RETINA IMPLANT ADJUSTMENT WITH REINFORCEMENT LEARNING

M. Becker, M. Braun, R. Eckmiller

Universität Bonn, Informatik VI (Neuroinformatik)
D - 53117 Bonn, F. R. Germany
email: becker@nero.uni-bonn.de, URL: http://www.nero.uni-bonn.de

ABSTRACT

A tuning method with reinforcement learning (RL) for the Retina Encoder (RE) of a Retina Implant (RI) as a visual prosthesis for blind subjects with retinal degenerations is proposed. RE simulates retinal information processing in real time by means of spatio-temporal receptive field (RF) filters and generates electrical signals for stimulation of several hundreds of ganglion cells (GC) to regain a modest amount of vision. For each contacted GC, RE has to be optimized with regard to the patient's perception. The patient's (for the present simulated) evaluative feedback is applied here in a dialog module as a reinforcement signal to train several RL agents in a neural network learning process (see also http://www.nero.uni-bonn.de).

1. INTRODUCTION

Retina Implants for blind subjects with retinal degenerations are currently under development [2, 3]. An eye-glass mounted photosensor device transforms the visual scene into electrical signals. The signals are processed according to a suitable model of retinal information processing (Retina Encoder, RE) to bridge the defect retinal layers. Intact retinal ganglion cells (GC) are stimulated electrically with the simulation outputs to elicit visual perceptions [6]. RE consists of receptive field filters (RF), one for each GC to be stimulated. Each RF has several spatial and temporal parameters depending on the implemented model to enable simulation of different GC-types obeserved in retina physiology. The parameters are continuous variables which span the RE state space. The GC-types of the contacted cells during stimulator implantation are unknown and must be determined in a dialog with the patient. RE parameters have to be optimized for each RF with respect to the patient's perception.

Supported by Federal Ministry for Education, Science, Research, and Technology (BMBF)

We expect the patient to be able to give evaluative feedback as a perception quality measure for consecutive RE states in the iterative RE-tuning procedure. It is very appealing to use the patient's feedback as a direct reinforcement signal in a reinforcement learning (RL) [8] task to find the optimal RE state.
Unlike other RL applications we use a subjective reinforcement signal which is not determined by heuristic rules that define the goal state. Since the reinforcement signal is influenced by the patient's psychological fluctuations the RL algorithm has to face uncertainty. We investigate on- and off-policy methods in TD(λ) training to achieve sufficient control policies for successful RE adjustment.

2. RETINA MODEL AND PERCEPTION

The implemented retina model is an intermediate version towards the DSP based software RE *Mark II* [2], which is currently being developed in our group. Retinal information processing is simulated with an input-output model from systems theory [4]. Figure 1 explains the inseparable spatio-temporal model. Each RF of RE is apportioned a spatial receptive field as two-dimensional input. Input data is fed into 2 distinct filter pathways, one for the center computation and the other for the surround. Each pathway performs a spatial scalar product of the pixel data and a rotationally symmetric two dimensional Gaussian (G) with the corresponding widths σ_c, σ_s. The values of the sigmas determine the spatial extent of the RF. The resulting scalar signals then each pass a temporal low pass (LP) , for the present of first order, with (inverse) time constants λ_c, λ_s. Additionally, the surround pathway signal can be delayed (D). Since the model is implemented as a digital software simulation the delay is restricted to multiples of the RE-calculation cycle τ. The signals from both pathways converge at the mixer component. Finally, a gain factor (g) enables range adaptation and switching between on-off and off-on behavior. The resulting signal simulates the GC membrane

