BRILL *Seeing and Perceiving (2011)* DOI:10.1163/187847611X595891 brill.nl/sp

Depth in Box Spaces

Sylvia C. Pont [1,*], **Harold T. Nefs** [2], **Andrea J. van Doorn** [1], **Maarten W. A. Wijntjes** [1], **Susan F. te Pas** [3], **Huib de Ridder** [1] **and Jan J. Koenderink** [2]

[1] π-lab, Industrial Design, Delft University of Technology, Landbergstraat 15, 2628 CE Delft, The Netherlands

[2] π-lab, Electrical Engineering, Mathematics and Computer Science, Mekelweg 4, 2628 CD Delft, The Netherlands

[3] Helmholtz Institute, Experimental Psychology, Utrecht University, Heidelberglaan 1, 3584 CS Utrecht, The Netherlands

Received 28 September 2011; accepted 26 July 2011

Abstract

Human observers adjust the frontal view of a wireframe box on a computer screen so as to look equally deep and wide, so that in the intended setting the box looks like a cube. Perspective cues are limited to the size–distance effect, since all angles are fixed. Both the size on the screen, and the viewing distance from the observer to the screen were varied. All observers prefer a template view of a cube over a veridical rendering, independent of picture size and viewing distance. If the rendering shows greater or lesser foreshortening than the template, the box appears like a long corridor or a shallow slab, that is, like a 'deformed" cube. Thus observers ignore 'veridicality'. This does not fit an 'inverse optics' model. We discuss a model of 'vision as optical user interface'.

Keywords

Pictorial shape, vision as user interface, perspective cues, pictorial depth

1. Introduction

The primordial 'deep space' in Western art is the box space. It appears before the convention of linear perspective, and indeed, these spaces are frequently 'out of perspective' (Panofsky, 1925). Despite this, they look perfectly acceptable to modern observers (Fig. 1), who apparently tolerate large deviations from veridicality.

It is often considered self-evident that human vision strives towards a *veridical* representation of the physical scene before the eyes (Palmer, 1999). Such a view is

* To whom correspondence should be addressed. E-mail: S.C.Pont@tudelft.nl

 DOI:10.1163/187847611X595891

Figure 1. At left a painting by Duccio di Buonsegna (1255–1318), at right a panel of the Merode altar piece by Robert Campin/the Master of Flémalle, c.1425. In Duccio's Annunciation Mary is seated in a box space of almost square opening. This space looks perhaps more like a shallow slab than like a cube. The perspective is evidently off, just notice Mary's seat! The Annunciation in the Merode altarpiece is also situated in a box space. The round table top is evidently out of perspective. This space looks perhaps more like a long corridor (notice the length of Mary's bench). The images were downloaded from: http://en.wikipedia.org/wiki/File:Robert_Campin_-_L%27_Annonciation_-_1425.jpg and http://www.art-wallpaper.com/6809/Di+Buoninsegna+Duccio/MaestàAltarretabel+of+Sienese+DomsVorderseiteAltarbekrönung+with+scenes+to+Marie?Width=1600&Height=1200.

at odds with biology (Lorenz, 1973; Tinbergen, 1951), for there is no evolutionary pressure towards veridicality. The evolutionary pressure is towards utility, that is, optimal efficaciousness in the interaction with the world. But a demand for optimal utility runs counter to a demand for veridicality (Hoffman, 2009; Mark *et al.*, 2010). Agents need a fast and reliable optical interface, rather than a system that builds a veridical representation, if such were possible at all (Koenderink, 2011).

The concept of perception as an optical user interface pertains universally to biological agents; there is no doubt that this concept captures the essence of the matter (Lorenz, 1973; Tinbergen, 1951). If human vision is indeed an optical user interface, then one expects to find vision rife with idiosyncrasies and non-veridicalities.

