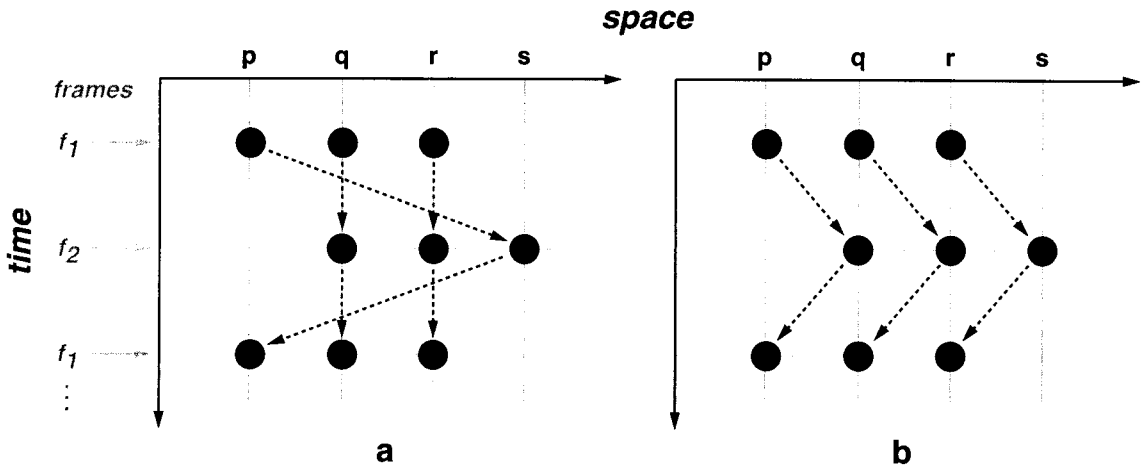

Fig. 1. Ternus display. The dotted arrows show the directions of perceived motion. (*a*) Element motion. (*b*) Group motion.

Motion Lattices. To refute the sequential models, we will show that spatial and temporal grouping mechanisms interact even when simple matching between the successive frames is possible. We created spatiotemporal dot lattices (motion lattices) in which we could independently vary the strength of spatial and temporal grouping by manipulating the spatial proximity of concurrent and successive dots [a generalization of the stimuli used by Burt and Sperling (16)]. We varied the strength of temporal grouping by manipulating the spatial proximity between successive dots. To avoid ambiguities of data interpretation that beset the Ternus display (i.e., the ISI effect), we held ISI constant. In our displays, as in the Ternus display, observers see either *e*-motion or *g*-motion. The advantage of motion lattices over the Ternus display is that in lattices the directions of *e*-motion and *g*-motion differ. The direction of *e*-motion is determined by matching individual dots in successive frames of the display. The direction of *g*-motion is determined by the matching of dot groupings in successive frames.

If we show an observer a single frame of a motion lattice, its spatial grouping is determined by the relative distance between concurrent dots (17). Although we hold the ISI constant, temporal grouping can be determined only by spatial distances between successive dots. According to the sequential model, the propensity of dots to group within frames—and thus yield *g*-motion—is independent of the determinants of temporal grouping. To test this prediction, we ask whether the frequency of *g*-motion changes when we hold constant the relative distance between concurrent dots and vary the spatial distance between successive dots. We find that it does, and on this basis we will claim that we have refuted the sequential model in favor of the interactive model.

General Methods

A motion lattice is a lattice of locations, whose rows we call baselines ($b_1, b_2, b_3, \ldots$; Fig. 2*a*), displayed in two frames, f_1 and f_2. In f_1, dots occupy the locations of the odd-numbered baselines; in f_2, dots occupy the locations of the even-numbered

