In: Proc. International Joint Conference on Neural Networks (IJCNN), 2012

A SOM-Based Model for Multi-Sensory Integration in the Superior Colliculus

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

***Abstract*—We present an algorithm based on the self-organizing map (SOM) which models multi-sensory integration as realized by the superior colliculus (SC). Our algorithm differs from other algorithms for multi-sensory integration in that it learns mappings between modalities' coordinate systems, it learns their respective reliabilities for different points in space, and uses mappings and reliabilities to perform cue integration. It does this in only one learning phase without supervision and such that calculations and data structures are local to individual neurons. Our simulations indicate that our algorithm can learn near-optimal integration of input from noisy sensory modalities.**

I. Introduction

The superior colliculus (SC) is a mid-brain area which receives input from a number of sensory modalities and integrates it to localize a possible common origin of the individual cues. The SC uses this integrated localization to generate motor commands. For instance, hearing a bird sing and perceiving its motion in the periphery of our visual field e.g. might lead to a saccade, i.e. a fast eye movement which directs our eyes towards the origin of the stimuli, the bird [1].

Two main challenges arise in this task: first, the conceptual coordinate systems in which sensory modalities code their output must be aligned. Second, noisy, sometimes contradictory information from different modalities needs to be integrated taking into account their reliabilities.

In the context of the SC, uni- and multi-sensory localizations are coded spatiotopically: ganglion cells from the retina, e.g., project to different locations in the SC depending on where in the retina they originate. Auditory localization cues also reach the SC spatiotopically organized via the external nucleus (ICx) of the inferior colliculus (IC), with the different cues leading to horizontal and vertical localization already integrated [2]. Thus, a biological formulation of the problem of alignment is how to group neurons pertaining to different sensory modalities by the direction in real space for which they stand. For example, there may be two neurons, one carrying information from the retina, one from the ICx, which both fire most strongly whenever a stimulus in their modality is located at 5° right and 10° upwards from the center of the field of view. In order for their signals to be integrated, these neurons must project to the same location in the SC.

A more abstract, algorithmic formulation assigns each neuron from one modality coordinates in a modality-specific coordinate system. The problem then becomes transforming coordinates from these modality-specific coordinate systems into one unified coordinate system.

The problem of integrating cues from different, noisy sensory modalities can, again, be described taking vision and hearing as an example: say, an audio-visual stimulus is right in front of us. Then vision will provide us with a very exact estimate of the location of that, while hearing may be off even in the best of cases by a degree or more [3]. If both cues are to be integrated—and they should be, because hearing does provide valuable information [4]—then they need to be weighted differently. Exactly how reliable cues from different modalities are is not equal throughout space, as e.g. accuracy of vision is much greater at the center of the visual field than in the periphery [5].

The effects of sensory deprivation on the spatial organization of the SC described by e.g. Knudsen and Brainard [6], the observation that sensory accuracy is different between individuals [7], [8], and the fact that neurophysiological and behavioral evidence for cue integration are not found until considerable time after birth [9], [10] suggest that both coordinate transformation and integration of multi-sensory cues are subject to learning in the early stages of development.

Various modeling approaches have been proposed for coordinate transformation, integration, and their learning, at different levels of abstraction from biological reality. Mathematically, the maximum likelihood estimator (MLE) has been a very successful means of modeling and explaining biological multi-sensory integration in a variety of tasks including in object localization [4], [10]–[12]. In that setting, and under certain simple and plausible assumptions, its application amounts to a simple linear combination of two or more sensory estimates of an object's location. For the MLE to be used in this way, the reliability of each sensory modality must be known and be equal at every point in space. Depending on the situation, statistical learning methods like the EM algorithm [13], [14] can be used to learn these reliabilities.

Most of the current artificial neural network (ANN) models at levels of high biological plausibility focus on replicating neurophysiological observations like the well-known suppression and enhancement effects and inverse effectiveness [1], spatiotopic organization, as well as interaction of the SC with higher cortical areas ([15]–[18], see also Rowland and Stein [19] for a review). Some develop coordinate transformation using SOM or SOM-like algorithms [17], [18]. None of

these models, however, consider the different reliabilities of the input modalities.

Weisswange et al. [20] on the other hand apply ANN learning to the problem of learning optimal cue combination. Their version of the cue combination task is more complex than the one we are considering in this paper in that cues from different modalities may or may not originate from one event. Weisswange et al. show that their learner performs similar to Bayesian model averaging in this task. Their approach, however, requires a teaching signal in order to develop spatial organization, and learn coordinate transformation and modalities' reliabilities.

Ghahramani [21] proposes an unsupervised learning algorithm in which units in a grid—like the neurons in the SC—adapt their Gaussian receptive fields with respect to two sensory modalities such that they map and integrate input from these modalities. His algorithm is firmly grounded in information theory and, like ours, motivated by multi-sensory integration in the SC. However, it requires knowledge of coordinate transformation as input. Also, as the author notes, calculations are quite involved and non-local in this algorithm and it does not immediately support more than two sensory modalities.

We will show in this paper how an unsupervised neural network algorithm based on Kohonen's SOM can learn to integrate cues from three or more sensory modalities for object localization, modeling multi-sensory integration in the SC. The algorithm will be introduced in Section II. In Section III, we will report on our simulations which show that, our algorithm can learn statistically near-optimal integration. The discussion in Section IV of our model, our results, and how our approach may be extended with ideas from related work will conclude this paper.

II. Multi-Sensory Integration Using SOMs

After briefly reviewing the SOM algorithm in Section II-A, establishing some of the concepts and notation used throughout this paper, we will, in Sections II-B and II-C, describe how it can in general solve the two great problems arising in multi-sensory localization: transformation between coordinate systems, and optimally combining signals from modalities with different reliabilities. In Section II-D, we will present the main contribution of this article: an extension of the original SOM algorithm which lets the SOM learn what the reliability of each input modality is and how it varies in space.

