Blind subjects construct conscious mental images of visual scenes encoded in musical form

J. Cronly-Dillon^{1*}, K. C. Persaud² and R. Blore¹

¹*Department of Optometry and Neuroscience, and* ²*Department of Instrumentation and Analytical Science, University of Manchester Institute of Science and Technology, Manchester M60 1QD, UK*

Blind (previously sighted) subjects are able to analyse, describe and graphically represent a number of high-contrast visual images translated into musical form *de novo*. We presented musical transforms of a random assortment of photographic images of objects and urban scenes to such subjects, a few of which depicted architectural and other landmarks that may be useful in navigating a route to a particular destination. Our blind subjects were able to use the sound representation to construct a conscious mental image that was revealed by their ability to depict a visual target by drawing it. We noted the similarity between the way the visual system integrates information from successive fixations to form a representation that is stable across eye movements and the way a succession of image frames (encoded in sound) which depict different portions of the image are integrated to form a seamless mental image. Finally, we discuss the profound resemblance between the way a professional musician carries out a structural analysis of a musical composition in order to relate its structure to the perception of musical form and the strategies used by our blind subjects in isolating structural features that collectively reveal the identity of visual form.

Keywords: intermodality information transfer; blind; hearing shapes; mental images; music

1. INTRODUCTION

A musical score provides a pictorial representation of a melody in the form of a graph that displays notes as a spatial distribution. Conversely, a set of points that define the outline of a visual object when mapped as a spatial distribution of notes onto a musical staff and then scanned by the tracker bar of an automatic piano produce a melody that may serve as a musical signature for the shape in question. Alternatively, a device based on a digital video camera linked to a PC with a suitable sound card can be programmed to scan and translate a digitized visual image into an equivalent musical form (Mansur *et al.* 1985; Meijer 1992; Cronly-Dillon *et al.* 1999). We have previously demonstrated the ability of blind (previously sighted) subjects to reconstruct and identify high-contrast line drawings of simple geometrical and other figures translated into musical form (Cronly-Dillon *et al.* 1999). However, when dealing with a musical representation of complex images which depict natural or urban scenes problems arise because the brain is faced with the task of extracting meaningful percepts from what may appear to be a confusing array of sounds within which are embedded elements of structure. Similar problems face the visual system when presented with a complex image containing a multitude of lines and other features from which it has to segment and extract meaningful visual components in order to retrieve the information necessary to carry out the visual task at hand successfully.

(a) *The building blocks of perception*

The advantages of simplifying a representation of a complex visual form by considering it as a combination of simple shape primitives (i.e. icon building blocks) was

recognized by the artist Paul Cézanne who, in 1904, wrote that ‘...treat nature by means of the cylinder, sphere, the cone...everything brought into proper perspective’ (quoted in Gregory 1998, p. 79). The notion of representing complex visual forms as combinations of basic shape primitives that are detected and isolated by specific feature-detecting cells and/or circuits within the brain was revived by Marr (1982) and Biederman (1987). The latter and others extended this idea by allowing gnostic modules (Konorski 1969), established through learning, to play a fundamental role as higher-order primitives/building blocks in the perceptual representation of visual forms (Stryker 1992).

Applying the same principles to the analysis and reconstruction of complex visual images represented in musical form requires processing before the image is translated into sound. This generates a set of simplified representations (see § 2) in which different portions and features of the image are sequestered in different packages. The latter are then encoded in sound and presented as a musical melody with many voices, where each voice encodes the information contained in a particular package (Cronly-Dillon *et al.* 1999). Subjects analyse the component voices, either separately, in sequence or simultaneously in various combinations, in a manner that enables them to patch together and bind the different features of a visual image into a mental reconstruction of a single visual percept. This process of orchestration is comparable to the binding together of the different voices in a chorale or fugue in a polyphonic work of music so that the combination is perceived as a single entity. For vision and visual images mapped into sound, the overall process that leads to stimulus binding also involves the use of working and long-term memory (Desimone *et al.* 1996), mental imagery (Kosslyn & Sussman 1996), selective attention (Triesman 1998) and hypothesis testing (Gregory 1998).