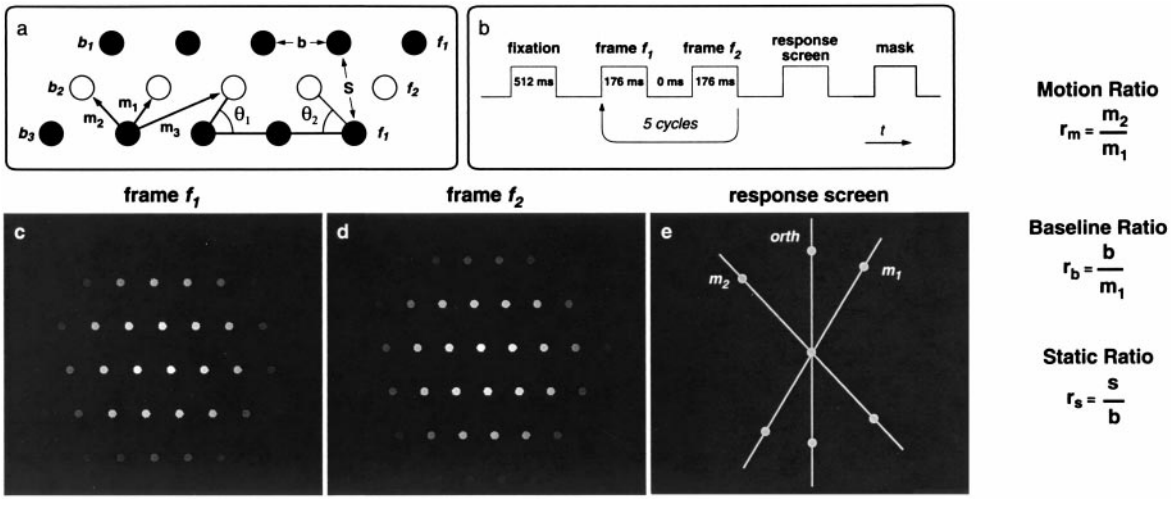

Fig. 2. (*a*) Three rows (baselines) of a motion lattice. The solid and open circles stand for dots that appear in frames f_1 and f_2, respectively. The three spatial parameters of motion lattices are: $|\mathbf{b}|$, the distance between adjacent dots in a baseline; $|\mathbf{m}_1|$, the shortest distance between successive dots; and θ_1, the acute angle between the orientations **b** and $\mathbf{m}_1$. The acute angle between **b** and $\mathbf{m}_2$ is θ_2. **S** is the second (after **b**) shortest distance within a frame. To minimize edge effects, we modulated the luminance of lattice dots [radius = 0.3 degrees of visual angle (dva)] according to a Gaussian distribution (σ = 1.5 dva). We held $\mathbf{m}_1$ (= 0.9 dva) constant. (*b*) Time line of each trial. We kept the total duration of each presentation at 1.76 s to prevent the percept of oscillatory motion, which sometimes is seen with longer presentations. (*c* and *d*) Two successive frames of a motion lattice captured from the computer screen (not to scale). (*e*) A response screen with three response options. (Labels were not presented in the display.)

baselines. When these frames rapidly alternate (f_1, f_2, f_1, ...) under appropriate spatial and temporal conditions, motion lattices are perceived as a continuous flow of apparent motion. Motion lattices are specified by two temporal and three spatial parameters. The two temporal parameters were kept constant: ISI (= 0) and frame duration (= 176 ms). The three spatial parameters are: $|\mathbf{b}|$, the distance between adjacent dots in a baseline; $|\mathbf{m}_1|$, the shortest distance between a dot in an f_1-frame and a dot in an f_2-frame; θ_1, the angle between the orientations of $\mathbf{b}$ and $\mathbf{m}_1$.

Two kinds of apparent motion can be seen in motion lattices:

- *e*-motion is based on element-to-element matching. In *e*-motion, each dot in a baseline is matched with a dot in an adjacent baseline. The three shortest distances between successive dots are $|\mathbf{m}_1| \leq |\mathbf{m}_2| \leq |\mathbf{m}_3|$ (Fig. 2*a*). The shorter the distance between dots, the more often the dots are linked in apparent motion (18, 19). Motion along $\mathbf{m}_1$ is seen more often than along $\mathbf{m}_2$ and never along $\mathbf{m}_3$. The strength of grouping between successive elements is called affinity (2), which is inversely related to interdot distance.
- *g*-motion is based on grouping-to-grouping matching. That is to say, in *g*-motion matching occurs between dot groupings (called virtual objects) as wholes. For example, in Figs. 2 *c–d*, suppose that the observer sees the lattice organized into virtual objects that coincide with the baselines, i.e., into horizontal strips of dots. If the observer sees vertical motion, it must be *g*-motion, because neither $\mathbf{m}_1$ nor $\mathbf{m}_2$ is vertical. In other cases, the virtual objects may not coincide with the baselines, but *g*-motion is always orthogonal to the virtual objects.[‡]

The sequential model predicts (for our stimuli) that the direction of motion will be determined solely by spatial grouping. To test this prediction, we measure the relative frequency of *e*-motion and *g*-motion, where we hold the determinants of spatial grouping constant and vary the spatial determinants of temporal grouping.

