Temporal Structure in the Input to Vision Can Promote Spatial Grouping

Randolph Blake and Sang-Hun Lee

Vanderbilt Vision Research Center,
Vanderbilt University, Nashville TN 37240, USA
randolph.blake@vanderbilt.edu

1 Introduction

Humpty Dumpty sat on a wall
Humpty Dumpty had a great fall
All the King's horses and all the King's men
Couldn't put Humpty together again.

This familiar "Mother Goose" nursery rhyme captures the essence of what has become a central problem in the field of visual science: how can an aggregate of individuals working together reassemble a complex object that has been broken into countless parts? In the case of vision, the "horses and men" comprise the many millions of brain neurons devoted to the analysis of visual information, neurons distributed over multiple areas of the brain (Van Essen et al. 1992; Logothetis, 1998). And, in the case of vision, the source to be reassembled corresponds to the panorama of objects and events we experience upon looking around our visual world. In contemporary parlance, this "reassembly" process has been dubbed the *binding problem*, and it has become a major focus of interest in neuroscience (Gray, 1999) and cognitive science (Treisman 1999).

We have no way of knowing exactly why the King's men and horses failed to solve their particular version of the binding problem, but perhaps they possessed no "blueprint" or "picture" of Humpty Dumpty to guide their efforts. (Just think how difficult it would be to assemble a jig-saw puzzle without access to the picture on the box.) Of course, the brain doesn't have blueprints or pictures to rely on either, but it can -- and apparently does -- exploit certain regularities to constrain the range of possible solutions to the binding problem (Marr, 1982). These regularities arise from physical properties of light and matter, as well as from relational properties among objects. Among those relational properties are ones that arise from the temporal structure created by the dynamical character of visual events; in this chapter we present psychophysical evidence that the brain can exploit that temporal structure to make educated guesses about what visual features go with one another. To begin, we need to back up a step and review briefly how the brain initially registers information about the objects and events we see and then outline several alternative theories of feature binding.

S.-W. Lee, H.H. Bülthoff, T. Poggio (Eds.): BMCV 2000, LNCS 1811, pp. 635-653, 2000.

2 Early Vision and Possible Binding Mechanisms

It is generally agreed that early vision entails local feature analyses of the retinal image carried out in parallel over the entire visual field. By virtue of the receptive-field properties of the neurons performing this analysis, visual information is registered at multiple spatial scales, ranging from coarse to fine, for different contour orientations (DeValois and DeValois, 1988). Of course, we are visually unaware of this multiscale dissection of the retinal image: from a perceptual standpoint, coarse and fine spatial details bind seemlessly into harmonious, coherent visual representations of objects. In addition to spatial scale, different qualitative aspects of the visual scene -- color, form, motion -- engage populations of neurons distributed among numerous, distinct visual areas (Zeki, 1993). But, again, the perceptual concomitants of those distributed representations are united perceptually; it's visually impossible to disembody the "red" from the "roundness" of an apple. Evidently, then, the process of binding -- wherein the distributed neural computations underlying vision are conjointly represented -- transpires automatically and efficiently.

Several possible mechanisms for feature binding have been proposed over the years. These include:

- coincidence detectors - these are neurons that behave like logical AND-gates, responding only when specific features are all present simultaneously in the appropriate spatial arrangement (Barlow, 1972). This idea is implicit in Hubel and Wiesel's serial model of cortical receptive fields (Hubel and Wiesel, 1962), in which a simple cell with an oriented receptive field is activated only when each of its constituent thalamic inputs is simultaneously active. Coincidence detection has also been proposed as the neural basis for registration of global, coherent motion (Adelson and Movshon, 1982) and for the integration of information within the "what" and "where" visual pathways (e.g., Rao et al, 1997). A principal objection to coincidence detection as a mechanism of binding is the combinatorial problem: there are simply not enough neurons to register the countless feature combinations that define the recognizable objects and events in our world (e.g., Singer and Gray, 1995; but see Ghose and Maunsell, 1999, who question this objection).
- attention - distributed representations of object features are conjoined by the act of attending to a region of visual space (Treisman and Gelade, 1980; Ashby et al, 1996). This cognitively grounded account, while minimizing the combinatorial problem, remains ill-defined with respect to the actual neural concomitants of attentional binding.
- temporal synchronization - grouping of object features is promoted by synchronization of neural activity among distributed neurons responsive to those various features (von der Malsburg, 1995; Milner, 1974; Singer, 1999). On this account, the coupling of activity among neurons can occur within aggregates of neurons within the same visual area, among neurons within different visual areas and, for that matter, among neurons located in the separate hemispheres of the brain.

This third hypothesis has motivated recent psychophysical work on the role of temporal structure in spatial grouping, including work in our laboratory using novel displays and tasks. In the following sections this chapter examines the extent to which visual grouping is jointly determined by spatial and temporal structure. From the

outset we acknowledge that there is considerable skepticism within the field of biological vision about the functional significance of synchronized discharges among neurons. This skepticism exists for several reasons. Some have argued that synchronized activity may be an artifact that simply arises from the spiking behavior of retinal and/or thalamic cells unrelated to stimulus-driven phase locking (Ghose and Freeman 1997) or from the statistical properties of rapidly spiking cells (Shadlen and Movshon 1999). Others question whether the noisy spike trains from individual neurons possess the temporal fidelity for temporal patterning to be informationally relevant (Shadlen and Newsome 1998). Still others argue that neural synchrony, even if it were to exist, provides a means for *signalling* feature clusters but not a means for *computing* which features belong to an object, therefore leaving the binding problem unsolved (Shadlen and Movshon, 1999). Finally, some have found stimulus induced synchrony is just as likely between neurons responding to figure and background elements as between neurons responding to figure elements only (Lamme and Spekreijse 1998). Balanced overviews of these criticisms are provided in recent reviews by Gawne (1999) and by Usrey and Reid (1999).

