PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Perceptual categories for spatial layout

Daniel Kersten

Phil. Trans. R. Soc. Lond. B 1997 **352**, 1155-1163
doi: 10.1098/rstb.1997.0099

Perceptual categories for spatial layout

DANIEL KERSTEN

N218 Elliott Hall, Psychology Department, University of Minnesota, 75 East River Road, Minneapolis, MN 55455, USA

SUMMARY

The central problems of vision are often divided into object identification and localization. Object identification, at least at fine levels of discrimination, may require the application of top-down knowledge to resolve ambiguous image information. Utilizing top-down knowledge, however, may require the initial rapid access of abstract object categories based on low-level image cues. Does object localization require a different set of operating principles than object identification or is category determination also part of the perception of depth and spatial layout? Three-dimensional graphics movies of objects and their cast shadows are used to argue that identifying perceptual categories is important for determining the relative depths of objects. Processes that can identify the causal class (e.g. the kind of material) that generates the image data can provide information to determine the spatial relationships between surfaces. Changes in the blurriness of an edge may be characteristically associated with shadows caused by relative motion between two surfaces. The early identification of abstract events such as moving object/shadow pairs may also be important for depth from shadows. Knowledge of how correlated motion in the image relates to an object and its shadow may provide a reliable cue to access such event categories.

1. INTRODUCTION

Determining the identity and spatial location of objects are often cited as the two great divisions of problems in vision. The challenge in either division is how to deal with the enormous variability in the image for various occurrences of an object, or spatial arrangement of objects. Much recent work in object recognition has focused on how to cope with variation in viewpoint (Biederman 1987; Bülthoff & Edelman 1992; Tarr & Bülthoff 1995) and in illumination (Adini *et al*. 1995; Hallinan 1995; Tarr *et al*. 1997). But vision has to deal with more than variability in illumination and viewpoint—due to the fact that the image is a complex function of *all* of the scene variables. A particular visual task requires the estimation of useful subsets of scene parameters, leaving the remainder as confounding variables. Unravelling the causes seems even more complex when one realizes that some scene causes, such as light source position or the location of a shadow casting object, can produce large changes distant from the corresponding region in the image, in contrast to other changes such as shape or material which produce primarily local changes in the corresponding region of the image. Dealing with variability is also difficult because almost any given image measurement seems to be ambiguous about its cause in the scene.

A careful analysis of visual task requirements provides part of the solution to the dilemma of complexity and ambiguity. Specifically, one should seek out image measurements that are appropriate to the visual task at its level of abstraction. This notion is familiar to theories of recognition, although robust image measurements supporting task-specific classification have been hard to come by. For example, finding an object 'key' is a crucial first step in the recognition of specific objects using alignment (Ullman 1996), and serves to select potential object models to test whether they can be transformed to fit the image data. The invariant features of 'geons' were proposed as reliable indicators of the qualitative shape of parts and their relationships, well-suited to entry-level recognition of object classes rather than specific instances (Biederman 1987).

Cavanagh (1991) has suggested that although edge maps confound shadows with illumination-invariant features, they may nevertheless be reliable enough to access the prototype for a human face, with that and other knowledge subsequently brought to bear to discern the specific face. Schyns & Oliva (1994) have suggested that coarse scale spatial information may provide diagnostic information for scene recognition. One consequence of such strategies is that not all image edges have to be parsed into their respective causes in the scene (i.e. whether they are shadows, specularities, material changes or occlusions) for useful visual decisions. This has led to a revival of appearance-based algorithms in computer vision (Murase & Nayar 1995). In general, we see a common theme that certain image measurements provide useful features for object categories at different levels of abstraction. A firm decision about a perceptual category has the potential to reduce ambiguity which can then be resolved with other sources of knowledge. Are categories also useful for determining object location?

Spatial relationships between objects can be specified by visual direction and depth. Visual direction is related to retinal position in a straightforward way whereas depth is less direct. Vision uses depth information both for orientation and shape within surfaces (derivatives of depth), as well as relative position between surfaces. In this paper we consider depth between surfaces. Both stereo and motion parallax provide information about relative depth in terms of differences in retinal positions over space and time, but perception is also able to make reliable inferences about spatial relationships from the pictorial cues. Human vision shows interactions between a number of pictorial cues and the more 'direct' information from stereo and motion. These interactions have been documented for occlusion and stereo (Nakayama *et al*. 1989), transparency and stereo (Trueswell & Hayhoe 1993), and transparency and structure from motion (Kersten *et al*. 1992). For shape, there are interactions between stereo and shading (Bülthoff & Mallot 1988), texture and stereo (Johnston *et al*. 1994), and texture and motion (Landy *et al*. 1995). Below we look in detail at the interaction between cast shadow information for depth and motion between surfaces, and show that reliable inferences of depth require the resolution of ambiguity of the causes of intensity change. It is argued that this resolution, like that for object recognition, may also take advantage of image measurements that are reliable for perceptual categories at various levels of abstraction. In particular, we will look at how the identification of material category is related to shadow labelling, and how identifying object/shadow event categories may explain the robustness of perception of depth from shadows.

