Learning Multi-Sensory Integration with Self-Organization and Statistics

Johannes Bauer, Stefan Wermter
Department of Informatics
University of Hamburg
Germany
Email: {bauer,wermter}@informatik.uni-hamburg.de

Abstract

Recently, we have presented a self-organized artificial neural network algorithm capable of learning a latent variable model of its high-dimensional input and to optimally integrate that input to compute and population-code a probability density function over the values of the latent variables of that model. We did take our motivation from natural neural networks and reported on a simple experiment with simulated multi-sensory data. However, we focused on presenting the algorithm and evaluating its performance, leaving a comparison with natural cognition for future work. In this paper, we show that our algorithm behaves similar, in important behavioral and neural aspects, to a prime example of natural multi-sensory integration: audio-visual object localization.

1 Introduction

Imagine you are given a sheet of paper with unlabeled numbers, told that these numbers contain information about the value of some quantity, and asked what you think the value of that quantity is. This is clearly an impossible task. What if you are repeatedly shown such numbers, for different values of the quantity in question, but never given the right answer? Biological neurons face a similar situation. It is their function to produce activity corresponding to some quantity in the outside world. And all they have, to estimate that quantity, is the activity at their incoming synapses, which carries no information about its origin. A comparable situation also exists in unsupervised machine learning. Models of unsupervised neural learning have therefore found applications in general machine learning.

We have recently presented an artificial neural network (ANN) algorithm with possible applications in general machine learning based on the self-organizing map (SOM) [Bauer and Wermter, 2013]. This algorithm was inspired by the apparent ability of humans to utilize the sensory information they get in a statistically optimal fashion in many situations [Ernst and Banks, 2002; Landy *et al.*, 2011]. In particular, it aims to model how neural populations learn to make sense of uni- and cross-sensory stimuli. The result is an algorithm which takes high-dimensional data as input, learns a low-dimensional latent-variable model and computes for a given input a probability density function (PDF) over the values of the latent variables.

We have shown that our algorithm can perform near-optimally on uni-sensory input and we have shown that it can handle multiple sensory modalities with different response and noise characteristics. However, we have not fully closed the loop in 1) trying to understand a problem faced by a natural system, 2) trying to find a solution 3) comparing that solution to the one found, tried, and tested by nature [Jacobs and Kruschke, 2011; Landy *et al.*, 2011]. In this paper, we will therefore be concerned with the last step in this process, comparing our model's behavior to a biological example.

The superior colliculus (SC), a mid-brain region found in all vertebrates, is a prime candidate for this last step for a number of reasons: First, its deeper levels integrate information from vision, hearing, and touch to localize objects in space. Like our network, it thus uses high-dimensional input—the responses of uni-sensory input populations—to infer about a latent variable: the location of an object. Second, its physiology has been studied intensively and the relationship between input stimuli and overt behavior caused by SC activity is comparatively well-understood. Knowledge about and models of the SC can therefore serve as starting points for investigating other brain regions with similar tasks [Stein, 2012]. Third, as mentioned before, the SC is present in all vertebrates (being called optic tectum (OT) in non-mammals) and evolutionarily highly stable [Stein and Meredith, 1993]. The strategies it employs are therefore tried and tested indeed and likely to approach optimality.

We will compare our model's performance to two well-established principles of natural audio-visual object localization: On the neurophysiological level, we show that our network reproduces the effects of enhancement and depression depending on the spatial relation between auditory and visual stimuli found in single SC neurons [Stein and Meredith, 1993]. On the behavioral level, we demonstrate comparability with maximum likelihood estimation on the basis of uni-sensory localizations, shown for multi-sensory localization and several other cases of multi-sensory integration in humans [Alais and Burr, 2004; Ernst and Banks, 2002; Hillis *et al.*, 2004].

