Proc. SOAVE 2004, Ilmenau September 2004, Fortschrittberichte VDI, Reihe 10, Nr. 743, pp. 262 - 270, VDI-Verlag, Duesseldorf, 2004

Application of Action Dependent Heuristic Dynamic Programming to Control an Industrial Waste Incineration Plant *

V. Stephan and F. Wintrich
Powitec Intelligent Technologies GmbH
Am Teelbruch 134b,
45129 Essen-Kettwig,
volker.stephan@powitec.de

A. König and K. Debes
Technische Universität Ilmenau,
Fachgebiet Neuroinformatik
98684 Ilmenau, PF 100565,
klaus.debes@tu-ilmenau.de

Abstract

In this paper, we describe our application of a neurocontroller based on Action Dependent Heuristic Dynamic Programming (ADHDP) to optimize the combustion-process for an industrial hazardous waste incineration plant. This ADHDP-controller originally was designed for online learning. That implies, that this controller starts with a randomly initialized policy and improves its performance while interacting with the process. This learning scheme could not be used in our case, since the plant operators would not allow long training periods with inevitable poor performance. We describe, how this problem can be solved by a modified training procedure for the action network. Finally, we present first and promising results for the optimization of action net output in order to improve the waste incineration process with respect to the targets defined by the plant operator.

1 Introduction

The AVG Abfall-Verwertungs-Gesellschaft mbH in Hamburg (Germany) is one of the largest companies in Germany for disposal of hazardous waste. Every year about 100000 tons of solid, pasty, liquid and waste packed in containers are treated. The incineration at temperatures higher that 1100°C is a very safe and effective technique to dispose that waste, since it reduces the quantity and also destroys harmful substances.

The center of the incineration plant is a rotary kiln with a length of 12m and a diameter of 5.4m. Solid wastes are brought into the kiln via a falling shaft (see figure 1, part 1), liquid and slurry type wastes are pumped via lances and burners, and drums are brought into the kiln via special lifts. By slowly rotating the kiln (figure 1, part 2), the transport and complete combustion of the waste material is ensured. In the secondary combustion chamber (figure 1, part 3), exhaust fumes are after-burnt at temperatures higher than 1100°C to destroy organic compounds like dioxins or furans. Afterwards, the hot flue gas is used to generate steam, which is fed to a local district heating network.

2 The Control Problem

Additionally to the continuously incoming liquid and pasty waste, from time to time also solid material from the waste pit or packages in form of barrels filled with waste are fed into the kiln. These

*sponsored by AIF, Project-Nr.: KF 0363802KLF2

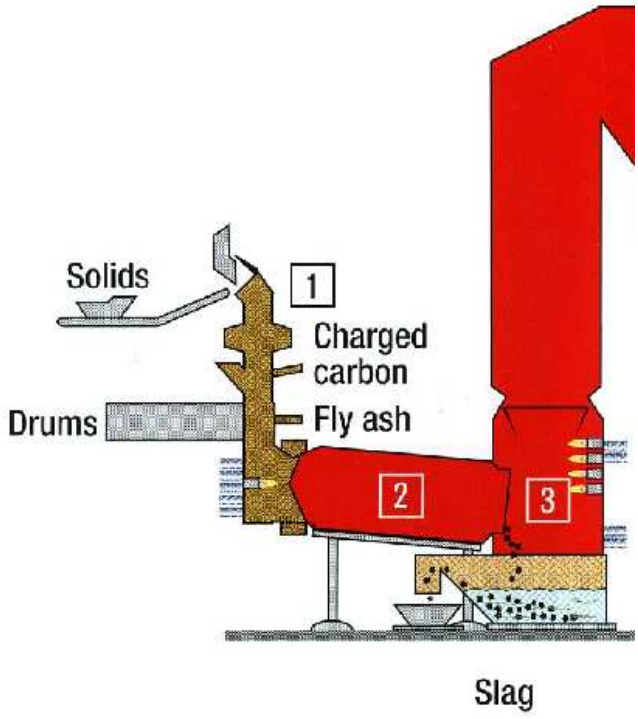

Figure 1. Schematic view of the central part of the hazardous waste incineration plant.

discontinuously incoming fuels cause problems for several reasons:

- The content of the barrels has to be declared by the delivering customer. But, since it is for safety reasons not allowed to open these packages, their real content is unknown and the customers declaration cannot be verified.
- The incineration of especially large barrels is usually controlled manually in order to avoid high CO-emissions. This is done by setting the combustion air flow to a safe level for the current category of incoming packages.
- But, if a package unexpectedly contains waste material with a higher caloriphic value, its combustion would have a significant air deficiency. That results in a violation of the hard statutory requirements for emissions and environmental pollution.