While not ignoring these controversies, our approach has been to study the effect of externally imposed synchrony on visual perception. We reason that if temporal synchronization were to provide a means for binding and segmentation, psychophysical performance on perceptual grouping tasks should be enhanced when target features vary synchronously along some dimension over time, whereas performance should be impaired when features vary out-of-phase over time. These predictions are based on the assumption that temporal modulations of an external stimulus produce modulations in neural activity, within limits of course.[1] We give a more detailed exposition of this rationale elsewhere (Alais et al 1998). The following sections review evidence that bears on the question: "To what extent do features changing together over time tend to group together over space?" Next, we turn to our very recent studies using stochastic temporal structure, the results from which provide the most compelling demonstrations to date for the role of temporal structure in spatial grouping. The chapter closes with speculative comments about whether the binding problem really exists from the brain's perspective.

3 Temporal Fluctuation and Spatial Grouping

Periodic Temporal Modulation The simplest, most widely used means for varying temporal structure is to repetitively flicker subsets of visual features, either in-phase or out-of-phase. If human vision exploits temporal phase for spatial organization, features flickering in synchrony should group together and appear segregated from those flickering in different temporal phases. The following paragraphs briefly summarize studies that have utilized this form of temporal modulation.

[1] 1. Most versions of the temporal binding hypothesis posit the existence of intrinsically mediated neural synchrony engendered even by static stimulus features. On this model, synchrony is the product of neural circuitry, not just stimulation conditions. Hence, one must be cautious in drawing conclusions about explicit mechanisms of "temporal binding" from studies using externally induced modulations in neural activity.

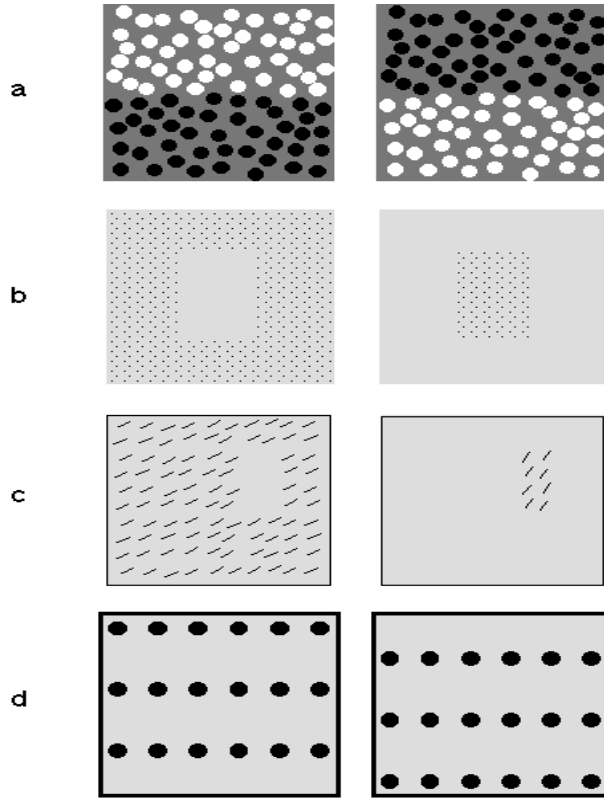

Fig. 1. Examples of visual displays used to assess the contribution of repetitive flicker on spatial grouping (redrawn from originals). A. Roger-Ramachandran and Ramachandran (1998). Black and white spots formed a texture border in the first and second frames of a two-frame "movie" -- the contrast of all spots reversed each frame. B. Fahle (1993). Rectangular arrays of small dots forming the "figure" and the "background" were rapidly interchanged. C. Kiper et al (1996). Oriented contours form successively presented "background" and "figure" frames that were rapidly interchanged. D. Usher & Donnelly (1998). A square lattice of elements (shown here as filled circles) was presented with alternating rows presented successively or with all elements presented simultaneously (not shown)

Rogers-Ramachandran and Ramachandran (1991, 1998) created an animation consisting of two frames (Figure 1a). Black and white dots were spatially distributed against a grey background to create a texture border in the first frame. Then, in the second frame, the luminance of all spots was reversed (black turned to white and vice versa). Repetitively alternating the two frames created counterphase flicker of the two groups of dots. When the display flickered at 15 hz, that is, the phase difference between the two groups of spots was 15 msec, observers could still perceive the texture boundary but could not discern whether any given pair of spots was flickering in-phase or out-of-phase. Rogers-Ramachandran and Ramachandran termed this

boundary a "phantom contour" because clear texture segregation was perceived even though texture elements themselves were indistinguishable.

Using a similar strategy, Fahle (1993) manipulated the stimulus onset asynchrony between two groups of flickering dots and measured the clarity of figure-ground segregation. He tested arrays of regularly and randomly spaced dots (Figure 1b). The dots within a small "target" region (denoted by a rectangle in Figure 1b) were flickered in synchrony while dots outside the target region flickered at the same rate but with temporal phase the flicker delayed relative to that of the target dots. Observers could judge the shape of the target region with temporal phase shifts as brief as 7 msec under optimal conditions. Fahle concluded that the visual system can segregate a visual scene into separate regions based on "purely temporal cues" because the dots in figure and those in ground were undifferentiated within any single frame, with temporal phase providing the only cue for shape. Kojima (1998) subsequently confirmed this general finding using band-passed random dot texture stimuli.

The results from these two studies imply that very brief temporal delays between "figure" and "background" elements provide an effective cue for spatial grouping and texture segregation. However, other studies using rather similar procedures have produced conflicting results. Kiper,

Gegenfurtner and Movshon (1991, 1996) asked whether onset asynchrony of texture elements influences performance on tasks involving texture segmentation and grouping. Observers discriminated the orientation (vertical vs horizontal) of a rectangular region containing line segments different in orientation from those in a surrounding region (Figure 1c). Kiper et al varied the angular difference in orientation between target and background texture elements, a manipulation known to affect the conspicuity of the target shape. They also varied the onset time between target and background elements, reasoning that if temporal phase is utilized by human vision for texture segmentation, performance should be superior when "target" texture elements are presented out of phase with the "background" elements. To the contrary, however, temporal asynchrony had no effect on the ease of texture segmentation; performance depended entirely on the spatial cue of orientation disparity.

Using a novel bistable display (Figure 2), Fahle and Koch (1995) also failed to find evidence that temporal cues promote spatial organization. When the two identical Kanizsa triangles formed by illusory contours partially overlapped, observers typically experienced perceptual rivalry: one triangle appeared nearer than the other, with the depth order reversing spontaneously every several seconds. When one triangle was made less conspicuous by misaligning slightly the inducing elements, the other, unperturbed triangle dominated perceptually. However, introducing temporal offsets between the different inducing elements of a triangle had no significant effect on its perceptual dominance.