A. SOM

Self-organizing feature maps (SOM), or Kohonen maps ([22], [23]), are an abstract neural network algorithm which has been very successfully applied to various problems since it was introduced in 1982. In short, a SOM is a mapping $\mathfrak{i}$ from some input domain D with a distance measure dist between elements into a one- or multidimensional, finite grid U of SOM units, and a mapping $\mathfrak{e}$ from U to D such that

$$\mathfrak{e}(u) = v \quad \Rightarrow \quad \mathfrak{i}(v) = u.$$

and, for all $u, u' \in U$, $v \in D$

$$\mathfrak{i}(v) = u \quad \Rightarrow \quad \mathsf{dist}(\mathfrak{e}(u), v) \leq \mathsf{dist}(\mathfrak{e}(u'), v).$$

We say that the unit $u_B = \mathfrak{i}(v)$ is the best matching unit (BMU) for the input v. The symbols $\mathfrak{i}$ and $\mathfrak{e}$ are chosen to reflect the intuition about them in this paper: integration and extrapolation.

A SOM's mapping is learned from a set of data points by repeatedly selecting a data point v, finding the BMU u_B for v, and updating u_B and units within a certain neighborhood around u_B such that they are closer to v wrt. dist. During training, the size of the neighborhood is shrunk and the amount by which neighbors are updated is decreased. This unsupervised algorithm generally tends to generate a mapping which preserves the structure [24] of the input space in that, for three data points v_1, v_2, v_3, it tends to be true that

$$|\mathfrak{i}(v_1) - \mathfrak{i}(v_2)| \leq |\mathfrak{i}(v_1) - \mathfrak{i}(v_3)|$$

if

$$\mathsf{dist}(v_1, v_2) < \mathsf{dist}(v_1, v_3).$$

In this work, we are particularly interested in SOMs as a means of manifold learning [24], [25] or function approximation [26].

B. Coordinate Transformation

Integration of signals from n sensory modalities as realized by the SC can be described as mapping input tuples from $(\mathbb{Q}^2)^n$ to $\mathbb{Q}^2$ and SOMs are an algorithm for learning structure-preserving mappings from some domain to a grid. Together, this motivates the concept of modeling the SC using SOMs. A simple SOM-based architecture can take n inputs $v_1, v_2, \ldots, v_n$, one per sensory modality, each its respective modality's estimate of the location of the current stimulus. It can turn them into one n-dimensional vector $v = (v_1, v_2, \ldots, v_n)$ and feed that to the SOM (see Figure 1). The SOM will then learn which coordinates in the combined vector go together and thus how to convert between the individual modalities' coordinate systems regardless of different scales or orientations (see Figure 1).

C. Noise and the Distance Measure

Let us first define how we model noise. For our purposes, we assume noise in modalities to be independent and governed by Gaussian distributions, i.e. assuming a noise standard deviation for some modality $\mathcal{M}_i$ is σ_i, then the probability density for the distance d between a perceived point and its real origin is

$$N(d, \sigma_i) = \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{d^2}{2\sigma_i^2}}. \tag{1}$$

We call σ_i the intensity of the noise in modality $\mathcal{M}_i$. The reliability of $\mathcal{M}_i$ is defined as $r_i = \frac{1}{\sigma_i^2}$ [4], [21], [27].

In the following, we will show how a SOM can in principle deal with noise in the input data. We will limit ourselves to 1-dimensional SOMs and coordinate mapping. Extending the

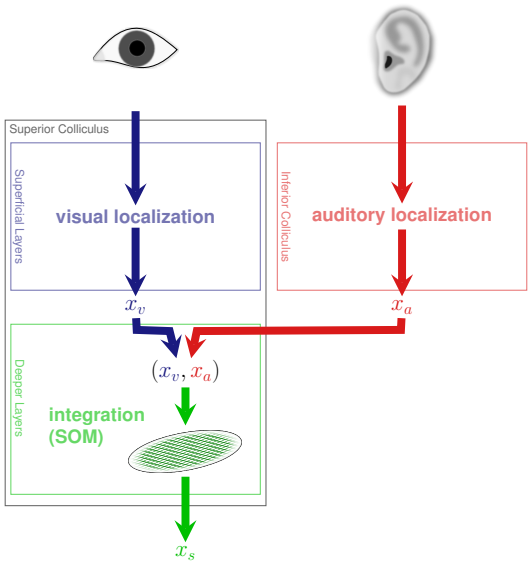
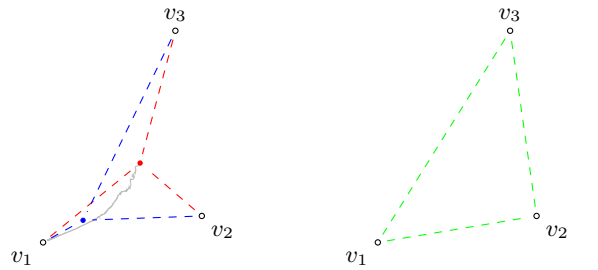


Fig. 1: Learning Simple Coordinate Transformation Using a SOM



(a) Fitting Distances between Estimates and SOM Units

(b) Fitting Distances between Estimates

Fig. 2: Ideas for Learning Sensory Reliabilities (Explanations in Text)

algorithm to more dimensions is straightforward although not without implications, as will be described in Section II-E.

Recall that there needs to be a distance function defined between elements of the domain of the learned mapping. Quite often, when implementing SOMs, this distance function is simply the Euclidean distance between vectors. If, however, the distance function uses knowledge of the sensory modalities' reliabilities to weight the input's components, then it can not only convert between modalities, but also make a better guess at the common origin of a number of signals than any modality does on its own.