Spatial Grouping. The determinants of spatial grouping in static lattices are known: the dots group by proximity alone independent of lattice configuration. The strength of this grouping is called attraction. Attraction decreases exponentially, as the relative distance between dots increases (17). Analogously, within each frame of a motion lattice, the likelihood of different spatial groupings is controlled by the ratio of the two shortest interdot distances, $r_s = (|\mathbf{s}|/|\mathbf{b}|)$ (Fig. 2*a*). ($|\mathbf{S}|$ is equal to the shortest of the two distances: $\sqrt{4|\mathbf{m}_1|^2 + |\mathbf{b}|^2 \pm 4|\mathbf{m}_1||\mathbf{b}| \cos \theta_1}$.) We call r_s the static ratio. If we hold this ratio constant, the likelihood of different virtual objects within a frame will remain constant.

Temporal Grouping. In motion lattices temporal grouping is the outcome of two competitions:

- *e*-motion vs. *g*-motion: Because we are holding the temporal parameters constant, the outcome of this competition is controlled by the baseline ratio, $r_b = (|\mathbf{b}|/|\mathbf{m}_1|)$. For example, if r_b is low, then attraction within the baselines is high; therefore, observers are likely to see *g*-motion orthogonal to the baseline. If r_b is higher, then attraction within the baselines is lower; therefore, observers are less likely to see *g*-motion orthogonal to the baseline and—depending on the parameters of the motion lattice—they will see either *e*-motion or *g*-motion orthogonal to other virtual objects.
- between alternative *e*-motions: The outcome of this competition is controlled by the motion ratio, $r_m = (|\mathbf{m}_2|/|\mathbf{m}_1|)$. If r_m is high, then temporal grouping favors m_1 motion over m_2 motion.

The time line of a trial is shown in Fig. 2*b*. After fixating a central dot, observers viewed the motion sequence. In the trials with three response options, the response screen displayed three radial lines (Fig. 2*e*): two were parallel to the most probable orientations of *e*-motion (m_1 and m_2), and one was orthogonal to the baselines, *orth*. In the trials with two response options, the response screen displayed two radial lines, parallel to the orientations of m_1 and m_2. Observers reported the orientation of the motion by clicking on one of the circles attached to the lines on the response screen. The mask was an array of randomly moving dots. We randomized the orientation of the lattice between trials to minimize carryover from trial to trial.

Experiment 1

In Experiment 1, which is a control experiment, we show that changes in the frequency of *g*-motion are not attributable to observers' confusion between the response options. This possibility arises because as the orientations of $\mathbf{m}_1$ and $\mathbf{m}_2$ vectors approach the orthogonal to the baseline, confusion between *orth* and affinity responses (m_1, m_2) might increase. Response confusion might mimic an effect of interaction between the two types of grouping.

Methods. We held θ_1 (= 60°) constant and chose six motion lattices, each of which is defined by a (r_m, r_b) pair: (1.00, 1.00), (1.07, 1.13), (1.15, 1.24), (1.23, 1.35), (1.30, 1.44), (1.38, 1.54). Because of the geometry of motion lattices, when one holds θ_1 constant, r_m and r_b covary: $r_m = r_b\ (\sqrt{|\mathbf{m}_1|^2 + |\mathbf{b}|^2 - 2|\mathbf{m}_1||\mathbf{b}|\cos \theta_1}/|\mathbf{b}|)$.

Our six observers had normal or corrected-to-normal vision and were naive as to the purpose of the experiment. Each observer went through 100 trials per motion lattice (i.e., 600 trials) for each of two 1-hour sessions—one with two (m_1 and m_2) and the other with three (m_1, m_2, and *orth*) response options in the response screen.

