Temporal synchrony in perceptual grouping: a critique

Hany Farid
Department of Computer Science and Center for Cognitive Neuroscience
Dartmouth College
Hanover, NH 03755

It has been hypothesized that the human visual system can use temporal synchrony for the perceptual grouping of image regions into unified objects, as proposed in some neural models. It is argued here, however, that previous psychophysical evidence for this hypothesis is due to stimulus artifacts, and that earlier studies do not, therefore, support the claims of synchrony sensitive grouping mechanisms or processes.

The sequence of dots ● ●● ●● ●● ●● ● is typically perceived as a series of loosely spaced dot pairs ●●. Why don't we perceive this sequence as a series of the equally plausible tightly spaced dot pairs ● ●? Because elements in close proximity are more likely to be perceived as belonging together. In the sequence, ● ○● ○● ○● ○● ○, however, proximity and color compete against one another, so that we might perceive a series of dot pairs ○●, or two interlaced sequences of ●'s and ○'s. Proximity and color are just two possible cues that play a role in perceptual grouping [1], that is, the ability of the visual system to organize a multitude of parts into a unified entity or object.

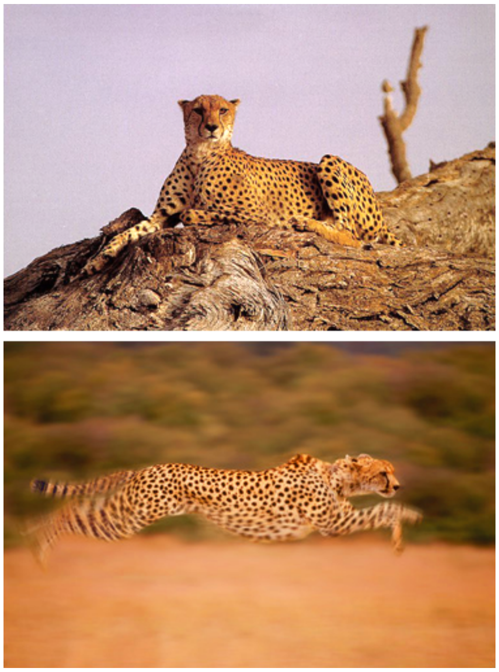


Figure 1: Static and dynamic cues help the human visual system organize complex visual scenes into coherent objects.

In complex visual scenes similar cues play a role in the perceptual grouping of objects. For example the mostly uniform color and texture of the cheetah in Figure 1 helps us in organizing its various parts into a single object. In addition to such static cues, dynamic cues also play a role in perceptual grouping. For example, in the context of a stationary background, the largely uniform translational (common-fate) motion of the running cheetah's body and head in Figure 1 help us in grouping its various parts. These and similar cues probably cannot explain all of perceptual grouping, but it is generally believed that our visual system uses these cues, most likely in concert with other cues.

Consider now the perceptual grouping of a flock of birds. All the birds typically fly at the same speed and direction and may suddenly change directions. Why is the flock seen as a single entity? The common-fate motion certainly is a grouping cue, and recently it has been suggested that simultaneous motion changes may also be a grouping

in: Trends in Cognitive Sciences 6(7):284-288, 2002

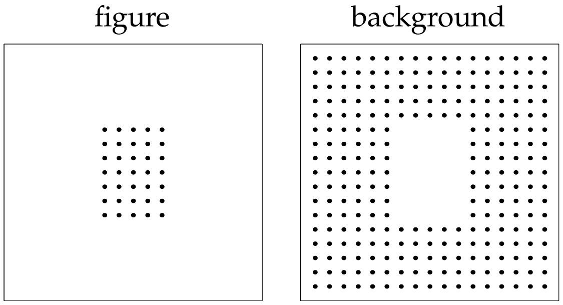


Figure 2: A periodic synchrony stimulus. Figure and background frames are shown in rapid alternation.

cue. That is, all the birds changing directions at precisely the same time is presumably not accidental, but rather suggests that the birds belong to a single entity. This theory, referred to as grouping by temporal synchrony, posits that the human visual system can measure and correlate fine-grained changes in motion across the visual field. This theory is particularly intriguing given its link to the claims of neural binding based on temporal synchrony (e.g., [2]). The hotly debated topic of synchrony and neural binding will not be addressed here (see, for example, [3] and [4] for opposing views). It is argued here, however, that previous psychophysical evidence for this hypothesis is due to stimulus artifacts, and that earlier studies do not support the claims of synchrony sensitive grouping mechanisms or processes.

