This manuscript appears as:

Scholl, B. J. (2001). Spatiotemporal priority and object identity. *Cahiers de Psychologie Cognitive*, *20(5)*, 359 - 371.

Spatiotemporal Priority and Object Identity

[Commentary on F. Bedford's "A Law of Numerical Object Identity"]

BRIAN J. SCHOLL
Yale University

1. Introduction

What is an object? This question can only be addressed by relativizing it to a particular type of cognitive process; for example, the 'objects' of lower-level visual processes may not always correspond to higher-level intuitions about the nature of object-hood. Here I will be concerned with the nature of *visual* objects (keeping in mind that there may also be different types of these, relative to different visual processes — e.g. those which do and do not involve attention). So what is a visual object? Perhaps the most natural way to ask this question is to explore the types of feature clusters in a scene which are segmented and processed together — i.e. to ask what types of stimuli 'count' as objects for the visual system in the first place (e.g. Scholl, Pylyshyn, & Feldman, 2001). Another way of asking about visual objecthood, however, introduces a temporal component: what properties mediate the representation of some portion of the visual field as the *same* object, persisting through time and perhaps motion (e.g. Scholl & Pylyshyn, 1999). Answering such questions is a critical project for vision science, since these properties essentially determine the 'currency' over which many later processes operate.

This 'temporal' aspect of objecthood is one of the many related problems that is dealt with in Bedford's 'nested geometries' model of object identity (Bedford, this issue). This model attempts to account for a wide variety of phenomena in which a decision must be made about whether two states of the world constitute or derive from the same object. Bedford suggests that a host of such 'correspondence' problems — including those in apparent motion, prism adaptation, ventriloquism, priming, and stereopsis — have a common solution. She identifies five types of progressively more constrained geometries (topology, projective, affine, similarity, and Euclidean; see Klein, 1893), and claims that the corresponding objects in any two samples are those whose transformation (from one to the other) satisfy the most constrained geometry. For example, a square may be more likely to be matched with a rectangle than with a circle in certain cases of apparent motion, since the transformation of a square to a rectangle is allowed in affine geometry, whereas the transformation from a square to a circle is not (and is allowed only in the less restrictive topological geometry). This general framework, if correct, is both elegant and incredibly powerful. It is also sociologically unusual: given the incredible over-specialization which infects cognitive psychology, it is rare for a theory to attempt to encompass such a wide variety of phenomena under a single set of explanatory principles.

Though I will not question the details of the nested-geometries framework here, I propose that there is another type of factor — in particular, a type of spatiotemporal priority — which can trump even the object identities which are computed on the most constrained Euclidean geometry. In general, Bedford recognizes that factors beyond the nested geometries, such as higher-level knowledge, can affect judgments of object identity. However, she claims that these other factors are less foundational and consistent across observers, whereas the nested-geometries framework constitutes the primitive 'core' of object identity decisions. In this context, I suggest that spatiotemporal matching constitutes an even more primitive core of such decisions, such that matching according to transformations within the nested geometries may only play a role when these spatiotemporal properties do not provide a unique match.

This work was supported by NIMH #R03-MH63808-01. Address correspondence and reprint requests to Brian Scholl, Department of Psychology, Yale University, Box 208205, New Haven, CT, 06520-8205. Email: Brian.Scholl@yale .edu. Web: http://pantheon.yale.edu/~bs265 .

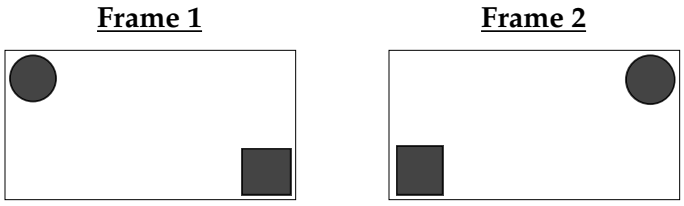

Figure 1. A simple apparent motion display in which vertical motion is typically preferred to horizontal motion. This correspondence minimizes the distance of apparent motion, but violates all but the least constrained topological geometries.

In the remainder of this commentary, I point to several different phenomena which require that persisting object identities be computed among visual objects, including apparent motion, 'launching vs. passing' in causal event perception, and multiple object tracking. In each case I suggest that spatiotemporal factors can trump even the most constrained geometric factors. I begin with apparent motion, since that is one of the domains in which Bedford's framework is most developed and tested.