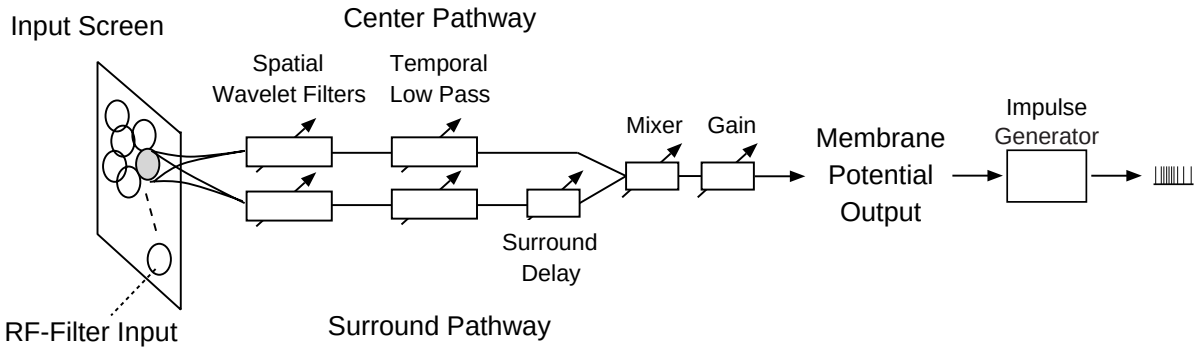

Figure 1: *Receptive field filter (RF) of RE. RF receives input data from the photosensor device. Data is processed via two distinct spatio-temporal pathways for center and surround computation. Pathways converge at the mixer-stage. The resulting signal is finally multiplied with a gain factor. Each component is individually tunable to achieve a wide range of GC-types observed in retina physiology.*

potential (MP) and is used to generate stimulation signals. Equations 1 - 4 summarize the model operators.

$$Path_c = LP(\lambda_c)\, G(\sigma_c) \quad (1)$$

$$Path_s = D(n\tau)\, LP(\lambda_s)\, G(\sigma_s) \quad (2)$$

$$Mixer = mPath_c - (1-m)Path_s \quad (3)$$

$$m \in [0,1]$$

$$MP = g \cdot Mixer \quad (4)$$

The model is able to simulate a wide range of primate ganglion cell types observed in retina physiology. There are 2 basic types of GC: sustained P-cells and transient M-cells. Cells in each of these classes show antagonistic center-surround weighting with either the center positively and the surround negatively weighted (on-off), or vice versa (off-on). Thus, RE states that actually match physiological findings will be located around 4 focal points in the RE state space: P-on, P-off, M-on, M-off. Note that states of cells in each class can deviate significantly from the corresponding focal point state. Individual RF tuning within each GC class is a separate problem not treated here. Additionally, a comprehensive RE adjustment procedure will also take care of the positions of all RFs on the input screen, since stimulation contacts are fixed on the patient's retina and properties of the contacted GC cannot be predicted or even controlled during stimulation device implantation. A procedure for raw classification of the contacted cells in terms of the four basic classes will be described elsewhere. Here, we restrict the task to find optimal parameters for the four basic classes, to which each cell in the simulation will be attached to. Stimulated GC signals are conveyed on the optic nerve to higher cortical centers. High level information processing is performed in several cortex areas and visual perceptions are elicited. A model of perception with GC signals as input and perception as output does not yet exist. Therefore error information for RE state optimization must be provided by the implant carrying patient in an iterative procedure. Objective (VEP/EEP) as well as subjective (dialog) information can be used for RE adjustment.

3. REINFORCEMENT LEARNING FORMULATION

Our current investigation applies a dialog concept with a simple approximation of perception quality to adjust RE. We use a mathematical measure (RMS) to describe the error between the current RE state and the arbitrary goal state which is chosen on initialization. Therefore we measure the RE output sequence $\vec{Curr}$ of length N_k in response to a moving light-stimulus for each of the N_c RFs of RE. The corresponding sequence $\vec{Goal}$ for the RE goal state is measured once on initialization and stored.

$$Err = \frac{1}{N_c N_k} \cdot \sqrt{\sum_c^{N_c} \sum_k^{N_k} (Goal_k^c - Curr_k^c)^2} \quad (5)$$

This quantitative error should not be incorporated completely for RE adjustment since it is only available in this approximation but not in the future dialog with implant carrying subjects. Therefore, an immediate reinforcement signal is calculated from consecutive errors. It is set zero in case of decreasing error and -1 otherwise. A dialog method in which a normal sighted person simulates the behaviour of the inplant carrying subject is currently under development[1] [2].

[1] A problem with these methods is that we cannot proof equivalence of simulated and future, actual evaluative feedback from the patient. In both cases, the response will be equivalent for

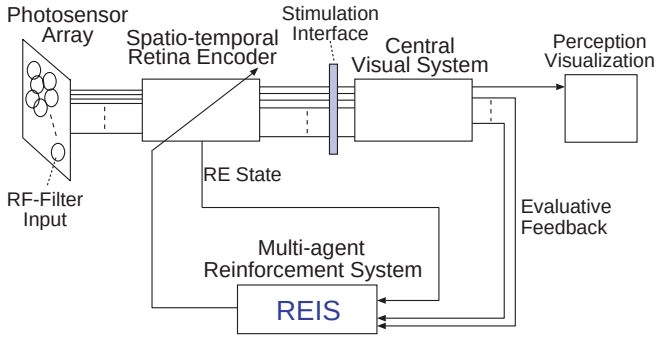

Figure 2: *Scheme of RE adjustment with RL agents as part of a dialog module using the patient's evaluative feedback.*

The RL environment comprises RE, stimulation interfaces (SI) and the patient's central visual system (CVS). RE state can be modified by altering parameters of the RFs. A complete description of the environment would include CVS and SI, but no model exists for these components, and furthermore corresponding state variables would be hardly observable. When we assume stationary SI and CVS (low CVS-plasticity) we can approximate environment situations by RE states. With this simple CVS approximation used here RE and environment states are indeed equivalent.