Periodic synchrony stimuli

One of the earliest suggestions for the theory of grouping by temporal synchrony is based on a simple temporally periodic stimulus. In its simplest form, this stimulus consists of two frames: a figure and background, Figure 2. The frames are displayed in rapid alternation with each frame displayed for 7-15 ms. Subjects are readily able to judge the shape of the figure. Since the figure and background elements are identical except with respect to their temporal presentations, it is concluded that the human visual system can segregate regions based solely on temporal information on the scale of 10 ms. A number of independent studies are consistent with these findings [5, 6, 7, 8]. Additional studies have also found increased performance in the presence of both spatial and temporal grouping cues [9]. At least two other studies, however, have found that when both spatial and temporal synchrony cues are present performance is no better than with the spatial cue alone [10, 11].

A static form confound in periodic synchrony stimuli

A potentially troubling aspect of the temporally periodic stimulus is that the figure frame contains, in isolation, an obvious static form cue. The concern then is that an even brief exposure to the figure frame might provide a simple form cue. Beaudot suggested that it is such a static cue that is responsible for grouping percepts in at least some of these stimuli [12]. Using the stimulus of [7], Beaudot showed that when the figure frame is presented first in the temporal sequence, subjects are better able to distinguish the location of the figure than when the background frame is presented first. This ordering effect suggests that a simple priming cue is at least partially responsible for the grouping percept.

A clever variation on this periodic stimulus was devised to remove this potential static form cue [13]. This stimulus consists of a sea of colons (:) with random orientations. Each colon flips by 90 degrees about its midpoint. The colons forming the background and rectangular figure regions flip orientations on alternate frames. In this way a form cue no longer exists on any single static frame. The authors found that this stimulus promotes grouping only with temporal delays on the order of 25-30 ms. Since there is no form cue on any single frame, at least two frames are necessary for any percept to be possible, thus yielding a minimal stimulus period of 50-60 ms. This is nearly five times longer than the 7-15 ms times claimed for the two frame periodic stimuli. It is noteworthy that a temporal integration time of 50-60 ms is consistent with the well established visual processing mechanisms of early visual processing (see below), and hence synchrony-based mechanisms need not be invoked.

Combined, these studies cast into doubt the claims for the existence of temporally sensitive mecha-

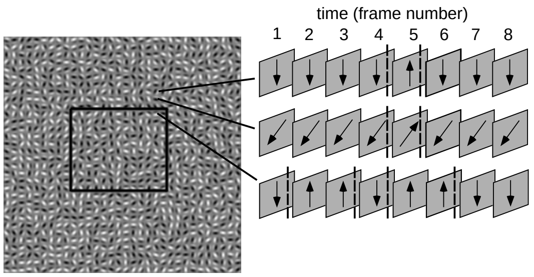


Figure 3: A stochastic synchrony stimulus. The motion reversals of all elements in the background and central figure region are synchronized to different random processes. The dashed vertical lines mark motion reversals.

nisms and processes on the scale of 10 ms.

Stochastic synchrony stimuli

In another attempt to overcome some of the potential static cues in the two frame periodic stimuli, Lee and Blake constructed a stochastic texture stimulus composed of randomly oriented Gabor elements [14]. On each frame the phase of each Gabor shifts forwards or backwards according to a random process, Figure 3. One random process is used for all the Gabors in a central figure region and a different process for all the Gabors in the background region. In this way, no form cue exists on any single or pair of frames, and the authors claimed that the only remaining form cue is defined solely by temporal synchrony. Subjects are readily able to judge the shape of the central figure. In the purported absence of classic static or dynamic grouping cues, the authors concluded that the human visual system can segregate regions based solely on temporal information on the scale of 10 ms (see also [15, 9]).