Below, we will first examine the relationship between scene complexity and tasks, and in particular get an overview of how a visual task determines the division of the explicit and implicit (or generic) variables that contribute to the image. Here the importance of robust image features for a visual task will be stressed. Then, several experiments involving the perception of depth from moving cast shadows will be described. An analysis of the visual decisions to resolve shadow ambiguity illustrate the theoretical ambiguities in apparently unambiguous scenes. The subsequent section will argue that both local and global image features determine the perceptual categories that are useful for determining object trajectory using cast shadow information.

2. SCENE COMPLEXITY AND VISUAL TASKS

In addition to determining object identity and location, seeing solves other problems such as identifying object material properties, object classes, configurations of objects and events. In doing this, vision does more than discount variability in viewpoint and illumination. Different visual tasks require a more explicit representation of some classes of scene parameters than others. Because the image at the eye is a function of all of the scene variables, it is important to consider both the variables relevant to the visual task (the 'explicit variables'), as well as the confounding variables not relevant, but which nonetheless contribute to variability in the images received (the 'generic variables', cf. Freeman (1994)). Identifying the appropriate scene parameters to estimate, and those to discount constrains the kinds of image measurements that are reliable for those explicit variables. One classification of scene variables that are useful for several tasks is shown in figure 1. The choice of classes of scene parameters is characteristic of those found in three-dimensional (3D) computer graphics programs (Kersten 1997). A typical computer graphics system would employ an object modelling module that could have the following three components: a shape editor to specify the geometry; a material editor to specify what the object is made of at a texture scale finer than the object, and articulation tools to control, for example, the joints in the human body. Articulation parameters could also include those required to model variation over facial expression, or the shape changes in a soft-sided brief-case. A computer graphics system would also have separate components to position the objects relative to each other, and set the camera and light positions. It is no doubt a matter of debate as how best to partition and prioritize the scene variables, but the classes used by graphics animators seem to provide a natural and useful classification.

Vision's job, of course, is to do something with the image data received. In particular, the job of visual perception is to make explicit the scene parameters appropriate for a task, despite the variations in the image caused by the generic variables. Two broad distinctions are tasks involving single objects (object perception) and tasks requiring information about the relationships between objects and/or the viewer (figure 1). An example of a subtask for object perception (discussed above) is entry-level recognition, where shape may be the crucial information to make explicit (Biederman 1987). Other contributions such as material or articulation variation need to be discounted at some level, perhaps only to be explicitly estimated if finer-level discriminations are required. Spatial layout tasks can be distinguished based on whether an observer-centred or world-centred representation is most useful. Spatial layout subtasks for which viewpoint is generic include spatial planning and scene recognition, and it is in this context that we will discuss depth from cast shadows.

In actual practice, the division between explicit and generic may not be entirely clear for a task. For example, material category may be important for some entry-level recognition decisions. Further, material classification is a useful task in its own right (e.g. picking matching clothing). One can soften the boundaries between explicit and generic variables using statistical decision theory to weight differentially the loss of getting bad estimates of the scene parameters (Yuille & Bülthoff 1993; Freeman & Brainard 1995). The extent to which human vision actually achieves invariance over a generic variable is an experimental question. Departures from ideal invariance provide clues as to the mechanisms of perception (e.g. viewpoint and illumination dependency in object

	visual tasks image data = f(shape, articulation, material, illumination, viewpoint, relative position)				
	object perception		spatial layout		
	object-centred		world-centred	observer-centred	
	recognition		scene recognition/ spatial planning	action	
	entry-level	subordinate-level		reach	grasp
(*a*) explicit variables	shape	shape material articulation?	relative position	viewpoint	shape articulation
(*b*) generic variables	articulation material illumination viewpoint relative position	illumination viewpoint relative position	shape articulation material illumination viewpoint	shape articulation material illumination relative position	material illumination viewpoint relative position

Figure 1. Task-dependent classification of scene variables.

recognition (Bülthoff & Edelman 1992; Tarr & Bülthoff 1995; Tarr *et al*. 1997).

Note that the one variable that is generic across all tasks is illumination. A consequence of this is that a general-purpose visual system can begin to discount some of the effects of illumination, such as variation over the mean level, right at the input. Other effects, such as shadows or highlights, may be more difficult to discount, and may in fact be useful (Tarr *et al*. 1997). Earlier we noted that viewpoint variation poses particularly challenging problems for object recognition. But viewpoint is also a problem for certain aspects of depth perception. Viewpoint is an explicit variable for determining viewer–object relations, but a generic variable for determining relationships between objects, as well as for shape. Later we will exploit the genericity of viewpoint in an analysis of what may constitute a reliable feature for determining the presence of an object and its cast shadow.

