Usage of General Developmental Principles for Adaptation of Reactive Behavior

Inna Mikhailova **Werner von Seelen** **Christian Goerick**
Honda Research Institute Europe GmbH,
Carl-Legien-Strasse 30,
63073 Offenbach / Main, Germany
inna.mikhailova@honda-ri.de

Abstract

We propose a design of a self-motivational system which would be suitable for both adaptation and development scenarios. The core features of our design are: the usage of general rewards and building up on a reactive system. As a test case for our design we use the human-robot interaction via shared attention. We show that with addition of a general-purpose self-motivational mechanism the system can aquire a behavior for distinguishing between objects presented by a human and those that just happen to be close to the robot.

1 Introduction

The role of developmental robotics is two-fold. On the one hand it strives for the implementation of general principles of child development and the validation of psychological theories on robots. On the other hand it aims at increasing the autonomy and adaptivity of robots. However, it remains difficult to combine both issues in a really synergetic fashion. The general developmental principles are applied mostly to open-ended scenarios but less to scenarios where a robot has to learn a particular skill for a particular task. In this article we discuss the question if and how a general intrinsic developmental algorithm can be used for adaptation towards specific strategies.

We discuss first the difference between specific and unspecific rewards. Subsequently we argue that exploration should not be an independent module, but a parameter of the behavior in order to allow for adaptation.

Finally the theoretical ideas are tested in an experiment on a robot head which can move in pan and tilt directions and is equipped with stereo-vision cameras. The context of the experiment is human-robot interaction for object recognition and learning. We start with reactive gaze selection used in (Goerick et al., 2005) as innate behavior of the robot. The extension of the reactive control by a self-motivation system allows for the adaptation of the innate behavior.

2 Specific Versus Unspecific Reward

The developmental approach supposes that the intrinsic developmental system consists at least of a self-motivation or value system, an abstraction system and an anticipation system, as well as innate behaviors, see e.g. (Lungarella et al., 2003), (Blank et al., 2005). The design of these parts constraints what the system can learn and what the resulting behavior will be.

In the literature several proposals for the design of reward for motivation have been made varying from very sensor-close and specific, e.g. red objects, (Sporns and Alexander, 2002), to very sensor-far and unspecific, e.g. learning progress, (Oudeyer and Kaplan, 2006), novelty and predictability, (Marshall et al., 2004).

The specific rewards are easy to implement, but if they are employed for signal-symbol mapping then they can lead to symbol grounding problems. For example in (Sawada et al., 2004) a "social reward" is given if the human comes closer to the robot. However, in a natural environment the human could come closer to the robot because the human is angry and not because he wants to reward the robot.

An additional issue of specific reward is its locality. It rewards the end-point without rewarding the way to it. Thus it can not help to find the strategy. For example associating red objects with reward does not help to grasp a red object from an inaccessible position.

In contrast, unspecific rewards cover large parts of the behavior space. These rewards can provide the evaluation of unknown situations not forseen by the designer. Such an evaluation can considerably speed up the exploration compared to a pure random search. In this paper we use some simple heuristics for deciding on the direction of the exploration. An alternative would be to let the exploration strategy emerge from the progress evaluation

(Oudeyer and Kaplan, 2006). The results of this approach confirm that specific behaviors can emerge from unspecific reward.

Except for qualitative reviews of approaches in developmental robotics, e.g. (Lungarella et al., 2003), there exists yet no well-established methodology for comparing different designs. We propose a simple empirical consideration which can help to design a motivational system.

It makes no difference whether one pre-designs some reactive, innate, task-specific solution or a self-motivation system if the reward chosen for motivation occurs only in the situations which correspond to this very task-specific solution.

In other words: through the introduction of the self-motivation system one can gain more adaptivity only if the reward used in this system covers a larger sub-space in the behavior space than the subspace covered by the reactive behavior.

From the point of view of adaptation it means that the system can adapt if it can go back to the general evaluation (general description of "good" and "bad") once the specific evaluation (specific strategy) turns out to be wrong.

Consequently for our goal of increasing adaptivity of a reactive system unspecific, general, grounded rewards are a better design choice. This is validated by an experiment described in section 4. It does not necessarily mean that there is no need for specific rewards, it only means that they are probably playing a different role, e.g. for conditioning.