How can we reconcile these conflicting results? In those studies where temporal phase mattered (Roger-Ramachandran and Ramachandran, 1991, 1998; Fahle, 1993; Kojima, 1998), there were no spatial cues for segmentation -- all elements in the displays were identical in form, orientation, disparity and other static properties. In contrast, obvious spatial cues were present in displays revealing little or no effect of temporal phase (Kiper et al, 1991, 1996; Fahle and Koch, 1995). Perhaps, then, the salience of temporal structure on spatial grouping is modulated by the presence and

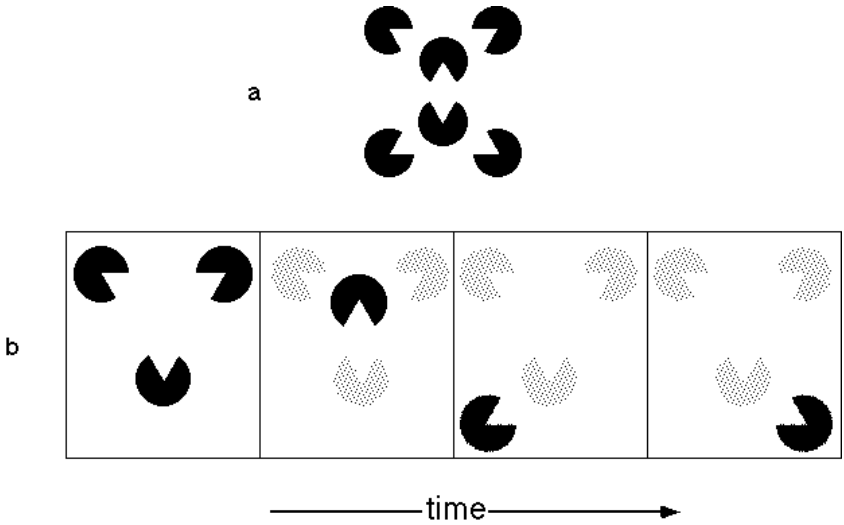

Fig. 2. Display used by Fahle & Koch (1995). When viewed without flicker, the two superimposed illusory Kanizsa triangles fluctuate in perceptual dominance, with one appearing in front of the other for several seconds at a time. The temporal configuration of the components of the illusory triangles was manipulated by briefly presenting all three pacmen for one triangle simultaneously followed by brief, sequential presentation of the three pacmen forming the other triangle. (In this schematic, the lightly stippled pacmen were not actually presented during the sequence and are shown here as reference for the positions of the single pacmen)

strength of spatial cues. Leonards, Singer and Fahle (1996) explicitly tested this idea using texture arrays like those used by Kiper et al (1991; 1996). Figure and background could be defined by a difference in temporal phase alone, by a difference in orientation alone, or by both temporal phase and orientation differences (Figure 1c). In line with Roger-Ramachandran and Ramachandran (1991, 1998), Fahle (1993) and Kojima (1998), the effect of temporal cues on texture segmentation was significant when figures were defined only by temporal phase difference or when temporal cues and spatial cues defined the same figures. When figures were well defined by spatial cues, temporal phase had no influence on figure/ground segregation. Based on their results, Leonards et al proposed a flexible texture segmentation mechanism in which spatial and temporal cues interact.

Comparable results have been reported by Usher and Donnelly (1998), who used a square lattice display (Figure 1d) in which elements appear to group into either rows or columns. When the elements in alternating rows (or columns) of the lattice were flickered asynchronously, the display was perceived as rows (or columns) correspondingly. In the second experiment, Usher and Donnelly presented arrays of randomly oriented lines segments and asked observers to detect collinear target elements. Performance was better when target and background line segments flickered asynchronously than when they flickered in synchrony. Under this condition, it should be noted, temporal and spatial cues were congruent. However, the efficacy of temporal phase waned when the same target elements were randomly

oriented and no longer collinear. Now the temporal lag had to be extended to about 36 msec before target elements were segregated from background elements. These results, together with those of earlier works, indicate that the potency of temporal information for spatial segmentation and grouping depends on the salience of available spatial cues.

Random Contrast Modulation The studies summarized above all used periodic temporal modulation in which luminance values fluctuated predictably between levels. In an effort to create more unpredictable temporal modulation, our laboratory developed displays in which the contrast levels of spatial frequency components comprising complex visual images (e.g., a face) are modulated over time. With this form of temporal modulation, the amplitude of the contrast envelope increases and decreases by random amounts over time without changing the space-average luminance of the display. In our initial work (Blake and Yang, 1997), we found that observers were better able to detect synchronized patterns of temporal contrast modulation within hybrid visual images composed of two components when those components were drawn from the same original picture. We then went on to show that "spatial structure" coincides with the phase-relations among component spatial frequencies (Lee and Blake, 1999a). These two studies set the stage for our more recent experiments examining the role of synchronous contrast modulations in perception of bistable figures.

In one study (Alais et al, 1998), we created a display consisting of four spatially distributed apertures each of which contained a sinusoidal grating that, when viewed alone, unambiguously appeared to drift in the direction orthogonal to its orientation (Figure 3). When all four gratings were viewed together, however, they intermittently grouped together to form a unique "diamond" figure whose global direction of motion corresponded to the vector sum of the component motions. Thus upon viewing this quartet of gratings, observers typically experience perceptual fluctuations over time between global motion and local motion. We found that the incidence of global motion increased when contrast modulations among the gratings were correlated and decreased when the component contrast modulations were uncorrelated. Similar results were obtained using a motion plaid in which two gratings drifting in different directions are spatially superimposed. Under optimal conditions, this display is also bistable, appearing either as two transparent gratings drifting in different directions or as one plaid moving in a direction determined by a vector sum of the velocities of component gratings (Adelson and Movshon, 1982).