Results and Discussion. The pie charts in Fig. 3*a* represent the relative proportion of m_1, m_2, and *orth* responses as a function of r_m, r_b. Because θ_2 varied, we represent this variation by changing the diameters of the pie charts in proportion to e^{r_θ}, where $r_\theta = [(\tan \theta_2)/(\tan \theta_1)]$. We split this three-way variation into two independent parts:

- Affinity function: In Fig. 3*b*, we plot the logarithm of odds (log-odds) with which observers chose *e*-motion along m_2 over m_1 for the two sessions. First, we note that because log-odds(m_2, m_1)[§] decrease linearly as a function of $r_m = (|\mathbf{m}_2|/|\mathbf{m}_1|)$, the odds $[p(m_2)]/[p(m_1)]$ decrease exponentially as a function of r_m. Further, we note that the confusion between the *orth* responses and the affinity responses (m_1, m_2) had little effect on response frequencies in the three-response sessions. The affinity function with three response options (thick line, "∘" symbols) differed little from the affinity function with two response options (thin line, "+" symbols), compared to the effect of $\mathbf{r}_\mathrm{m}$.
- Objecthood function: In Fig. 3*c*, we plot the log-odds with which observers chose *g*-motion over *e*-motion for the three-response session. This function reflects the tradeoff between group-to-group and element-to-element matching. Because

[‡]Orthogonal motion in motion lattices is an outcome of the so-called aperture problem (20–22). The aperture problem arises, for example, when a bar is moved behind an aperture so that its terminators are hidden: only the motion orthogonal to the bar is visible, even if the true motion of the bar is different. However, when the bar has a gap (or some other conspicuous feature), the visual system's solution is surprising: the gap appears to slide along the bar and does not disambiguate the direction of motion (20, 21, 23). In motion lattices, when dot grouping within the virtual objects is strong, the dots appear to move along the objects, like beads on a string.

[§](log-odds (m_2, m_1) = $\ln[p(m_2)]/[p(m_1)]$ = s_m (r_m − 1), where s_m is the slope.

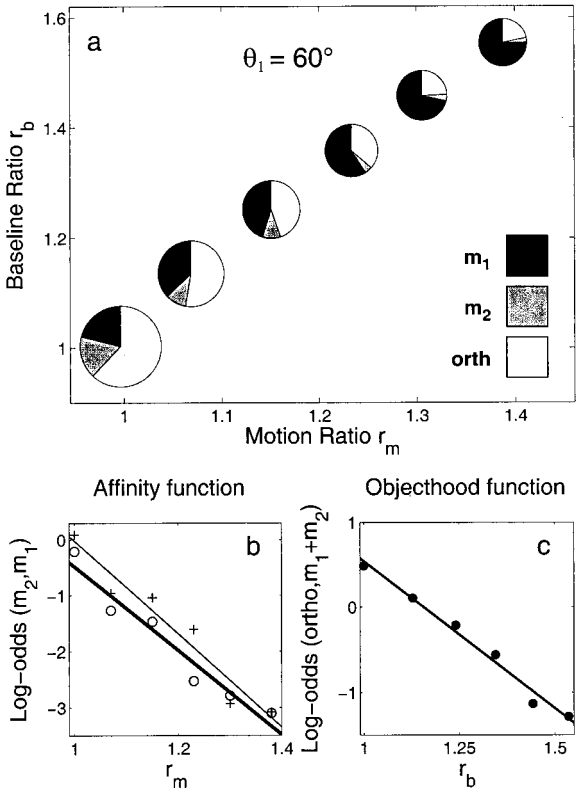

Fig. 3. Results of Experiment 1. (*a*) The pie charts (averaged over observers) show how the distribution of the responses (m_1, m_2, and *orth*) varied as a function of the motion ratio, r_m, and the baseline ratio, r_b, in the three-response sessions. These data are decomposed in (*b* and *c*). (*b*) Affinity function: the tradeoff between the probabilities of two alternative *e*-motions. Two affinity functions were obtained in different sessions: with three response options (the thick line through "○" symbols) and with two response options (the thin line through "+" symbols). (*c*) Objecthood function: the tradeoff between the probability of *e*-motion (m_1 or m_2) and the probability of *g*-motion (*orth*).

log-odds ($orth$, $m_1 \vee m_2$)[¶] decrease linearly as a function of $r_b = (|b|/|\mathbf{m}_1|)$, the odds $p(orth)/[p(m_1) + p(m_2)]$ decrease exponentially as a function of r_b.