A spatiotemporal contrast confound in stochastic synchrony stimuli

The stimulus of Lee and Blake provided perhaps the most compelling evidence for synchrony sensitive mechanisms and processes. We have previously argued, however, that their stimulus contains an unintended artifact [16]. Due to the stochastic nature of the reversal sequences, there are moments when one region rapidly alternates between forward and backward shifts (thus undergoing little overall motion), while the other region has a run of all forward or all backward shifts (thus undergoing a large amount of overall motion). These coarse motion differences can occur over a relatively long time scale, on the order of 10 frames (70-100 ms). This is long enough to be seen by the well established mechanisms of early visual processing. Specifically, we showed that a physiologically plausible temporal bandpass filter [1] (with a temporal integration window on the order of 70-100 ms [17, 18]) can convert the relatively large-scale temporal change differences into a classic spatiotemporal contrast cue (Box 1). This led us to conclude that this unintended cue, not a finer temporal synchrony cue, is most likely responsible for the grouping percepts in this stimulus. We also showed that a temporal lowpass filter [2] reveals a contrast cue. Lee and Blake showed, however, that a slightly modified stimulus that removes this cue from a lowpass filter still promotes grouping [19]. These perturbations do not, however, remove the cue from a temporal bandpass filter, hence the contrast cue is still available to the visual system.

While at first glance they may not appear so, the confounds present in this stochastic stimulus are similar to those in the periodic stimuli. The two frame periodic stimulus, Figure 2, contains a cue on any single frame. The subsequent multi-frame periodic stimulus of [13] contains a cue on any pair of frames, and the stochastic stimulus contains a cue on ten (or so) frames. In all cases, these stimuli contain an unintended grouping cue, in addition to a temporal synchrony cue, that confounds each experiment.

Synchrony with no spatiotemporal contrast confound

[1] The bandpass impulse response is: $h(t) = (kt/\tau)^n e^{-kt/\tau}[1/n! - (kt/\tau)^2/(n+2)!]$, with $\tau = 0.01$, $k = 2$ and $n = 4$, and $t \in [0, 10]$ frames. The temporal filter alone is sufficient to reveal the contrast cue, and as such no spatial filtering is necessary.

[2] The lowpass impulse response is: $h(t) = (t/\tau)^2 e^{-t/\tau}$, with $\tau = 0.01$ and $t \in [0, 10]$ frames.

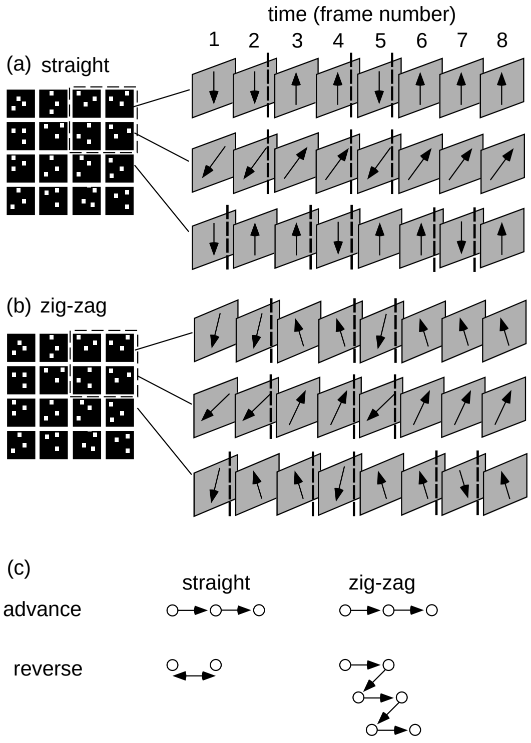


Figure 4: Temporal properties of the dot stimulus. **(a)** A schematic of a portion of the dynamic dot stimulus. The motion reversals of all the dots in the background and within the dashed rectangle are synchronized to different random processes. The vertical dashed lines mark reversal points. **(b)** In the zig-zag condition, the synchronous motion reversals are preserved while slightly altering the reversal direction. In so doing, this stimulus no longer contains a classic temporal contrast cue. **(c)** A schematic of a dots advance/reverse path.