Again, the incidence of coherent motion was enhanced by correlated contrast modulations and suppressed by uncorrelated contrast modulations.A second study assessed the role of temporal patterning of contrast modulation in grouping visual features using another bistable phenomenon, binocular rivalry. When dissimilar patterns are imaged on corresponding areas of the retinae in the two eyes, the patterns compete for perceptual dominance (Breese, 1899; Blake, 1989). Since binocular rivalry is strongly local in nature, however, predominance become piecemeal with small zones of suppression when rival targets are large or when small targets are distributed over space (Blake, O'Shea and Mueller, 1992). Alais and Blake (1999) investigated the potency of correlated contrast modulation to promote conjoint dominance of two, spatially separated rival targets pitted against random-dot patches

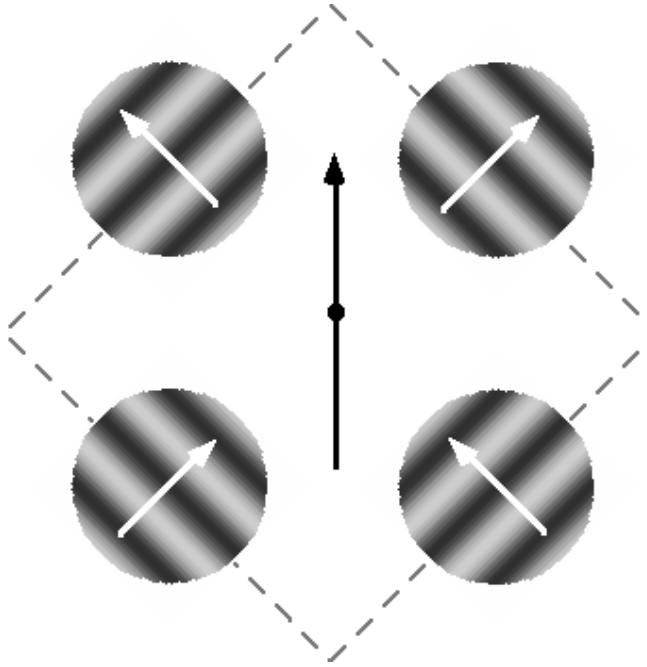

Fig. 3. Display used by Alais et al (1998). Gratings drifted smoothly in the direction indicated by the white arrows. When all four gratings are viewed simultaneously, the four occasionally group to form a partially occluded 'diamond' whose direction of motion corresponds to the vector sum of the component motions ("upward" in this illustration)

presented to corresponding portions of the other eye. The orientations of the two gratings were either collinear, parallel or orthogonal, and they underwent contrast modulations that were either correlated or uncorrelated. Correlated contrast modulation promoted joint grating predominance relative to the uncorrelated conditions, an effect strongest for collinear gratings. Joint predominance depended strongly on the angular separation between gratings and the temporal phase-lag in contrast modulations. Alais and Blake (1999) speculated that these findings may reflect neural interactions subserved by lateral connections between cortical hypercolumns.

Generalizing from the studies summarized in these last two sections, it appears that temporal flicker cannot overrule explicit spatial structure - flicker, in other words, does not behave like "super glue" to bind spatial structures that do not ordinarily form coherent objects. Nor, for that matter, does out-of-phase flicker destroy spatial structure defined by luminance borders. Based on the above results, one would conclude that temporal flicker promotes grouping primarily when spatial structure is ambiguous or weak. The generality and implications of these earlier studies must be qualified in two ways:

- All the studies cited above utilized local stimulus features whose luminance or contrast was modulated periodically, with only the rates and phases of flicker varying among different components of the display. The use of periodic flicker is

grounded in linear systems analysis, which has played an important role in shaping techniques used in visual science. Still, periodic flicker constitutes a highly predictable, deterministic signal which could be construed as a somewhat artificial event. In the natural environment, there exists considerable irregularity in the temporal structure of the optical input to vision: objects can move about the visual environment unpredictably, and as observers we move our eyes and heads to sample that environment. In contrast, repetitively flashed or steadily moving stimuli, so often used in the laboratory, are highly predictable events which, from an information theoretical perspective, convey little information. In fact, periodically varying stimulation may significantly underestimate the temporal resolving power of neurons in primate visual cortex (eg Buracas et al 1998).

- In those studies where evidence for binding from temporal synchrony was found, successive, individual frames comprising a flicker sequence contained visible luminance discontinuities that clearly differentiated figure from background; spatial structure was already specified in given, brief "snapshots" of those dynamic displays. Thus these studies do not definitively prove that human vision can group spatial features based purely on temporal synchrony.

Stochastic Temporal Structure The two considerations presented in the previous paragraph motivated us to develop stochastic animations in which individual frames contained no static cue about spatial structure, leaving only temporal information to specify spatial grouping. How does one remove all static cues from individual frames? A hint for tackling this challenge came from an expanded notion of 'temporal structure' conveyed by time-varying signals. There are three alternative ways to define temporal structure (see Appendix, section A): (i) a time series of absolute quantity, (ii) a collection of distinctive times when a particular event occurs (point process), and (iii) a collection of times for events and magnitudes associated with those events (marked point process). Among of these representations, the point process contains information only about 'time' and not about magnitude. Thus we reasoned that if it's possible to create an animation display in which groups of local elements differ only in terms of their respective point processes, but do not differ in other stimulus properties when considered on a frame by frame basis, that display would be devoid of static spatial cues.

To create these conditions, we created animation displays in which each individual frame consisted of an array of many small sinusoidal gratings each windowed by a stationary, circular gaussian envelope -- such stimuli are termed 'Gabor patches'. All Gabor patches throughout the array had the same contrast, and their orientations were randomly determined. As the animation was played, grating contours within each small, stationary Gabor patch moved in one of two directions orthogonal to their orientation. Each grating reversed direction of motion irregularly over time according to random (Poisson) process. Temporal structure of each Gabor element was described by a point process consisting of points in time at which motion reversed direction (Figure 4). When all Gabor elements within a virtual 'figure' region reversed their direction of motion simultaneously while Gabor elements outside of this area changed direction independently of one another, the 'figure' region defined by temporal synchrony stood out conspicuously from the background. Here, the 'figure' region and the 'ground' region differed only in point process and they were

