Noname manuscript No.
(will be inserted by the editor)

A generative learning approach to sensor fusion and change detection

Alexander R.T. Gepperth · Thomas Hecht · Mandar Gogate

the date of receipt and acceptance should be inserted later

Abstract We present a system for performing multi-sensor fusion that learns from experience, i.e., from training data and propose that learning methods are the most appropriate approaches to real-world fusion problems, since they are largely model-free and therefore suited for a variety of tasks, even where the underlying processes are not known with sufficient precision, or are too complex to treat analytically. In order to back our claim, we investigate two simulated fusion problems which are representative of real-world problems and which exhibit a variety of underlying probabilistic models and noise distributions. To perform a fair comparison, we study two other ways of performing optimal fusion for these problems: empirical estimation of joint probability distributions and direct analytical calculation using Bayesian inference. We demonstrate that near-optimal fusion can indeed be learned, and that learning is by far the most generic and resource-efficient alternative. In addition, we show that the generative learning approach we use is capable of improving its performance far beyond the Bayesian optimum by detecting and rejecting outliers, and that it is capable to detect systematic changes in the input statistics.

1 INTRODUCTION

This study is situated in the context of biologically motivated sensor fusion (often also denoted multi-sensory or multi-modal integration). This function is a necessity for any biological organism, and it seems that, under certain conditions, humans and other animals can perform statistically optimal multi-sensory fusion[1]. As to how this is achieved, many questions remain: mainly, one can speculate whether there is a generic,

A.Gepperth, Thomas Hecht
ENSTA ParisTech 828 Boulevard des Marechaux
91762 Palaiseau, France
E-mail: alexander.gepperth@ensta.fr

Mandar Gogate
BITS Pilani - K K Birla Goa Campus
NH 17B, Zuarinagar, Goa India - 403726
E-mail: mandarvinyakgogate@gmail.com

sensor-independent fusion mechanism, operating on probability distributions and taking into account the basic statistical laws such as Bayes rule at some neural level, or whether optimal fusion, where it occurs, is something that is fully learned through experience.

In this article, we investigate the matter by comparing several feasible possibilities for implementing multi-sensor fusion in embodied agents, namely generative learning, inference based on estimated joint probabilities, and model-based Bayesian inference. The first method is purely learning-driven, knowing nothing of Bayes' law and using adaptive methods both for representing data statistics and inference, whereas the second estimates joint probability statistics from data but uses Bayes' law for inference. The third method is not adaptive at all but uses models (which have to be known beforehand) of the data generation process, based on which it performs Bayesian inference. To perform this comparison in a meaningful way, allowing each approach to play its strengths, we consider two very different simulated fusion tasks: on the one hand the "standard model" of multi-sensor fusion, where two (or more) physical sensor readings are modeled by a single underlying "true" value that is corrupted by Gaussian noise, and on the other hand a more realistic process modeling, e.g., the estimation of depth measurements in which one sensor reading depends non-linearly upon the "true" value and the other reading(s). Testing all three approaches on the same fusion tasks allows a meaningful quantitative comparison, giving indisputable results.

1.1 Overview of biological literature

The multisensory processes going on in mammalian brains are implied in maximizing information gathering and reliability by the effective use of a set of available observations from different sensors [2]. Multi-sensory fusion aims at providing a robust and unified representation of the environment through multiple modalities or features [3]. This sensory synergy provides speed in physiological reactions [4], accuracy in detection tasks, adaptability in decision making and robustness in common abstract representations. It also allows to distinguish relevant stimuli from each other, which permits quick behavioural and cognitive responses. Hence, it helps avoiding saturation and infobesity traps in tasks like motion orientation towards auditory signal source or focusing visual attention for object recognition.

Multi-sensory fusion has been widely studied at different levels (i.e. from particular cortical cells to individual psycho-physiological behaviour) and into different scientific fields (e.g. neurophysiology, neuroimagery, psychophysiology or neurobiology). Several works in neuroscience have already described multi-sensory fusion in individual neurons [5] or in various cortical areas [6]. Since the 1960's it has been demonstrated that multi-sensory fusion is carried out hierarchically, layer after layer, within cortical areas containing cells which respond to signals from multiple modalities[7,8]. Regarding the superior colliculus (SC), a mid-brain structure that controls changes in orientation and attention focusing towards points of interest [9], *Multisensory enhancement* (MSE) refers to a situation in which a *cross-modal stimulus* (i.e. from two or more sensory modalities) provokes a response greater than the response to the most effective of its component stimuli [5]: this effect increases as the cues strength decreases (*inverse effectiveness rule*). *Multisensory depression* (MSD) hints at the opposite phenomenon[5]. In psychology, following the idea of a certain harmony across senses [10], several well-known experiments have underlined multi-sensory fusion causing so-called

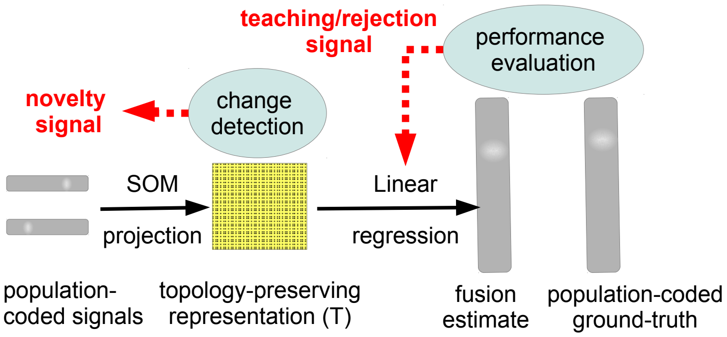

Fig. 1 Schematic overview of the learning system for multi-sensory fusion proposed in this study.

"multi-sensory illusions" like the ventriloquism effect [11], the McGurk effect [12], the rubber-hand illusion [13] or the sound-induced flash illusion [14].

It stands to reason that multi-sensory fusion in biological systems is not generally innate but learned [15,16]. Thus not only the question of how multi-sensory fusion is carried out is of importance, but also of how it is acquired and how it can be adapted if the statistics of the environment change.