*Author for correspondence (john.cronly-dillon@umist.ac.uk).

2. METHODS

(a) *Subjects*

The primate brain has evolved different mechanisms for selectively segregating and processing the specific items of information needed to carry out the different cognitive tasks involved in the perception of visual form, e.g. depth, spatial relationships, movement, etc. (Zeki 1978). When translated into musical form, each of the latter requires systematic investigation for determining whether the subject is able to use the information encoded in this transmodal form in order to carry out the tasks that would normally be performed using vision. In order to allow us to survey a small set of topics chosen from the most important areas of vision (i.e. form, movement depth and location) and to complete these within a reasonable period of time, we decided to limit our investigation to a few blind subjects who were able and willing to participate in the project over an extended period. The study presented here is a follow-up to experiments reported in an earlier paper (Cronly-Dillon *et al.* 1999), which tested a large number of subjects (i.e. 42). The latter included those who were sighted but prevented from using vision and blind subjects who had been previously sighted. We have previously examined the usefulness of image segmentation and feature extraction in the analysis of high-contrast line drawings of visual forms translated into sound. Our present paper extends the application of these principles to the analysis of static photographic images of urban scenes. These were processed using a Sobel filter in order to produce a high-contrast, edge-enhanced image in which energy peaks (dark points) map the major features of the image. In the present study, testing was carried out using two experienced blind subjects (B. K. and M. S.) who had previously carried out the full series of tests described in our earlier paper (Cronly-Dillon *et al.* 1999). Both B. K. and M. S. lost all form vision through retinitis pigmentosa over 20 years ago. B. K. (a 52-year-old male) also lost all light sense at the age of 27 years, whereas M. S. (a 49-year-old female) retains a faint light sense.

(b) *Apparatus*

The apparatus which was used to transform visual images into musical form consisted of a digital camera for capturing images of the visual scene or target object and relaying this information to a computer that displayed the image on a monitor screen. Otherwise, the method used for transforming images displayed on the screen into musical form was the same as described in our previous paper (Cronly-Dillon *et al.* 1999). On command of the subject, the computer processed and deconstructed the image, producing a set of simpler ones, which extracted and displayed a preselected set of features and then translated each of these 'filtered' images into an equivalent musical form. Subjects were provided with a choice of 'exploratory strategies' that allowed them 'to interrogate the image' in order to gather information that might assist in identification of its contents. These included the following.

- (i) *Testing for symmetry*. This is the ability to test for bilateral symmetry by scanning the image in both directions (along the vertical or horizontal axis).
- (ii) *Selected masking of portions of the target*. Subjects were supplied with a 3 × 4 keypad, which enabled them to select and isolate various square or rectangular strips of the image for detailed investigation and analysis, and, if necessary, to repeat the process again on an isolated patch in order to

segregate even smaller patches of the image for examination.