2. Object Identity in Apparent Motion

The claim of the nested-geometries model for apparent motion is that the perceived correspondence will involve the transformation which is allowed under the most constrained of the geometries — so that, e.g., a square will more likely pair with a rectangle than with a circle, *all other things being equal* (more on this italicized part in a moment). However, a square is more likely to be paired with a circle than with a square-with-a-hole-in-it — since this last transformation violates even topological geometry. (This unintuitive prediction has been confirmed by Chen, 1985; for similar results see Prazdny, 1986, and Ramachandran et al., 1983.) According to Bedford, this framework begins to explain "contradictions" in the apparent motion literature concerning the role of shape: some authors claim that types of shape-based similarity can affect motion correspondence (e.g. Green & Odom, 1986; Mack et al., 1989), while others find no such effects (e.g. Burt & Sperling, 1981; Kolers & Pomerantz, 1971; Navon, 1976; Ramachandran et al., 1983; Schecter et al., 1988). According to Bedford, these conflicts are intelligible from within the nested-geometries model: some shape transformations will play a role (when they adjudicate between different geometries) and some will not (when they do not differ in terms of which geometries they satisfy).

This conclusion seems too simple, though. A more global reading of the research on apparent motion correspondence, I suggest, is that effects of *all* types of shape transformations (whatever their geometries) are minimal compared to spatiotemporal factors such as proximity (á la the 'nearest neighbor principle'; see Dawson, 1991, and Ullman, 1979). Roughly, shape-similarity will bias motion correspondence only when spatiotemporal factors (primarily the items' relative proximities) are balanced.[1] Accordingly, some simple cases of apparent motion appear to conflict with the predictions of the nested-geometries model. The simplest case is just when a circle and a square alternate corners of a rectangle, as in Figure 1. Here observers are more likely to perceive cross-shape vertical motion because of the shorter resulting proximities. Such transformations are illicit in Euclidean, similarity, affine, and projection geometries, whereas adding only a modestly greater spatial offset would give rise to horizontal motion and satisfy even Euclidean geometry. Despite this, the visual system appears to cleave to the spatiotemporal factors, in cases (such as this one) where such factors provide an unambiguous match.

Bedford address a similar case in Section 3.2: "Consider an apparent motion experiment where there are two figures on the right hand side competing to capture a rectangle on the left hand side. . . . [D]oesn't the theory predict that a very distant rectangle, otherwise identical, will capture the left hand rectangle rather than a much nearer rectangle that has been shrunk slightly?" Bedford escapes such a prediction by appealing to the overall

[1] Note that the proximity computations might still be fairly complex. For example, proximity might be measured not in global 2D or 3D coordinates, but as the distance along a surface (He & Nakayama, 1994). Also, in some cases the visual system might prefer local correspondences which violate the 'nearest neighbor' principle, in order to minimize motion at a more global level (e.g. Dawson & Pylyshyn, 1988).

