IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.Doi Number

An online control approach for forging machine using reinforcement learning and taboo search

Dapeng Zhang[1], Zhiwei Gao[2], Senior Member, IEEE, and Zhiling Lin[3]

[1]School of Electrical and Information Engineering, Tianjin University, Tianjin, 300072 CHINA
[2]Faculty of Engineering and Environment, University of Northumbria, Newcastle upon Tyne, NE2 8ST UK
[3]School of Electrical Engineering, Tianjin University of Technology, Tianjin, 300384 CHINA

Corresponding author: Dapeng Zhang (e-mail: zdp@tju.edu.cn).

This work was supported in part by the Tianjin University Innovation Fund under Grant 2020XT-0024.

ABSTRACT It is noticed that offline-training and online-implementation method is dominant in the data-driven control. However, the inconsistence existing in offline data and online data may degrade the control performance. To address the aforementioned issue, an online control strategy is developed so that the control parameters can be updated online based on the real-time data measured to ensure satisfactory control performance in this study. Specifically, an online control algorithm is addressed to control the pressing-down speed of the forging machine based on the framework of the reinforcement learning that has a capability of building a complete mapping from state space to action space only according to the neighbour samples. Rather than using the way of trials and errors which is too slow to be online implementation, a taboo search is addressed to speed up the learning-working process by directly searching the control on the current states, followed by the stability conditions, derived from Lyapunov stability theory. A coarse model that is limited to get the cost information of the reinforcement learning is used to make the best of mechanism information, which prevents the occurrence of the invalid states that do not conform to system characteristics. The effectiveness of the algorithm is demonstrated by an ultra-low forging machine, which outperforms the conventional approaches such as PID and neural network control approaches.

INDEX TERMS Online control, Reinforcement learning, Taboo search, Forging machine

I. INTRODUCTION

Forging machine, as an electro-hydraulic hybrid system with nonlinearity and multi-field coupling, is an essential equipment in forging industry [1]. The control on the forging machine is the guarantee of the quality for the forgings production which is vital for the high reliability areas such as in aviation, space exploration and nuclear industry. To meet the needs of the precise forgings, some advanced algorithms such as the sliding mode control [2,3], back-stepping control [4], feedback linearization [5] were used in the control on forging machine instead of the conventional PID-based control [6] and fuzzy-based control [7,8]. However, the aforementioned approaches [2-5] are model-based control algorithms, which strongly depend on the accuracy of model. Unfortunately, it is hard to build an accurate model in a complex engineering practice. For example, the viscosity of hydraulic oil is prone to be influenced by the temperature, which will lead to the model bias. On the other hand, the forging machine is usually facing the different forging batches, which further increases the difficulty in producing an accurate model.

Compared with the established model, the collected data will be better to reflect the real states which are interacted with the system and the surroundings. Therefore, the data-driven approaches [10,11] based on the fact that advanced measurement techniques [9] have made it easy to obtain the large-scale data online have been introduced to the forging machine field in recent years. Reference [12] developed two online updated backpropagation (BP) neural network algorithms to accurately control the die forging hydraulic press machine. The weights of the neural networks were initially trained offline and then updated online according to an error backpropagation algorithm. A novel least squares support vector machine (LS-SVM) control method was addressed in [13] for general unknown nonlinear systems, which was further proved that the control error was fully equal to the LS-SVM modeling error. In [14] a novel online probabilistic extreme learning machine (ELM) method was proposed to model batch forging processes. By using the characteristics of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3020550, IEEE Access

the online ELM, a strategy was developed to update the distribution model as new forging process data were collected. In [15] a combination of the neural network and genetic algorithms had been employed to optimize the forging force.

These data-driven approaches are always working on a way of offline-learning and online-working. The offline-learning forms an implication relation according to the historical data. After this implication is obtained by learning with the ways of supervisory or un-supervisory it will be used for the online-working as a black box model. Either supervisory learning or un-supervisory learning requires a large volume of data as the training dataset, however, it is difficult to get them as the forging machine is often to deal with different forging batches. Firstly, for a new forging process, the training data are empty due to the lack of historical process, while for some special forging processes, the training data are not available due to the differences of the experimental conditions or tests. Secondly, it is inevitable that the working situation is not consistent with the training condition which leads to the performance degradation, even mistake of the forging machine under the function of the previous well-trained controller. As a result, it is a challenge to develop a control strategy for forging machine without depending on the accurate model and the way of offline-learning and online-working in traditional data-driven approach.

Using a way of online-learning and online-working is a feasible solution for the forging machine control because the forging machine is always working at a slow process due to the machine's large mechanical inertia and slow hydraulic activity. Compared with this slowness, the computer shows an amazing computing ability which makes the way of online-learning online-working become possible. All the methods concerning the accurate model and an amount of historic data are forced to be abandoned due to the aforementioned limitations of the forging machine.

To our best knowledge, the reinforcement learning (RL) [16] is able to support the offline learning (Q algorithm) and online learning (e.g., Sarsa algorithm) by the means of approaching to the stage reward with adjusting the action based on the difference of the adjacent sampling time series as an error rectification. The RL does not need an accurate model and it just needs an effect from the action which reduces the requirement of the precision for traditional model. Now the RL has been extended to the deep reinforcement learning (DRL) with the development of deep learning technique. Reference [17] developed a novel artificial agent, termed a deep Q-network, that can learn successful policies directly from high-dimensional sensory inputs using end-to-end reinforcement learning. Reference [18] presented a brief survey on the advances that have occurred in the area of deep learning. From engineering application aspect, the RL/DRL showed an excellent performance after a good training in UAV [19], air-conditioning refrigeration [20], smart power control [21], fault tolerant control [22,23] and so forth [24,25].

The difficulty of RL in applying to the practical system is its slow training speed whether offline nor online. The RL aims to build a complete mapping between the state space and the action space by training with trial and error in order to deal with the unknown environment. The training is divided into the value-based method and the policy-based method. Compared with the value-based method, the policy-based method is dominated due to its simplicity and intuition. Most algorithms for policy optimization can be classified into three broad categories [26]: (1) policy iteration methods, which alternate between estimating the value function under the current policy and improving the policy [27]; (2) policy gradient methods, which use an estimator of the gradient of the expected return (total reward) obtained from sample trajectories [28]; and (3) derivative-free optimization methods, such as the cross-entropy method (CEM) and covariance matrix adaptation (CMA), which treat the return as a black box function to be optimized in terms of the policy parameters [29]. Generally, both methods spend a long time to train which is often unbearable for real-time system. Concurrently the trial actions in training process maybe bring to the system risk of out of control because no one knows the effect of actions on the system in advance.

In fact it is not necessary to spend too much time to build a complete mapping from states space to action space because the succeeding states will follow up the occurred states under the control which is a subset of the complete mapping. Searching a control in this subset will speed up the train owning to removing the large redundant states. The taboo search (TS) proposed by Glover [30] is an effective stochastic optimization method [31] which gets rid of the historical data in training of the data-driven methods. The TS has an efficient search capability by avoiding circuitous search with introducing a flexible storage structure and corresponding Taboo criteria. It also escapes the local extremum by extending the local optimization to the global optimization. As a result, the TS algorithm is selected as a substitution for trials and errors.

The above discussions show an evolution of control on the forging machine from the model-based control to the data-driven strategy in which most studies focused on the way of offline-learning and online-working. Motivated by overcoming the difficulties of the inconsistence between the training and the working for forging machine, a novel approach is proposed to implement an online control of the forging machine in this study. By integrating reinforcement learning with taboo search, the RL is taken as the evaluation of the actions, and the taboo search is used to improve the learning efficiency. On the other hand the computer simulation technology provides the way of forecasting the system state without a real action on the system which avoids the danger of system out of order from the training actions. The advantages of proposed approach are summarized as follows:

(1) This is an online approach with the combination of the data and model which breaks through the conventional mode

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3020550, IEEE Access

of offline learning and online working. The optimal control will be achieved in the common process of the learning and working.

(2) All the control vectors are limited within the range of requirements based on the current states which guarantees the system stability in the learning process.

(3) The learning process is speeding up to meet the real-time requirement by bringing to the taboo search which abandons the redundant states independent of the current states.