Assume we are integrating percepts from n modalities $\mathcal{M}_i$, $1 \leq i \leq n$ and we know that for each modality $\mathcal{M}_i$, σ_i^2 is the variance of the Gaussian distribution governing the distance between the actual and perceived origin of a stimulus. Given two vectors $v_a = (v_{a1}, v_{a2}, \ldots, v_{an})$ and $v_b = (v_{b1}, v_{b2}, \ldots, v_{bn})$, this could be a suitable distance function:

$$\mathsf{dist}(v_a, v_b) = 1 - \prod_{i=1}^{n} N(|(v_{ai} - v_{bi})|, \sigma_i) \qquad (2)$$

With this function, the distance between two elements would be determined by the likelihood of one being a noisy version of the other under the given noise probability distribution. As a result, for some input element v, the mapping $\mathfrak{i}(v)$ realized by the SOM would yield that element $u \in \mathsf{range}(\mathfrak{i})$ such that $\mathfrak{e}(u)$ is the most likely true origin of the signals encoded in v. Of course this still assumes that we know the senses' reliabilities and that they are the same at all points in space.

In order to see how the latter problem can be solved, suppose for now that our SOM units not only maintain n-dimensional weight vectors m, but are in fact tuples (m, σ) where $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_n)$ is a vector of standard deviations describing the senses' reliabilities. This suffices for now; the actual structure of our SOM units will be explained a little later, in Section II-D. Note, however, that this is already a slight departure from mainstream SOM formulations where the structure of SOM units is usually the same as that of input vectors. Accordingly, the distance function dist must not be defined symmetrically between input vectors v and v' but between input vectors $v = (v_1, v_2, \ldots, v_n)$ and SOM units $u = (m_u, \sigma_u)$. The distance function we will be using in our algorithm is similar to the one from (2), but uses each SOM unit's individual σ_u:

$$\mathsf{dist}(v, u) = 1 - \prod_{i=1}^{n} N(v_i - m_{ui}, \sigma_{ui}) \qquad (3)$$

The question is how to make our SOM units learn the vector σ. Figure 2a shows the effect of simply fitting Gaussian distributions to the distances between m_i's in SOM units (m, σ) and v_i values of the data points $v = (v_1, v_2, \ldots, v_n)$ that were merged into them (blue, red lines in Figure 2a). If we start with equal σ_i for all modalities $\mathcal{M}_i$, then, initially, the distance function dist will weight each v_i equally and the BMU will be somewhere in the middle between them (red dot in Figure 2a). Estimates from the most reliable modality, say $\mathcal{M}_1$, will tend to be closest to their corresponding values in m and thus the corresponding σ_1 will shrink. At one point, the estimated reliabilities will actually be very close to the real ones, and the BMU will be close to the optimal one (blue dot in Figure 2a). However, therefore, data points $v = (v_1, v_2, \ldots, v_n)$ will continue to be merged into SOM units (m, σ) whose m_1 is very close to v_1, leading to an even smaller σ_1 in SOM units. In the end, this cycle will result in our algorithm favoring almost exclusively one modality (the best one, most of the time), and only learning the reliabilities of the others as predictors of that modality's guesses.

Our solution to this problem is deriving σ not from the distribution of data points around the SOM units they are merged into, as above, but from the distances between each modalities' estimates from each other (green lines in Figure 2b). The intuitive reason why we require input from more than two sensory modalities becomes clear at this point: since the distance between two modalities' estimates is symmetric, i.e. $d(v_1, v_2) = d(v_2, v_1)$, calculations based solely on the

distribution of distances between estimates from only two modalities cannot result in different reliabilities for the individual modalities. In the next section, we will show how estimates from three or more modalities, however, can be used to learn their reliabilities.

D. The Learning Rule

Let a SOM unit be a tuple $u = (m, c, \mathcal{V})$, where $m = (m_1, m_2, \ldots, m_n) \in \mathbb{Q}^n$ is the unit's weight vector, $c \in \mathbb{Q}$ is a generalized counter, and $\mathcal{V}$ is a symmetric $n \times n$ matrix whose diagonal elements are all 0. We will use $\mathcal{V}$ to record *the (weighted) average of squared differences* between modalities' guesses:

$$\mathcal{V} = \begin{pmatrix} 0 & \nu_{1,2} & \cdots & \nu_{1,n} \\ \nu_{2,1} & 0 & \cdots & \nu_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ \nu_{n,1} & \nu_{n,2} & \cdots & 0 \end{pmatrix}, \tag{4}$$

where $\nu_{i,j} = \nu_{j,i}, \forall i, j$. This matrix $\mathcal{V}$ will be used later to approximate the σ_i needed for our distance function.

The BMU and neighboring SOM units are updated with an update strength given by a Gaussian neighborhood interaction function h of their distance from the BMU. Let $u, u' \in \mathsf{range(i)}$, then

$$h(u, u') = N(d(u, u'), \sigma_h), \tag{5}$$

where $d(u, u')$ is the Euclidean distance between two SOM units in the SOM's grid and σ_h is called the width of the neighborhood interaction function.

Let $s \in \mathbb{Q}^+$ be the update strength given by h for a BMU u_B and some SOM unit $u = (m, c, \mathcal{V})$. Then u is updated with a data point $v = (v_1, v_2, \ldots, v_n)$ as follows:

$$c' = c + s \tag{6}$$

$$m' = \frac{1}{c'}(cm + sv) \tag{7}$$

$$\nu'_{i,j} = \nu_{i,j} + s\left[(m_i - v_i) - (m_j - v_j)\right]^2 \tag{8}$$

This update rule does three things. First of all, it lets the SOM units learn a mapping between coordinate systems through the fairly standard update of the weight vector in (7). Second, it organizes the SOM units spatially, through the neighborhood interaction in (5) and (7), which is standard, again. Third, and most importantly, in (8), it lets each unit learn the variances of the differences between the modalities' predictions as will be explained in the following.