Because of these problems, the plant operator is interested in an automation of the incineration of these waste packages.

An automatic control of that package incineration is difficult, since properties like caloriphic value of the waste within the package is not known exactly. Thus, a conventional control strategy for that combustion must rely on the emission-values itself in order to detect significant changes and to try to compensate. The problem of this strategy is, that on one hand these emissions are measured with delay and on the other hand, if a significant change of these parameters can be detected, in most cases a violation of the legal requirements already happened.

A new approach to solve this problem is to directly and automatically observe the combustion process with a special video-camera system provided by our project partner Powitec. These special camera systems allow an observation of the interior of the rotary kiln despite of the extreme conditions with respect to temperature, dust and so on. But, since there is no prior knowledge about an

interpretation of that new visual information to control the process, we decided to investigate an adaptive and self learning control approach based on reinforcement learning (RL). Recent RL-approaches following the actor-critic design seem to be very promising and applicable for our control problem, because of their ability to operate in continuous state and action spaces.

3 The ADHDP-Neurocontroller

Action Dependent Heuristic Dynamic Programming (ADHDP) was first presented by [6] as a promising neuro-control approach for high-dimensional, non-linear and noisy control problems. Its performance has been published for various applications like automated aircraft landing [2] or stabilization of the Apache helicopter [1]. ADHDP as a member of the adaptive critic design family was inspired by biological information processing, where on one hand a motoric subsystem generates actions based on the current system state, and on the other hand an emotional subsystem learns to estimate the long term consequences of that policy. As to be seen in figure 2, in ADHDP, these two biological subsystems can be represented by two neural networks: the actor net (motor-system) and the critic net (emotional subsystem) [4].

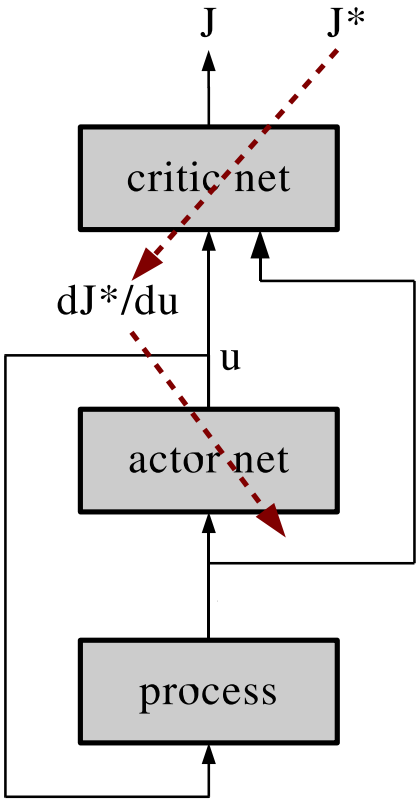

Figure 2. Principal structure of an ADHDP-controller. Inspired by biological information processing for control, two subsystems for generation of actions (action net) and prediction of the outcomes of that policy for the system (Critic net) are used. The action net observes the state $\underline{x}$ of the process and computes an action u that optimizes the process. The critic net estimates the value J of these state-action-pairs. After defining an optimal value J^* for the current state-action-pair it is possible to calculate the derivation $\partial J^*/\partial u$, which is used to adapt the action net. That gradient yields information about how the action net should change its output (policy u) in order to obtain higher (better) evaluations J by the critic net.

The training of that system can be divided into two subprocesses. On one hand, the critic net has to learn a value function for the process states $\underline{x}$ while applying the current policy, which is defined by the action network. This value function is defined by the traditional Bellman equation. On the other hand, the action network has to be adapted in order to optimize the output of the critic net. That corresponds to an optimization of the behavior of the agent with respect to the predefined reward function.

In contrast to conventional reinforcement learning approaches like Q-learning [5], for ADHDP there is no need to store or approximate all values for every state-action pair. Instead, the actions are generated by a separate network and in consequence, the critic network only has to approximate the values for each process situation given the current policy. That becomes an advantage if the state space becomes to large to be explored and stored in memory.