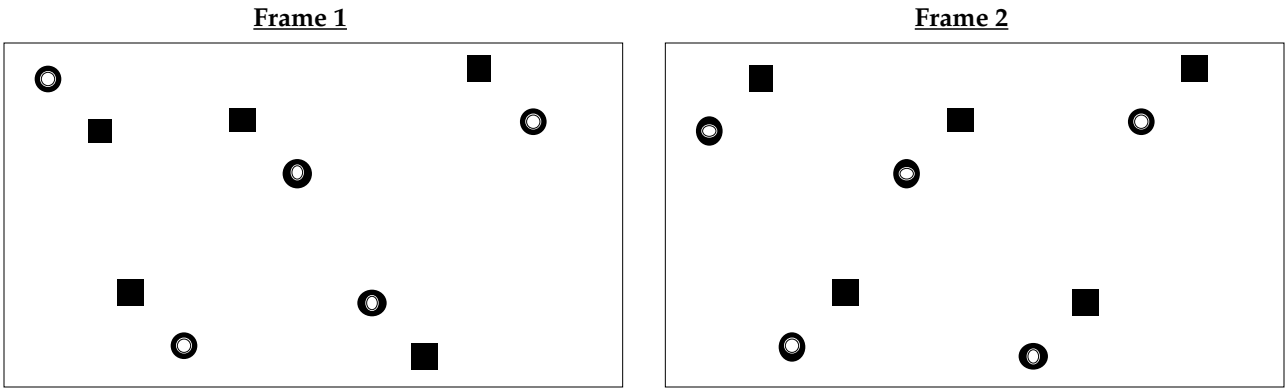

Figure 2. An apparent motion display containing several local 'quartets', modeled after Ramachandran and Anstis (1983). Percepts of such displays are constrained by a type of global spatiotemporal coherence: motion in all quartets occurs in the same direction (either horizontal or vertical), even though this entails gross (and otherwise unnecessary) geometric violations within half of the quartets. (Note that to see the intepretation preferred by the nested-geometries model, you would have to see vertical motion in the upper-left and lower-right quartets, and horizontal motion in the others.).

magnitude of the two transformations: "Comparisons between the levels require equating the amount of transformation from each level." In this case, a "very distant" Euclidean transformation is pit against a "slight" similarity transformation, and since they aren't comparable, the Euclidean interpretation doesn't trump the similarity transformation. Bedford acknowledges the difficulty entailed here, requiring comparisons of qualitatively different kinds of magnitudes ('geometric apples and oranges'), but she suggests that some psychophysical procedures might be able to solve such problems. The real problem here, however, is that the strength of shape-based effects is not even close to the strength of spatiotemporal effects such as proximity (and others such as relative velocity). Indeed, Dawson (1991) estimates that in some contexts shape-based effects are up to 15 times weaker than proximity-based effects, and as a result many models of motion correspondence explicitly ignore shape information (e.g. Burt & Sperling, 1981; Dawson, 1991). As such, the "very distant" vs. "slight" comparison in Bedford's example is unnecessarily extreme, and effects such as that depicted in Figure 1 are possible, wherein only a modest difference in proximity can trump even geometrically-enormous shape changes. Overall, this pattern seems consistent with a core spatiotemporal bias in apparent motion correspondence, which defers to the nested-geometry model only when spatiotemporal information fails to provide a unique match.

3. Global Spatiotemporal Effects in Apparent Motion

A primary role for spatiotemporal information in apparent motion is also indicated in other higher-order effects, beyond simple local correspondences. One such example is the global coherence which is obtained with multiple apparent motion stimuli. Consider, for example, the apparent motion resulting from bistable 'quartets' (in which dots are presented on alternate pairs of diagonals in a square, and the motion can be seen as vertical or horizontal). When multiple quartets of this type fill a display, the display as a whole remains bistable (with local motion being either horizontal or vertical), but each of the individual quartets tends to shift in the same direction as all the others, such that all motion in the display is seen in a single direction (Ramachandran & Anstis, 1983). Figure 2 depicts a variation on such a display. Of particular interest is the fact that this global entrainment effect can override local shape-based differences, even those which violate topological geometries. Thus, when viewing the display depicted in Figure 2, half of the quartets will typically involve correspondence between a square and a donut (because they will all be in the same direction), despite the gross (and otherwise unnecessary) geometric violation. Moreover, note that this effect is inherently asymmetric: while in this case a global spatiotemporal bias ('same direction') can override local shape-based information ('same shape'), the opposite is not true. In Figure 2, for example, it is conceptually possible that an overall entrainment of