Assume for now that the modalities' reliabilities are invariant across space and that the coordinate systems of the different modalities coincide. In a data point $v = (v_1, v_2, \ldots, v_n)$, every v_i can then be seen as the sum of a random variable X, which is the true origin of the signal, and another random variable N_i, which models the noise in modality $\mathcal{M}_i$.

After k updates to a SOM unit $(m, c, \mathcal{V})$, $\nu_{i,j}$ is then

$$\begin{aligned} \nu_{i,j} &= \nu_0 + \sum_{l=1}^{k} s_l[(x_l + \rho_{il}) - (x_l + \rho_{jl})]^2 \\ &= \nu_0 + \sum_{l=1}^{k} s_l[\rho_{il} - \rho_{jl}]^2, \end{aligned}$$

for values x_l, ρ_{il}, and ρ_{jl} of X, N_i, and N_j, respectively; for an initial small, non-zero ν_0; and for update strengths s_l, $1 \leq l \leq k$.

Since the noise processes are assumed to be independent, and because the variance of the difference of two independent random variables is the sum of their variances, $\nu_{i,j}$ will, for large k, eventually approach the sum of the variances of the random variables N_i and N_j, scaled by c.

$$\nu_{i,j} \approx c(\sigma_i^2 + \sigma_j^2) \tag{9}$$

Using this, we are now able to approximate the σ_i needed for our distance function defined in (3) from $\mathcal{V}$. For any sequence of an odd length $p \geq 3$ of integers $s = i_1, i_2, \ldots, i_p$ between 1 and the number n of sensory modalities where $i_j \neq i_{j'}, \forall j, j'$, let

$$\begin{aligned} \nu^s = \quad & \nu_{i_1,i_2} - \nu_{i_2,i_3} \\ & + \nu_{i_3,i_4} - \nu_{i_4,i_5} \\ & \quad \cdots \\ & + \nu_{i_{p-2},i_{p-1}} - \nu_{i_{p-1},i_p} \\ & + \nu_{i_{p,1}}. \end{aligned}$$

Using (9), we then get

$$\begin{aligned} \frac{1}{c}\nu^s \approx \quad & (\sigma_{i_1}^2 + \sigma_{i_2}^2) - (\sigma_{i_2}^2 + \sigma_{i_3}^2) \\ & + (\sigma_{i_3}^2 + \sigma_{i_4}^2) - (\sigma_{i_4}^2 + \sigma_{i_5}^2) \\ & \quad \cdots \\ & + (\sigma_{i_{p-2}}^2 + \sigma_{i_{p-1}}^2) - (\sigma_{i_{p-1}}^2 + \sigma_{i_p}^2) \\ & + (\sigma_{i_p}^2 + \sigma_{i_1}^2) \\ = \quad & 2\sigma_{i_1}^2. \end{aligned} \tag{10}$$

For any $i, 1 \leq i \leq n$, an approximation for σ_i can be computed by choosing such a sequence s starting with i and combining the elements of $\mathcal{V}$ as above.

We now have the definition of the distance function between SOM units and data points, we have a rule for updating SOM units, and we showed how SOM units can estimate the strength of the noise in each modality, which is needed by the distance function. This completes the description of our algorithm.

Lifting now the restriction that the modalities' reliabilities be equal at all points in space, it becomes clear why each SOM unit must maintain its own $\mathcal{V}$ and why we need the weighting factor s: suppose, a SOM is updated with data points from a region in space where some modality $\mathcal{M}_i$ is particularly unreliable. Then this will have a greater effect on the reliability attributed to $\mathcal{M}_i$ by SOM units updated strongly with these data points than those updated weakly. Since the

update strength depends on the distance from the BMU, SOM units responsible for different points in space, where $\mathcal{M}_i$ may actually be very reliable, will hardly be affected. The SOM thus learns the modalities' reliabilities at different points in space.

Next, we drop the requirement that modalities' coordinate systems coincide. Since the origin of their percepts is still the same, one can say that there are functions $t_1, t_2, \ldots, t_n$ such that, for a data point $v = (v_1, v_2, \ldots, v_n)$, $v_i = t_i(x) + \rho_i$, where x is the real origin of the signal and ρ_i is the output of some noise process. For each SOM unit $u = (m, c, \mathcal{V})$, m can then be understood as approaching $(t_1(z), t_2(z), \ldots, t_n(z))$ for some coordinate z in real space. Each entry $\nu_{i,j}$ of $\mathcal{V}$ is

$$\nu_{i,j} = \sum_{l=1}^{k} s_l[(m_{li} - t_i(x_l) - \rho_{il}) - (m_{lj} - t_j(x_l) - \rho_{jl})]^2,$$

where k is the number of updates to the SOM unit, m_{il} is the i^{th} entry of m before the l^{th} update, and x_l, ρ_{il}, ρ_{jl} are the l^{th} values of the random variables X, N_i, and N_j, respectively.

If the coordinate systems are merely shifted wrt. each other, then, after coordinate transformation has been learned and the SOM is sufficiently organized, it is true that

$$m_{li} - t_i(x_l) \approx m_{lj} - t_j(x_l), \tag{11}$$

and therefore

$$\nu_{i,j} \approx \sum_{l=1}^{k} s_l(\rho_{jl} - \rho_{il})^2 \approx c(\sigma_i^2 + \sigma_j^2), \tag{12}$$

which means that our earlier considerations based on (9) hold again.