(*a*) *Perceptual categories*

A profound problem of vision is how knowledge about image behaviour can be integrated with knowledge regarding scene structure to enable reliable inferences. Computationally, we now know how to integrate image constraints and prior knowledge for the fine-grained estimation of scene parameters for certain well-defined vision tasks (e.g. surface orientation, optic flow and reflectance, (cf. Clark & Yuille *et al*. 1990). Perhaps a more difficult problem is to understand how to identify more abstract classes or categories of scene properties from image measurements. The power and flexibility of human vision may arise through its ability to organize prior knowledge of the world into useful categories, and make reliable inferences about these categories at multiple levels of abstraction.

Categories can be defined in terms of a set of instances that are associated with a common label. A crucial decision is whether to represent instances in terms of image features or scene properties. Here we assume categories are defined in terms of scene properties, because it is the properties of scenes that are useful for visual function. An economical way of representing a set of instances is in terms of a prototype together with some rules for allowable transformations. An eye can be characterized in terms of the mean values of a set of spline points representing explicit variables for points of high curvature, together with some model of variation about those points (Yuille *et al*. 1988). For the purposes of this paper, I would like to use an intuitive notion of a category as a label that tags a range of scene-variable properties together with some rules that specify how that category relates to other aspects of the scene. Categories useful for depth could include opaque surface, transparent surface, shiny surface and cast shadow. An opaque surface can have any reflectance value, but the critical property is that it occludes light from surfaces that are behind. A cast shadow is darker than the surround, should be coplanar with its receiving surface, and have a companion casting surface. There are also categories of events (Jepson *et al*. 1996) that may be useful for depth perception. Examples include states of rest or motion of an object, the interaction between objects, immanent contact with the viewer or a moving object/shadow pair.

For a perceptual category to be useful for vision, there must be image measurements or features that reliably support the inference of the category. The features indexing the category must be robust with respect to the generic variables in the sense that large changes in the generic variables should have little effect on the image features supporting the category hypothesis (cf. Freeman 1994; MacKay 1992). An important aspect of categories that will not be

addressed is their overall structure, which involves a balanced trade-off between maximizing the differences between classes, while minimizing the differences within a class (Bobick 1987).

3. DEPTH FROM MOVING CAST SHADOWS

We are usually sure (enough) of what we see—whether an object is in front or behind another, or whether it is headed this way or that way. This section describes simple computer-synthesized movies of 3D scenes involving moving objects and their cast shadows. The movies usually have an unequivocal perceptual interpretation, despite the fact that, in theory, there are alternative scene constructions that could have produced the same image data. An advantage of computer graphics is that while the images are realistic and complex, there is sufficient simplicity to analyse the theoretical ambiguities of material, location and movement.

(*a*) *Square-over-checkerboard shadow movies*

The first movies consisted of a stationary central square in front of a checkerboard, and a shadow moving diagonally and away from and then towards the square (figure 2, top panel; Kersten *et al*. (1996)) Using orthographic projection, the size and position of the square were fixed in time relative to the background. By keeping the square a constant size in the image, the motion cues of object expansion and contraction indicated no motion in depth. Assuming a general viewpoint, the lack of any translational motion of the square in the image provided further evidence that the square was stationary in 3D. In fact, requiring robustness over viewpoint as a generic variable says that the square should be coplanar with the checkerboard. The cast shadow was generated either with an extended light source (like a fluorescent panel) or with a point source. An extended light source at a finite location produces a penumbra that gets fuzzier as the square gets further away from the background.

Despite the presence of strong cues to stationarity, all observers who viewed the movie with a fuzzy shadow reported an initial strong perception of a square moving in depth. When the shadow boundary was sharp, some observers reported seeing the target square move back and forth in depth; other observers report seeing a dark patch sliding back and forth across the checkerboard background. Below, I will return to the reasons for this perceptual ambiguity. There are several conditions that facilitate apparent motion in depth: (i) a penumbra which gets more blurry (as the square rises off the surface) is better than a sharp shadow; (ii) a dark shadow is better than a (physically unnatural) light one (Kersten *et al*. 1997); and (iii) a shadow below the square is better than one above the square (Kersten *et al*. 1996). We will return to the importance of a dynamically changing shadow penumbra.

