ISSN 1975-8359 [Print] / ISSN 2287-4364 [Online]
The Transactions of the Korean Institute of Electrical Engineers, vol. 72, no. 12, pp. 1705~1713, 2023
https://doi.org/10.5370/KIEE.2023.72.12.1705

Sim-to-Real Reinforcement Learning Techniques for Double Inverted Pendulum Control with Recovery Property

Sim-to-Real 강화학습 기법을 활용한 Recovery 특성을 갖는 2단 도립진자 제어

Taegun Lee · Doyoon Ju · Young Sam Lee
이태건* · 주도윤* · 이영삼†

Abstract

In recent years with the rapid advancement of artificial intelligence, there has been extensive research to address control problems, which was previously unsolvable with traditional control techniques, using reinforcement learning-based controllers. This paper discusses a challenge in controlling a double inverted pendulum system. With the commonly used 2-DOF control technique, once the swing-up control is performed and a strong disturbance is applied, the system becomes uncontrollable and fails to perform another swing-up. However, the reinforcement learning-based controller proposed in this paper overcomes this limitation using the Sim-to-Real learning technique. To ensure successful application of Sim-to-Real learning, this paper proposes a design method for the real-world system that minimizes the reality gap, a chronic issue with the Sim-to-Real technique. Utilizing these techniques, we introduce a characteristic termed 'recovery property' denoting the ability to recover from strong disturbances, a feature difficult to achieve with traditional control methods. We design a controller with this characteristic and validate its successful operation in a real-world system.

Key Words

Reinforcement Learning, Double Inverted Pendulum, Sim-to-Real Learning, Recovery Property

1. 서 론

도립진자 시스템은 제어공학적인 측면에서 불안정한 동특성과 비선형 모델 방정식, 그리고 비최소 위상이라는 난도 높은 특성을 모두 함유하는 시스템이다. 이러한 특성 때문에 해당 시스템은 오랜 기간 다양한 제어 기법의 성능을 평가하기 위한 테스트베드로써 널리 사용되어 왔다. 도립진자 시스템을 활용한 주요 연구 분야는 진자를 도립시키기 위한 swing-up 제어이며, 이를 해결하기 위해 다양한 제어 기법을 사용한 연구가 활발히 진행되었다[1-2]. 특히 진자의 단수가 증가한 형태인 2단 도립진자의 경우에는 swing-up 제어 문제 자체의 난도가 높아 2007년에 와서야 Graichen에 의해 2자유도 구조의 효과적인 swing-up 제어 기법이 제시되었다[3].

또한, 최근 인공지능 기술의 급격한 발전에 따라 심층 신경망을 활용한 강화학습을 제어공학 분야에 적용하는 연구가 활발히 진행되고 있다[4]. 강화학습은 에이전트가 관측한 환경의 상태 정보에 따라 자신의 행동 정책에 기반하여 행동을 수행하고, 그로 인해 변화한 환경으로부터 얻어지는 보상이 최대가 되도록 자신의 행동 정책을 반복적으로 개선하여 학습하는 기법이다. 강화학습 기반의 제어기는 주어진 시스템의 상태 정보를 입력으로 받아 학습된 행동 정책에 따른 최적의 행동, 즉 제어량을 출력하게 된다. 이러한 융합적인 연구 분야에 있어서도 여전히 도립진자 및 다단 도립진자는 인공지능 기반 제어기의 효용성을 검증하기 위한 테스트베드로서 활용되고 있다. 기존의 전통적인 제어기를 강화학습 기반의 제어기로 대체하여 앞서 언급한 swing-up 문제를 해결하거나[5-6], 혹은 새롭게 제안되는 인공지능 학습 기법의 성능을 검증하기 위한 연구 등의 다양한 분야에서 활발히 사용되고 있다[7-8].

하지만 강화학습에서 학습의 주체가 되는 에이전트가 실물 시스템과 직접 상호작용하며 학습을 진행하는 경우, 몇가지 문제점이 발생한다. 여기엔 상호작용에 요구되는 물리적인 시간의 소요와 데이터 획득 비용의 증가, 그리고 실험 중 발생할 수 있는 물리적인 위험 등의 요소들이 포함된다[9]. 상기된 문제점들로 인해 강화학습을 이용하여 제어기를 설계하는 연

† Corresponding Author : Dept. of Electrical and Computer Engineering, Inha University Incheon, Korea
E-mail:lys@inha.ac.kr
https://orcid.org/0000-0003-0665-1464
* Dept. of Electrical and Computer Engineering, Inha University, Incheon, Korea
https://orcid.org/0009-0007-3107-2735
https://orcid.org/0000-0001-7011-6779
Received: Oct. 12, 2023 **Revised:** Nov. 21, 2023 **Accepted**: Nov. 24, 2023

구는 주로 실물 시스템의 동특성을 묘사하는 시뮬레이션 환경을 구축하고, 이를 바탕으로 강화학습 알고리즘을 적용하는 방식의 실험을 통해 이루어지고 있다[10]. 이렇게 시뮬레이션 환경에서 학습이 이루어지고, 학습이 완료된 후 이를 실제 시스템에 적용하는 방식을 Sim-to-Real 학습 기법이라고 통칭한다.

