Audio-Visual Localization with Hierarchical Topographic Maps: Modeling the Superior Colliculus

M.C. Casey[1,*], A. Pavlou[1], A. Timotheou[1]

Department of Computing, University of Surrey, Guildford, Surrey, GU2 7XH, UK

Abstract

A key attribute of the brain is its ability to seamlessly integrate sensory information to form a multisensory representation of the world. In early perceptual processing, the superior colliculus (SC) takes a leading role in integrating visual, auditory and somatosensory stimuli in order to direct eye movements. The SC forms a representation of multisensory space through a layering of retinotopic maps which are sensitive to different types of stimuli. These eye-centered topographic maps can adapt to crossmodal stimuli so that the SC can automatically shift our gaze, moderated by cortical feedback. In this paper we describe a neural network model of the SC consisting of a hierarchy of nine topographic maps that combine to form a multisensory retinotopic representation of audio-visual space. Our motivation is to evaluate whether a biologically plausible model of the SC can localize audio-visual inputs live from a camera and two microphones. We use spatial contrast and a novel form of temporal contrast for visual sensitivity, and interaural level difference for auditory sensitivity. Results are comparable with the performance observed in cats where coincident stimuli are accurately localized, while presentation of disparate stimuli causes a significant drop in performance. The benefit of crossmodal localization is shown by adding increasing amounts of noise to the visual stimuli to the point where audio-visual localization significantly out performs visual-only localization. This work demonstrates how a novel, biologically motivated model of low level multisensory processing can be applied to practical, real-world input in real-time, while maintaining its comparability with biology.

Keywords: Audio-visual localization, Multisensory integration, Topographic maps, Superior colliculus, Computational neuroscience

1. Introduction

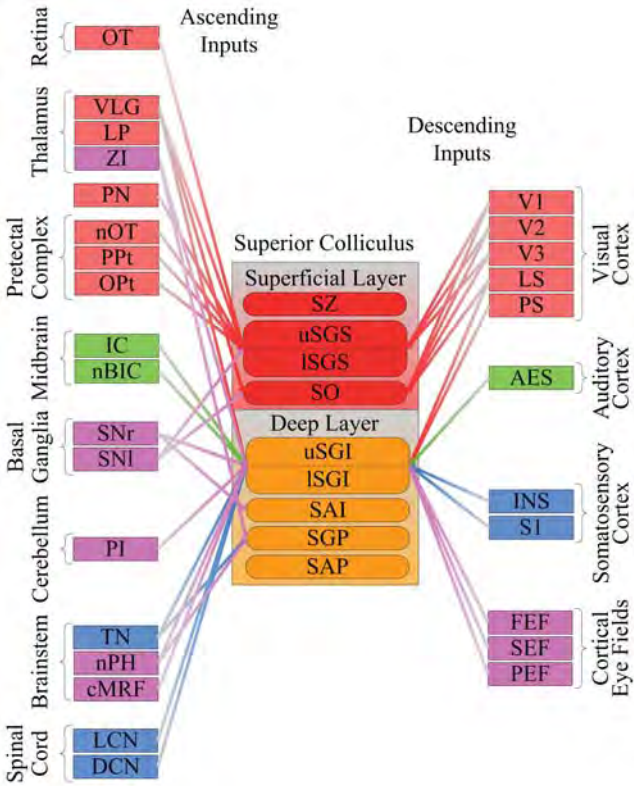
A key attribute of the brain is its ability to seamlessly integrate sensory information to form a multisensory representation of the world. Examples such as the McGurk and MacDonald [1] and rubber hand [2] effects and other perceptual phenomena [3, 4] demonstrate that the processing of one sense can influence significantly what we perceive in another. This influence occurs in both late and early stages of perceptual processing: as a result of semantic similarity in crossmodal stimuli, or because of spatial and temporal similarity [5]. Understanding how such seamless fusion of sensory information is achieved is therefore an important goal, and one which may have a significant impact on computational systems if we can replicate it.

In early perceptual processing, the SC takes a leading role in this *multisensory integration* [6]. The SC is a paired structure found in the midbrain of vertebrates, also known as the optic tectum in non-mammalian species [7]. As an evolutionary stable structure, the SC successfully combines sensory information in order to direct eye movements. In mammals, this involves combining visual, auditory and somatosensory stimuli through layers of aligned topographic maps [5] with a retinotopic organization [8]. The neurons which represent this multisensory space are significant in that they have a superadditive response to weaker crossmodal stimuli [9], which makes the SC particularly important, for example, from a survival perspective. Imagine you are being stalked by a predator which is moving quietly and is camouflaged. The weak crossmodal audio-visual stimuli detected from the quiet sounds and camouflaged motion will provoke a disproportionately strong response from the SC (called multisensory enhancement), thus directing your eyes to the approaching predator [7] even if conflicting loud sounds are present elsewhere (multisensory suppression). As a pivotal structure in the saccadic system with significant influence over a large number of subcortical and cortical areas (Fig. 1), the SC is therefore an important structure to study in neuroscience and computationally.

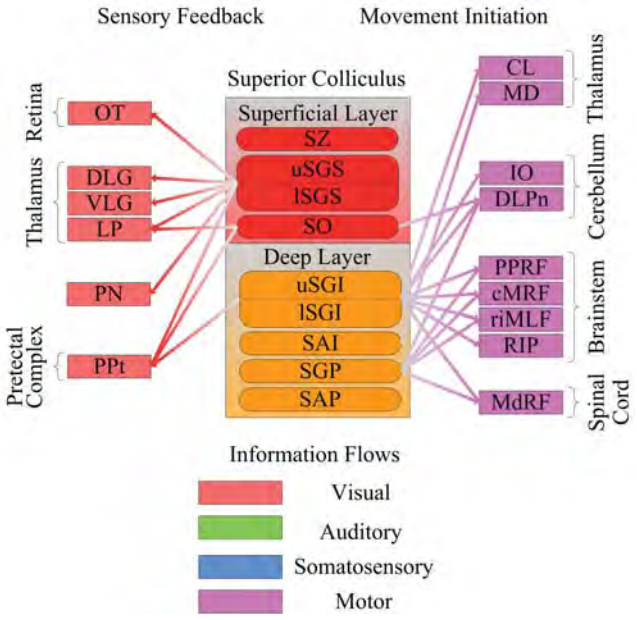
Computationally the SC has four key principles. First, the SC uses aligned topographic maps that *develop* spatial

*Corresponding author. Tel: +44 (0)1483 689635; fax: +44 (0)1483 686051.

Email addresses: m.casey@surrey.ac.uk (M.C. Casey), a.pavlou.csp@googlemail.com (A. Pavlou), at00087@surrey.ac.uk (A. Timotheou)



(a) Ascending and Descending Inputs



(b) Sensory and Movement Outputs

Superficial and deep layers of the SC

SZ	stratum zonale
uSGS	upper stratum griseum superficiale
lSGS	lower stratum griseum superficiale
SO	stratum opticum
uSGI	upper stratum griseum intermediale
lSGI	lower stratum griseum intermediale
SAI	stratum album intermediale
SGP	stratum griseum profundum
SAP	stratum album profundum

Ascending connections

CL	central lateral nucleus
cMRF	central mesencephalic reticular formation
DCN	dorsal column nuclei
DLG	dorsal lateral geniculate
DLPn	dorsolateral pontine nucleus
IC	inferior colliculus
IO	inferior olive
LCN	lateral cervical nucleus
LP	lateral posterior nucleus
MD	medial dorsal nucleus
MdRF	medullary reticular formation
nBIC	nucleus of the brachium of the IC
nOT	nucleus of the optic tract
nPH	nucleus prepositus hypoglossi
OPt	olivary pretectal nucleus
OT	optic tract
PI	posterior interposed nucleus
PN	parabigeminal nucleus
PPRF	paramedian reticular formation of the pons
PPt	posterior pretectal nucleus
riMLF	rostral interstitial nucleus of the medial longitudinal fasciculus
RIP	nucleus raphe interpositus
SNl	pars lateralis
SNr	substantia nigra pars reticulata
TN	trigeminal nucleus
VLG	ventral lateral geniculate
ZI	zona incerta

Descending connections

AES	anterior ectosylvian sulcus
FEF	frontal eye field
INS	insular cortex
LS	lateral suprasylvian cortex
PEF	parietal eye field
PS	posterior suprasylvian cortex
S1	primary somatosensory cortex
SEF	supplemental eye field
V1	primary visual cortex
V2	secondary visual cortex
V3	third visual complex

Figure 1: Known major connectivity of the mammalian SC showing its wide ranging influence: (a) ascending and descending inputs, and (b) sensory feedback and movement initiation outputs. For convenience, ipsilateral and contralateral areas are shown combined. The superficial layers of the SC consist of the SZ which is cell free, uSGS and lSGS, and SO, all processing visual stimuli only. The deep layers of the SC are the uSGI and lSGI, SAI, SGP, and SAP, of which the SGI and SGP are known to process multisensory stimuli. Connections within the layers of the SC are not shown but are known to descend from the SGS, through the SO to the SGI. The ipsilateral and contralateral areas of the SGS, SO, SGI and SGP are also connected to the deep SC of their opposing areas, terminating mostly in the SGI. Constructed from [10].

representations of each different sensory stimuli. Each map is aligned to form an eye-centered (retinotopic) representation with vision driving the alignment [7]. Second, during normal postnatal development, the SC gains the ability to integrate the senses after the first few months. Crucially, however, in abnormal development where one sense is deprived, the ability to learn how to integrate is not lost, and integration is gained after only a short period of exposure to crossmodal stimuli [11]. Here then, the SC not only demonstrates automatic development, detection and alignment of sensory information, but it also shows post-developmental *adaptation*. Third, while the localization of

crossmodal stimuli in the SC is autonomous, superadditive multisensory integration is moderated by cortical feedback from unisensory areas [12], demonstrating how different parts of the perceptual system work together through both *feed forward* and *feedback* connections. Fourth, the SC is a prime example of a *real-time sensorimotor system* that fuses crossmodal stimuli to react rapidly; in as little as 80ms for express saccades [13].