The remainder of this article is organized as follows: In section 2 the forging machine model is addressed and the relation between the states and controls is derived under the stability condition. Section 3 descripts the proposed approach including the reinforcement learning, the taboo search, the structure and algorithm. The case studies are illustrated in Section 4, followed by conclusions in section 5.

II. FORGING MACHINE AND STABILITY CONDITION

A. THE MODEL OF FORGING MACHINE

The ultra-low forging machine with the heavy force and the slow speed is equipment for a semi-solid metallic confectioning constant-speed isothermal forging which is an important forging technique, particularly for light-weight alloy confectioning in the aerospace industry. The typical structure of the ultra-low forging machine is depicted in figure 1.

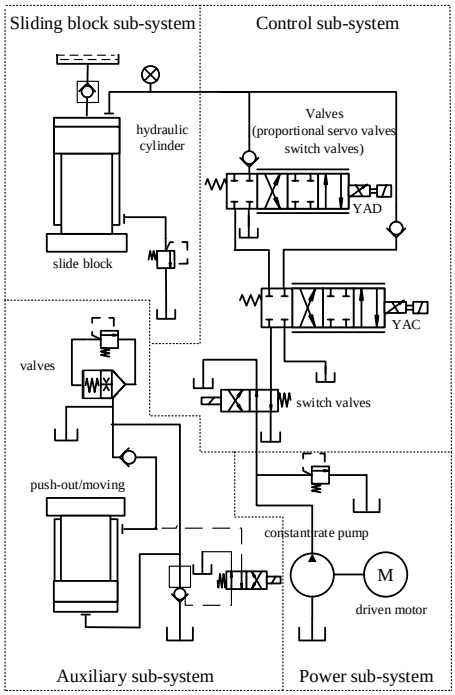

FIGURE 1. The typical structure of ultra-low forging machine

The forging machine is divided into a power sub-system, a sliding block sub-system, a control sub-system and an auxiliary sub-system. A power sub-system consists of an oil resource that forms the high pressure working oil through a constant rate pump with a driven motor and the pipe that delivers the high pressure working oil to the operating mechanisms. The sliding block sub-system is made up with a hydraulic cylinder that produces the high pressing force at a sliding velocity and a huge slide block that directly acts on the forgings. The control sub-system includes all kinds of valves, sensors and control algorithms, in which the switch valves complete the logic function of the process, and the proportional servo valves control the speed of the slide block by adjusting the valve openings. The auxiliary sub-system is used to implement the additional functions except for the pressing process such as push-out, moving and so on.

The pressing-down phase is the key process in the semi-solid metallic confectioning constant-speed isothermal forging process which usually includes six phases: fast-down phase, slow-down phase, pressing-down phase, keep-pressure phase, fast-up phase and slow-up phase. This pressing-down phase is made up with a long pipe-line with working oil, a proportional servo valve, and a hydraulic cylinder.

For a long pipe-line with working oil, the dynamic process can be described by [5]:

$$\frac{dq_1}{dt} = \frac{S_1(p_1 - p_s)}{\rho l} + \frac{32\rho^2\mu}{D^2} q_1 \tag{1}$$

$$q_2 - q_1 = \frac{S_1 l}{K}\frac{dp_1}{dt} \tag{2}$$

where q_1 is the oil flow of pipe; p_1 is the inlet pressure of proportional servo valve; q_2 is the flow of proportional servo valve, and the other parameters are defined in Table 1.

For a proportional servo valve, the dynamics can be described by [5]:

$$\frac{1}{\omega_n^2}\frac{d^2q_2}{dt^2} + \frac{2\xi}{\omega_n}\frac{dq_2}{dt} + q_2 = K_n\sqrt{\frac{p_2 - p_1}{\Delta p_n}} \cdot O_p \tag{3}$$

where the symbols in (3) is shown in Table 1.

For a hydraulic cylinder, the dynamic processes can be illustrated by [5]:

$$\frac{K}{V_c}q_2 - \frac{KS_2}{V_c}v - \frac{K\lambda_c}{V_c}p_2 = \frac{dp_2}{dt} \tag{4}$$

$$\frac{S_2}{m}p_2 - \frac{B}{m}v - \frac{F_l}{m} = \frac{dv}{dt} \tag{5}$$

where $\dot{q}_2$ is the flow velocity, v is the speed of slide block and the other parameters are explained in Table 1.

In terms of equations (1)-(5), the compact state-space model can be given as follows:

$$\dot{x} = Ax + g(x)u \tag{6}$$

where

$x = [x_1, x_2, x_3, x_4, x_5, x_6] = [q_1, p_1, \dot{q}_2, q_2, p_2, v]^T$,

$u = [O_p]$, $\mathrm{A} = \begin{bmatrix} \frac{32\rho^2\mu}{D^2} & \frac{S_1}{\rho l} & 0 & 0 & 0 & 0 \\ -\frac{K}{S_1 l} & 0 & 0 & \frac{K}{S_1 l} & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & -2\xi\omega_n & -{\omega_n}^2 & 0 & 0 \\ 0 & 0 & 0 & \frac{K}{V_c} & -\frac{K\lambda_c}{V_c} & -\frac{KS_2}{V_c} \\ 0 & 0 & 0 & 0 & \frac{S_2}{m} & -\frac{B}{m} \end{bmatrix}$,

$\mathrm{g} = [0,0,0,{\omega_n}^2 K_n\sqrt{\frac{x_2 - x_5}{\Delta p_n}}, 0, -\frac{F_l}{B}]^T$.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3020550, IEEE Access

The states q_1 is the oil flow of pipe; p_1 is the inlet pressure of proportional servo valve; The meanings of model parameters are table I.

TABLE I
THE MEANINGS OF MODEL PARAMETERS

Symbol	Meanings	Symbol	Meanings
ρ	The density of oil	λ_c	The leak coefficient of hydraulic cylinder
μ	The friction coefficient of pipeline	S_2	The plunger's sectional area of exporting cavity of hydraulic cylinder
D	The diameter of oil pipe	m	The mass of slide block
S_1	The sectional area of pipe	B	The viscous damping coefficient
l	The length of oil pipe	K_n	The rated flow gain
K	The young's modulus of oil equal volume	Δp_n	The valve port pressure drop
ξ	The damping rate of propositional servo valve	F_l	The load resistance
ω_n	The inherent frequency of propositional servo valve	P_s	The pressure from a constant rate pump
V_c	The oil volume of the upper cavity of hydraulic cylinder	O_p	The opening of proportional servo valve

B. THE CONDITION OF STABILITY

The relation between the states and control variables are gotten according to the Lyapunov stability condition. Let

$$V = x^T P x \tag{7}$$

where P is a semi definite matrix with the form of

$$P = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & p_{22} & 0 & 0 & 0 & 0 \\ 0 & 0 & p_{33} & 0 & 0 & 0 \\ 0 & 0 & 0 & p_{44} & 0 & 0 \\ 0 & 0 & 0 & 0 & p_{55} & 0 \\ 0 & 0 & 0 & 0 & 0 & p_{66} \end{bmatrix} \tag{8}$$

with $p_{22} > 0, p_{33} > 0, p_{44} > 0, p_{55} > 0$ and $p_{66} > 0$.

According to the physical meaning of states $x_i \neq 0$ ($i = 2,3,4,5,6$), one has

$$V = x^T P x > 0 \tag{9}$$

$$\begin{aligned} \dot{V} &= \dot{x}^T P x + x^T P \dot{x} \\ &= (Ax + g(x)u)^T P x + x^T P (Ax + g(x)u) \\ &= x^T \underbrace{(A^T P + PA)}_{I} x + \underbrace{u^T g^T(x) P x + x^T P g(x) u^T}_{II} \end{aligned} \tag{10}$$

If $I \leq 0$ and $II < 0$ there exits $\dot{V} < 0$ which means the system is Lyapunov stability.