1.2 Goals of the study

This study aims at providing a purely computational perspective on which multi-sensory fusion strategy might be most appropriate to implement in future artificial cognitive agents. Criteria we use are precision, generality, outlier detection capacity and resource efficiency, the last point being important because resources are usually severely limited in artificial cognitive agents. Lastly, we wish to show that learning approaches offer ways to improve fusion performance beyond the limit imposed by Bayes' law: this seeming paradox can be resolved when accepting to process only those samples which are particularly suited for fusion, based on simple criteria made available by learning approaches. In this "super-Bayesian fusion" setting one may trade reaction speed for precision, where it depends on the frequency of incoming samples which compromises can be made.

1.3 Approach

1.4 Related work

Multi-sensory fusion can be modeled in various ways, the most relevant ones being self-organized learning and Bayesian inference. The former regroups a set of bio-inspired, unsupervised learning algorithms while the latter argues that mammals can combine sensory cues in a statistically optimal manner.

1.4.1 Self-organizing topology-oriented algorithms and multimodal fusion

Among artificial neural networks, self-organizing maps (SOMs) perform clustering while preserving topological properties of the input space. Studies applying SOMs to multimodal fusion are usually focused on reproducing properties one can find in biology: continuous unsupervised processes, adaptation and plasticity, topological preservation of the input space relationship, or dimensionality reduction. Self-organized approaches have the potential to establish a transition from high-dimensional, noisy and modality-specific sensor readings to abstract, multimodal and symbolic concepts, whereas they are considered less appropriate for reproducing statistical optimality which should be respected by any fusion process.

Most SOM-based approaches imitate, more or less closely, the hierarchical and layered structure of cortical areas, especially the superior colliculus (SC). In [17] the authors use one basic non-layered self-organizing map to simulate biological MSE and inverse effectiveness from artificial combinations of sensory input cues with Gaussian neighborhood and Manhattan distance. Even though their data are low-dimensional and non-realistic, the authors confirm that the (nearly) original SOM algorithm can lead to meaningful multisensory cue combination. As in [18], they emphasize the positive impact of a non-linear transfer function applied to map outputs (in these cases, a sigmoid function). [19] design a SOM-like feed-forward layered architecture which uses Hebbian learning to associate maps. The study shows how uni-sensory maps can be aligned by a system that learns coordinate transformations, and automatically integrated into a multisensory representation, i.e. a third map. It focuses on comparing simulated and biological responses to spatially coincident or non-coincident stimuli and achieves the reproduction of simplistic MSE and MSD effects. [20] models the learning of words and objects relations by young children from a psychological point of view using two unimodal SOMs. Hebbian learning is used to model the mapping between unimodal spaces so that the presentation of an object activates the correct corresponding word and vice versa. [21] also deals with imitating SC multi-sensory fusion. The study designs a SOM-based model which is task-oriented and aims at autonomously finding a way to reach a specific goal (in this case object localization). The model tries to characterize the reliability of sensory modalities when dealing with noisy artificial stimuli. Using a single SOM, it uses a custom distance function that measures the likelihood of one input vector to be a noisy version of a known prototype. To this effect, the original Kohonen algorithm is adapted so that the metric takes into account the estimation of each sensor modality's reliability. Nevertheless, noise for different modalities is assumed to independent and normally distributed, and the analogy with the superior colliculus is rather superficial.

Finally, several studies aim at designing models inspired by recent neurophysiological findings without fully copying biological architectures or processes, focusing on well-defined applications. Following [22], [23] proposes a learning algorithm based on a variant of the original SOM, ViSOM [24], which forces the inter-unit distances in a map to be proportional to those in the input space. As for multi-expert techniques in supervised learning [23], a single final map is learned that takes into account multiple SOMs. Each unit of the final map corresponds to a fusion by Euclidean distance, and a voting process is performed on neurons at the same position of the grid in each unimodal SOM. Although the model does not really focus on bio-plausibility and prefers enhancing topology preservation and data visualization properties, it is remarkable that it allows autonomous artificial concentration on interesting features in each of the unimodal SOMs. [25] is one of the most ambitious works using self-organized ANNs for multi-sensory fusion without only reproducing known biological phenomena. With a hierarchical lattice of SOMs (two unimodal maps integrated in a multimodal map), they confirm abilities of SOMs and feedforward connections in integrating unimodal artificial percepts into multimodal unified representations, as [26] already did. This study above all puts forward the effective role of feedback connections in artificial attentional mechanisms: without feedback, hierchical SOM networks might only achieve multi-sensory fusion for elementary artificial stimuli. They apply their (strongly fine-tuned) model to produce bimodal fusion of phonemes and letters but do not provide a clear task-oriented (e.g. speech recognition) evaluation.

1.4.2 Bayesian inference as a model of multi-sensory fusion

Several psychophysiological studies have shown that mammalian brains, and in particular humans ones, fuse multiple sensory signals in an statistically optimal way by a weighted linear combination of estimates of the individual measurements [2]. The intrinsic uncertainty of sensory cues make them more or less reliable and this has to be taken into account by the fusion process [27]. Most of the time, these observations are conducted using animals at a behavioural level. However, probabilistic inference in neural circuits is still not well understood. Among fusion methods relying on probability distributions or density functions to express data uncertainty, Bayesian fusion is one of the best-known techniques and is known for having strong links to biological multi-sensory fusion [28]. It consists of weighting each sense according to its known variance by applying *maximum likelihood estimator* (MLE) or *maximum a posteriori* (MAP) techniques. A couple of assumptions are usually made when performing optimal fusion according to Bayesian statistics [29–35], namely the assumption of Gaussian noise that is applied independently to each sense, and the assumption of known variances. These assumptions, while acceptable on a theoretical level, prevent the use of such techniques in domains such as developmental learning since evidently neither the variances nor the theoretical distribution and independence properties of signals should be known in advance.

2 METHODS

In this section, we will give all necessary details of the used learning approach (Sec.2.4), Bayesian inference based on estimated joint probabilities (Sec. 2.2) and model-based Bayesian inference (Sec. 2.3). In all cases, the setting is identical: a single "true" value r and two noisy sensor readings s_1 and s_2. The way of obtaining s_1 and s_2 from r depends on the particular problem that is treated, as do the noise distributions that additionally perturb the sensor readings.