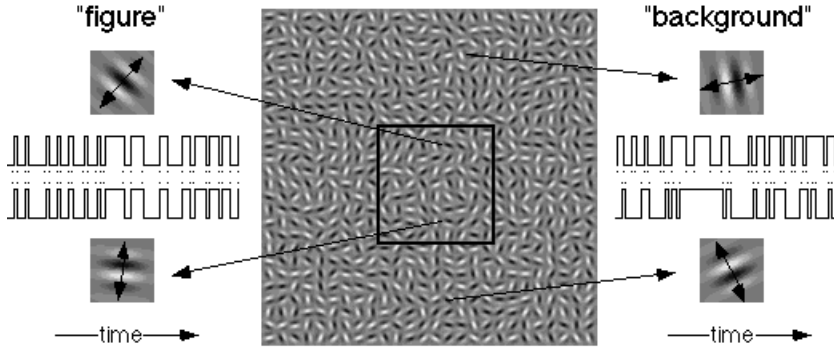

Fig. 4. Display used by Lee and Blake (1999). One frame from an animation sequence consisting of an array of small Gabor patches within which contours move in one of two directions orthogonal to their orientation; from frame-to-frame motion direction changes irregularly. Shown schematically on either side of the square array of Gabor patches are enlarged pictures of several Gabor patches with double-headed arrows indicating the two possible directions of motion. The time series indicate direction of motion, and the small dots associated with each time series denote that time series' point process - points in time at which direction changed. Gabor patches within a virtual region of the array (dotted outline) have point processes that are correlated while Gabor patches outside this virtual area are uncorrelated.(Reprinted by permission from Science, volume 5417, p. 1166. Copyright 1999 © American Association for the Advancement of Science)

not defined by static cues such as luminance, contrast and orientation in individual frames. Furthermore, there was no information about the figure in any two successive frames of the animation, for all contours moved from frame to frame. Therefore, only the difference in point process (which specifies 'when' direction of motion changes) distinguished figure from background.

Because all local elements throughout the display changed directions of motion 'irregularly' over time, we were able systematically to manipulate two variables that reveal just how sensitive human vision is to the temporal structure contained in these dynamic displays. One variable is the 'predictability' (or 'randomness') of temporal structure conveyed by individual elements -- this is easily manipulated by changing the probability distribution associated with the two directions of motion. Borrowing ideas from information theory, this 'predictability' was quantified by computing the entropy of temporal patterns of elements (see Appendix, section B). Since time-varying signals with high entropy convey more dynamic, fine temporal structure, the systematic manipulation of the entropy of all the elements made it possible to examine how accurately human vision can register fine temporal structure contained in contours irregularly changing direction of motion. The second variable was the temporal relationship among elements in the 'figure' region -- this was manipulated by varying the extent to which all possible pairs of those elements are correlated (see Appendix, section C). Since the time points at which the elements change direction of motion could be represented by point processes as described earlier, the index of temporal relationship among those elements could be quantified by computing the correlations among their point processes. By varying this index systematically, the efficiency of human vision in utilizing temporal structure was examined.

Lee and Blake (1999b) found that these two factors (entropy and correlation) both systematically affected the perceptual quality of spatial structure; the clarity of perceived figure/ground segregation increased with increases in the entropy of the entire display and with the correlation among elements within the 'figure' region. Our results clearly show that human vision can register fine temporal structure with high fidelity and can efficiently construct spatial structure solely based on the temporal relationship among local elements distributed over space.

From the outset of this work, we carefully tried to identify and evaluate possible stimulus artifacts that could underlie perception of shape in these displays. An artifact would be any cue that does not depend on correlations among point processes for shape definition. In the displays utilizing an array of Gabor patches, for example, reversals in direction of motion mean that contours return to the same positions they occupied one frame earlier. If contrast were summed over multiple frames, it is possible that these 2-frame change sequences, when they occurred, could produce enhanced contrast of that Gabor. When all Gabors in the figure region obey the same point process, these "pulse" increases in apparent contrast could define the figure against a background of irregular contrast pulses occurring randomly throughout the background where changes were unsynchronized. To counteract this cue, we randomly varied the contrast values of Gabor patches from frame-to-frame throughout the array, thereby eliminating contrast as a potential cue. Spatial form from temporal synchrony remains clearly visible. When the elements defining the figure are all synchronized but the background elements are not, the background contains a more varied pattern of temporal change, which means that the temporal frequency amplitude spectra of the background - considered across all background elements - may differ in detail from that of the figure, although both would be quite broad. However, this difference does not exist when background elements all obey the same point process (albeit one different from the figure) - still, shape discrimination performance remains quite good for this condition, ruling out differences in temporal frequency amplitude spectra as an alternative cue.

Another possible artifact arises from the possibility of temporal integration over some number of consecutive animation frames in which no change in direction occurs. If the grating contours were to step through one complete grating cycle without change in direction and if the integration time constant were to match the time taken for that event, the net luminance of pixels defining that grating would sum to levels near the mean and effectively produce a patch of approximately uniform brightness. For the synchronized condition, all Gabors in the figure would assume this brightness level simultaneously while those in the background would not because they are changing direction at different times on average - with time-integrated signals, the figure region could occasionally stand out from the background simply based on luminance. In fact, this is not at all what one sees when viewing these displays - one instead sees smooth motion continuously throughout the array of Gabor elements. Still, it could be argued that this putative integrated luminance cue is being registered by a separate, temporally sluggish mechanism. To formalize this hypothesis, we have simulated it in a MatLab program. In brief, the model integrates individual pixel luminance values over *n* successive animation frames, using a weighted temporal envelope. This procedure creates a modified sequence of animation frames whose individual pixels have been subjected to temporal integration, with a time constant

explicitly designed to pick up *n*-frame cycles of no change. We then compute the standard deviation of luminance values within the figure region and within the background and use those values to compute a signal/noise index. To the extent that luminance mediates detection of spatial structure in these dynamic displays, psychophysical performance should covary with this "strength" index. We have created sequences minimizing the incidence of "no change" sequences by manipulating entropy and by selectively removing "no change" sequences in an animation - both manipulations affect the "strength" index. When we compare performance on trials where this cue is potentially present to trials where it is not, we find no difference between these two classes of trials.

Our work to date has focused on changes in direction of translational motion and changes in direction of rotational motion. These were selected, in part, because registration of motion signals requires good temporal resolution. There is no reason to believe, however, that structure from temporal synchrony is peculiar to motion. Indeed, to the extent that temporal structure is fundamentally involved in feature binding, any stimulus dimension for which vision possesses reasonable temporal resolution should in principle be able to support figure/ground segmentation from synchrony. There must, of course, be limits to the abstractness of change supporting grouping. Consider, for example, a letter array composed of vowels and consonants. It would be very unlikely that irregular, synchronized changes from vowels to consonants, and vice versa, in the array would support perceptual grouping (unless those changes were accompanied by prominent feature changes, such as letter size or font type). Grouping should be restricted to image properties signifying surfaces and their boundaries.