We might exploit these principles in numerous ways. For example, as a real-time sensorimotor system, computational principles extracted from the SC may have a number of applications, such as controlling binocular gaze in robotic systems [14] or enhancing camera surveillance to focus on small-scale anomalies or multisensory cues [15]. This is further enhanced with the ability to adapt, which could be used to make sensory systems more robust or allow them to be trained to discriminate between different types of crossmodal stimuli, such as for audio-visual speech recognition [16]. The ability to automatically align sensory representations could be applied in medical systems, such as in operating theaters to combine, say, fMRI and EEG data to aid surgery [17] in real-time. Lastly, the feedback and cooperation between subcortical and cortical areas that the SC exhibits may demonstrate how we can include top-down influences in bottom-up processing [18] using neural computation. Learning from the SC through modeling could therefore prove pivotal.

A number of different models of the SC have been developed to explore its properties. In general, models have been used to explore how sensory stimuli are integrated in order to characterize multisensory enhancement and suppression. For example, Bayes' rule and simple perceptrons have been applied to describe the computational properties of enhancement and suppression [19, 20]. Other mathematical models of integration have also been defined and compared with biological data [21]. Similar analysis has been conducted using neurobiologically motivated models with explicit cortical feedback [22, 23, 24].

A small number of studies have modeled the development and adaptation of sensory alignment. Simple Hebbian association with rate-coded neurons has been used to align abstract audio and visual stimuli to produce a multisensory representation of space [25]. Other studies have used a more neurobiologically motivated spiking model of the inferior and superior colliculi coupled with spike-timing dependent plasticity (STDP) to learn how auditory stimuli can be re-aligned to displaced visual cues [26, 27].

By far the most extensive models have simulated alignment, multisensory integration and cortical feedback together because of their interdependence. An early model by Grossberg et al. [28] examined the interaction between visual and multimodal cues for reactive and planned eye movement using a neurobiologically motivated model of burst and buildup neurons found in the deep SC. With this model Grossberg et al. explored how different senses can be automatically aligned and how enhancement and suppression can be achieved through cortical feedback. A similar model has also been used to explore the interaction between reactive and planned movement for anti-saccades [29]. Perhaps the most capable neurobiological model to date has been applied within a robot simulation that learns to saccade to a target using head, neck and eye movements [30]. However, despite this model's complexity, there has been no exploitation of the potential for the SC to form the basis of a real-time fusion system going beyond mere simulations. For example, no model of the SC has been applied to real-time video and audio input. This is in contrast to the more established models of cortical visual processing, such as those which explore bottom-up [31] and top-down visual [18] attention. These models combine visual-only topographic maps sensitive to color, intensity, orientation and change in intensity over time to highlight salient areas in images. They therefore use very similar computational techniques to those employed in neurobiologically motivated models of the SC, while they have also been deployed to operate in real-time in, for example, robot platforms [32], albeit using only a single modality.

While neurobiologically motivated models of the SC have been used to explore key attributes of multisensory integration, none have done this on a large-scale or in real-time in the same way that visual-only attention models have been applied. Applying a model of the SC to real-world audio-visual stimuli may therefore offer us insight into how such principles might be used for practical computational intelligence (CI) tasks. An application of a neuroscience model to a practical task would be significant in that it would help us understand if computational principles of low level sensorimotor brain structures can be utilized. The development and application of practical subcortical models is also significant for large-scale cortical models [33, 34] as they will allow us to explore how such models behave when faced with real-world stimuli, aiding the analysis of causal flow, such as in Darwin X [35], and perhaps providing the hybrid architecture needed for us to bridge the gap between large-scale brain simulations [36] and cognitive architectures [37].

In this paper, we describe a behavioral model of the SC that has sensory representations of visual and audio space and combines these into a multisensory representation. While on the one hand this attempts to provide a biologically plausible model of the SC by starting with its connectivity and functional specialization, on the other hand this is balanced with the desire to allow the model to operate on real-time video and audio signals. Although there are necessary abstractions made to achieve this, notably the use of over 12,000 rate-coded neurons in a fixed hierarchy, we demonstrate for the first time how a subcortical computational neuroscience model can be applied to real-world stimuli, and hence how such models may be applied to practical CI problems. The model also includes a novel method using temporal inhibition to discriminate between visual stimuli moving in different directions.

In section 2 we explore the computational properties

of the SC in relation to known physiology and behavior to drive the development of the model. In section 3 we describe the hierarchical topographic map architecture we use to model the SC, including properties of the maps and the required inhibition between neurons. In section 4 we evaluate the model on real-time audio-visual localization and discuss the implications of the results. Finally in section 5 we summarize the contributions of this work and relate it to the wider context.

2. Properties of the Superior Colliculus

The SC is a pivotal structure in the control of saccadic eye movements [7] with connections to and from a large number of brain areas [10] as shown in Fig. 1. As such it takes direct input from the retina for visual stimuli, auditory input from the inferior colliculus (IC) and somatosensory input from the spinal cord and brainstem (Fig. 1a). Additional feedback is also provided from the visual, auditory and somatosensory cortical areas, some of which are crucial for multisensory integration [12].

The main function of the SC is to move the sensory organs of the head to focus on interesting stimuli. The principal output from the SC is therefore to the brainstem in order to initiate eye movement. This is achieved by connections from neurons in the deep SC (SGI and SGP) which are sensitive to both uni- and multisensory stimuli to the PPRF and riMLF in the brainstem (Fig. 1b) to initiate eye movement. Similar connections also exist for head movements via the spinal cord and with other motor systems since gaze changes are important to many different functions [10].

Although we have a good understanding of the anatomy and connectivity of the SC, we know less about the detailed operation of each layer and instead have a broader understanding of the overall functionality and what it is sensitive to, especially for visual processing in the superficial layers and auditory processing in the deep layers. For the purposes of this paper, we will not explore somatosensory processing further since we are focused only on audio-visual integration.

2.1. Visual Sensitivity

Lettvin et al. [38] provided a detailed insight into the visual stimuli key to the operation of the superficial layers of the optic tectum of a frog. Their examination determined that retinal fibers terminate in the tectum to provide a layered continuous map of the retina. There were four types of discrimination performed by these fibers and their corresponding tectal layers. The first layer has receptive fields which detect static contrast at a sharp edge. The second layer, perhaps highly specialized for a frog, detects the motion of convex-shaped dark objects, such as might be obtained from a fly. The third more general layer responds to moving edges, while the fourth, again more specialized, detects a sudden reduction in illumination, such as might occur if a predator's shadow became visible.

These findings are backed-up by studies on the mammalian SC. Sterling and Wickelgren [39] found that the cat SC combined eye-centered information from both eyes to detect motion, especially if the motion was horizontal toward the periphery of the visual field. Similarly, Rauschecker and Harris [40] found that cat SC neurons were selective for the direction of motion. In a study on the rhesus monkey, Wallace et al. [41] also saw sensitivity to moving stimuli from the whole visual field, together with an expanded representation of the foveal region. A more recent study on the rat also shows that the SC is selective for high contrast moving stimuli [42]. In humans, these mammalian studies have been confirmed with data from an fMRI study showing that the superficial layers are sensitive to stimulus contrast and motion [43].

These insights show that across species, the SC 1) uses retinotopic information from both eyes, and that the receptive fields within this visual representation must be sensitive to 2) stimulus contrast and 3) motion, especially in particular directions for some species. Although there is little evidence to understand where this information is processed except within the superficial layers of the SC, it presumably occurs in the SGS, which receives input from the optic tract and visual areas of the thalamus [10]. This visual sensitivity is then used in the deep layers of the SC via the SO, as well as fed back to the optic tract to ignore saccade related stimulus changes, and areas of the visual thalamus, such as LP (Fig. 1b), in order to influence later visual processing, including attention.

2.2. Auditory Sensitivity

Auditory input to the SC principally comes from the inferior colliculus. Sound localization in the IC results in a spatial map varying by azimuth and elevation [44, 45]. This representation is formed using spectral cues for vertical localization and through a comparison of binaural signals using interaural level and time difference (ILD and ITD) for horizontal localization over a range of sound frequencies [46]. Projections from the IC terminate in the SGI, with a topographic map of the auditory space formed varying by azimuth and elevation [47, 48].

Despite ILD, ITD and spectral cues being used in the auditory system for localization, sound representation in the deep SC appears only to be generated from ILD and spectral cues [49], with auditory neurons in the SC rapidly (7 to 27ms) sensitive to a range of sound frequencies, with white noise being most effective at producing a response [47]. Response rates of these neurons also increase with greater sound intensity, although this increase is not linear, with rapid increases if the intensity just exceeds the neuron's observed threshold.

Perhaps the most striking observation of the auditory representation in the SC is that it is eye-centered [7], which is in contrast to a head-centered representation that we

Table 1: Visual and auditory representation and sensitivity

	Visual	Auditory
Representation	Binocular Retinotopic Eye-centered	Binaural Topographic Eye-centered
Sensitivity	Contrast Motion	ILD for horizontal Spectral cues for vertical

would expect given that mammalian ears are either fixed with respect to head position, or move independently to the eyes. Here then, whenever the eyes move, proprioceptive feedback causes the auditory eye fields to shift to maintain correspondence [50]. This eye-centered representation allows the different sensory modalities in the SC to be integrated, with evidence suggesting that visual processing in the superficial layers drives the development of eye-centered auditory representation [51].

In summary, auditory sensitivity in the SC 1) uses a binaural, topographic and eye-centered representation of the auditory space, with receptive fields rapidly sensitive to 2) horizontal location using ILD and 3) vertical location using spectral cues. However, a key part of the topographic representation is that it remains aligned with the visual representation even when the eyes move relative to the ears.

2.3. Properties of Integration in the SC

Table 1 provides a summary of the visual and auditory representation and sensitivity attributes of the SC. These sensory representations in the superficial and deep layers are brought together to form an eye-centered multisensory representation of space. Multisensory enhancement and suppression then operate within this representation through cortical feedback [9] in order to prioritize output from the SC to drive an appropriate saccade or fixate response [52]. Computationally, we are therefore interested in whether we can construct a model which implements these uni- and multisensory representations in order to localize to an audio, visual or audio-visual stimulus in real-time. For simplicity we do not implement cortical feedback or motor control, using instead a simple neural mechanism for integration providing a localization output for a fixed audio-visual input.