For I and II

$$I = A^T P + PA \leq 0 \tag{11}$$

$$II = u^T g^T(x) P x + x^T P g^T(x) < 0 \tag{12}$$

Using (6) and (11), one can obtain

$$\begin{bmatrix} \frac{32\rho^2\mu}{D^2} & \frac{S_1}{\rho l} & 0 & 0 & 0 & 0 \\ -\frac{K}{S_1 l} & 0 & 0 & \frac{K}{S_1 l} & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & -2\xi\omega_n & -{\omega_n}^2 & 0 & 0 \\ 0 & 0 & 0 & \frac{K}{V_c} & -\frac{K\lambda_c}{V_c} & -\frac{KS_2}{V_c} \\ 0 & 0 & 0 & 0 & \frac{S_2}{m} & -\frac{B}{m} \end{bmatrix}^T \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & p_{22} & 0 & 0 & 0 & 0 \\ 0 & 0 & p_{33} & 0 & 0 & 0 \\ 0 & 0 & 0 & p_{44} & 0 & 0 \\ 0 & 0 & 0 & 0 & p_{55} & 0 \\ 0 & 0 & 0 & 0 & 0 & p_{66} \end{bmatrix}$$

$$+ \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & p_{22} & 0 & 0 & 0 & 0 \\ 0 & 0 & p_{33} & 0 & 0 & 0 \\ 0 & 0 & 0 & p_{44} & 0 & 0 \\ 0 & 0 & 0 & 0 & p_{55} & 0 \\ 0 & 0 & 0 & 0 & 0 & p_{66} \end{bmatrix} \begin{bmatrix} \frac{32\rho^2\mu}{D^2} & \frac{S_1}{\rho l} & 0 & 0 & 0 & 0 \\ -\frac{K}{S_1 l} & 0 & 0 & \frac{K}{S_1 l} & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & -2\xi\omega_n & -{\omega_n}^2 & 0 & 0 \\ 0 & 0 & 0 & \frac{K}{V_c} & -\frac{K\lambda_c}{V_c} & -\frac{KS_2}{V_c} \\ 0 & 0 & 0 & 0 & \frac{S_2}{m} & -\frac{B}{m} \end{bmatrix}$$

$$\leq 0 \tag{13}$$

Formula (13) shows a part expansion of the Lyapunov stability based on the forging machine model. Solving formula (13), one can obtain:

$$\begin{cases} 0 \leq 0 \\ 0 \cdot p_{22} \leq 0 \\ 0 \cdot p_{33} \leq 0 \\ -{\omega_n}^2 p_{44} \leq 0 \\ -\frac{K\lambda_c}{V_c} p_{55} \leq 0 \\ -\frac{B}{m} p_{66} \leq 0 \end{cases} \tag{14}$$

As a result, P can be selected as follows:

$$P = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{15}$$

Substituting (15) into (12), one can have

$$u^T \begin{bmatrix} 0,0,0,{\omega_n}^2 K_n \sqrt{\frac{x_2 - x_5}{\Delta p_n}}, 0, -\frac{F_l}{B} \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{bmatrix}$$

$$+ [x_1, x_2, x_3, x_4, x_5, x_6] \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0,0,0,{\omega_n}^2 K_n \sqrt{\frac{x_2 - x_5}{\Delta p_n}}, 0, -\frac{F_l}{B} \end{bmatrix}^T u < 0 \tag{16}$$

Solving (16), one can have

$$u \cdot \left({\omega_n}^2 K_n \sqrt{\frac{x_2 - x_5}{\Delta p_n}} \cdot x_4 - \frac{F_l}{B} \cdot x_4 \right) < 0 \tag{17}$$

As a result, the Lyapunov stability condition is satisfied which means the system is stable subject to formula (17).

Remark 1: Formula (17) shows the relationship between the control variable, the states and the load under the stability of system, which is regarded as a restraint condition in taboo search. The states x_2, x_5 and x_4 in formula (17) are measurable by the flow sensors and pressure sensors, and the parameters are obtained from the design of forging machine.

III. THE PROPOSED APPROACH

A. REINFORCEMENT LEARNING

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3020550, IEEE Access

The basic idea of the reinforcement learning is simply to capture the most important aspects of the agent which includes sensation, action, and goal. The basic frame of reinforcement learning is shown in Figure 2 [16].

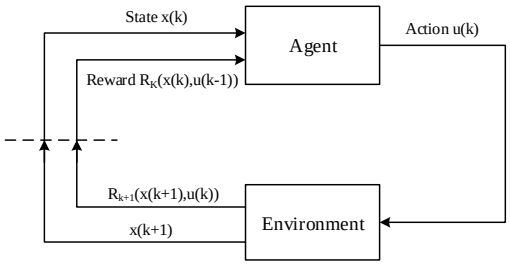

FIGURE 2. A basic frame of reinforcement learning

An agent will get the evaluation of good or bad behavior on environment and learn through experience without a teacher who teaches how to do. In each training session, named episode, the agent explores/exploits the environment by changing action $u(k)$ and receives the states $x(k+1)$ and the immediate cost $R_{k+1}(x(k+1),x(k),u(k))$ based on $x(k)$. The purpose of the training is to enhance the 'brain' of agent. The goal of an agent is to minimize/maximize the immediate cost $\sum_{i=k}^{k+T} R_i(x(i+1),x(i),u(i))$ which is received in the long run. This process is considered as a decision process MDP $(\mathcal{X},\mathcal{U},\mathcal{P},\mathcal{R})$ with a control u and cost R in which $\mathcal{X}$ is a set of states, $\mathcal{U}$ is a set of controls, $\mathcal{P}$ is the transition probabilities $\mathcal{P}: \mathcal{X}\times\mathcal{U}\times\mathcal{X}\to[0,1]$ and $\mathcal{R}$ is the cost function $\mathcal{R}: \mathcal{X}\times\mathcal{U}\times\mathcal{X}\to\mathcal{R}$.

In order to evaluate the good or bad behavior (often named action or control) the value of a control $V_k^u(x(k))$ is defined as

$$V_k^u\big(x(k)\big) = E_u\left\{\sum_{i=k}^{k+T}\gamma^{i-k}R_i\right\}$$
$$= \sum_u u(x,u)\sum_{x(k+1)} P\big(x(k+1),x(k),u(k)\big)$$
$$\times\left[R_k\big(x(k+1),x(k),u(k)\big) + \gamma E_\pi\left\{\sum_{i=k+1}^{k+T}\gamma^{i-(k+1)}R_i\right\}\right]$$
$$= \sum_u u(x,u)\sum_{x(k+1)} P\big(x(k+1),x(k),u(k)\big)$$
$$\times\left[R_k\big(x(k+1),x(k),u(k)\big) + \gamma V_{k+1}^u\big(x(k+1)\big)\right] \quad (18)$$

Where $R_i(x(i+1),x(i),u(i))$ is abbreviated by R_i because we do not stress the relation of $x(k+1), x(k)$, and $u(k)$.

The optimal controls will be achieved by carrying an alternation of the policy evaluation and policy improvement using the formulas as follows:

$$V_k\big(x(k)\big) =$$
$$\sum_u u_k(x,u)\sum_{x(k+1)} P(x(k+1),x(k),u(k)))$$
$$\times\left[R_k(x(k+1),x(k),u(k)+\gamma V_k(x(k+1))\right] \quad (19)$$
$$u_k(x,u) = \underset{u}{argmin}\sum_{x(k+1)} P(x(k+1),x(k),u(k))$$
$$\times\left[R_k(x(k+1),x(k),u(k)) + \gamma V_k(x(k+1))\right] \quad (20)$$

where γ is a discount factor with $0\le\gamma<1$ in order to converge.

For a deterministic system, it is evident that:

$$\sum_u u_k(x,u)\sum_{x(k+1)} P\big(x(k+1),x(k),u(k)\big) = 1. \quad (21)$$

Therefore, the formulas (19) and (20) can be simplified as:

$$V_k\big(x(k)\big) = R_k(x(k+1),x(k),u(k)) + \gamma V_k(x(k+1)) \quad (22)$$

$$u_k(x,u) = \underset{u}{argmin}\, R_k(x(k+1),x(k),u(k)) + \gamma V_k(x(k+1)) \quad (23)$$

Remark 2: The optimal control $u(k)$ can be obtained only using the state information and the immediate cost because there are only $x(k)$, $x(k+1)$ and R_k in formulas (22) and (23).

A general approach is to adopt iterative method until it is convergent. It is a time-consumption process due to a large number of iterations which form the disadvantage of RL. In fact once $x(k)$ is determined, $u(k)$ will be within a feasible space due to the system limitation. One can directly seek an appropriate control $u(k)$ to maximize the cost function, which can be solved by the technology of the random optimization search. Here we chose the taboo search owing to its high search efficiency.

