Top-down processes and the visual perception of shape from motion

Ivan Lamouret, Valérie Cornilleau-Pérès and Jacques Droulez

An important challenge of contemporary studies on visual perception is to understand how three-dimensional (3D) shape is recovered from two-dimensional (2D) images. When an object moves, the pattern of retinal velocities provides monocular depth information. This 'kinetic depth effect' (KDE) was discovered about 45 years ago by Wallach and O'Connell[1]. Recently, Sinha and Poggio[2] have provided evidence for a strong influence of learning and memory on the visual analysis of 3D shape from motion.

In the experiments conducted by Sinha and Poggio stimuli consisted in a set of paired wireframes rotating back and forth around a fronto-parallel axis. In each pair, the objects had an identical 2D projection in their median position, but differed in their 3D structure. After 60 seconds viewing of the first object of a given pair, subjects were presented with either the first ('familiar') or the second ('novel') object of the pair for a few seconds. Subjects were asked to report whether this test object appeared as rigid or not. Novel objects were classified as non-rigid about half the time, whereas familiar objects were almost always seen as rigid. This effect was long lasting (it remained for 24 hours after the training) and persisted when the objects were scaled in the test sequence. The effect decreased when the rotation axis was changed, and vanished when the training and test axes differed by 90°. For control subjects who did not undergo training, all objects were rated as equally rigid. The authors interpret their results as an effect of memorizing associations between 2D views and 3D shapes during the visual analysis of 3D shape from motion. When tested with novel objects, subjects associated the median view with the 3D shape of the familiar objects, and the mismatch which occurred between this shape and the subsequent motion pattern resulted in a non-rigid percept.

The idea that memory and knowledge can influence 3D space perception is not new. For example, Helmholtz[3] proposed that 'in many instances it is sufficient to know or assume that the object perceived has a certain regular form, in order to get a correct idea of its material shape from its perspective image as presented to us either by the eye or in an artificial drawing. If the objects portrayed are man's handiwork, such as a house or a table, we may presume that the angles are right angles, and the surfaces are flat or cylindrical or spherical'. However, studies on depth perception from stereopsis[4] or motion[1] demonstrated that object identification is not a prerequisite to the perception of 3D shape. Rather, objects can be isolated from background and identified geometrically, solely from a distribution of binocular disparities, or image velocities. In order to account for this ability of the visual system, theoretical studies of 3D shape from motion have proposed that assumptions about the physical properties of the world such as smoothness of surfaces or the rigidity of objects are integrated in the process[5,6]. These assumptions can be formalized mathematically, and are often expressed as constraints in regularization algorithms[7]. Experiments like the KDE suggest that the rigidity hypothesis is indeed used to some extent by the visual system. Hence, the idea that an a priori knowledge of shape underlies depth perception has been replaced by schemes integrating more generic assumptions about the visual world.

However, rigid objects are not always perceived as rigid. For instance Ames[8] showed that a trapezoidal window tends to be seen as rectangular in monocular viewing, and is perceived as deforming when rotating around a vertical axis. This effect, once taken as an example of the influence of object recognition on perception, has been reinterpreted as a cue conflict between static perspective and kinetic information about 3D shape[9,10]. Sinha and Poggio's important contribution is to demonstrate that beside generic assumptions about the world (that can be portrayed through the use of static depth cues), knowledge about the relationship between a 3D shape and a sequence of 2D views, as acquired during a learning phase, can dominate the rigidity hypothesis. In this sense it demonstrates that the computational approach to the 3D analysis of visual scenes through the use of competing constraints on the physical properties of objects is not sufficient. This analysis can be modified by prolonged presentations of a specific shape and motion, and the modifications are then specific for this configuration (learning effect is reduced when the direction of 3D motion differs for the training and test objects). The paradoxical point here is that a rigidity constraint seems necessary for the establishment of a correlation between 2D successive views and 3D shape during training, and that this correlation then dominates the rigidity constraint during the test phase.