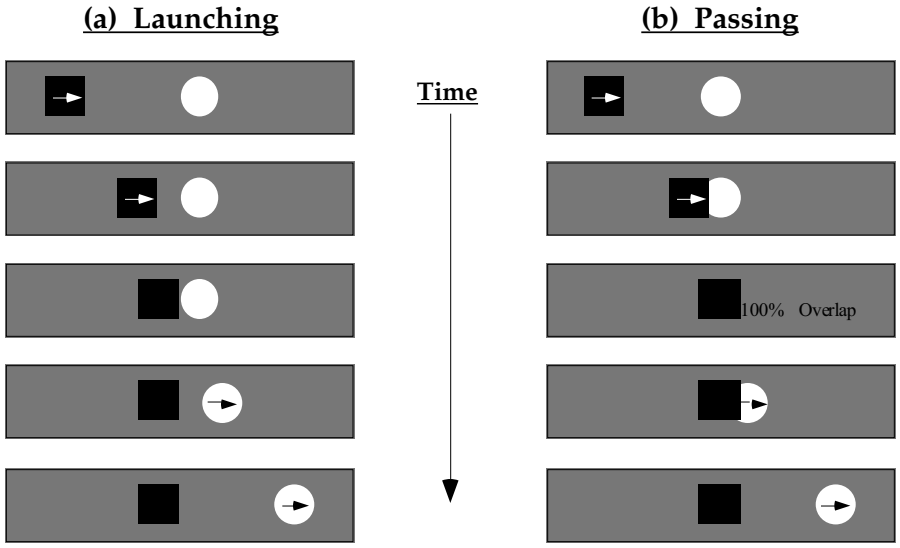

Figure 3. 'Launching' and 'Passing' displays, based on Scholl and Nakayama (under review a). (a) Here observers see the square cause the circle to move, as in a collision. (b) Because the two items overlap completely during their motion, observers typically see a completely non-causal 'pass': observers see a stationary item which changes from a circle to a square, and a moving item which traverses the display, changing from a square to a circle.

shape information could take place, such that circles throughout the display are always matched with circles, and donuts with donuts. That this does not occur is additional evidence that spatiotemporal information can trump even the strongest geometric evidence in apparent motion.[2]

This section and the previous section dealt with apparent motion because this is one of the central phenomena to which Bedford applies her theory. In the next two sections I briefly discuss some other domains (not discussed by Bedford) which require object-identity decisions, and which also seem to involve a spatiotemporal bias which can trump the match provided by the nested-geometries framework.

4. Launching vs. Passing in Event Perception

Consider a square which moves toward a circle until they are adjacent, at which point the square stops and the circle starts moving along the same path (Figure 3a). The perception of such collision events is striking: beyond these objective kinematics, we see the square *cause* the circle's motion (Michotte, 1946/1963; for a recent review, see Scholl & Tremoulet, 2000). Now consider a slight variation (Figure 3b; see Scholl & Nakayama, to appear), wherein the two shapes overlap completely before the square stops and the circle starts moving (it doesn't matter which occludes the other, or if one 'sticks out' slightly behind the other even at the moment of maximal overlap). (To be clear, these displays involve perceptually continuous smooth motion, unlike those discussed in previous sections.) Both of these are clear cases which require a computation of object identity over time. There are two shapes in both the first and last panels of each part of Figure 3; which goes with which? Here the intuitive match upon viewing the figure is also the one preferred by the nested-geometries theory, especially given the continuous motion: the circle goes with the circle, and the square goes with the square. This seems just obvious intuitively, and is certainly preferred by Bedford's theory, since the alternative requires a gross geometric transformation which violates Euclidean, affine, similarity, and projection geometries.

And indeed, observers always see this intuitive match in the launching display (Figure 3a). In the

---

[2]The observation reported in this section with the multi-shape display is my own, and has been confirmed with 6 observers, using the dynamic analogue of Figure 2. All observers immediately and continuously saw motion in the same direction for each quartet (i.e. the Ramachandran & Anstis entrainment effect), despite the featural differences. It is to the credit of Bedford's theory that it motivates such potentially important and previously untested observations.

passing display (Figure 3b), however, most naive observers see a completely non-causal 'pass' instead of a causal 'launch': they see (1) a single stationary shape which turns from a circle to a square; and (2) a single moving shape, which traverses the display, changing at one point from a square to a circle. Similar results occur when the items are different colors; see Scholl and Nakayama (to appear). This 'passing' percept can be seen even in the standard non-overlapping 'launch' if you fixate above or below the display to move it into the periphery (but still keeping the shape-changes visible). Such percepts are striking because you see the nonintuitive motion correspondence (i.e. the 'pass') despite clearly perceiving the shape (and possibly color) differences: e.g. you still *see* the stationary shape change from a square to a circle. (For additional related spatiotemporal manipulations, see Scholl & Naka-yama, to appear, and the related literature on 'bouncing vs. streaming', e.g. Sekuler & Sekuler, 1999.) Such percepts are presumably caused by the dominating role of spatiotemporal information in the form of a continuous motion signal, together with the general impotence of shape (and color) effects in event perception (Michotte, 1946/1963). In general, this seems to be another case where spatiotemporal information can trump even the most constrained geometric transformations. (As with other domains, however, the nested-geometries framework makes interesting untested predictions here too: e.g., this view predicts that the degree of eccentricity required to see a 'launch' as a 'pass' should depend on the nature of the shapes, with a greater eccentricity required to override changes which violate more geometries.)