4 Implementation and Results

For any optimization approach there is required a numerical description of the value of the process to be controlled. For reinforcement learning, this is done by a reward function.

4.1 The Reward-Function

The plant operator is interested in an incineration of the barrels without any increase of CO-emissions while keeping a given temperature level inside the kiln. These two subgoals are hard to achieve, because a certain flow of continuously incoming fuels has to be burned to maintain the desired kiln-temperature. The occasional, additional incineration of a barrel with unknown contents while running a high level of other fuels may cause a lack of oxygen inside the kiln, which results in an incomplete combustion of carbon to carbon monoxide (CO). We tried to fulfill these two subgoals by two controllers. First, a traditional PID-controller maintains the kiln-temperature by controlling the amount of the continuously incoming fuels. In parallel, the ADHDP-controller tries to avoid high CO-emissions by controlling the amount of incoming fresh air.

In consequence, the reward function for the ADHDP-controller is defined by equation 1.

$$J \;=\; 1 - x^{CO} \tag{1}$$

Because the input variable x^{CO} denotes the current CO-emission normalized to the interval $[0 \ldots 1]$, the value J is close to one without CO-emissions and decreases down to zero otherwise.

4.2 Input Preprocessing

After numerically describing the value of the incineration of a barrel, we had to define, which information should be used as input for the action and the critic network. Thereto, based on the experience and knowledge of the plant operators we defined a set of channels originating from the distributed control system (DCS) of the plant, that described the current state of the incineration process. We used the temperature inside the kiln, and the visual information generated by the Powitec-camera system to describe the current situation of the combustion process. Because the camera images are very high dimensional data, we applied a dimension reduction realized by a neural network based principal component analysis (PCA) [3]. Thus, the whole image showing the flames inside the kiln is represented by the two coordinates belonging to those two eigenvectors having the largest eigenvalues. In order to provide some information about the prior behavior of the flame, we apply a time delayed approach by using five past values for each PCA-coordinate covering a total time range of about eight minutes. Finally, the measurement from the DCS (kiln-temperature) together with the 10

visual features (first 2 PCA-coordinates from 5 different time steps each) are used as a description of the current process situation and are denoted as $\underline{x}'(t)$. After removal of outliers, low-pass filtering over time and scaling to the interval $[0 \ldots 1]$ the resulting vector $\underline{x}(t)$ is used as input for the action network. To control the incineration process, our ADHDP-system is allowed to control the amount of incoming fresh air u by modulating the set-point of a subsequent O_2-controller. Thus, the critic network additionally receives the control variable (real oxygen concentration as a measure for the amount of used fresh air) as input.

4.3 Training of Critic-Net

Usually, the critic net is trained to approximate the mapping $\underline{x}(t) \times u^{AN}(t) \mapsto J(t)$, where the current action $u^{AN}(t)$ is generated online by the action net's policy. In our case, we only had a data set, where the control actions $u^{OP}(t)$ show the behavior of the human plant operators. Typically, these actions are not always unique, that means that different plant operators reacted to the same situation with different control actions. Nevertheless, the causal relationship between a current situation $\underline{x}(t)$, the corresponding control action $u^{OP}(t)$, and the observed reaction $J(t)$ can be used to train a valid critic net.

We use a classical multilayer perceptron (MLP) with one hidden layer and gradient descent (see figure 3).

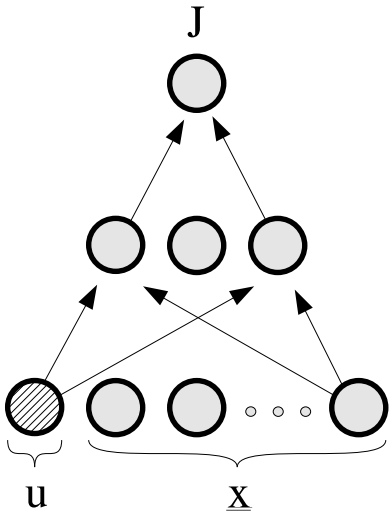

Figure 3. The critic net is realized by a two layered sigmoidal perceptron, which maps the current situation $\underline{x}$ and the current control action u to the value of the situation J.