3. Hierarchical Topographic Maps

In developing an appropriate method, we take inspiration from rate-coded techniques that have been used to model hierarchical processing with receptive fields [53, 54, 55, 56]. Although rate-coded techniques are not as biologically plausible as other spiking or dynamic models [22, 23, 26, 27, 28, 29, 30], they do allow us to build an appropriate architecture with clearly definable properties and apply appropriate learning algorithms. An example of this is the model developed by Armony et al. [54] who used a hierarchy of interconnected modules, each consisting of a single layer of neurons, to model auditory conditioning in the rat amygdala [55]. Receptive fields in this model were developed through a competitive learning algorithm [57] based on Hebbian principles [58]. The developed representations formed within the maps exhibited local properties of receptive fields, while the combination of modules exhibited the required global behavioral properties of the system. This work has been adapted to operate on two-dimensional, visual inputs [59] with lateral inhibition, and is similar to other competitive learning models of visual processing [53, 56]. Crucially, this approach has been validated against animal behavior [55, 24], while being simple enough to have the potential to operate computationally in real-time [60].

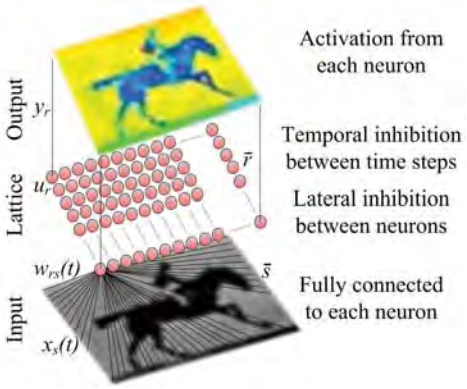


Figure 2: Neural architecture. Each neuron in the lattice is fully connected to the input. Lateral inhibition is used to suppress activation in neurons outside the winning neuron's neighborhood. Temporal inhibition across time steps is used to suppress activation of neurons in the non-preferred direction of motion for the specific map.

3.1. Neural Model

The model defined in Pavlou and Casey [59] arranges neurons into a hierarchy of maps, where each map is formed from a two-dimensional lattice of neurons trained using a competitive learning scheme with lateral inhibition. In this paper we extend this to implement temporal inhibition (Fig. 2).

Maps may take input direct from a source, such as an image, or from one or more maps, or from a combination. Each neuron within a map is fully connected with its input. Connectivity between maps is defined to model the required functionally, such as to model connectivity between the superficial or deep layers of the SC. The activation $y_{\bar{r}}$ of a neuron at the location $\bar{r}$ in the map given an input $x_{\bar{s}}(t)$ at time step t from a two-dimensional stimulus is calculated as:

$$u_{\bar{r}} = \alpha z_{\bar{r}}(t) + \sum_{\forall \bar{s}} x_{\bar{s}}(t) w_{\bar{r}\bar{s}}(t) \quad (1)$$

$$y_{\bar{r}} = \begin{cases} f(u_{\bar{r}}) & \text{if } \|\bar{r} - \bar{n}\| < h(t) \\ f(u_{\bar{r}} - \mu y_{\bar{n}}) & \text{otherwise} \end{cases} \quad (2)$$

$$f(u) = \begin{cases} 1 & u \geq 1 \\ u & 0 < u < 1 \\ 0 & u \leq 0 \end{cases} \quad (3)$$

where $w_{\bar{r}\bar{s}}(t)$ is the weight for input $\bar{s}$ to neuron $\bar{r}$.

Lateral inhibition is used so that the map acts as a set of topographic feature detectors. This is achieved by allowing neurons to compete over an input, such that the neuron with the highest activation, $y_{\bar{n}} = max_{\forall \bar{r}} f(u_{\bar{r}})$ at location $\bar{n}$, inhibits all other neurons outside of its immediate neighborhood. The amount of inhibition is moderated by a factor μ (2). Note that for multiple neurons with the same activation $f(u_{\bar{r}})$, ties are broken randomly.

The circular neighborhood radius $h(t)$ can vary in size, much like that used in Kohonen's Self-organizing Map (SOM) [61], in order to allow the map to be more or less selective in its response. This can be used during training to tune the receptive fields developed by the neuron weights, or during normal feed forward operation to expand or restrict the number of neurons that respond to an input. For example, with a neighborhood that restricts output to just the winning neuron, when activated, the winning neuron's location corresponds to the input that provokes the strongest activity. This enables the map to select the most salient input. If the neighborhood was larger, then the map selects the most salient area represented by a number of neurons. With inhibition $\mu = 0$ activity is enabled for all neurons in the map, and hence the map selects all locations to which its weights are sensitive.

Temporal inhibition across time steps is implemented in the $z_{\bar{r}}(t) \leq 0$ term in the weighted summation (1). This defines an amount by which the neuron is temporally inhibited, moderated by the temporal rate α. This temporal inhibition is determined by previous activity in the map, such that:

$$z_{\bar{r}}(t) = \beta z_{\bar{r}}(t-1) + y_{\bar{n}} I_{\bar{r}} \quad (4)$$

Here, temporal inhibition decays as a factor β over time, while a pre-determined pattern of inhibition $I_{\bar{r}}$ for each neuron is used as a template, and $z_{\bar{r}}(0) = 0$. This allows us to model dynamics observed in sensory neurons. For example, neurons achieve motion detection using pre-filtering, delay filtering and non-linear interactions [62]. One such non-linear interaction uses inhibition for direction sensitivity where a neuron produces an inhibitory response when motion is detected in a non-preferred direction. In essence, our model neurons can inhibit others within the map in the non-preferred direction in order to only permit responses from neurons in the preferred direction.

Competitive learning has been used to train maps to exhibit receptive fields [54, 53, 56]. For example, Linsker [53] trained a four-layered architecture using Hebbian learning. With random activity in the first layer, subsequent layers developed center-surround and orientation selective cells. We have followed a similar approach to train a hierarchy of topographic maps, coupled with exponentially decreasing neighborhood and learning rate functions [24]. With a sufficient period of training on randomly selected stimuli covering the input space, the neurons in each map can learn to form simple Gaussian receptive fields of appropriate sizes.

While previous work demonstrates how receptive fields may be developed using competitive learning, the end result is a uniform set of receptive fields defined by the weights of each neuron. Having determined how such weights can be trained and the form they take, we can therefore use a one-shot learning or fixed weight scheme to precisely define the required receptive fields for each neuron. For simplicity, we use a fixed weight scheme in this paper.

We arrange a series of these topographic maps into a hierarchy to model the visual and auditory processing in the superficial and deep layers of the SC, as shown in Table 1. The hierarchy consists of nine maps with a total of 12,240 neurons. Connectivity between the maps is shown in Fig. 3.

3.2. Modeling Visual Processing in the Superficial Layer

Visual input to the model is taken from a single camera. To reduce the input dimension we convert the color 320 by 240 pixel camera input into grayscale 40 by 30 pixel images with values scaled to range from 0 and 1. From a single camera the corresponding visual field is 72° azimuth by 55° elevation.

To detect spatial contrast we use center-on and center-off receptive fields. When a neuron is tuned to respond to a center-on receptive field, it will fire whenever the center is exposed to light but the surrounding area is not. For a center-off receptive field the neuron fires whenever the surround is exposed to light but not the center. With varying sized center areas, neurons can therefore detect different scales of contrast. An established model of these retinal center-on and -off receptive fields is the a difference of Gaussians (DoG) [63]. Here,

$$G(\bar{r}; \bar{s}; \sigma; \lambda) = \lambda e^{-\left(\frac{\|\bar{s}-\bar{r}\|^2}{2\sigma^2}\right)} \quad (5)$$

$$D(\bar{r}; \bar{s}; \sigma_c) = G(\bar{r}; \bar{s}; \sigma_c; \lambda_c) - G(\bar{r}; \bar{s}; \sigma_s; \lambda_s) \quad (6)$$

where

$$\sigma_s = \rho \sigma_c \quad (7)$$

$$\lambda_c = \frac{1}{2\pi\sigma_c^2} \quad (8)$$

$$\lambda_s = \frac{1}{2\pi\sigma_s^2} \quad (9)$$

define an appropriate DoG where the ratio of the standard deviations of the center and surround Gaussians ρ can be set to a suitable value to model the responses observed from retinal ganglion cells. Enroth-Cugell and Robson [63] provide a range of suitable parameters for the DoG as observed from ganglion neuron studies. Here we select the

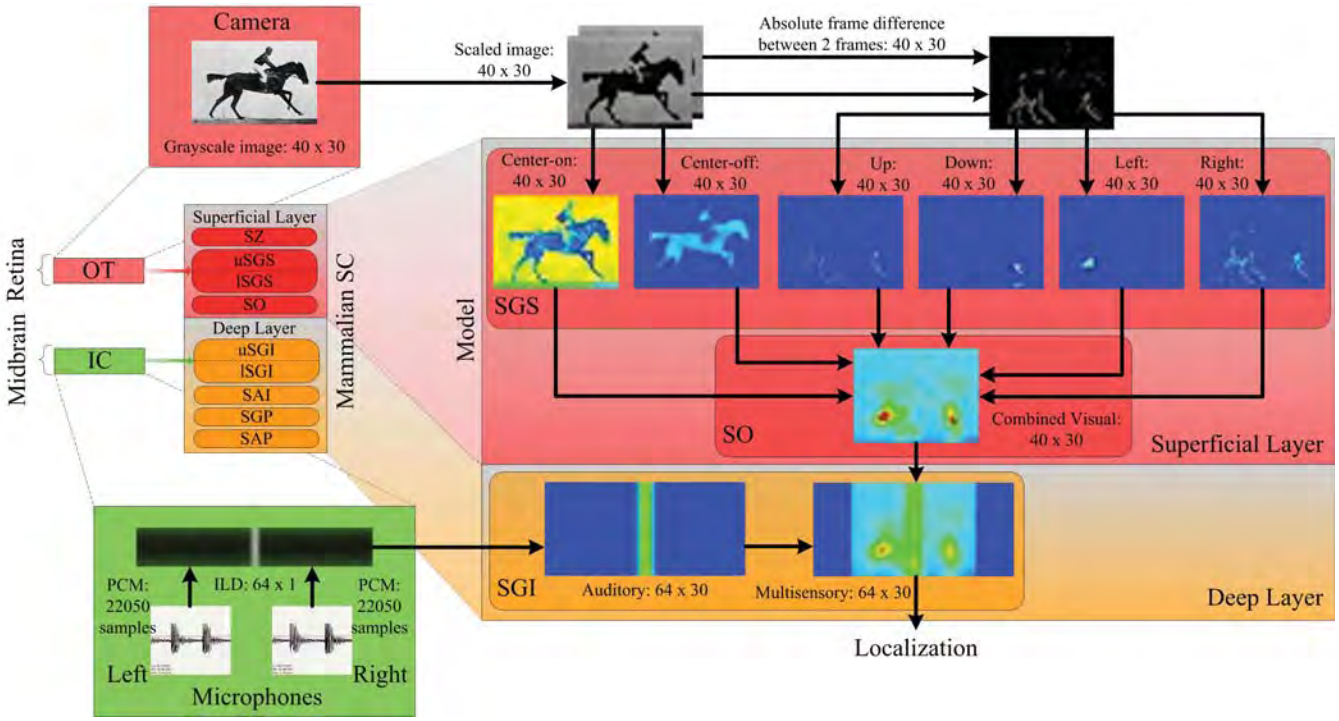


Figure 3: Mammalian SC compared to model architecture. The model consists of a hierarchy of topographic maps where each map consists of a lattice of neurons which are all fully connected to their input. Camera input (representing input from the optic tract) is scaled to form a grayscale image with values in the range 0 to 1. Frame differencing is used across one time step to highlight absolute change. In the superficial layer, maps are sensitive to spatial (center-on and center-off) and temporal (moving up, down, left and right) stimuli, modeling layers in the SGS. Outputs from each visual map are combined in the SO prior to being used within the deep layer. The SGI within the deep layer has audio input which is pre-processed to localize using ILD (representing input from the inferior colliculus). The ILD localization is then represented within an auditory map and this is combined with visual information in a multisensory map for audio-visual localization. Note that the auditory representation is wider (64 neurons) than the visual representation (40 neurons). One pixel in each map shown equates to the value of the output of one neuron.