4 Concluding Remarks

Our results, in concert with earlier work, lend credence to the notion that temporal and spatial coherence are jointly involved in visual grouping. Biological vision, in other words, interprets objects and events in terms of the relations -- spatial and temporal -- among features defining those objects and events. Our results also imply that visual neurons modulate their responses in a time-locked fashion in response to external time-varying stimuli. Theoretically, this stimulus-locked variation in neural response could be realized either of two ways depending on how information is coded in the spike train. One class of models emphasizes temporal coding wherein fine temporal structure of dynamic visual input is encoded by exact locations in time of individual neural spikes in single neurons; in effect, the "code" is contained in the point processes associated with trains of neural activity.

Alternatively, a second class of models assumes that stimulus features are encoded by ensembles of neurons with similar receptive field properties. The average firing rate within a neural ensemble can fluctuate in a time-locked fashion to time-varying stimuli with temporal precision sufficient to account for the psychophysical data presented here (Shadlen and Newsome, 1994). Although the temporal coding scheme allows neurons to more efficiently transmit information about temporal variations in

external stimuli than the rate coding scheme, psychophysical evidence alone does not definitively distinguish between the two models.

Regardless how this coding controversy is resolved, our results using stochastic animations convincingly show that temporal structure in the optical input to vision provides a robust source of information for spatial grouping. Indeed, one could argue that vision's chief job is extracting spatio-temporal structure in the interest of object and event perception. After all, the optical input to vision is rich in temporal structure, by virtue of the movement of objects and the movement of the observer through the environment. Consequently, our eyes and brains have evolved in a dynamic visual world, so why shouldn't vision be designed by evolution to exploit this rich source of information?

In closing, we are led to speculate whether the rich temporal structure characteristic of normal vision may, in fact, imprint its signature from the outset of neural processing. If this were truly the case, then concern about the binding problem would fade, for there would be no need for a mechanism to reassemble the bits and pieces comprising visual objects. Perhaps temporal structure insures that neural representations of object "components" remain conjoined from the very outset of visual processing. Construed in this way, the brain's job is rather different from that facing the King's horses and men who tried to put Humpty Dumpty back together. Instead of piecing together the parts of a visual puzzle, the brain may resonate to spatio-temporal structure contained in the optical input to vision.

References

Adelson EH and Movshon JA (1982) Phenomenal coherence of moving visual patterns. Nature, 300, 523-525.

Alais D and Blake R (1998) Interactions between global motion and local binocular rivalry Vision Research, 38, 637-644.

Alais D, Blake R and Lee S (1998) Visual features that vary over time group over space. Nature Neuroscience., 1, 160-164.

Ashby FG, Prinzmetal W, Ivry R, Maddox,WT (1996) A formal theory of feature binding in object perception. Psychological Review, 103, 165-192.

Barlow, HB (1972) Single units and sensation: a neuron doctrine for perceptual psychology? Perception, 1, 371-394.

Blake R (1989) A neural theory of binocular rivalry. Psychological Review 96, 145-167.

Blake R and Yang Y (1997) Spatial and temporal coherence in perceptual binding. Proceedings of the National Academy of Science, 94, 7115-7119.

Blake R, O'Shea RP and Mueller TJ (1992) Spatial zones of binocular rivalry in central and peripheral vision. Visual Neuroscience, 8, 469-478.

Breese, BB (1899) On inhibition. Psychological Monograph, 3, 1-65.

Brillinger, DR (1994) Time series, point processes, and hybrids. Canadian Journal of Statistics, 22, 177-206

Brook D and Wynne RJ (1988) Signal Processing: principles and applications. London: Edward Arnold.

De Coulon, F (1986). Signal Theory and Processing. Dedham MA: Artech House, Inc.

DeValois RL and DeValois KK (1988) Spatial vision. New York: Oxford University Press.
Fahle M (1993) Figure-ground discrimination from temporal information. *P*roceedings of the Royal Society of London, B, 254, 199-203.
Fahle M and Koch C (1995) Spatial displacement, but not temporal asynchrony, destroys figural binding. Vision Research, 35, 491-494.
Gawne T J (1999) Temporal coding as a means of information transfer in the primate visual system. Critical Reviews in Neurobiology 13, 83-101.
Ghose GM and Freeman RE (1997) Intracortical connections are not required for oscillatory activity in the visual cortex. Visual Neuroscience 14, 963-979.
Ghose GM and Maunsell J (1999) Specialized representations in visual cortex: a role for binding? Neuron 24 79-85.
Gray CM (1999) The temporal correlation hypothesis of visual feature integration: still alive and well. Neuron 24 31-47.
Hubel DH and Wiesel TN (1962) Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. Journal of Physiology, London, 160, 106-154.
Kiper DC, Gegenfurtner KR and Movshon JA (1996) Cortical oscillatory responses do not affect visual segmentation. Vision Research, 36, 539-544.
Kiper DC and Gegenfurtner KR (1991) The effect of 40 Hz flicker on the perception of global stimulus properties. Society of Neuroscience Abstracts, 17, 1209.
Kojima H (1998) Figure/ground segregation from temporal delay is best at high spatial frequencies. Vision Research, 38, 3729-3734.
Lamme VAF and Spekreijse H (1998) Neuronal synchrony does not represent texture segregation. Nature 396, 362-366.
Lee SH & Blake R (1999a) Detection of temporal structure depends on spatial structure. Vision Research, 39, 3033-3048.
Lee SH & Blake R(1999b) Visual form created solely from spatial structure. Science, 284, 1165-1168.
Leonards U, Singer W, & Fahle M (1996) The influence of temporal phase differences on texture segmentation. Vision Research, 36, 2689-2697.
Logothetis,NK (1998) Single units and conscious vision. Philosophical Transactions of the Royal Society, London B, 353, 1801-1818.
Mainen ZF and Sejnowski TJ (1995) Reliability of spike timing in neocortical neurons. Science, 268, 1503-1506.
Mansuripur M (1987) Introduction to Information Theory. Englewood Cliffs NJ: Basic Books.
Marr D (1982) Vision: A computational Investigation into the human representation and processing of visual information. San Francisco: WH Freeman.
Milner, P.M. (1974) A model for visual shape recognition. *Psychological Review*, 81, 521-535.
Rogers-Ramachandran DC and Ramachandran VS (1998) Psychophysical evidence for boundary and surface systems in human vision. Vision Research, 38, 71-77.
Rogers-Ramachandran DC and Ramachandran VS (1998) Phantom contours: Selective stimulation of the magnocellular pathways in man. Investigative Ophthalmology and Visual Science, Suppl., 26.
Rao SC, Rainer G, Miller EK (1997) Integration of what and where in the primate prefrontal cortex. Science, 276, 821-824