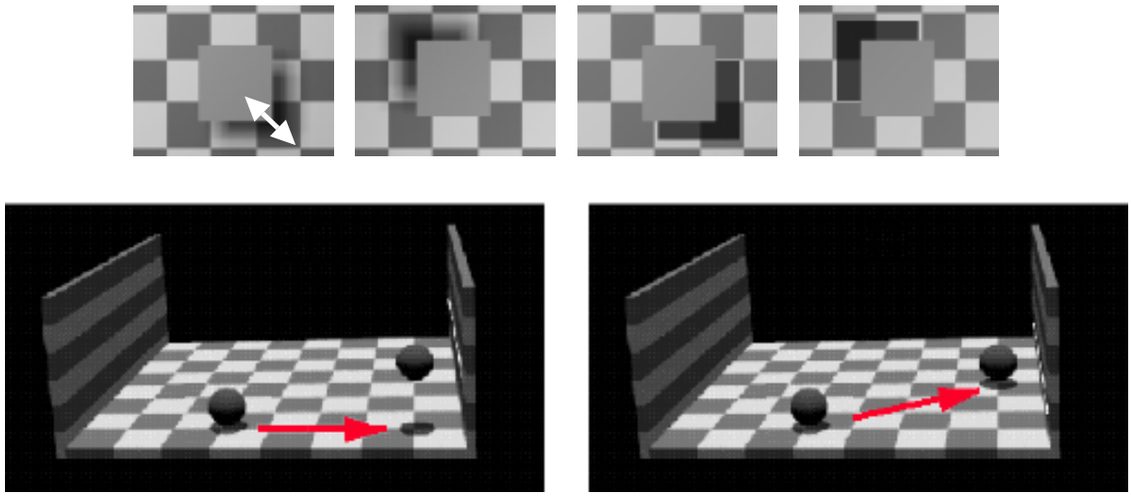

Figure 2. The top panel illustrates the four conditions for the square-over-checkerboard experiment in which the shadow was (from left to right): (i) blurry and below the square; (ii) blurry and above; (iii) sharp and below; and (iv) sharp and above (details in Kersten *et al*. (1996)). The central square was stationary, while the shadow moved diagonally as illustrated in the upper left. Observers report the central square to apparently move away from the checkerboard as the shadow moves away from the square. The percept is strongest for a blurry shadow below the square (condition on the upper left) (adapted from Kersten *et al*. 1996). The bottom panel illustrates the two main conditions of the ball-in-a-box movies. The ball repeated its motion back and forth between its left- and right-most positions in the image. The shadow moved so that it remained vertically below the ball in the image. Only the distance between the shadow and the ball varied as the shadow and ball moved. The shadow had either a horizontal trajectory (lower left) or a diagonal trajectory (lower right). The ball had a fixed size in the image and the same trajectory in both cases. Observers were asked to indicated the ball's height at the right-most portion of the trajectory using a cursor on the right-hand wall (details in Kersten *et al*. (1997)). Observers report seeing the ball rise above the floor of the box in a fronto-parallel plane for the horizontal shadow trajectory (lower left), and move across the floor to the back of the back for the diagonal shadow trajectory (lower right).

(*b*) *The ball-in-a-box shadow movie*

The square-over-checkerboard experiment provided a strong test of the strength of moving cast shadows as compared with motion cues for depth change. The fact that sharp shadows were not as effective as fuzzy shadows suggests that the general viewpoint constraint was strong enough to rule out the assignment of 'shadow' category to the dark patch. Therefore, we might find an even more robust effect of moving cast shadows if we eliminate the accidental view. This was done in 3D graphics simulations (figure 2, lower panel), in which a ball moved inside a box in such a way that it followed a diagonal trajectory in the image plane (Kersten *et al*. 1997). The shadow boundary was always sharp (from a point light source at infinity). The size of the ball's image remained fixed throughout the movie. There were two different movie sequences. In the first, the ball's cast shadow followed a horizontal trajectory in the image; in the second, it followed a diagonal trajectory identical to that of the ball's image. Despite the fact that the ball's image remained the same size and had an identical trajectory in the image plane in both movies, all observers saw the ball rise above the checkerboard floor when the shadow trajectory was horizontal, and recede smoothly in depth along the floor when the slope of the shadow trajectory matched that of the ball. In several experiments Kersten *et al*. (1997) found that (i) the observer's settings of apparent ball height are consistent with a fixed, but fictitious light source position which varied between individuals; (ii) nonlinear shadow motion can induce an apparent nonlinear ball trajectory; and (iii) changing shadow shape, opacity or contrast do not measurably affect motion in depth from moving 'shadows'.

The last point is illustrated in figure 3. Several manipulations of shadow contrast and opacity had no effect on observers' settings of apparent height of the ball as a function of shadow trajectory. These observations stand in contrast to those obtained for the interpretation of shadows in static images, which show that similar manipulations of shadow brightness and contrast strongly interfere with shape perception (Cavanagh & Leclerc 1989). Why are the ball-in-a-box percepts of motion in depth so robust over variations in shadow properties? We return to this question later. But first, let us examine the theoretical ambiguities present in the square-over-checkerboard movie.

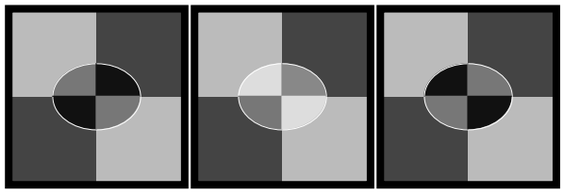

Figure 3. The original, natural shadow (left), followed by two shadow substitutions, a (physically unnatural) light 'shadow', and a patch with the opposite contrast of a transparent surface or shadow. The X-junction intensity relations on the left-most panel are consistent with natural shadows. The shadow substitutions behave like natural shadows and have no measurable effect on observers' settings of the apparent height in the ball-in-a-box experiments (figure 2, bottom panel).

