2023 62nd IEEE Conference on Decision and Control (CDC)
December 13-15, 2023. Marina Bay Sands, Singapore

Reinforcement Learning for Image-Based Visual Servo Control

Ashwin P. Dani, Shubhendu Bhasin

***Abstract*—In this paper, a continuous-time reinforcement learning (RL)-based controller is developed for image-based visual servoing (IBVS). The IBVS control dynamics is of the form where the drift term is absent and there is an uncertainty in the Jacobian matrix that is multiplied with the input. This poses a challenge for developing a continuous-time RL controller. The paper presents an actor-critic or synchronous policy iteration (PI)-based RL controller along with a parameter update law for the unknown parameter in the image Jacobian and proves closed-loop stability with the proposed controller. An infinite-horizon value function minimization objective is achieved by regulating the current image features to the desired with near-optimal control efforts. The proposed controller is tested using a simulation use case and the results validate the proposed theory.**

I. Introduction

Visual servo (VS) control uses image feedback to control the motion of a camera or a robot. There are several approaches to visual servoing: image-based visual servo control (IBVS), position-based visual servo control (PBVS), 2.5D approach, see [1], [2] for more details. PBVS requires knowledge of the camera pose and the error is computed between desired and current camera pose. IBVS uses image features to directly compute the camera velocities, thereby obviating the need to estimate pose. 2.5D visual servo control approaches use camera pose and image feature error. Other methods to visual servo control include partitioned approach [3], switched VS [4] that switches between IBVS and PBVS, and hybrid VS [5], which simultaneously minimizes the pose and image feature errors. Although several advanced methods have been developed for VS, few optimal control methods exist for VS, e.g., [6], [7], where an linear quadratic (LQ) controller is designed using the linearized system. The VS dynamics are nonlinear, where the drift dynamics term is not present and the control is multiplied by a Jacobian matrix which is not fully known due to the presence of uncertain parameters. Reinforcement learning (RL) has successfully provided a means to design optimal adaptive controllers for various classes of systems. The main objective of this paper is to design an RL-based controller for IBVS dynamics.

RL is a goal-oriented learning method, where the decision maker learns the optimal policy that maximizes a long term reward. By interacting with the environment the decision maker gets an evaluative feedback about its actions, which can be used to iteratively improve the control policy [8]. A class of iterative RL methods is adaptive dynamic programming (ADP), which was introduced by Werbos for discrete time (DT) systems [9], [10], and implemented in actor-critic framework. Extension of RL algorithms to continuous time (CT) systems is achieved in [11] by using Hamilton-Jacobi-Bellman (HJB) framework with known system dynamics, where a continuous-time version of the temporal difference error is employed. Several offline approaches of solving a generalized HJB equation are developed in [12], [13], using Galerkin's spectral approximation [12] and least-squares successive approximation solution [13] to HJB, which is then used to compute the optimal control.

To obviate the requirements of known system dynamics, among online approaches, an integral reinforcement learning (IRL) method is developed in [14], [15], which requires only partial knowledge of system dynamics. The approach called policy iteration (PI) is developed based on actor-critic structure, where the actor neural network (NN) is learned at a faster time scale than the critic NN. In [16] the IRL approach is extended to simultaneously learn both the actor and critic NNs, leading to a new method called synchronous PI. Further in [17], an actor-critic-identifier (ACI) approach is presented, which in addition to actor and critic NNs, uses an identifier network to identify the unknown drift term in the dynamics. The above mentioned methods require the knowledge of the input gain or control effectiveness matrix. The method in [18] identifies the complete nonlinear system dynamics using experience replay technique and learns the actor-critic NN using PI method for completely unknown dynamics. For linear systems, a completely model free RL method is developed in [19], which iteratively solves the algebraic Riccati equation using the online information of state and input. Using the IRL framework an on-policy model-free Q learning approach is developed in [20] for linear systems.

In this paper, an actor-critic RL algorithm is designed based on the development in [17] without the identifier part for the IBVS system. The uncertainty in the input gain matrix complicates the design of the RL algorithm for the IBVS setting, where the ACI approach of [17] is not directly applicable and therefore, motivates the contribution in this paper. The unknown depth parameters of the system dynamics (input gain/control effectiveness) is modeled as a constant parameter and an adaptive parameter update law is

A. P. Dani is with the Department of Electrical and Computer Engineering at University of Connecticut, Storrs, CT 06269. S. Bhasin is with the Indian Institute of Technology Delhi Email: ashwin.dani@uconn.edu, sbhasin@ee.iitd.ac.in

A.P. Dani was supported in part by a Space Technology Research Institutes grant (number 80NSSC19K1076) from NASA Space Technology Research Grants Program and in part by NSF grant no. SMA-2134367.

developed using Lyapunov-based stability analysis. The critic and actor NN approximate the optimal value function and optimal control. The critic NN weight update law is derived based on minimization of the Bellman error computed using optimal and approximate HJB equation. A least-squares weight update law is derived. Similarly, a gradient based NN weight update law is derived for actor NN based on minimization of Bellman error. The parameter and actor-critic weights are learnt simultaneously as new state and control input data becomes available. Lyapunov stability analysis shows an exponential convergence to an ultimate bound of the closed loop error system leading to uniformly ultimately bounded (UUB) stability. The proposed IBVS controller is validated using a simulation study.