If the sensory coordinate systems are scaled wrt. each other by a moderate factor, then (11) will still be true for the BMU and the units around it, which are updated most strongly. However, if the scale of the modalities' coordinate systems differs greatly, and especially if the scaling factor between two modalities is negative, then (11) and thus (12) will hold approximately only for the BMU and a very small neighborhood. In order to maintain the SOM's organization, however, a certain minimum size of the update neighborhood is needed. Thus, our algorithm works if the sensory coordinate systems are shifted wrt. each other, or scaled moderately, i.e. by scaling factors around 1.

E. Multi-Dimensional Input

As stated, so far we have been considering only one-dimensional localization. The obvious extension to two dimensions, in which each SOM unit contains *two* tuples $(m_x, c_x, \mathcal{V}_x)$ and $(m_y, c_y, \mathcal{V}_y)$, one per dimension, is straightforward and works quite well, except for one effect one might not immediately expect: Although scaling and shifting along either of the axes is supported by the two-dimensional SOM just as well as by the one-dimensional SOM, rotation strongly affects its performance.

This is clear from the following consideration: in a nutshell, our algorithm learns how well sensory modalities' guesses

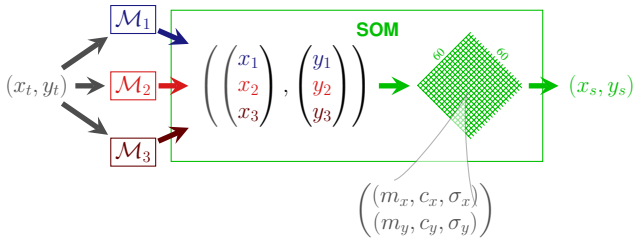


Fig. 3: Data Flow in the Simulation

predict each other. If extended like above, it can learn how well x-coordinates of one sensory modality predict x-coordinates of another. However, if the coordinate system of one modality, $\mathcal{M}_i$, is rotated with respect to a different modality, $\mathcal{M}_j$, then all the algorithm can learn is that x_i coordinates predict x_j coordinates very badly, which is because those latter x_j coordinates are actually a linear combination of the x_i and y_i coordinates, and x_i simply does not predict y_i.

III. Simulation

In order to validate the theoretical considerations about our algorithm, we implemented it and tested it in simulations. Sections III-A and III-B describe the details of learning and sampling in our reference simulation: integration of cues from three modalities with identical coordinate systems and different reliabilities, which are constant in space. Section III-C reports on the results of that simulation. Section III-D briefly summarizes three more simulations, which differ from the first one in that the sensory coordinate systems are shifted, they are scaled, and sensory reliability is variant in space.

A. Learning

We trained a SOM consisting of 60×60 2-dimensional units as described in the previous section (see Figure 3). The data points were created from a sequence of 100,000 random 2-D vectors with coordinates between 0 and 1 by adding independent noise of three different intensities to form a (2×3)-dimensional vector.

Each data point thus represented the combined estimates of the location of some stimulus from three different sensory modalities, each with its respective reliability. We will refer to these simulated modalities as $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_3$ from now on. The noise imposed on the vectors was Gaussian distributed (see (1)) with standard distributions of $\sigma_1 = 0.1$, $\sigma_2 = 0.2$, and $\sigma_3 = 0.3$, for $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_3$, respectively.

The neighborhood radius decreased linearly over the first 11,000 update steps from 90 units, spanning the full diagonal of the SOM in the beginning, down to 15 units, where it remained constant. The width of the neighborhood interaction function h (see (5)) was always a fifth of the neighborhood radius. These values were found empirically to lead to fast, smooth, and stable topological organization.

B. Sampling

After learning was finished, we sampled the SOM with 10 000 fresh data points which, again, were randomly generated. These data points were of the same quality as those with

which the SOM was trained, except that the coordinates of the true origin only spanned the interval $(0.33, 0.66)$ which was to prevent border effects: to see why this was necessary, suppose we had admitted true origins like $v = (1, 1)$. Noise could have led to the coordinates of a data point generated from such a v being outside the SOM's grid. However, in finding the BMU for this data point, the SOM would have chosen a unit whose coordinates are within the grid, and thus closer to v. Although it could be argued that the SOM would have learned the span of likely input origins, and thus this behavior is justified, it would make it difficult to evaluate the SOM's performance and compare it against the maximum likelihood estimator as below.

C. Results

Figure 4a shows a visualization of the learning results. The top row shows the mapping of the modality $\mathcal{M}_i$'s coordinate system into the SOM's grid for $i \in \{1, 2, 3\}$. Each pixel represents a SOM unit $((m_x, c_x, \sigma_x), (m_y, c_y, \sigma_y))$. Pixels' redness values correspond to the m_{xi} coordinate, their blueness values to the m_{yi} coordinate. The bottom row shows the reliability assigned by each SOM unit to the modalities. Black and white represent noise intensities of 0 and 1, respectively.

The smooth transitions from black to red, black to blue, and black to magenta show the smooth mapping of input coordinates across the SOM. The fact that the three squares in the top row of Figure 4a look almost identical indicates that the SOM learned to associate equal coordinates in the modalities' coordinate systems with each other (as opposed to the scaled and shifted cases, see Section III-D). This part of learning—coordinate transformation and spatial organization—was completed after the first few hundred learning steps in which the update radius was large.

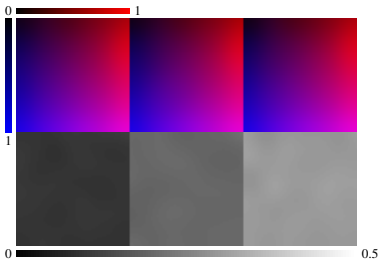
The noise intensity in the three modalities was learned to be constant across space, as can be seen in the bottom row of Figure 4a. The different shades of gray in the three squares stand for intensities of 0.1, 0.2, 0.3, respectively.