In the next section, we will first briefly review our algorithm, focussing on giving a good intuition of the main prin-

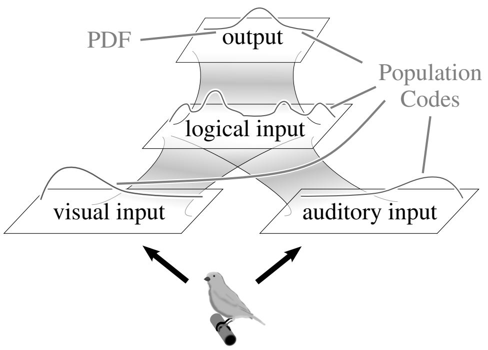


Figure 1: Visual and auditory input layers respond differently to multisensory input. They are concatenated into one logical input layer which is presented to our network, the output layer. Neurons in the output layer have no information about the origin of neurons in the logical input layer. The output layer learns to produce a population-coded PDF over the latent variable behind its input's activity.

ciples behind it, and referring to the original paper for details. We will then report on experiments we carried out in which we simulated various conditions of multi-sensory integration. We will review the known neurophysiological and psychophysical effects and compare them to our network's responses. Finally, we will interpret our results in the broader scope of computational neuroscience and its interaction with general artificial intelligence.

2 The Model

Before we start describing our solution, let us again look at the problem we are trying to solve. A population of neurons is to collaborate in learning to compute a PDF for the latent variables behind patterns of neural activity.[1] This input activity can be uni- or cross-sensory; conceptually, both can be treated the same by introducing a logical population which simply concatenates the separate input populations' activity vectors (see Fig. 1).

Our approach is to have the network learn to represent the PDF in a population code, where each neuron codes for the probability of a different value being the true value of the latent variable [Pouget *et al.*, 2003]. ANN models working with such population-coded PDFs have been proposed eg. by Cuijpers and Erlhagen [2008] and Beck *et al.* [2008]. Our model focuses on how computing such a PDF from arbitrary neural responses can be learned unsupervised and without heavy assumptions on the noise properties. The population code realized by our network is to be spatially organized, ie. close-by neurons code for similar values of the latent variable. This restriction on the more general definition of a population code seems natural and it also reflects biological evidence of topographic maps in various sensory brain areas [Stone, 2012; Hyde and Knudsen, 2000; Stein and Stanford, 2013; Kaas, 1997].

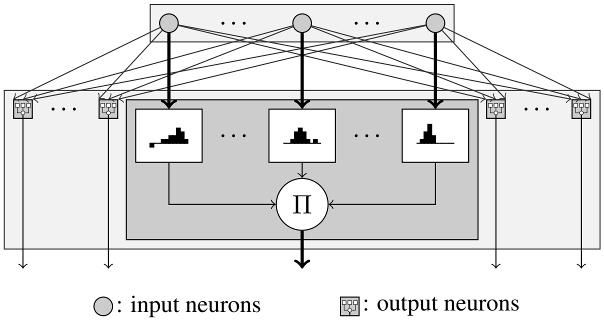


Figure 2: Structure of a single neuron within the output layer. Input neurons can be from any of the actual input populations, in no particular order. Each one is connected to all neurons in the output layer. Each output neuron learns 1) a preferred value of the latent variable, and 2) the statistics of its input given that preferred value of the latent variable. The network output is a PDF over the values of the latent variable.

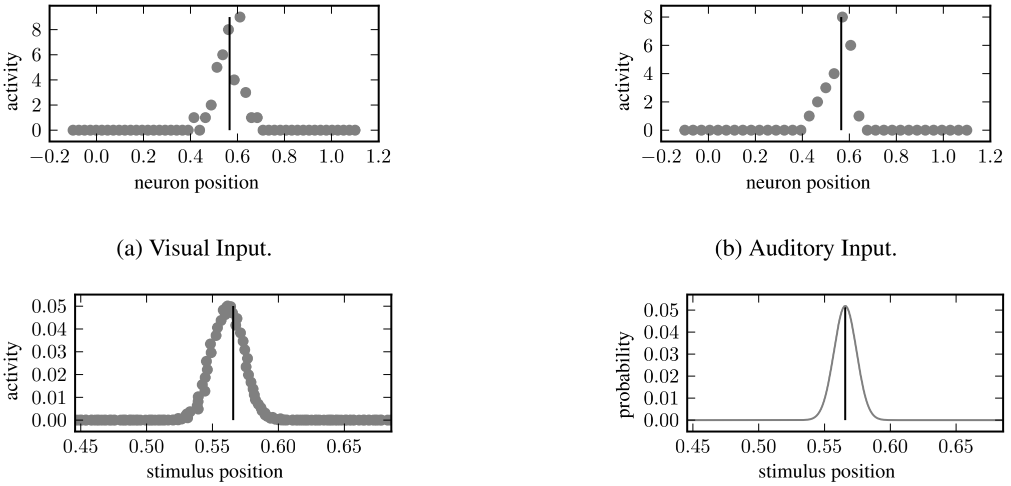
Kohonen's SOM [Kohonen, 1982] was inspired by the formation of such topographic maps in the brain. It is a self-organizing ANN algorithm which has been shown to be able to learn latent-variable models [Yin, 2007; Klanke, 2007]: It learns topography-preserving mappings from points in a data space to its spatially ordered units, or neurons. Since a SOM models a population of neurons and each neuron has a response to a stimulus, it is possible to read out a population code from a SOM [Zhou *et al.*, 2011].