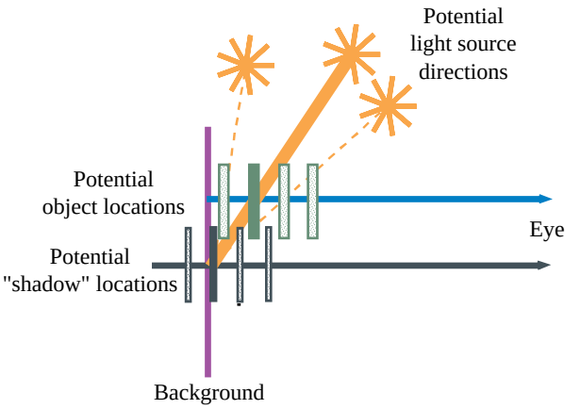

Figure 4. The possible spatial relationships between background, target square, dark 'shadow' patch and light source at a given moment. Without knowledge of the material properties of the 'shadow' and background (the dark 'shadow' patch could be due to an opaque surface behind a transparent background or an opaque surface in front of the background), the position in depth is ambiguous (thin stippled rectangles). If the dark region is known to be a shadow, then it is constrained to lie on the background (thin solid rectangle). Even if the position of the shadow is known, the position of the target square and light source direction are ambiguous (stippled rectangles and dashed lines). A commitment to the light source direction (solid line) specifies a depth for the target square (thick solid rectangle). Conversely, knowledge of the depth of the target square constrains the light source direction. (Reprinted from Kersten (1997).)

4. KNOWLEDGE REQUIRED TO RESOLVE SCENE AMBIGUITY

For a given individual, there is little subjective ambiguity about the trajectory of the objects in the shadow movies. What is not immediately apparent is that even when a scene is complex enough to provide a rich set of cues for spatial layout, there is a large set of ambiguities that must be resolved. How does perception come by its confidence?

Let us take a closer look at alternative physical scenes that could have produced the square-over-checkerboard movie (figure 4). Consider first the kinematic information, and note that shadow displacement can either be caused by movement of the light source or of the target square. One way to resolve this ambiguity is to incorporate *a priori* knowledge that light sources are much less likely to be moving than the objects they illuminate—a 'stationary light source constraint'. Some other interpretations are contingent on knowledge regarding material properties of regions in the image. For example, the dark shadow-like square could instead be a transparent surface film in front of the checkerboard, which is what observers often report when the shadow has a sharp edge. The image is also consistent with a scene in which the dark square, rather than being a shadow, is an opaque surface behind a transparent checkerboard. Even the change

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

in the degree of blur at the shadow edge, instead of being a penumbra, could, in theory, be due to the dark square falling outside the depth-of-field of the eye (no observers have reported this percept). A visual decision regarding the material (transparent, opaque surface or shadow) affects more or less how the motions are seen. Kersten *et al*. (1992) reported a movie with a bistable percept of apparent rigid or non-rigid motion of two square planar overlapping surfaces. The type of motion perceived (rigid or non-rigid) was contingent on which of the two surfaces appeared to be transparent.

In the next couple of sections we will take a closer look at local and global information that support the resolution of shadow ambiguity. But first, let us consider the visual decisions regarding shadow identification in a broader context of spatial layout.

(*a*) *Spatial layout: levels of abstraction, categories and visual decisions*

Think of a complex scene as a play at the theatre. Two main components to sort out are the stage set (context) and players (objects). At any given moment there is usually one player in the lead which captures the focus of attention. There are also light sources, visible props and invisible stage-hands. The first questions address categorical decisions regarding the properties and general identities of the players. To be specific, consider the square-over-checkerboard.

(i) *Context decision*

One of the first decisions to be made is: which of the objects provides the context with which to interpret the spatial layout of the other objects? which is the stage? This question is directly related to the visual task. A particularly critical choice is whether the objects should be represented in a world-centred or observer-centred frame of reference. Viewpoint is a generic variable for deciding whether the square is headed away from the checkerboard. Viewpoint is an explicit variable if the task is to reach to the square. Let us assume that the checkerboard provides a world-centred frame of reference useful for spatial planning. Note that the decision to measure motion with respect to the checkerboard is not without theoretical ambiguity. Perceptually the checkerboard is a suitable stationary frame of reference; however, under orthographic projection, it could have moved along the line of sight while the target square remained stationary.

(ii) *Player category decision*

The most salient players are the central square, the checkerboard background and especially if seen as an independent 'thing' itself, a dark transparent patch. These are the explicit 'players' ordered according to my guess of salience, where the central square is the lead player. But there are implicit elements ('stage props') as well. Phenomenally, a fuzzy faint shadow can influence scene structure, but be barely noticeable—after all, it is not a surface inviting action or attention. Light sources are not usually explicit players, but their stage properties certainly are manifest in how images are interpreted (a shadow above the object is less effective in producing apparent 3D motion than a shadow below). Below, we discuss the image measurements that support decisions about the object material category.