Figure 5 shows the progress of reliability learning throughout the simulation. As indicated, the graphs represent the mean over all SOM units of each modality's noise intensity as learned at a given step throughout the simulation.

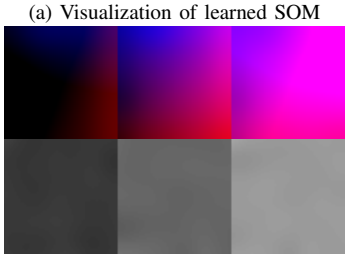
In order to evaluate the learned SOM's performance, we calculated the distance of each true input vector v_k to the SOM's guesses given the noisy input. Thus, if $v_k = (x, y)$ and the BMU was $((m_x, c_x, \mathcal{V}_x), (m_y, c_y, \mathcal{V}_y))$, we calculated $d_{ki} = |v_k - (m_{xi}, m_{yi})|$ for each modality $\mathcal{M}_i$. We obtained the reliability r_{SOM_i} of the SOM as an estimator of true $\mathcal{M}_i$ coordinates from

$$r_{SOM_i} = \frac{1}{\frac{1}{N}\sum_{k=1}^{N} d_{ki}^2},$$

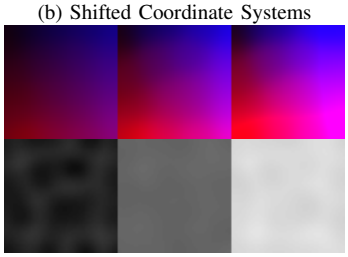
where $N = 10\,000$ was the size of the randomly generated test data set. Figure 6 shows the noise intensity $\sigma_{SOM_1} = \sqrt{\frac{1}{r_{SOM_1}}}$ in this estimator; the values for σ_{SOM_2} and σ_{SOM_3} were very similar (0.0855, 0.0852, respectively).



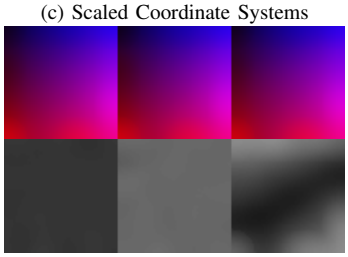
(a) Visualization of learned SOM



(b) Shifted Coordinate Systems



(c) Scaled Coordinate Systems



(d) Space-Variant Reliabilities

Fig. 4: Visualizations for three two-dimensional modalities. Top row: redness, blueness for x, y coordinates. Bottom row: black for high ($\sigma = 0$), white for low ($\sigma = 0.5$) reliability.

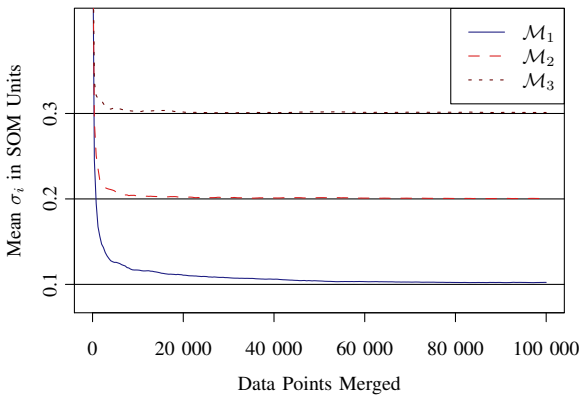


Fig. 5: The SOM's estimates of modalities' reliabilities during training: mean estimates over all SOM units

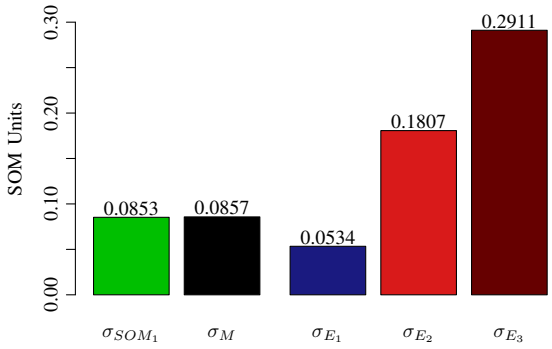


Fig. 6: Sampling Results:

σ_{SOM_1}: noise intensity in the SOM's estimate of true $\mathcal{M}_1$ coordinates,
σ_M: noise intensity expected from MLE,
σ_{E_i}: deviation of SOM's estimate of $\mathcal{M}_i$ coordinates from noisy $\mathcal{M}_i$ input, $i \in \{1, 2, 3\}$.

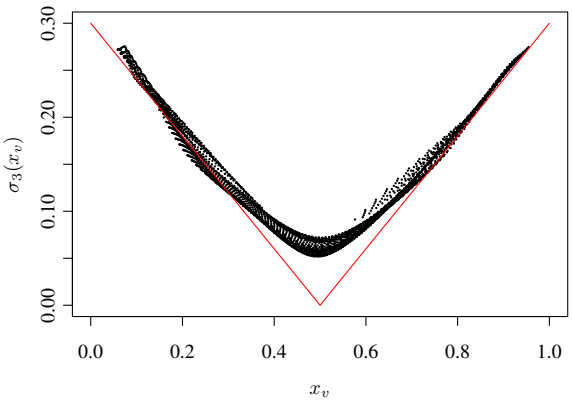


Fig. 7: Learned, Space-Variant Noise
Red line: actual noise intensity.
Dots: learned noise intensity.