As in previous studies of perceptual grouping in space–time, these data cannot adjudicate between the sequential model and the interactive model. The sequential model implies that the frequency of *g*-motion will decrease as a function of r_b because when baseline ratios are high, the tendency of concurrent dots to form virtual objects is low. The interactive model implies that the frequency of *g*-motion will decrease as a function of r_m because larger motion ratios strengthen temporal grouping and thus favor element-to-element matching at the expense of group-to-group matching.

Experiment 2

Methods. In Experiment 2, we crossed five values of r_m (1.0, 1.1, 1.2, 1.3, and 1.4) with four values of r_b (1.11, 1.43, 1.25, and 1.67) to create 20 types of motion lattices. This was possible because we allowed θ_1 to vary. Our seven observers had normal or corrected-to-normal vision and were naive as to the purpose of the experiment. On each trial, they were offered three response options—m_1, m_2, and *orth*. Every observer went through 40 trials per motion lattice, i.e., 800 trials in one session. In all other respects, the experiment was identical to Experiment 1.

¶(log-odds ($orth$, $m_1 \vee m_2$) = $\ln[p(orth)]/[p(m_1) + p(m_2)] = s_b r_b + k$), where s_b is the slope.

Results and Discussion. The distribution of the three responses varied systematically as a function of r_m, r_b, and r_θ (Fig. 4*e*). In Fig. 4*g*, we plot four linear objecthood functions to show the tradeoff between *g*-motion and *e*-motion as a function of both r_m and r_b. The statistical model used to fit these functions accounts for 98% of variance in the data. (We used this model to interpolate the response frequencies for Fig. 4*h*). The frames in Fig. 4 *a–d* illustrate different outcomes of spatial grouping within the frames. When spatial grouping favors the formation of salient virtual objects orthogonal to the baselines (Fig. 4*a*: low r_m and high r_b), *g*-motion orthogonal to the baselines is never seen, and *e*-motion dominates (Fig. 4*g*). As r_b decreases within the same r_m (Fig. 4, *a* and *c*), spatial grouping progressively favors the formation of objects within the baselines, and the frequency of *g*-motion grows relative to the frequency of *e*-motion. As r_m increases within the high r_bs (Fig. 4, *a* and *b*), the virtual objects not parallel to the baselines become less orthogonal to the baselines and less salient and thus allow a higher frequency of *g*-motion. Within the small r_bs, the growing r_m does not cause appreciable change in the high salience of virtual objects (Fig. 4, *c* and *d*), and the frequency of *g*-motion does not change.

The gray curves in the background of Fig. 4*e* are iso-r_s lines. Each of these curves represents the set of lattice configurations for which spatial grouping favors the formation of identical virtual objects. As Fig. 4*e* shows, the frequency of *g*-motion changes along iso-r_s lines. To make this observation explicit, we interpolated the empirical frequencies of *g*-motion and plotted them within the iso-r_s sets in Fig. 4*h*: as r_b increases, the frequency of *g*-motion within the iso-r_s sets drops rapidly. As temporal grouping progressively becomes stronger than spatial grouping, observers tend to see *g*-motion less frequently and *e*-motion more frequently. (Note that as one moves within each iso-r_s set from high to low *g*-motion frequencies, r_m grows, and therefore one of the *e*-motions, m_1 motion, becomes increasingly salient.) The interactive model can explain this result, for example by pitting against each other the two scales of spatial grouping (elements vs. element aggregates) *after* the temporal grouping operation (24): *g*-motion wins the competition with *e*-motion when the latter becomes ambiguous at low r_m, so that the dot aggregates (virtual objects), and not the individual dots, become the moving entities. (We thank one of the anonymous reviewers for suggesting we emphasize this point.)

We conclude that the sequential model does not hold: Invariant conditions for spatial grouping contribute to the perception of apparent motion differently, depending on the conditions for temporal grouping.