In previous years we have been investigating such cases of idiosyncratic user responses and extreme non-veridical perceptions. We find that these are indeed frequent (Koenderink *et al.*, 2009), although the literature tends to downplay their generic nature. Another exception to the mainstream literature is the work by Konkle and de Oliva (2010), who show that human observers prefer a consistent visual size of real-world objects depictions, which they term the canonical visual size. The present paper fits in with this quest: we investigate whether human observers apply generic laws of perspective, which they should if vision were 'inverse optics' (Forsyth and Ponce, 2002; Poggio, 1984), or whether they rely on some template.

The paradigm used in this study is simple, involving merely a view of a box space with square cross-section as viewed in frontal position. This 'degenerated

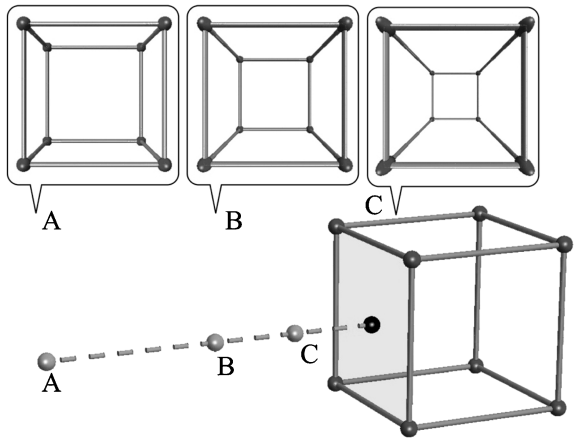

Figure 2. A wireframe cube (box space at aspect ratio one) as seen from various distances (viewpoints A, B and C, as shown in the lower figure) appears quite differently to the eye: the two-dimensional renderings A, B and C (upper three figures) are mutually quite distinct, despite the fact that they 'represent' the same object (the cube shown below).

viewpoint' is chosen intentionally in order to greatly simplify the task. In this case a change of viewing distance does not affect any of the angles in the picture. The only change is due to the size–distance effect, a special case of 'foreshortening'.

We address the topic by having the observers make a direct judgment of the ratio of fronto-parallel to orthogonal (orthogonals running straight into depth) extents. When these stretches are equal, the box space should appear as the degenerate view of a cube. Such a 'degenerated view of a cube' is perfectly suited for our purposes (Fig. 2).

The task is to set the aspect ratio to unity, which can indeed be described as 'making the view look like that of a cube' as opposed to 'a flattish slab of space' or a 'long corridor'. Thus the question could be reformulated as:

'Does the box look like a cube when the optical input is that of an actual physical cube, or does it correlate better with a mere template, or 'standard view of a cube'?

This setting is different, and much simpler, than general views of cuboids (Perkins, 1972, 1974; Sedgwick, 1991; Yang and Kubovy, 1999). We are not concerned with the problem of apparent cubicity as such, as these authors are. For instance, so-called 'Perkins' laws' do not apply to our design, whereas these are crucial to their studies. We merely use the box space in frontal view as an effective way to address our problem.

2. Theory

Any object projects differently on the retina as viewed from different distances. The instance that figures in this paper is the box, or, at the veridical setting, the cube. In Fig. 2, the change of the 'view' with distance is illustrated. A camera would record such variations, but it is an empirical issue whether human observers do.

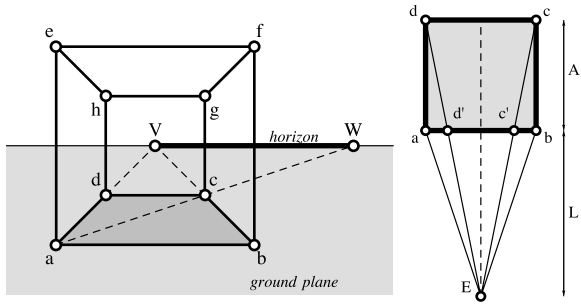

Figure 3. At left the frontal view of a cube, at right the scene is shown in the ground plan. The observer is situated at E. Since the front face is at a distance L of the observer, whereas the rear face is at a (larger) distance $L + A$, the front face appears larger than the rear face. In the perspective picture (left) a simple construction yields the 'correct' viewing distance *VW*.