The second bar in Figure 6 shows the noise intensity σ_M to be expected in a maximum likelihood estimator (MLE) given the noise intensities $\sigma_1 = 0.1$, $\sigma_2 = 0.2$, and $\sigma_3 = 0.3$ in our simulation: An MLE optimally combines uni-sensory estimates $x_1, x_2, \ldots, x_n$ from n sensory modalities with known reliabilities $r_i = \frac{1}{\sigma_i^2}, 0 \leq i \leq n$ into one estimate x_M:

$$x_M = \frac{1}{\sum_{i=1}^{n} \frac{1}{\sigma_i^2}} \sum_{i=1}^{n} \frac{1}{\sigma_i^2} x_i.$$

The variance σ_M^2 of the MLE, in turn is [21]:

$$\sigma_M^2 = \frac{1}{\sum_{i=1}^{n} \frac{1}{\sigma_i^2}}.$$

Bars three to five in Figure 6 display the difference between the noisy input coordinates from modalities $\mathcal{M}_i, i \in \{1, 2, 3\}$ and the SOM's estimates of the true coordinates in that modality's coordinate system. Similarly to above, we calculated $d'_{ki} = |v'_{ki} - (m_{xi}, m_{yi})|$, for each $\mathcal{M}_i$, where this time v'_{ki} was the noisy input from $\mathcal{M}_i$ received by the network. We plotted

$$\sigma_{E_i} = \sqrt{\frac{1}{N} \sum_{k=1}^{N} d'^2_{ki}},$$

which can be interpreted as the noise intensity in $\mathcal{M}_i$ perceived by the naïve version of our SOM, as discussed in Section II-C, once it approximates the true noise intensities.

D. Shifted & Scaled Coordinate Systems, and Space-Variant Reliabilities

Figures 4b, 4c, and 4d show visualizations analogous to Figure 4a for simulations in which the modalities' coordinate systems were shifted and scaled against each other, and one modality's reliability varied across space, respectively.

Without going too much into detail, Figure 4b shows a case identical to our reference case described in Section III-B, except that coordinates for $\mathcal{M}_1$ and $\mathcal{M}_3$ were shifted by constant vectors $(-0.5, -0.5)$ and $(0.5, 0.5)$, respectively, before the data points were fed to the SOM for learning and sampling. In Figure 4c, the coordinates for modalities $\mathcal{M}_1$ and $\mathcal{M}_3$ were scaled by factors of 0.5 and 1.5.

The difference in coordinate systems is learned by the SOM in each case, as can be seen from the different color intensities. The visualizations for learned reliabilities in Figure 4b are almost the same as in Figure 4a which shows that learning of reliabilities is not affected by the shift in coordinate systems.

In Figure 4c, the visualization for the learned reliability of $\mathcal{M}_1$ is a bit darker than that in the other images, and the one for the reliability of $\mathcal{M}_3$ is brighter. This is because the noise components in the input vectors were scaled along with the original values. In either case, the sampling performance as described above was unaffected.

Figure 4d shows the result of a learning process in which the noise in $\mathcal{M}_3$ depended on the x coordinate of the original signal v: Let $v = (x_v, y_v)$ be such an original signal. Then, the noise intensity was $\sigma_3(x_v) = 0$ for $x_v = 0.5$ and increased linearly to the sides up to $\sigma_3(x_v) = 0.3$ for $x_v = 0$ and $x_v = 1$. This noise distribution was also learned, which shows in Figure 4d by the correspondence of redness values in the upper third panel to brightness in the lower third panel.

Figure 7 shows this connection more clearly: the red line shows the actual strength of the noise as it varies with $\mathcal{M}_3$'s x coordinate, and the dots show the σ_3 derived from SOM units' $\mathcal{V}$s plotted by their x coordinates (see (10)).

IV. Discussion

The model presented in this paper shows how Kohonen's SOM can be extended to not only learn coordinate transformation between sensory maps, but also reliability of the sensory modalities whose input is to be integrated. It amounts to a model of multi-sensory integration and learning thereof in the SC at a comparatively high ANN level. Our simulations indicate that it is indeed able to learn sensory modalities' reliabilities not only globally, but as they vary across space, and that, learning being finished, it can perform near-optimally.

One aspect of our model is that it requires at least three different modalities for learning sensory reliabilities. This may

be surprising, at first, considering that e.g. humans' main sensory modalities are just vision and hearing. If learning would indeed be limited to events with visual, auditory, and, say, proprioceptive input, then that would mean it would probably have to make do with very scarce data. Introducing synthetic modalities, i.e. additional input from prediction processes or prior knowledge, on top of physical modalities like vision and hearing could be one remedy.

Previous work on multi-sensory integration in the SC with different foci provides leads for future work. As pointed out, a number of (low-level) ANN models have been devised which deal with the role of input to the SC descending from cortical regions, in particular on the phenomena of enhancement, depression, and inverse effectiveness [15]–[18]. Examining possible connections to our distance function could lead to a very interesting interpretation of these effects.

Another direction in a similar vein is designing a version of our algorithm which operates at a less abstract level, closer to biological plausibility. The model due to Ghahramani [21], the present model, and such a biologically plausible ANN model could be seen as three related approaches describing the same phenomenon observed in biological computation at the mathematical, the algorithmic, and the neuronal levels.

Researchers have argued that the brain may and should take into account not only the general reliability of sensory modalities, but also situational cues [28]. It would make sense e.g. to weight visual information about the location of a far-away object much less strongly in a dark or foggy environment than in a clear and well-lit one. This observation gives rise to yet one more improvement of our model: so far, we are not considering the amount of uncertainty of a stimulus. It is easy, within the Bayesian framework, to use that uncertainty, if it is available. Extending the present model such that it does would further improve its explanatory power and, indeed, its usefulness in real-world applications.

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

[1] B. E. Stein and M. A. Meredith, *The Merging Of The Senses*, 1st ed., ser. Cognitive Neuroscience Series. MIT Press, Jan. 1993.