General Discussion

Using spatiotemporal dot lattices, we varied the spatial distances between concurrent dots and spatiotemporal distances between successive dots. Each dot could be grouped either (*i*) with a successive dot to generate *e*-motion or (*ii*) with concurrent dots to form (virtual) objects that are matched to generate *g*-motion. According to the sequential model, the only factor that can determine which of these is seen is the attraction between concurrent dots. We held attraction between concurrent dots constant and found that affinity between successive dots determines whether *e*-motion or *g*-motion will be seen. Thus we have refuted the sequential model in favor of the interactive model. Our findings imply that matching units can arise at any level in the cascade of visual processes, as late as the level of complex objects (25), in contrast to the view that matching units were derived early in visual perception (2).

Current theories of motion perception distinguish between three systems that compute motion. The systems differ by the

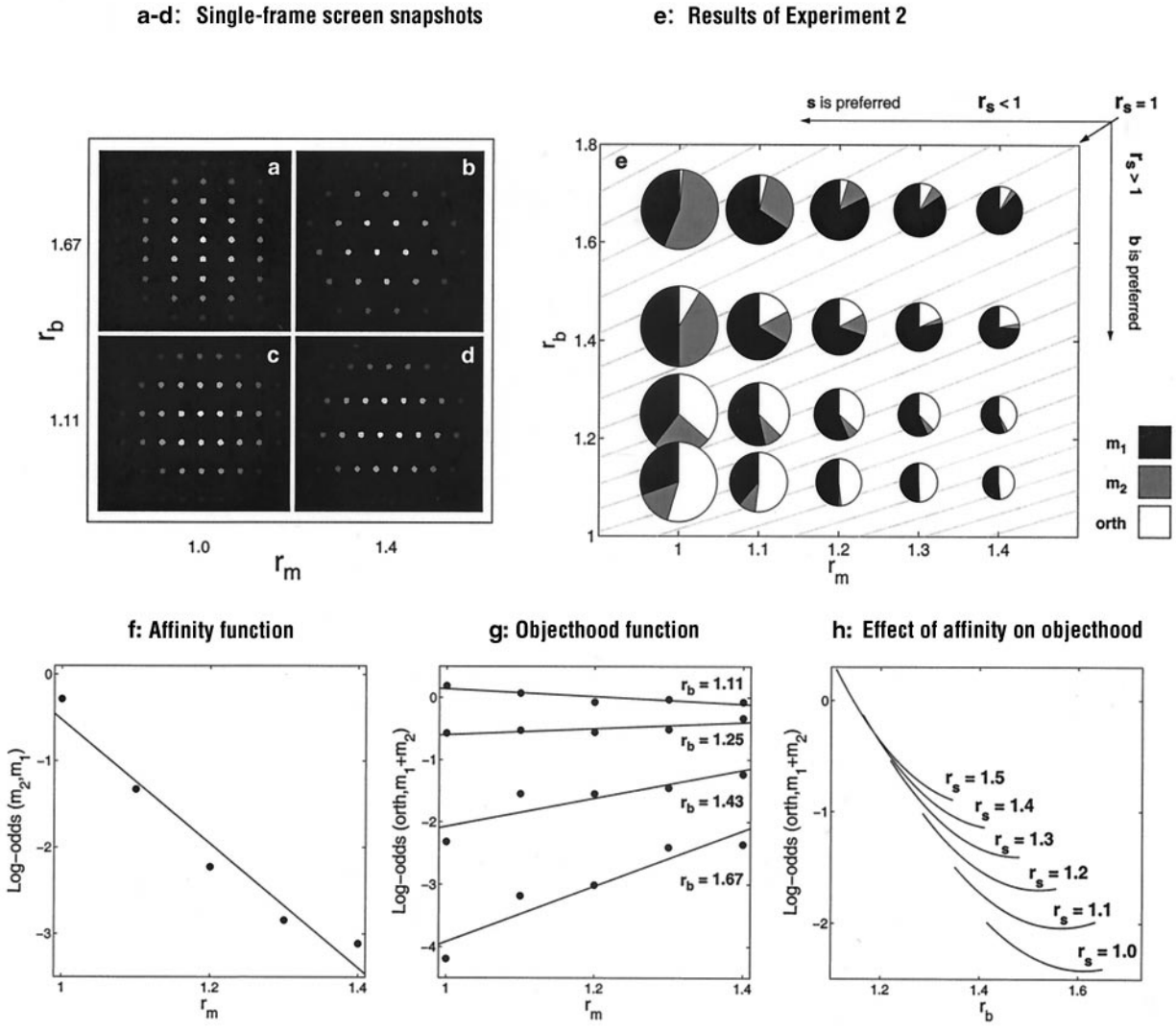

Fig. 4. (*a–d*) Single frames captured from the computer screen with the extreme values of $r_m = (|\mathbf{m}_2|/|\mathbf{m}_1|)$ and $r_b = (|\mathbf{b}|)/|\mathbf{m}_1|)$. The snapshots are arranged in the (r_m, r_b) space, parallel to the plot in e. (e) The pie charts show the distribution of the three responses in the 20 motion lattices. The gray lines on the background are the iso-r_s lines, where $r_s = (|\mathbf{s}|/|\mathbf{b}|)$, which are the contours for which within-frame spatial grouping should remain constant. For the isoline $r_s = 1.0$, the organizations along **s** and **b** are equiprobable. It is marked with an oblique arrow (*Upper Right*). Conditions that favor dot grouping within the baselines ($r_s > 1.0$; e.g., *c* and *d*) lie to the right of the isoline of $r_s = 1.0$; in the rest of the conditions ($r_s < 1.0$), dots tend to form groupings not along the baselines (as in *a*). (*f*) Affinity function collapsed across the r_b conditions. (*g*) Four objecthood functions summarize the effects of the baseline and motion ratios, r_b and r_m. The frequency of *g*-motion grows rapidly as r_b drops and groupings along the baselines become more prominent. This effect is evident both when r_b is low (high ambiguity of *e*-motion) and when r_b is high (m_1 wins the competition with m_2). The plot in *h* explicates the effect of r_s. Dot organizations within the baselines dissolve as r_s grows, which reduces the frequency of *g*-motion. In contrast to the prediction of the sequential model, *e*-motion and *g*-motion tradeoff within the iso-r_s sets.