- (iii) *Changing the resolution of the sound representation of the image (simulating the function of the fovea)*. Increased image resolution was achieved by decreasing the vocal pixel dimensions (i.e. changing the grid size). This increased the number of pixels within an isolated patch, thereby allowing the sound representation of the image displayed therein to be represented at a higher resolution.
- (iv) *Image simplification by feature extraction*. A further facility allowed subjects to generate a series of simplified images by feature extraction. For example, subjects could isolate a particular class of orientated line component so that they could analyse the image with all other types of orientated line components removed or, alternatively, analyse an image in which the selected class of orientated line elements was absent (Cronly-Dillon *et al.* 1999).
- (v) *Alter the scanning speed*. Subjects were also able to change the scanning speed with which the tracker bar traversed either the 'whole' or 'an isolated patch' of the image selected for analysis. In music, the perception of musical form is often dependent on the particular range of speeds at which a passage is played. This is indicated in the score by the tempo, which is indicated as either a metronome mark or a direction in words, e.g. *allegro*, *molto allegro*, *adagio*, *andante*, *lento*, *largo*, etc. A visual form (encoded in sound) being perceived as a single entity depends on a variety of factors, including the velocity selected for the scan and the complexity and length of the musical phrase used to depict the visual form. In general, a slow scan is required for revealing the structural details that allow a subject to distinguish, for example, one architectural form of building from another. A much faster scan is required for generating a global percept of a musical texture that enables subjects to partition an image into separate zones that demarcate where, for example, 'building-like structures' as opposed to 'open grassland and vegetation' predominate (for further details see §§ 3 and 4). Our subjects were able to select from a variety of sweep speeds in order to traverse an area of the screen subdivided into vocal pixels. The sweep speeds of the tracker used in these experiments ranged from 460 ms (for the slowest scan) to 25 ms per vocal pixel (for the fastest scan used in textural analysis).

3. RESULTS

The assortment of 21 (stationary) test images used in these experiments included a series of scenic views and other items taken around a university campus and residential areas (figures 1–3), e.g. trees, a set of railway arches, buildings of various sort (modern halls of residence, a university staff house, Victorian buildings, modern terraced and Edwardian houses, a church, etc.), a wide, curving, concrete staircase (with metal handrails), various types of car, a pavement wastebin and a number of architecturally different entrances to buildings, etc. Among the buildings and other architectural features selected we included a number that might serve as useful landmarks for assisting in navigation from a particular starting location on the university campus to the laboratory where our experiments were conducted. (In homage to Paul Cézanne, we also tested subject B. K. with a print reproduction of one of Cézanne's paintings (figure 4)

Figure 1. Original photographic and processed images analysed by blind subjects B. K. and M. S. Patch analysis was the main strategy used by the subjects here, with individual patches isolated and examined separately. Both subjects identified the images as 'vehicles'. The drawings were produced by B. K. who also verbally described (a) as '... a vehicle ... not a bicycle', then after further analysis '... it is like a Land Rover', and (b) as '... it seems to me, probably to be like what you might call a hatchback ...'.

depicting the roofs of Paris, i.e. 'Les Toits' (*Vue de Paris* 1880). Subjects were instructed to analyse the musical signature of each item depicted separately and then describe and report their relative locations within each image. Prior to being tested, subjects were given no information about the nature or content of the images presented to them other than that they were the sort of items or structures they may encounter while walking in an urban environment. They partitioned and interrogated the images (or portions thereof) during an experimental trial using the tools and strategies available to them for exploring and investigating their contents and structure. Subjects gave their responses by either describing them verbally or producing a sketch of the item, the musical representation of which they had analysed. A video recording was taken of each experimental session in order to provide a permanent record of subject performance. Figure 1 shows the sketches of two types of car produced by B. K. that he was able to identify as being 'rather like a Land Rover' (figure 1a) and 'a kind of hatchback' (figure 1b). M. S. also identified these items as 'vehicles' but was unwilling to be more precise saying '... Sorry ... I am not really into vehicles'.

(a) *Global recognition*

Once our blind subjects had successfully analysed a building (e.g. figure 2a) using a combination of strategies, which included 'patch by patch' analysis and 'feature extraction', they were able to describe it more precisely as 'a modern high-rise building with many windows'. After listening to the target played at different speeds, they also learned to recognize the global signature (sound texture) that characterizes 'buildings in general' (i.e. structures with a predominance of vertical and horizontal components). They were subsequently able to extend this knowledge to recognize buildings in general in other images

presented *de novo* (i.e. figures 2b and 3b,c) and others (not shown) of different architectural type, e.g. a modern, two-storey terraced house, an isolated Edwardian house and a variety of Victorian buildings.