B. TABOO SEARCH

There are more complex versions of the taboo search which improve its searching capability. Here the basic taboo search algorithm is applied to demonstrate its application in finding the optimal solution. For an element x in the discrete space X, the goal is

$$\min C(x)$$
$$\text{s.t. } x\in X \quad (24)$$

and the optimal states are solved by neighbor moving continuously

$$s(x) = \{s|s = x + wd, s\in X\} \quad (25)$$

where w is the step length, d is the direction. A taboo list whose goal value is updated according to the first input first output (FIFO) rule is designed to prevent the loop search. But the aspiration $A(s,x)$ that records the best solution of history is not limited by the taboo list.

The basic taboo search is summarized as procedure 1.

Procedure 1

Step1: Generate an initial $x, x\in X$, then let the optimal $x^* = x$ and set a null of the taboo list $T = \emptyset$

Step2: Choose a neighbor solution $s(x)$ according to formula (25).

Step3: If $s(x) = opt\{s(x), s(x)\in S(x) - T\}$, let $x = s(x)$ and update $C(x)$.

Step4: If $C(s(x)) < A(s,x)$, $s(x)\in T$ and $C(s(x)) < C(x)$, let $x = s(x)$ and $A(s,x) = C(s(x))$.

Step5: If $C(x) < C(x^*)$, let $x^* = x$, $C(x^*) = C(x)$.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3020550, IEEE Access

Step6: Update taboo list by storing x to the last place of taboo list T.

Step7: Repeat step2 to step6 until one of termination conditions is met, that is, (a) the predetermined times of the moves; or (b) no improvement in the goal with adding the times of the moves

C. THE PROPOSED APPROACH

The structure of the proposed approach is shown in figure 3. Beginning with the states $x(k)$ and $x(k+1)$ at sample time k and $k+1$, the optimal control u^* is found by adjusting the u in order to target on the minimization of $C(x)$ according to the RL. Instead of the policy iteration of the gradient method, the taboo search is used to find the optimal action in the action space which is a table in the discrete system.

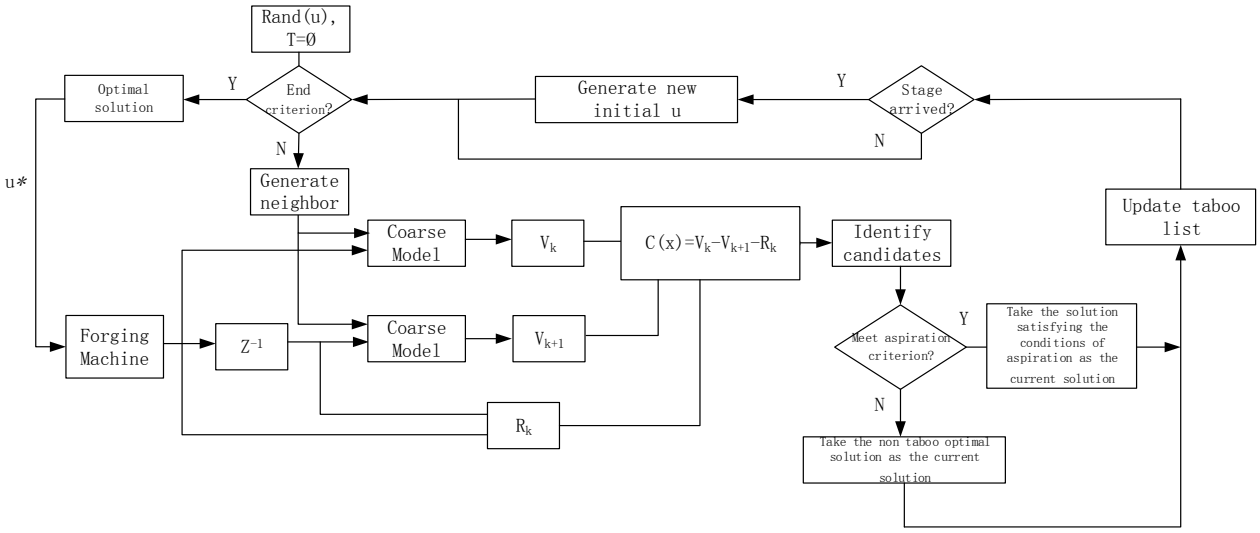

FIGURE 3. The structure of proposed approach

I) ACTION SAPCE, VALUE FUNCTION AND REWARD

The values of the control variable are limited to the analog-to-digital (DA) conversion accuracy. For a n -bit DA converter, the action space is within the range of $[2^{-n}, 2^n]$.

The forging machine's velocity is determined according to the properties of the forging materials which requires a constant pressing speed during a certain temperature range or a given curve of speed. Therefore, the immediate cost is selected as the absolution value of the error between the actual speed and the reference speed

$$R(k) = \left||v(k) - v_{\text{set}}(k)| - |v(k+1) - v_{\text{set}}(k+1)|\right| \quad (26)$$

Based on the coarse model (6) and formula (18), the cost functions $V_k(x,u)$ and $V_{k+1}(x+1,u)$ are prone to obtain,.

$$V_k(x,u) = E_u\left\{\sum_{i=k}^{k+T} \gamma^{i-k} R(i)\right\} \quad (27)$$

$$V_{k+1}(x+1,u) = E_u\left\{\sum_{i=k+1}^{k+T} \gamma^{i-k} R(i)\right\} \quad (28)$$

Noticed that the coarse model is better to express the tendency than a state expression, the time series error with TD(0) is selected as the immediate cost which is the goal of taboo search

$$\min C(x) = \min_u \left(V_k(x,u) - V_{k+1}(x+1,u) + R(k)\right) \quad (29)$$

II) NEIGHBORHOOD FUNCTION, TABOO OBJECT, TABOO LIST AND ASPIRATION CTITERION

Formula (25) provides a neighbour search but it will cause the curse with the increase of dimension. The mode of the coding and crossing changing position is usually used to avoid the curse of dimensionality in the taboo search. Let $s_i = u_i$ where $u_i \epsilon [2^{-n}, 2^n]$, this mode of the neighbour rule is given in the following:

$$[s_i,\ s_j] = [s_j,\ s_i], i \neq j. \quad (30)$$

The taboo object is selected as the current control variable u_i that is put into the taboo list. If the length l of the taboo list is too long it is prone to trap in the local optimization. If the length l of taboo list is too short it is prone to trap in the loop. Here the length of taboo list is selected as a constant of 200. The aspiration $A(s,x)$ is selected as the best states of history in order to unlock the process when all the candidates are locked.

III) LONG TERM LIST AND STRICT LIST

The basic TS has an excellent local search ability but a worse global search. A long term list that stores the initial values of each stage is proposed to improve the TS global search ability by generating the initial values as far as the past stages, this is

$$K^* = Argmax\{D(k)|D(k) = \sum_{L\in B}\sum_{i-1}^{n}(x_i^k - x_i^l)^2\} \quad (31)$$

where B is a set of selected initial solutions, and K is a set of initial values randomly generated, $K \in \mathcal{R}$.

In order to reduce the search range and speed the search velocity, a strict list is built based on the result of the system stability in section 2.2

$$\left\{u|u \cdot \left(\omega_n{}^2 K_n \sqrt{\frac{x_2 - x_5}{\Delta p_n}} \cdot x_4 - \frac{F_l}{B} \cdot x_4\right) > 0\right\} \quad (32)$$

If this condition cannot be met in the process of neighbor searching, the u will be abandoned immediately without further work.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3020550, IEEE Access

IV) THE PROCESS OF METHOD

The proposed algorithm is summarized as procedure 2.

Procedure 2:

Step 1: Give a state $x(k)$.

Step 2: Select an action $u(k)$ randomly.

Step 3: Observe the next state $x(k+1)$.

Step 4: Receive immediate reward $R\big(x(k),u(k)\big)$ according to formula (26).

Step 5: Compute the cost $V_k(x,u)$ and $V_{k+1}(x+1,u)$ according to (27) and (28) based on the coarse model (6).