The choice of reward signals or the evaluation metric is one important design issue. Another important issue is the design of interfaces from the self-motivation or value system to the rest of the system.

3 Integration of Self-Motivation into the Reactive Behavior System

The architecture of an autonomous robot can use self-motivation on different levels: on the level of emotional evaluation, goal selection, or action selection. Following the developmental approach, we start with the minimal architecture without an explicit representation of goals. We are interested in the question how self-motivation can be used for the exploration of new behaviors on a lower level, because the adaptation also presupposes an exploration strategy for finding an appropriate behavior.

The exploration is often decoupled from the rest of the system both from architectural and functional points of view. Sometimes it is considered as a default behavior and is implemented as an isolated module, e.g. (Sawada et al., 2004). Sometimes it acts on the level of the action selection, but independently, e.g. (Oudeyer and Kaplan, 2006), (Singh et al., 2004).

We propose that the amount of exploration is a parameter of the system which describes the possible deviation from the known behavior. The advantage of this approach is the natural transition from task execution to exploration. Both in case of failure and in case of absence of the specific task we gradually increase the allowed deviation from the known behavior. This means that we can keep the stable reactive architecture, do the monitoring in parallel and do the exploration by influencing the main sensor-actuator loop. As we start without goal representations, it is the self-motivational system that controls the allowed deviation from the known behavior. If the needs/motivations exceed the "viability ranges" a higher deviation is allowed.

This approach requires that behaviors and behavior strategies are described by continuous parameters with accessible neighborhoods. In this case we have a second advantage: the exploration can profit from the known behaviors. Think for example about discovering an asking gesture in the proximity of a pointing gesture as proposed in (Kaplan and Hafner, 2004). The system needs to learn primarily to recognize the right context, to slightly modify pointing and to combine it with vocalization. In contrast, the random rediscovery of coordination between hand, object and vocalization is very costly.

4 Experiment: Discover Appropriate Interaction Behavior

The goal in the following experiment is to establish a robust human-robot interaction for online learning of object recognition. The system should make a visual search between all interesting (regarding color, contrast or structure) locations, but fixate the location of an object shown to the system by a user.

4.1 Innate Behavior: Reactive Gaze Selection and Peripersonal Space

In (Goerick et al., 2005) this task was solved with the help of reactive behaviors and the concept of peripersonal space, which is defined as a particular volume in front of the body. The objects inside of this volume are perceived differently than distant objects. In this way shared attention between the system and the human can be created on a very low level without any psychological concepts.

The system selects its gaze direction according to a saliency map in the spirit of (Itti et al., 1998). This map is a weighted sum of visual saliency, disparity saliency selection, and motion saliency selection as illustrated by the upper part of Figure 2. The visual saliency computation provides a map of "interesting" (regarding color, contrast, or structure) locations. The disparity saliency selection computes

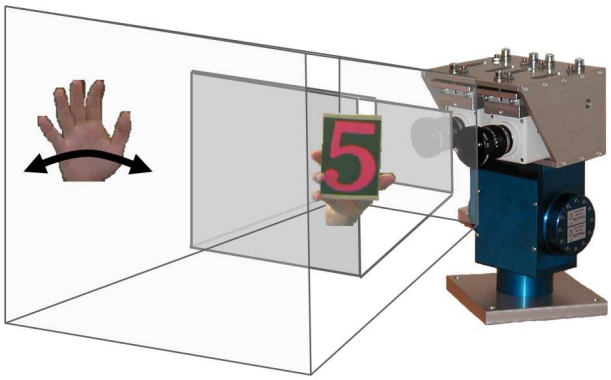

Figure 1: Schematic visualization of the peripersonal space approximation. The inner volume represents the peripersonal space, the outer volume the complete field of view with the sensitivity to visual and motion cues.

disparities and selects the closest region within a specific distance range and angle of view. This simple mechanism represents a first approximation to the concept of the peripersonal space (see Figure 1). The motion saliency selection produces a map with an activation corresponding to the largest connected area of motion within a defined time-span.