In a standard SOM, the response is just the Euclidean distance of the stimulus as a vector from the preferred stimulus of each neuron. This response is used to find the best-matching unit (BMU), the neuron with the strongest response and the neuron the stimulus is mapped to. In our algorithm, the network simultaneously learns the latent-variable model and the statistics (and noise properties) of the input (see Fig. 2). The response of each neuron then is an estimate of the probability of the neuron's preferred value being the actual value of the latent variable, given what the network knows about the noise. This is done basically by keeping weighted counts of previous activities at each input connection of each neuron. Input activities at the neurons' synapses are discrete and treated non-metrically. Therefore, the algorithm lends itself to learning problems where data points have nominal dimensions. The main benefit over previous approaches [Zhou *et al.*, 2011; Bauer *et al.*, 2012b] is that our approach does not assume any specific noise model. We refer to our original paper [Bauer and Wermter, 2013] for details on the learning algorithm and its motivation and derivation.

3 Experiments

In the experiments described below, we will compare our network's performance and response properties to those found in psychophysical and neurophysiological studies. Specifically, we will examine the responses of our network in light of biological data about the SC. The SC is an evolutionarily stable midbrain structure concerned with localizing objects in space on the basis of visual, auditory, and somatosensory stimuli. It is involved in generating orienting movements on the basis

[1] From now on, we will assume wlog. a single latent variable. Note that strictly speaking a combination of latent variables can be seen as a single complex latent variable.



(a) Visual Input.

(b) Auditory Input.

(c) Network Output.

(d) PDF Recovered from Input.

Figure 3: Input, Network Response, and PDF Recovered From Input.

of uni- and cross-sensory stimuli [Stein and Stanford, 2013]. We choose it as a standard to which to compare our model because its input-output behavior is relatively well-understood and because it is likely that the general principles of SC functioning are realized similarly in other brain regions with comparable tasks [Stein, 2012].

For our simulations, we used a network of 500 output neurons connected to two populations $\mathbf{i}_{a,1}, \mathbf{i}_{a,2} \ldots, \mathbf{i}_{a,35}$ and $\mathbf{i}_{v,1}, \mathbf{i}_{v,2} \ldots, \mathbf{i}_{v,50}$ of input neurons. The two input populations each represented one sensory modality. Each input neuron $\mathbf{i}_{m,k}$ had a preferred value of $\mathfrak{p}_{m,k}$ such that the $\mathfrak{p}_{m,k}$ were evenly distributed over the interval $[-0.125, 1.125]$, for $m \in \{a, v\}$. The neuron $\mathbf{i}_{m,k}$ responded to a stimulus $\mathfrak{p} \in [0,1]$ according to a poisson-noisy Gaussian:

$$\mathbf{a}_{m,k}(\mathfrak{p}) \sim \Pr(a_m \exp(-(\mathfrak{p}_k - \mathfrak{p})^2/\sigma_m^2)), \qquad (1)$$

for $a_a = 4, \sigma_a^2 = 0.01$ and $a_v = 7, \sigma_v^2 = 0.005$. This models receptive fields of neurons in the visual and auditory layers of the SC. There, neurons respond most strongly to stimuli from some direction, depending on their position in these layers, and less or not at all to stimuli from other directions [Stein and Stanford, 2013]. This spatiotopic organization is a feature found throughout the brain, and it is present also in neural populations projecting to the SC like the retina, the lateral geniculate nucleus, and the external nucleus of the inferior colliculus [Stone, 2012; Gutfreund and King, 2012]. In the following, the neurons $\mathbf{i}_{a,k}$ and $\mathbf{i}_{v,k}, k = 1, 2, \ldots$ will be referred to as ‘auditory’ and ‘visual’, respectively, to make presentation more intuitive.