Step 6: Compute the time series error $C(x)$ according to

$$C(x) = V_k(x,u) - V_{k+1}(x+1,u) + R(k)$$

Step 7: Search the neighbor based on $u(k)$ and find a new action $u(j)$ according to formula (30)

Step 8: If $u(j)$ satisfies a strict list of formula (32), then go to step 7, else repeat step 5 to 6

Step 9: Carry out the taboo search according to procedure 1

Step 10: If it achieves the stage of long term list, then reset $u(k)$ according to formula (31), else go to step 7

Step 11: Repeat steps 7 to 10 until it satisfies the terminate condition and finally gets the optimal $u^*(k)$

Step 12: Set the next state $x(k+1)$ as the current state $x(k)$ and the optimal $u^*(k)$ as $u(k)$

Step 13: Repeat steps 3 to 12 until it ends

IV. CASES STUDIES

An ultra-low forging machine is used as the test bed which is controlled by the combination of S7-300 PLC that completes the electric logic control for the process and a trio-MC224 as a special controller that implements the pressing-down phase by the proposed approach. We proposed this special controller as an addition embedded in S7-300 PLC because the PLC cannot complete this complex algorithm due to its limited computation capability. The MC224 and the PLC shared the collected data by a Modbus connection and commutated with the supervisory computer through the Profibus. The structure of test bed is indicted in figure 4.

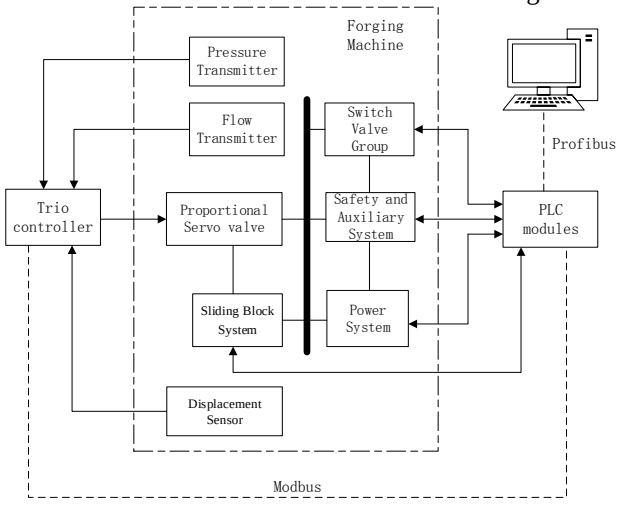

FIGURE 4. The structure of test bed

The pressure transmitter is selected as YN-type fog-proof pressure gauge with the accuracy of class 0.1. The flow transmitter is LWGYC-type with the accuracy of class 0.5. The displacement sensor is selected as the MTS production with a minimum resolution of 0.002mm. The proportional servo valve is Rexroth with the responding time less than 10ms. An ultra-low forging machine is working at the slow or ultra-low speed which will spend hours to complete a forging production. In the long pressing process, the forging is keeping the suitable temperature by the mold heating technology as dictated in figure 5.

FIGURE 5. Mold heating

According to the assembly drawing of the ultra-low forging machine, the main oil pipe is almost keeping the same diameter of 0.042m and there are protective measures on the turns in order to reduce the pressure loss of the pipeline, therefore, the actual main pipe is supposed as an ideal long pipeline. The pipe between the proportional servo valve and the hydraulic cylinder is omitted because the proportional servo valve is close to the hydraulic cylinder which leads to little pressure loss. The mass of the slide block, the plunger's size of hydraulic cylinder and the geometric parameters of the oil pipe such as the diameter and the length are obtained from the drawing annotation. The properties of the matter come from the design handbook such as the young's modulus of the oil equal volume and the density of the oil. The parameters of the proportional servo valve are obtained from the chart of the product manual. The other physical parameters are responding to the designed working point. For example, the Ps is guaranteed to the designed 32MPa with adjusting the set value of the relief valve. The friction coefficient is determined according to the criterion of the machine design. The parameters of the coarse model are indicted in table II.

TABLE II
THE VALUES OF PARAMETERS

Parameters	Values	Unit	Parameters	Values	Unit
μ	0.174	$Pa\cdot s$	S_2	0.02463	m^2
ρ	870	kg/m^3	B	2*10e4	
l	7	m	K_n	2*10e-4	
S_1	0.0138	m^2	P_s	32	MPa
K	10e9		Δp_n	3.5	MPa
ξ	0.7		m	10e3	kg
ω_n	70		λ_c	0	
V_c	4.9e-3	m^3	D	0.042	m

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3020550, IEEE Access

It is noticed that there is an implicit condition of the sampling time being small enough in formula (17) which means the states during two adjacent samplings should change a little enough. In practice, the interval of the adjustment on the ultra-low forging machine should not exceed 5 minutes for ensuring the forging quality. As a result, the ultra-low forging is always suffering a slow change. It is indicated that the practical machine is working in consistence with the assumptions of formula (17) though there is no theoretical proof. The interval of 2 minutes is chosen as the sampling time because this is the minimum time to get a valid control in our computer although the transmitters and actuators have the abilities to speed them up.

A. SCENARIO OF A CONSTANT SPEED

A pressing-down process of the slide block working at an ultra-speed of 0.03mm/s is used to test the proposed approach. In this scenario, only a few oil flow through the servo valve will pump to the upper chamber of the hydraulic cylinder to achieve the ultra-speed of the slide block. It will bring the pressure loss due to the small opening of the servo valve which causes an insufficient pressure acting on the forgings. As a result, the control of the servo valve is a compromise of the pressure loss and the working pressure. The proposed approach is following the *procedure 2*. However, TS is a random search in essence though it is an efficient searching algorithm. In order to verify the results obtained are reliable, the experiments of the pressing-down process are repeated 7 times. Figures 6 and 7 show the results of the speed and output under control at each experiment with different color curves. In figure 6 the speed is around 0.03mm/s with a little fluctuation and the maximum spikes are 0.0302mm/s (at the first experiment) and 0.0298 mm/s (at the fourth experiment) with the relative errors are 0.7%. It is seen from figure 7 that the different curves are not overlapped with each, showing some differences at each control. However, they all converge around 20.5 with fluctuations, and these differences between them do not affect to meet the need of set speed.

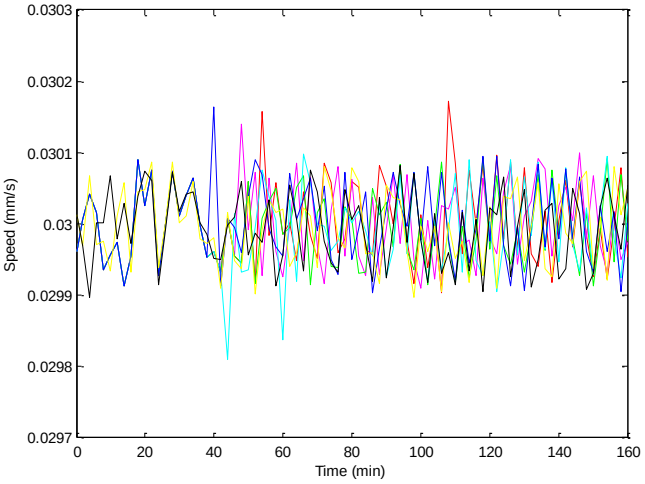

FIGURE 6. The speed of pressing-down at a constant speed

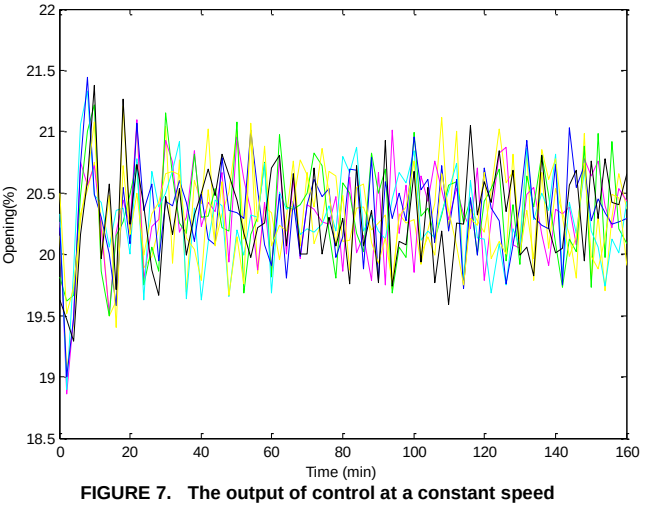

FIGURE 7. The output of control at a constant speed

As aforementioned the control of the servo valve is a compromise of the pressure loss and the working pressure. However it is difficult for the different forging processes to find this compromise due to the influences of resistance and the machine character. A practical approach is to look for the appropriate parameter values by trials and errors during the equipment debugging. All these parameters are recorded as a table and call it when required. For example, the resistance of titanium alloy is always changing with pressing speed, whose relation is following a curve according to the information of related field. Therefore, some typical speeds from the curve will be controlled as the key indicators in the debugging process and the others are determined by interpolation method and improved by fine-tune based on the working conditions. This debugging process will spend a long time (often achieves several months even years) by the conventional PID because there are many scenarios to be tested one by one. The fuzzy based approaches were applied to improve this parameter values, but failed to the requirement of accuracy. With the data increasing it is feasible to introduce the NN as a tool due to its excellent nonlinear fitting function. So for comparison, conventional PID and neural network (NN) are applied in this study.