complexity of the spatial representations on which they are based (26–29). The first-order system can detect the temporal modulation of raw luminances but is insensitive to spatial configuration. In our stimuli, this system can detect *e*-motion, because the displacement of each dot is a temporal modulation of luminance, but it cannot detect *g*-motion, because the direction of *g*-motion does not correspond to the direction of motion of any dot. The perception of *g*-motion requires a system that can take advantage of the spatial organization of the stimulus. Thus, *g*-motion is detected either by a second-order system that matches spatial features or by a third-order system that derives motion of more complex visual constructs.

How does vision decide which of the spatial representations will determine what is moving? According to the sequential model, the alternative spatial representations compete before motion matching, whereas according to the interactive model these representations compete after motion matching, i.e., among the outputs of the alternative motion systems: first-, second-, or third-order (30).

Most theories of motion perception are versions of the sequential model (30). A notable theory that agrees with the interactive model has been proposed by Wilson, Ferrera, and Yo (24). According to this theory, matching is applied in parallel to the raw visual input (the first-order system) and to the output of a preprocessor (the second-order system). When the outputs of the two motion systems support different motion directions, they compete, and the winner takes all. Wilson *et al.*'s model is interactive because the competition between different spatial representations occurs after motion computation. As it happens, the evidence that inspired their model is actually consistent with

a sequential model.[‖] Our data, on the other hand, do support the model of Wilson *et al.* (24).

To summarize, the interactive model holds that spatial organization and motion matching are tightly integrated. The identity of moving visual entities is determined both by spatial proximities between elements at each moment and by spatial proximities between elements that occur at successive moments. Thus visual objects emerge when motion matching between element aggregates (Gestalts) is stronger than motion matching between elements.

[‖]The two sources of evidence for the model of Wilson *et al.* (24) are: (*i*) Physiological evidence that the cortical areas responsible for motion perception receive both first- and second-order spatial information (31). This evidence is consistent with the sequential model because the two sources of information could compete before motion computation. (*ii*) Psychophysical evidence that two superimposed moving gratings may be perceived as either independently moving gratings or as a plaid moving in an orientation different from the motion of the individual gratings (20, 22). This evidence is also consistent with the sequential model because the gratings can group and form a plaid before motion computation (27).