2.1 Fusion problems

Two fusion problems are considered in this study, in both of which the goal is to infer the true value r from noisy sensor readings s_1, s_2. Several types of noise $\tilde{\epsilon}(\sigma)$ are considered for corrupting sensor readings (Gaussian, uniform, triangular), all of which have a single parameter σ that models their "strength". In the case of Gaussian

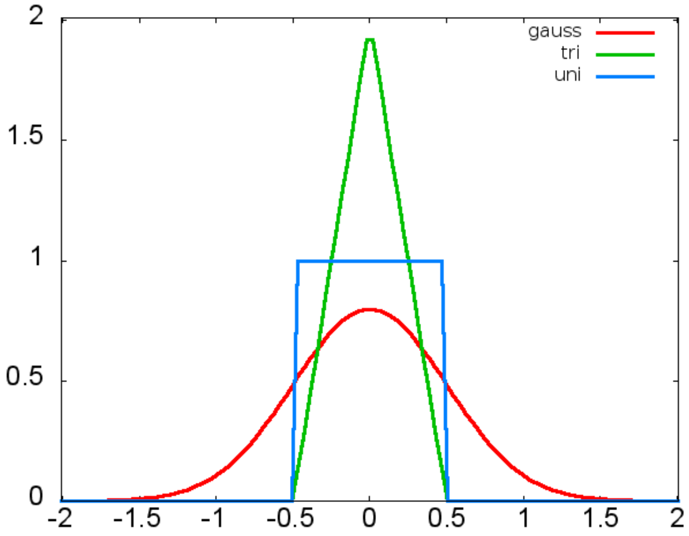

Fig. 2 Probability density functions for the three parametrized noise types used in this article: Gaussian, uniform and triangular. For all curves, a parameter value of $\sigma = 0.5$ has been used.

noise, σ would correspond to the standard deviation, for uniform and triangle noise it is the half-width of the interval with nonzero probability. For every noise type in both problems, we vary the parameter σ in order to determine how this impacts fusion accuracy. The probability density functions for all three (additive) noise types are as follows (see also Fig.2):

$$p^{\text{gauss}}(x,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp -\frac{x^2}{2\sigma^2} \tag{1}$$

$$p^{\text{uni}}(x,\sigma) = \begin{cases} \frac{1}{2\sigma} & \text{if } x \in [-\sigma,\sigma] \\ 0 & \text{else} \end{cases}$$

$$p^{\text{tri}}(x,\sigma) = \frac{1}{\sigma}\left(1 - \frac{1}{\sigma}|x|\right) \tag{2}$$

Problem I This first "family" of problems follows the "classic" Bayesian framework: a single "true" value r that gives rise to several noisy sensor readings s_i. We suppose that the sensor readings s_i are from an unique r value by adding independent, parametrized noise $\tilde{\epsilon}(\sigma)$.

$$r \sim p^{\text{uni}}_{a,b}(x)$$
$$s_i = r + \tilde{\epsilon}(\sigma)$$

Problem II A more realistic setting is where the sensory readings s_i and the underlying true value r are more tightly coupled. In this second problem family, we suppose that the s_i are no longer class-conditionally independent and depend on r as well as each other. In other words, *all* of the s_i have to be considered simultaneously for inferring r. Formally, we express this by drawing r independently from a bounded uniform distribution, and then making $\tilde{s}_2$ a deterministic function of r and $\tilde{s}_1$: $\tilde{s}_2 = f(\tilde{s}_1, r)$. Afterwards, the $\tilde{s}_i$ are subjected to additive Gaussian noise in order to produce the real sensor values s_i:

$$\begin{aligned} r &\sim p^{\text{uni}}_{a,b}(x) \\ \tilde{s}_1 &\sim p^{\text{uni}}_{a,b}(x) \\ \tilde{s}_2 &= f(\tilde{s}_1, r) \\ s_i &= \tilde{s}_i + \tilde{\epsilon}(\sigma) \end{aligned}$$

2.2 Bayesian inference based on estimated joint probabilities

The basic idea of this approach is to empirically estimate the conditional probability distribution $p(r|s_1 s_2)$ during a training phase where r is available with each sample, and to perform Bayesian inference using this conditional probability distribution in the subsequent evaluation phase where r is not available. First of all, it is straightforward to see that the maximum of $p(r|s_1 s_2)$ w.r.t. r is equivalent to the maximum of the joint probability $p(rs_1 s_2)$, so it is sufficient to estimate this quantity. In order to estimate joint probabilities in practice, all variables must be discretized to n bins using an invertible function $b_{\mu,n}(x) \rightarrow i \in \mathbb{N}$, where obviously a finer discretization implies higher precision but also higher memory and execution time demands. for variables in the $[0, 1]$ interval, we chose b such that it pads the encoded scalar value with borders of width μ, which is necessary because random variables might fall outside the $[0, 1]$ interval depending on noise, and still need to be represented properly:

$$b_{\mu,n}(x) = \text{floor}\left(n - (\mu + (1 - 2\mu)x\right) \tag{3}$$

$$b^{-1}_{\mu,n}(i) = \frac{\frac{i}{n} - \mu}{1 - 2\mu} \tag{4}$$

For three discretized variables, the estimated joint probability matrix $\hat{p}_{ijk}$ has n^3 entries and requires roughly n^3 samples to be filled properly. During training, samples (r, s_1, s_2) are received on by one, and for each sample the matrix is updated as follows:

$$\begin{aligned} p_{ijk}(0) &\equiv 0 \\ p_{b(r)b(s_1)b(s_2)}(t + 1) &= p_{b(r)b(s_1)b(s_2)}(t) + 1 \end{aligned} \tag{5}$$

At the end of the training phase, p_{ijk} is normalized to have a sum of 1. When performing inference during the evaluation phase, only the two sensor readings s_1 and s_2 are available, and the task to infer the underlying value r^* that best matches s_1 and s_2 amounts to finding the matrix bin for r with the highest probability:

$$\begin{aligned} i^* &= \max\nolimits_i \; p_{i\,b(s_1)\,b(s_2)} \\ r^* &= b^{-1}(i^*) \end{aligned} \tag{6}$$

2.3 Model-based Bayesian inference

Similar in spirit to the preceding section, model-based Bayesian inference aims to find the most probable value of r given the observations s_1 and s_2:

$$r^* = \text{arg max}_r p(r|s_1 s_2) \sim \text{arg max}_r p(s_1 s_2|r)p(r) \tag{7}$$