Figure 2 explains the setup for RE adjustment. Suitable spatio-temporal light stimuli are projected onto the (virtual) photosensor device which serves as RE input. RE output is calculated from the input according to the RE state. RE state is given by the parameters of each RF constituting RE. In the special case of finding optimal parameter sets for the four basic GC classes P-on, P-off, M-on, M-off there are four different RF parameter sets. Each RF of RE is attached to one of the four basic classes. RE output innervates the (simulated) patient's central visual system and produces visual perceptions. The virtual screen in the figure is for alternative use with normal sighted persons as mentioned above. Evaluative feedback from one of these sources is fed into a multi-agent reinforcement system (REIS) which performs actions on the RE state, given by the RF parameters of the four basic classes. Iteratively, REIS moves the RE system towards the predefined goal state.

REIS consists of one or more agents, each of which alters certain parameters in the four base classes by performing *actions (a)*. The actions shift parameters by pre-defined values, typically $\frac{1}{100}$ of the total parameter range. Two important aspects result from this quasi-discretization: It is unlikely that the goal state is exactly matchable in the adjustment procedure. This is not a severe problem since we can assume continuity and narrowness in the *mapping* from RE states to perception for neighbouring RE states. The second aspect is that we can enumerate RE states along the adjustment trajectory. With the step size given above the total number of visitable states [2] becomes 100^{28}. It is obvious that exploration will be local in this large state space. Exploration is achieved by a simple random component in choosing actions rather than stricly following the current action policy. The probability $p(a_{rand})$ of choosing a random action is directly controlled by a temperature value which can be decreased with time. Each agent has access to view the complete state of RE (s) and receives the global reinforcement signal r on each iteration. Additionally, each agent maintains an action value function approximation (Q). TD(λ) is applied to improve consecutive predictions of Q.

$$\begin{aligned} \Delta\vec{w}_t &= \alpha \cdot (r_{t+1} + \gamma Q_t(s_{t+1}, \tilde{a}) - Q_t(s_t, a_t)) \cdot \vec{e_t} \\ \vec{e_t} &= \gamma\lambda \vec{e_{t-1}} + \nabla_w Q_t(s_t, a_t) \end{aligned} \quad (6)$$

Equations 6 specify the approximators learning rule. α is a positive step size constant. γ discounts the influence of past weight alterations ($0 \leq \gamma < 1$). The agent's action policy is immediately derived from Q: Take the action a_t that maximizes Q_t and include a random component. For off-policy training, $\tilde{a}$ equals $arg\,max_a(Q_t(s_{t+1}, a))$, the action that maximizes the current action-value approximation of the next RE state (Q-learning, [5]). For on-policy training, there is an additional random component in chosing $\tilde{a}$, but in contrast to off-policy action $\tilde{a}$ is actually performed on the next iteration, $a_{t+1} = \tilde{a}$.

4. EXPERIMENTS AND RESULTS

Figure 3 shows a typical experimental situation to test REIS's ability to adjust RE. Here, RE consists of nine RFs distributed on a hexagonal grid. Each RF belongs to one the four basic classes P-on, P-off, M-on, M-off. P-cells have smaller spatial receptive field extensions than M-cells and are drawn with a smaller radius, accordingly. The vertical bar on the right denotes the light stimulus. It is moved over the screen (see arrows) with sinusoidal speed profile. RF output data are measured for a complete spatio-temporal stimulus presentation. The error signal is calculated from the

the goal state, but this is not necessarily true on the trajectory towards the goal.

[2] for comparison: Elevator scheduling 10^{22} states [1], Cellular phone channel allocation 49^{49} states [7].