II. System Model and Control Objective

A. System Dynamics

Consider the following system representing the evolution of the points as a function of the camera velocity. The system model can be written as

$$\dot{x} = J(x,\theta)v \tag{1}$$

where $x(t) = [x_1^T, ..., x_n^T]^T \in \mathbb{R}^{2n}$ denotes the set of feature points with $x_i(t) = [s_{xi}, s_{yi}]^T$ denoting the i^{th} feature point's position with respect to a fixed frame, $v(t) \in \mathbb{R}^m$ represents the input velocity vector, where m denotes the number of DOF. The Jacobian matrix $J(x,\theta) \in \mathbb{R}^{2n\times m}$ is expressed as [5]

$$\begin{aligned} J(x,\theta) &= [J_v(x)\theta \; J_\omega(x)] \\ &= Y_r(x)\theta + [\mathbf{0}_{2n\times\frac{m}{2}}, \; J_\omega(x)] \end{aligned} \tag{2}$$

where $\theta \in \mathbb{R}$ is an unknown parameter related to the depth and $J_v \in \mathbb{R}^{2n\times\frac{m}{2}}$ and $J_\omega \in \mathbb{R}^{2n\times\frac{m}{2}}$ and $Y_r = [J_v(x), \; \mathbf{0}_{2n\times\frac{m}{2}}] \in \mathbb{R}^{2n\times m}$.

Remark 1. *The parameter θ can be related to the unknown depth using Homography matrix decomposition, refer to [5] for further details.*

B. Controller Objective

The control objective is to regulate the current locations of the image features to the desired positions given by a reference image.

For the control design, the IBVS regulation error $\bar{x}(t) \in \mathbb{R}^{2n}$ is defined as

$$\bar{x}(t) \triangleq x(t) - x_d \tag{3}$$

where $x_d \in \mathbb{R}^{2n}$ are the desired locations of the feature points and the parameter estimation error $\tilde{\theta}(t) \in \mathbb{R}$ is defined as

$$\tilde{\theta}(t) \triangleq \theta - \hat{\theta}(t). \tag{4}$$

Since the optimal regulation objective is to bring the state $x(t)$ to a non-zero desired state x_d, the system model (1) is first written in terms of $\bar{x}$

$$\dot{\bar{x}} = J(x,\theta)v = J(\bar{x}, x_d, \theta)v \tag{5}$$

A continuous RL controller is now designed using the system in (5) with the objective to optimally regulate the state $\bar{x}(t)$ to 0 with the minimum control effort $v(t)$.

C. Continuous RL-based Controller Design

An RL-based controller is designed to achieve the desired control objective given by the optimal value function is defined as

$$V^*(x(t)) = \min_{u(\tau)\in\Theta(\mathcal{X})} \int_t^\infty r(s)ds \tag{6}$$

where $\Theta(\mathcal{X})$ is a set of admissible policies, $r(x,u) = Q(x) + u^T(x)Ru(x) \in \mathbb{R}$ is the local cost, $Q(x)$ is a positive definite function and $R = R^T > 0$. Without loss of generality, the system of the form $\dot{x} = J(x,\theta)u$ is considered for further development. Given the dynamics (5) and the value function (6), the optimal control is given by

$$u^*(x) = -\frac{1}{2}R^{-1}J^T(x,\theta)\frac{\partial V^{*\,T}}{\partial x} \tag{7}$$

where V^* is continuously differentiable and satisfies $V^*(0) = 0$.

The Hamiltonian of the system is given by

$$H(x,u,V_x) = \frac{\partial V}{\partial x}J(x,\theta)u + r(x,u) \tag{8}$$

The optimal Hamiltonian associated with the optimal cost and control is given by

$$H(x,u^*,V_x^*) = \frac{\partial V^*}{\partial x}J(x,\theta)u^* + r(x,u^*) = 0 \tag{9}$$

Computing the value function $V(x)$ and the optimal controller requires solution to the HJB which is a partial differential equation and hence hard to solve. The value function is approximated using a neural network called as a critic NN and the corresponding optimal control is approximated using actor NN. Using the approximated cost and the controller the approximated Hamiltonian is computed as

$$H(x,\hat{u},\hat{V}_x) = \frac{\partial \hat{V}}{\partial x}J(x,\hat{\theta})\hat{u} + r(x,\hat{u}) \tag{10}$$

Using the optimal and approximated Hamiltonian, a temporal difference or Bellman error δ is computed as follows

$$H(x,\hat{u},\hat{V}_x) - H(x,u^*,V_x^*) = \delta \tag{11}$$

$$\delta = \frac{\partial \hat{V}}{\partial x}J(x,\hat{\theta})\hat{u} + r(x,\hat{u}) \tag{12}$$

because the value of optimal Hamiltonian is 0. Bellman error, δ, in Hamiltonian is used to learn the critic and actor NN weights.

Assumption 1. *The function $J(x,\theta)$ is second order differentiable. Moreover, J is bounded $0 < J(x,\theta) < \bar{j}$ with a known bound.*

Assumption 2. *For a given NN, $L(x) = W^T\sigma(V^Tx) + \epsilon(x)$, the ideal NN weights W and V are bounded by known*

positive constants, i.e., $\|W\| \leq \bar{W}$, $\|V\| \leq \bar{V}$ [21]. The NN activation function σ and σ' are bounded.

Assumption 3. *Using the universal approximation property of NN, the function reconstruction error $\|\epsilon(x)\| \leq \bar{\epsilon}$ and its derivative $\|\epsilon(x)'\| \leq \bar{\epsilon}'$ are bounded [22].*