In Fig. 3 we illustrate the familiar case of the frontal view of a box space of unit aspect ratio, that is, a cube. The box is seen from the direction defined by one set of four parallel edges; the fixation point is the centre of the box. The picture plane is parallel to the faces orthogonal to the viewing direction.

The primary vanishing point V is taken as the centre of the picture. The horizon divides the picture into two equal parts. The box sits on the ground plane with its bottom face *abcd*. The projection of the rear face of the box is smaller than that of its frontal face.

In the experiment, the observers adjusted the size of the rear edges *dc*, *hg*, *cg* and *dh*, whereas the size of the frontal edges *ab*, *ef*, *bf* and *ae* was kept constant. The task was to make the more distant frontal edge of apparently equal length to the closest frontal edge (*cd* to *ab* in Fig. 2) or, equivalently, to equate the lengths of the orthogonal edges to those of the frontal edges (*ad* to *ab* or *cd* in Fig. 2). This effectively requires the observer to make the box space look like a cube. This 'changes the perspective', essentially the 'correct viewing distance', as argued below. In cases in which the box space fails to look like a cube, it either appears as a flattish fronto-parallel slab or as an elongated corridor running into depth.

Let A be the length of the front side edges (*ab*, *bf*, *ef* or *ae* in Fig. 3), B of the backside (*dc*, *cg*, *hg* or *dh* in Fig. 3). The ratio $f = B/A \leqslant 1$ is the foreshortening. Let L be the viewing distance. Then, in the case of a perspective cube, one has

$$f = \frac{L}{A + L}$$

and

$$L = \frac{Af}{1 - f}.$$

These two relations are sufficient for the analysis of the experimental data.

In the experiment the observer has direct control over the foreshortening f. In the analysis it is convenient to compute a factor $g = f/(1 - f)$. The veridical relation

then becomes simply $g = L/A$, or (most conveniently)

$$\log g = \log L - \log A,$$

a relation that can be conveniently tested via multiple linear regression.

The ground face diagonal *ac* defines the righthand 'distance point' W, so W is in a direction at 45° from the viewing direction. Notice that this implies that the viewing distance is equal to the stretch *VW*. (The diagonal *bd* defines the left distance point; it has the same distance to the vanishing point.) The triangles *aWV* and *acd* are similar in the perspective drawing, and so are the triangles *abV* and *dcV*. Thus, $dc : ab = dV : aV$ and $dc : VW = ad : aV = (aV - dV) : aV$ from which one once again obtains the relation $L = Af/(1 - f)$, where L is the viewing distance ($= VW$), A is the edge length ($= ab$) and f is the foreshortening factor. The factor f is set by the observer: from this one finds the veridical viewing distance L. If the actual viewing distance is H, then the observer is veridical if $L/H = 1$.

3. Experimental

3.1. Method

3.1.1. Stimuli

The front and back faces of the box space are visible in wire frame representation. The background was white and the stimuli were drawn in black. All stimuli were similar to the left cube shown in Fig. 3, without the circles, letters, horizon, construction lines and ground plane. Stimuli were generated on an Apple MacBook Pro, running OS-X 10.5.

The stimulus parameter was the ratio of the sizes of the back and front faces. The presentation parameters were viewing distance (30, 60, 120, 240 and 480 cm) and absolute size (2.6, 5.2, 10.5, 20.9 and 41.9 cm wide front edges). Note that this stimulus range is *huge*. In terms of the visual angle the front side of the box ranges from 0.62° (somewhat larger than a full moon) to 109° (the diagonal a factor of 1.4 more, so the stimulus covers a very large part of the viewing field).

