Reward-based learning of optimal cue integration in audio and visual depth estimation

Cem Karaoguz, Thomas Weisswange, Tobias Rodemann, Britta Wrede, Constantin Rothkopf

2011

Preprint:

This is an accepted article published in Proceedings of ICAR 2011. The final authenticated version is available online at: https://doi.org/[DOI not available]

Submitted to ICAR 2011, June 20 – 23, 2011 in Tallinn, Estonia

Reward-based learning of optimal cue integration in audio and visual depth estimation

Cem Karaoguz[1,2,*], Thomas H Weisswange[3,*], Tobias Rodemann[2,*], Britta Wrede[1] and Constantin A Rothkopf[3]

***Abstract*— Many real-world applications in robotics have to deal with imprecisions and noise when using only a single information source for computation. Therefore making use of additional cues or sensors is often the method of choice. One examples considered in this paper is depth estimation where multiple visual and auditory cues can be combined to increase precision and robustness of the final estimates. Rather than using a weighted average of the individual estimates we use a reward-based learning scheme to adapt to the given relations amongst the cues. This approach has been shown before to mimic the development of near-optimal cue integration in infants and benefits from using few assumptions about the distribution of inputs. We demonstrate that this approach can substantially improve performance in two different depth estimation systems, one auditory and one visual.**

I. INTRODUCTION

The combination of different cues to improve the performance in tasks like segmentation [1], [2], object identification [3], or object tracking [4] is a common method in robotics. Merging different cues with complementary or partially redundant characteristics has a good potential to improve both precision and robustness (for example reduce mean and maximum estimation error). The optimal integration is theoretically well defined and straightforward in a Bayesian framework. However, for real world applications, performing this computation is usually intractable. A commonly used approximation is a weighted sum of the maximum likelihood estimates of the different cues [5]. Such an approach is computationally efficient but is only guaranteed to be close to the optimal solution for the idealized case of cues with uncorrelated Gaussian noise and knowledge of error variances and potential biases. Unfortunately for many practical applications these conditions are not necessarily met. Additionally fixed weights can lead to problems when the environment changes. An adaptive approach has been presented in [4], where an optimal weighting of different cues in an object tracking task is learned online.

In this work we present a different approach that uses a general reward-based learning scheme for training a neural network to combine depth estimations from multiple cues. The approach was developed to model the development of optimal cue integration in infants [6], [7] without making assumptions about the characteristics of the cues. We test this approach in two different robotics tasks. The first is auditory depth estimation using stereo recordings from a humanoid robot, the second one is a visual depth estimation task in a stereo camera setup with vergence. Such depth estimation is a basic prerequisite for important behaviors like navigation, grasping or verbal interaction. In both sensory domains this task is challenging if only standard sensors (i.e. two microphones or two cameras) are available.

Both systems have been described in previous papers [8], [9] and will only be explained briefly here. The cue integration method was outlined in detail in [6], [7].

For both applications learning is done offline in a standard training session using only part of the recorded sensory data. The reward signal used to adapt the neural network is based on the accuracy of its response to an input. Since we have labeled data, this accuracy simply depends on the difference between estimated and true depth, but in general could relate to a behavioral outcome, e.g. the success of a grasping movement. The weight are updated using a gradient descent method. After training, performing cue integration can be done with minimal computational effort.

For both sensory domains our results show a substantial reduction in mean and maximum depth estimation error compared with those of the best individual cue. This was possible although the quality of individual cues varied heavily across the input space, they were strongly correlated and showed significant biases. For these reasons the new method was able to also outperforme standard weighted cue averaging.

II. AUDIO DEPTH ESTIMATION

Estimating the depth of a sound source is notoriously difficult, especially when only one or two microphones are available. If no triangulation is possible (either by moving the robot or by using several pairs of microphones) no direct, unambiguous cue to depth is available. In a previously described system [8] we therefore used a combination of many different depth cues (outlined below) that were computed and averaged over a complete sound segment. In this framework sounds are considered as (proto) objects with a set of attached audio features. Each of these audio features i is mapped to a depth estimation $D^i_{audio}(z)$, where $D^i_{audio}(z)$ is the evidence for one of 9 different depths (z)

1 Research Institute for Cognition and Robotics (CoR-Lab), Bielefeld University, 33594 Bielefeld, Germany ckaraogu@cor-lab.uni-bielefeld.de

2 Honda Research Institute Europe GmbH, Carl-Legien-Str. 30, 63073 Offenbach, Germany tobias.rodemann@honda-ri.de

3 Frankfurt Institute for Advanced Studies, Ruth-Moufang-Strasse 1, 60438 Frankfurt, Germany weisswange@fias.uni-frankfurt.de

∗ These authors contributed equally

based on the current values of cue i . The mapping from audio cues to depth evidences is learned in a calibration session using more than 600 sounds for each distance. Due to technical constraints, we only tested the distances $z \in [0.5m, 1m, 1.5m, 2m, 2.5m, 3m, 4m, 5m, 6m]$.