The restriction of the learning effect to the case of similar training and test motion is interpreted by the authors as favouring an implicit viewer-centred over an explicit object-centred coding of 3D shape. In our view, this point needs to be clarified. The concept of viewer- or object-centred representation refers to the referential in which variables are coded, whereas the concept of implicit or explicit coding concerns the nature of the coded variables. The two concepts are independent: for example a viewer-centred depth map is an explicit coding of a 3D structure. Actually Sinha and Poggio seem to discuss mainly the question of implicit versus explicit coding. Indeed if an explicit coding of object shape was learned during training, then learning should transfer to any type of 3D motion in the test phase. Although one cannot discard the possible existence of an explicit object representation further in the process, the effect exhibited here seems to occur at an intermediate step where object shape is not yet explicitly coded.

Independently of the question of shape representation, a major problem raised by this study concerns the concept of reference in non-rigidity ratings. A possible interpretation of Sinha and Poggio's result is that the learned pattern served as a reference for rigidity ratings of similar stimuli (in terms of 2D positions and motions). In this case, judgements of non-rigidity may not reflect a difference in the output of the process (a description of the 3D object), but rather that subjects notice a small departure from the learned baseline. A possible scenario, then, is that when the test configuration differs too much from the learned pattern (in position or motion) the subject uses a different strategy, discarding any a priori knowledge of the object shape. For instance it could be that after prolonged inspection of a rotating object seen under parallel projection, the same object

*I. Lamouret,
V. Cornilleau-Pérès,
and J. Droulez are at
the Laboratoire de
Physiologie de la
Perception et
de l'Action CNRS
Collège de France,
11 place Marcelin
Berthelot, 75005
Paris, France.*

tel: +33 144 27 1624
fax: +33 144 27 1382
e-mail: lamouret@
cdf-lppa.in2p3.fr

1364-6613/97/$17.00 PII: S1364-6613(97)01017-6

seen in polar projection (the configuration that normally elicits the most rigid percept) would be seen as deforming.

Hence Sinha and Poggio's results clearly demonstrate that future psychophysical investigation on the perception of 3D shapes will have to take into account learning processes that can take place on relatively short time scales. More generally they also lead to reconsider classical schemes of 3D shape perception in terms of: (1) the type of object representation involved in visual processes, and (2) the existence of top-down control of these processes.

References

1 Wallach, H. and O'Connell, D.N. (1953) The kinetic depth effect *J. Exp. Psychol.* 45, 205–217

2 Sinha, P. and Poggio, T. (1996) Role of learning in three-dimensional form perception *Nature* 384, 460–463

3 Ittelson, W.H. (1960) Size, shape, perspective, in *Visual Space Perception*, p. 80, Springer

4 Julesz, B. (1971) Cyclopean perception in perspective, in *The Foundation of Cyclopean Perception*, pp. 300–314, The University of Chicago Press

5 Longuet-Higgins, H.C. and Prazdny, K. (1980) The interpretation of a moving retinal image *Proc. R. Soc. London Ser. B* 208, 385–397

6 Ullmann, S. (1979) The interpretation of structure from motion, in *The Interpretation of Visual Motion*, pp. 133–175, MIT Press

7 Poggio, T., Torre, V. and Koch, K. (1985) Computational vision and regularization theory *Nature* 317, 314–319

8 Ames, A. (1951) Visual perception and the rotating trapezoidal window *Psychol. Monogr.* 65, 1–32

9 Mingolla, E. and Todd, J. (1981) The rotating square illusion *Percept. Psychophys.* 29, 487–492

10 Cornilleau-Pérès, V., Marin, E. and Droulez, J. (1996) The dominance of static depth cues over motion parallax in the perception of surface orientation *Perception* 25 (suppl.), 40

Response from Sinha and Poggio

P. Sinha and T. Poggio are at E25-201, Center for Biological and Computational Learning, Department of Brain and Cognitive Sciences, MIT, 45 Carelton Street, Cambridge, MA 02142, USA.

tel: +1 617 253 0547
fax: +1 617 253 2694
e-mail: sinha@ai.mit.edu

Lamouret, Cornilleau-Pérès and Droulez raise a number of very interesting points in their **Comment** article. We should like to take this opportunity to emphasize one of these issues that we find particularly important and intriguing but could not dwell upon adequately in the original paper, for lack of space.