III. Actor-Critic based Optimal Control Design

A. Approximate Optimal Control

Using NN representation the optimal value function and the optimal control is written as

$$V^*(x(t)) = W_c^T \phi(x) + \epsilon_c(x)$$
$$u^*(x) = -\frac{1}{2} R^{-1} J^T(x,\theta)(\phi'(x)^T W_c + \epsilon_c'(x)^T) \quad (13)$$

where $W_c \in \mathbb{R}^{n_c \times 1}$, $\phi(x) : \mathbb{R}^{2n} \to \mathbb{R}^{n_c}$ are the basis functions, $\epsilon(x) \in \mathbb{R}$ is the function error. Due to the function approximation error these cannot be implemented in practice, for that the approximated value function and the optimal control laws are designed as

$$\hat{V}(x(t)) = \hat{W}_c^T \phi(x)$$
$$\hat{u}(x) = -\frac{1}{2} R^{-1} J^T(x,\hat{\theta})(\phi'(x)^T \hat{W}_a) \quad (14)$$

where $\hat{W}_c \in \mathbb{R}^{n_c \times 1}$ and $\hat{W}_a \in \mathbb{R}^{n_a \times 1}$ are the estimated critic and actor weights, along with the parameter update law for $\hat{\theta}(t)$ defined as

$$\dot{\hat{\theta}} = \text{proj}\left(-\gamma_\theta \hat{u}^T Y_r^T \phi'^T \hat{W}_c - \sigma\hat{\theta}\right) \quad (15)$$

where (2) is used and γ_θ, σ are constant gains and proj is a smooth projection operator to keep the parameter estimates bounded [23].

Assumption 4. *The parameter estimate $\hat{\theta}(t) \neq 0$ for any time t.*

Remark 2. *Assumption 4 is made to avoid loss of control or to ensure $\dot{x} \neq 0$ for non-zero velocity control u and can be guaranteed by a smooth projection operator that ensures that the estimates are projected to a region away from zero. For IBVS, $\hat{\theta}(t) = 0$ would imply that the feature points are infinitely far away from the camera.*

Remark 3. *A similar parameter update law for $\theta \in \mathbb{R}^p$ can be derived for $p > 1$.*

Let the actor-critic NN approximation errors are defined as $\tilde{W}_c = W_c - \hat{W}_c(t)$ and $\tilde{W}_a = W_a - \hat{W}_a(t)$. The actor and critic NN weights are updated using weight update laws that minimize the error between approximated Hamiltonian and the optimal one, given by Bellman error. The Bellman error in a measurable form in terms of actor and critic NN weights is written as

$$\delta = \hat{W}_c^T \phi'(x) J(x,\hat{\theta})\hat{u} + r(x,\hat{u}) \quad (16)$$

For the analysis, another form of Bellman error based on (11) is derived as follows

$$\delta = \hat{W}_c^T \phi'(x) J(x,\hat{\theta})\hat{u} + \hat{u}^T R \hat{u}$$
$$- W_c^T \phi'(x) J(x,\theta) u^* - u^{*T} R u^* - \epsilon_c J(x,\theta) u^* \quad (17)$$

which can be simplified to

$$\delta = -\tilde{W}_c^T \phi'(x) J(x,\hat{\theta})\hat{u} - W_c^T \phi'(x)\tilde{J}(x,\theta)\tilde{u}$$
$$- \epsilon_c' J(x,\theta) u^* + \frac{1}{4}\tilde{W}_a^T \phi' J(x,\theta) R^{-1} J^T(x,\theta) \phi'^T \tilde{W}_a$$
$$- \frac{1}{2}\tilde{W}_a^T \phi' J R^{-1} J^T \phi'^T W_c - \frac{1}{4}\epsilon_c' J R^{-1} J^T \epsilon_c'^T$$
$$- \frac{1}{2}\epsilon_c'^T J R^{-1} J^T \phi'^T W_c \quad (18)$$

where $\hat{a}R\hat{a} - a^{*T}Ra^* = \tilde{a}^T R \tilde{a} - 2\tilde{a}^T R a^*$ is used for $-u^{*T}Ru^* + \hat{u}^T R \hat{u}$.

B. Critic NN Weight Update Laws

The least-squares update law for the critic NN can be derived by minimizing the integral Bellman error $E_{cr} = \int_0^t \frac{1}{2}\delta^2(\tau)d\tau$. Taking the time derivative of E_{cr} with respect to $\hat{W}_c$

$$\frac{\partial E_{cr}}{\partial \hat{W}_c} = 2\int_0^t \delta \frac{\partial \delta}{\partial \hat{W}_c} d\tau = 0 \quad (19)$$

where $\frac{\partial \delta}{\partial \hat{W}_c} = \hat{u}^T J^T(x,\theta)\phi'(x)^T = w^T$

$$\frac{\partial E_{cr}}{\partial \hat{W}_c} = 2\int_0^t \hat{W}_c^T w w^T + r(x,\hat{u}) w^T d\tau = 0 \quad (20)$$

The least squares solution can be derived as

$$\dot{\hat{W}}_c = \text{proj}\left(-\gamma_c \Gamma \frac{w}{1 + c_1 w^T \Gamma w}\delta\right) \quad (21)$$

where $c_1 \in \mathbb{R}$ and $\gamma_c \in \mathbb{R}$ are constant gains, $\Gamma(t) = (\int_0^t w(\tau)w(\tau)^T d\tau)^{-1} \in \mathbb{R}^{n \times n}$ is a symmetric estimation gain matrix, which is computed using the differential equation

$$\dot{\Gamma} = -\gamma \Gamma \frac{w w^T}{1 + c_1 w^T \Gamma w}\Gamma, \qquad \Gamma(0) = c_2 I \quad (22)$$

for positive constants γ and c_2.