To further study the relative importance of synchrony and spatiotemporal contrast we devised a stochastic motion stimulus that allows for the independent control of the synchrony and spatiotemporal contrast cues [20]. The basic stimulus consists of an array of small windows each containing dots drifting with a constant speed and direction. Across windows, the speed is constant, but the direction is randomized. On each frame the dots move randomly forward or backward along their specified direction, Figure 4(a). As with the original Gabor stimulus, a form cue defined by temporal synchrony is introduced: the motion reversals of all the dots in the central figure and background regions are synchronized to different random processes. As with the original Gabor stimulus, this dot stimulus contains a confounding spatiotemporal contrast cue. This cue can be eliminated by simply changing the reversal angle, so that reversing dots no longer fall back onto themselves, Figure 4(b). With respect to spatiotemporal contrast, the overall motion of a region rapidly reversing directions is now largely the same as an area repeatedly shifting along the same direction. At the same time the temporal synchrony cue that purportedly gives rise to the perception of form is preserved, Figure 4(c). Under these conditions subjects are unable to reliably perceive or judge the shape of the figure region. Consistent with these findings, Morgan and Castet have also found that when the basic Gabor stimulus is manipulated to reduce the unintended spatiotemporal contrast cue, subjects are unable to reliably judge the shape of the figure region [21].

Combined, these studies suggest that the coarse motion differences (on a scale of 100 ms), not the finer temporal synchrony cue (on a scale of 10 ms), are responsible for the perception of form in the stochastic synchrony stimuli. Standard mechanisms and processes known to exist in early visual processing are sufficient to explain these results, while a synchrony-based explanation is both unnecessary and insufficient.

Energy versus synchrony

More than 15 years ago Adelson and Bergen presented a simple and elegant spatiotemporal energy

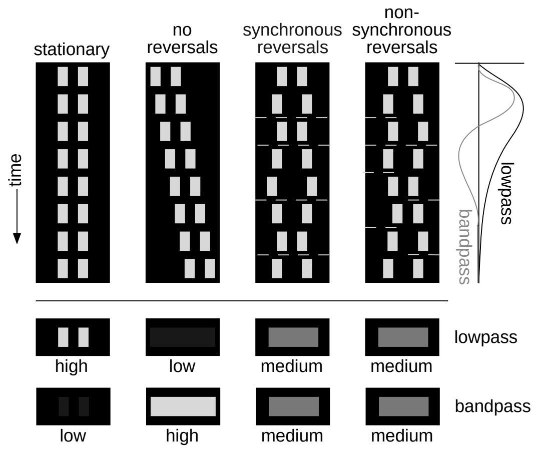


Figure 5: Shown are, from left to right, two bright bars on a black background undergoing no motion, continuous right-ward motion, repeated and synchronous reversals, and repeated but non-synchronous reversals (dashed lines denote motion reversals). Shown below are idealized outputs from temporal lowpass (average) and temporal bandpass (difference) filtering (see also Box 1). Overall motion differences (smooth vs. jittering motion) across several frames yields a classic spatiotemporal contrast cue. The synchronization of motion reversals is irrelevant.

model for the perception of motion [22] (see also [23]). The first stage of the model consists of a quadrature pair of linear filters oriented in space-time and tuned in spatial frequency. The second stage squares and sums the output of these filters to give a measure of motion energy. This model is consistent with a wide range of known physiology and psychophysics.

It is precisely classic motion energy that is the relevant cue in the stimuli described above - that is, a coarse motion difference between different image regions (i.e., motion versus little or no motion, on the scale of 100 ms). Shown in Figure 5, for example, is an illustration of how this cue might arise. Shown are four different motion sequences for two bright bars on a black background. In the first column, the bars do not move, while in the second column the bars move continuously to the right. Note how a temporal lowpass (average) or bandpass (difference) filter responds differently to these different motion sequences (Box 1). For example, the temporal lowpass yields a pair of high-contrast bars in the absence of motion, and a low contrast blurred-out region for the continuous motion. In the third column the bars repeatedly and synchronously reverse directions. The bars in the fourth column repeatedly but non-synchronously reverse directions. In both of these cases, the temporal lowpass filter yields a medium contrast region, whether the motion reversals are synchronous or not. It is simply the large-scale motion pattern over several frames, that of the bars jittering in place or repeatedly advancing that yields a spatiotemporal contrast cue. The synchronization is simply irrelevant. The ability to perceive such coarse motion differences is consistent with the well known and accepted spatiotemporal energy models, and is therefore neither surprising nor suggestive of novel mechanisms or processes based on temporal synchrony.

Conclusions

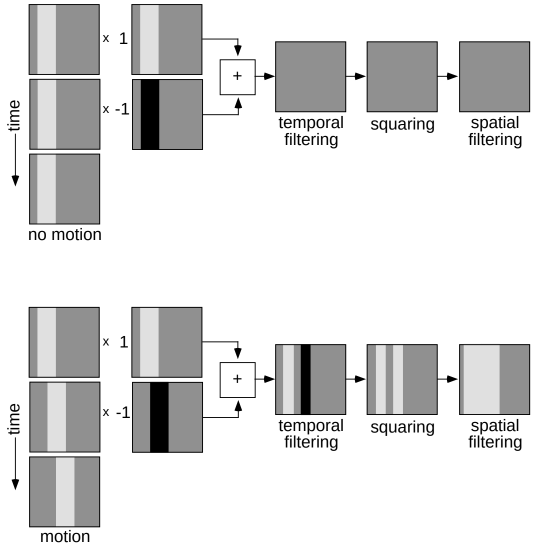
Can the human visual system use fine-grained temporal synchrony to bind image regions into unified objects? There is, as of yet, no evidence to suggest so. The psychophysical evidence to support this theory contains unintended classic grouping cues that are themselves responsible for any grouping percepts. These grouping cues are consistent with well established grouping mechanisms and processes. Furthermore, when these cues are removed, while preserving the temporal synchrony cue, the resulting stimuli no longer promote grouping. There is, therefore, no reason to posit the existence of novel synchrony-sensitive mechanisms or processes.

Acknowledgments

HF is supported by an Alfred P. Sloan Fellowship, a National Science Foundation CAREER Award (IIS-99-83806), and a departmental National Science Foundation Infrastructure Grant (EIA-98-02068).

Box 1: Spatiotemporal filtering: from motion to contrast

Shown below are three frames from a movie of a stationary bar (top) and a moving bar (bottom). Our visual system easily distinguishes between these two different temporal stimuli. By simply filtering in both space and time, spatiotemporal models are sufficient to explain the perception of motion (or lack thereof). These models convert motion into a temporal contrast cue in three basic stages: (1) the stimulus is temporally filtered by multiplying each frame by the corresponding filter value and then summing. In the figure below the filter is a simple difference (bandpass) filter with only two values, 1 and -1 (note how this is a crude approximation to the temporal bandpass filter of Figure 5); (2) the output of the temporal filtering is squared so as to be invariant to the sign of the filter response; and (3) a final spatial filtering is applied, typically with an averaging (lowpass) filter, as shown below. This filtering, of course, occurs throughout the entire temporal sequence by sliding the temporal filter down one frame at a time and repeating the entire calculation. A full blown spatiotemporal energy model [22] employs the same basic principles, differing only in the specific choice of temporal and spatial filters. As shown below, this simple model is sufficient to differentiate between a moving and stationary stimulus, without the need to explicitly track or correlate features across multiple frames. By converting coarse motion differences (e.g., motion vs. little or no motion) into a contrast difference across the visual field, the visual system is able to perceptually group regions undergoing different motion patterns.