ratio $\rho = 5.95$ for center-on, which is the median of the values reported from their study, with $1/\rho$ for center-off.

With a map that has the same number of neurons as the input, we can define each neuron to correspond to the equivalent location in the input by defining $\bar{r}$ the center of the receptive field within the range of input values $\bar{s}$. With this set-up we could therefore have a number of different maps with varying DoG radii to select center-on and center-off contrast at different scales [31]. However, as we shall demonstrate in section 4, for our selected image size, one radii is sufficiently sensitive for a range of scales. We therefore used one map for center-on and one for center-off sensitivity with appropriate radii, each taking input direct from the scaled image.

For motion detection, we use frame differencing as a pre-filter and neural inhibition to promote sensitivity to particular directions [62]. Frame differencing allows us to model time within a rate coded model. The absolute difference between the current and last frame is calculated and scaled to be from 0 (no change) to 1 (maximum change). This is then input to four maps with center-on receptive fields as described above that use inhibition to prefer up, down, left and right motion. Center-on fields are used to select the areas of change in the frame difference, while appropriate inhibition patterns are defined for each of the preferred directions through evaluation (further detail in section 4). We note that these four direction sensitive maps could be replaced by a single map sensitive to any motion without direction specific inhibition. However, in the model we wish to test whether direction sensitivity can be achieved since this is found in the SC [39, 40].

The output from each of these six maps represents processing in the SGS, which are then combined into a single visual map representing the SO. The output from each map is weighted such that motion detection takes a higher priority over static contrast. The fixed weight for the combination is formed such that the weighted output from a map for neuron $\bar{r}$ is added to the corresponding weighted outputs from each map. For generality, the output from each neuron is averaged over its neighboring neurons using a Gaussian with $\lambda_v = 1$ and $\sigma_v = 1$ so that broad patterns of activity can provoke a response in the combination.

3.3. Modeling Auditory Processing in the Deep Layer

The SC takes input from the IC for sound localization. Here ILD information is used for horizontal localization and spectral cues for vertical localization. Localization using ILD is less accurate than spectral methods, but this

reflects the crude but rapid level of processing needed for low-level gaze shifts. Note that consumer hardware, such as Microsoft's Kinect [64], is already capable of real-time sound source localization using techniques such as beamforming [65], which is an analogue of ITD in that waveforms are correlated to determine the likely direction of a sound. Unlike these waveform techniques, ILD is relatively simple to implement, rapid and matches closely to the required functionality. We therefore use ILD for horizontal localization with two microphones, which can be used to localize on a single plane. For simplicity, we do not implement spectral techniques for vertical localization.

We implement the ILD algorithm defined by Birchfield and Gangishetty [66] which allows us to localize by comparing the energy received between two microphones. The root mean square (RMS) of a single frame of pulse-code modulation (PCM) input from each microphone is calculated to give a sound pressure level (db SPL) which is converted into sound intensity (db SIL) for each microphone. The intensity is then used to describe the radius of a circle centered on each microphone. By assuming that sound emanates from a plane with base defined by the straight line joining each microphone, an estimate for the sound location via the intersection of the circles can be calculated. We also assume that the location falls at the point on the circle which is closest to the center point between the two microphones and ignore the so-called "cone of confusion". Full details of the algorithm can be found in [66].

The output location from the ILD algorithm is represented as a one-dimensional Gaussian input with maximum amplitude 1, standard deviation 1. The amplitude varies in proportion to the maximum energy detected so that the loudest sound gains amplitude 1. This is input to the auditory map representing parts of the SGI. The map is defined such that a vertical strip of neurons is active to represent the input location. The weights for this map are fixed to be Gaussian $G(\bar{r};\bar{s};\sigma;\lambda)$ with $\lambda_a = 1$ and $\sigma_a = 1$ and are set to be the same for each row of neurons.

3.4. Modeling Multisensory Processing in the Deep Layer

In order to form a multisensory representation of audio-visual space, the output from the visual and auditory maps are combined to provide the desired localization. The SC provides feedback to the IC to translate the incoming head-centered auditory space into an eye-centered space [50]. In our model, we fix the camera and microphone positions so that the eye- and head-centered representations correspond. However, because the auditory space is larger (spanning 114°) than the visual space (72°), the visual and auditory map outputs are combined in a multisensory map which has the same size as the auditory representation, but with the visual output placed centrally within this. The visual and auditory map outputs are weighted equally after scaling from their maximum output to 1 and combined using a Gaussian function in the same way as for the visual map.

Table 2: Input Parameters

Input	$\bar{s}$	Value	Range
Camera	40×30	Grayscale	0 to 255
Scaled Image	40×30	$x^v_{\bar{s}}(t)$	0 to 1
Frame Difference	40×30	$\lvert x^v_{\bar{s}}(t) - x^v_{\bar{s}}(t-1)\rvert$	0 to 1
Left Audio	22050×1	PCM 44.1kHz	16 bit signed
Right Audio	22050×1	PCM 44.1kHz	16 bit signed
ILD	64×1	$x^a_{\bar{s}}(t)$	0 to 1

4. Audio-Visual Localization

In order to evaluate the model, we first describe the analysis we conducted to parameterize its visual sensitivity. A key part of this is evaluating the performance of motion detection using neural inhibition. Second, we test the model on audio-visual stimuli. These latter tests mimic those of Stein et al. [67] when exploring with cats how coincident or disparate auditory stimuli changed visual localization. Video and audio data for the experiments were captured using a Java implementation of the model at two frames per second[1]. This was run on a Dell Precision M2400 laptop with an Intel Core 2 Duo P8600 (2.40GHz) processor, 4GB of RAM, Logitech Live! Cam Voice USB camera and two Logitech USB desktop microphones. The laptop was running Windows XP Professional SP3, Live! Cam driver version 1.1.2.410, Java Runtime Environment 1.6.0_26-b03 and Java Media Framework 2.1.1e. Table 2 shows the size ($\bar{s}$), value and range of each input and input filter used in the model.

4.1. Visual Sensitivity Analysis

To select receptive field sizes for the center-on and center-off DoGs, we tested a range of Gaussian radii so that the receptive fields spanned from 1 up to 15 pixels (half the image height). Images were used with varying contrast (6 variations of background and foreground pixel intensity) which contained a single target (contrasting squares at 5 locations). We found that a radius of 2 pixels for center-on and 64 pixels for center-off gave the best localization rates (93% and 53% respectively) across all the different sized targets. This rate of localization on various sized targets is sufficient to use a single map for center-on and another for center-off. An alternative approach is to use multiple maps, each with different radii in a similar way to the Gaussian pyramid used by Itti et al. [31].

For motion detection, we tested a range of inhibition patterns against moving Gaussian blobs with effective radii varying from 1 to 12 pixels (bandwidths 0.25, 1, 2 and 4). Each target was presented on its own, moving in one of the four preferred directions. The inhibition pattern for a neuron was determined by the map's preferred direction. The pattern was a simple rectangular strip with no inhibition enabled ($I_{\bar{r}} = 0$) starting from the adjacent neuron

[1] A Java demonstration of the system is available at `http://mypages.surrey.ac.uk/css1mc/superior_colliculus.zip`.

and extending in the preferred direction (Fig. 4). Strip widths of 1, 3, 15 and 41 pixels wide were tested to evaluate the best performing area. All other neurons outside of the strip were inhibited ($I_{\bar{r}} = -1$). Note from (4) that the pattern of inhibition is moderated by the winning neuron's output. This pattern therefore concentrates detection on stimuli evoking the highest response.

By varying the strip width from a focused, small strip to a strip which encompasses the whole width of the map, we found that a strip of 3 neurons wide was the best at detecting patterns moving up, down, left or right in the single target motion tests with 94% accuracy. Here we define a correct localization as the highest activity in the map coinciding with the center of the target blob ± 1 standard deviation of the blob's bandwidth (equivalent to the winning neuron exactly locating the smallest, 1 pixel target, up to a radius of 5 neurons locating the largest, 12 pixel target). Since frame differencing is being used, the first frame was ignored as this gave a biased correct localization because of the large change between no input and the first frame. For the larger blobs, localization performance dropped to 76% because of the significant overlap between successive frames preventing the frame differencing from easily highlighting the center of the movement.

The method of inhibition is therefore effective at localizing single moving targets within a static background with a range of target sizes. However, we are also interested in its effectiveness in more complex scenes. To analyze this performance, we constructed scenes consisting of four blobs, each moving in one of the four preferred directions. For the largest radii, the blobs overlapped significantly. An example frame with blobs of bandwidth 1 is shown in Fig. 4.