Assumption 5. *The normalized critic regressor $\xi = \frac{w}{\sqrt{1+c_1 w^T \Gamma w}}$ is bounded and is persistently exciting (PE), i.e.,*

$$\mu_a I \leq \int_{t_0}^{t_0+T} \xi(\tau)\xi^T(\tau)d\tau \leq \mu_b I, \quad \forall t_0 > 0 \quad (23)$$

where μ_a, μ_b, T are positive constants. [17]

C. Actor NN Weight Update Laws

The least-squares gradient-based update law for the actor NN is derived using the squared Bellman error $E_a = \delta^2$. Computing the gradient of E_a, we get

$$\frac{\partial E_a}{\partial \hat{W}_a} = 2\frac{\partial \delta}{\partial \hat{W}_a}\delta = 2\left(\hat{W}_c^T\phi'(x)J(x,\theta)\frac{\partial \hat{u}}{\partial \hat{W}_a} + 2\hat{u}^T R\frac{\partial \hat{u}}{\partial \hat{W}_a}\right)\delta \quad (24)$$

Setting the above inequality to 0, the gradient weight update law for $\hat{W}_a$ can be derived as

$$\dot{\hat{W}}_a = \text{proj}\left(-\frac{2\gamma_a}{\sqrt{1+w^Tw}}\left(\hat{W}_c^T\phi'(x)J(x,\theta)\frac{\partial \hat{u}}{\partial \hat{W}_a}\right)^T\delta - \frac{4\gamma_a}{\sqrt{1+w^Tw}}\frac{\partial \hat{u}}{\partial \hat{W}_a}^T R\hat{u}\delta - \gamma_{a2}(\hat{W}_a - \hat{W}_c)\right) \quad (25)$$

where γ_a and γ_{a2} are constant gains.

IV. Stability Analysis

Theorem 1. *Given that the Assumptions 1-5 hold and the following sufficient condition is satisfied $\eta_2 > 0$,[1] the actor-critic controller (14)-(15) along with the weight update laws in (21)-(25) guarantee that the signals $x(t)$, $\tilde{W}_a(t)$ and $\tilde{W}_c(t)$ are uniformly ultimately bounded.*

Proof. Consider a positive definite continuously differentiable Lyapunov function $V : \mathcal{X}\times\mathbb{R}^{n_c}\times\mathbb{R}^{n_a}\times\mathbb{R}\times[0,\infty) \to \mathbb{R}^+$

$$V_z(x,\tilde{W}_c,\tilde{W}_a,t) = V^*(t) + V_{wc}(t,\tilde{W}_c) + \frac{1}{2}\tilde{W}_a^T\tilde{W}_a + \frac{1}{2\gamma_\theta}\tilde{\theta}^T\tilde{\theta} \quad (26)$$

where $V^*(t)$ is the optimal value function, V_{wc} is a Lyapunov function corresponding to $\tilde{W}_c$ defined further in the proof, γ_θ is a constant gain. Since the optimal value function $V^*(t)$ is continuously differentiable and positive definite, there exists class $\mathcal{K}$ functions such that $\alpha_1(\|x\|) \leq V^*(t) \leq \alpha_2(\|x\|) \quad \forall x \in \mathcal{B}_a \subset \mathcal{X}$. Let us define $z(t) = [x(t)^T, \tilde{W}_c(t)^T, \tilde{W}_a(t)^T, \tilde{\theta}]^T \in \mathbb{R}^{n+n_c+n_a+1}$. Based on the bounds on $V^*(t)$, the following bounds can be derived

$$\alpha_3(\|z\|) \leq V_z(x,\tilde{W}_c,\tilde{W}_a,t) \leq \alpha_4(\|z\|) \quad \forall x \in \mathcal{B}_z \quad (27)$$

where $\alpha_3(\cdot)$ and $\alpha_4(\cdot)$ are the class $\mathcal{K}$ functions. Taking the time derivative of the Lyapunov function

$$\dot{V}(x,\tilde{W}_c,\tilde{W}_a,t) = \frac{\partial V^*}{\partial x}J(x,\hat{\theta})\hat{u} + \dot{V}_{wc} + \tilde{W}_a^T\dot{\tilde{W}}_a + \frac{1}{\gamma_\theta}\tilde{\theta}^T\dot{\tilde{\theta}} \quad (28)$$

[1] η_2 is defined in the proof.

Utilizing linear-in-the-parameter property of the dynamics $J(x,\theta)$ we get

$$\dot{V}(x,\tilde{W}_c,\tilde{W}_a,t) = \frac{\partial V^*}{\partial x}J(x,\theta)\hat{u} - \frac{\partial V^*}{\partial x}Y_r(x)\hat{u}\tilde{\theta} + \dot{V}_{wc} + \tilde{W}_a^T\dot{\tilde{W}}_a + \frac{1}{\gamma_\theta}\tilde{\theta}^T\dot{\tilde{\theta}} \quad (29)$$

Using $\dot{\tilde{W}}_a = -\dot{\hat{W}}_a$, and $\dot{\tilde{\theta}} = -\dot{\hat{\theta}}$, we get

$$\dot{V}(x,\tilde{W}_c,\tilde{W}_a,t) = \frac{\partial V^*}{\partial x}J(x,\theta)\hat{u} - \tilde{W}_a^T\dot{\hat{W}}_a + \dot{V}_{wc} - \frac{1}{\gamma_\theta}\tilde{\theta}^T\dot{\hat{\theta}} - \frac{\partial V^*}{\partial x}Y_r(x)\hat{u}\tilde{\theta} \quad (30)$$

Adding and subtracting $\frac{\partial V^*}{\partial x}J(x,\theta)u^*$ term yields

$$\dot{V}(x,\tilde{W}_c,\tilde{W}_a,t) = \frac{\partial V^*}{\partial x}J(x,\theta)u^* - \frac{\partial V^*}{\partial x}J(x,\theta)(u^*-\hat{u}) + \dot{V}_{wc} - \tilde{W}_a^T\dot{\hat{W}}_a - \frac{\partial V^*}{\partial x}Y_r(x)\hat{u}\tilde{\theta} - \frac{1}{\gamma_\theta}\tilde{\theta}^T\dot{\hat{\theta}} \quad (31)$$