We trained the network with congruent stimuli until it had developed spatial organization and learned the simulated input noise statistics. Figs. 3a, 3b, and 3c show typical input in the visual and auditory input populations, and the network’s response, respectively, after training. Fig. 3d shows a PDF we recovered from the input using knowledge of the response properties of visual and auditory input neurons. The network did not have access to this knowledge. We then started simulating our chosen psychophysical and neurophysiological experiments. For these experiments, we changed the simulated stimulus conditions as will be described below.

Enhancement/Depression. It is a well-established fact that multi-sensory SC neurons tend to react more strongly to cross-sensory stimuli in their receptive fields than to uni-sensory stimuli [Stein and Meredith, 1993]. This effect is called enhancement. Depression, on the other hand, occurs when stimuli are temporally or spatially incongruent: In the spatial case, this means that the reaction of a multi-sensory neuron to a visual stimulus in its receptive field will actually be weaker if that stimulus is accompanied by a sound coming from a different point in space (and vice-versa).

We simulated this condition by presenting, in each trial, one random stimulus $\mathfrak{p}_a, \in [0,1]$ to neurons $\mathbf{i}_{a,k}$ and a different stimulus $\mathfrak{p}_v \in [0,1]$ to neurons $\mathbf{i}_{v,k}$. We recorded the network’s response to the combined, incongruent, cross-sensory input population response. Fig. 4 shows the mean response over all trials of the output neuron at whose center was the visual stimulus depending on the absolute distance of the incongruent auditory stimulus. Although somewhat noisy, the graph clearly shows that congruent stimuli elicit much stronger responses than incongruent responses, which is in accordance with the phenomena of enhancement and depression explained above.

MLE. The effects discussed so far are on the level of single multi-sensory neurons. Since these neurons are part of a sensory-motor processing circuit, it is to be expected that they manifest themselves in observable behavior. Alais and Burr found that their test subjects’ performance in an audio-visual localization task was consistent with a maximum likelihood estimator (MLE) model of multi-sensory integra-

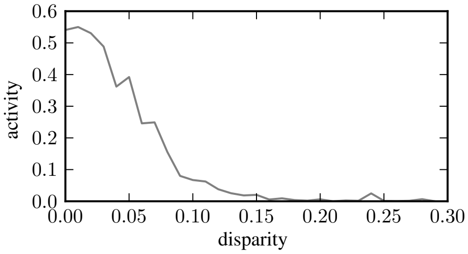


Figure 4: Response of an Output Neuron to a Visual Stimulus in its Receptive Field and an Auditory Stimulus at Various Distances from the Visual Stimulus.

tion [2004]. Other authors have found similar results for different combinations of sensory cues [Ernst and Banks, 2002; Hillis *et al.*, 2004].

The model used by Alais and Burr assumes Gaussian noise in sensory localizations. Under this assumption, MLE integrates two uni-sensory localizations l_a, l_v optimally according to a linear combination:

$$l_{MLE} = \frac{r_a}{r_a + r_v} l_a + \frac{r_v}{r_a + r_v} l_v, \tag{2}$$

where $r_m = \frac{1}{\sigma^2}$ is the reliability of a modality m, if σ is the standard error of localizations by that model, that is, the mean absolute error between the localization and the actual location of the target.

The expected reliability r_{MLE} of the combined result is given by:

$$\frac{1}{r_{MLE}} = \frac{1}{r_a} + \frac{1}{r_v}. \tag{3}$$

The distribution of errors of the combined estimator, like the assumed distribution of errors of the individual modalities' estimators, is Gaussian.

First, we determined the distributions of errors of our model given uni-sensory and cross-sensory stimuli. To do that, we fed our network with input in which either only auditory neurons $\mathbf{i}_{a,k}$, or only visual neurons $\mathbf{i}_{v,k}$ had non-zero activity (according to Eq. 1), or both, as usual. Figure 5 shows histograms of errors (mislocalizations) for uni- and cross-sensory localization, as well as Gaussian functions fitted to these errors. It can be seen that the distribution of errors is Gaussian-like, and that combined localization has much greater reliability than either of the individual localizations. Closer analysis reveals that the standard deviations of auditory-only, visual-only, and cross-modal localization are $\sigma_a = 4.594 \times 10^{-4}, \sigma_v = 1.061 \times 10^{-4}$, and $\sigma_m = 8.135 \times 10^{-5}$, respectively. The expected cross-modal localization error according to Eq. 3 would be $\sigma_{m,e} = 4.185 \times 10^{-5}$. We attribute this discrepancy to sampling error,[2] outliers, and actual learning errors. All in all, both visual inspection of the distribution of errors and this analysis demonstrate that our network effectively integrates the information in its multi-sensory input.