To detect up, down, left and right motion in the scenes, we use four maps each selective for one of these preferred directions. The maps achieved a combined mean localization accuracy of 38% across the four directions and four

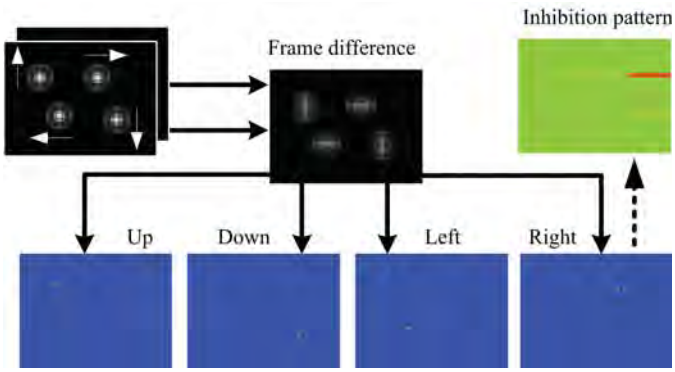


Figure 4: Example output from maps selective for up, down, left and right motion using neural inhibition. Over successive frames the maps filter motion in their preferred direction. Targets are Gaussian-shaped blobs (bandwidth 1 pixel) in the input moving in the directions indicated by the arrows. For the 'right' map, the temporal inhibition pattern is shown, highlighting that all activity outside of the strip which is 1 neuron wide is inhibited.

Table 3: Localization of Four Moving Gaussian Blobs

Blob Bandwidth	Up	Down	Left	Right	Mean
0.25	100%	17%	0%	59%	44%
1.00	66%	86%	0%	17%	42%
2.00	66%	52%	59%	59%	59%
4.00	62%	45%	38%	0%	36%
Mean	73%	50%	24%	34%	45%

blob sizes, using the same criteria for correct localization as above. Blobs with a bandwidth matching closely to the center-on receptive field size evoked the best localization across all directions (45%), while the largest blobs achieved the lowest localization (26%). The best performing strip radii were 1 and 3 neurons wide, both achieving 45% correct localization.

To illustrate the performance, the results for the 3 neuron strip radius are shown in Table 3. The ability of the model to detect all four moving blobs at any one time is lower than detection of a single target. This is because of the difficulty each map has in initially selecting the correct stimulus. Prior to any inhibition being applied, the most active neurons in the map are those that correspond to the largest change in pixel intensity between frames. With four identical blobs, all have an equal change in intensity and therefore all have an equal chance of being selected as the origin for the temporal inhibition even if the chosen blob is not moving in the preferred direction for the map. With ties of winning neurons broken randomly, these equally likely situations can result in low detection rates. Once the inhibition pattern coincides with the correct blob this situation resolves itself. In Table 3 we can see that detection of up and down were typically resolved correctly and achieved a mean of 73% and 50% respectfully over all of the tested blob sizes, while left and right were more difficult to resolve with some blob bandwidths (0.25, 1 and 4) not localized at all for different directions.

A second difficulty that the maps had in resolving the preferred motion was due to the size of the blobs. Blobs with bandwidth 4 overlapped in the input and therefore frame differencing did not highlight the motion of the blobs. This caused the overall performance to drop to 36% across all directions for this bandwidth. What these demonstrate is that, while single target blobs are localized with 94% accuracy, once the scene becomes more complex the accuracy depends significantly on the size of stimulus and how the temporal inhibition patterns overlap if the stimuli have an equal amount of change.

These experiments allowed us to tune the parameters required to localize using either spatial or temporal contrast. Table 4 shows the size ($\bar{s}$), how the weights are constructed ($w_{\bar{r}\bar{s}}(t)$), maximum map output observed during testing, and combination map scaling for each map. Temporal inhibition uses a rectangular strip of 3 neurons for the up, down, left and right maps. All map outputs are normalized from their maximum to 1. Maximum values

Table 4: Map Parameters

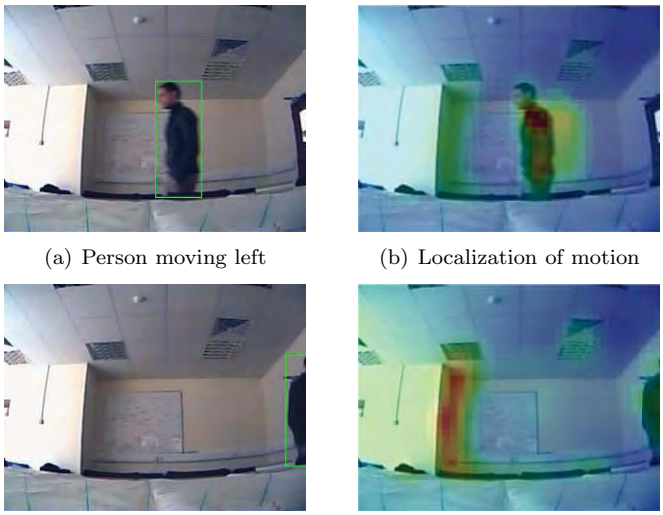
Map	$\bar{s}$	$w_{\bar{r}\bar{s}}(t)\ \forall t$	Max	Scale
Center-on	40×30	$D(\bar{r}; \bar{s}; 0.125)$	0.969	1.5
Center-off	40×30	$D(\bar{r}; \bar{s}; 4.000)$	0.203	3.0
Up	40×30	$D(\bar{r}; \bar{s}; 0.125)$	0.969	6.0
Down	40×30	$D(\bar{r}; \bar{s}; 0.125)$	0.969	6.0
Left	40×30	$D(\bar{r}; \bar{s}; 0.125)$	0.969	6.0
Right	40×30	$D(\bar{r}; \bar{s}; 0.125)$	0.969	6.0
Visual	40×30	$G(\bar{r}; \bar{s}; 1; 1)$	1.000	2.0
Auditory	64×30	$G(\bar{r}; \bar{s}; 1; 1)$	0.791	2.0
Multisensory	64×30	$G(\bar{r}; \bar{s}; 1; 1)$	-	-

were determined using data which provoked a peak response from the maps. Visual testing data consisting of 20 samples generated with a constant dark background (intensity 13) with multiplicative (speckle) noise with pixel intensity variance 256. Auditory test data had a maximum constant intensity of 1.

Once each map output is normalized it is then multiplied by the combination scale prior to input to the relevant visual or multisensory map (Fig. 3). These combination scales have been defined to ensure that any one input can cause a response in the combination map based upon the average activity of the map (and not just the peak). For example, the center-on map (weighting 1.5) responded with a higher overall activity in the map in response to the test patterns compared to the center-off map (weighting 3.0). Although average activity in the up, down, left and right maps varied slightly, we have chosen a single weighting for all for simplicity. The weight value (6.0) was chosen such that any detected motion results in a peak in the visual combination map overriding the peak center-on and -off activities. In the multisensory map, visual and auditory stimuli combine together equally.

With this set-up we tested the visual capabilities of the model on a real-world scene. We collected 30 seconds of data at two frames per second from a camera set up in a room with non-uniform lighting conditions and motion. The scene provided examples of static center-on and -off contrast with motion from a person walking from the right hand side of the scene to the left and then back. For all frames target bounding boxes were defined manually with the person as the target. The person was moving in 43 of the 60 frames but was otherwise visible but stationary.

Fig. 5 shows two example frames from this scene with the corresponding combined visual map output. Across all frames the person was localized with an accuracy of 87%. Of these 52 localizations, 41 were when the person was moving and 11 when stationary, leaving the remaining localizations focused on other areas of static contrast in the frame. This demonstrates that the contrast and motion maps can be combined effectively to localize in real-world scenes to a high degree of accuracy.



(a) Person moving left (b) Localization of motion

(c) Person stationary (d) Localization of contrast

Figure 5: Example frames from a real-world scene testing for motion detection: a) and b) showing a scene where the person is walking left with the corresponding localization (red showing the highest map response); c) and d) showing the person stationary on the right with the corresponding localization of static contrast from the wall.

4.2. Spatial Coincidence and Disparity

To test multisensory integration in cats Stein et al. [67] positioned a cat at the center of a semi-circular screen which had lights and speakers arranged around it at 30° intervals from −90° (left) to 90° (right). The cat was trained to fixate at 0° until a stimulus was presented. For spatial coincidence, the cat had to orient and move towards a coincident light and sound. For spatial disparity, the cat had to orient and move towards the light only, ignoring the sound. The experiments were designed to compare the cat's performance when congruent or conflicting sensory stimuli were received.

To evaluate our model, we positioned the camera centrally to take the place of the cat. The two microphones were placed 0.25m in front of the camera in parallel with the visual plane, and separated by 1m (Fig. 6). The microphones were calibrated so that their response to maximum energy (db SIL) for the same intensity input are equal. Unlike the cat, our camera and microphone position were fixed so that localization could only take place between ±45°. Audio and visual stimuli were therefore placed on the arc at 15° intervals 1m in front of the camera starting at −30° (left) and ending at 30°. Auditory-only stimuli were placed at ±45° 0.25m in front of the microphones. The auditory stimulus was 45 db SPL of white noise, which registered from 13 db SPL to 32 db SPL by either microphone at the various test positions. The light was from a small bulb so that with the room lights turned off, this was the only light visible and hence an easily identifiable target. Since our set-up differs sufficiently to that of the set-up used by Stein et al. [67] (such as in the positioning of the microphones compared to ears) we cannot compare directly to the experiments on cats. Nonetheless, the com-

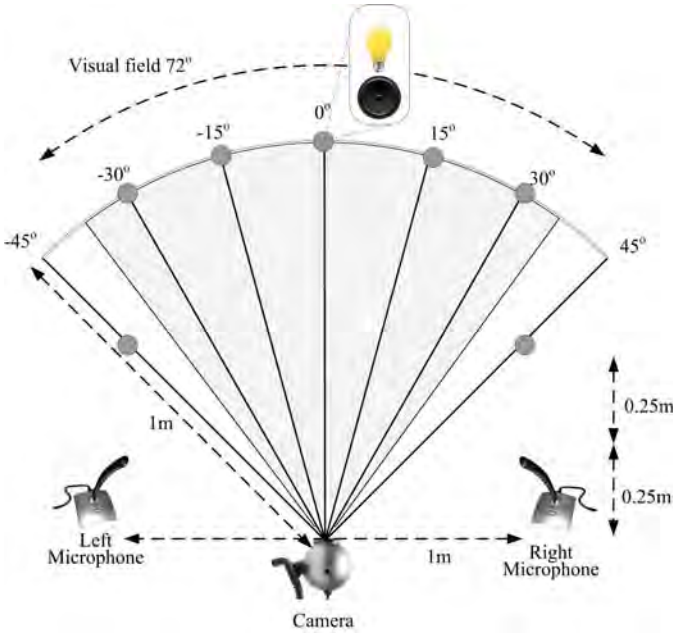


Figure 6: Experimental setup for live audio-visual localization.

putational principles remain the same in that we are combining crossmodal stimuli for localization in a biologically plausible manner.