This amounts to a maximization problem:

$$\begin{aligned} \partial_r p(s_1 s_2|r)p(r) = 0 \\ \Leftrightarrow \partial_r \left(p(s_1 s_2|r)\right) p(r) + p(s_1 s_2|r)\partial_s p(r) = 0 \end{aligned} \tag{8}$$

Eqn. (8) has trivial solutions outside the interval $]a, b[$ where both $p(r)$ and $\partial_s p(r)$ vanish. However they *minimize* $p(s_1 s_2|r)p(r)$ (inserting an appropriate r always gives a value of 0), and are thus excluded from our considerations. If, however, a solution exists inside $[a, b]$, it must obey the simplified equation

$$\partial_s \left(p(s_1 s_2|r)\right) = 0 \tag{9}$$

On the other hand, if eqn.(9) has a non-trivial solution outside the interval $]a, b[$ then it must be either $s = a$ or $s = b$, depending on which is closer, because the infinities in the derivatives of $p(r)$ achieve a ”clamping” of obtained fusion results to the known interval $[a, b]$. This can be implemented very efficiently, without solving any equations at all, as a post-processing step of fusion.

Evidently, eqn.(9) needs to be solved both for problem I and II separately, and in general this approach requires that the data generation model be known. So, we present two different solutions for problem I and problem II.

2.3.1 Problem I

Corrupting a clean variable like r, here supposed deterministic so its distribution $p(x|r) = \delta(x - r)$, by additive noise drawn from a distribution $p(x, \sigma)$, implies the convolution of the probability densities of clean and noise variables from which the resulting noisy variables are effectively drawn. The result of the convolution is thus the conditional distribution $p(s_i|r)$ of the form:

$$p(s_i|r) = p(s_i - r, \sigma) \tag{10}$$

For making the link to eqn.(9), we observe that the $p(s_i|r)$ are class-conditionally independent, and we can thus express their joint probability $p(s_1 s_2|r)$ by the product of individual probabilities:

$$p(s_1 s_2|r) = p(s_1|r)p(s_2|r) \tag{11}$$

This leads to the following estimates for the underlying value r:

$$r^*_{\text{gauss}} = \sum_i \frac{1/\sigma_i^2}{\sum_j 1/\sigma_j^2} s_i \tag{12}$$

$$r^*_{\text{uni}} = \frac{A + B}{2} \tag{13}$$

where $[A, B] = \bigcap_i [s_i - \sigma_i, s_i + \sigma_i]$. For triangle noise, an analytical treatment is difficult and is therefore omitted from these considerations.

Problem II The only tricky point consists here in computing the quantity $p(s_1 s_2|r)$ required by eqn. (8), which remains valid as $p(r)$ is still uniformly distributed. Still, the calculation is a little more cumbersome since the factorization $p(s_1 s_2|r) = \Pi_i p(s_i|r)$ no longer holds:

$$\begin{aligned} p(s_1 s_2|r) &= \int\int d\tilde{s}_1 d\tilde{s}_2 p(s_1 s_2|\tilde{s}_1 \tilde{s}_2 r) p(\tilde{s}_1 \tilde{s}_2|r) \\ &= \int\int d\tilde{s}_1 d\tilde{s}_2 p(s_1 s_2|\tilde{s}_1 \tilde{s}_2) p(\tilde{s}_1 \tilde{s}_2|r) \\ &= \int\int d\tilde{s}_1 d\tilde{s}_2 p(s_1|\tilde{s}_1) p(s_2|\tilde{s}_2) \delta(f(r, \tilde{s}_1) - \tilde{s}_2) \\ &= \int d\tilde{s}_1 p(s_1|\tilde{s}_1) p(s_2|f(\tilde{s}_1, r)) \end{aligned} \tag{14}$$

where the first transformation follows from the law of total probability: we insert a complete set of disjunct states $\tilde{s}_1 \tilde{s}_2$. In the second line, the factor r has been removed from the conditional probability $p(s_1 s_2|\tilde{s}_1 \tilde{s}_2 r)$ as it can be deduced from $\tilde{s}_1$ and $\tilde{s}_2$. Later, the conditional probability has been split as s_i depends only on $\tilde{s}_i$.

The optimal fused value of r in the interval $[a, b]$ is obtained as before by maximizing eqn. (9). As the resulting expression is in general intractable analytically, we resort to numerical methods to solve it for r, which do work well for Gaussian noise but not for other forms of noise due to numerical problems for uniform noise and analytical intractability for triangular noise.

2.4 Learning approach

The learning approach is schematically depicted in Fig. 1. It is essentially a three-layer neural network that learns a set of plastic, topologically organized prototypes in its hidden layer. A read-out mechanism between hidden and output layer maps the set of graded prototype activities to output values using simple linear regression learning.

2.4.1 Population encoding

In order to increase the computational power of the employed algorithms (see [36]), we adopt a population-coding approach[36] where continuous values of the input and target variables (i.e., the noisy sensor readings s_1, s_2 and r) are represented by placing a Gaussian of variance σ_p onto a discrete position in an one-dimensional activity vector such that the discrete center position is in a linear relationship with the encoded continuous value. As in Sec. 2.2, this discretization is associated with a loss of precision, thus a sufficiently large size of the activity vector must be chosen. Furthermore, the activity vector must have a sufficiently large margin μ around the interval to be encoded because random variables can fall outside this interval and need to be represented as well. The precise way of encoding a scalar value $x \in [0, 1]$ into a vector $\mathbf{v}$ of size n is as follows, using :

$$c = b_{\mu,n}(x) \tag{15}$$

$$v_i = \exp - \frac{(i - c)^2}{2\sigma_p^2} \tag{16}$$

where we have used the discretizing function b from eqn.(3). As a final step in population encoding, the vector $\mathbf{v}$ is normalized to have an L_2 norm of 1.

2.4.2 Neural learning architecture

The architecture is essentially depicted in Fig. 1 and consists essentially of the layers I, P and E, representing the input layer layer obtained by concatenating two population-coded sensor values, a hidden layer and a fusion estimate, respectively. in addition, there is a target layer T that represents the "true" sensor value r.

Generally, we denote neural activity vector in a 2D layer X by $z^X(\mathbf{y}, t)$, and weight matrices feeding layer X, represented by their line vectors attached to target position $y = (a, b)$, by $w^X_{\mathbf{y}}(t)$. For reasons of readability, we often skip the dependencies on space and time and include them only where ambiguity would otherwise occur. Thus we write z^X instead of $z^X(\mathbf{y}, t)$ and w^X instead of $w^X_{\mathbf{y}}(t)$. Using this notation, the two weight matrices that are subject to learning in this architecture are the connections from I to T, w^{SOM} and the weights from T to E, w^{LR}.

$$
\begin{aligned}
\bar{z}^P(\mathbf{y}) &= w^{\mathrm{SOM}}_{\mathbf{y}} \cdot z^I && (17)\\
z^P &= \mathrm{TF}\left(\bar{z}^P\right) && (18)\\
z^E &= w^{\mathrm{LR}} \cdot z^P && (19)\\
w^{\mathrm{LR}}_{\mathbf{y}}(t+1) &= w^{\mathrm{LR}}_{\mathbf{y}} + 2\epsilon^{\mathrm{LR}} z^I \left(z^E(\mathbf{y}) - z^T(\mathbf{y})\right) && (20)\\
w^{\mathrm{SOM}}_{\mathbf{y}}(t+1) &= \mathrm{norm}\ \left(w^{\mathrm{SOM}}_{\mathbf{y}} + \epsilon^{\mathrm{SOM}} g_\sigma(\mathbf{y} - \mathbf{y}^*)(z^I - w^{\mathrm{SOM}}_{\mathbf{y}})\right) && (21)\\
& && (22)
\end{aligned}
$$

where $g_s(x)$ is a zero-mean Gaussian function with standard deviation s and $\mathbf{y}^*$ denotes the position of the best-matching unit (the one with the highest similarity-to-input) in P. In accordance with standard SOM training practices, the SOM learning rate and radius, ϵ^{SOM} and σ, are maintained at ϵ_0, σ_0 for $t < T_1$ and are exponentially decreased afterwards in order to attain their long-term values $\epsilon_\infty, \sigma_\infty$ at $t = T_{\mathrm{conv}}$. The learning rate of linear regression, ϵ^{LR} remains constant during at all times. TF represents a monotonous non-linear transfer function, $\mathrm{TF} : [0,1] \rightarrow [0,1]$ which we model as follows with the goal of maintaining the BMU value unchanged while non-linearly suppressing smaller values:

$$
\begin{aligned}
m_0 &= \max_{\mathbf{y}} \bar{z}^P(\mathbf{y}, t)\\
m_1 &= \max_{\mathbf{y}} \left(z^P(\mathbf{y}, t)\right)^{20}\\
\mathrm{TF}\left((z^P(\mathbf{y})\right) &= m_0 \frac{\left(z^P(\mathbf{y})\right)^{20}}{m_1} && (23)
\end{aligned}
$$

2.4.3 Rejection strategy for super-Bayesian fusion

In this setting, we simply reject an incoming sample, i.e., take no decision, if the simple criterion

$$\max z^E > \theta \tag{24}$$

$$\text{with } \theta(t+1) = (1-\alpha)\theta(t) + \alpha \max z^E(t). \tag{25}$$

is fulfilled. simply put, we check whether the highest activated unit in the output layer E has an activity that is higher than the temporal average of past maximal activities, calculated by exponential smoothing. This is pretty ad hoc and not rigorously justified, but we find that in practice this strategy gives significant performance improvements and in no case that we could observe deteriorates performance.

2.4.4 Novelty detection

As the hidden SOM layer implements a generative model of the sensory input, it should be able to recognize out-of-the-ordinary samples, i.e., outliers. This is particularly important for detection persistent changes in input statistics which must be countered by adapting the fusion model. Here, a change detection mechanism could provide, first of all, a means to detect when a model should be adjusted to new realities, and furthermore to stop fusion until this has been successfully done. Such an ability is therefore imperative for life-long learning in embodied agents and should be considered a significant advantage. We approach change detection by simply monitoring the temporal average activity of the best-matching until (BMU) in the hidden SOM layer P. This is done because we assume that the SOM prototypes represent the joint distribution of s_1 and s_2 in input space; any significant deviation from this distribution should therefore result in lower input-prototype similarity which results in lower activity. Again, the temporal average is calculated by exponential smoothing and thus requires no memory overhead. The smoothing constant β has to be set such that short-term random fluctuations are smoothed away whereas long-term systematic changes are retained.

3 EXPERIMENTS

For all experiments, we use an interval of $[0, 1]$ for r, s_1 and s_2. Each experiment is repeated 25 times, each time with a different pairing of standard deviations which are chosen for each sensor from the following fixed set: 0.016, 0.032, 0.048, 0.064 and 0.08.

Joint probability estimation parameters The discretization step size is set to $n = 100$, and the joint probability matrix is built for n^3 iterations. The margin parameter μ is set to $\mu = 0.2$.

Model-based Bayesian inference parameters This method is parameter-free in the sense that it only uses the parameters contained in the data generation model. There are a few parameters tied to the numerical solution of integrals but the standard values of the numerical solvers always work well so it is not necessary to include them here.

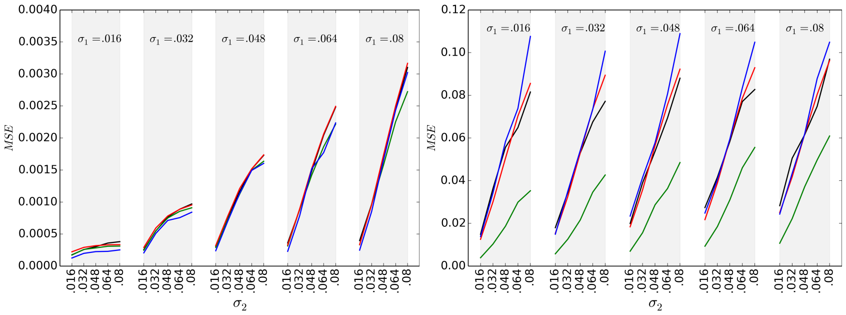

Fig. 3 Comparison of fusion performance under Gaussian noise for four methods on problem I (left) and problem II (right): Bayesian inference using estimated joint probabilities (red), model-based Bayesian inference (blue), learning approach(black) and learning approach using super-Bayesian fusion(green).

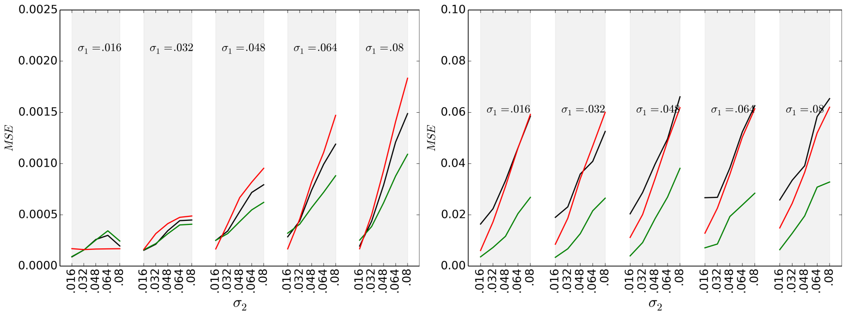

Fig. 4 Comparison of fusion performance under uniform noise for four methods on problem I (left) and problem II (right): Bayesian inference using estimated joint probabilities (red), learning approach(black) and learning approach using super-Bayesian fusion(green). Model-based Bayesian inference is not practicable for uniform noise and therefore not shown.

Learning approach Here, several parameters need to be fixed: the hidden layer contains 15x15=225 units, the output layer has $n = 100$ units. The margin parameter for population encoding is set to $\mu = 0.2$. The variance of Gaussians for population encoding is fixed at $\sigma_p = 3$ pixels. The LR learning rate is $\epsilon^{\text{LR}} = 0.01$ and the parameters for decreasing SOM learning rate and radius are: $\epsilon_0 = 0.1$, $\sigma_0 = 5$, $\epsilon_\infty = 0.01$, $\sigma_\infty = 0.5$, $T_{\text{conv}} = 5000$, $T_1 = 1000$. Total training time is always 20.000 iterations unless otherwise mentioned, and testing is conducted subsequently for 20.000 iterations for calculating performance statistics, with learning turned off. The smoothing parameter for super-Bayesian fusion is $\alpha = 0.001$, and the smoothing parameter for change detection is $\beta = 0.001$ as well.

3.1 Comparison of fusion performances

In this experiment, we compare the performances of all three fusion methods (Bayesian inference by joint probability estimation, model-based Bayesian inference and our learning approach) for problem I and problem II, each time using three noise types (Gaussian, uniform and triangular noise) as described in detail in Sec. 2.1. For each method,

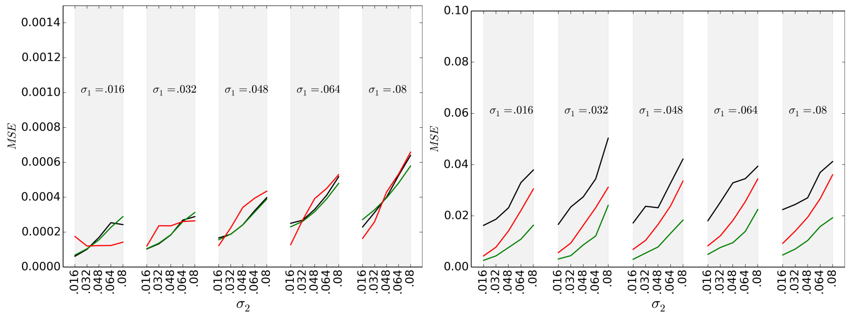

Fig. 5 Comparison of fusion performance under triangular noise for four methods on problem I (left) and problem II (right): Bayesian inference using estimated joint probabilities (red), learning approach(black) and learning approach using super-Bayesian fusion(green). Model-based Bayesian inference is not practicable for triangular noise and therefore not shown.

problem and noise type we conduct 25 separate experiments, corresponding to all possible combinations of standard deviations given above. In this way, the behavior of each fusion methods is sampled uniformly in a representative range of noise strengths in a way that can be directly compared. Model-based Bayesian inference runs into problems when using uniform noise because the numerical solution of the involved integrals becomes numerically unstable, requiring interval discretization that render the problem intractable. As for triangular noise, it poses severe problems for the analytical derivation due to its non-diffentiability at $x = 0$, see Fig. 2. Model-based Bayesian inference is therefore conducted for Gaussian noise only. We observe from figs. 3, 4, 5 that fusion performances are always very similar for problem I regardless of noise, probably because this problem is rather simple in nature. For problem II, apart from an overall decrease in precision, we observe that super-Bayesian fusion always performs best, sometimes by large margins. in the parametrization used here, around 50% of samples were rejected for the latter. The only noticeable deviations for problem II occur for triangular noise, for unknown reasons. As we measure the mean squared error here, however, and not is square root, these differences are less than they appear, and overall we can state that all fusion methods independently obtain very similar results on problem II, too, once more supporting the global consistency of our experiments.

Memory usage is non-existent for model-based Bayesian fusion, negligible for the learning approach and enormous for joint probability estimation, as a matrix of n^3 elements is to be represented for $n = 100$. In terms of computation speed, model-based Bayesian inference is efficient for training as it does not need to be trained, whereas a training session for the learning approach takes approximately 1 minute and 5 minutes for inference by joint probability estimation. Execution of models is rather similar for all three methods, where around 100 estimations can be performed per second.

3.2 Change detection performance

For this experiment, we introduce the concepts of a slight, medium and severe change in input statistics. Working on problems I and II using Gaussian noise, we train the system with noise standard deviations of $\sigma_1 = \sigma_2 = 0.016$, and then change these parameters after the first half of the testing phase, i.e., without retraining the system.

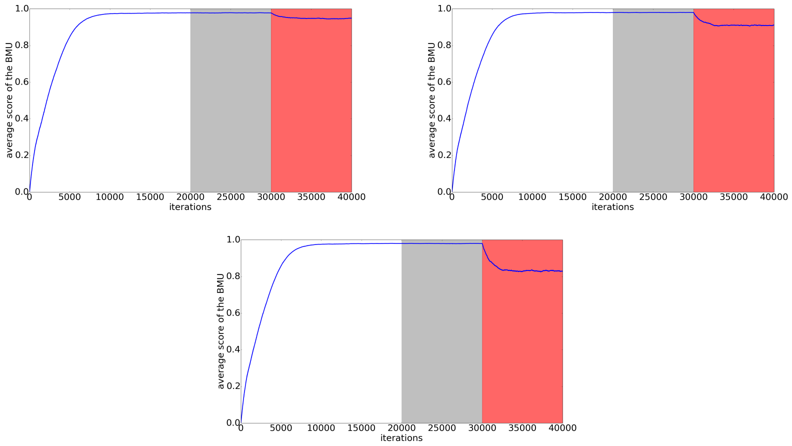

Fig. 6 Change detection performance of the learning approach, evaluated for problem I. Beginning of the testing phase, with unmodified standard deviations, is marked in grey, whereas the second half of the testing phase where a change of statistics occurs is marked in red. Change conditions are slight (upper left), medium (upper right) and severe (centered).

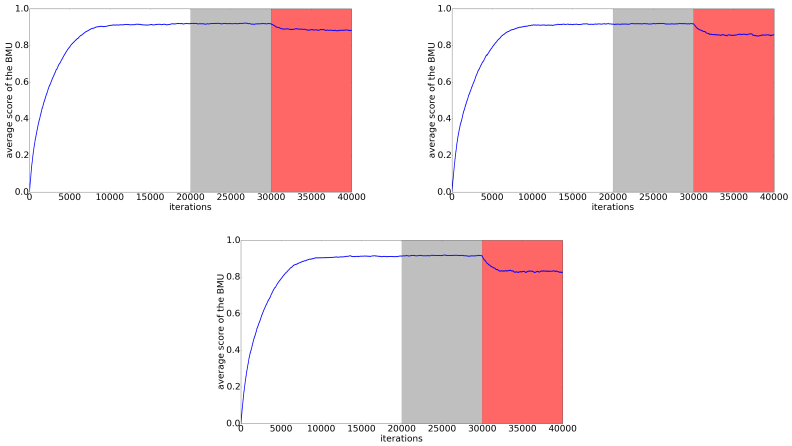

Fig. 7

Fig. 8 Change detection performance of the learning approach, evaluated for problem II. Beginning of the testing phase, with unmodified standard deviations, is marked in grey, whereas the second half of the testing phase where a change of statistics occurs is marked in red. Change conditions are slight (upper left), medium (upper right) and severe (centered).

We therefore expect a reaction to this change, as described in Sec. 2.4.4. Each change condition differs only in the new values for the standard deviations: $\sigma_1 = \sigma_2 = 0.032$ (slight), $\sigma_1 = \sigma_2 = 0.048$ (medium) and $\sigma_1 = \sigma_2 = 0.08$ (severe). From Figs. 8, 8 we can observe that even slight changes can be reliably detected by the system since there is always a significant decease in average BMU activity.

4 DISCUSSION

When we compare the three fusion methods we investigated in the light of the criteria we put forward in Sec. 1.2, namely precision, generality, change detection capacity and resource-efficiency, we find a mixed picture at fist glance.

All proposed methods are pretty equal in terms of precision, except notably the super-Bayesian fusion by the learning approach, although this is not completely fair as it cannot treat all samples. Still, it should be noted that the Bayesian "optimal" fusion can be surpassed significantly by a rather simple and efficient approach, which can be very useful to any artificial cognitive agent.

In terms of generality, it is clearly the learning approach and Bayesian inference by joint probability estimation that are most generally applicable as they can cope with any problem under any type of noise, which is something that model-based Bayesian inference is incapable of.

Regarding resource efficiency, especially in the light of an application in artificial cognitive agents, it is clearly the learning approach that is most favorable: it has both a favorable execution and training time, and it is very memory-efficient. In fact, by reducing the size of the hidden SOM layer, one can gradually trade memory usage for precision, making use of the graceful decay property of SOMs for this purpose. Bayesian inference by joint probability estimation has the problem of training time and memory usage that grow cubically with the discretization step n, quickly rendering this approach impracticable where high precision is needed. Model-based Bayesian inference is memory and time-efficient as well but is not very suited t artificial agents as it is incapable of adapting.

For change detection capacity, it is the learning approach that wins the competition because it offers a very efficient-to-compute criterion to detect even rather slight changes in input statistics, using only quantities like the BMU score that are calculated anyway and thus do not impose a computational burden.

Based on all these points, we may safely conclude that the two learning approaches to multi-sensory fusion are certainly preferable due to their generality and resource efficiency. The learning architecture we presented possesses the additional capacity to perform change detection and super-Bayesian fusion which are very important points in their own right, and it rather more resource-efficient than Bayesian inference by joint probability estimation.

5 Summary, conclusion and future work

We have presented a comparison of three methods for perform multi-sensory fusion in a simulated setting that is nevertheless very closely modeled after real tasks. We compared these methods in terms of precision, generality, change detection capacity and resource-efficiency, and found that the self-organized neural network was most suited, in summary, for application in artificial cognitive agents, thus making a very strong statement in favor of learning method in multi-sensory fusion. We furthermore investigated a simple way to improve fusion performance between the Bayesian optimum and found it both practicable and beneficial for performance under the condition that one accepts to ignore a certain percentage of incoming samples. As a last, we investigated how fusion might be continuously updated and re-calibrated by detecting significant changes in input statistics, and found that the detection of such changes is feasible and simple for the presented system.

In future work, we wish to investigate the issue of incremental learning for multisensory fusion, meaning that upon the detection of changed input statistics, the learned fusion model should be adapted in a way that allows stable life-long learning. In addition, verifying these algorithm on a real-world fusion task will be an important validation of the presented theoretical work.

6 Compliance with ethical standards

This article does not contain any studies with human participants or animals performed by any of the authors. Thomas Hecht has received a research grant from the Direction Générale de l'Armement (DGA), France. Alexander Gepperth, Thomas Hecht and Mandar Gogate declare that they have no conflict of interest.