4.4 Training of Action-Net

The training of the action net becomes more difficult, since the control actions of the data set $u^{OP}(t)$ do not follow a unique policy. This is a problem, because the original ADHDP-approach backpropagates a desired optimal output J^* through the critic net to adapt the action net. In our case, that is not possible, since a randomly initialized action net would not generate the same actions as did by the plant operators. Thus, the backpropagation $\partial J^* / \partial u^{OP}(t)$ would mislead the adaptation of the action net, since the action net probably has generated a different action $u^{AN}(t) \neq u^{OP}(t)$. In consequence, also another and wrong teacher for the action net adaptation would be calculated.

To solve this problem, we do not start the ADHDP-training with a randomly initialized action net. Instead, we first trained the action net to copy the behavior of the plant operators as good as possible and afterwards entered the usual ADHDP-training.

As for the critic net, we use a standard multilayer perceptron with one hidden layer and gradient descent (see figure 4).

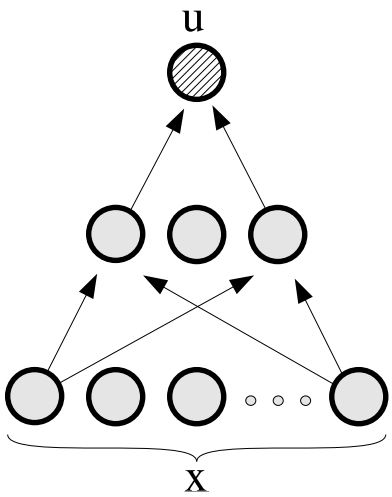

Figure 4. The action net is realized by a two layered sigmoidal perceptron, which maps the current situation $\underline{x}$ onto the desired control action u.

4.5 Results of the ADHDP-Training

After training the action net to copy the plant operator and training the critic net to predict the value of the control actions of the data set, we applied the ADHDP-approach to optimize the output of the critic net J by changing the output of the action net u^{AN} based on a gradient ascent defined by $\partial J^*/\partial u^{OP}(t)$.

Figure 5 depicts the result of that process. As can be seen, the output of the critic net tries to approximate the value of the state-action pair defined by the reward function (equation 1). Furthermore, the action net tries to optimize the output of the critic net by changing the action performed by the plant operator. In case of an high CO-emission, which could be predicted by the critic-net based on visual information, the action-net generated an increased set-point for a subsequent O_2-PID-controller. In consequence , that O_2-PID-controller fed more fresh air into the kiln in order to provide more oxygen and thus to avoid the production of CO.

Figure 6 shows one example of the online-behavior of that ADHDP-agent in detail. Based on the combustion describing data from the Powitec-video system, the ADHDP-controller was able to predict the oncoming high CO-emission (figure 6, lower diagram) starting at 12:43 about 3 minutes earlier at 12:40 (figure 6, upper diagram)! Based on that early prediction, the ADHDP-controller immediately increased the set-point of the subsequent O_2-PID-controller in order to increase the amount of incoming fresh air (figure 6, middle) and thus to avoid the oncoming high CO-emission. As to be seen, despite of the correct control action of the ADHDP-agent, a high CO-emission could not be avoided completely, but significantly reduced. Usually, these CO-emissions are up to ten times larger than the CO-peak depicted in figure 6.