To test the influence of crossmodal stimuli on localization, the visual target was systematically obscured by increasing levels of noise (described below). Throughout all experiments, the following parameters are fixed for all maps: neighborhood radius $h(t) = 1\ \forall t$, temporal inhibition rate $\alpha = 1$ and decay $\beta = 0.4$. For the multisensory map the inhibition rate $\mu = 1$ so that only one neuron is active to provide a single localization, while for all other maps $\mu = 0$.

We first tested the ability of the model to locate single modality stimuli. Visual- or auditory-only stimuli were presented to the model and the localization recorded. Each test consisted of 20 frames (10 seconds of data). The number of frames was selected to provide sufficient examples for each location and to allow us to evaluate if any change occurred in localization across multiple frames. The model's localization was taken to be the neuron with the maximum output in the multisensory map (the winner). Following the method by Stein et al. [67], the target localization was the visual stimuli. We therefore used the peak visual intensity as the target. This target was specified as a 3×3 square of pixels centered on the peak (covering the bulb's illumination), and constitutes just 0.75% of the audio-visual field. However, because we are combining horizontal and vertical visual localization with horizontal-only auditory localization it is difficult to compare individual modality performance. Therefore to aid comparison we also report performance for horizontal-only visual localization. Since we use frame differencing to detect change in the visual input, the first frame will always detect a change between no input (darkness) and the turning on of the light source. Consequently in reporting the results below, we discard the response from the first frame to avoid experimental bias in the results. Results are therefore reported against 19 frames.

Table 5: Confusion Matrix for Stimuli against Visual Targets

Target	Model Response							
	−45°	−30°	−15°	0°	15°	30°	45°	NK
−45°	**19**	17	2	0	0	0	19	76
−30°	19	**19**	18	0	0	0	19	58
−15°	19	19	**12**	0	0	0	19	64
0°	19	4	10	**18**	19	19	19	25
15°	19	0	0	13	**19**	19	19	44
35°	19	0	0	0	13	**19**	19	63
45°	19	0	0	0	0	0	**19**	95

Visual-only localization successfully located 66% of all audio-visual possible targets (93% of the visual-only targets). 7 frames at −15° were not localized because of slight reflections of the light from the floor surface. In contrast, auditory localization was 99% successful for all targets. When both auditory and visual stimuli are presented to the model and are coincident, the combination of audio and visual cues gives 94% localization. Here, auditory localization makes up for the lack of visual input at ±45°, although visual input dominates still at −15°. Comparing this with horizontal-only visual performance, 71% of all possible audio-visual targets were located using visual-only stimuli (100% of visual-only targets). When audio-visual stimuli were combined, 99% of all targets were located. This demonstrates that the vertical visual component is accounting for just 7 out of 133 incorrect localizations and that the horizontal component dominates.

Table 5 shows the confusion matrix for coincident and disparate stimuli with visual targets (rows) against the model's corresponding localization response (columns). For example, when a visual target of 15° is presented, the model responds with 19 correct localizations when the visual and auditory stimuli are both at 15° (coincident results are shown on the diagonal in the table with a maximum of 19 correct responses each). When the stimuli are disparate, 114 incorrect localizations are recorded, such as when the auditory stimulus is at −45° resulting in 19 localizations at −45° because the auditory stimulus confuses the response. In some cases neither of the conflicting stimuli dominate but instead are combined in the maps to form an intermediate localization. In these cases where the intermediate localization is not at one of the known target points, this is labeled as "not known" (NK). For the 15° example this occurs 44 times. In total there are $7 \times 19 = 133$ possible coincident (19) and disparate (114) sample pairs.

When presented with coincident and disparate audio-visual stimuli, the ideal outcome is for a diagonal confusion matrix with 133 for each diagonal element. We can

see from Table 5 that disparate stimuli produce adverse localizations. From 94% correct for coincident-only stimuli, the performance drops to 13% (for horizontal-only visual targets this goes from 99% to 14%). However, we can see that there is a greater likelihood of responding with a location near the target for closer audio-visual stimuli, versus no localization for stimuli further apart. The exception to this is $\pm 45^\circ$ where there is only an auditory stimulus. A similarly profound decrease in localization was observed in cats when disparate stimuli were presented.

With clear single and coincident modality visual and auditory stimuli being localized with greater than 93% accuracy, what benefit does audio-visual integration bring except to make up for localization outside of the available visual field? Recalling the example of being stalked by a predator which is moving quietly and is camouflaged [7], crossmodal localization is useful when one or both stimuli are weak or difficult to discern. To test this in our model we can therefore weaken one of the stimuli to determine if the deficit is compensated for via the other modality. Since vision tends to dominate in the SC [68] we tested the model on a series of degraded visual stimuli but kept the auditory stimuli unchanged (note that only a slight amount of noise in the auditory modality would cause an immediate failure in sound localization using ILD). Multiplicative noise (speckle) was added to each recorded frame with variance increasing from pixel intensity 16 to 256 in steps of 16, such that the location of the light in the input was completely obscured with noise variance 192 and above. The resulting coincident audio-visual localization performance is shown in Fig. 7 compared against the single-modality visual localization.

By combining audio-visual stimuli, localization performance does not drop below 41% (17% for visual-only targets). This demonstrates how the auditory stimulus provides a horizontal location, while what remains of the original visual target (or random noise) provides the vertical localization. Using crossmodal stimuli is clearly of benefit. Note that random noise in the visual stimuli can be selected by the maps for the localization. This results in the minimum 17% localization on visual-only stimuli.

4.3. Discussion

These results demonstrate that our model of the SC is capable of processing visual and auditory stimuli in real-time to form a localization to relevant multimodal stimuli. This localization performs well on single-modality stimuli, and is significantly enhanced for coincident multimodal stimuli, especially when the visual stimulus is weakened. However, the increase in performance is not a straightforward combination of single-modality performance since audio-visual localization achieves a maximum of 94% and a minimum of 41% when noise (variance 176 or 240) is applied. This compares to 99% for auditory stimuli alone. If we assume that each stimulus is treated independently in combined localization as per Stein et al. [67], then the probability of correctly locating an audio-visual stimulus

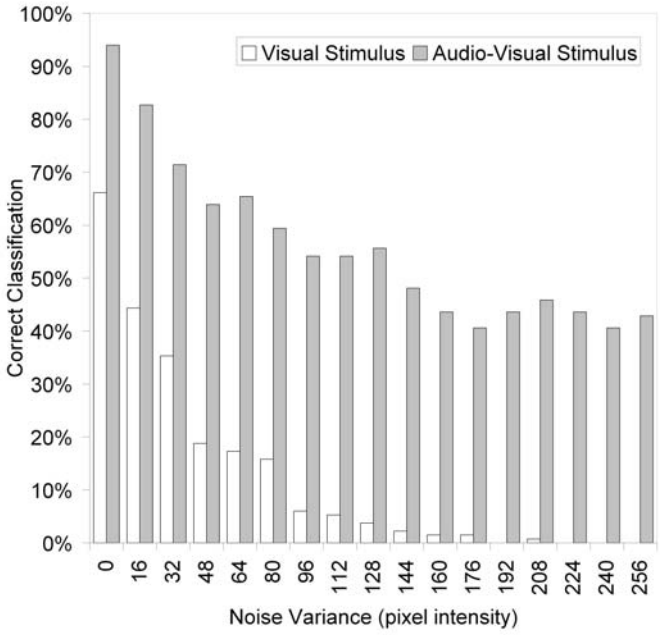


Figure 7: Visual vs. audio-visual localization performance for the 7 target locations with increasing levels of multiplicative noise applied. Audio-visual localization is maintained at a level above 41% even when the visual stimulus is obscured by noise from variance 192 and above.

P_{av} can be calculated using the independent probabilities of correctly responding to audio P_a and visual P_v stimuli:

$$P_{av} = (P_a + P_v) - (P_a \times P_v) \tag{10}$$

Since $P_a = 1$ for all targets, $P_{av} = 1$. This holds for any level of visual noise. However, we can see from our results that the actual performance does not achieve this. Stein et al. [67] found a similar response in cats, and concluded that the combination must therefore be multiplicative, rather than additive, and this led to investigations into multisensory enhancement and suppression [9].

In addition to this, two general models of sensory integration for spatial localization have been proposed. Knudsen and Brainard [68] reviewed audio-visual integration in the tectum and the forebrain to conclude that vision dominates whenever visual information is available. In contrast, Ernst and Banks [69] proposed a maximum-likelihood integrator in which the brain's estimate of the reliability of a given modality is used to weight the relative contribution the modality has to the integration. Both models have been evaluated by Battaglia et al. [70] who conclude that each is partially correct in that noise within a modality is taken into account in the integration, but that one sense still tends to dominate.

In our model the combination of responses is achieved through a linear weighting of the outputs from independent modality topographic maps, yet the resulting localization performance appears to be non-linear. For exam-

ple, for stimuli at $-15°$, we get 63% visual-only localization, 100% audio-only, but 63% combined (Table 5). This effect is caused by the competition between neurons in the multisensory map. With conflicting audio-visual localizations, the map forms a combined response by selecting a winning neuron which has the highest combined input. The net effect is that regions with lower responses which overlap can have higher combined responses than a single modality. It is therefore the overlap which becomes as important as a large single-modality response. This is a crude form of multisensory enhancement of coincident low intensity cues, versus suppression of high intensity single-modality cues and is demonstrated in the way that localization performance is maintained in our model even when the visual stimulus is degraded by noise.

Although our model can successfully combine audio-visual stimuli, the combination is fixed and is not based on maximum-likelihood. We selected weightings of each map within the hierarchy (Table 4) to ensure that any one map can provoke a multisensory response when presented with test stimuli consisting of noise. A limitation of our choice of weights was to bias activity towards motion detection, and hence our model is dominated by vision as per Knudsen and Brainard [68]. However, current work on the mechanisms underlying enhancement and suppression have shown that integration is more complex in that it depends upon coincident activity in sensory cortical areas feeding back into the SC [12]. This feedback may explain the findings of Battaglia et al. [70] such that the SC internally provides a visual bias, while cortical feedback performs the likelihood estimate to override the integration. In this paper we modeled integration without feedback, but we have explored how such feedback might be modeled with topographic maps representing cortical structures [24]. Alternate approaches using spiking neurons have also been explored in higher level vision [71]. So far such techniques are too crude to shed any further light on this phenomenon, yet they show promise in that they can be trained using conditioning-type feedback to respond differently to varying stimuli [54, 59], perhaps showing how coincident cues could be recognized and enhanced cortically.