Lamouret *et al.* remark on how learning initially depends on bottom-up sensory-information processing that uses generic biases such as those favoring object rigidity. However, the learning subsequently can overwhelm the results of such bottom-up processing. The percept, apparently, is controlled to different extents at different times by the generically processed sensory information on the one hand and object-specific learned expectations on the other. The big question is: How does the brain strike a compromise between sensation and, for want of a better term, hallucination? The parameters determining the relative contributions of the two quantities to the overall percept are likely to be a function of time in two ways. (1) Expectations will exercise greater control in determining percepts the longer the training time. (2) The bottom-up sensory information will become increasingly evident the greater the stimulus inspection time. The well-known hollow-mask illusion serves as a nice illustration of this point. The illusion often persists even under binocular viewing. If one subscribes to the accounts of the illusion that are based on familiarity, then it is reasonable to suggest that the greater the familiarity of an observer with faces, the more susceptible the observer will be to perceiving the illusion. On the other hand, the longer one binocularly inspects the hollow mask, the more likely one is to perceive its correct (hollow) structure. Our experimental results follow a similar pattern. The key question that needs to be addressed to explain these empirical observations is how expectations are combined with sensory information to yield the overall percept. It seems to be a rather involved question, given that the combination strategy is a function of at least two temporal variables. Among others, it is likely to prove interesting to colleagues who have been studying so fruitfully the issue of cue-combination, except that one of the cues would now be 'internal' to the visual system. Work on this problem holds the exciting potential of bringing together two big, and so far largely independent, streams of research – one examining 'bottom-up' processing and the other 'top-down' strategies.

On another issue, Lamouret *et al.* correctly point out that the shape representation schemes we discuss are better characterized as implicit versus explicit, with emphasis on the nature of the coded variables. The visual system might also possess some limited ability to extract viewer-centered depth information, which, though an 'explicit' encoding of shape, cannot readily be subjected to arbitrary projectional transformations.

Lamouret *et al.* deserve thanks for summarizing our results so clearly and for suggesting and highlighting some of the important issues that need to be tackled next.

Higher-order processes in auditory-change detection

Risto Näätänen and Kimmo Alho

R. Näätänen and K. Alho are at the Cognitive Brain Research Unit, Department of Psychology, University of Helsinki, PO Box 13 (Meritullinkatu 1), FIN 00014, Finland.

tel: +358 9 1912 3445
fax: +358 9 1912 2924
e-mail: risto.naatanen@helsinki.fi

The paper by Schröger and Wolff[1] is, perhaps, the first study that has clearly succeeded in demonstrating what is memory-related and what is memory-unrelated (as we interpret the results) in the enhancement of an electric brain response to an infrequent stimulus change. In this study, a sound (the 'standard') with a certain apparent location (manipulated by the interaural time difference) was repeated at short intervals, and was occasionally replaced by an identical sound, which had a slightly different apparent location (the 'deviant'), whilst the attention of the subject under investigation was directed elsewhere. These deviants elicited an event-related potential (ERP), which was enhanced relative to that elicited by the standard. This enhancement emerged as a negative shift, at the time region of 100–250 ms from stimulus onset, in the deviant–standard difference wave.

To account for this enhancement, firstly one needs to consider the fact that the sound-location specific afferent

Copyright © 1997, Elsevier Science Ltd. All rights reserved. 1364-6613/97/$17.00 PII: S1364-6613(97)01013-9