(iii) *Motion event category decision*

Given the checkerboard as reference frame, the visual system can make some categorical decisions about the motion class itself. We have several possible categories of motion to choose from for the moving square. Natural choices would be stationary, constrained to move within the plane of the checkerboard or moving in depth. Image information would seem to rule out all but the first, but as we have seen, the third is also seen. How do we choose the appropriate event category? A measurement of zero object motion relative to a background is insufficient to correctly categorize the motion class of the square and shadow.

(iv) *Parameter estimation: where are the objects headed and how fast?*

Finally, a visual decision can be made as to where the square is headed and how fast. In other words, given a motion category or model, what values should be fit to the parameters?

We could describe the last two types of visual decisions as (i) finding an appropriate model to describe the category of motion and (ii) fitting appropriate parameters to that model. Different types of measurements reliably support different kinds of decisions depending on the level of abstraction. What kinds of image measurements can be used to support the decision regarding the player category, motion category or velocity? The next section discusses the cues that support shadow identity. The subsequent section discusses a global cue that may be important for determining the event category of object/shadow pair. Finally, the information required for velocity estimation is discussed.

(*b*) *Shadow and material categories for depth: square-over-checkerboard*

Kersten *et al*. (1997) discuss a number of local cues that support shadow identification. For the square-over-checkerboard movies, there was a particularly diagnostic cue in the dynamically changing penumbral blur caused by an extended light source. Such a local image measurement has less ambiguity with other scene causes (e.g. it is likely to be confused with a material change, although it could result from surface edge motion out of the depth-of-field range, or a spreading stain). This cue is also robust over viewpoint and a large range of types of illumination. In contrast, the sharp shadow is often seen as a transparent surface—a decision supported by transparency constraints (Metelli 1975).

Even if a shadow has been identified, vision faces the problem of determining which surface it belongs to.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

This problem raises the possibility that vision may sometimes make the decision that a particular event is occurring and then determine the target object's motion. Shadow identification may only be implicit. The findings from the ball-in-a-box experiments illustrate this argument.

(*c*) *Event categories for depth: ball-in-a-box*

Above we noted that apparent motion of the stationary square is sensitive to the specifics of shadow properties. In contrast, for the ball-in-a-box, an object's cast shadow does not have to be physically reasonable—it can have the wrong contrast polarity or lightness—for observers to see consistently different motions in depth which depend on shadow trajectory. Further although the shadow had a sharp edge, no observers ever reported the shadow to appear as an independent surface patch coincidentally moving below the ball. Why do the ball-in-a-box demonstrations produce a strong percept of motion in depth, even when several properties of shadows such as contrast polarity, and transparency are wrong? The answer may be that the detection of a particular kind of correlated motion provides diagnostic information that an event category corresponding to an object/shadow pair has occurred. The informativeness of correlated motion depends on the assumptions of a stationary light source and a general viewpoint. A stationary light source implies that the line connecting the shadow and object always terminates at the fixed light source. This constraint is preserved in the image projection and thus defines the correlated motion. If the light source is at infinity (as in the ball-in-a-box movies), the line makes a fixed angle in the image. If the light source is at a finite position, the line sweeps through an angle anchored to a fixed location in the image plane (figure 5).

The interpretation of this correlated motion is robust over viewpoint. There are competing events (e.g. the ends of a stick moving rigidly in space) which could under some circumstances (swinging pendulum) mimic object/shadows, which seem less likely, but may pose some ambiguity. Although correlated motion may provide a low-ambiguity initial index to object/shadow pair, knowledge of this event does not indicate which image patch corresponds to the object and which to the shadow, nor does it specify where the object is headed. Shadow identity could be resolved by assuming that the light source is above, and thus the lower region is the shadow. But we still require knowledge of the shadow location to compute a unique depth trajectory.

(*d*) *Where is the ball heading? parameter estimation*

Figure 6 illustrates some of the geometric ambiguities present in the ball-in-a-box movies. Three pieces of information that constrain the ball's location are (i) the location of the shadow; (ii) the direction of the light source; and (iii) the viewpoint. The first two pieces of information constrain the ball to be on a line between the shadow and light source. The third specifies a line from the eye through the ball. The intersection of these two lines could, in theory, determine the ball's apparent position. Kersten *et al.* (1997) showed that the visual system assumes a specific fixed