5. Tracking Multiple Objects Through Occlusion vs. Implosion

Another experimental task which requires the visual system to individuate objects through time is *multiple object tracking* (MOT; Pylyshyn & Storm, 1988). In this paradigm subjects must track a number of independently and unpredictable moving identical items in a field of identical distractors. In a typical experiment, a display will start with 8 static objects, 4 of which will blink several times to indicate their status as targets. The blinking will then stop, and all items will move randomly about the screen for 10 s, after which the subject must use the computer mouse to indicate the 4 targets. Since all items are identical during the motion phase, subjects can only succeed by picking out the targets when they were initially flashed, and then using attention to track them through the motion interval, maintaining their enduring identities as the 'same' objects. Subjects can successfully perform this task (with over 85% accuracy) when tracking up to five targets in a field of ten identical items, and several lines of evidence indicate that the items are being attentionally pursued as distinct objects (see Scholl, 2001).

Scholl and Pylyshyn (1999) adapted this task to demonstrate that such tracking was still possible even when the items periodically disappeared behind occluders, demonstrating that occlusion is taken into account when computing enduring perceptual objecthood. (Online demonstrations are available at http://pantheon.yale.edu/~bs265/bjs-demos.html .) Moreover, subjects were able to track successfully even when the occluders were invisible, but still functionally present via the accretion and deletion cues along the (invisible) occluders' borders. Crucially, however, such accretion and deletion needed to be along those borders: Performance was significantly impaired when items were present on the visual field at the same times and to the same degrees as in the occlusion conditions, but disappeared and reappeared in ways which did not implicate the presence of occluding surfaces — e.g. by imploding and exploding into and out of existence, instead of accreting and deleting along a fixed contour. (This distinction is exactly Gibson's distinction between going in and out of *sight*, vs. going in and out of *existence*.)

Consider the geometric transformations involved in each case: occlusion and disocclusion involve an affine transformation, whereas implosion and explosion involve a similarity transformation (see Bedford, this issue, for details). In these terms, observers in this experiment could track continuing objects through affine transformations (e.g. occlusion) but not (other types of) similarity transformations (e.g. implosion) — despite the fact that similarity transformations are supposed to be more constrained and thus preferred. Does this result directly conflict with the nested-geometries theory? Perhaps not, since it involves continuous dynamic transformations rather than inferred transformations (as in most of Bedford's other examples). Still, this result highlights another critical spatiotemporal factor which is not embodied in the nested-geometries framework: the local spatiotemporal dynamics of items during brief disappearances help

define what 'counts' as an enduring dynamic visual object. Unlike other types of extra-theoretic factors considered by Bedford (e.g. higher level knowledge), however, I would maintain that this constraint is also a 'core' principle, embedded in the visual system.

6. Conclusions

The nested-geometries model of object-identity proposed by Bedford (this issue) is one of the most ambitious, rigorous, and elegant theories of perception to be proposed in some time. It may help explain many types of (intuitively different) phenomena, using only a small set of basic principles, and it generates several new testable predictions in each of these domains. We should all aspire to such theories.

While none of the arguments I have made here directly conflict with the details of the nested-geometries model, they do attempt to situate it more globally. Bedford implicitly argues for a 'meta-hierarchy' of object-identity principles. In particular, she acknowledges the existence of other object-identity principles based on higher-level knowledge, but she claims that these are more peripheral than the core nested-geometry principles. Here I have suggested that there is another category of object-identity principles based on spatiotemporal information, and that in the 'meta-hierarchy' of object-identity principles, these spatiotemporal biases are even more primary than the nested-geometries discussed by Bedford. I have tried to show how these spatiotemporal biases cannot be subsumed by the nested-geometries model, yet can play a critical role in the perception of object-identity in apparent motion, causal event perception, and object-based attention. Collectively, these examples suggest that the true 'core' of (at least visual-) object-identity judgments may consist of primarily spatiotemporal biases. And because a similar spatiotemporal priority seems to occur in other contexts (e.g. in infant object cognition), this raises the possibility that the underlying mechanisms in each case may be identical (Scholl & Leslie, 1999).