Approximate Bayesian cue integration [5] was used to combine the estimations from different cues. This is formulated as:

$$z^{all}_{audio} = \sum_i w_i \cdot z^i_{audio} \quad , \tag{1}$$

where z^{all}_{audio} is the estimated distance using all audio cues and z^i_{audio} the maximum likelihood distance estimatie of audio cue i. The weight w is expressed as:

$$w_i = \frac{1/\sigma_i^2}{\sum_j 1/\sigma_j^2}, \tag{2}$$

where σ_i is the mean localization error of feature i and j iterates over all depth cues. Errors are computed as the difference (in meters) between the true position and the maximum-likelihood depth.

A. Localization cues

In the following we provide a short explanation of different localization cues. A detailed description is not in the focus of this article. Rather we want to stress the different performance characteristics of the different cues (see Fig. 3), and note that errors are in general not Gaussian and unbiased. Furthermore we expect substantial correlations between different cues. The system uses the following depth estimation cues (see also Fig. 1):

1) Mean envelope amplitude: This cue represent the mean amplitude (related to mean energy of the sound) with roughly a $1/z$ relation over distance z. Since the measured signal amplitude also depends on the production amplitude (which is unknown), this cue depends on the distribution of production amplitude values in training and test datasets. Despite of this we got good results for very close and far distances.

2) Spectral envelope: This cue measures the mean amplitude (energy) of the sound in different frequency bands. It is known that higher frequency bands are more strongly attenuated with distance than lower frequency bands. This cue is very weak over the depth range tested in our set-up.

3) Binaural cues (IID and ITD): Interaural Intensity Difference (IID) and Interaural Time Difference (ITD) are two standard cues for horizontal sound localization, showing a strong dependency on the azimuth angle of the sound relative to the robot's head. They also exhibit a weak dependency on the sound's elevation (see [10]) and also the sound source's depth. These cues are quite useful especially for shorter depths, but decrease in performance when e.g. the robot is moving [8].

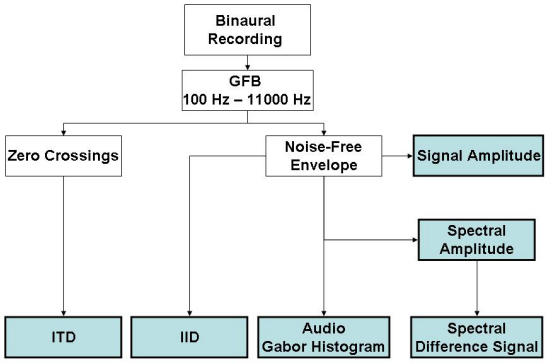

Fig. 1. Basic flowchart of the audio cue processing system.

4) Binaural spectral difference cues: Similar to IID and ITD this cue is based on differences in the signal recorded in two different microphones. However, while IID and ITD operate at single frequency channels, binaural spectral cues look at the distribution of binaural differences over a range of frequencies. This cue shows a similar performance as the above mentioned binaural cues.

5) Audio Gabors: Here we compute histograms of filter responses. These filters are 2D Gabor filters known from image processing, applied on the spectral envelope of the audio signal. Our assumption was that with changing depth filter response histograms might vary. However, on the tested range of depths, results were not very promising. We included this cue into our test to evaluate if the cue integration can also deal with low-performance cues.

B. Comparison to Related Work

Most approaches for audio depth estimation in robots are based on either motion triangulation [11], [12] or larger microphone arrays [13]. The first approach requires a motion of the robot while the sound source stays active. Using larger microphone arrays limits the robot mobility and will probably only be effective when the source is relatively close to the array. Animals in turn seem to employ a totally different approach using only two ears and without relying on any ego-motion (see e.g. [14], [15]).

C. Experimental setup

The sounds to be localized were recorded in our robot lab, a room of dimension (12 x 11 x 2.8 m) with a substantial amount of echo ($T_{60} = 810\ ms$). A loudspeaker set at different depths from the robot head generated in total 68 different sounds (speech, environmental sounds, music). We also rotated the head to 19 different pan positions. The database therefore consists of 19*68 = 1292 recorded sounds for each of the 9 distances. The recording set-up is sketched in Fig. 2.