Utilizing $\frac{\partial V^*}{\partial x}J(x,\theta) = -2u^{*T}R$ and $\frac{\partial V^*}{\partial x}J(x,\theta)u^* = -Q(x) - u^{*T}Ru^*$, $\dot{V}$ becomes

$$\dot{V}(x,\tilde{W}_c,\tilde{W}_a,t) = -Q(x) - u^{*T}Ru^* + 2u^{*T}R(u^*-\hat{u}) + \dot{V}_{wc} - \tilde{W}_a^T\dot{\hat{W}}_a - \frac{\partial V^*}{\partial x}Y_r(x)\hat{u}\tilde{\theta} - \frac{1}{\gamma_\theta}\tilde{\theta}^T\dot{\hat{\theta}} \quad (32)$$

Substituting u^* from (13), $\hat{u}$ from (14) and simplifying few terms yields

$$\dot{V} = -Q(x) + \frac{1}{4}\left(\frac{\partial V^*}{\partial x}J(x,\theta)R^{-1}J(x,\theta)^T\frac{\partial V^*}{\partial x}\right) - \frac{1}{2}\frac{\partial V^*}{\partial x}J(x,\theta)R^{-1}J^T(x,\theta)\phi'^T\hat{W}_a + \dot{V}_{wc} - \tilde{W}_a^T\dot{\hat{W}}_a - \frac{\partial V^*}{\partial x}Y_r(x)\hat{u}\tilde{\theta} - \frac{1}{\gamma_\theta}\tilde{\theta}^T\dot{\hat{\theta}} \quad (33)$$

Substituting the NN form of the V^* from (13) we get

$$\dot{V} = -Q(x) + \frac{1}{4}\epsilon_c'J_r\epsilon_c'^T - \frac{1}{4}W_c^TJ_\phi W_c - \frac{1}{4}\epsilon_c'J_rW_c + \frac{1}{2}W_c^TJ_\phi\tilde{W}_a + \frac{1}{2}\epsilon_c'J_r\phi'^T\tilde{W}_a + \dot{V}_{wc} - \tilde{W}_a^T\dot{\hat{W}}_a - \hat{W}_c^T\phi'Y_r(x)\hat{u}\tilde{\theta} - \tilde{W}_c^T\phi'Y_r(x)\hat{u}\tilde{\theta} - \epsilon_c'^T\phi'Y_r(x)\hat{u}\tilde{\theta} + \tilde{\theta}^T\hat{u}^TY_r^T\phi'^T\hat{W}_c - \sigma_1\tilde{\theta}^T\tilde{\theta} + \sigma_1\tilde{\theta}^T\theta \quad (34)$$

where $J_r = J(x,\theta)R^{-1}J(x,\theta)^T$, $J_\phi = \phi'J(x,\theta)R^{-1}J(x,\theta)^T\phi'^T$ and $\sigma_1 = \sigma/\gamma_\theta$.

To further simplify the $\dot{V}$ expression (34) consider $-\tilde{W}_a^T\dot{\hat{W}}_a$ term after substituting actor weight update law from (25)

$$-\tilde{W}_a^T\dot{\hat{W}}_a = -C_1\tilde{W}_a^T\left(\hat{W}_c^TJ_\phi\right)^T\delta + C_1\tilde{W}_a^T\left(J_\phi\hat{W}_a\right)\delta + C_2\tilde{W}_a^T(\hat{W}_a - \hat{W}_c) \quad (35)$$

where $C_1 = -\frac{2\gamma_a}{\sqrt{1+w^Tw}}$ and $C_2 = \gamma_{a2}$. The term in (35) can be written as

$$-\tilde{W}_a^T\dot{\hat{W}}_a = C_1\tilde{W}_a^T\left(J_\phi(\hat{W}_a - \hat{W}_c)\right)\delta - C_2\tilde{W}_a^T\tilde{W}_a + C_2\tilde{W}_a^T(W_c - \hat{W}_c) \quad (36)$$

Consider $\dot{\hat{W}}_c$ term after substituting δ from (18)

$$\dot{\hat{W}}_c = -\gamma_c \Omega w^T \tilde{W}_c + \gamma_c \Omega \left(-W_c^T \phi' \tilde{J} \tilde{u} + \frac{1}{4} \tilde{W}_a^T J_\phi \tilde{W}_a - \frac{1}{2} \tilde{W}_a^T J_\phi W_c - \epsilon_c' J u^* - \frac{1}{4} \epsilon_c' J_r \epsilon_c'^T - \frac{1}{2} \epsilon_c'^T J_r \phi'^T W_c \right) \tag{37}$$

where $\Omega = \Gamma \frac{w}{1+c_1 w^T \Gamma w}$. Under Assumption 5, a nominal system formed using first term of (37) is globally exponentially stable [17], [24], which according to converse Lyapunov Theorem induces a Lyapunov function $V_{wc}(t, \tilde{W}_c)$ with following properties $\gamma_1 \|\tilde{W}_c\|^2 \leq V_{wc}(\tilde{W}_c, t) \leq \gamma_2 \|\tilde{W}_c\|^2$, $\frac{\partial V_{wc}}{\partial t} + \frac{\partial V_{wc}}{\partial \tilde{W}_{wc}}(-\gamma_c \Omega w^T \tilde{W}_c) \leq -\eta_1 \|\tilde{W}_c\|^2$ and $\|\frac{\partial V_{wc}}{\partial \tilde{W}_c}\| \leq \bar{\gamma} \|\tilde{W}_c\|$ where $\gamma_1, \gamma_2, \eta_1, \bar{\gamma}$ are positive constants. Substituting (36) and the bounds on Lyapunov function V_{wc} into (34), $\dot{V}$ can be written as