3.1.2. Setup

All stimuli were shown on a 37" wide-screen (aspect ratio 16/9), LCD Philips Cineos television. The television screen was set to a resolution of 1280 by 1024 pixels and a vertical refresh rate of 75 Hz.

3.1.3. Participants

All seven authors (four male, three female with ages ranging between 30 and 67) took part in the experiment. All participants had normal or corrected-to normal vision. All participants were highly trained in psychophysical experiments. The experiment was done in accordance with local ethical guidelines, Dutch Law, and with the Declaration of Helsinki.

3.1.4. Procedure

Participants were seated in a normally lit room in front of the television set. We instructed the participants to set the foreshortening of the box space such that it looked like a cube. The foreshortening could be adjusted interactively by pressing the *LEFT/RIGHT* arrow keys on the computer keyboard. The size of the simulated front face of the box in the picture was held constant: only the size of the back (most distant) face was affected by adjusting the perspective-foreshortening. Once participants were satisfied with their settings they pressed the <*ENTER*> key on the keyboard to go to the next trial.

Before each trial, participants were instructed on the screen of the MacBook Pro at which distance they should sit from the screen (30, 60, 120, 240 or 480 cm). A chair was placed at each of these distances before the beginning of the experiment. We did not impose any head or viewing restrictions on the participant except that they wore an eye patch in front of their non-dominant eye.

There were five different viewing distances (see above) and five different sizes of the simulated front face of the box (2.6, 5.2, 10.5, 20.9 and 41.9 cm). All combinations were tested, except the smallest size at a distance of 480 cm, because in this case the observers were not able to adjust the foreshortening with confidence. The remaining 24 combinations were repeated five times, leading to $24 \times 5 = 120$ trials and a lot of walking up and down the room for each participant. The order of stimulus presentation was randomized anew for all participants.

3.2. Results

The data consisted of 840 trials altogether. The veridical foreshortenings range from 0.4 to almost one. The stimulus set is slightly skewed towards large values (close to one). Experimentally we find settings of the foreshortening around a value of 0.63, in a unimodal distribution. The responses are better described as 'constant' than as 'equal to the true foreshortening'. A direct comparison of the adjusted foreshortenings as a function of distance and size confirms this finding (see Fig. 4).

The predicted values according to 'inverse optics' follow the relation $f = L/(A + L)$, with A the size and L the distance (the drawn gray line in Fig. 4). The predicted values for a template model are simply constant (the dashed gray line in Fig. 4). We take the mean of the data for the value of this constant.

The medians of the settings increase somewhat as a function of the true foreshortening (the black line in Fig. 4), but the data are not even close to veridical. We find that the data are almost completely described by a constant, since the slope of the best fit is only 0.28, and the coefficient of variation 0.34. Thus, there is only a slight trend towards veridicality.

There are some small idiosyncratic differences if one splits out the individual observers. The coefficients of variation (R^2) of the straight correlations of the predicted and experimental foreshortenings range from 0.00 (SP) to 0.66 (MW). For most observers we find very weak correlations between the predicted and experimental data. All observers show an appreciable offset (median 0.40, interquartile