References

[1] M. Wertheimer, "Untersuchungen zur lehre von der gestalt ii," *Psycologische Forschung* **4**, pp. 301–350, 1923. Translation published in Ellis, W. (1938). A source book of Gestalt psychology (pp. 71-88). London: Routledge & Kegan Paul.

[2] W. Singer and C. Gray, "Visual feature integration and the temporal correlation hypothesis," *Annu. Rev. Neurosci.* **18**, pp. 555–586, 1995.

[3] W. Singer, "Neuronal synchrony: a versatile code for the definition of relations," *Neuron* **24**, pp. 49–65, 1999.

[4] M. Shadlen and J. Movshon, "Synchrony unbound: a critical evaluation of the temporal binding hypothesis," *Neuron* **24**, pp. 67–77, 1999.

[5] D. Rogers-Ramachandran and V. Ramachandran, "Psychophysical evidence for boundary and surface systems in human vision," *Vis. Res.* **38**, pp. 71–77, 1991.

[6] M. Fahle, "Figure-ground discrimination from temporal information," *Proc. R. Soc. Lond, B* **254**, pp. 199–203, 1993.

[7] M. Usher and N. Donnelly, "Visual synchrony affects binding and segmentation in perception," *Nature* **394**, pp. 179–182, 1998.

[8] H. Kojima, "Figure-ground segregation from temporal delay is best at high spatial frequencies," *Vis. Res.* **38**, pp. 3729–3734, 1998.

[9] S. Lee and R. Blake, "Neural synergy in visual grouping: when good continuation meets common fate," *Vis. Res.* **41**, pp. 2057–2064, 2001.

[10] M. Fahle and C. Koch, "Spatial displacement, but not temporal asynchrony, destroys figural binding," *Vis. Res.* **35**, pp. 491–494, 1995.

[11] D.C. Kiper, et al., "Cortical oscillatory responses do not affect visual segmentation," *Vis. Res.* **36**, pp. 539–544, 1996.

[12] W. Beaudot, "Role of onset asynchrony in contour integration," *Vis. Res.* **42**(1), pp. 1–9, 2002.

[13] F. Kandil and M. Fahle, "Purely temporal figure-ground segregation," *Euro. J. of Neurosci.* **13**, pp. 2004–2008, 2001.

[14] S. Lee and R. Blake, "Visual form created solely from temporal structure," *Science* **284**, pp. 1165–1168, 1999.

[15] R. Blake and S. Lee, *Biologically Motivated Computer Vision*, ch. Temporal Structure in the Input to Vision Can Promote Spatial Grouping, pp. 635–653. Springer-Verlag, 2000.

[16] E. Adelson and H. Farid, "Filtering reveals form in temporally structured displays," *Science* **286**, p. 2231a, 1999.

[17] A. Watson, *Handbook of Perception and Human Performance*, ch. Temporal Sensitivity. John Wiley and Sons, 1986.

[18] W. Geisler and D. Albrecht, "Visual cortex neurons in monkeys and cats: detection, discrimination, and identification," *Vis. Neurosci.* **14**(5), pp. 897–919, 1997.

[19] S. Lee and R. Blake, "Reply to: Filtering reveals form in temporally structured displays," *Science* **286**, p. 2231a, 1999.

[20] H. Farid and E. Adelson, "Synchrony does not promote grouping in temporally structured displays," *Nature Neurosci.* **4**(9), pp. 875–876, 2001.

[21] M. Morgan and E. Castet, "High temporal frequency synchrony is insufficient for perceptual grouping," *Proc. Roy. Soc.* **269**, pp. 513–516, 2002.

[22] E. Adelson and J. Bergen, "Spatiotemporal energy models for the perception of motion," *J. Opt. Soc. Am. A* **2**(2), pp. 284–299, 1985.

[23] A. Watson and A. Ahumada, "Model of human visual-motion sensing," *J. Opt. Soc. Am. A* **2**(2), pp. 322–341, 1985.