$$\begin{aligned}
\dot{V} = & -Q(x) - \eta_1 \|\tilde{W}_c\|^2 - C_2 \|\tilde{W}_a\|^2 + \frac{1}{4} \epsilon_c' J_r \epsilon_c'^T \\
& - \frac{1}{4} W_c^T J_\phi W_c - \frac{1}{4} \epsilon_c' J_r W_c + \frac{1}{2} W_c^T J_\phi \tilde{W}_a + \frac{1}{2} \epsilon_c' J_r \phi'^T \tilde{W}_a \\
& + \bar{\gamma} \|\tilde{W}_c\| \left(- W_c^T \phi'(x) \tilde{J} \tilde{u} + \frac{1}{4} \tilde{W}_a^T J_\phi \tilde{W}_a - \frac{1}{2} \tilde{W}_a^T J_\phi W_c \right. \\
& \left. - \epsilon_c' J(x, \theta) u^* - \frac{1}{4} \epsilon_c' J_r \epsilon_c'^T - \frac{1}{2} \epsilon_c'^T J_r \phi'^T W_c \right) \\
& + C_1 \tilde{W}_a^T \left(J_r \phi'^T (\hat{W}_a - \hat{W}_c) \right) \delta - C_2 \tilde{W}_a^T \tilde{W}_a \\
& + C_2 \tilde{W}_a^T (W_c - \hat{W}_c) - \sigma_1 \|\tilde{\theta}\|^2 + \sigma_1 \tilde{\theta}^T \theta
\end{aligned} \tag{38}$$

Let

$$\|\frac{1}{4} \epsilon_c' J_r \epsilon_c'^T - \frac{1}{4} W_c^T J_\phi W_c - \frac{1}{4} \epsilon_c' J_r W_c\| \leq \kappa_1 \tag{39}$$

$$\|\frac{1}{2} W_c^T J_\phi\| \leq \kappa_2, \quad \|\frac{1}{2} \epsilon_c' J_r \phi'^T\| \leq \kappa_3 \tag{40}$$

$$\|\frac{1}{4} \tilde{W}_a^T J_\phi \tilde{W}_a\| \leq \kappa_4, \quad \overline{w} = \left\| J_r \phi'^T (\hat{W}_a - \hat{W}_c) \right\| \tag{41}$$

$$\| - W_c^T \phi' \tilde{J} \tilde{u} - \epsilon_c' J u^* - \frac{1}{4} \epsilon_c' J_r \epsilon_c'^T - \frac{1}{2} \epsilon_c'^T J_r \phi'^T W_c\| \leq \kappa_5 \tag{42}$$

Utilizing the bounds in (39)-(42), $\dot{V}$ can be upper bounded as

$$\begin{aligned}
\dot{V} \leq & -Q(x) - \eta_1 \|\tilde{W}_c\|^2 - 2C_2 \|\tilde{W}_a\|^2 + \kappa_1 \\
& + (\kappa_2 + \kappa_3) \|\tilde{W}_a\| + \bar{\gamma} \kappa_5 \|\tilde{W}_c\| + \frac{1}{4} \bar{\gamma} \kappa_4 \|\tilde{W}_c\| \\
& + \frac{1}{2} \bar{\gamma} \kappa_2 \|\tilde{W}_c\| \tilde{W}_a\| - C_1 \tilde{W}_a^T \left(J_\phi (\hat{W}_a - \hat{W}_c) \right) \delta \\
& + C_2 \|\tilde{W}_a\| \|\tilde{W}_c\| - \sigma_1 \|\tilde{\theta}\|^2 + \sigma_1 \|\tilde{\theta}\| \|\theta\|
\end{aligned} \tag{43}$$

which can further be simplified to

$$\begin{aligned}
\dot{V} \leq & -Q(x) - \eta_1 \|\tilde{W}_c\|^2 - 2C_2 \|\tilde{W}_a\|^2 - \sigma_1 \|\tilde{\theta}\|^2 + \kappa_1 \\
& + (\kappa_2 + \kappa_3) \|\tilde{W}_a\| + \bar{\gamma} \kappa_5 \|\tilde{W}_c\| + \frac{1}{4} \bar{\gamma} \kappa_4 \|\tilde{W}_c\| \\
& + \frac{1}{2} \bar{\gamma} \kappa_2 \|\tilde{W}_c\| \tilde{W}_a\| + C_1 \kappa_5 \overline{w}_j \left\| \tilde{W}_a \right\| + C_1 \kappa_4 \overline{w}_j \left\| \tilde{W}_a \right\|^2 \\
& + C_1 \kappa_2 \overline{w}_j \left\| \tilde{W}_a \right\| + C_2 \|\tilde{W}_a\| \|\tilde{W}_c\| + \sigma_1 \|\tilde{\theta}\| \|\theta\|
\end{aligned} \tag{44}$$

Completing the squares, we get

$$\begin{aligned}
\dot{V} \leq & -Q(x) - (1 - \theta_1) \eta_1 \|\tilde{W}_c\|^2 \\
& - (1 - \theta_1) \left(2C_2 - C_1 \kappa_4 \overline{w}_j \right) \|\tilde{W}_a\|^2 - (1 - \theta_3) \sigma_1 \|\tilde{\theta}\|^2 \\
& + \kappa_1 + (\kappa_2 + \kappa_3 + C_1 (\kappa_5 + \kappa_2) \overline{w}_j) \|\tilde{W}_a\| \\
& + \left(\bar{\gamma} \kappa_5 + \frac{1}{4} \bar{\gamma} \kappa_4 \right) \|\tilde{W}_c\| + \frac{\sigma_1 \|\theta\|^2}{4\theta_3}
\end{aligned} \tag{45}$$

where $1 - \theta_1 > 0$, $1 - \theta_3 > 0$, $\eta_2 = 2C_2 - C_1 \kappa_4 \overline{w}_j$ and the final bound on $\dot{V}$ can derived as