[2] Y. E. Cohen and E. I. Knudsen, "Maps versus clusters: different representations of auditory space in the midbrain and forebrain," *Trends in Neurosciences*, vol. 22, no. 3, pp. 128–135, Mar. 1999.

[3] R. M. Stern, G. J. Brown, and D. Wang, "Binaural sound localization," in *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, D. Wang and G. J. Brown, Eds. Wiley-IEEE Press, 2006, ch. 5, pp. 147–185.

[4] M. S. Landy, M. S. Banks, and D. C. Knill, "Ideal-observer models of cue integration," in *Sensory Cue Integration*, J. Trommershäuser, K. Körding, and M. S. Landy, Eds. Oxford: Oxford University Press, Aug. 2011, ch. 1, pp. 5–29.

[5] C. Weber and J. Triesch, "Implementations and implications of foveated vision," *Recent Patents on Computer Science*, vol. 2, no. 1, pp. 75–85, Jan. 2009.

[6] E. I. Knudsen and M. S. Brainard, "Creating a unified representation of visual and auditory space in the brain," *Annual Review of Neuroscience*, vol. 18, no. 1, pp. 19–43, 1995.

[7] C. A. Curcio, K. R. Sloan, R. E. Kalina, and A. E. Hendrickson, "Human photoreceptor topography." *The Journal of Comparative Neurology*, vol. 292, no. 4, pp. 497–523, Feb. 1990.

[8] J. C. Middlebrooks and D. M. Green, "Sound localization by human listeners," *Annual Review of Psychology*, vol. 42, no. 1, pp. 135–159, 1991.

[9] A. J. King, "Development of multisensory spatial integration," in *Crossmodal Space and Crossmodal Attention*. Oxford University Press, USA, May 2004.

[10] M. Gori, M. Del Viva, G. Sandini, and D. C. Burr, "Young children do not integrate visual and haptic form information," *Current Biology*, vol. 18, no. 9, pp. 694–698, May 2008.

[11] P. W. Battaglia, R. A. Jacobs, and R. N. Aslin, "Bayesian integration of visual and auditory signals for spatial localization," *Journal of the Optical Society of America A*, vol. 20, no. 7, pp. 1391–1397, Jul. 2003.

[12] M. O. Ernst and M. S. Banks, "Humans integrate visual and haptic information in a statistically optimal fashion." *Nature*, vol. 415, no. 6870, pp. 429–433, Jan. 2002.

[13] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977.

[14] J. A. Bilmes, "A gentle tutorial of the EM algorithm and its application to parameter estimation for gaussian mixture and hidden markov models," International Computer Science Institute, Tech. Rep., Apr. 1998.

[15] B. A. Rowland, T. R. Stanford, and B. E. Stein, "A model of the neural mechanisms underlying multisensory integration in the superior colliculus," *Perception*, vol. 36, no. 10, pp. 1431–1443, 2007.

[16] C. Cuppini, B. E. Stein, B. A. Rowland, E. Magosso, and M. Ursino, "A computational study of multisensory maturation in the superior colliculus (sc)," *Experimental Brain Research*, vol. 213, no. 2, pp. 341–349, Sep. 2011.

[17] T. J. Anastasio and P. E. Patton, "A two-stage unsupervised learning algorithm reproduces multisensory enhancement in a neural network model of the corticotectal system," *The Journal of Neuroscience*, vol. 23, no. 17, pp. 6713–6727, Jul. 2003.

[18] A. Pavlou and M. Casey, "Simulating the effects of cortical feedback in the superior colliculus with topographic maps," in *The 2010 International Joint Conference on Neural Networks (IJCNN)*. IEEE, Jul. 2010, pp. 1–8.

[19] B. A. Rowland and B. E. Stein, "Computational models of multisensory integration in the cat superior colliculus," in *Sensory Cue Integration*, J. Trommershäuser, K. Körding, and M. S. Landy, Eds. Oxford: Oxford University Press, Aug. 2011, ch. 18, pp. 333–344.

[20] T. H. Weisswange, C. A. Rothkopf, T. Rodemann, and J. Triesch, "Bayesian cue integration as a developmental outcome of reward mediated learning," *PLoS ONE*, vol. 6, no. 7, pp. e21 575+, Jul. 2011.

[21] Z. Ghahramani, "Computation and psychophysics of sensorimotor integration," Ph.D. dissertation, Massachusetts Institute of Technology, Sep. 1995.

[22] T. Kohonen, "Self-organized formation of topologically correct feature maps," *Biological Cybernetics*, vol. 43, no. 1, pp. 59–69, Jan. 1982.

[23] ——, *Self-Organizing Maps*, 3rd ed., ser. Springer series in information sciences, 30. Springer, Dec. 2000.

[24] O. J. Vrieze, "Kohonen network artificial neural networks," in *Artificial Neural Networks*, ser. Lecture Notes in Computer Science, P. Braspenning, F. Thuijsman, and A. Weijters, Eds. Berlin/Heidelberg: Springer, 1995, vol. 931, ch. 5, pp. 83–100.

[25] S. Klanke, "Learning manifolds with the parametrized self-organizing map and unsupervised kernel regression," Ph.D. dissertation, University of Bielefeld, Mar. 2007.

[26] J. Göppert and W. Rosenstiel, "Varying cooperation in SOM for improved function approximation," in *IEEE International Conference on Neural Networks*, vol. 1. IEEE, Jun. 1996, pp. 1–6 vol.1.

[27] A. Zaidel, A. H. Turner, and D. E. Angelaki, "Multisensory calibration is independent of cue reliability," *The Journal of Neuroscience*, vol. 31, no. 39, pp. 13 949–13 962, Sep. 2011.

[28] R. A. Jacobs, "What determines visual cue reliability?" *Trends in Cognitive Sciences*, vol. 6, no. 8, pp. 345–350, Aug. 2002.