REFERENCES

Bedford, F. L. (2001, this issue). Towards a general law of numerical/object identity. *Cahiers de Psychologie Cognitive/Current Psychology of Cognition*, *20(3)*, 113 - 175.

Burt, P., & Sperling, G. (1981). Time, distance, and feature trade-offs in visual apparent motion. *Psychological Review*, *88*, 171 - 195.

Chen, L. (1985). Topological structure in apparent motion. *Perception*, *14*, 197 - 208.

Dawson, M. (1991). The how and why of what went where in apparent motion: Modeling solutions to the motion correspondence problem. *Psychological Review*, *98*, 569 - 603.

Dawson, M., & Pylyshyn, Z. W. (1988). Natural constraints on apparent motion. In Z. W. Pylyshyn (Ed.), *Computational processes in human vision* (Chapter 5). Ablex.

Green, M., & Odom, J. (1986). Correspondence matching in apparent motion: Evidence for three-dimensional spatial representation. *Science*, *233*, 1427 - 1429.

He, Z., & Nakayama, K. (1994). Apparent motion determined by surface layout not by disparity of three-dimensional distance. *Nature*, *367*, 173 - 175.

Kolers, P., & Pomerantz, J. (1971). Figural change in apparent motion. *Journal of Experimental Psychology*, *87*, 99 - 108.

Mack, A., Klein, L., Hill, J., & Palumbo, D. (1989). Apparent motion: Evidence of the influence of shape, slant, and size on correspondence. *Perception & Psychophysics*, *46*, 201 - 206.

Michotte, A. (1946/1963). *La perception de la causalité*. Louvain: Institut Supérior de Philosophie, 1946. English translation of updated edition by T. Miles & E. Miles, *The perception of causality*, Basic Books, 1963.

Navon, D. (1976). Irrelevance of figural identity for resolving ambiguities in apparent motion. *Journal of Experimental Psychology: Human Perception & Performance*, *2*, 130 - 138.

Prazdny, K., (1986). What variables control (long-range) apparent motion? *Perception*, *15*, 37 - 40.

Pylyshyn, Z. W., & Storm, R. W. (1988). Tracking multiple independent targets: Evidence for a parallel tracking mechanism. *Spatial Vision*, *3*, 179 - 197.

Ramachandran, V. S., & Anstis, S. M. (1983). Perceptual organization in moving patterns. *Nature*, *304*, 829 - 831.

Ramachandran, V. S., Ginsburg, A., & Anstis, S. (1983). Low spatial frequencies dominate apparent motion. *Perception*, *12*, 457 - 461.

Schecter, S., Hochstein, S., & Hillman, P. (1988). Shape similarity and distance disparity as apparent motion correspondence cues. *Vision Research*, *28*, 1013 - 1021.

Scholl, B. J. (2001). Objects and attention: The state of the art. *Cognition, 80(1/2)*, 1 - 46.

Scholl, B. J., & Leslie, A. M. (1999). Explaining the infant's object concept: Beyond the perception/cognition dichotomy. In E. Lepore & Z. Pylyshyn (Eds.), *What is cognitive science?* (pp. 26 - 73). Oxford: Blackwell.

Scholl, B. J., & Nakayama, K. (in press). Causal capture: Contextual effects on the perception of collision events. *Psychological Science*.

Scholl, B. J., & Pylyshyn, Z. W. (1999). Tracking multiple items through occlusion: Clues to visual objecthood. *Cognitive Psychology*, *38*, 259 - 290.

Scholl, B. J., & Tremoulet, P. D. (2000). Perceptual causality and animacy. *Trends in Cognitive Sciences*, *4(8)*, 299 - 309.

Scholl, B. J., Pylyshyn, Z. W., & Feldman, J. (2001). What is a visual object? Evidence from target merging in multiple-object tracking. *Cognition, 80(1/2)*, 159 - 177.

Sekuler, A., & Sekuler, R. (1999). Collisions between moving visual targets: What controls alternative ways of seeing an ambiguous display? *Perception*, *28*, 415 - 432.

Ullman, S. (1979). *The interpretation of visual motion*. Cambridge, MA: MIT Press.