Shadlen M and Movshon JA (1999) Synchrony unbound: a critical evaluation of the temporal binding hypothesis. Neuron, 24 67-77.
Shadlen MN and Newsome WT (1994) Noise, neural codes and cortical organization. Current Opinion in Neurobiology, 4, 569-579.
Shadlen MN and Newsome WT (1998) The variable discharge of cortical neurons: implications for connectivity, computation, and information coding. Journal of Neuroscience, 18, 3870-3896.
Shannon CE and Weaver W (1949) The mathematical theory of communication. Urbana: Univ. of Illinois Press.
Singer W and Gray CM (1995) Annual Review of Neuroscience, 18, 555-586.
Singer W.(1999) Striving for coherence. Nature, 397, 391-392.
Treisman A (1999) Solutions to the binding problem: progress through controversy and convergence. Neuron 24 105-110.
Treisman A and Gelade G (1980) A feature-integration theory of attention. Cognitive Psychology, 12, 97-136.
Usher M and Donnelly N (1998) Visual synchrony affects binding and segmentation in perception. Nature, 394, 179-182.
Usrey WM and Reid RC (1999) Synchronous activity in the visual system. Annual Review of Physiology, 61, 435-456.
Van Essen DC, Anderson CH and Felleman DJ (1992) Information processing in the primate visual system, an integrated systems perspective. Science 255, 419-423 .
von der Malsburg,C (1995) Binding in models of perception and brain function. Current Opinions in Neurobiology, 5, 520-526.
Zeki S (1993) A vision of the brain. Cambridge MA: Blackwell Scientific.

Appendix: Time-Varying Signals and Information Theory

Throughout this chapter we employ the term “temporal structure” in reference to spatial grouping. In this appendix, we define “temporal structure” using information theory and signal processing theory as conceptual frameworks.

A. Time-Varying Signals

According to signal processing theory, time-varying signals are defined by variations of some quantity over time which may or may not be detectable by a given system (Brook and Wynne, 1988). For the visual system, time-varying signals would be the temporal fluctuations of visual properties of external stimuli. Thus, the temporal structure of visual stimuli can be carried by any property which is potentially detectable by human vision, including luminance, color, stereoscopic disparity, and motion. When a vertically oriented sinusoidal grating shifts its spatial phase either right-ward or left-ward over time, for instance, time-varying signals can be defined by the temporal fluctuations of direction of motion.

Deterministic vs Random Signals Time-varying signals can be distinguished depending on whether they are deterministic or random. Deterministic signals are perfectly predictable by an appropriate mathematical model, while random signals are unpredictable and can generally be described only through statistical observation (de Coulon, 1986). Figure 5a shows an example of a deterministic signal and Figure 5b shows a random signals. In Figure 5a, a deterministic time-varying signal, directions of motion of a grating can be exactly predicted on the basis of time points where the grating shifts its spatial phase because the two opposite directions of motion alternate in a deterministic fashion. For the time-varying signals in Figure 5b, one cannot predict in which direction the grating will be moving at time $t + 1$ by knowing in which direction it is moving at time *t*, because direction of motion changes randomly over time. Instead, one only can make a statistical prediction (e.g., "the grating is more likely to move left-ward than right-ward") by estimating the probabilities of the two motion directions. Unlike most engineered communications systems that use well-defined frequencies to transmit information, many biological systems must deal with irregularly fluctuating signals.

Representation of the temporal structure of time-varying signals The temporal structure contained in time-varying visual signals can be described in any of several ways, depending upon what properties one wishes to emphasize.
(1) Time series of absolute quantity (Figure 5c). The time series plot of absolute quantity is the most direct and simple description of temporal signals (Brillinger, 1994). Here, simply the absolute quantity of visual stimuli is plotted against time;

$$Y(t),\ -4 < t < +4$$

where t is time and Y represents the quantity of stimulus. When a dot changes in luminance over time, for example, its temporal structure can be represented by plotting the absolute level of luminance over time (Figure 5c).
(2) Marked point process (Figure 5d). An irregular time series representing both the times at which events occur as well as the quantities associated with those times:

$$\{(\tau_j, M_j)\},\ j = "\ 1, "\ 2, "\ 3, \ldots$$

where M_j represents the magnitude of change associated with the jth event. In Figure 5d, the locations of vertical lines represent time points when a stimulus changes in luminance and the directions and lengths of vertical lines represent the magnitude and direction (increase or decrease in luminance) of changes, respectively.
(3) Point process (Figure 5e). If we are interested only in 'when' a given stimulus quantity changes or 'when' events occur, we may specify a point process which is the collection of distinctive times when a stimulus changes its quantity;

$$\{\tau_j\},\ j = 0, "\ 1, "\ 2, "\ 3, \ldots$$

where the t_j is a time point for the jth event. For instance, asterisks in Figure 5e denote the times when a stimulus element changes its luminance, without regard to the value of luminance itself. In a point process, the characteristics of temporal structure in time-varying signals can be described by analyzing the distribution of events in a given time period and also the distribution of time intervals between events.