To test whether the behavior of our network is consistent with the MLE model described above, we conducted another experiment. As in the first experiment, we again chose one auditory stimulus $\mathfrak{p}_a$ and a visual stimulus $\mathfrak{p}_v$ in every trial. This time, each trial consisted of three conditions: an auditory, a visual, and a cross-sensory condition. In the auditory condition, we combined the normal auditory population response (Eq. 1) with a flat response of all-zero activity. The visual condition was analogous and in the cross-sensory condition, the population responses were combined as usual. In each trial n, the input was presented to the model and the auditory, visual, and cross-sensory localizations $l_{a,n}, l_{v,n}, l_{c,n}$ were recorded.

We then computed the least-squares solution to the equation

$$p_a \begin{pmatrix} l_{a,1} \\ l_{a,2} \\ \vdots \\ l_{a,N} \end{pmatrix} + p_v \begin{pmatrix} l_{v,1} \\ l_{v,2} \\ \vdots \\ l_{v,N} \end{pmatrix} = \begin{pmatrix} l_{m,1} \\ l_{m,2} \\ \vdots \\ l_{m,N} \end{pmatrix},$$

where $N = 10\,000$ is the number of trials. We found $p_a = 1.680 \times 10^{-1}$ and $p_v = 8.272 \times 10^{-1}$ which is close to the optimal values $\hat{p}_a = 1.876 \times 10^{-1}$, $\hat{p}_v = 8.124 \times 10^{-1}$ obtained by inserting σ_a and σ_v into Eq. 2.

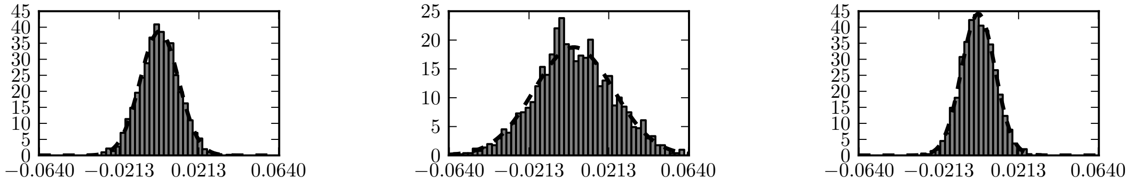
Together, these results show that our algorithm is not only statistically well-motivated and shows response characteristics similar to that of a biological information processing system, the SC, as was found in the first experiment: Its behavior on the functional level is also comparable to the optimal cue combination behavior demonstrated in human multi-sensory integration. This is especially interesting for our algorithm as a general machine learning algorithm.

4 Discussion

In this paper, we have shown that the neural learning algorithm introduced previously is able not only to integrate multi-sensory input, but also mimics biology both on the single-neuron and behavioral level. We can therefore interpret our network as a model of the SC, as it develops a representation of sensory input space, integrates percepts from different modalities depending on their reliabilities, uses the statistics of the input to learn this, and incorporates concepts known to be key in the SC, like population coding, winner-take-all, and local interactions.

We strongly believe that both fields, life sciences and artificial intelligence, will benefit from the approach of modeling observed biology to generate biological research questions, and implementing models in technical systems to validate their fitness and real-world applicability (see Fig. 6). Therefore, the next step will be validating our model's functionality and resemblance of biology in a robotic implementation. Initially, our experiments will mimic classical experiments like the ones due to Stein and Meredith [1993], which originally established the properties of multisensory integration in the cat SC: In our versions of these experiments, a robot will take the place of the feline or human subjects. It will be exposed to very similar multi-sensory stimuli as the

[2] Estimation is done by selecting the winner neuron and choosing its preferred value as the estimate. Since there are only finitely many neurons but infinitely many rationals in $[0, 1]$, estimation is bound to make sampling errors.



(a) Visual.

(b) Auditory.

(c) Audio-Visual.

Figure 5: Histograms of Distances between Visual, Auditory, and Audio-Visual Localization and Stimulus.