$$\begin{aligned}
\dot{V} \leq & -Q(x) - (1 - \theta_1 - \theta_2) \eta_1 \|\tilde{W}_c\|^2 \\
& - (1 - \theta_1 - \theta_2) \eta_2 \|\tilde{W}_a\|^2 - (1 - \theta_3) \sigma_1 \|\tilde{\theta}\|^2 \\
& + \kappa_1 + \frac{(\bar{\gamma} \kappa_5 + \frac{1}{4} \bar{\gamma} \kappa_4)^2}{4(1 - \theta_1 - \theta_2) \eta_1} \\
& + \frac{(\kappa_2 + \kappa_3 + C_1 (\kappa_5 + \kappa_2) \overline{w}_j)^2}{4(1 - \theta_1 - \theta_2) \eta_2} + \frac{\sigma_1 \|\theta\|^2}{4\theta_3}
\end{aligned} \tag{46}$$

where $1 - \theta_1 - \theta_2 > 0$. Since $Q(x)$ is positive definite, Lemma 4.3 of [25] can be utilized to derive

$$\begin{aligned}
\alpha_5(\|z\|) \leq & Q + (1 - \theta_1 - \theta_2) \eta_1 \|\tilde{W}_c\|^2 + (1 - \theta_3) \sigma_1 \|\tilde{\theta}\|^2 \\
& + (1 - \theta_1 - \theta_2) \eta_2 \|\tilde{W}_a\|^2 \leq \alpha_6(\|z\|)
\end{aligned} \tag{47}$$

Using (47), the expression (48) can be upper bounded as

$$\begin{aligned}
\dot{V} \leq & -\alpha_5(\|z\|) + \kappa_1 + \frac{(\bar{\gamma} \kappa_5 + \frac{1}{4} \bar{\gamma} \kappa_4)^2}{4(1 - \theta_1 - \theta_2) \eta_1} \\
& + \frac{(\kappa_2 + \kappa_3 + C_1 (\kappa_5 + \kappa_2) \overline{w}_j)^2}{4(1 - \theta_1 - \theta_2) \eta_2} + \frac{\sigma_1 \bar{\theta}}{4\theta_3}
\end{aligned} \tag{48}$$

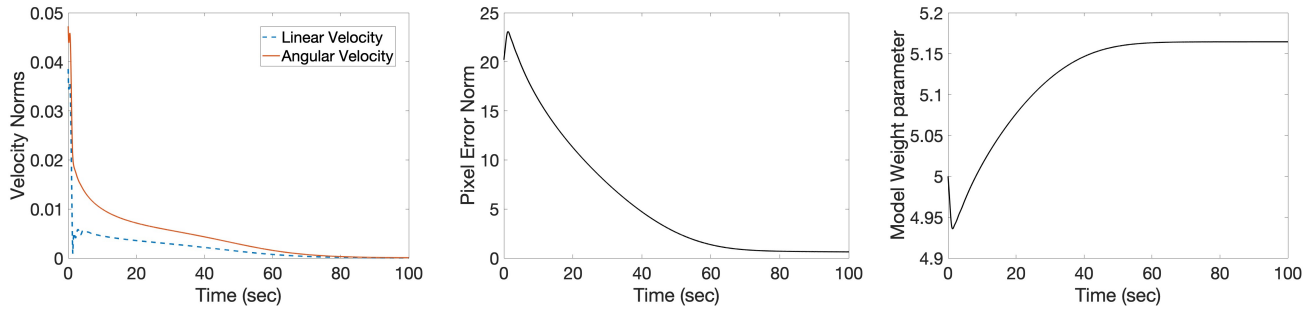
which proves that $\dot{V}$ is always negative whenever $z(t)$ is outside the compact set

$$\begin{aligned}
\bar{\Omega} = \{ z : \|z\| \leq \alpha_5^{-1} (& \kappa_1 + \frac{(\bar{\gamma} \kappa_5 + \frac{1}{4} \bar{\gamma} \kappa_4)^2}{4(1 - \theta_1 - \theta_2) \eta_1} \\
& + \frac{(\kappa_2 + \kappa_3 + C_1 (\kappa_5 + \kappa_2) \overline{w}_j)^2}{4(1 - \theta_1 - \theta_2) \eta_2} + \frac{\sigma_1 \bar{\theta}}{4\theta_3}) \}.
\end{aligned} \tag{49}$$

Invoking Theorem 4.18 of [25] $\|z\|$ is uniformly ultimately bounded (UUB). □

V. Simulation

Simulations are carried out to test the performance of the proposed RL-based IBVS control law. Camera and image frame motion is simulated using MATLAB, where four points were selected on the image plane with the initial pixel values of $[50\ 50; 100\ 50; 100\ 100; 50\ 100]^T$ and the desired pixel values of $[825\ 790; 860\ 825; 825\ 860; 790\ 825]^T$. The controller parameters are selected as $Q = 1000 \mathbb{I}_{8\times8}$ and $R = 800 \mathbb{I}_{6\times6}$. The basis functions for approximating the value function V is selected to be polynomial combinations of elements of $\phi = [\bar{x}_1, \bar{x}_2, .. \bar{x}_n]^T$, the initial weights for the critic NN are set to $W_c(t_0) = 3 \mathbb{I}_{36\times36}$. The parameters of critic NN are found by empirical tuning as $\gamma_c = 0.0006$, $\Gamma = 20$ and $c_1 = 100$. The actor NN weights are initialized



(a) (b) (c)

Figure 1. Results of simulation 1: (a) control velocity, (b) pixel error norm, (c) model parameter weights.

to $W_a(t_0) = 3.5\mathbb{I}_{36\times 36}$. The parameters of actor NN are selected as $\gamma_1 = 0.01$ and $\gamma_{a2} = 0.05$. Fig, 1(a) shows the norm of camera linear and angular velocities generated by the RL-IBVS controller. The velocities are bounded and are generated in an optimal manner based on the minimization of the value function. The image pixel error norms are shown in Fig. 1(b). The model parameter estimation $\hat{\theta}(t)$ is shown in Fig. 1(c), which is generated using the parameter estimation update law. The model parameter is bounded and converges to a value, not necessarily the true parameter. However, the RL-IBVS controller drives the error to a small ball around zero in the presence of uncertainties in the image Jacobian.