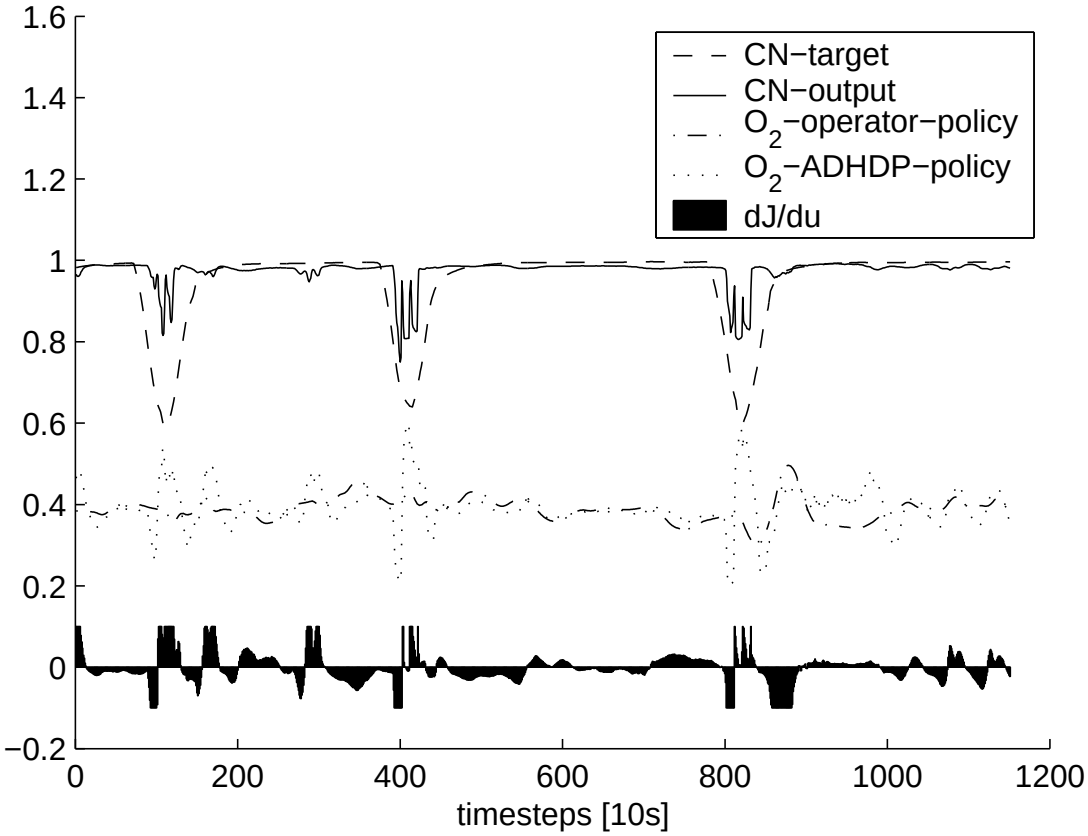

Figure 5. Result of the training process of both action an critic net. The critic net output (solid line) approximates the value of control actions based on the reward function (CN-target, dashed line) for about 3 hours. Based on that value prediction, the derivation $\partial J^*/\partial u^{AN}$ (black bars) defines, how to change the output of the action network in order to optimize the plant behavior. As can be seen, every high CO-emission (lower CN-target, dashed line) could be correctly predicted by the critic-net by a lower CN-output (solid line). In contrast to the plant operator, the ADHDP-controller increased the set-point for oxygen (dotted line) in order to avoid high CO-emissions.

5 Conclusions & Outlook

In this paper we applied a neurocontroller based on Action Dependent Heuristic Dynamic Programming (ADHDP) to optimize the combustion-process for an industrial hazardous waste incineration plant. The problem of online learning of the ADHDP-controller could be solved by a pre-training of the action net to copy the behavior of the plant operators. Afterwards, the usual ADHDP-training could be applied offline using a data set. Thus, we will be able to start the period of online mode for optimization of the ADHDP-controller with a policy, witch is as good as the policy of the plant operators. There is no learning period with probably very bad initial behavior required.

We could demonstrate the functioning of the original ADHDP learning scheme after pre-training of the action network. It could be shown, that the critic network was able to predict oncoming CO-peaks and that the derivation $\partial J^*/\partial u^{AN}$ can be used to adapt the action network so, that the critic network would predict higher values for the adapted policy. That means, the ADHDP-controller is able to find out how to change its policy to optimize the incineration process. Finally, we could present promising first results of successful online-operation of the ADHDP-controller at the industrial hazardous waste incineration plant in Hamburg (AVG).