Here an ultra-speed of 0.03mm/s are taken as an example. The parameters of the PID are adjusted by trials in order to achieve better performance as possible. A three-layer feed-forward backpropagation network with an input layer, a hidden layer and a output layer is chosen as a NN controller, whose input layer include the states $(q_1, p_1, \Delta q_2, q_2, p_2, v)$, and the output layer is the control variable (O_p). The hidden layer consists of 20 nodes full connect to the input layer and output layer by trials because there is no mature theory to follow. The NN is trained by a classical Levenberg-Marquardt method with random weights initialization. The training database is built based on the selected 4000 data from the fine control by PID in order to make sure of the excellent training database, in which 3500 as training and 500 as testing. After many times of trying to select different weights initialization, the well-trained NN is fine as a controller.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3020550, IEEE Access

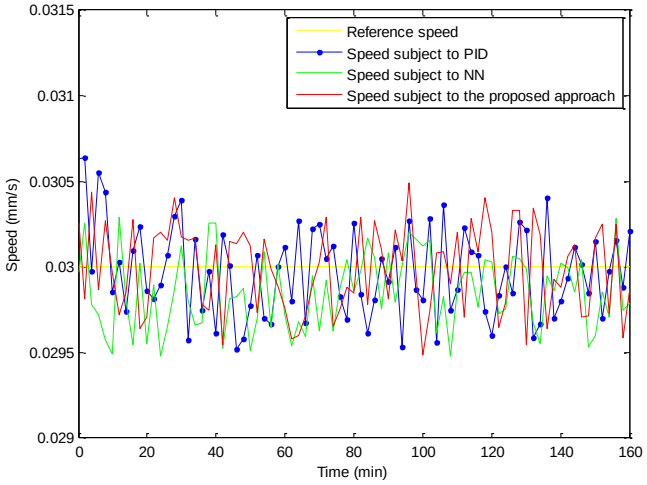

FIGURE 8. The results of pressing-down speed under different controls

The mean $\bar{x}$ and the variance σ according to formulas (32-33)

$$\bar{v} = \frac{1}{n}\sum_{k=1}^{n} v(k) \tag{32}$$

$$\sigma = \frac{1}{n}\sum_{k=1}^{n}(v(k) - \bar{v})^2 \tag{33}$$

are used to evaluate the performance. The relative error δ between the mean $\bar{v}$ and the reference v_r is according to formula (34)

$$\delta = (\bar{v} - v_r)/v_r \tag{34}$$

The results are shown in table III

Table III

THE MEAN, THE RELATIVE ERROR AND THE VARIANCE AT CONSTANT SPEEDS

	mean	variance	relative error
The PID	**0.0307**	**2.7934e-004**	**0.0233(2.33%)**
The NN	**0.0305**	**2.7393e-004**	**0.0167(1.67%)**
The proposed approach	**0.0301**	**7.5976e-008**	**0.0033(0.3%)**

It is seen from figure 8 and table III that all three methods including the traditional PID, the NN and the proposed approach have abilities to achieve the requirement of the speed accuracy (the relative errors<3%). In fact, even after the debugging stage, more parameter values to respond to the practical different cases are being collected in order to deal with the difference between offline-training and online-implementation. In the whole process, it is difficult for the PID to adjust the parameters, and the NN highly depends on an excellent training database and weights initialization. In contrast, the proposed approach can well realize the automatic control according to the current states. As a result, the proposed algorithm is in a superior position.

B. SCENARIO OF VARIANT SPEEDS

A variant speed with the range from 0.08mm/s to 0.06mm/s via 0.04mm/s is to test the proposed approach with sampling times of 2 minutes. The reference v_r follows the following formula according to the craft requirement.

$$v_r = \begin{cases} 0.08 & k \le 30 \\ -0.002(\mathrm{x} - 30) + 0.08 & 30 < k < 50 \\ 0.04 & 50 \le k \le 80 \\ 0.002(\mathrm{x} - 80) + 0.04 & 80 < k < 90 \\ 0.06 & 90 \le k \le 100 \end{cases} \tag{33}$$

This kind of pressing-down process is seldom in ultra-low forging and there is no effective approach to implement until now. In practice an experimental engineer is required to monitor this process and adjust the PID parameters online to meet the craft curve based on the experiments.

First the result of proposed approach is presented. The pressing-down process is repeated 5 times in order to verify the reliability of proposed approach due to the random essence of the TS. The results of the speed and output undercontrol are shown in figure 9 and figure 10. The cyan color, pink color, green color, red color and blue color represent the results from tests 1 to 5 respectively.

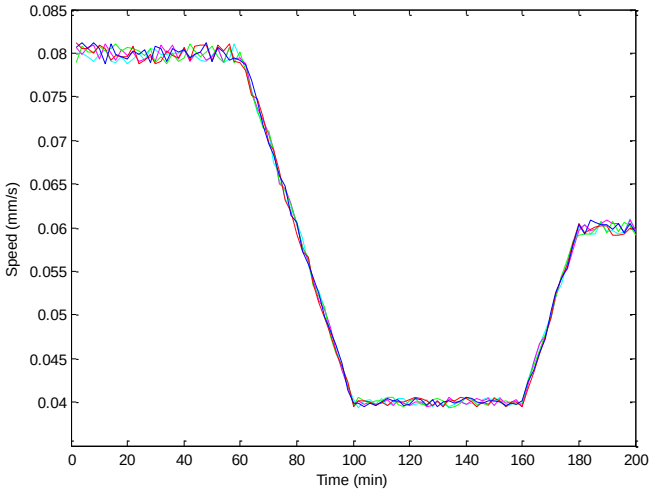

FIGURE 9. The speed of pressing-down at variant speed

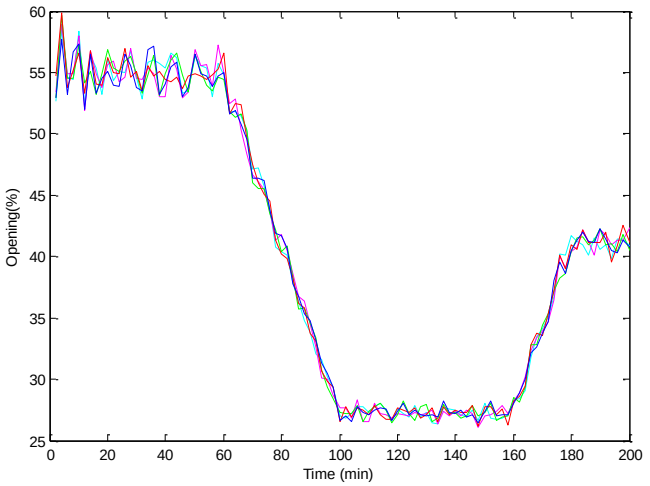

FIGURE 10. The output of control at variant speed

It is seen from figure 9 that the curves with different colors have the same tendency which achieves the reference speed under the different constant level and the changing speed period. During the interval from 1min to 30 mins, the maximum speed spikes are 0.0812 mm/s (at the 5th test) and 0.0788 mm/s (at the 1st, 3th, 4th and 5th tests) with relative

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3020550, IEEE Access

errors of 1.5%. The maximum peak speeds are 0.0406 mm/s (at the 3rd test) and 0.0394 mm/s (at the 1st, 2nd, 3th and 4th tests) during the interval between 50 mins and 80 mins, while the speeds vary between 0.0609 mm/s (max) and 0.0394 mm/s (min) during the interval from 80 mins to 100 mins. All the relative errors are less than 1.5%. Figure 10 shows the output under control with different colors at each test. The blue curve is taken for a further analysis based on the points representing the samples. The variance at the different intervals of 1-30 mins, 50-80 mins, and 90-100 are respectively 231.87, 20.5296, and 38.7686. The variance reduces as the reference speeds are down. The similar case happens on the other curves. One can find the reason from the working principle of the pressing-down process. The pressing-down speed is determined by the load resistance and the upper chamber pressure of the hydraulic cylinder. The upper chamber pressure is the rest of the pressure of the power sub-system taking away the pressure loss of servo valve (the pressure loss of the pipe is omitted because it is far less than that of the servo valve). On the other hand, the slide block is pressing down as a result of the space expansion of the upper chamber with the accumulation of hydraulic oil which can be controlled through the opening of the servo valve. Bigger is the opening of servo valve, less is the pressure loss of the servo valve, and more hydraulic oil will pump into the upper chamber of the hydraulic cylinder. This will widen the tuneable range and lead to a relatively easy control. The means and variance of the speed, and control output are shown in table IV.