The weights of maps define the behavior of the system. In (Goerick et al., 2005) the weights (W_V for visual saliency, W_D for disparity, and W_M for Motion) are set according to the information priority (from highest to lowest): disparity, motion, visual saliency ($W_D = 3.0$, $W_M = 2.0$, $W_V = 1.0$). Without any interaction the gaze selection is autonomously driven by the visual saliency and the memory of the gaze selection. A natural way for humans to raise the attention is to step into the field of view and wave at the system. Due to the chosen weights the system will immediately gaze in the direction of the detected motion. The motion cue can be used continously in order to keep the gaze direction of the system oriented towards the hand until the hand enters the peripersonal space. Again, due to the chosen weights the signal from the peripersonal space will dominate the behavior of the system. This means that the system will continously fixate the hand and what is in the hand of the user. Finally the object recognition learns whatever is shown to the system in this way.

This pre-designed solution provides very robust and natural means of interacting with a robot. However it has a small drawback: if an object is not presented by a human, but is close to the system, then it will also be fixated. This can be interpreted as a symbol grounding problem. The mapping from the depth signal to the interaction hypothesis is created by the designer. The reality does not always correspond to this mapping but the system can not find out the discrepancy on its own. Nothing would change if instead of the reactive system we would take a self-motivated system which gets rewards from the described sensory maps with the depth as the highest reward. This would be an example of a too specific reward which does not help adaptation as discussed in section 2. We propose instead to use "consistency" and "activity" as rewards. The next section describes how these rewards can be used for self-motivation.

4.2 Extension of the Reactive System by a Self-Motivation System

Our implementation of the self-motivation system follows ideas of homeostatic regulation presented in (Cos-Aguilera et al., 2003). The difference is that we use unspecific rewards instead of specific ones and that we use slightly different dynamics.

The self-motivation system consists of two needs $N_i, i \in \{1, 2\}$. The needs are satisfied if their values are close to zero. If the needs are below a chosen threshold $N_0 > 0$ they are set to this threshold. Otherwise they change according to dynamics of the Lotka-Volterra type:

$$\begin{cases} \tau_1 dN_1/dt = N_1(t)\,(R_{01} - R_1(t) - W_1 * N_2(t)) \\ \tau_2 dN_2/dt = N_2(t)\,(R_{02} - R_2(t) - W_2 * N_1(t)) \end{cases}$$

where τ_i are time constants, R_{0i} characterize the speed of the need growth in absence of rewards, W_i are the coupling weights between the different needs, and $R_i(t)$ are the corresponding rewards. The coupling weights are positive so that a high value of one need prevents the growth of the others.

The needs in our experiment are: a need of "consistency" and a need of "activity". In general the "consistency" reward measures the quality of interaction with the environment. It measures if the action of the robot leads to consistent sensory observations. In our example the "consistency" reward is directly measured as the correlation between the gaze direction and the entries in the sensory maps described in the last section. The correlation is calculated as a total sum of a element-wise product of the sensory maps and the gaze selection map accumulated over time. For normalization we use the maximum of the total sum of elements in the sensory map and total sum of the elements in the gaze selection map.

While the "consistency" reward characterizes if interaction with the environment is favorable for learning, the "activity" reward measures if the system acts. In our experiment "activity" reward is derived from the difference between the new and the old gaze direction. For this purpose we calculate the total sum of the absolute values of the difference between two successive gaze selection maps.

Similar concepts of opposite motivations were proposed in psychology and robotics for a long time,

e.g. adaptation/expansion, safety/curiosity, or predictivity/novelty desires. We have chosen an implementation that stays as close as possible to the sensors in order not to introduce too much design at a too early stage.

4.3 Exploration of Possible Behaviors

The space of possible behaviors of our system is spanned by the 3 weights $\vec{w} = (W_D, W_M, W_V)$ which couple sensory maps to the gaze selection as described in section 4.1 and Figure 2. We add a simple monitoring to the reactive system. The observed combination of reward and weights are stored into the table at index $i \in [0 \dots M]$ in the form $[\vec{w}_f^i, \vec{R}(\vec{w}_f^i)]$. The entries in the table characterize the situation comprising the reward and the active behavior. A new entry to the table is added whenever the distance from the observed situation to the entries of the table is larger than a threshold T_d : $\forall i \in [0 \dots M]\,,\ ||\vec{w}_f^i - \vec{w}|| + ||\vec{R}(\vec{w}_f^i) - \vec{R}|| > T_d$. The best known weight from the table is defined by

$$\vec{w}_f = \arg\max_i(R^n(\vec{w}_f^i))\,,$$

with n as an index of the most urgent need: $n = \arg\max_i(N_i - N_0)$. The table provides a primitive quantization of the behavior space for exploration only, not for action selection. The action selection needs a more robust quantization, for example with help of self-organizing maps as described in next section. At the actual stage of research we don't use any action selection algorithm other than the exploration described bellow.