5. Conclusion

We set out in this paper to demonstrate whether a biologically motivated, behavioral model of the SC could be applied to localize real-time audio-visual stimuli. With 12,240 neurons arranged into nine topographic maps, the model is capable of localizing in real-time (at two frames per second using a Java implementation). The localization performed is deliberately simple in order to match the computational capabilities of the SC. For visual sensitivity we focus on spatial sensitivity to detect broad patterns of high contrast, and temporal sensitivity to detect motion in preferred directions from the most intense changes extracted by frame differencing. Our auditory sensitivity is based upon locating the loudest sound. In the SC, this simple but fast sensitivity determines the focus of subsequent midbrain through to cortical sensory processing. For example, through a subcortical pathway to the amygdala, relevant multimodal stimuli can be assessed in more detail to detect potential threats [72].

The principle we are following is therefore to construct a hierarchy of increasingly more complex *functional* processing. Hierarchical approaches are often used in computational neuroscience models, but few aim at achieving functionality. For example, the large-scale model developed by Izhikevich and Edelman [34] is impressive in its scale and use of biologically plausible computational units (a model thalamocortical column). While the model is not designed to be functional, for us a key objection to the approach is the smaller amount of detail in modeling low level structures in comparison with the detail applied to the cortical neurons. This imbalance reflects the focus on cortical operation, which is an understandable focus in exploring higher cognition, but at the expense of low level detail. Without low level detail or the influence of real-world stimuli, such models may not be able to demonstrate comparative cognitive properties. Darwin X [35] is a good example of how a combination of levels of processing on real-world stimuli can be utilized, and this is something our model aspires to by showing how low level processing can be used practically, despite its deliberate simplicity. A natural next step in our model development is to explore whether a spiking model of the SC could achieve similar results to our rate-coded model and then use this to enhance models of the thalamocortical column to make them functional. The potential for this has been demonstrated with a hierarchical model of auditory localization in the inferior colliculus using leaky integrate-and-fire neurons [73].

To improve localization, we could implement different spatial maps with varying contrast sensitivity. The model of visual attention developed by Itti et al. [31] is a good example of this with a real-time implementation. Our approach is very similar to theirs in that we also use a hierarchy of topographic maps to select salient locations. Since they use cortical visual structures for motivation, their model is sensitive to color, intensity and orientation. With our focus on the SC, we model visual contrast and motion. In other respects our models are also very similar, such as the combination of hierarchical processing and competitive selection of saliency. Yet our model is distinguished by its more biologically motivated implementation of maps, plus the combination of auditory sensitivity which we use to explore how multimodal cues interact.

Our biological motivation is evident in the way we use temporal inhibition to detect motion. Sensory neurons use pre-filtering, delay filtering and non-linear interactions [62] to detect motion. We use simple frame differencing to pre-filter incoming stimuli, temporal inhibition between neurons to concentrate sensitivity on stimuli moving in preferred directions, and non-linear interactions between neurons to select the highest response. This combination of

simple mechanisms achieves high detection rates in simple scenes, but lower rates in scenes with conflicting motion stimuli. Further evaluation of this mechanism is needed to understand how it performs against computational, rather than biologically motivated, approaches [cf. 74].

For auditory localization we use ILD. This intensity-based approach suffers from a number of limitations. First, in order to detect energy differences between the two fixed microphones, they need to be calibrated. We chose two identically branded microphones which reduced the need for calibration, yet both varied in their performance. Reverberation within the room also impacts on localization. Second, the energy recorded at a microphone represents all locations on the radius of a sphere centered on the microphone, where the radius is determined by the comparative energy received by the second microphone. In order to localize we have to assume that the sound source is horizontal between the microphones (to avoid the 'cone of confusion'). With more microphones, some of these assumptions can be avoided. Third, if more than one sound source is present, this can disrupt the localization because the energies compared may come from different locations. To overcome these limitations a range of techniques can be used across frequency bands, or more accurate localization via calibrated equipment such as Microsoft's Kinect [64].

Returning back to the mammalian localization, an important aspect of the SC is its ability to automatically translate auditory spatial coordinates into visual coordinates. This means that we can move our eyes relative to our ears and still combine coincident cues. This is achieved through feedback from the SC to the IC. In our model we have assumed a fixed camera to microphone setup and a purely feedforward operation. Feedback from the multisensory space could be used to adjust the relative position of the eyes in the auditory space. An example of how this might be achieved is provided by Huo and Murray [27]. Notably their model uses spiking neurons with spike timing dependent plasticity to learn the alignment between multimodal stimuli. This type of approach, using more biologically plausible models of neurons and learning, coupled with cortical feedback and a movable camera position is the next step in our work.

Acknowledgments

We would like to thank Barry Stein for his helpful comments regarding the function and connectivity of the SC and the three anonymous reviewers for their detailed and helpful comments.

References

[1] H. McGurk, J. MacDonald, Hearing lips and seeing voices, Nature 264 (1976) 746–748.

[2] M. Botvinick, J. Cohen, Rubber hands 'feel' touch that eyes see, Nature 391 (1998) 756–756.

[3] V. S. Ramachandran, D. Rogers-Ramachandran, S. Cobb, Touching the phantom limb, Nature 377 (1995) 489–490.

[4] L. Shams, Y. Kamitani, S. Shimojo, Modulations of visual perception by sound, in: G. A. Calvert, C. Spence, B. E. Stein (Eds.), The Handbook of Multisensory Processes, A Bradford Book, MIT Press, 2004, pp. 27–33.

[5] G. A. Calvert, T. Thesen, Multisensory integration: Methodological approaches and emerging principles in the human brain, Journal of Physiology - Paris 98 (2004) 191–205.

[6] B. E. Stein, M. A. Meredith, The Merging of the Senses, A Bradford Book, MIT Press, Cambridge, MA., 1993.

[7] A. J. King, The superior colliculus, Current Biology 14 (2004) R335–R338.

[8] S. Katyal, S. Zughni, C. Greene, D. Ress, Topography of covert visual attention in human superior colliculus, Journal of Neurophysiology 104 (2010) 3074–3083.

[9] B. E. Stein, T. R. Stanford, Multisensory integration: Current issues from the perspective of the single neuron, Nature Reviews Neuroscience 9 (2008) 255–266.

[10] P. J. May, The mammalian superior colliculus: Laminar structure and connections, in: J. A. Buttner-Ennever (Ed.), Progress in Brain Research: Neuroanatomy of the Oculomotor System, volume 151, Elsevier, 2006, pp. 321–378.

[11] L. Yu, B. A. Rowland, B. E. Stein, Initiating the development of multisensory integration by manipulating sensory experience, Journal of Neuroscience 30 (2010) 4904–4913.

[12] J. C. Alvarado, T. R. Stanford, B. A. Rowland, J. W. Vaughan, B. E. Stein, Multisensory integration in the superior colliculus requires synergy among corticocollicular inputs, The Journal of Neuroscience 29 (2009) 6580–6592.

[13] A. K. Moschovakis, C. A. Scudder, S. M. Highstein, The microscopic anatomy and physiology of the mammalian saccadic system, Progress in Neurobiology 50 (1996) 133–254.

[14] E. K. C. Tsang, B. E. Shi, Active binocular gaze control inspired by superior colliculus, in: Proceedings of the International Joint Conference on Neural Networks (IJCNN) 2006, IEEE, 2006, pp. 7–14.

[15] M. C. Casey, D. L. Hickman, J. R. E. Sadler, Small-scale anomaly detection in panoramic imaging using neural models of low-level vision, in: Proceedings of SPIE Defense, Security, and Sensing Conference 2011 on Enhanced and Synthetic Vision, volume 8042B, SPIE, 2011.

[16] J. S. Lee, C. H. Park, Adaptive decision fusion for audio-visual speech recognition, in: F. Mihelic, J. Zibert (Eds.), Speech Recognition, InTech, 2008, pp. 275–296.

[17] P. Ritter, A. Villringer, Simultaneous EEG-fMRI, Neuroscience and Biobehavioral Reviews 30 (2006) 823–838.

[18] V. Navalpakkam, L. Itti, Modeling the influence of task on attention, Vision Research 45 (2005) 205–231.

[19] T. J. Anastasio, P. E. Patton, K. Belkacem-Boussaid, Using bayes' rule to model multisensory enhancement in the superior colliculus, Neural Computation 12 (2000) 1165–1187.

[20] P. E. Patton, T. J. Anastasio, Modeling cross-modal enhancement and modality-specific suppression in multisensory neurons, Neural Computation 15 (2003) 783–810.

[21] B. A. Rowland, T. R. Stanford, B. E. Stein, A model of the neural mechanisms underlying multisensory integration in the superior colliculus, Perception 36 (2007) 1431–1443.

[22] E. Magosso, C. Cuppini, A. Serino, G. D. Pellegrino, M. Ursino, A theoretical study of multisensory integration in the superior colliculus by a neural network model, Neural Networks 21 (2008) 817–829.

[23] C. Cuppini, M. Ursino, E. Magosso, B. A. Rowland, B. E. Stein, An emergent model of multisensory integration in superior colliculus neurons, Frontiers in Integrative Neuroscience 4 (2010).

[24] A. Pavlou, M. C. Casey, Simulating the effects of cortical feedback in the superior colliculus with topographic maps, in: Proceedings of the International Joint Conference on Neural Networks (IJCNN) 2010, IEEE, 2010.

[25] M. C. Casey, A. Pavlou, A behavioral model of sensory alignment in the superficial and deep layers of the superior colliculus, in: Proceedings of the International Joint Conference on Neural Networks (IJCNN) 2008, IEEE, 2008, pp. 2751–2756.