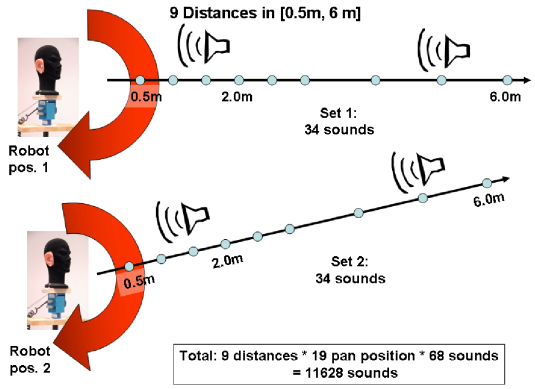

Fig. 2. Outline of the experimental sessions to record sound data. Note that the two sets were recorded at slightly different positions of robot and speakers. Training and test data consists of half of the data from set 1 and set 2, each.

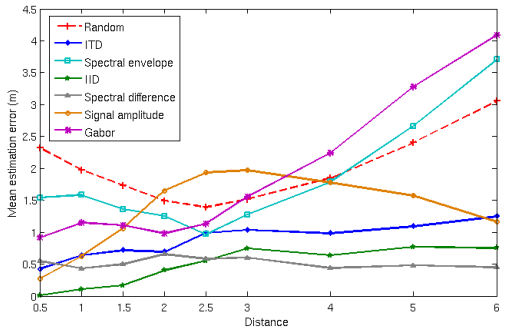

Fig. 3. Mean estimation error of all auditory cues for different depths averaged over all sounds.

Cue	mean error	rel. error	off-target	conf. near/far
Random	2.0	1.22	89%	9.5%
Amplitude	1.33	0.56	74%	2.8%
Spectral	1.71	0.98	78%	9.8%
IID	0.46	0.15	28%	1.5%
ITD	0.86	0.41	50%	2.3%
Gabor	1.82	0.78	85%	12.1%
Spec. Diff.	0.52	0.32	27%	1.6%
Combined	0.51	0.25	44%	0.34%

TABLE I

AUDIO DEPTH ESTIMATION PERFORMANCE . COMBINED RESULTS ARE COMPUTED AS DESCRIBED IN EQN. 1. FOR THE COMBINED CUES THE OFF-TARGET VALUE WAS COMPUTED BY BINNING THE ESTIMATED DISTANCE TO THE MEASUREMENT DISTANCES.

D. Baseline results

Results for the first four cues on a slightly different data set have been presented in detail in [8]. Here we just summarize the main results of individual cues (see Tab. I). The table shows the mean localization error in meters, the relative error (mean error devided by distance), the probability of a mis-localization, and the probability of a severe mislocalization, defined as instances where the estimated distance was more than 4 m away from the true depth. Also for the first time we report on the results for the Binaural Spectral Difference cue and the Audio Gabor filter response histograms.

Figure 3 shows the estimation errors of the single cues for objects at different depths. The accuracy of most of the cues decreases with depth, only the spectral difference cue keeps it performance at all depths.

III. VISUAL DEPTH ESTIMATION

In [9], a series of visual depth estimation experiments have been undertaken comparing three different cues (stereo disparity, vergence and familiar size). The goal of this study was to determine how accurately depth can be estimated, which cue is more accurate in what depth region, what are the main error sources and how can the estimation be improved by combination of the cues. Based on the statistical data obtained from the experiments, approximate Bayesian cue integration, explained in Eq. 1 and Eq. 2, is used to combine estimations from different cues.

A. Localization cues

The following cues are used for visual depth estimation:

1) Vergence: A visual system capable of changing its camera parameters can achieve stereo fixation on an object by positioning the intersection point of the line of sight of the two cameras on the surface of the object. The distance to the fixation point can be derived from the triangulation using a pinhole camera model (Fig. 4) as:

$$z = \frac{b}{2 \cdot tan(\frac{\Theta_v}{2})}, \tag{3}$$

where Θ_v is the vergence angle and b is the baseline (Fig. 4). The vergence angle is computed from left and right camera angles as $\Theta_v = \Theta_{left} + \Theta_{right}$. We use symmetric vergence ($|\Theta_{left}| = |\Theta_{right}| = \Theta$).

2) Stereo Disparity: In a visual system using vergence, points belonging to objects that are residing out of the stereo fixation point project to different locations on the left and right camera images. This difference is referred to as stereo disparity. In active vision systems disparity information is relative to the fixation point (i.e. points belonging to an object under fixation have disparities of zero or very close to zero). In order to obtain absolute disparity information (i.e. the distance from the baseline to the stereo fixated object) an active rectification process [16] is used. As shown in Fig. 4 this process epipolarly rectifies images from an active stereo camera configuration (cameras with solid lines) to virtual image planes of a parallel stereo camera configuration (cameras with red dotted lines). After applying rectification, depth from disparity can be computed as:

$$z = \frac{bf}{d} + r + f, \tag{4}$$

where d is the horizontal disparity (defined as $d = x_{VL} - x_{VR}$, x_{VL} and x_{VR} being the projections of the object on the