The continuous exploration of the behavior space is defined by the strength S_e and direction $\vec{d}_e$. Starting from weight $\vec{w}_s$ the system tries out the weight $\vec{w}_e = \vec{w}_s + S_e\vec{d}_e$. At the start of the exploration the direction is chosen either randomly or towards the weight predicting the best reward if such information is available from the table of known behaviors: $\vec{d}_e = \vec{w}_f - \vec{w}_l$, were the $\vec{w}_f$ is the best known weight as described above and $\vec{w}_l$ is the last used weight.The direction is kept for a while in order to have a hysteresis. If the need continues to grow then the direction is changed to the opposite. The starting point is the best known weight at the beginning of the exploration and the last used weight during exploration.

The strength S_e is controlled by two factors: by the allowed deviation Δ_e from the known behavior and by the time interval ΔT since the last change of the exploration: $S_e = \Delta_e(1.0 - \exp(-\tau_e \Delta T))$, where the τ_e defines the speed of the exploration. The allowed deviation is increased if the needs described in the last section leave the "viability range": $\Delta_e \sim \sum_{i=1,2}(N_i - N_0)^2$.

In the next section we show how through the exploration the system can discover an "avoiding" behavior with a negative weight of the disparity channel: (W_D = -2.0, $W_M = 2.0$, $W_V = 1.0$). With these weights the system turns away from the near object. If the object is just "background", then it does not react and there is no correlation between the action of the system and the sensory map. If the object is shown by a user, then it is natural for the user to slightly follow the head movement of the robot in order to stay in interaction. The "consistency" reward is thus provided only in interaction with the user and the system can discriminate the disparity signal induced by a user from the disparity signal induced by a static "background" object.

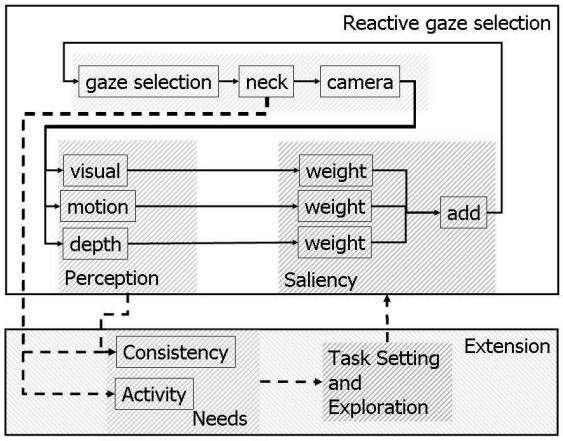

Figure 2: The extension of the gaze selection by self-motivation. The innate behavior of the system is reactive gaze selection driven by a weighted sum of visual, depth and motion saliency maps. This innate behavior is extended by exploration of the weights controlled by homeostatic regulation of "consistency" and "activity" needs.

4.4 Experimental Results

Figure 3 shows a typical run of our experiment. The parameters are set as follows: $\tau_1 = \tau_2 = 0.05$, $R_{01} = 0.8$, $R_{02} = 0.3$, $W_1 = W_2 = 0.1$, $\tau_e = 0.2$. The needs are initially at the lowest level $N_0 = 0.05$. We start with the zero disparity weight. Thus the gaze direction of the system is guided by the visual saliency only. At step 100 the user enters an object into the peripersonal space. The user follows slightly with his object the gaze direction of the robot, thus creating a consistency reward. The system explores the disparity weight into the positive direction. This direction was chosen randomly, as the system has no particular knowledge about existing rewards. Approximatively at step 150 the weight is sufficiently high for fixating the object. As long as the user moves the object the system performs the tracking. At this point the system has the same behavior as the reactive setup in (Goerick et al., 2005). At about