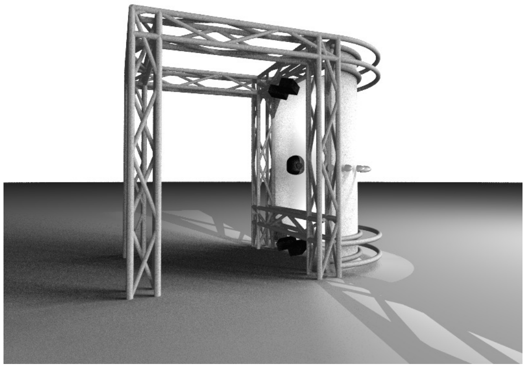


Figure 7: 3D Model of the Virtual Reality Robot Environment: A multi-purpose aluminium structure holds four projectors and a 180° projection screen. Around the screen, there is an array of speakers. The robot head is placed at the center of the half-cylinder spanned by the screen.

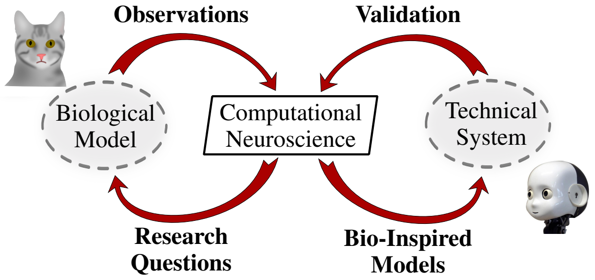


Figure 6: Research Cycle in Computational/Robotic Neuroscience

subjects in the original experiments. And its behavior and simulated neural processing will be monitored and compared to the original findings.

For these experiments, we will use the virtual reality lab infrastructure recently implemented ([Bauer *et al.*, 2012a], see Fig. 7). Designed and built as a basis for robotic sensory and cognitive experiments, this Virtual Reality for Robots Lab features a 180° projection screen and a matrix of 18 speakers. It allows us to precisely control the conditions of audio-visual localization experiments and still deliver rich and life-like stimuli to the iCub robot head which is placed at its center [Beira *et al.*, 2006]. With feedback from these experiments, we will extend our model and aim to accommodate attentional effects. This will give our model greater explanatory power and, at the same time, make it more flexible and more widely applicable in robotic and other AI systems. Again, biological experiments like those by Spence *et al.* [2004], which studied the effects of priming on multi-sensory integration, will guide our modeling efforts and serve as models for robotic experiments.

Acknowledgements

This work is funded by the DFG German Research Foundation (grant #1247) – International Research Training Group CINACS (Cross-modal Interactions in Natural and Artificial Cognitive Systems).

References

[Alais and Burr, 2004] David Alais and David Burr. The ventriloquist effect results from near-optimal bimodal integration. *Current Biology*, 14(3):257–262, February 2004.

[Bauer and Wermter, 2013] Johannes Bauer and Stefan Wermter. Self-organized neural learning of statistical inference from high-dimensional data. In *Proceedings of the International Joint Conference of Artificial Intelligence 2013*, 2013. To appear.

[Bauer *et al.*, 2012a] Johannes Bauer, Jorge Dávila-Chacón, Erik Strahl, and Stefan Wermter. Smoke and mirrors — virtual realities for sensor fusion experiments in biomimetic robotics. In *Multisensor Fusion and Integration for Intelligent Systems (MFI), 2012 IEEE Conference on*, pages 114–119. IEEE, 2012.

[Bauer *et al.*, 2012b] Johannes Bauer, Cornelius Weber, and Stefan Wermter. A SOM-based model for multi-sensory integration in the superior colliculus. In *The 2012 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, June 2012.

[Beck *et al.*, 2008] Jeffrey M. Beck, Wei J. Ma, Roozbeh Kiani, Tim Hanks, Anne K. Churchland, Jamie Roitman, Michael N. Shadlen, Peter E. Latham, and Alexandre Pouget. Probabilistic population codes for bayesian decision making. *Neuron*, 60(6):1142–1152, December 2008.

[Beira *et al.*, 2006] Ricardo Beira, Manuel Lopes, Miguel Praça, José Santos-Victor, Alexandre Bernardino, Giorgio

Metta, Francesco Becchi, and Roque Saltarén. Design of the robot-cub (icub) head. *Robotics and Automation, 2006. ICRA 2006. Proceedings 2006 IEEE International Conference on*, pages 94–100, May 2006.