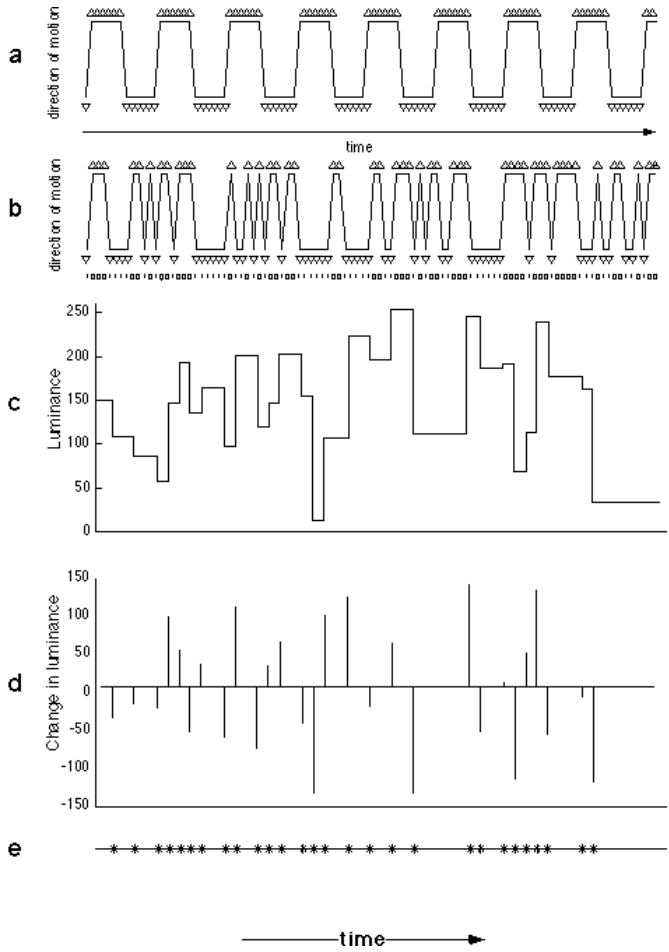

Fig. 5. Time series indicating direction of visual motion over time. (a) Deterministic time-varying signal. Direction of motion is entirely predictable and specifiable mathematically. (b) Random time-varying signal. Direction of motion reverses randomly over time according to a Poisson process. (c) - (e) show three types of representations of temporal structure. (c) Time series of absolute quantity. The absolute luminance level of a given visual stimulus is plotted against time. (d) Marked point process. The locations of vertical lines represent points in time when the stimulus changes in luminance, and the directions and lengths of vertical lines represent the magnitude and direction of change in luminance, respectively. (e) Point process. Asterisks denote times when the luminance value changes.

B. Time-Varying Signals as Information Flow

How well does human vision respond to the temporal structure generated by unpredictable time-varying visual signals? Information theory (Shannon and Weaver, 1949; Mansuripur, 1987) sheds light on how we can conceptualize this problem. According to information theory, 'time-varying visual signals' can be treated as information flow over time as long as they are an ensemble of mutually exclusive and statistically independent messages. For example, the time series of motion direction in Figure 5b can be understood as an ensemble of 'left-ward motion' messages and 'right-ward motion' messages. These messages are mutually exclusive because the grating cannot move in two opposite directions simultaneously. They are also statistically independent since the probability of one message does not depend on the probability of the other. If we assign a symbol '0' to the left-ward motion signal and '1' to the right-ward motion signal, the time plot in the top part of Figure 5b can be translated into an information flow composed of '0' and '1' as illustrated in the bottom portion of Figure 5b.

Furthermore, information theory allows us to have a measure of uncertainty or randomness about the temporal structure of visual stimuli. We can quantify the amount of information as 'entropy' if the probability distribution of messages comprising time-varying signals is known. Suppose that a time series of signals, S, is an ensemble of n different messages, $\{m_1, m_2, \ldots, m_i, \ldots, m_n\}$, and the probabilities for those messages are $\{p_1, p_2, \ldots, p_i, \ldots, p_n\}$. Then the information (expressed in binary units called "bits") attributed to each message is defined by

$$H_i = -\log_2(p_i)$$

and the averaged amount of information,

$$H(S) = -\sum_{i=1}^{N} p_i * \log_2(p_i)$$

is the probability-weighted mean of information contained in all the messages. This equation implies that the averaged amount of information is maximized when all of the possible messages are equally probable. In the earlier example of time-varying stimuli, the direction of motion is the most unpredictable when the two directions of motion are equally probable.

While H(S) represents the uncertainty of time-varying signals, the temporal complexity of signals can be evaluated by the rate of flow of information

$$H(S)/T$$

where T is the average duration of messages from the ensemble, weighted for frequency of occurrence, that is,

$$T = \sum_{i=1}^{N} p_i * T_i$$

The high value of H(S)/T indicates that the relatively large amount of information (high value of H(S)) flow in a short duration (low value of T). Thus, human vision's ability to process the temporal structure of signals can be evaluated by finding the maximum information flow rate which can be processed by human

vision. If the visual system successfully processes time-varying signals with a specific information flow rate, it should (i) reliably generate identical outputs in response to identical input signals and (ii) accurately discriminate among different signals. By measuring these abilities, reliability and accuracy, while varying the rate of information flow, the capacity of human vision for time-varying signals can be determined (e.g., Mainen and Sejnowski, 1995).

C. Characterization of Relationship among Spatially Distributed Temporal Structures

In the visual environment, the sources of time-varying signals are often distributed over space. It seems reasonable to assume that time-varying visual signals arising from the same object are likely to have related temporal structures. Since this is the case, is the visual system able to detect important relations among temporal structures of visual signals? If so, what kinds of relationship among temporal structures is human vision sensitive to?

Once the temporal structures of time-varying signals are defined quantitatively by one of the ways mentioned above (time series of absolute quantity, point process and marked point process), the relationship among stochastic time series can be quantitatively expressed by computing correlation coefficients in the time domain. Suppose we have N observations on two series of temporal signals, x and y,

$$\{x_1, x_2, x_3, ..., x_N\}$$

$$\{y_1, y_2, y_3, ..., y_N\}$$

at unit time interval over the same time period. Then, the relationship between those series can be characterized by the cross-correlation function,

$$\rho_{xy}(k) = \gamma_{xy}(k) / \sqrt{[\gamma_{xx}(0)\gamma_{yy}(0)]} \quad , \quad k = -1, -2, -3, ..., -(N-1)$$

$$\gamma_{xy}(k) = \mathrm{Cov}(x_t, y_{t+k})$$

$$\gamma_{xx}(0) = \mathrm{Cov}(x_0, x_{0+k})$$

$$\gamma_{yy}(0) = \mathrm{Cov}(y_0, y_{0+k})$$

where k is a temporal phase lag between two signals.