We thank D. R. Proffitt, M. Shiffrar, J. Wagemans, and S. Yantis for valuable discussions; W. Epstein, J. Hochberg, C. Von Hofsten, and three anonymous reviewers for helpful comments on an earlier version of the manuscript; and D. M. Johnson and S. C. Haden for assistance in running the experiments. This work was supported by National Eye Institute Grant 9 R01 EY12926-06.

1. Neisser, U. (1967) *Cognitive Psychology* (Appleton Century Crofts, New York).
2. Ullman, S. (1979) *The Interpretation of Visual Motion* (MIT Press, Cambridge, MA).
3. Julesz, B. (1971) *Foundations of Cyclopean Perception* (Univ. of Chicago Press, Chicago, IL).
4. Wertheimer, M. (1936; originally published in 1923) in *A Source of Gestalt Psychology*, ed. Ellis, W. D. (Routledge & Kegan Paul, London), pp. 71–88.
5. Ternus, J. (1936; originally published in 1926) in *A Source of Gestalt Psychology*, ed. Ellis, W. D. (Routledge & Kegan Paul, London), pp. 149–160.
6. Pantle, A. J. & Picciano, L. (1976) *Science* **193,** 500–502.
7. Kramer, P. & Yantis, S. (1997) *Percept. Psychophys.* **59,** 87–99.
8. He, Z. J. & Ooi, T. L. (1999) *Perception* **28,** 877–892.
9. Kaplan, G. A. (1969) *Percept. Psychophys.* **6,** 193–198.
10. Kellman, P. J. & Cohen, M. H. (1984) *Percept. Psychophys.* **35,** 237–244.
11. Tse, P., Cavanagh, P. & Nakayama, K. (1998) *High-Level Motion Processing*, ed. Watanabe, T. (MIT Press, Cambridge, MA), pp. 249–266.
12. Shipley, T. F. & Kellman, P. J. (1993) *Spatial Vision* **7,** 323–339.
13. Cicerone, C. M., Hoffman, D. D., Gowdy, P. D. & Kim, J. S. (1995) *Percept. Psychophys.* **57,** 761–777.
14. Guzman, A. (1968) in *Automatic Interpretation and Classification of Images*, ed. Griselli, A. (Academic, New York), pp. 243–276.
15. Nakayama, K., He, Z. J. & Shimojo, S. (1995) *Visual Cognition*, eds. Kosslyn, S. M. & Osherson, N. (MIT Press, Cambridge, MA), pp. 1–70.
16. Burt, P. & Sperling, G. (1981) *Psychol. Rev.* **88,** 171–195.
17. Kubovy, M., Holcombe, A. O. & Wagemans, J. (1998) *Cognit. Psychol.* **35,** 71–98.
18. von Schiller, P. (1933) *Psychologische Forschung* **17,** 179–214.
19. Ramachandran, V. S. & Anstis, S. M. (1983) *Nature (London)* **304,** 529–531.
20. Wallach, H. (1935) *Psychologische Forschung* **20,** 325–380.
21. Wuerger, S., Shapley, R. & Rubin, N. (1996) *Perception* **25,** 1317–1367.
22. Adelson, E. H. & Movshon, J. A. (1982) *Nature (London)* **300,** 523–525.
23. Castet, E. & Wueger, S. (1997) *Vision Res.* **37,** 705–720.
24. Wilson, H. R., Ferrera, V. P. & Yo, C. (1992) *Visual Neurosci.* **9,** 79–97.
25. Ramachandran, V. S., Armel, C. & Foster, C. (1998) *Nature (London)* **395,** 852–853.
26. Cavanagh, P. & Mather, G. (1990) *Spatial Vision* **4,** 103–129.
27. Smith, A. T. (1994) in *Visual Detection of Motion*, eds. Smith, A. T. & Snowden, R. J. (Academic, New York), pp. 145–176.
28. Lu, Z.-L. & Sperling, G. (1995) *Vision-Res.* **35,** 2697–2722.
29. Wilson, H. R. & Wilkinson, F. (1997) *Perception* **26,** 939–960.
30. Kubovy, M. & Gepshtein, S. (2000) in *Perceptual Organization for Artificial Vision Systems*, eds. Boyer, K. & Sarkar, S. (Kluwer, Dordrecht, The Netherlands), pp. 41–71.
31. Maunsell, J. H. R. & Newsome, W. T. (1987) *Annu. Rev. Neurosci.* **10,** 363–401.