TABLE IV
THE MEANS AND VARIANCE OF SPEED AND CONTROL OUTPUT AT VARIANT SPEEDS

		Speed		Output of control	
		Mean	Variance	Mean	Variance
1-30	Time1	0.0797	3.2192e-07	5.5143e+02	2.3664e+02
	Time2	0.0799	4.4986e-07	5.5083e+02	2.5443e+02
	Time3	0.0800	4.8825e-07	5.5070e+02	2.1223e+02
	Time4	0.0800	6.1320e-07	5.5076e+02	1.7084e+02
	Time5	0.0801	5.1593e-07	5.4841e+02	2.3087e+02
50-80	Time1	0.0399	1.2163e-07	2.7259e+02	24.9263
	Time2	0.0400	8.4246e-08	2.7215e+02	25.5312
	Time3	0.0400	1.2380e-07	2.7324e+02	30.3035
	Time4	0.0400	1.0079e-07	2.7278e+02	26.2325
	Time5	0.0400	8.4627e-08	2.7274e+02	20.5276
90-100	Time1	0.0598	3.5491e-07	4.0973e+02	35.0187
	Time2	0.0601	2.5978e-07	4.1319e+02	56.0678
	Time3	0.0598	4.1441e-07	4.1149e+02	35.8819
	Time4	0.0598	2.1885e-07	4.1205e+02	69.1209
	Time5	0.0601	2.7346e-07	4.1173e+02	38.7686

Then the conventional PID is used under test to control the speed of the pressing-down process. The neural network is abandoned here due to 1) lack of good training database; 2) it is an offline control strategy. The results are shown in figure 8. The red curve, the blue curve, and the green curve are results of the reference speed, the PID control, and the online control approach respectively.

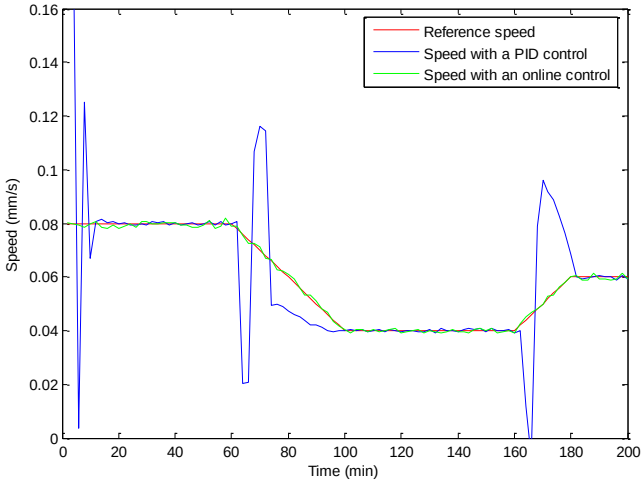

FIGURE 11. The results of pressing-down speed under different controls

It is seen from figure 11 that PID can achieve good control accuracy during the period from15 mins to 60 mins, from 100 mins to 160 mins, and from 180 mins to 200 mins when the speed is stable. The mean, the relative error and the variance at stable speeds are shown in table V. (The data of proposed approach is based on time3.)

Table V
THE MEAN, THE RELATIVE ERROR AND THE VARIANCE AT STABLE SPEEDS

		Mean	Variance	Relative error
The PID	1-30	**0.0799**	**2.2054e-007**	**0.0013**
The proposed approach	1-30	**0.0800**	**4.8825e-07**	**0**
The PID	50-80	**0.0401**	**1.9598e-007**	**0.0025**
The proposed approach	50-80	**0.0400**	**1.2380e-07**	**0**
The PID	90-100	**0.0607**	**6.8762e-006**	**0.0117**
The proposed approach	90-100	**0.0598**	**4.1441e-07**	**0.0033**

Table V shows both PID and proposed approach can provide a fine control with the relative error <3%. However figure 11 shows the PID has a worse performance during the transient process because it is difficult to get appropriate PID parameters. In contrast to the flawed PID control, the proposed online control shows a perfect effect throughout the whole process.

C. INFLUENCES OF SAMPLING PERIOD

In this subsection, sampling times are tested to show their effects on the speed under control. The sampling periods are chosen from 2 minutes (the minimun interval time for obtaining the right control) to 5 minutes (the maximum

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3020550, IEEE Access

interval time for the forging quality). The reference speed is set as 0.04mm/s. The RL selected a random action at the beginning and then go into the autonomous control according to *procedure 2*. Figure 12 shows the speed of the pressing-down during different sampling periods. Figure 13 shows the outputs of the controller during different sampling periods.

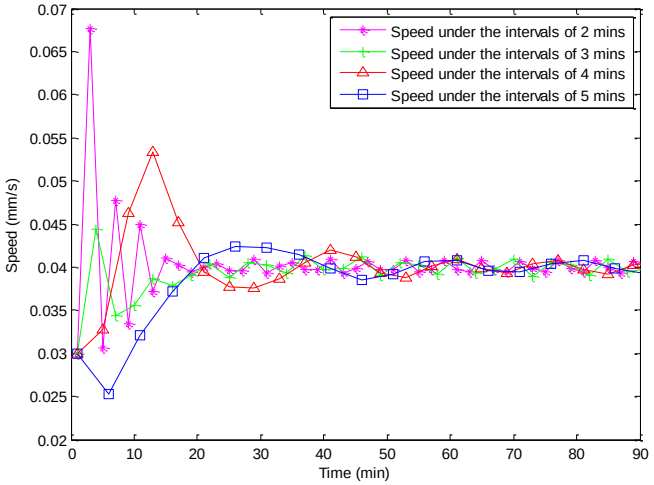

FIGURE 12. The speed of pressing-down during different sampling periods

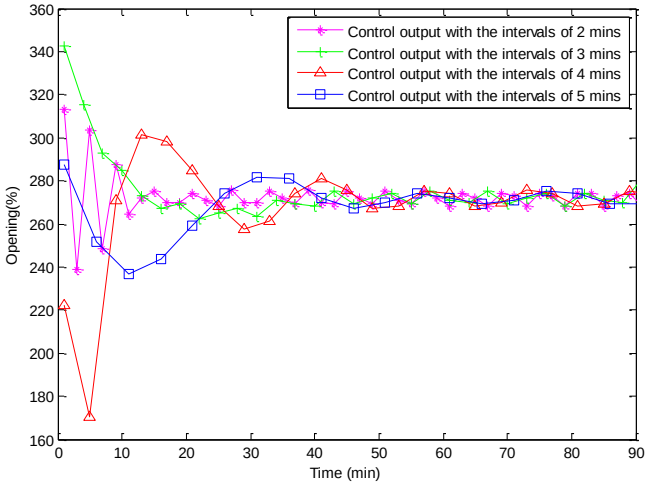

FIGURE 13. The output of control during different sampling periods

In figure 12, the pink curve, the green curve, the red curve and the blue curve represent the speed of the slide block at the sampling period of 2 minutes, 3 minutes, 4 minutes, and 5 minutes respectively. The stars, the crossings, the triangles, and the squares are the sample points. All four curves can approach to the reference speed (0.04mm/s) after a transient process. The mean and the variance in the stable process are shown in table VI. There are some differences in the transient process. The transient of speed1 (lasts about 18 minutes) is shorter than the others (about 25 minutes for the green curve, about 50 minutes for the red curve, and about 70 minutes for the blue curve). It is the reason that it can adjust the output of control in a shorter time which weakens the accumulative effects of forging machine for a longer period based on the previous moment.