step 220 the user stops the interaction and at step 280 an object is put on the table so that it enters into the peripersonal space statically. Due to the high disparity weight the system keeps fixating this background object. At the step 379 due to the internal dynamics described in section 4.2 the need of activity is going out of the viability range and the system changes the exploration in order to fulfill this need. At the start of the experiment the system observed that for the zero disparity weight W_D there was a high reward in the activity. Thus the exploration goes into the direction of the zero weight. The negative exploration direction is first kept by hysteresis and later by the fact that the activity need is decreasing. At steps 500-600 the user comes back. He takes the former "background" object and follows with the object the escaping gaze direction of the head signaling the start of an interaction. The system observes that in the case of an "escaping" behavior it is also possible to get the consistency reward.

Figure 4 shows another run with different initial conditions: disparity weight $W_D = 4.0$ and needs $N_1(0) = N_2(0) = 1.0$. The exploration (steps 200-350) changes the direction several times. In the previous example it was not the case because there at the start of the exploration the system already uses the information from the table of know behaviors and explores in the direction of $W_D = 0.0$ with the high activity reward.

For the off-line analysis the data observed during the run are mapped with a self-organizing map (SOM) presented in Figure 5. Results are qualitatively the same for both experiments. It can be seen that the SOM finds well separated clusters both in observable behavior and in the input space. Our future research intends to use these clusters for the behavior quantization. For example we can see from the SOM that the "tracking" behavior with a positive disparity weight leads to rewards only in the context where both motion and disparity are present. These two signals are better at predicting the existence of a human user than the disparity alone. But the user can also show an object to the system without moving it. Thus we have to learn the strategy of combining "avoiding" and "tracking" to check the responsiveness of the environment. Then the system can discriminate the background activity from a response of the environment to the system's action.

5 Discussion

With the "consistency" reward proposed in this article we aim at putting the system into situations where it can learn from its interaction with the environment. Our measurement of interaction quality is implemented via a simple action/sensor correlation in space. This works well for our type of interaction via the peripersonal space. For vocal interaction one would need the measurement of responsiveness of the environment not over space but over time.

A more elaborate measurement of the quality of system/environment interaction is proposed in (Oudeyer and Kaplan, 2006). There the reward to the system is proportional to the progress of learning to predict the sensory input. Such a reward system explains well the dynamics of the development and transitions from simple to more complex situations. In our work we investigated if a similar rewarding is appropriate for adaptation. Our reward system is independent of the chosen learning algorithm. Because we do not use reinforcement learning we are not forced to build only one value function. We work with multiple rewards which are closer to the sensory input than the measure of progress in learning, but are still unspecific. (Oudeyer and Kaplan, 2006) also argue against the specific social reward. While our argumentation is based on the general difference between unspecific and specific reward for seeking adaptation, the progress drive hypothesis aims at explaining the communication development.

In (Oudeyer and Kaplan, 2006) the response from the communication partner is seen as a communication affordance. However the responsiveness depends on the context and the state of the partner, which are not known a-priori. Instead the system needs an active strategy to detect the responsiveness of the partner. An example of such a strategy is developed in (Movellan, 2005) for vocalization on the base of a pre-designed model for timing of self- and other-responses. Our approach also starts with a predesigned model of interaction captured in reactive behaviors and the concept of the peripersonal space. However with the help of self-motivation the system can discover the discrepancy between the model and reality and find the escaping-behavior for active checking of the responsiveness of the environment. The strategy to use both escaping and tracking (or asking and responding in terms of vocal interaction) is subject to our future research.

Generally we believe that development is not restricted to refinement of internal representations of contexts and behaviors. Additionally the developing system creates means to actively check the content of the representations. We propose that the checking should use the most basic perceptual elements like time-space correlation or causality. The grounding is then provided not only through the self-creation of representation, but also by the possibility to use the physical constraints to define the content and to test the actual instance. In this work we gave an example of monitoring the existence of responsive partner. In human daily life one can find many more examples: shifting an object before grasping (checking out that the object is separable from the supporting plane), greeting a person, before asking a question (checking