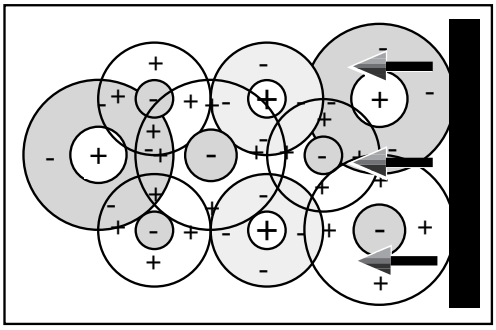

Figure 3: *Receptive fields of RE with nine hexagonally arranged RF filters. RFs are attached to one the four different GC classes P-on, P-off, M-on, M-off. Arrows indicate the moving bar used as light stimulus with the experiments.*

data according to equation 5. Comparison of current and previous error signal yields the immediate reinforcement value as discussed in section 3. The reinforcement signal is propagated to REIS's agents. One of the agents performs an action on RE according to the action policy (section 3).

Possible actions are positive or negative steps for each RF parameter of the four basic GC classes. The total number of actions becomes $7 \cdot 4 \cdot 2 = 56$ actions. In case parameter range borders are violated, REIS is punished and no state alteration is performed. Note that this knowledge is not limited to the current simple patient approximation. Also with implant carrying subjects RE parameter range violations are detectable without incorporating the patient. Step sizes are individually tunable, usually values are $\approx \frac{1}{100}$ of the normalization ranges. Figure 4 shows a typical error time course for RE adjustment. For these first experiments we applied an off-policy temporal difference method to train REIS's agents. This performance measure is available with the simple patient simulation only, and the same is true for the actual deviation in RE state between the defined goal state and the current RE state. Mean deviation is about 18% at the final step in figure 4, which has to be improved by optimizing training parameters.

5. CONCLUSION

We have shown that a modified reinforcement learning method is capable of adjusting a retina encoder as a part of a learning visual prosthesis. Our results on temporal difference learning for value-function approximation clearly show the ability of the combined learning and control architecture to move the RE state toward an arbitrarily defined goal. Further experiments are necessary to optimize reinforcement learning algorithms and patient models for more accurate evaluative feedback.

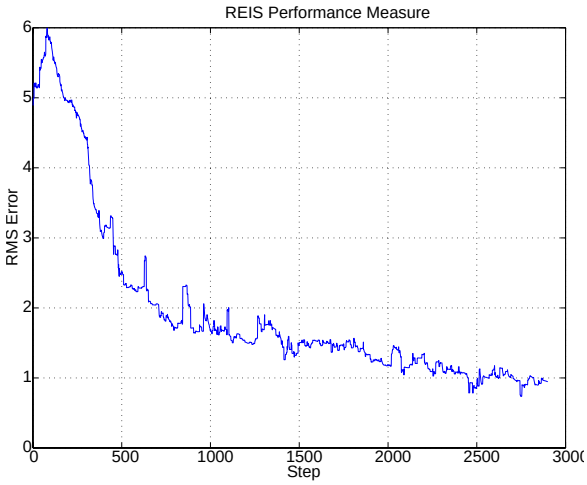

Figure 4: *Typical error time curve (RMS) for RE adjustment. Off-policy TD(λ) (Q-Learning) is applied to 56 RL agents.*

6. REFERENCES

[1] R. H. Crites, A. G. Barto, "Improving elevator performance using reinforcement learning", Adv. in Neural Inform. Proc. Sys. 8, pages 1017-1023, MIT Press, 1996

[2] R. Eckmiller, M. Becker, R. Hünermann, "Dialog concepts for learning retina encoders", Proc. ICNN'97, pages 2315-2320, Houston, Texas, June 1997

[3] Jahresbericht Retina Implant Projekt 96/97, BMBF/DLR, Oktober 1997 (in press)

[4] D. J. Fleet, P. E. Hallett, A. D. Jepson, "Spatiotemporal inseparability in early visual processing", Biol Cybern 52:153-164, 1985

[5] G. A. Rummery, M. Niranjan, "On-line Q-Learning using connectionist systems", CUED/F-INFENG/TR 166, 20 pages, Cambridge University (UK), 1994

[6] A. Santos, M. S. Humayun, E. de Juan, R. J. Greenburg, M. J. Marsh, I. B. Block, A. H. Milam, "Preservation of the inner retina in retinitis pigmentosa", Arch. Opthalmol. 115:511-515, 1997

[7] S. Singh, D. Bertsekas, "Reinforcement learning for dynamic channel allocation in cellular telephone systems", NIPS 9, Denver, Dec. 1997 (in press)

[8] R. S. Sutton, "On the significance of Markov decision processes", Proceedings of ICANN 97, pages 273-282, Springer, 1997

[9] R. Tesauro, "Practical issues in temporal difference learning", Maschine Learning 8:257-277, 1992