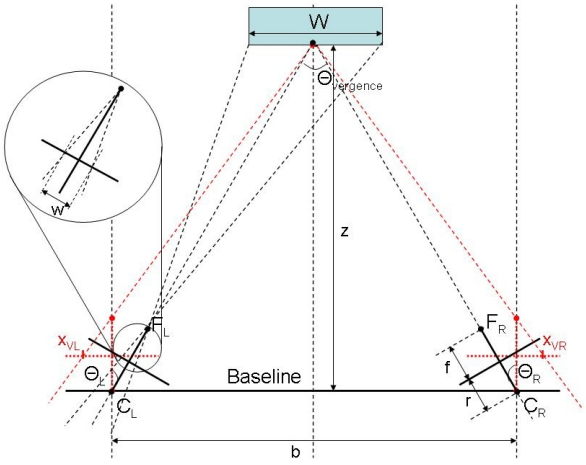

Fig. 4. Analytical model of the active vision system and depth estimation methods. Depth z is defined by the distance from the baseline to the object. F_L and F_R denote focal points, C_L and C_R denote center of rotations of the left and right cameras respectively. The static parameters of our system are as follows: r = 18.75 mm, f = 5.4 mm, b = 65 mm.

virtual left and right image planes), f is the focal length of the cameras and r is the distance from the center of rotation of the cameras to the image planes (Fig. 4).

The OpenCV version 2.0 [17] block matching algorithm is used for disparity computation. This algorithm provides a dense disparity map of given left and right camera image pairs without any post-processing applied. The disparity search range is set to 32 pixels (-16 to +15). A simple color based segmentation process is used to distinguish the disparities corresponding to the object in the disparity maps and the average of these disparities is taken for depth estimation in each frame.

3) Familiar Size: If the real size of an object is known, the depth of the object can be estimated from the size of its projection on the camera images. Using a pinhole camera model (Fig. 4) the depth can be derived as:

$$z = \left(\frac{fW}{w} + r + f\right) cos(\Theta), \tag{5}$$

where Θ is the camera angle and $cos(\Theta) \approx 1$ due to the small baseline. The physical size W for all objects used in the experiments is measured beforehand, the retinal size w is computed using a simple color based segmentation process (the same used for the stereo disparity method). The width of the objects is used for estimation since it showed better overall accuracy compared to the height. More advanced methods (e.g. [18]) could be used to improve the precision of estimations.

B. Related Work

A limited comparison between vergence and photogrammetry methods was done in [19]. In [20] and [21] reviews on different stereo disparity computation methods were presented. However these were restricted to a static-parallel stereo camera setup. We examine three depth estimation methods (stereo disparity, vergence and familiar size) with an active vision setup in an extensive test setting.

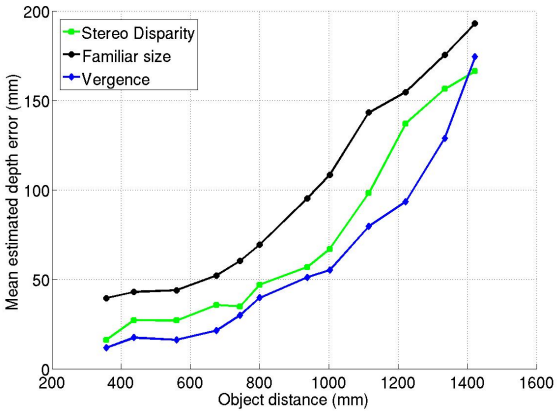

Fig. 5. Mean estimation errors of the methods. Plot shows the running average over 8 data points.

C. Experimental Setup

An experimental stereo vision head with 4 DoF (2 DoF for head and 1 DoF for each camera) and a baseline of 65 mm is used as a platform for the examined depth estimation methods. All experiments are performed using images with a resolution of 400x300 pixels, which is the standard resolution for most of our vision applications [22]. A linear unit that moves a small object platform on a linear axis is utilized to rapidly and autonomously generate data for depth estimation algorithms and acquire *ground-truth* depth information. The object platform is moved via a stepper motor within an error of 0.1%. 11 objects were selected from the HRI150 database [23]. One of the objects is used for calibration purposes.

D. Baseline results

Mean estimation errors[1] of individual methods are shown in Fig. 5. The error is defined as the absolute value of the difference between the estimated depth and actual depth. Tab. II shows mean estimation errors of individual cues and their combinations by weighted averaging in three ranges. Overall comparison of results shows that combinations of methods via Bayesian Cue Integration did not produce the best results in all ranges. The reason for this might be that the weighted averaging requires unbiased and uncorrelated signals to be guaranteed to be close to the optimal solution and these requirements might not be met.

IV. REWARD-BASED LEARNING OF CUE INTEGRATION

As we can see from the previous sections, using weighted averaging to combine estimates of multiple cues does not necessarily improve performance that much. This is due to the fact that the requirements for this approximation to be valid are not met, mostly the independence assumption for the cue's noise, and also due to different values of the optimal

[1]The mean estimation error is averaged over all 10 objects.

TABLE II

MEAN (AND STANDARD DEVIATION) OF ESTIMATION ERRORS (IN MM) FOR ALL OBJECTS AT DIFFERENT RANGES.

Methods	Near	Middle	Far
Vergence	16.52 (6.65)	44.46 (13.18)	131.31 (46.99)
FS	44.64 (6.11)	86.09 (23.81)	175.38 (33.27)
SD	27.15 (13.91)	53.02 (21.13)	141.14 (32.47)
Combinations via approximate Bayesian cue integration			
Vergence+FS	17.63 (9.60)	36.64 (14.55)	123.62 (109.52)
SD+FS	26.69 (19.60)	46.11 (29.03)	134.63 (82.24)
Vergence+SD	19.26 (12.85)	43.83 (16.60)	120.73 (78.03)
Vergence+FS+SD	19.56 (13.68)	37.52 (19.47)	121.16 (77.92)

weights for different input regions. Therefore we decided to use a method that learns how to best combine the given cues from data.

This method was developed to try to explain psychophysical findings showing that human infants often combine multiple cues sub-optimally or not at all [24], [25]. In contrast, experiments could show that adult performance in many multi-cue tasks is close to predictions from the optimal Bayesian model. More details on this and references can be found in previous publications [6], [7]. The principle behind the approach is learning a mapping from an input of multiple cue estimates to reward predictions of the possible responses of the robot. We use the reward as a loose signal for the quality of our final estimate inspired by reinforcement learning theory [26]. In our case an action is equivalent to a specific depth estimate and is chosen based on the reward predictions. The error between the prediction and the true received reward is used to train the model.

In [7] we could show that this model is indeed able to perform as good as a full, numerically simulated, Bayesian observer, which uses explicit knowledge of all prior and likelihood distributions as well as the reward function. Additionally these results could be extended to cases where inputs could originate from a single or multiple objects in the scene, a problem known as causal inference [27]. In this paper we show the applicability of the framework to real world data from two different robotic setups. While for artificially generated data one can provide idealized input parameters (e.g. using Gaussian distributions, independent noise) to allow for straightforward integration, real data does often not fulfill those assumptions. For that reason it is usually not feasible to compute the optimal Bayesian solution numerically or analytically. Also common fast approximations like weighted averaging were only shown to be accurate given those idealized assumptions [5]. Since our method does not require any knowledge about the input, it is possible to outperform such approximations given complex data. Here we test this prediction for the two datasets described in the previous sections: visual and auditory depth perception.

A. The model

We use a three-layer neural network to approximate the mapping function. The input layer encodes the estimates of the different depth cues into a concatenation of vectors (one for each cue) of neurons i with binary activity x_i. An entry in each one of these vectors represents a range of depth positions, so each vector has only one active neuron at depth estimated by the corresponding cue. The input neurons are all-to-all connected with weights $v_{i,j}$ to j neurons in the hidden layer.

A sigmoidal transfer function on the sum of the weighted inputs gives the outputs y_j of the hidden neurons:

$$y_j = \frac{1}{1 + e^{-\sum_i v_{i,j} x_i}} \tag{6}$$

The hidden neurons are fully connected to output neurons k with weights $w_{j,k}$. Each output unit represents an action, and its activation z_k is the reward expected when performing this action. All weights are drawn from uniform distributions, V between -0.1 and 0.1, W between -1 and 1.

Based on these outputs we choose one action $\hat{k}$ by using the softmax function on all reward predictions:

$$P(\hat{k} = k|X) = \frac{e^{z_k/\tau}}{\sum_{\bar{k}} e^{z_{\bar{k}}/\tau}}. \tag{7}$$

We start with a high temperature parameter $\tau = \tau_0 = 10$, so that the learner chooses his actions only weakly influenced by the initial reward expectations. τ then decreases exponentially with learning time (with $\tau(t) = \tau_0^{\frac{\nu_\tau - t}{\nu_\tau}}$), passing 1 after a given number of steps ν_τ. At smaller values of τ the selection favors more and more the action with highest expected reward, thus exploiting the environment.