References

1. Marc O Ernst and Martin S Banks. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, 415(6870):429–433, Jan 2002.
2. Dora E. Angelaki, Yong Gu, and Gregory C. DeAngelis. Multisensory integration: psychophysics, neurophysiology, and computation. *Current opinion in neurobiology*, 19(4):452–458, 2009.
3. Marc O. Ernst and Heinrich H. Blthoff. Merging the senses into a robust percept. *Trends in cognitive sciences*, 8(4):162–169, 2004.
4. Michael S. Beauchamp. See me, hear me, touch me: multisensory integration in lateral occipital-temporal cortex. *Current opinion in neurobiology*, 15(2):145–153, 2005.
5. Barry E. Stein and Terrence R. Stanford. Multisensory integration: current issues from the perspective of the single neuron. *Nature Reviews Neuroscience*, 9(4):255–266, 2008.
6. Jon Driver and Toemme Noesselt. Multisensory interplay reveals crossmodal influences on sensory-specificbrain regions, neural responses, and judgments. *Neuron*, 57(1):11–23, 2008.
7. Mark T. Wallace. The development of multisensory processes. *Cognitive Processing*, 5(2):69–83, 2004.
8. Gemma A. Calvert and Thomas Thesen. Multisensory integration: methodological approaches and emerging principles in the human brain. *Journal of Physiology-Paris*, 98(1):191–205, 2004.
9. Asif A. Ghazanfar and Charles E. Schroeder. Is neocortex essentially multisensory? *Trends in cognitive sciences*, 10(6):278–285, 2006.
10. George M. Stratton. Vision without inversion of the retinal image. *Psychological review*, 4(4):341, 1897.
11. Ian P. Howard and William B. Templeton. Human spatial orientation. 1966.
12. Harry McGurk and John MacDonald. Hearing lips and seeing voices. 1976.
13. Matthew Botvinick and Jonathan Cohen. Rubber hands' feel'touch that eyes see. *Nature*, 391(6669):756–756, 1998.
14. Ladan Shams, Yukiyasu Kamitani, and Shinsuke Shimojo. What you see is what you hear. *Nature*, 2000.
15. Andrew King. Development of multisensory spatial integration. 2004.
16. Monica Gori, Michela Del Viva, Giulio Sandini, and David C. Burr. Young children do not integrate visual and haptic form information. *Current Biology*, 18(9):694–698, 2008.
17. Jacob G. Martin, M. Alex Meredith, and Khurshid Ahmad. Modeling multisensory enhancement with self-organizing maps. *Frontiers in computational neuroscience*, 3, 2009.
18. Thomas J. Anastasio and Paul E. Patton. A two-stage unsupervised learning algorithm reproduces multisensory enhancement in a neural network model of the corticotectal system. *The Journal of neuroscience*, 23(17):6713–6727, 2003.

19. Athanasios Pavlou and Matthew Casey. Simulating the effects of cortical feedback in the superior colliculus with topographic maps. In *Neural Networks (IJCNN), The 2010 International Joint Conference on*, pages 1–8. IEEE, 2010.
20. Julien Mayor and Kim Plunkett. A neurocomputational account of taxonomic responding and fast mapping in early word learning. *Psychological review*, 117(1):1, 2010.
21. Johannes Bauer, Cornelius Weber, and Stefan Wermter. A som-based model for multisensory integration in the superior colliculus. In *Neural Networks (IJCNN), The 2012 International Joint Conference on*, pages 1–8. IEEE, 2012.
22. Apostolos Georgakis, Haibo Li, and Mihaela Gordan. An ensemble of SOM networks for document organization and retrieval. In *Int. Conf. on Adaptive Knowledge Representation and Reasoning (AKRR05)*, page 6, 2005.
23. Bruno Baruque and Emilio Corchado. A bio-inspired fusion method for data visualization. In *Hybrid Artificial Intelligence Systems*, pages 501–509. Springer, 2010.
24. Hujun Yin. ViSOM-a novel method for multivariate data projection and structure visualization. *Neural Networks, IEEE Transactions on*, 13(1):237–243, 2002.
25. Tamas Jantvik, Lennart Gustafsson, and Andrew P. Papliski. A self-organized artificial neural network architecture for sensory integration with applications to letter-phoneme integration. *Neural computation*, 23(8):2101–2139, 2011.
26. Valentina Gliozzi, Julien Mayor, Jon-Fan Hu, and Kim Plunkett. The impact of labels on visual categorisation: A neural network model. 2008.
27. Michael S. Landy, Martin S. Banks, and David C. Knill. Ideal-observer models of cue integration. *Sensory cue integration*, pages 5–29, 2011.
28. David C. Knill and Alexandre Pouget. The bayesian brain: the role of uncertainty in neural coding and computation. *TRENDS in Neurosciences*, 27(12):712–719, 2004.
29. Robert A. Jacobs. Optimal integration of texture and motion cues to depth. *Vision research*, 39(21):3621–3629, 1999.
30. Peter W. Battaglia, Robert A. Jacobs, and Richard N. Aslin. Bayesian integration of visual and auditory signals for spatial localization. *JOSA A*, 20(7):1391–1397, 2003.
31. Marc O. Ernst. A bayesian view on multimodal cue integration. *Human body perception from the inside out*, pages 105–131, 2006.
32. Hannah B. Helbig and Marc O. Ernst. Optimal integration of shape information from vision and touch. *Experimental Brain Research*, 179(4):595–606, 2007.
33. Mustapha Makkook, Otman Basir, and Fakhreddine Karray. A reliability guided sensor fusion model for optimal weighting in multimodal systems. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 2453–2456. IEEE, 2008.
34. Xuan Song, Jinshi Cui, Huijing Zhao, and Hongbin Zha. Bayesian fusion of laser and vision for multiple people detection and tracking. In *SICE Annual Conference, 2008*, pages 3014–3019. IEEE, 2008.
35. Lasse Klingbeil, Richard Reiner, Michailas Romanovas, Martin Traechtler, and Yiannos Manoli. Multi-modal sensor data and information fusion for localization in indoor environments. In *Positioning Navigation and Communication (WPNC), 2010 7th Workshop on*, pages 187–192. IEEE, 2010.
36. A Gepperth, B Dittes, and M Garcia Ortiz. The contribution of context information: a case study of object recognition in an intelligent car. *Neurocomputing*, 2012.