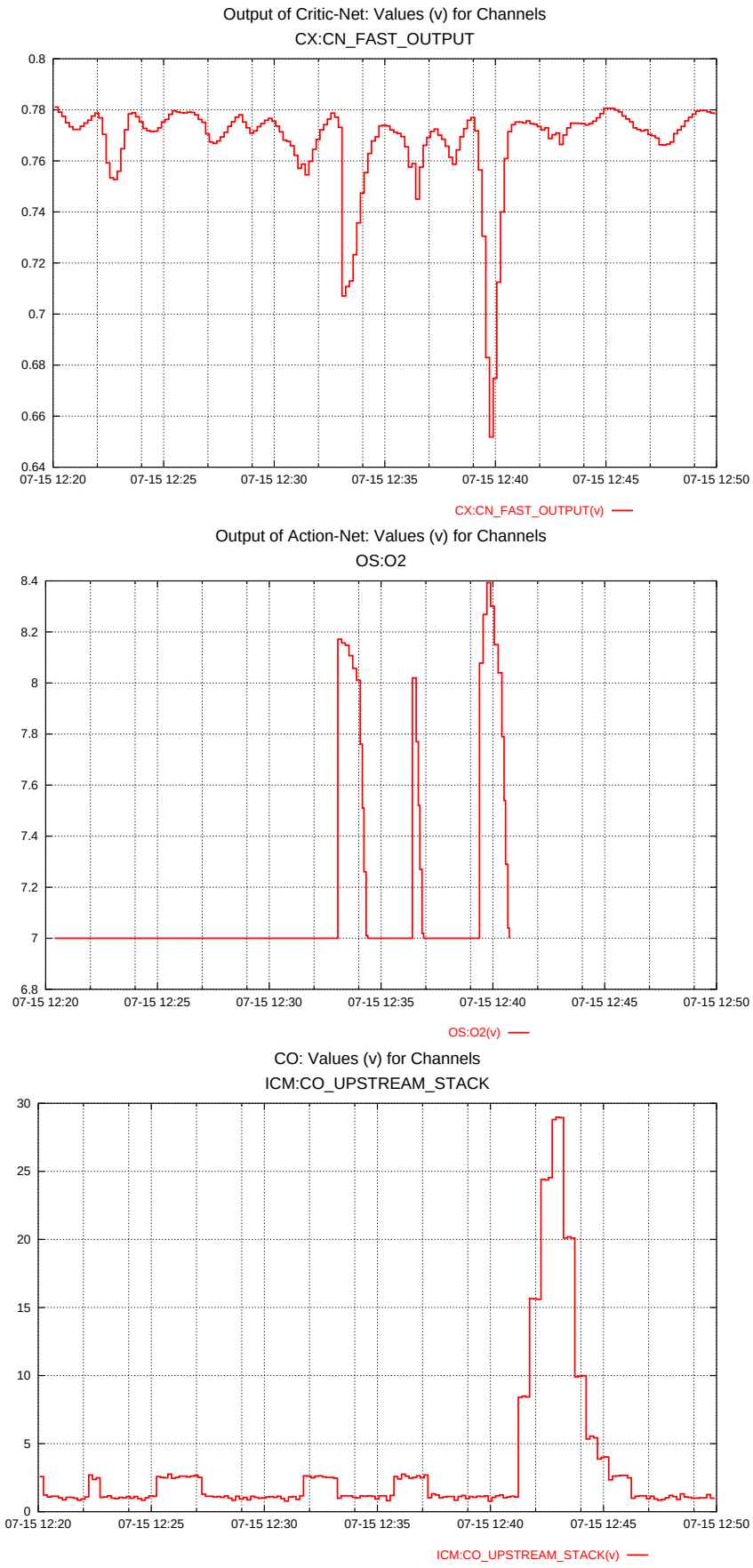

Figure 6. Online behavior of the ADHDP-agent. For explanation see text.

Future work addresses further optimization of the presented control approach with respect to accuracy and speed of the ADHDP-controller. Additionally, more comprehensive comparisons between the conventional and ADHDP-controlled mode of plant operation have to be maid.

References

[1] R. Enns and J. Si. Apache helicopter stabilization using neuro-dynamic programming. Technical report, Arizona State University, Tempe, AZ 85287, 2002.

[2] D. Prokhorov, R. Santiago, and D.C. Wunsch. Adaptive critic designs: a case study for neurocontrol. *Neural Networks*, 8:1367–1372, 1995.

[3] T. Sanger. Optimal unsupervised learning in a single layer linear feedforward neural network. *Neural Networks*, 2:459–473, 1989.

[4] J. Si and Y.-T. Wang. On-line learning control by association and reinforcement. *IEEE Transactions on Neural Networks*, 12:264–276, 2001.

[5] C. J. C. H. Watkins. *Learning from Delayed Rewards*. PhD thesis, Cambridge University, Cambridge, England, 1989.

[6] P.J. Werbos. Neurocontrol and supervised learning: An overview and evaluation. In White, editor, *Handbook of intelligent control*. 1992.