After performing the selected action $\hat{k}$, the learner receives the true reward $r(\hat{k})$. We use a reward function that is maximal if $\hat{k}$ equals the true object position k_t, decaying quadratically with increasing distance within a surrounding area (with radius ρ) and zero otherwise.

$$r(\hat{k}|X) = \max(0, (\rho + 1 - |\hat{k} - k_t|)^2) \tag{8}$$

We use gradient descent on the weights of the network to minimize the error between predicted and received reward. After each training step the new weights are:

$$v_{i,j}^{\text{new}} = v_{i,j}^{\text{old}} + \Delta v_{i,j} \tag{9}$$

$$w_{j,\hat{k}}^{\text{new}} = w_{j,\hat{k}}^{\text{old}} + \Delta w_{j,\hat{k}}, \tag{10}$$

with

$$\Delta v_{i,j} = -\varepsilon(r_{\hat{k}} - z_{\hat{k}})(-w_{j,\hat{k}}) y_j (1 - y_j) \sum_i v_{i,j} x_i, \tag{11}$$

$$\Delta w_{j,\hat{k}} = -\varepsilon(r_{\hat{k}} - z_{\hat{k}})(-y_j) \tag{12}$$

for all i and j. Note that we can only update the output weights connected to the winning output unit $\hat{k}$, since its action is the only one for which we get a true reward. ε is an exponentially decreasing learning rate: $\varepsilon(t) = 10^{\log(\varepsilon_0) - \frac{t}{\nu_\varepsilon}}$, with $\varepsilon_0 = 0.05$.

The computation of a single combined estimate requires only a single propagation through the network, which means

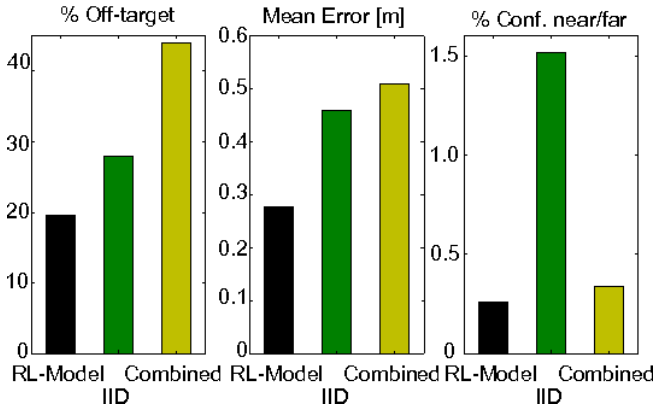

Fig. 6. Audio distance estimation: The plots show from left to right the percentage of errors, the mean estimation error in meters and the fraction of near/far confusions.

two matrix multiplications and calculating the sigmoidal function of the hidden neurons. This is fast and can therefore be considered a useful approximation of optimal Bayesian inference.

V. RESULTS

A. Audio domain

One training sample consists of the depth estimate of each of the six cues for the given auditory signal encoded into a binary vector of length 6x9 (#cues x #depth). We set $\rho = 3$ and $\nu_\varepsilon = \nu_\tau = 10,000$ and use half of the stimuli for training, the other half for testing. Figure 6 shows the results on the test sounds after 10,000 training steps compared to the performance of the best cue (IID) as well as of the model averaging approach (*Combined*) discussed in section II. The plots show from left to right the percentage of correct estimations, the mean estimation error in meters and the fraction of near/far confusions (errors of more than 4 m).

For all measures the model is better than the best single cue and better/equal the weighted average of all cues. One reason for that can be seen when looking at the spatial change of the mean error for each single cue (see Fig 3). Some cues are for example very accurate at short depth but performance decreases with increasing depth. From that one can easily predict that a single set of weights can not lead to an optimal integration at all depths. The neural network instead can easily learn to integrate the cues differently depending on the input pattern and thus performs almost equally well at all distances (Fig. 7).

B. Visual domain

For the visual depth estimation task we use a very similar setup, but the depth estimates of the single cues are continuous so that each neuron will now be active for inputs within a certain range. The same is true for the output units, the integrated estimate can only be as accurate as allowed by the binning size. It is worth mentioning that the coding range of the input and output neurons does not have to be the same, but for simplicity we set both to $10mm$ for all results shown here. We chose $\rho = 15$ and $\nu_\varepsilon = \nu_\tau = 50,000$ and

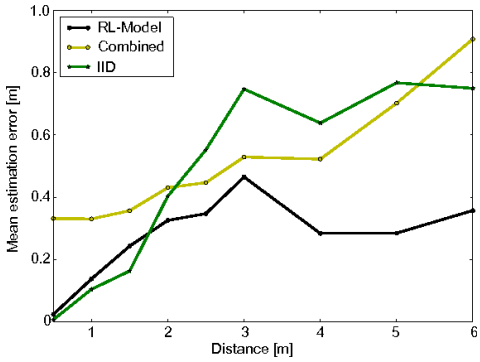

Fig. 7. Mean error for different depths of the weighted averaging (yellow) and the reward-based learning approach (black) for the auditory task in comparison with the best single cue (green).