(b) *Elemental (i.e. analytical) processing*

In order to distinguish one building from another it was often necessary for subjects to use a slow scan and resort to patch analysis combined with feature extraction, and sometimes a change in resolution, in order to gain detailed information about the specific features that made one building visually distinct from another. For example, the university staff house depicted in figure 2b was described as a three-storey building with large windows and, thus, was distinguishable from the buildings in figures 2a and 3, etc. For other features such as vegetation, subjects would initially investigate the foliage of trees and bushes without success but, once the experimenter informed them of what they were investigating, they learned the latter's sound signature and, later, were generally able to use this knowledge, though not always reliably, in distinguishing zones where foliage was present. B. K. was able to analyse the top portion of the Cézanne painting *de novo* (figure 4) using his textural knowledge of buildings and identify it as a city skyline. However, while B. K. was able to identify the sloping and horizontal lines of the roof of a building in the immediate foreground and even draw it, he was unable to identify what it was because he had not gathered enough information to recognize that what he was analysing depicted a view 'seen' from the roof of one of the buildings.

(i) *Problems of scale and perspective*

Both subjects were able to identify the presence of arches in figure 3 successfully. Subject B. K. identified the structure in figure 3c as 'a series of arches' which he

Figure 2. Buildings investigated by blind subjects using patch analysis combined with feature extraction (here vertical versus horizontal lines) in differentiating one building from another. (a) This building was identified as ‘... a high-rise building with many windows’. (b) This building described as ‘a three-storey building with large windows’ (the subjects missed the ground floor).

depicted in a drawing and described as ‘probably some form of viaduct extending into the distance’ (in fact it was a railway viaduct), but he was only able to do this after he was informed that the photograph presented a side view of the structure. Subject M. S. was able to identify the first two arches of ‘the viaduct’ but, even when supplied with the same information as B. K., was unable to take account of perspective and so failed to identify or give a full description of the structure. Both subjects experienced problems of scale except when a recognizable target occluded part of another target situated behind (e.g. figure 3a where part of the building is occluded by the railway arch and viaduct). Without the benefit of scale B. K. mistook a ‘sidewalk wastebin’ for ‘some sort of circular telephone kiosk’. Other difficulties arose with some of the other items, such as the photographic image of a curving staircase. Both blind subjects had easily identified a simple ‘skeleton’ drawing showing a straight staircase in a previous study (Cronly-Dillon *et al.* 1999). We suspect that, in this and other cases, difficulties sometimes arose because our subjects could not investigate the item from different perspectives.

4. DISCUSSION

(a) *Orchestration of the mind*

(i) *Top-down rules and working memory are required to produce a seamless mental image*

The fact that our blind subjects were able to produce relatively faithful drawings of the images presented to them is a strong indication that the latter were based on a

conscious (visual) mental image derived from the serial patch by patch analysis and reconstruction of the visual scene encoded in musical form. A mental image of a complex visual object or scene, whether experienced by a sighted or a blind (previously sighted) subject, is a global percept. When subjects first encounter a novel visual target (encoded in sound) the information contained in the image is segmented and presented sequentially as a set of separate items of packaged information. How is the information from the successive presentation of such patches orchestrated in order to form a seamless mental image? The problem is not dissimilar to that encountered in humans and other animals, which sample their visual worlds with a series of discrete fixations separated by saccades, in which the snapshots generated by successive fixations are integrated in order to form a representation that is stable across eye movements. The problem of integrating successive, slightly different snapshots is also analogous to the problem of fusing two disparate images when stereopsis is achieved. In all three cases, the brain must discover which aspects of the image correspond to the same objects and match them accordingly. Thus, if a snapshot from one discrete sample is to be matched with a second slightly different view obtained at a later time, there must be some sort of memory in order to preserve the first for comparison with the second, etc. Of the several theoretical schemes proposed, those described by Hochberg (1968) and Minsky (1977) are particularly relevant. Hochberg (1968) introduced the idea that schematic maps are built up not only from the successive views of a given object, such as a line drawing of an object displayed

Figure 3. Analysis of a visual scene. (a) The subjects identified the building, archway, tree and path. Patch analysis was combined with the use of global texture in partitioning the image into zones of 'foliage' versus 'buildings'. In addition, some feature extraction was used to analyse the 'arch'. The drawing was produced by a blind subject. (b) View from Sackville Street, Manchester. The subjects identified the tree, arch, building and road/path. The drawing on the extreme right was produced by B. K. The zigzag scrawl in the top left was used by B. K. to denote foliage on which he also inscribed the word tree. He verbally identified the structure on the extreme right as a building and started to write it in but his pen ran off the page. (c) Side view of a railway viaduct with drawing produced by a blind subject.

section by section through a slit, but from previous experiences as well. Minsky (1977) addressed the broader problem of integrating successive views of an object or scene as the observer moves around or within it. Essentially, he proposed that prior experience and knowledge of a target object, such as a cube, leads to the establishment of a gnostic circuit consisting of a coherent inter-linked system of frames. As the observer moves around the cube different faces of it become visible and others become concealed from view. Minsky (1977) suggested that, rather than recomputing the description of each viewpoint anew, an encoded representation of the correct frame is accessed when information about an impending movement becomes available. In our previous paper (Cronly-Dillon *et al.* 1999) we commented that, on first

presentation with a complex and unfamiliar line figure (encoded in sound), it might take a long while (6 min or more) for the subject to complete the analysis and reconstruction of the target (unfamiliar images often require segmentation and repeated analysis of separate patches). However, once subjects have successfully analysed and identified a target, they are frequently able to repeat the same task successfully, even as long as six months later, in record time (e.g. 25 s instead of 6 min in the case of a line drawing of a house). Such a substantial saving in time is consistent with the suggestion that recognition on the second occasion involves some form of global process mediated by one (or more) gnostic circuits established through learning, where each combines a coherent system of frames.

Figure 4. Analysis of the Paul Cézanne painting 'Les Toits' (*Vue de Paris* 1880) by B. K. B. K. was informed that the image depicted a painting but received no information concerning either the subject matter or the artist. Separate tracings of vertical, horizontal and oblique components of the image were made on acetate sheets. B. K. described the scene as 'a city skyline', but was unable to identify the structure in the foreground as the roof of a building, although he was able to reproduce it in a rough sketch.

(ii) *Global versus elemental processing*

Global processing of images encoded in sound requires a sufficiently fast scan over the visual target in order to 'telescope' the envelope of the sound representation so that it is perceived as a Gestalt pattern. The speed of traversal of the scan bar across the image, in addition to the patch and pixel size and the properties of the human ear and auditory system, are all factors that determine image resolution (for images translated into sound).

(iii) *The structure and perception of music*

Similar considerations apply to analysis, learning and recognition of the musical phrases that make up a melody, where recognition or non-recognition of individual musical phrases can be greatly affected by articulation and by the speed at which the work is performed. If notes which constitute a phrase are spaced too far apart or too close or there is insufficient contrast with neighbouring phrases, they fail to segregate as a group that forms a Gestalt pattern (Deutsch 1975). We attempt to illustrate some of these points in figure 5 by drawing attention to the similarity between the way our blind subjects set about analysing a visual image translated into musical form and the way a professional musician analyses a musical score in order to reveal the patterns within. In order to perform this comparison, we carried out a simple structural analysis of a single bar of music based on an approach described by Tureck

(1996), the internationally renowned pianist and Bach scholar.

The process of 'simplification' of a musical work follows a similar set of procedures of segmentation and feature extraction to those used by our blind subjects in investigating the musical representations of the photographic images depicted in figures 1–4. Once segregated, the characteristics of individual notes (pitch value) and their relative positions can be investigated by adjusting the patch size and scan speed across the image (this effectively changes the pixel size and resolution). Increasing scan velocity reduces the resolution of fine detail but, in a certain critical range of velocities, can help reveal the existence of a global feature, i.e. a Gestalt pattern, which is perceived as the latter's sound envelope (without the subject necessarily being able to identify the individual notes). The Gestalt signature may later serve as a musical icon that signals the presence of a familiar structural feature, the fine details of which only become apparent when analysed more slowly at higher resolution. Second, the range of critical velocities required for perception of different 'musical icons' depends on a number of features, e.g. the complexity of the musical pattern, its length or duration, etc., so that, although the different critical scan velocity ranges required to detect different musical icons may overlap, fine tuning may often enable a subject to 'highlight' or 'home in' selectively on a particular icon of interest.

Figure 5. Relating musical structure to the perception of musical forms (Gestalt grouping) based on a system developed by Tureck (1996). The system illustrates the same principles of feature extraction employed by our blind subjects in their analysis of musical representations of visual images. (a) A short two-bar treble section (J. S. Bach, F Minor Keyboard Concerto BWV 1056). (b) Vertical structure (pitch variation) omitting the time-values of separate notes. (c) Grouping according to time-values; the notes are separated and grouped solely according to their time-values without regard to pitch. (d) Grouping according to beats within a bar (disregarding pitch-values but retaining time-values) and grouping according to beats within a bar (original score with notes retaining original pitch- and time-values).

(iv) *Orchestrating the 'parts' in order to generate perception of the 'whole'*

In the case of more complex musical sequences involving several voices, the next level of analysis takes account of the combination of two or more simplified 'voices' played at the same scan velocity and pixel size (and, therefore, at the same level of resolution). The task is to tune the scan velocity to a suitable level so that it lies within a range that allows the combination to reveal the musical icons perceived in each segmented portion when each was analysed on its own. Of course, combining the separate portions in this way may expose other more complex Gestalt structures, i.e. higher-order musical icons, with their own critical ranges of scan velocities, which only emerge when the combined subsections are orchestrated and played together. This allows the subject to relate the parts to the whole and corresponds to the way in which a musician may analyse the separate voices in a complex fugue (or chorale) in order to determine their contribution to the global musical effect, i.e. unfolding higher-order Gestalt, which is generated by performance of the whole work. It is significant that, when listening to such a work, we generally find it difficult to tune in to the combined global effect and simultaneously listen to one of the voices in the composition because of the difference in the critical velocity ranges required to accomplish these tasks. Global processing interferes with local processing and vice versa. Recent positron emission tomography and magnetic resonance

imaging studies have also shown this to be true for the neural mechanisms involved in the processing of global and local aspects of hierarchically organized visual stimuli (Fink *et al.* 1997).

(v) *Texture processing in visual perception*

The global processing of a musical phrase via its overall sound envelope has an interesting parallel with the role of texture processing in vision where several levels of textural process may be used in order to partition the image of a countryside or urban landscape into a hierarchy of distinct topographical domains (Watt 1995). A similar strategy based on a process of 'melodic texture discrimination' applied to images transformed into musical form was used by our blind subjects in distinguishing built-up areas (the sound signature of which reveals a predominance of vertical and horizontal components) from areas of open grass or woodland where the power spectrum signature is characteristically different. Indeed this approach capitalizes on the fact that scene classification by power spectrum shapes can provide a useful means of categorizing different natural and artificial scenes for human and machine recognition (Oliva *et al.* 1998).

5. CONCLUSIONS

Our experiments revealed that blind (previously sighted) subjects are able to use musical representations of

visual images in carrying out some of the visual tasks performed by sighted individuals. As with normal vision, the analysis and identification of stationary objects and their relative spatial distributions within an image may be facilitated by patch and textural analysis (combined with selective feature extraction). These procedures allow partitioning of the picture into distinct topographical domains (textural analysis) or the resolution of fine detail through elemental processing (i.e. by combining patch analysis with feature extraction). This enables the subject to acquire information about the character and distribution of important local cues, which may assist in identifying and reconstructing the form of the visual item. Finally, we noted that the degree of agreement between the actual image and the drawings produced by our subjects provides a strong indication that the latter were based on a conscious (visual) mental image that the subject derived from the musical representation of the visual scene. If this is indeed the case, it presents the tantalizing possibility that, with practice and habitual use, blind (previously sighted) subjects may eventually be able to use visual information encoded in sound in generating a running set of mental images that they experience as a daydream sequence depicting scenes and events in the visual world.

We thank the Royal Scottish Asylum and School for the Blind and Smartsight (UMIST Ventures) for financial assistance, Professor Marcus Jacobson for advice and assistance in processing the Cézanne print and Dr R. P. F. Gregory, Dr A. Bunting and Ms Krystine Harvey for assistance in the laboratory.

REFERENCES

- Biederman, I. 1987 Recognition by components: a theory of human understanding. *Psychol. Rev.* **94**, 115–147.
- Cronly-Dillon, J., Persaud, K. & Gregory, R. P. F. 1999 The perception of visual images encoded in musical form: a study in cross modality information transfer. *Proc. R. Soc. Lond. B* **266**, 2427–2433.
- Desimone, R., Miller, E. K., Chelazzi, L. & Lueschow, A. 1996 Multiple memory systems in the visual cortex. In *The cognitive neurosciences* (ed. M. Gazzaniga), pp. 475–486. Cambridge, MA: MIT Press.
- Deutsch, D. 1975 Musical illusions. *Sci. Am.* **October**, 92–98.
- Fink, G., Halligan, P., Marshall, J., Frith, C., Frackowiak, S. & Dolan, R. 1997 Neural mechanisms involved in the processing of global and local aspects of hierarchically organised visual stimuli. *Brain* **120**, 1779–1791.
- Gregory, R. L. 1998 *Eye and brain*, 5th edn. Oxford University Press.
- Hochberg, J. 1968 In the mind's eye. In *Contemporary theory and research in visual perception* (ed. R. M. Haber), pp. 337–354. London: Holt, Rinehart & Winston.
- Konorski, J. 1969 *Integrative action of the brain*. Amsterdam, The Netherlands: Elsevier.
- Kosslyn, S. M. & Sussman, A. L. 1996 Roles of imagery in perception. In *The cognitive neurosciences* (ed. M. Gazzaniga), pp. 1035–1041. Cambridge, MA: MIT Press.
- Mansur, D. L., Blattner, M. M. & Joy, K. I. 1985 Sound graphs: a numerical data analysis method for the blind. *Int. J. Med. Syst.* **9**, 163–174.
- Marr, D. 1982 *Vision*. San Francisco, CA: W. H. Freeman.
- Meijer, P. 1992 An experimental system for auditory image representation. *IEE Trans. Biomed. Engng* **39**, 112–121.
- Minsky, M. 1977 Frame-system theory. In *Thinking: readings in cognitive science* (ed. P. N. Johnson-Laird & P. C. Wason). Cambridge, MA: MIT Press.
- Oliva, A., Guerin-Dugue, A. & Torralba, A. 1998 Scene classification from power spectrum shapes. Proceedings of the European Conference on Visual Perception, August 1998. *Perception* (Suppl.) **27**.
- Stryker, M. P. 1992 Elements of visual perception. *Nature* **360**, 301–302.
- Treisman, A. 1998 Feature binding, attention and object perception. *Phil. Trans. R. Soc. Lond. B* **353**, 1295–1306.
- Tureck, R. 1996 Cell, functions, relationships in musical structure and performance. *Proc. R. Inst.* **67**, 277–318.
- Watt, R. 1995 Some speculations on the role of texture processing in visual perception. In *Early vision and beyond* (ed. T. V. Papathomas, C. Cubb, A. Gorea & E. Kowler), pp. 59–68. Cambridge, MA: MIT Press.
- Zeki, S. M. 1978 Functional specialization in the visual cortex of the rhesus monkey. *Nature* **274**, 423–428.

As this paper exceeds the maximum length normally permitted, the authors have agreed to contribute to production costs.