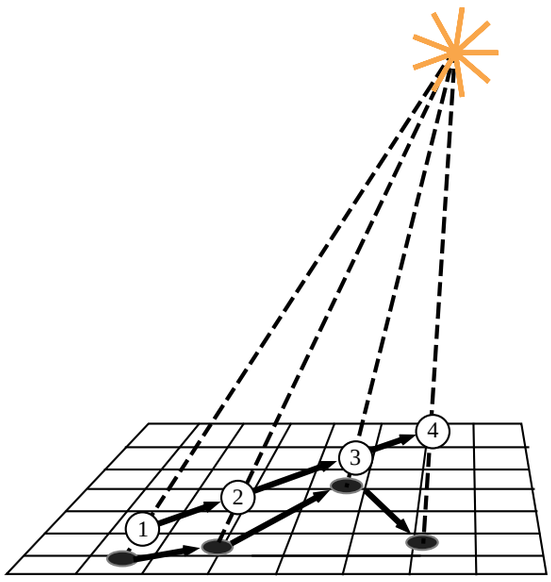

Figure 5. The correlated motion between an object and its shadow. A fixed light source constrains the image of the ball and shadow to lie along a straight line anchored at a point which may be in the image, or beyond (as with the infinite light source used in the ball-in-a-box simulations). When observers view a smooth animation of a ball moving through points 1, 2, 3 and 4, the ball first appears to rise above the ground plane and move away from the viewer (points 1, 2 and 3); however, at point 3, the ball appears to bounce back towards the viewer. This apparent nonlinear trajectory persists even though the trajectory of the ball has a constant velocity in the image, and the ball has a constant image size.

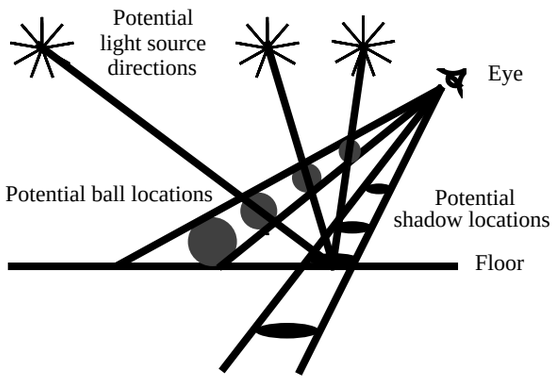

Figure 6. The positional ambiguities in the ball-in-box shadow scene at a given moment in time. Once the light source direction is fixed, and the dark region is categorized as a shadow (and therefore on the floor of the box), the height of the ball above the floor is determined. As long as the light source is assumed to remain fixed, the trajectory of the ball is determined over time. (Reprinted from Kersten *et al.* (1997).)

(albeit incorrect) light source position. But what information determines the location of the shadow? Local photometric constraints could contribute to labelling a region as a shadow, which is by necessity on the receiving surface. But photometrically wrong shadows have no significant effect on the apparent trajectory. Another source of information is the non-accidental alignment of the canonical axis of the shadow patch with that of the floor. An economical explanation for this coincidence is that the shadow and floor are coplanar (cf. Richards *et al*. 1996), thus resolving the remaining ambiguity required to compute the position and trajectory.

5. DISCUSSION

(*a*) *Knowledge required for spatial layout*

We have used three types of knowledge to resolve ambiguity in the perception of depth from shadows: (i) prior assumptions (e.g. stationary light source) provide high probability default constraints; (ii) image information constrains the category of a local image patch or motion event type (changing penumbral blur, correlated motion); and (iii) the visual task determines the category as well as the generic variables over which perception's model of the image data should be robust (e.g. correlated motion with respect to viewpoint). More specifically, vision may resolve ambiguity in spatial layout by incorporating knowledge at several different levels of abstraction defined by categorical structure. Categories would serve several functions. First, because they are task-specific, category decisions provide the means to reduce the complexity of treating vision as a full-blown inverse optics problem in which one explicitly estimates all of the scene parameters. Second, categories abstract properties (e.g. opaqueness or shadow) which are sufficient to determine 'player roles' and thus relative depth relationships. Exact values of reflectance or transparency are not needed for such inferences. Third, it makes sense to use image information to support a reliable high-level category decision (e.g. correlated motion for an object/shadow event), when such information is not immediately (i.e. bottom-up) and reliably available for the estimation of specific parameters (e.g. 3D velocity of the ball).

(*b*) *Confidence-driven perceptual decisions*

Thinking in terms of the reliability of image measurements for visual decisions regarding categories cuts across the usual debate of bottom-up versus top-down. Instead, one asks what image measurements are 'good features' for various classes of categories. A strong hypothesis is perceptual decisions are 'confidence-driven' and it is the quality of the image data that determines the level of abstraction that is first accessed. At any given time, vision has a set of image measurements and a set of hypotheses spread over various levels of abstraction. Fast visual decisions (bottom-up) are those for which the probability of a category conditional on the image feature is sufficiently high as compared with the probabilities of the alternatives at that level. A category commitment signals a firm conclusion upon which early (less abstract) but more ambiguous decisions can subsequently rely via the top-down flow of information.

This work was supported by the National Science Foundation (BNS-9109514) and NIH (RO1 EY11507-001).

REFERENCES

Adini, Y., Moses, Y. & Ullman, S. 1995 *Face recognition: the problem of compensating for changes in illumination direction*. The Weizmann Institute of Science.