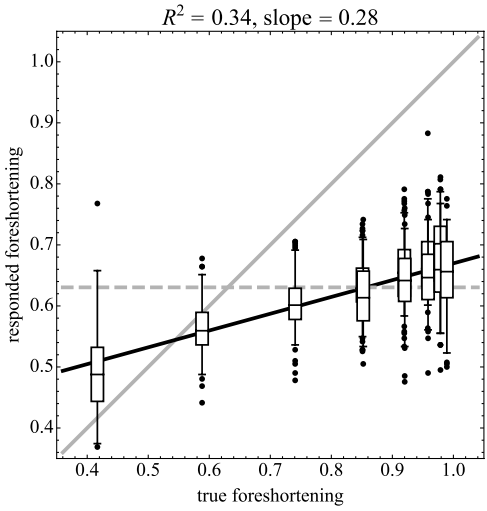

Figure 4. Scatterplot of all data, a total of 840 trials. In the whisker-dot-plots the box indicates the interquartile range, the line the median, whereas the whiskers denote the 5 and 95% quantiles. The dots indicate all outliers. The drawn gray line indicates the expectation for veridical perception, it has slope 1.0. The dashed gray line indicates the expectation for a template model (height taken as the mean of the data), it has slope 0.0. The black line indicates the regression line. The coefficient of variation is 0.34, thus the responses are only weakly correlated with the true perspective. The slope of the regression line is 0.28, thus very shallow. The response is better explained with a constant (implying a template model) than with an equality (implying veridical perception).

range 0.33–0.45) and a very shallow slope (median 0.27, interquartile range 0.21–0.35). Thus, the responses are *very* non-veridical.

A more detailed analysis regresses on both distance L and size A, instead of the mere foreshortening $f = L/(A + L)$, which combines these. *A priori* one expects a linear dependence of $\log g$ on $\log L$ and $\log A$. In the case of complete veridicality one would expect

$$\log g = \log L - \log A,$$

whereas in the case of the application of a simple template one expects

$$\log g = \log g_0,$$

where the constant g_0 is idiosyncratic. In practice we expect (in the simplest case) a linear combination of these dependencies, say

$$\log g = \xi[\log g_0] + (1 - \xi)[\log L - \log A]$$

for some weighting factor $0 \leqslant \xi \leqslant 1$.

Linear regression of the pooled data gives the best fit (in terms of the natural logarithm)

$$\log g = -0.00157 + 0.142[\log L - 0.677 \log A],$$

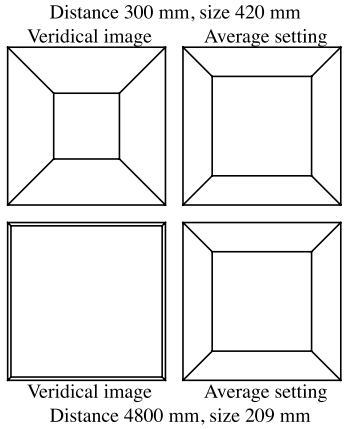

Figure 5. Various cases compared. In the left column the veridical setting, in the right column the overall average response. The two veridical settings are very different. The difference between the two columns neatly summarizes the major result. Apparently, observers refer the optical structure to some 'cube template' and if the stimulus conforms to it will 'look like a cube' to them, otherwise they will experience it as 'deformed' (either shallow slab or deep corridor).

where distance and size are reckoned in millimeters. Apparently one has $\xi = 0.858$.

Thus, the pooled data can be explained for 86% with a template. These numbers differ among observers, for instance, observer SP appears to use pure template matching. The weighting factor ξ ranges from 0.82 to 0.97.

The bottom line is that all observers prefer a template over veridicality. How little the true perspective influences the observer's settings is illustrated in Fig. 5.

4. Conclusions and Discussion

The data clearly show that even for a simple stimulus such as a wire frame cube viewed frontally, observers prefer a template image over a veridical representation (perspective projection) of its shape. Informal interviews confirmed that indeed most people prefer such a 'template' over the veridical rendering of the cube. Observers apparently assume a 'standard view' and hardly bother with perspective at all. Such a technique can be successful as is evident from experiences in computer vision (Murase and Nayar, 1995; Tarr and Buelthoff, 1998).

If the foreshortening of the rendering exceeds that of the template, observers experience an 'elongated corridor' instead of a 'cube'; and if the foreshortening of the rendering falls short of that of the template, observers experience a 'shallow depth slice' instead of a 'cube'.

This is exactly the type of complaint often heard when people casually peruse photographs or movies (Cutting, 1986; Hecht *et al.*, 1999). Wide angle shots are experienced as 'deformed' because showing 'too much depth', whereas tele-shots also look deformed, but now because they look 'flattened', that is to say, they show a lack of depth, see Fig. 6.