[26] J. Huo, A. Murray, L. Smith, Y. Zhijun, Adaptation of barn owl localization system with spike timing dependent plasticity, in: Proceedings of the International Joint Conference on Neural Networks (IJCNN) 2008, IEEE, 2008, pp. 155–160.
[27] J. Huo, A. Murray, The adaptation of visual and auditory integration in the barn owl superior colliculus with spike timing dependent plasticity, Neural Networks 22 (2009) 913–921.
[28] S. Grossberg, K. Roberts, M. Aguilar, D. Bullock, A neuronal model of multimodal adaptive saccadic eye movement control by superior colliculus, Journal of Neuroscience 17 (1997) 9706–9725.
[29] V. Cutsuridis, N. Smyrnis, I. Evdokimidis, S. Perantonis, A neural model of decision-making by the superior colliculus in an antisaccade task, Neural Networks 20 (2007) 690–704.
[30] N. Srinivasa, S. Grossberg, A head-neck-eye system that learns fault-tolerant saccades to 3-d targets using a self-organizing neural model, Neural Networks 21 (2008) 1380–1391.
[31] L. Itti, C. Koch, E. Nieber, A model of saliency-based visual attention for rapid scene analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (1998) 1254–1259.
[32] C. Siagian, C. K. Chang, R. Voorhies, L. Itti, Beobot 2.0: Cluster architecture for mobile robotics, Journal of Field Robotics 28 (2011) 278–302.
[33] H. Markram, The blue brain project, Nature Reviews Neuroscience 7 (2006) 153–160.
[34] E. Izhikevich, G. M. Edelman, Large-scale model of mammalian thalamocortical systems, Proceedings of the National Academy of Sciences of the USA 105 (2008) 3593–3598.
[35] A. K. Seth, G. M. Edelman, Distinguishing causal interactions in neural populations, Neural Computation 19 (2007) 910–933.
[36] H. de Garis, C. Shuo, B. Goertzel, L. Ruiting, A world survey of artificial brain projects, part i: Large-scale brain simulations, Neurocomputing 74 (2010) 3–29.
[37] B. Goertzel, R. Lian, I. Arel, H. de Garis, S. Chen, A world survey of artificial brain projects, part ii: Biologically inspired cognitive architectures, Neurocomputing 74 (2010) 30–49.
[38] J. Y. Lettvin, H. R. Maturana, W. S. Mcculloch, W. H. Pitts, What the frog's eye tells the frog's brain, Proceedings of the IRE 47 (1959) 1940–1951.
[39] P. Sterling, B. G. Wickelgren, Visual receptive fields in the superior colliculus of the cat, Journal of Neurophysiology 32 (1969) 1–15.
[40] J. P. Rauschecker, L. R. Harris, Auditory and visual neurons in the cat's superior colliculus selective for the direction of apparent motion stimuli, Brain Research 490 (1989) 56–63.
[41] M. T. Wallace, J. G. McHaffie, B. E. Stein, Visual response properties and visuotopic representation in the newborn monkey superior colliculus, Journal of Neurophysiology 78 (1997) 2732–2741.
[42] F. Prévost, F. Lepore, J. P. Guillemot, Spatio-temporal receptive field properties of cells in the rat superior colliculus, Brain Research 1142 (2007) 80–91.
[43] K. A. Schneider, S. Kastner, Visual responses of the human superior colliculus: A high-resolution functional magnetic resonance imaging study, Journal of Neurophysiology 94 (2005) 2491–2503.
[44] E. I. Knudsen, M. Konishi, A neural map of auditory space in the owl, Science 200 (1978) 795–797.
[45] J. C. Middlebrooks, E. I. Knudsen, A neural code for auditory space in the cat's superior colliculus, Journal of Neuroscience 4 (1984) 2621–2634.
[46] B. Grothe, M. Pecka, D. McAlpine, Mechanisms of sound localization in mammals, Physiological Reviews 90 (2010) 983–1012.
[47] A. J. King, A. R. Palmer, Cells responsive to free-field auditory stimuli in guinea-pig superior colliculus: Distribution and response properties, Journal of Physiology 342 (1983) 361–381.
[48] S. Sterbing, K. Hartung, K. P. Hoffmann, Representation of sound source direction in the superior colliculus of the guinea pig in a virtual auditory environment, Experimental Brain Research 142 (2002) 570–577.
[49] R. A. A. Campbell, T. P. Doubell, F. R. Nodal, J. W. H. Schnupp, A. J. King, Interaural timing cues do not contribute to the map of space in the ferret superior colliculus: A virtual acoustic space study, Journal of Neurophysiology 95 (2006) 242–254.
[50] J. C. Zella, J. F. Brugge, J. W. H. Schnupp, Passive eye displacement alters auditory spatial receptive fields of cat superior colliculus neurons, Nature Neuroscience 4 (2001) 1167–1169.
[51] A. J. King, W. H. Schnupp, I. D. Thompson, Signals from the superficial layers of the superior colliculus enable the development of the auditory space map in the deeper layers, Journal of Neuroscience 18 (1998) 9394–9408.
[52] A. J. V. Opstal, D. P. Munoz, Auditory-visual interactions subserving primate gaze orienting, in: G. A. Calvert, C. Spence, B. E. Stein (Eds.), The Handbook of Multisensory Processes, A Bradford Book, MIT Press, 2004, pp. 373–393.
[53] R. Linsker, Self-organization in a perceptual network, Computer 21 (1988) 105–117.
[54] J. L. Armony, D. Servan-Schreiber, J. D. Cohen, J. E. LeDoux, An anatomically constrained neural network model of fear conditioning, Behavioral Neuroscience 109 (1995) 246–257.
[55] J. L. Armony, D. Servan-Schreiber, L. M. Romanski, J. D. Cohen, J. E. LeDoux, Stimulus generalization of fear responses: Effects of auditory cortex lesions in a computational model and in rats, Cerebral Cortex 7 (1997) 157–165.
[56] R. Miikkulainen, J. A. Bednar, Y. Choe, J. Sirosh, Computational Maps in the Visual Cortex, Springer Science+Business Media, New York, 2005.
[57] D. E. Rumelhart, D. Zipser, Feature discovery by competitive learning, in: D. E. Rumelhart, J. L. McClelland (Eds.), Parallel Distributed Processing: Explorations in the Microstructure of Cognition, volume Volume 1: Foundations, MIT Press, 1986, pp. 151–193.
[58] D. O. Hebb, The Organization of Behavior: A Neuropsychological Theory, John Wiley and Sons, New York, 1949.
[59] A. Pavlou, M. C. Casey, Identifying emotions using topographic conditioning maps, in: M. Koeppen, N. Kasabov, G. Coghill (Eds.), Advances in Neuro-Information Processing: Proceedings of the 15th International Conference on Neuro-Information Processing, Lecture Notes in Computer Science 5506, Springer-Verlag, 2009, pp. 40–47.
[60] M. C. Casey, A. Pavlou, A. Timotheou, Mind the (computational) gap, in: Proceedings of the UK Workshop on Computational Intelligence (UKCI 2010), IEEE, 2010.
[61] T. Kohonen, Self-organized formation of topologically correct feature maps, Biological Cybernetics 43 (1982) 59–69.
[62] C. W. G. Clifford, M. R. Ibbotson, Fundamental mechanisms of visual motion detection: Models, cells and functions, Progress in Neurobiology 68 (2002) 409–437.
[63] C. Enroth-Cugell, J. G. Robson, The contrast sensitivity of retinal ganglion cells of the cat, Journal of Physiology 187 (1966) 517–552.
[64] Microsoft Corporation, Kinect for Windows SDK from Microsoft Research [WWW page], URL http://research.microsoft.com/en-us/um/redmond/projects/kinectsdk/, 2011.
[65] H. Cox, R. Zeskind, M. Owen, Robust adaptive beamforming, IEEE Transactions on Acoustics, Speech and Signal Processing 35 (1987) 1365–1376.
[66] S. T. Birchfield, R. Gangishetty, Acoustic localization by interaural level difference, in: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2005), volume 4, pp. iv/1109–iv/1112.
[67] B. E. Stein, M. A. Meredith, W. S. Huneycutt, L. McDade, Behavioral indices of multisensory integration: Orientation to visual cues is affected by auditory stimuli, Journal of Cognitive Neuroscience 1 (1989) 12–24.
[68] E. I. Knudsen, M. S. Brainard, Creating a unified representation of visual and auditory space in the brain, Annu. Rev. Neurosci. 18 (1995) 19–43.
[69] M. O. Ernst, M. S. Banks, Humans integrate visual and haptic information in a statistically optimal fashion, Nature 415 (2002)

429–433.
[70] P. W. Battaglia, R. A. Jacobs, R. N. Aslin, Bayesian integration of visual and auditory signals for spatial localization, J. Opt. Soc. Am. A 20 (2003) 1391–1397.
[71] S. Schrader, M. O. Gewaltig, U. Krner, E. Krner, Cortext: A columnar model of bottom-up and top-down processing in the neocortex, Neural Networks 22 (2009) 1055–1070.
[72] C. Shi, M. Davis, Visual pathways involved in fear conditioning measured with fear-potentiated startle: Behavioral and anatomic studies, The Journal of Neuroscience 21 (2001) 9844–9855.
[73] J. Liu, D. Perez-Gonzalez, A. Rees, H. Erwin, S. Wermter, A biologically inspired spiking neural network model of the auditory midbrain for sound source localisation, Neurocomputing 74 (2010) 129–139.
[74] G. Monaci, P. Vandergheynst, F. T. Sommer, Learning bimodal structure in audiovisual data, IEEE Transactions on Neural Networks 20 (2009) 1898–1910.

Matthew Casey received his BSc in Mathematics and Computer Science from the University of Kent in 1992. He then worked for 10 years for Data Sciences, IBM and Anite Telecoms as a software engineer and consultant. In 2004 he was awarded his PhD after studying part-time. His PhD evaluated neural network combinations for classification and for behavioral modeling, attempting to bridge the gap between computational neuroscience and computational intelligence. His research interests follow this theme with the development of more complex and biological plausible models of low-level sensory processing and integration, which have been applied to real-world problems. He is a Senior Lecturer in the Department of Computing at the University of Surrey.

Athanasios Pavlou received his PhD on Biologically Inspired Modeling and Applications from the University of Surrey in 2009. His research was focused on modeling neural structures of low-level visual processing aiming to develop computational frameworks that would be utilized within real-world applications and further the understanding of neurobiological behaviors. He is an IT consultant for BAE Systems Detica.

Anthony Timotheou is currently in his final year of study for a BSc in Computer Science from the University of Surrey. He has had a year on placement in industry writing software for government sectors such as the Department of Energy and Climate Change and the Marine Management Organization. His current dissertation is looking at modeling the auditory mid-brain using spiking neural networks. His research interests are in modeling the brain in order to gain an understanding of its processing by applying current and novel models to real-world problems.