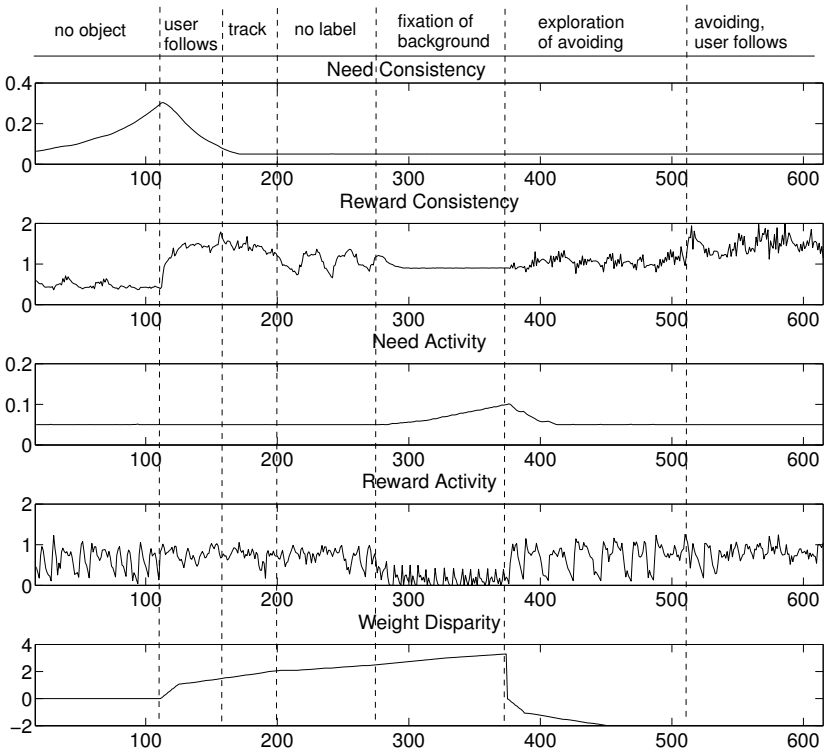

Figure 3: The experimental run: exploration of disparity weight. The duration of one time step varies from 300ms to 1300ms dependent on the amplitude of neck movement. The dashed vertical lines approximatively show the phases of different behavior as judged by the observer. Key time-steps are: 100 - user enters an object into the peripersonal space; 220 - user removes the object; 280 - an object is put on the table; 500 - user restarts the interaction. For more details see the description in the text.

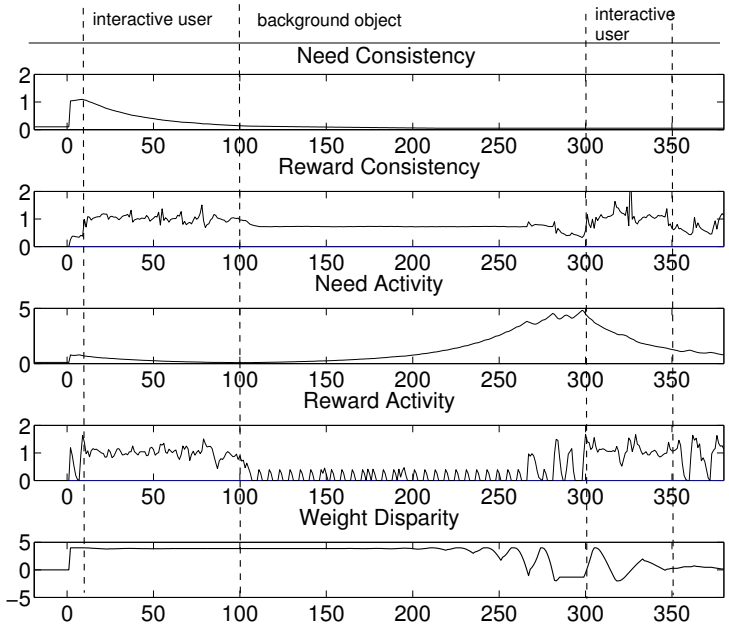

Figure 4: The experimental run: exploration of disparity weight starting from fixation setting $W_D = 4.0$.