Biederman, I. 1987 Recognition-by-components: a theory of human image understanding. *Psychol. Rev.* **94**, 115–147.

Bobick, A. F. 1987 Natural object categorization. Technical report **1001**. Cambridge, MA: MIT Artificial Intelligence Laboratory.

Bülthoff, H. H. & Edelman, S. 1992 Psychophysical support for a two-dimensional view interpolation theory of object recognition. *Proc. Natn. Acad. Sci. USA* **89**, 60–64.

Bülthoff, H. H. & Mallot, H. A. 1988 Integration of depth modules: stereo and shading. *J. Opt. Soc. Am.* A **5**, 1749–1758.

Cavanagh, P. 1991 What's up in top-down processing? In *Representations of vision: trends and tacit assumptions in vision research* (ed. A. Gorea), pp. 295–304. Cambridge University Press.

Cavanagh, P. & Leclerc, Y. G. 1989 Shape from shadows. *J. Exp. Psychol., Human Perception and Performance* **15**, 327.

Clark, J. J. & Yuille, A. L. 1990 *Data fusion for sensory information processing*. Boston: Kluwer Academic Publishers.

Freeman, W. T. 1994 The generic viewpoint assumption in a framework for visual perception. *Nature* **368**, 542–545.

Freeman, W. T. & Brainard, D. H. 1995 Bayesian decision theory, the maximum local mass estimate, and color constancy. *Proc. Fifth Int. Conf. Computer Vision*, pp. 210–217.

Hallinan, P. W. 1995 A deformable model for the recognition of human faces under arbitrary illumination. Ph.D. thesis, Division of Applied Sciences, Harvard University.

Jepson, A., Richards, W. & Knill, D. 1996 Modal structure and reliable inference. In *Perception as Bayesian inference* (ed. D. C. Knill & W. W. Richards), pp. 63–92. Cambridge University Press.

Johnston, E. B., Cumming, B. G. & Landy, M. S. 1994 Integration of stereopsis and motion shape cues. *Vision Res.* **34**, 2259–2275.

Kersten, D. 1997 Inverse 3D graphics: a metaphor for visual perception. *Behav. Res. Meth. Instr. Comp.* **29**, 37–46.

Kersten, D., Mamassian, P. & Knill, D. C. 1997 Moving cast shadows induce apparent motion in depth. *Perception* **26**, 171–192.

Kersten, D., Bülthoff, H. H., Schwartz, B. & Kurtz, K. 1992 Interaction between transparency and structure from motion. *Neural Comput.* **4**, 573–589.

Kersten, D., Knill, D. C., Mamassian, P. & Bülthoff, I. 1996 Illusory motion from shadows. *Nature* **379**, 31.

Landy, M. S., Maloney, L. T., Johnston, E. B. & Young, M. J. 1995 Measurement and modeling of depth cue combination. *Vision Res.* **35**, 389–412.

MacKay, D. J. C. 1992 Bayesian interpolation. *Neural Comput.* **4**, 415–447.

Metelli, F. 1975 Shadows without penumbra. In *Gestaltentheorie in der modernen psychologie* (ed. S. Ertel, L. Kemmler & L. Stadler), pp. 200–209. Darmstadt: Dietrich Steinkopff.

Murase, H. & Nayar, S. 1995 Visual learning and recognition of 3-D objects from appearance. *Int. J. Comp. Vision* **14**, 524.

Nakayama, K., Shimojo, S. & Silverman, G. H. 1989 Stereoscopic depth: its relation to image segmentation,

grouping, and the recognition of occluded objects. *Perception* **18**, 55–68.

Richards, W., Jepson, A. & Feldman, J. 1996 Priors, preferences and categorical percepts. In *Perception as Bayesian inference* (ed. D. C. Knill & W. W. Richards). Cambridge University Press.

Schyns, P. G. & Oliva, A. 1994 From blobs to boundary edges: evidence for time- and spatial-scale-dependent scene recognition. *Psychol. Sci.* **5**, 195–200.

Tarr, M. J. & Bülthoff, H. H. 1995 Is human object recognition better described by geon-structural-descriptions or by multiple-views? *J. Exp. Psychol., Human Perception and Performance* **21**, 1494–1505.

Tarr, M., Kersten, D. & Bülthoff, H. H. 1997 Why the visual recognition system might encode the effects of illumination. (Submitted.)

Trueswell, J. C. & Hayhoe, M. M. 1993 Surface segmentation mechanisms and motion perception. *Vision Res.* **33**, 313–328.

Ullman, S. 1996 *High-level vision*. Cambridge, MA: The MIT Press.

Yuille, A. L., Cohen, D. & Hallinan, P. 1988 Facial feature extraction by deformable templates. Technical report, **88-2**, Harvard Robotics Laboratory.

Yuille, A. L. & Bülthoff, H. H. 1996 Bayesian decision theory and psychophysics. In *Perception as Bayesian Inference* (ed. D. C. Knill & W. W. Richards). Cambridge University Press.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES