OPINION

A computational perspective on the neural basis of multisensory spatial representations

Alexandre Pouget, Sophie Deneve and Jean-René Duhamel

We argue that current theories of multisensory representations are inconsistent with the existence of a large proportion of multimodal neurons with gain fields and partially shifting receptive fields. Moreover, these theories do not fully resolve the recoding and statistical issues involved in multisensory integration. An alternative theory, which we have recently developed and review here, has important implications for the idea of 'frame of reference' in neural spatial representations. This theory is based on a neural architecture that combines basis functions and attractor dynamics. Basis function units are used to solve the recoding problem, whereas attractor dynamics are used for optimal statistical inferences. This architecture accounts for gain fields and partially shifting receptive fields, which emerge naturally as a result of the network connectivity and dynamics.

In multisensory integration, signals that come from distinct sensory systems, but which might be related to the same physical object, are combined[1]. For example, you might try to locate an object on the basis of both the image and the sound it generates, or to recognize a word on the basis of its sound and the speaker's lip movements. There are three main aspects to such a process. The first is the assignment problem — determining which sensory stimuli pertain to the same object, which is a difficult task, given that the senses often respond to multiple objects simultaneously. Second, sensory signals — for example, the sound of a word and the image of a person's moving lips — must be recoded into a common format before they can be combined, because sensory modalities do not use the same representations. We call this the recoding problem. Finally, combining multimodal cues involves statistical inferences, because sensory modalities are not equally reliable and their reliabilities can vary depending on the context. For example, vision is usually more reliable than audition for localizing objects in daylight, but not at night. For best performance, the statistical inference must assign greater weights to the most reliable cues, and adjust these weights according to the context[2,3].

This article focuses on the recoding and statistical aspects of multisensory representations, because they can be considered within a common theoretical framework. We also restrict ourselves to multisensory integration in the context of spatial representations — specifically, to the issue of localizing an object using the visual, auditory and somatosensory modalities. We chose this problem because there is a substantial body of neurophysiological data available, allowing detailed comparisons to be made of predictions from the models against the responses of actual neurons.

We start by reviewing the standard theory of multisensory integration. We then summarize the main features of our model[4] and discuss its implications for our understanding of the neural basis of multisensory integration and the notion of frame of reference, particularly in the context of reaching.

The standard theory

A natural way to encode the position of an object is to specify its location with respect to a frame of reference — that is, an origin and a set of axes. For example, the position of the image of an object can be specified with respect to the centre of the retina and its vertical and horizontal axes (a retinal, or eye-centred, frame of reference). Typically, the frame of reference used by a modality is imposed by the geometry of the sensory organs, but it can also result from neural computation. Vision uses an eye-centred frame of reference because of the topographic organization of the retina, but the auditory system uses a head-centred frame of reference that arises from the computations performed early in the auditory system[5].

What frame of reference is used to integrate these sensory signals? The current perspective (FIG. 1) is that multisensory integration involves multiple neural structures and multiple frames of reference. At the bottom of the figure, each modality encodes stimulus positions in its natural frame of reference. At the top, several multimodal modules encode object positions in a frame of reference that is specific to the motor or cognitive function of each module. In between, a single structure — possibly the posterior parietal lobe (for example, see REF. 6) — is suggested to be responsible for mapping from the sensory to the motor modules. In a variation of this view, the parietal lobe is subdivided into several modules, each of which is also associated with particular motor streams and their corresponding frames of reference[7–9].

This approach to multisensory integration conforms roughly to what we know of brain anatomy and physiology, but leaves crucial questions unanswered. First, it sheds no light on the neural mechanisms or computational principles that underlie the mappings. Second, in each output module, all neurons

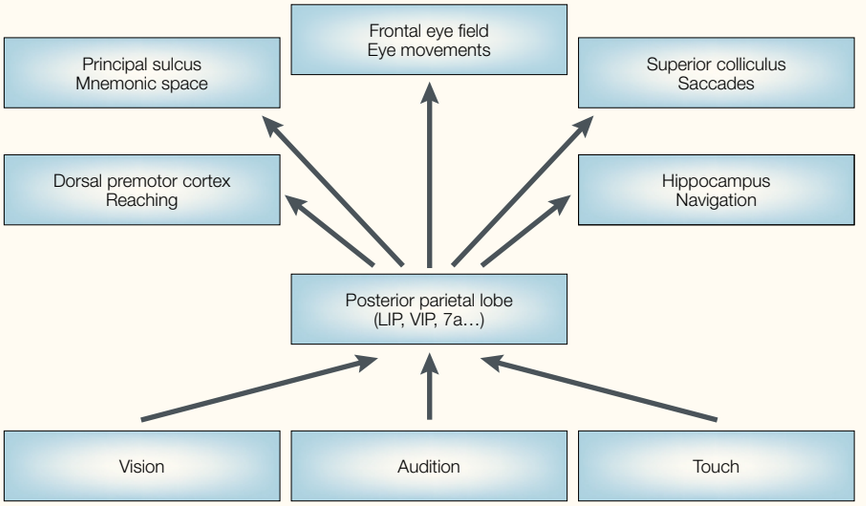

Figure 1 | **A schematic representation of the standard theory for multisensory spatial integration and sensorimotor transformations.** Sensory modalities encode the locations of objects in frames of reference that are specific to each modality. Multisensory integration occurs in multiple modules within the parietal cortex (LIP, lateral inferior parietal; VIP, ventral inferior parietal), which project to a set of motor modules. The multisensory motor modules encode the locations of objects in frames of reference specific to the task controlled by each system.

are supposed to show multimodal response fields that are perfectly remapped in the frame of reference used by the modules.

Surprisingly, the evidence for such complete remapping is scant. For instance, it is often claimed that neurons in the superior colliculus use eye-centred coordinates to specify the locations of visual and auditory stimuli. In this view, bimodal collicular neurons have spatially aligned visual and auditory receptive fields, the positions of which are invariant in eye-centred coordinates. For a receptive field to be eye-centred, it must be anchored to the eyes — every time the eyes move, the receptive field must move by the same amount. However, auditory receptive fields in the superior colliculus show an average displacement of just 52% of the amplitude of the eye displacement[10] — a far cry from the predicted 100%. In other words, the auditory receptive fields are not fully remapped in eye-centred coordinates — they shift, but only partially. The visual receptive fields of the same cells are eye-centred, so the visual and auditory receptive fields cannot be spatially aligned for all eye positions. Similar partially shifting receptive fields have been reported in other areas — such as the ventral inferior parietal area (VIP)[11], the lateral inferior parietal area (LIP)[12], the parietal reach region[13] and dorsal area 5 (REF. 14) — and are likely to occur in area V6a (C. Galletti, personal communication) and in the ventral premotor cortex[15].

It seems difficult to reconcile this behaviour with the idea that each motor module uses one specific frame of reference. These neurons might represent an intermediate stage in the overall process, but it is disconcerting to find so many of them in neural structures that are otherwise believed to specify motor commands (such as the superior colliculus for saccadic eye movements).

A third unresolved point is that multisensory integration involves important and difficult statistical issues, because sensory modalities are not equally reliable and their reliability can change with context. To address these issues, it is natural to adopt a probabilistic approach — sometimes known as a Bayesian approach because it makes use of Bayes' theorem. Localizing an object that can be seen and heard consists of computing, for each location in space, the probability that the object is at that location, given its image and sound. This method weights each cue in proportion to its reliability, and adjusts these weights dynamically when the context modifies cue reliability[2]. Behavioural results are consistent with the idea that the brain performs Bayesian inferences[3,16,17], and there is preliminary evidence that collicular neurons might compute probability distributions[18,19]. Nevertheless, none of these studies provides a theory of Bayesian inferences in neural circuits that can also solve the recoding problem.

In the following section, we review a neural theory of multisensory integration that attempts to address all three issues simultaneously. This theory, which we first proposed in REF. 4, is based on the combination of basis function networks, which provide a biologically plausible solution for performing spatial transformations (for example, see REFS 20–22), and line attractor networks, which have optimal statistical properties[4,23,24]. When these two ideas are combined, they lead to a neural architecture with intermediate stages that contain neurons with partially shifting receptive fields.

The basis function framework

Coordinate transformation. Because an important function of multisensory spatial integration is to enable us to manipulate objects, it is helpful to consider this process from the perspective of sensorimotor transformations. Sensorimotor transformations can often be thought of in terms of coordinate transformations. For example, when reaching for a visual object, it is necessary to find the set of joint angles of the arm that brings the hand to the location of the target — that is, to convert eye-centred target coordinates into joint-centred coordinates.

Coordinate transformation is also at the heart of other aspects of multisensory integration. For instance, predicting the auditory location of an object from its visual location requires a transformation from eye-centred to head-centred coordinates. Likewise, if an object is heard and seen at the same time, it is useful to perform both predictions — from vision to audition, and vice versa — and to compare the predictions with the actual inputs. The result of these comparisons can be used to determine whether the two sensory signals are likely to belong to the same object (the assignment problem). A coordinate transformation is also required when trying to predict the sensory consequences of a motor action — for example, to predict the visual, auditory and proprioceptive location of the hand on the basis of a motor command executed by the arm. In control theory, this is known as a forward model[25–27]. In essence, it is the inverse of a sensorimotor transformation — it requires transformation from motor to sensory coordinates.

Note that these coordinate transformations are generally nonlinear, which means that the output coordinates cannot be obtained through a simple weighted sum of the input coordinates[20,28]. This fact imposes an important constraint on the types of neural architecture that can implement these transformations — the network must contain three, or more, layers of units. In other words, at least one more layer is needed beyond the input and the output layers. What kind of unit should be used in this intermediate layer? One approach uses basis function units — a

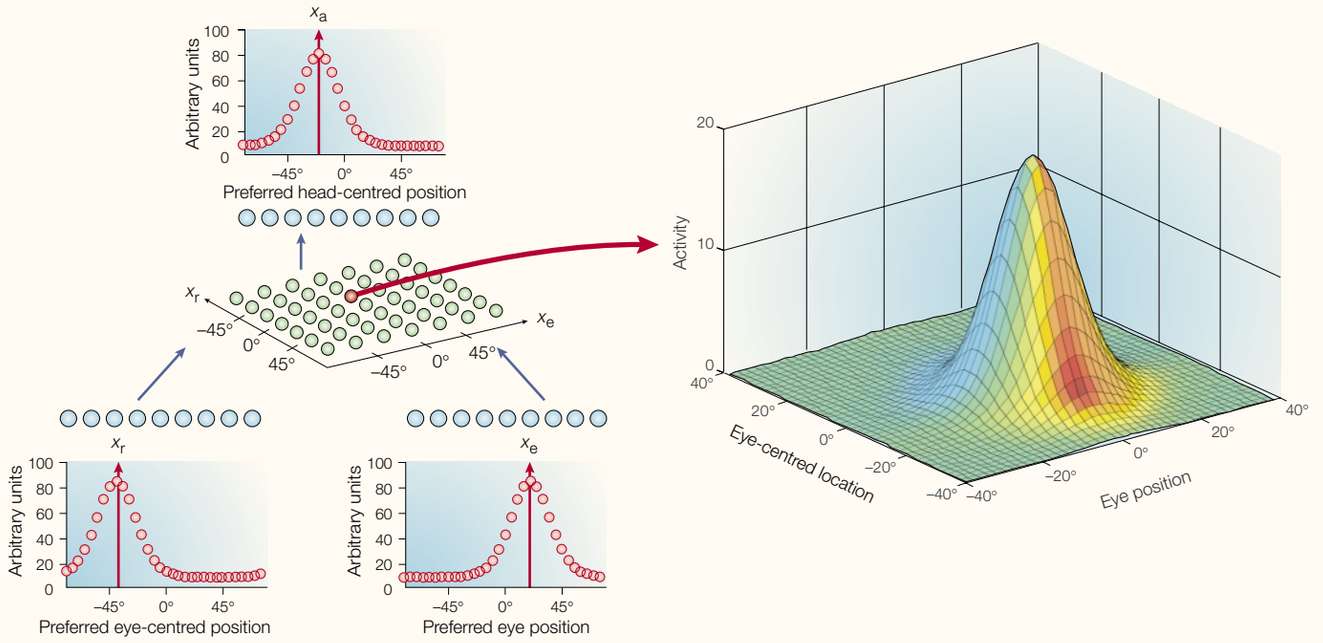

Figure 2 | **A neural network for coordinate transformations using basis functions.** The network contains two input layers. One input layer consists of a retinotopic map of 32 units (bottom left; only nine units shown) that encode the eye-centred locations of objects. The graph shows the activity of the units as a function of their preferred eye-centred location for a visual object at location x_r. The other input layer uses the same type of code but for the position of the eyes (x_e). These two sources of information are combined in the intermediate basis function layer. Each basis function unit computes the product of a pair of eye-centred and eye-position units, from which it inherits a preferred eye-centred location and a preferred eye position. The basis function map is organized in a two-dimensional map in which the preferred positions vary systematically from one unit to the next. The plot on the right shows the response of a typical basis function unit as a function of the eye-centred location of an object and the position of the eyes. The response of the output units is computed by a linear combination of the activities of the basis function units. More specifically, a head-centred unit with preferred head-centred location x_a^k receives connections with a weight of 1 from all the basis function units, the preferred eye-centred location (x_r^i) and eye position (x_e^j) of which are such that $x_a^k = x_r^i + x_e^j$. This pattern of connectivity ensures that the output layer computes a population code for the head-centred location of the stimulus.

solution that is computationally sound and neurally plausible[20–22]. The name 'basis function' is derived from the concept of basis vectors in linear algebra. A set of vectors forms a basis if any other vector can be approximated as a linear combination of these vectors. Likewise, a set of basis functions allows a very large class of functions — notably nonlinear functions — to be approximated as linear combinations of these basis functions.

To understand how this approach can be applied to coordinate transformations, it is helpful to consider a specific example. The network in FIG. 2 computes a population code (the activity of a population of units with bell-shaped tuning curves) for the head-centred location of an object, x_a, from population codes for the eye-centred location of the object, x_r, and the current position of the eyes, x_e. As a first approximation, the head-centred location of an object is equal to the sum of its eye-centred location and the current eye position. The network, however, does not compute $x_a = x_r + x_e$ directly; instead, it computes a population code for x_a given population codes for x_r and x_e. In other words, each head-centred unit computes a bell-shaped tuning curve for x_a that is a Gaussian function of $x_r + x_e$. As Gaussians are nonlinear functions, the computation performed by the output units is nonlinear, and therefore not simply $x_a = x_r + x_e$.

The network has two input layers. The first contains units with bell-shaped receptive fields in eye-centred coordinates (an eye-centred population code). The second contains similar units for the current position of the eyes. The output layer uses a similar scheme to represent the head-centred location of the target. The intermediate layer — the basis function layer — contains one unit for each pair of input units. Each of these units combines the activity of one eye-centred unit and one eye-position unit. FIG. 2 shows the response function of a typical basis function unit as a function of eye-centred location and eye position. Because the input units have bell-shaped tuning curves, the response function of the basis function units is a two-dimensional bell-shaped function. This function reaches its maximal value for a particular pair of eye-centred and eye-position coordinates, which are known as the preferred eye-centred and eye-position values for this unit, and are denoted as x_r^i and x_e^j for unit *ij*.

The basis function units are connected to the output layer to generate the appropriate response function. Our goal is to obtain output units with bell-shaped tuning curves to the head-centred location of the target. We denote x_a^k as the preferred head-centred location of unit *k* (where *k* varies from 1 to *N*, *N* being the total number of output units). To ensure that unit *k* will show bell-shaped tuning with preferred head-centred location x_a^k, it simply needs to receive connections from all the basis function units with preferred eye-centred and eye-position values, x_r^i and x_e^j, such that $x_a^k = x_r^i + x_e^j$. Once these connections are properly set for all output units, the network computes the desired mapping — that is, it generates an output hill of activity at position x_a in response to two input hills at positions x_r and x_e, where $x_a = x_r + x_e$ (FIG. 2). Several variations of this architecture exist[29–33], all relying on similar principles.

The network described so far can perform one coordinate transformation, from eye-centred to head-centred coordinates. As discussed earlier, however, it is often important to be able to perform coordinate transformations in both directions. Fortunately, we can

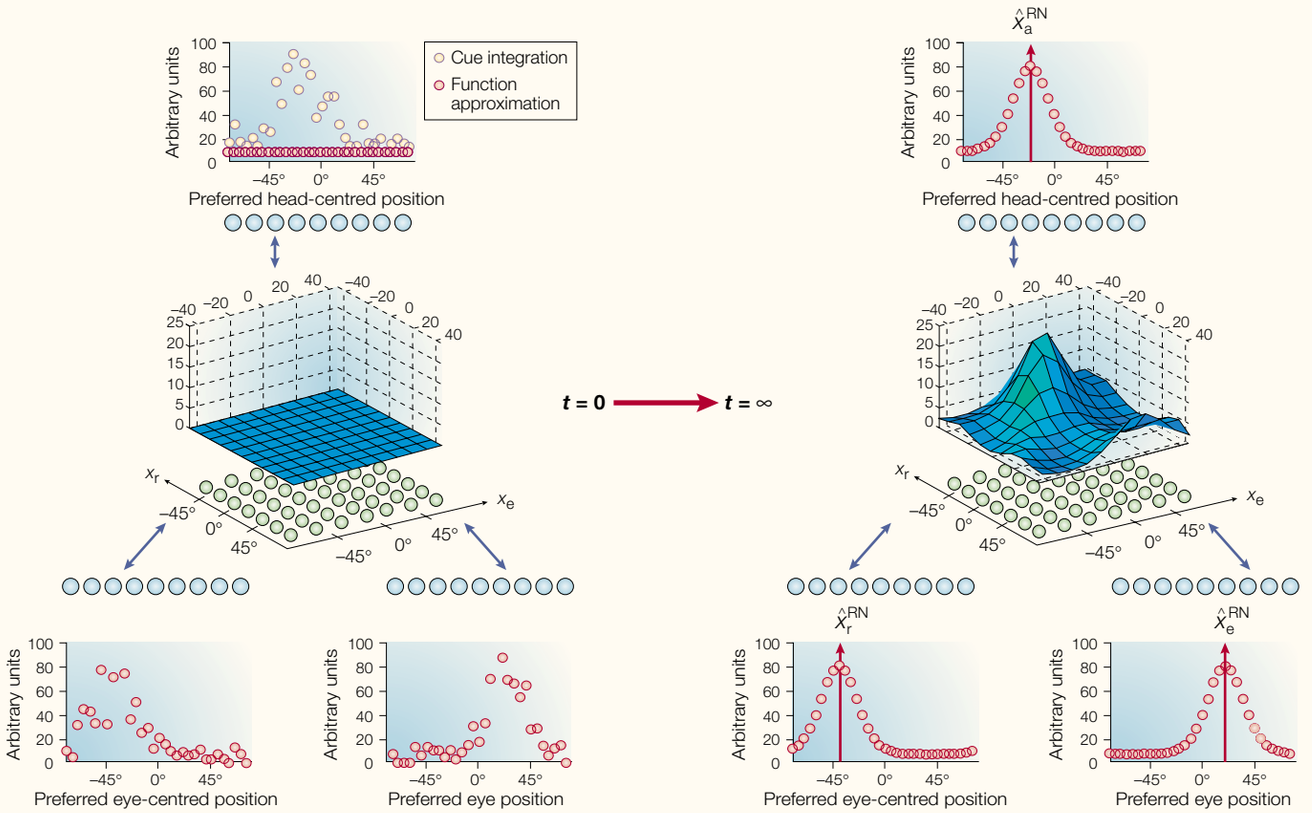

Figure 3 | **A recurrent basis function layer with attractor dynamics.** The network is similar to the one shown in FIG. 2, but all connections are bidirectional. The connections are set to ensure that smooth hills of activity are stable states for the network. Consequently, when the network is initialized with noisy hills of activity (left), it settles onto three smooth hills of activity (right), the positions of which are related through the function $\hat{x}_a^{RN} = \hat{x}_r^{RN} + \hat{x}_e^{RN}$ (RN indicates that these estimates are derived from the recurrent network). When only two hills of activity are provided as inputs, the network must compute the third hill. This could correspond to a sensorimotor transformation in which the sensory input is in eye-centred coordinates and the motor command is in head-centred coordinates. When three noisy hills are presented, the positions of the three smooth hills are the result of integrating the information provided by all three noisy hills. This is what happens in multisensory integration. In all cases, the positions of the smooth hills of activity are very close to the positions predicted by a maximum-likelihood estimator, showing that this architecture is statistically optimal — in particular, for multisensory integration.

re-use the basis function units to map in the other direction. We simply add connections in the opposite direction to the ones we have already established. Hence, if a basis function unit sends a connection to a head-centred unit, we add a connection from the head-centred unit to the basis function unit, and so on throughout the rest of the network. The result is a recurrent network that can perform bidirectional coordinate transforms.

As activity can flow in any direction, it makes little sense to refer to the eye-centred and eye-position layers as the input layers, and the head-centred layer as the output layer. For simplicity, we will refer to these three layers as 'input layers'.

Because the individual neurons in the input layers have bell-shaped tuning curves, patterns of population activity across any of the input layers should also have a bell-shaped profile. The activation function of the basis function units and the weights between layers were chosen to ensure that such bell-shaped profiles of activity were stable states for the recurrent network (for details, see REF. 24). This choice guarantees that, when the network is initialized with two hills in any pair of input layers, it eventually stabilizes onto three hills that peak at locations x_a, x_r and x_e, linked through the relation $x_a = x_r + x_e$ (FIG. 3). In the jargon of dynamical systems, the stable network states — that is, the smooth hills — are called attractors[4]. We therefore refer to our network as an interconnected basis function network or, alternatively, a basis function network with attractor dynamics.

Statistical issues. Now we can focus on the statistical issues involved in multisensory integration. The statistical issue arises because the sensory signals are typically corrupted by noise, either because of the stimulus itself, or because of neural noise within the central nervous system. Given this uncertainty, one can only estimate the location of an object from the evoked sensory activity. The challenge is to come up with the best possible estimate, given the available data.

When multiple sources of information are available, the redundancy between them can be used to refine the estimate. In particular, the most likely location of an object can be computed, given all the available observations. This is known as a maximum-likelihood estimate and is optimal for the problem we are considering[4], in the sense that it leads to an unbiased and efficient estimate. Unbiased means that, over many trials, the average of the maximum-likelihood estimates of an object's location converges on the true location. Efficient indicates that the variance of the estimate is as small as possible, given the noise in the sensory inputs; an efficient estimate is as reliable, or as accurate, as possible.

It is possible to tune the weights of an interconnected basis function network to

implement a close approximation of a maximum-likelihood estimator. This tuning involves adjusting only one parameter — the span of the weights sent by each unit, which, in turn, determines the width of the stable smooth hills (for details, see REF. 4). For equivalence with maximum likelihood, the estimates obtained from a basis function network must have the same mean and variance as those predicted by maximum likelihood. The network estimates are obtained by initializing the network with noisy hills of activity and iterating the network until it stabilizes to three smooth hills. The positions of the smooth hills can then be used as estimates of x_r, x_e and x_a. By repeating this process for many noisy hills, the mean and variance of the network estimates can be computed.

We have performed this test while systematically varying the span of the weights, and have found that, for a particular value of this parameter, the network estimate is unbiased and the variance is only 3.5% larger than predicted by maximum likelihood[4]. We have also been able to confirm these results mathematically, rather than by empirical simulation. Moreover, the network estimate remains close to the maximum-likelihood estimate even when the reliability of the modalities varies between trials. This result was obtained without any further adjustment of the network weights[4]. Recent experimental data suggest that humans can also perform optimal cue integration in conditions in which the reliability of the cue varies between trials[3].

So, we have managed to kill two birds with one stone. The same network architecture can deal not only with the recoding problem, reduced here to a change of coordinates, but also with the statistical inference issue. This is because the recoding is a statistical inference in disguise. If the input neurons in FIG. 3 were noiseless, the statistical inference would reduce to a simple recoding problem; we can think of recoding as estimating the head-centred location of a stimulus from noiseless hills of activity that encode the eye-centred location and the current eye position.

Partially shifting receptive fields. We now consider whether this neural architecture can account for the partially shifting receptive fields that have been found in multisensory areas. FIGURE 4 shows the visual receptive field of a typical basis function unit (from the intermediate layer of the network shown in FIG. 2) for three eye positions, plotted against eye-centred coordinates. The unit has a partially shifting receptive field — the eye-centred location of the receptive field changes when the eyes are moved, but the shift is equal to only half of the change in eye position. Similar results would be obtained by mapping the auditory receptive field of the same unit. Note that the height of the receptive field is also modulated by eye position — a phenomenon that is sometimes referred to as a gain field. Gain fields are common throughout the cortex and are often reported in combination with partially shifting receptive fields[11,34–36].

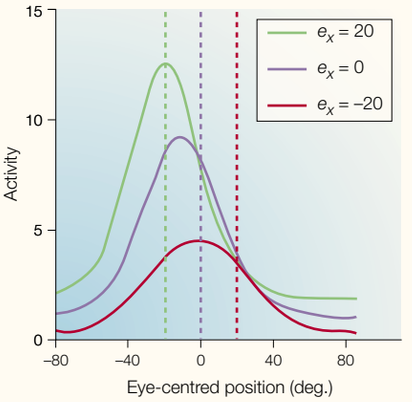

Figure 4 | **A partially shifting receptive field.** The figure shows the visual receptive field of a typical basis function unit in the recurrent network shown in FIG. 3 as a function of eye position (e_x). The three curves correspond to three positions of the eyes. The receptive field is not purely eye-centred, as its position in eye-centred coordinates shifts with the position of the eyes. The shift is only half of that predicted for a head-centred receptive field (vertical lines). The amplitude of the shift can be modulated by changing the connections between the input layers and the basis function layer. The case shown here was obtained in a network with equal visual and auditory weights.

From a computational point of view, there is no fundamental difference between gain fields and partially shifting receptive fields; both provide basis functions. Which of these is used by a network depends on its architecture. In artificial neural networks, each unit computes a nonlinear function of the sum of its inputs. To a first approximation, this operation amounts to a multiplication. Therefore, in FIG. 2, each basis function unit computes a multiplication of the eye-centred and eye-position inputs, resulting in an eye-centred receptive field with a gain modulated by eye position (a gain field). In FIG. 3, the situation is more complex because the basis function units multiply three inputs. The contribution from eye position accounts for the gain field. The multiplication between the eye-centred and head-centred inputs leads to the partial shift; the receptive field is a compromise between the two frames of reference (this can be shown analytically, but is beyond the scope of the present paper).

Therefore, the key for the partial shifts is the convergence of two inputs in distinct frames of reference. For the same reason, partially shifting receptive fields have been found in feedforward neural networks (with no recurrent connections, either lateral or feedback), with inputs that contain visual and auditory inputs, and they are trained on spatial transformations with the backpropagation algorithm[37]. Nevertheless, recurrent connections and attractor dynamics are crucial for efficient multisensory integration — they allow our network to perform multidirectional computations in a statistically optimal way.

In the network in FIG. 3, the partial shift is equal to half the change in eye position for all basis function units. This number can be modulated between 0 and 1 by changing the strength of the input connections onto the basis function units. For instance, if a unit receives a stronger weight from the visual units than from the auditory units, the amplitude of the shift of the visual and auditory receptive fields in eye-centred coordinates will tend to be close to 1. By modulating the strength of the network connection, we can generate a range of shifts, which appears to be consistent with experimental data. For example, Duhamel *et al.*[11] reported an even distribution of shift amplitudes for neurons in area VIP.

The existence of partially shifting receptive fields throughout the brain supports our architecture, but our model makes other experimental predictions. In particular, the presence of hills of activity (the attractor) entails specific patterns of correlations between pairs of cells, which are determined by the product of the derivative of their tuning curves[23]. As multi-electrode recordings become more readily available, we look forward to seeing this prediction put to the test.

Frames of reference

In the recurrent basis function network (FIG. 3), the 'visual' layer is, in fact, a multimodal area, as it also receives connections from the auditory units through the basis function layer. This layer uses eye-centred coordinates to encode both visual and auditory signals. Likewise, the input auditory layer is multimodal and uses a head-centred frame of reference to encode the locations of multimodal targets. These layers resemble the convergence areas predicted by the standard theory (FIG. 1).

However, most units are in the basis function layer. This layer has partially shifting receptive fields and, as such, cannot be assigned a single frame of reference. It uses both eye-centred and head-centred reference frames, allowing it to project to both

input layers in eye-centred and head-centred coordinates. In other words, multimodal integration in the basis function layer does not take place in one frame of reference, but involves a mixture of frames of reference.

Recordings from a cortical area that contained the entire network shown in FIG. 3 would reveal a heterogeneous mixture of multimodal neurons. Some neurons would show invariant multimodal response fields in eye-centred or head-centred coordinates, but most would be in between. This has been reported in area VIP[11] and in dorsal area 5 (REF. 14), and there is evidence for a similar situation in the superior colliculus[10] and the ventral premotor cortex[15].

The typical interpretation of this type of result is that each area uses a single frame of reference, but the responses of the cells appear less precise than expected, because the nervous system uses distributed representations[38] and experimental measurements are noisy. Our theory offers a different perspective. We argue that the partially shifting receptive fields are not the result of crude and noisy experimental measurements, nor of distributed representations in the brain (a statement that, in any case, would need to be clarified to be useful). Instead, they might reflect the use within each neural area of a well-defined class of networks known as recurrent basis function networks, the computational properties of which are well suited to the recoding and statistical problems that arise in multisensory integration.

Partially shifting receptive fields are also sometimes interpreted to be the neural correlates of partial remappings at the behavioural level. For example, when spatial attention is primed with tactile stimulation, the location of the attentional spotlight is only partially remapped in visual coordinates[39]. Although it would be tempting to draw a parallel between this behavioural observation and the partially shifting receptive fields, network models like the one shown in FIG. 3 show that the latter does not imply the former. The network fully remaps head-centred coordinates into eye-centred coordinates (and vice versa), while having partially shifting receptive fields.

Our approach also suggests a new perspective on the related issue of multimodal spatial representations for reaching. Recent studies have suggested that reaching motor commands are specified in eye-centred coordinates, regardless of the modalities in which the reaching target is defined[13,40–43]. Other frames of reference, such as hand-centred and joint-centred, are likely to be involved, but the fact that a frame of reference typically

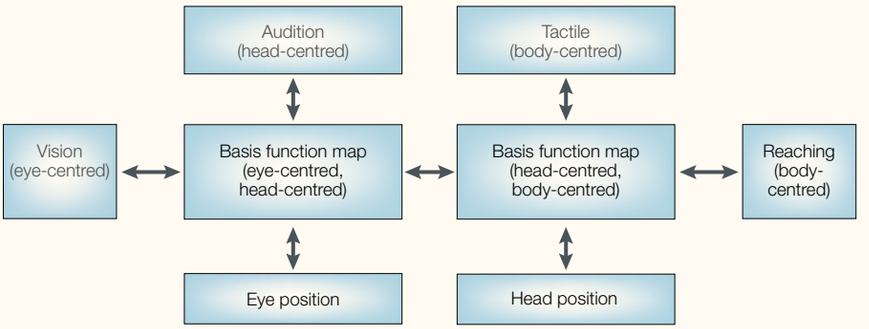

Figure 5 | **A schematic representation of a basis function network for reaching towards visual, auditory and tactile targets.** The network has connections in all directions, allowing it to perform sensorimotor transformations, predictions from one sensory modality to another and predictions of the sensory consequences of a motor action. As a result, the first basis function map encodes auditory and tactile targets in visual coordinates. This could explain why human observers encode auditory targets for reaching in eye-centred coordinates, even though reaching towards an auditory target by itself does not require an eye-centred stage.

associated with the visual system is involved in motor planning is surprising, particularly for auditory targets. In principle, the joint-centred coordinates of an auditory (or proprioceptive) stimulus could be computed from its head-centred location without having to remap it in eye-centred coordinates. Why would the nervous system use eye-centred coordinates for all modalities, even when they are not necessarily required?

Perhaps this simply reflects the dominant role of vision in human behaviour. We believe, however, that the reason for this shared representation is computational; it is a direct consequence of designing a network that can perform multiple tasks at once. FIGURE 5 shows a basis function network that is designed to do three tasks — reaching towards objects regardless of the sensory modality, predicting the position of a target in a modality on the basis of its position in another modality, and predicting the consequences of a reaching movement in all sensory modalities (the forward model). The left side is identical to the network shown in FIG. 3. In fact, the left basis function is the same as in FIG. 3, so it encodes auditory targets in eye-centred and head-centred coordinates. This stage would not be required for reaching towards auditory targets, but it is essential for the network to implement forward models and sensory predictions between modalities.

The question now arises of whether the notion of Euclidean frame of reference, which has been central to how we think about spatial representations, is the best way to characterize these neural representations.

What frame of reference is used by this network to specify reaching motor commands? It is clear that the body-centred frame of reference is involved because it is used in the reaching module (FIG. 5). However, this is not the only one. Owing to the recurrent connections, all layers are eventually activated, regardless of the modality used to specify the target location. As a result, the motor command is encoded in all the frames of reference available in the network (eye-, head- and body-centred in this case, although the list would be longer in a more complete network).

Therefore, the fact that reaching motor commands for auditory targets are specified in eye-centred coordinates is not so surprising when considering a multipurpose architecture like that shown in FIG. 5. It would be just as valid to claim that the first basis function map specifies reaching motor commands in head-centred coordinates for both visual and auditory targets, as the eye- and head-centred coordinates coexist in this map. The question now arises of whether the notion of Euclidean frame of reference, which has been central to how we think about spatial representations, is the best way to characterize these neural representations. We are advocating instead a computational approach, the goal of which is to specify what single cells compute (for example, basis functions, probability distributions, maximum-likelihood estimates) and how those single cells contribute to the overall computation performed by the cortical streams.

Conclusions

The standard theory of multisensory integration (FIG. 1) predicts the existence of convergence zones in which all modalities are remapped into a common frame of reference. The predictions are that cells in these areas should have receptive fields with positions that are invariant in the frame of reference used by the area, and that, for a given cell, the receptive fields for all modalities should be spatially congruent. Most cells in areas such as the superior colliculus, the premotor cortex and area VIP do not meet these requirements, and they have partially shifting receptive fields. Moreover, the standard theory does not provide a complete account of the recoding and statistical problems.

We have argued that developing a neurally plausible and computationally sound solution to the recoding and statistical problems is a key step towards understanding multisensory integration. The solution we have proposed relies on basis function units for the recoding problem and on attractor dynamics for optimal statistical inferences. Basis function units in a recurrent network have partially shifting receptive fields similar to those that have been reported in multisensory cortical areas. These partially shifting receptive fields are a direct consequence of building a network that can perform multidirectional computations, such as from one set of sensory coordinates to another (inter-sensory predictions), from any sensory to any motor coordinates (sensorimotor transformations), and from any motor to any sensory coordinates (forward models).

However, our model has limitations that should be addressed in future studies. In particular, our network assumes that all sensory signals come from the same object. In the presence of several objects, the final estimate of the network would correspond to a single 'phantom' position (a compromise between the positions of the different objects), which would clearly be wrong. This issue is closely related to the assignment problem (how does the nervous system determine which signals, among all those sensed at any given time, pertain to the same object?). One approach to this question consists of computing the likelihood that a set of signals from distinct modalities come from the same object, which is closely related to the ability of these signals to predict each other's values over time and space. It might, therefore, be possible to use a version of the basis function architecture designed for the spatial predictions to solve the assignment problem. It remains to be explored whether our architecture could also perform this prediction over time — that is, whether it could take advantage of temporal coincidences, which are also a potent cue for sensory fusion[44]. Further work will be required to tackle these complex problems, and to uncover their neural and computational basis.

Alexandre Pouget and Sophie Deneve are at the Department of Brain and Cognitive Sciences, University of Rochester, Rochester, New York 14627, USA.

Sophie Deneve and Jean-René Duhamel are at the Institut des Sciences Cognitives, UMR 5015 CNRS-Université Claude Bernard Lyon 1 67, Boulevard Pinel, 69675 Bron Cedex, Lyon, France.

Correspondence to A.P.
e-mail: alex@bcs.rochester.edu

doi:10.1038/nrn914

1. Spence, C. & Driver, J. *Crossmodal Space and Crossmodal Attention* (Oxford Univ. Press, Oxford, UK, in the press).
2. Yuille, A. L. & Bulthoff, H. H. in *Perception as Bayesian Inference* (eds Knill, D. C. & Richards, W.) 123–162 (Cambridge Univ. Press, New York, 1996).
3. Ernst, M. O. & Banks, M. S. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* **415**, 429–433 (2002).
4. Deneve, S., Latham, P. & Pouget, A. Efficient computation and cue integration with noisy population codes. *Nature Neurosci.* **4**, 826–831 (2001).
5. Pena, J. L. & Konishi, M. Auditory spatial receptive fields created by multiplication. *Science* **292**, 249–252 (2001).
6. Gross, C. & Graziano, M. S. A. Multiple representations of space in the brain. *Neuroscientist* **1**, 43–50 (1995).
7. Andersen, R. Multimodal integration for the representation of space in the posterior parietal cortex. *Phil. Trans. R. Soc. Lond. B* **352**, 1421–1428 (1997).
8. Rizzolatti, G., Luppino, G. & Matelli, M. Organization of the cortical motor system: new concepts. *Electroencephalogr. Clin. Neurophysiol.* **106**, 283–296 (1998).
9. Colby, C. & Goldberg, M. Space and attention in parietal cortex. *Annu. Rev. Neurosci.* **22**, 319–349 (1999).
10. Jay, M. F. & Sparks, D. L. Sensorimotor integration in the primate superior colliculus. I. Motor convergence. *J. Neurophysiol.* **57**, 22–34 (1987).
11. Duhamel, J., Bremmer, F., BenHamed, S. & Graf, W. Spatial invariance of visual receptive fields in parietal cortex neurons. *Nature* **389**, 845–848 (1997).
12. Stricanne, B., Andersen, R. & Mazzoni, P. Eye-centered, head-centered, and intermediate coding of remembered sound locations in area LIP. *J. Neurophysiol.* **76**, 2071–2076 (1996).
13. Cohen, Y. & Andersen, R. Reaches to sounds encoded in an eye-centered reference frame. *Neuron* **27**, 647–652 (2000).
14. Buneo, C. A., Jarvis, M. R., Batista, A. P. & Andersen, R. A. Direct visuomotor transformations for reaching. *Nature* **416**, 632–636 (2002).
15. Graziano, M., Hu, X. & Gross, C. Visuospatial properties of ventral premotor cortex. *J. Neurophysiol.* **77**, 2268–2292 (1997).
16. Knill, D. C. Ideal observer perturbation analysis reveals human strategies for inferring surface orientation from texture. *Vision Res.* **38**, 2635–2656 (1998).
17. Jacobs, R. A. & Fine, I. Experience-dependent integration of texture and motion cues to depth. *Vision Res.* **39**, 4062–4075 (1999).
18. Anastasio, T. J., Patton, P. E. & Belkacem-Boussaid, K. Using Bayes' rule to model multisensory enhancement in the superior colliculus. *Neural Comput.* **12**, 1165–1187 (2000).
19. Colonius, H. & Diederich, A. in *Advances in Neural Information Processing Systems* (eds Dietterich, T. G., Becker, S. & Ghahramani, Z.) (in the press).
20. Poggio, T. A theory of how the brain might work. *Cold Spring Harb. Symp. Quant. Biol.* **55**, 899–910 (1990).
21. Pouget, A. & Sejnowski, T. A neural model of the cortical representation of egocentric distance. *Cereb. Cortex* **4**, 314–329 (1994).
22. Pouget, A. & Sejnowski, T. Spatial transformations in the parietal cortex using basis functions. *J. Cogn. Neurosci.* **9**, 222–237 (1997).
23. Pouget, A., Zhang, K., Deneve, S. & Latham, P. E. Statistically efficient estimation using population codes. *Neural Comput.* **10**, 373–401 (1998).
24. Deneve, S., Latham, P. & Pouget, A. Reading population codes: a neural implementation of ideal observers. *Nature Neurosci.* **2**, 740–745 (1999).
25. Ito, M. Neurophysiological aspects of the cerebellar motor control system. *Int. J. Neurol.* **7**, 162–176 (1970).
26. Kawato, M., Furawaka, K. & Suzuki, R. A hierarchical neural-network model for control and learning of voluntary movement. *Biol. Cybern.* **57**, 169–185 (1987).
27. Jordan, M. & Rumelhart, D. Forward models: supervised learning with a distal teacher. *Cogn. Sci.* **16**, 307–354 (1992).
28. Pouget, A. & Snyder, L. Computational approaches to sensorimotor transformations. *Nature Neurosci.* **3**, 1192–1198 (2000).
29. Zipser, D. & Andersen, R. A back-propagation programmed network that stimulates response properties of a subset of posterior parietal neurons. *Nature* **331**, 679–684 (1988).
30. Burnod, Y. *et al.* Visuomotor transformations underlying arm movements toward visual targets: a neural network model of cerebral cortical operations. *J. Neurosci.* **12**, 1435–1453 (1992).
31. Groh, J. & Sparks, D. Two models for transforming auditory signals from head-centered to eye-centered coordinates. *Biol. Cybern.* **67**, 291–302 (1992).
32. Salinas, E. & Abbot, L. Transfer of coded information from sensory to motor networks. *J. Neurosci.* **15**, 6461–6474 (1995).
33. Grossberg, S., Roberts, K., Aguilar, M. & Bullock, D. A neural model of multimodal adaptive saccadic eye movement control by superior colliculus. *J. Neurosci.* **17**, 9706–9725 (1997).
34. Andersen, R., Essick, G. & Siegel, R. Encoding of spatial location by posterior parietal neurons. *Science* **230**, 456–458 (1985).
35. Boussaoud, D., Barth, T. & Wise, S. Effects of gaze on apparent visual responses of frontal cortex neurons. *Exp. Brain Res.* **93**, 423–434 (1993).
36. Trotter, Y. & Celebrini, S. Gaze direction controls response gain in primary visual-cortex neurons. *Nature* **398**, 239–242 (1999).
37. Xing, J. & Andersen, R. A. Models of the posterior parietal cortex which perform multimodal integration and represent space in several coordinate frames. *J. Cogn. Neurosci.* **12**, 601–614 (2000).
38. Robinson, D. A. Implications of neural networks for how we think about brain function. *Behav. Brain Sci.* **15**, 644–655 (1992).
39. Driver, J. & Spence, C. Crossmodal links in spatial attention. *Phil. Trans. R. Soc. Lond. B Biol. Sci.* **353**, 1319–1331 (1998).
40. Enright, J. The non-visual impact of eye orientation on eye–hand coordination. *Vision Res.* **35**, 1611–1618 (1995).
41. Lewald, J. & Ehrenstein, W. The effect of eye position on auditory lateralization. *Exp. Brain Res.* **104**, 1586–1597 (1996).
42. Henriques, D., Klier, E., Smith, M., Lowy, D. & Crawford, J. Gaze-centered remapping of remembered visual space in an open-loop pointing task. *J. Neurosci.* **18**, 1583–1594 (1998).
43. Pouget, A., Ducom, J. C., Torri, J. & Bavelier, D. Multisensory spatial representations in eye-centered coordinates. *Cognition* **83**, B1–B11 (2002).
44. Bertelson, P. in *Cognitive Contributions to the Perception of Spatial and Temporal Events* (eds Ashersleben, G., Bachmann, T. & Müsseler, J.) 347–362 (Elsevier, Amsterdam, 1999).

Online links

FURTHER INFORMATION
MIT Encyclopedia of Cognitive Sciences:
http://cognet.mit.edu/MITECS/
multisensory integration | spatial perception
Access to this interactive links box is free online.