TABLE VI
THE MEANS AND VARIANCE IN DIFFERENT SAMPLING PERIODS

	Sampling periods	Stable process	Mean	Variance
Speed1	2 min	18min-90min	0.0400	2.9115e-07
Speed2	3 min	25min-90min	0.0401	5.6880e-07
Speed3	4 min	50min-90min	0.0399	4.6431e-07
Speed4	5 min	70min-90min	0.0399	3.9806e-07

V. CONCLUSIONS

A data-driven online control strategy has been proposed for the control of the forging machine in order to deal with the difficulties in parameters adjustment of large batch change. This online-learning and online working algorithm has been carried out by reinforcement learning that can get the control only with two consecutive samples and the learning process is based on the computer simulation instead of trials and errors. The mapping space between the state and control has been reduced to a local space by developing the relationship between the states and controls according to the Lyapunov stability theory based on the coarse model, ensuring the system to be stable and preventing the system risk of out of control. The taboo search has been used to overcome the difficulty of the requirement of the historical data, which can find the control directly. Compared with the fine-parameters PID and well-trained NN controller the proposed approach can well realize the automatic control according to the current states, without the trouble of parameters adjustment that keeps tracing the working condition to get a good performance. The disadvantage is that taboo search would still spend the time to obtain an optimization, therefore the proposed approach can only be applied to the slow physical processes. The next step is to speed up the search to meet the need of the general real time control system.

REFERENCES

[1] https://www.forging.org/producers-and-suppliers/technology/vision-of-the-future#importance

[2] C. Jia, A. Wu, C. Du, and D. Zhang, "Variable structure control with sliding mode for a class of hydraulic nonlinear system", Proceedings of the World Congress on Intelligent Control and Automation (WCICA), p 3383-3387, 2010.

[3] T. Ho, and K. Ahn, "Speed control of a hydraulic pressure coupling drive using an adaptive fuzzy sliding-mode control", IEEE-ASME Trans Mech, vol.17, no.5, pp.976–986, 2012.

[4] C. Li, A.Wu, and C. Du, "Speed control of hydraulic press via adaptive back-stepping", Applied Mechanics and Materials, vol.40-41, pp.46-51, 2011.

[5] D. Zhang, A. Wu, G. Zhang, and C. Du, "Application of the differential geometric feedback linearization to the speed control of forging machine", 2010 Chinese Control and Decision Conference, pp.3185-3189, 2010.

[6] X. Lu and M. Huang, "System-Decomposition-Based Multilevel Control for Hydraulic Press Machine", IEEE Transactions on Industrial Electronics, vol. 59. No. 4, pp. 1980-1987, APR 2012.

[7] Y. Lee and R. Kopp, "Application of fuzzy control for a hydraulic forging machine", Fuzzy Sets and Systems, vol.118, no. 1, pp. 99-108, FEB 16 2001.

[8] X. Duan, H. Deng, and H. Li, "A Saturation-Based Tuning Method for Fuzzy PID Controller", IEEE Transactions on Industrial Electronics, vol. 60, no. 11, pp.5177-5185 , NOV 2013.

[9] M. Hawryluk, J. Ziemba, and P. Sadowski, "A Review of Current and New Measurement Techniques Used in Hot Die Forging Processes", Measurement & Control, vol. 50, no. 3, pp. 74-86, APR 2017.

[10] S. Yin, X. Li, and H. Gao, "Data-based techniques focused on modern industry: an overview", IEEE Transactions on Industrial Electronics, Vol. 62,No.1,pp.657-667, JAN 2015.

[11] W. Dai, T. Chai, and S. Yang, "Data-driven optimization control for safety operation of hematite grinding process", IEEE Transactions on Industrial Electronics, Vol.62, No.5, pp.2930-2941,MAY 2015.

[12] Y. Lin, D. Chen, M.S. Chen, X.M. Chen, and J. Li, "A precise BP neural network-based online model predictive control strategy for die forging hydraulic press machine", Neural Computing & Applications, vol. 29, no. 9, SI, pp. 585-596, MAY 2018.

[13] B. Fan, X. Lu, and M. Huang, "A novel LS-SVM control for unknown nonlinear systems with application to complex forging process", Journal of Central South University, vol. 24, no. 11, pp. 2524-2531, NOV 2017.

[14] X. Lu, C. Liu, and M. Huang, "Online Probabilistic Extreme Learning Machine for Distribution Modeling of Complex Batch Forging Processes", IEEE Transactions on Industrial Inforrmatics, vol. 11, no. 6, pp. 1277-1286, DEC 2015.

[15] M.Sedighi and M. Hadi, "Preform optimization for reduction of forging force using a combination of neural network and genetic algorithm", Proceedings of the Institution of Mechanical Engineers Part B-Journal of Engineering Manufacture, vol. 224, no. B11, pp.1717-1724, 2010.

[16] R. Sutton, and A. Barto, "Reinforcement Learning: An Introduction", The MIT Press Cambridge, Massachusetts. London, England. 2005.

[17] V. Mnih, K. Kavukcuoglu, D. Silver, *et. al*. "Human-level control through deep reinforcement learning", Nature, vol. 518, no. 7540, pp.529-533 , FEB 2015

[18] M.Z. Alom, T.M. Taha, C. Yakopcic, *et.al*. "A State-of-the-Art Survey on Deep Learning Theory and Architectures", Electronics, vol.8 , no.3 , DN:292, MAR 5 2019.

[19] J. Peng, Z. Zhang, and Q. Wu, "Anti-Jamming Communications in UAV Swarms: A Reinforcement Learning Approach", IEEE ACCESS, vol.7, pp.180532-180543, 2019.

[20] D. Zhang, and Z. Gao, "Improvement of Refrigeration Efficiency by Combining Reinforcement Learning with a Coarse Model", Processes, vol.7, no.12, Do: 967, DEC 2019.

[21] T. Zhang, and S. Mao, "Smart Power Control for Quality-Driven Multi-User Video Transmissions: A Deep Reinforcement Learning Approach", IEEE ACCESS, vol.8, pp. 611-622, 2020.

[22] L. Liu, Z. Wang, and H. Zhang, "Adaptive Fault-Tolerant Tracking Control for MIMO Discrete-Time Systems via Reinforcement Learning Algorithm With Less Learning Parameters", IEEE Transactions on Automation Science and Engineering, vol.14, no. 1, pp.299-313, JAN 2017

[23] D. Zhang, and Z. Gao, "Reinforcement learning-based fault-tolerant control with application to flux cored wire system", Measurement & Control, vol. 51, no. 7-8, pp.349-359, SEP-OCT 2018.

[24] Q. Wei, and D. Liu, "A Novel Iterative theta-Adaptive Dynamic Programming for Discrete-Time Nonlinear Systems", IEEE Transactions on Automation Science and Engineering, vol.11, no. 4, pp.1176-1190, OCT 2014

[25] D. Zhang, Z. Lin, and Z. Gao, "A Novel Fault Detection with Minimizing the Noise-Signal Ratio Using Reinforcement Learning", Sensors, vol.18, no. 9, Do: 3087, SEP 2018.

[26] J. Schulman, S. Levine, P. Moritz, M. Jordan, and P. Abbeel. "Trust region policy optimization", 32nd International Conference on Machine Learning, ICML 2015, v 3, p 1889-1897, 2015.

[27] D.Vrabie, O. Pastravanu, M. Abu-Khalaf, and F.L. Lewis, "Adaptive optimal control for continuous tiem linear systems based on policy iteration", Automatic, Vol.45, No. 2,pp. 477-484,FEB 2009.

[28] P. Jan, and S. Stefan, "Natural actor-critic", Neurocomputing, 71(7):1180–1190, 2008b.

[29] I. Szita, and A. Lorincz, "Learning tetris using the noisy cross-entropy method", Neural computation, 18(12): 2936–2941, 2006.

[30] F. Glover, and M. Laguna, "Tabu Search", Kluwer Academic Publishers,1997.

[31] D. Fouskakis, and D. Draper, "Stochastic optimization: a review", International Statistical Review, vol. 70, no. 3, pp.315-349, DEC 2002.