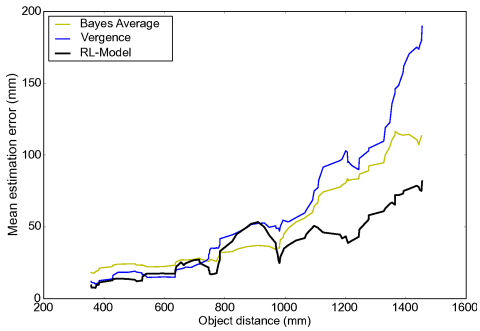

Fig. 8. Mean error for different depths of the reward-based learning approach, the Bayesian average and the best single cue (vergence) for the visual task. Mean values of 10 repetitions of each leave-one-out training trial. Plot shows the running average over 8 data points.

randomly selected five of the objects for training, the rest for testing.

As can be seen from Fig. 5 and Table II the three cues change in quality relative to each other, similar to the cues in the auditory task. It has been explained that estimation errors from individual methods can be reduced by improving the accuracy of the visual system however, trends will stay the same [9]. Therefore, it will not affect the cue integration process. The error after weighted averaging of multiple cues still has a strong tendency to increase with distance (Fig. 8 yellow curve). This is an evidence for the hypothesis that the cues show a depth dependent bias. Finally we can also expect correlations in the noise distribution of different cues, since for example both stereo disparity and familiar size use the same segmentation method. Figure 8 plots the performance of the neural network after 100,000 training steps as a function of depth. Again we get an error smaller or equal to both best cue estimate and weighted averaging, with a much lower increase with depth.

This can also be seen if we separately compute the errors for near, middle and far distance to compare it with the results shown in Table II. We find mean values of 16.6, 34.6, and 60.5 respectively.

VI. SUMMARY AND OUTLOOK

Different cues can be used to achieve auditory and visual depth estimation which are important aspects of scene perception in robotics applications. By collecting real world data for these tasks together with the corresponding responses of all these cues we were able to compare performance of single cue estimators with that after integration of multiple information sources. The common approach for such an integration is using methods from the Bayesian framework, usually combining cues by a reliability-weighted average [5]. Unfortunately this approximation is only shown to be optimal for specific environmental statistics. These requirements are often not met in real world applications, as can be exemplarily shown by the two datasets presented in this work. Therefore we proposed an alternative method which learns to combine multiple cue estimates mediated by a reward signal. We could show that this approach outperforms both the best single cue and the weighted average cue combination in both tasks. Additionally the input space dependent fluctuations in both these methods could be significantly reduced.

The reward signal we used in this paper is computed based on the true depth, which is known from the way the data was generated. For the method to be generally applicable, it would be beneficial to use reward signals that do not require hand-labeled stimuli. The quality of an action though could be measured in many different ways. One example for the depth estimation tasks would be the success of a grasping movement based on these estimations. If such an online reward signal is available, the model could adapt the optimal cue integration even during the operation of the robot assuming that tasks are executed frequently. In [7] we also demonstrated that the re-adaptation to new cue reliabilities can be done very quickly.

Training the model takes many iterations, but after completion each estimation can be computed very efficiently. Input cues are in no way limited to the ones presented here but could even act in different coordinate systems. As we could show in a previous paper [7] the model can also learn to only integrate stimuli that have a common cause but not those that come from different objects. This is particularly interesting in natural environments were there are usually multiple events happening in parallel.

VII. ACKNOWLEDGMENTS

This work was supported by the Honda Research Institute Europe. Authors T.W. and C.R. were supported by the German Federal Ministry of Education and Research (BMBF) within the Bernstein Focus: Neurotechnology through research grant 01GQ0840.

REFERENCES

[1] E. Hayman and J.-O. Eklundh, "Probabilistic and Voting Approaches to Cue Integration for Figure-Ground Segmentation," in *7th European Conference on Computer Vision (ECCV 2002)*, vol. 2352 of *Lecture Notes in Computer Science*, (Copenhagen, Denmark), pp. 469–486, Springer, Apr. 2002.

[2] S. Khan and M. Shah, "Object based segmentation of video using color, motion and spatial information," in *2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2001)*, (Kauai, USA), pp. II–746 – II–751, IEEE Computer Society, 2001.

[3] J. Triesch and C. Eckes, "Object Recognition with Multiple Feature Types," in *8th International Conference on Artificial Neural Networks (ICANN'98)*, (Skövde, Sweden), pp. 233–238, 1998.