Figure 6. Left an extreme wide angle shot and right an extreme tele-shot. One sees both deformations of space and of shape. Space is deformed because the size-distance relations are off in the wide angle shot the mug is gigantic with respect to the head, and in the tele shot people hardly grow smaller as they become clearly more distant. Shape is also deformed, as apparent in the mug (cylindrical?) or the floor tiles (square?) in the wide angle shot, or the flatness of the cars in the tele shot.

From such common observations (see any introductory text on photography) it may be concluded that visual awareness when looking at photographs is frequently non-veridical. Thus, it appears perhaps surprising that most studies in visual perception suggest that visual awareness in the case of perspectively correct renderings is almost invariably veridical. Of course such conclusions are usually based on extremely simplified, not very 'ecologically valid' stimuli (e.g., Palmer, 1999; Perkins, 1972, 1974; Sedgwick, 1991; Yang and Kubovy, 1999). Our stimulus is also extremely simple, so strictly speaking our conclusions only apply to rectangular rooms or corridors seen from a central position. However, because of the aforementioned observations we believe that a more general interpretation is reasonable.

The perspective of a three-dimensional scene depends crucially on the range of relative distances in the scene with respect to the mean viewing distance. Photographs of a fronto-parallel wall taken with wide-angle from near, or tele from far, turn out identical. 'Perspective effects' require an *articulated scene*.

For a viewing distance that is short with respect to the relative distance range of the scene, various objects of the scene are depicted in unfamiliar relative sizes, because of the strong size–distance effect. For instance, in Fig. 6(left), the coffee mug appears huge when compared to the head or the feet. There seems little doubt that this is the major reason to call the picture 'deformed'. Of course this is due to the fact that the observer has certain default expectations.

It can indeed not be otherwise, for there could always be a scene that would yield the same photograph but taken from another distance (think of a bas-relief). Thus the perception of deformation necessarily implies default assumptions. Our experimental results show that such expectations are rather strict and rigid. So the 'deformations' stubbornly remain, even if the observer assumes the 'correct' vantage point.

An example that might serve as a convincing demonstration for vision scientists would be the photograph of an empty Ames' room (Ittelson, 1952) from the correct viewpoint. Such a photograph is of course identical to a photograph of some suitable rectangular room. It is an easy guess what the generic observer is likely to report. In such cases there can be no doubt that observers apply a template.

It is often suggested that wide angle and tele shots are typically seen from the 'wrong' viewing distance (which is true). This is perhaps the standard explanation to be found in text books. The 'right' viewing distance would be the one that reproduces the angular extent at the time of the exposure. In practice, this implies that tele shots 'should' be seen from an impractically long distance and wide angle shots from an equally impossible short distance. The present data show that such ideas do not apply, at least not in the case of the frontally viewed square box or, for that matter, the frontally viewed cube.

Human observers apparently do not use perspective in the formal sense at all.

Instead, they maintain notions of how things should look in a 'canonical view' (Konkle and de Oliva, 2010). It can be seen as an instance of the 'beholder's share' (Gombrich, 1961). If the rendering deviates from the canonical view people will complain (Pirenne, 1970), in apparent disregard of such 'crucial' parameters as viewing distance and picture size.

This finding is hard to explain in the 'vision as inverse optics' paradigm, but fits neatly in the 'vision as optical user interface' concept (Hoffman, 2009; Koenderink, 2011; Mark *et al.*, 2010).

Acknowledgements

This work was supported by Grants from NiRict (SEL-D) and NWO.

References

Cutting, J. E. (1986). The shape and psychophysics of cinematic space, *Behav. Res. Methods, Instrum. Comput.* **18**, 551–558.

Forsyth, D. A. and Ponce, J. (2002). *Computer Vision: A Modern Approach*. Prentice Hall, Englewood, NJ, USA.