[Cuijpers and Erlhagen, 2008] Raymond H. Cuijpers and Wolfram Erlhagen. Implementing Bayes' rule with neural fields. In *Proceedings of the 18th international conference on Artificial Neural Networks, Part II*, ICANN '08, pages 228–237, Berlin, Heidelberg, 2008. Springer-Verlag.

[Ernst and Banks, 2002] Marc O. Ernst and Martin S. Banks. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, 415(6870):429–433, January 2002.

[Gutfreund and King, 2012] Yoram Gutfreund and Andrew J. King. What is the role of vision in the development of the auditory space map? In Barry E. Stein, editor, *The New Handbook of Multisensory Processing*, chapter 32, pages 473–587. MIT Press, Cambridge, MA, USA, June 2012.

[Hillis *et al.*, 2004] James M. Hillis, Simon J. Watt, Michael S. Landy, and Martin S. Banks. Slant from texture and disparity cues: Optimal cue combination. *Journal of Vision*, 4(12):967–992, December 2004.

[Hyde and Knudsen, 2000] Peter S. Hyde and Eric I. Knudsen. Topographic projection from the optic tectum to the auditory space map in the inferior colliculus of the barn owl. *The Journal of Comparative Neurology*, 421(2):146–160, May 2000.

[Jacobs and Kruschke, 2011] Robert A. Jacobs and John K. Kruschke. Bayesian learning theory applied to human cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2(1):8–21, 2011.

[Kaas, 1997] Jon H. Kaas. Topographic maps are fundamental to sensory processing. *Brain Research Bulletin*, 44(2):107–112, January 1997.

[Klanke, 2007] Stefan Klanke. *Learning Manifolds with the Parametrized Self-Organizing Map and Unsupervised Kernel Regression*. PhD thesis, University of Bielefeld, March 2007.

[Kohonen, 1982] Teuvo Kohonen. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43(1):59–69, January 1982.

[Landy *et al.*, 2011] Michael S. Landy, Martin S. Banks, and David C. Knill. Ideal-observer models of cue integration. In Julia Trommershäuser, Konrad Körding, and Michael S. Landy, editors, *Sensory Cue Integration*, chapter 1, pages 5–29. Oxford University Press, Oxford, August 2011.

[Pouget *et al.*, 2003] Alexandre Pouget, Peter Dayan, and Richard S. Zemel. Inference and computation with population codes. *Annual review of Neuroscience*, 26(1):381–410, 2003.

[Spence *et al.*, 2004] Charles Spence, John McDonald, and Jon Driver. Exogenous spatial-cuing studies of human crossmodal attention and multisensory integration. In Charles Spence and Jon Driver, editors, *Crossmodal Space and Crossmodal Attention*, chapter 11, pages 277–320. Oxford University Press, USA, May 2004.

[Stein and Meredith, 1993] Barry E. Stein and M. Alex Meredith. *The Merging Of The Senses*. Cognitive Neuroscience Series. MIT Press, 1 edition, January 1993.

[Stein and Stanford, 2013] Barry E. Stein and Terrence R. Stanford. *Development of the Superior Colliculus/Optic Tectum*, pages 41–59. Elsevier, 2013.

[Stein, 2012] Barry E. Stein. Early experience affects the development of multisensory integration in single neurons of the superior colliculus. In Barry E. Stein, editor, *The New Handbook of Multisensory Processing*, chapter 33, pages 589–606. MIT Press, Cambridge, MA, USA, June 2012.

[Stone, 2012] James V. Stone. *Vision and Brain: How We Perceive the World*. The MIT Press, 1 edition, September 2012.

[Yin, 2007] Hujun Yin. Learning nonlinear principal manifolds by self-organising maps. In *Principal Manifolds for Data Visualization and Dimension Reduction*, Lecture Notes in Computational Science and Engineering, chapter 3, pages 68–95. Springer, Dordrecht, 2007.

[Zhou *et al.*, 2011] Tao Zhou, Piotr Dudek, and Bertram E. Shi. Self-organizing neural population coding for improving robotic visuomotor coordination. In *Neural Networks (IJCNN), The 2011 International Joint Conference on*, pages 1437–1444. IEEE, July 2011.