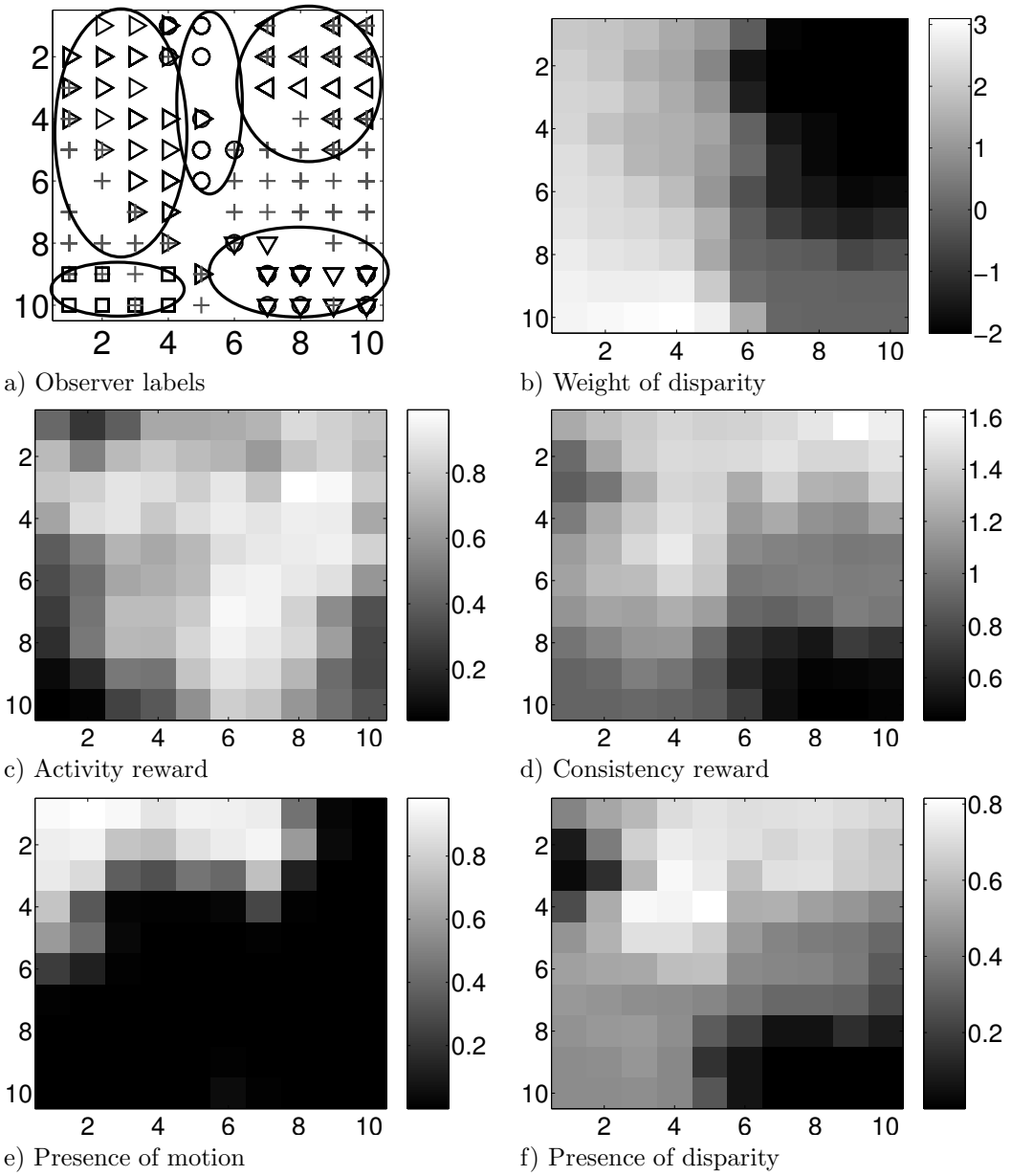

Figure 5: The clustering of behaviors with the help of a self-organizing map. The indices correspond to the indices of the two dimensional SOM. a) The labeling of node activation of the SOM during an experiment as manually assigned by human observer. Down-triangle: search driven by visual saliency only; circle: user follows with the object the movement of the head; right-triangle: tracking of the object; square: fixation of the not moving object; left-triangle: avoiding of the close object; cross: behavior not labeled by the user. The only markers which overlap with others are the cross (not-labeled) and the circle (ambiguous labeling). Thus the SOM builds well separated clusters. b)-f) The topology created by the SOM in the 5-D input space (Activity reward; Consistency reward; Weight of disparity; Presence of disparity; Presence of motion). b) The SOM preserves the topology by placing the weight of disparity channel along the x axis. e),f) The motion and disparity are very well clustered. The overlap of motion and disparity gives the best guess about the presence of the interacting user. This region corresponds to the high consistency and activity rewards in c),d).

out if the person can hear), and so on.

6 Conclusions

The goal of our work is to provide the system with the means to adapt its reactive behavior and to discriminate the cases of background activity from the activity coming from the interaction with the environment. Our system can show adaptation because the implementation follows the principles of the developmental approach:

- starting with a robust interaction with the environment as innate behavior,
- usage of environment compliance (here: user's adaptation),
- continuous behavior parameterization on the lowest level, and
- self-motivation with task-unspecific rewards.

The chosen experiment is a simple one. Still it shows that the developmental approach achieves more than simply learning behaviors which could be efficiently pre-designed in a reactive system. The core point is a careful design of a self-motivation system, its integration into the reactive system and its rewarding signals. We do not lose the advantages of reactive systems (high speed, robustness) while introducing the "reflective" self-motivation if the latter interferes only in the situations where the fast, known solution did not work or the system was free to play. In return we gain an advantage from the self-motivational part: if it operates with general, grounded needs then it can help the reactive part to discriminate in the cases of symbol-grounding problems.

Finally, we would like to emphasize that the type of context discrimination tackled in our experiment is not restricted to the social context. In learning generally one needs to check if sensory changes are related to one's action or if it is a background activity. Our future research will consider the usage of the self-motivational system presented here for learning of object manipulations.

Acknowledgments

The authors would like to thank all their collaborators at the Honda Research Institute Europe who contributed to the creation of the gaze selection system used in the experiment.

References

Blank, D., Kumar, D., Meeden, L., and Marshall, J. (2005). Bringing up robot: Fundamental mechanisms for creating a self-motivating, self-organizing architecture. *Cybernetics and Systems*, 36(2).

Cos-Aguilera, I., Cañamero, L., and G., H. (2003). Motivation-driven learning of object affordances: First experiments using a simulated khepera robot. *The Logic of Cognitive Systems: Proc. 5th Intl. Conference on Cognitive Modeling (ICCM'03), Bamberg, Germany*.

Goerick, C., Wersing, H., Mikhailova, I., and Dunn, M. (2005). Peripersonal space and object recognition for humanoids. In *Proceedings of the IEEE/RSJ International Conference on Humanoid Robots (Humanois 2005), Tsukuba, Japan*.

Itti, L., Koch, C., and Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259.

Kaplan, F. and Hafner, V. (2004). The challenges of joint attention. In Berthouze, L., Kozima, H., Prince, C., Sandini, G., Stojanov, G., Metta, G., and Balkenius, C., (Eds.), *Proceedings of the 4th International Workshop on Epigenetic Robotics: Modeling Cognitive Development in Robotic System*, pages 67–74. Lund University Cognitive Studies 117.

Lungarella, M., Metta, G., Pfeifer, R., and Sandini, G. (2003). Developmental robotics: A survey. *Connection Science*, 15(4):151–190.

Marshall, J., Blank, D., and Meeden, L. (2004). An emergent framework for self-motivation in developmental robotics. In *Third International Conference on Development and Learning*.

Movellan, J. R. (2005). Infomax control as a model of real time behavior. In *MPLab Tech Report 2005-01*.

Oudeyer, P.-Y. and Kaplan, F. (2006). Discovering communication. *Connection Science*, 18(2):189–206.

Sawada, T., Takagi, T., Hoshino, Y., and Fujita, M. (2004). Learning behavior selection through interaction based on emotionally grounded symbol concept. In *IEEE-RAS/RSJ International Conference on Humanoid Robots, Los Angeles, CA, USA*.

Singh, S. P., Barto, A. G., and Chentanez, N. (2004). Intrinsically motivated reinforcement learning. In *NIPS*.

Sporns, O. and Alexander, W. H. (2002). Neuromodulation and plasticity in an autonomous robot. *Neural Netw.*, 15(4):761–774.