Gombrich, E. H. (1961). *Art and Illusion: A Study in the Psychology of Pictorial Representation*. Princeton University Press, Princeton, NJ, USA.

Hecht, H., Van Doorn, A. and Koenderink, J. J. (1999). Compression of visual space in natural scenes and in their photographic counterparts, *Percept. Psychophys.* **61**, 1269–1286.

Hoffman, D. (2009). The interface theory of perception: natural selection drives true perception to swift extinction, in: *Object Categorization: Computer and Human Vision Perspectives*, S. Dickinson, M. Tarr, A. Leonardis and B. Schiele (Eds), pp. 148–165. Cambridge University Press, Cambridge, UK.

Ittelson, W. H. (1952). *The Ames Demonstrations in Perception*. Princeton University Press, Princeton, NJ, USA.

Koenderink, J. J. (2011). Vision and Information, in: *Perception Beyond Inference. The Information Content of Visual Processes*, L. Albertazzi, G. J. Van Tonder and D. Vishnawath (Eds), pp. 27–57. MIT Press, Bradford Book, Cambridge, MA, USA.

Koenderink, J. J., van Doorn, A. J. and Todd, J. T. (2009). Wide distribution of external local sign in the normal population, *Psychol. Res.* **73**, 14–22.

Konkle, T. and de Oliva, A. (2011). Canonical visual size for real-world objects, *J. Exper. Psych.: Human Percept. Perform.* **37**, 23–37.

Lorenz, K. (1973). *Die Rückseite des Spiegels: Versuch einer Naturgeschichte des menschlichen Erkennens*. Piper-Verlag, München, Germany.

Mark, J. T., Marion, B. B. and Hoffman, D. D. (2010). Natural selection and veridical perceptions, *J. Theoret. Biol.* **266**, 504–515.

Murase, H. and Nayar, S. K. (1995). Visual learning and recognition of 3-D objects from appearance, *Int. J. Comput. Vis.* **14**, 5–24.

Palmer, S. E. (1999). *Vision Science: Photons to Phenomenology*. MIT Press, Cambridge, MA, USA.

Panofsky, E. (1925). *Die Perspektive als symbolische Form*. Vorträge der Bibliothek Warburg.

Panofsky, E. (1927). *Die Perspektive als symbolische Form*. B. G. Teubner, Leipzig and Berlin;

Panofsky, E. (1991). *Perspective as Symbolic Form*. Zone Books, New York (English trans.).

Perkins, D. N. (1972). Visual discrimination between rectangular and non-rectangular parallelepipeds, *Percept. Psychophys.* **12**, 396–400.

Perkins, D. N. (1974). Compensation for distortion in viewing pictures obliquely, *Percept. Psychophys.* **14**, 13–18.

Pirenne, M. H. (1970). *Optics, Painting and Photography*. Cambridge University Press, London, UK.

Poggio, T. (1984). Low-level vision as inverse optics, in: *Proc. Sympos. Computat. Models Hearing and Vision*, M. Rauk (Ed.), pp. 123–127. Academy of Sciences of the Estonian S.S.R., Estonia.

Sedgwick, H. A. (1991). The effects of viewpoint on the virtual space of pictures, in: *Pictorial Communication in Virtual and Real Environments*, S. R. Ellis, M. K. Kaiser and A. J. Grunwald (Eds), pp. 460–479. Taylor and Francis, London, UK.

Tarr, M. J. and Buelthoff, H. H. (1998). Image-based recognition in man, monkey, and machine, *Cognition* **67**, 1–20.

Tinbergen, N. (1951). *The Study of Instinct*. Clarendon Press, Oxford, UK.

Yang, T. and Kubovy, M. (1999). Weakening the robustness of perspective: evidence for a modified theory of compensation in picture perception, *Percept. Psychophys.* **61**, 456–467.