VI. Conclusion

In this paper, a reinforcement learning based IBVS controller is developed using the RL method for continuous time systems. The IBVS control dynamics is of the form where the uncertainty is present in the Jacobian matrix with unknown depth parameter. The RL controller is developed based on a continuous-time version of the PI architecture, and a proof of stability is provided using Lyapunov stability analysis along with the proposed parameter update law.

References

[1] F. Chaumette and S. Hutchinson, "Visual servo control. I. basic approaches," *IEEE Robotics & Automation Magazine*, vol. 13, no. 4, pp. 82–90, 2006.

[2] ——, "Visual servo control, part ii: Advanced approaches," *IEEE Robotics and Automation Magazine*, vol. 14, no. 1, pp. 109–118, 2007.

[3] P. I. Corke and S. A. Hutchinson, "A new partitioned approach to image-based visual servo control," *IEEE Transactions on Robotics and Automation*, vol. 17, no. 4, pp. 507–515, 2001.

[4] N. R. Gans and S. A. Hutchinson, "Stable visual servoing through hybrid switched-system control," *IEEE Transactions on Robotics*, vol. 23, no. 3, pp. 530–540, 2007.

[5] N. R. Gans, G. Hu, J. Shen, Y. Zhang, and W. E. Dixon, "Adaptive visual servo control to simultaneously stabilize image and pose error," *Mechatronics*, vol. 22, no. 4, pp. 410–422, 2012.

[6] G. Chesi and Y. S. Hung, "Global path-planning for constrained and optimal visual servoing," *IEEE Transactions on Robotics*, vol. 23, no. 5, pp. 1050–1060, 2007.

[7] K. Hashimoto and H. Kimura, "LQ optimal and nonlinear approaches to visual servoing," in *Visual Servoing: Real-Time Control of Robot Manipulators Based on Visual Sensory Feedback*. World Scientific, 1993, pp. 165–198.

[8] B. Kiumarsi, K. G. Vamvoudakis, H. Modares, and F. L. Lewis, "Optimal and autonomous control using reinforcement learning: A survey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 6, pp. 2042–2062, 2017.

[9] W. T. Miller, R. S. Sutton, and P. J. Werbos, "A menu of designs for reinforcement learning over time," in *Neural Networks for Control*. MIT press, 1995.

[10] P. Werbos, "Approximate dynamic programming for real-time control and neural modeling," *Handbook of intelligent control*, 1992.

[11] K. Doya, "Reinforcement learning in continuous time and space," *Neural computation*, vol. 12, no. 1, pp. 219–245, 2000.

[12] R. W. Beard, G. N. Saridis, and J. T. Wen, "Galerkin approximations of the generalized hamilton-jacobi-bellman equation," *Automatica*, vol. 33, no. 12, pp. 2159–2177, 1997.

[13] M. Abu-Khalaf and F. L. Lewis, "Nearly optimal control laws for nonlinear systems with saturating actuators using a neural network HJB approach," *Automatica*, vol. 41, no. 5, pp. 779–791, 2005.

[14] D. Vrabie and F. Lewis, "Neural network approach to continuous-time direct adaptive optimal control for partially unknown nonlinear systems," *Neural Networks*, vol. 22, no. 3, pp. 237–246, 2009.

[15] D. Vrabie, O. Pastravanu, M. Abu-Khalaf, and F. L. Lewis, "Adaptive optimal control for continuous-time linear systems based on policy iteration," *Automatica*, vol. 45, no. 2, pp. 477–484, 2009.

[16] K. G. Vamvoudakis and F. L. Lewis, "Online actor–critic algorithm to solve the continuous-time infinite horizon optimal control problem," *Automatica*, vol. 46, no. 5, pp. 878–888, 2010.

[17] S. Bhasin, R. Kamalapurkar, M. Johnson, K. G. Vamvoudakis, F. L. Lewis, and W. E. Dixon, "A novel actor–critic–identifier architecture for approximate optimal control of uncertain nonlinear systems," *Automatica*, vol. 49, no. 1, pp. 82–92, 2013.

[18] H. Modares, F. L. Lewis, and M.-B. Naghibi-Sistani, "Adaptive optimal control of unknown constrained-input systems using policy iteration and neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 24, no. 10, pp. 1513–1525, 2013.

[19] Y. Jiang and Z.-P. Jiang, "Computational adaptive optimal control for continuous-time linear systems with completely unknown dynamics," *Automatica*, vol. 48, no. 10, pp. 2699–2704, 2012.

[20] K. G. Vamvoudakis, "Q-learning for continuous-time linear systems: A model-free infinite horizon optimal control approach," *Systems & Control Letters*, vol. 100, pp. 14–20, 2017.

[21] F. L. Lewis, J. Campos, and R. Selmic, *Neuro-fuzzy control of industrial systems with actuator nonlinearities*. SIAM, 2002.

[22] G. Cybenko, "Approximation by superpositions of a sigmoidal function," *Mathematics of control, signals and systems*, vol. 2, no. 4, pp. 303–314, 1989.

[23] W. E. Dixon, A. Behal, D. M. Dawson, and S. Nagarkatti, *Nonlinear Control of Engineering Systems: A Lyapunov-Based Approach*. Birkhäuser: Boston, 2003.

[24] S. Sastry, M. Bodson, and J. F. Bartram, "Adaptive control: stability, convergence, and robustness," 1990.

[25] H. K. Khalil, *Nonlinear Systems*, 3rd ed. Prentice Hall, 2002.