[4] J. Triesch and C. von der Malsburg, "Democratic integration: Self-organized integration of adaptive cues," *Neural Computation*, vol. 13, no. 9, pp. 2049–2074, 2001.

[5] D. C. Knill and W. Richards, *Perception as Bayesian inference*. New York, New York, USA: Cambridge University Press, 1996.

[6] T. H. Weisswange, C. A. Rothkopf, T. Rodemann, and J. Triesch, "Can reinforcement learning explain the development of causal inference in multisensory integration?," in *8th International Conference on Development and Learning*, pp. 1–7, IEEE, June 2009.

[7] T. H. Weisswange, C. A. Rothkopf, T. Rodemann, and J. Triesch, "Bayesian cue integration as a developmental outcome of reward mediated learning," *PloS ONE*, submitted, 2011.

[8] T. Rodemann, "A study on distance estimation in binaural sound localization," in *Proceedings of the IEEE/RSJ conference on Intelligent Robots and Systems (IROS)*, 2010.

[9] C. Karaoguz, A. Dankers, T. Rodemann, and M. Dunn, "An analysis of depth estimation within interaction range," in *IEEE-RSJ International Conference on Intelligent Robot and Systems (IROS 2010)*, IEEE Press, 2010.

[10] T. Rodemann, G. Ince, F. Joublin, and C. Goerick, "Using binaural and spectral cues for azimuth and elevation localization," in *IEEE-RSJ International Conference on Intelligent Robot and Systems (IROS 2008)*, pp. 2185–2190, IEEE, 2008.

[11] E. Berglund and J. Sitte, "Sound source localisation through active audition," in *Proc. Int. Conf. Intelligent Robots and Systems (IROS) '05*, (Edmonton, Canada), pp. 509–514, 2005.

[12] Y. Sasaki, S. Kagami, and H. Mizoguchi, "Multiple sound source mapping for a mobile robot by self-motion triangulation," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 380–385, 2006.

[13] K. Nakadai, H. Nakajima, M. Murase, H. Okuno, Y. Hasegawa, and H. Tsujino, "Real-time tracking of multiple sound sources by integration of in-room and robot-embedded microphone arrays," in *Proceedings of the International Conference on Intelligent Robots & Systems (IROS)*, IEEE, 2006.

[14] M. Naguib and H. Wiley, "Estimating the distance to a source of sound: mechanisms and adaptations for long-range communication," *Animal Behaviour*, vol. 62, pp. 825–837, 2001.

[15] B. S. Nelson and P. K. Stoddard, "Accuracy of auditory distance and azimuth perception by a passerine bird in natural habitat," *Animal Behaviour*, vol. 56, pp. 467–477, 1998.

[16] A. Dankers, N. Barnes, and A. Zelinsky, "Active vision - rectification and depth mapping," in *Australasian Conf. on Robotics and Automation*, 2004.

[17] G. Bradski and A. Kaehler, *Learning OpenCV: Computer Vision with the OpenCV Library*. O'Reilly Media, Inc., 1st ed., October 2008.

[18] C. Zhang, V. Willert, and J. Eggert, "Tracking with depth-from-size," in *Neural Information Processing, 15th Int. Conference*, pp. 274–283, 2009.

[19] A. Gasteratos, R. Martinotti, G. Metta, and G. Sandini, "Precise 3d measurements with a high resolution stereo head," in *Image and Signal Processing and Analysis, First Int. Workshop*, pp. 171–176, 2000.

[20] M. Z. Brown, D. Burschka, and G. D. Hager, "Advances in computational stereo," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 8, pp. 993–1008, 2003.

[21] D. Scharstein, R. Szeliski, and R. Zabih, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *Int. J. Comput. Vision*, vol. 47, 2002.

[22] C. Goerick, J. Schmuedderich, B. Bolder, H. Janssen, M. Gienger, A. Bendig, M. Heckmann, T. Rodemann, H. Brandl, X. Domont, and I. Mikhailova, "Interactive online multimodal association for internal concept building in humanoids," in *IEEE-RAS International Conference on Humanoids 2009*, IEEE, 2009.

[23] S. Kirstein, H. Wersing, and E. Koerner, "A biologically motivated visual memory architecture for online learning of objects," *Neural Networks*, vol. 21, no. 1, pp. 65–77, 2008.

[24] 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, pp. 694–698, May 2008.
[25] M. Nardini, P. Jones, R. Bedford, and O. Braddick, "Development of cue integration in human navigation.," *Current Biology*, vol. 18, pp. 689–693, May 2008.
[26] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, UK: MIT Press, 1998.
[27] K. Körding, "Decision theory: what "should" the nervous system do?," *Science*, vol. 318, pp. 606–610, Oct. 2007.