그러나 Sim-to-Real 학습 기법에는 한 가지 큰 문제점이 존재하는데, 이는 시뮬레이션과 실물 시스템 간에는 항상 간극, 즉 현실 격차(reality gap)가 존재한다는 것이다. 두 환경 간의 현실 격차 크기에 따라 시뮬레이션에서 학습한 모델이 실제 시스템에서 원활하게 동작하지 않거나, 동작의 성능 저하 문제가 발생할 수 있다[11]. 이를 극복하기 위해 본 논문에서는 저자들이 속한 연구실에서 오랜기간 연구했던 도립진자 시스템에 대한 제어공학 및 기구학적 지식을 바탕으로, 실제 시스템을 시뮬레이션 환경에 사용되는 모델과 정합성이 우수하도록 설계하여 이 격차를 최소화 한다. 해당 시스템에 Sim-to-Real 학습 기법을 적용할 경우 현실 격차로 인한 성능 저하의 걱정 없이, 물리적인 제약으로부터 자유로운 시뮬레이션 환경을 활용해 폭넓은 데이터를 취득하고 학습할 수 있다. 이러한 방식으로 학습된 강화학습 기반 제어기는 기존의 전통적인 제어 기법으로는 도달하기 어려웠던 새로운 제어 방식의 구현이 가능해진다. 이를 통해 전통적인 제어 기법으로는 해결할 수 없던 문제를 강화학습 기반의 제어기로 해결할 수 있는 가능성이 제시된다.

이를 뒷받침 하기 위해 본 논문은 2단 도립진자의 swing-up 제어에서 가장 대표적으로 사용되는 2자유도 제어 기법[3]으로는 불가능한 제어 동작을 구현하는 것을 목표로 한다. 2장에서는 앞서 언급한 제어기법의 한계와 이를 Sim-to-Real 학습 기법으로 극복할 수 있는 방안에 대하여 구체적으로 서술한다. 이어지는 3장에서는 Sim-to-Real 학습을 위해 현실 격차를 최소화 하는 2단 도립진자 시스템의 설계 구조를 제안한다. 이후 4장에서 실험 및 결과를 기술하고, 이를 바탕으로 5장에서 결론을 다루는 구성을 갖는다.

2. Recovery 특성을 갖는 강화학습 기반의 제어기

서론에서 언급된 Graichen이 제시한 제어 기법은 오프라인 최적화를 통해 2단 도립진자의 swing-up 궤적을 미리 계산하여 이를 앞먹임(feedforward) 형태로 시스템에 인가하고, 해당 궤적과의 오차를 되먹임(feedback) 제어를 통해 보정하는 방식으로 이루어진다. 이러한 2자유도 제어 기법을 통해 2단 도립진자의 레일 길이 제약을 고려하면서도 swing-up 제어를 성공적으로 수행하는 제어 방식을 도입하였다. 2013년에는 다른 연구자가 동일한 기법을 이용하여 구조적으로 더 높은 난도를 갖는 3단 도립진자의 swing-up 제어를 성공적으로 수행함으로써 해당 제어 방식의 우수성을 다시 한번 검증하였다[12].

하지만 해당 제어 기법에는 치명적인 단점이 존재한다. 이는 강한 외란이 인가될 경우 제어가 불가능한 상태에 이르게 된다는 것이다. 일정 수준의 외란에 대해서는 되먹임 제어의 보정을 통해 강건성을 갖는 모습을 보여주지만, 일정 수준 이상의 강한 외란을 인가하게 될 경우 시스템이 미리 구해두었던 선행 궤적과 아예 궤가 달라지며 앞먹임 제어가 무의미해지게 된다. 이는 되먹임 제어로도 보정할 수 없는 상태가 되어 결국 제어 불능 상태에 이르게 되는 것이다. 이런 상태에 이르게 될 경우, 기존의 2자유도 제어 기법으로는 다시 swing-up 동작을 할 수 없게 된다. 선행 궤적은 오프라인 상황에서 미리 산출되는 값이기 때문에, 시스템이 동작하는 도중에는 다시 궤적을 구할 수 없기 때문이다.

본 연구는 Sim-to-Real 학습 기법을 활용하여 이러한 문제점을 해결하였다. 강화학습 에이전트는 환경과 상호작용하며 자신이 경험해본 상태 정보와 그 당시의 보상에 기반하여 행동 정책을 개선한다. 이 과정을 반복하여 최대한 많은 상태 정보를 축적하고, 각 상태마다 최선의 제어량을 도출하는 수준까지 학습을 진행한다. 해당 시점까지 학습된 제어기는 어떠한 상태에 도달해도 원하는 제어를 수행할 수 있게 되는 것이다.

이는 마치 미로 찾기 문제와 동일하게 생각할 수 있다. 목적지가 고정되어 있는 미로가 존재할 때, 학습의 시작 지점을 미로의 무작위한 곳으로 배치시키고 목적지를 탐색하도록 학습을 반복한다면, 학습이 완료된 이후에는 어떤 지점에서 탐색을 시작하더라도 바로 목적지를 찾아 갈 수 있는 것과 같다. 마찬가지로, 도립진자 또한 어떤 상태에 도달하더라도 swing-up 제어를 수행할 수 있게 되는 것이다. 도립진자에 강한 외란이 인가되었을 경우에도 강화학습 에이전트는 이를 단순히 환경의 상태 정보가 변화했다고 인식한 뒤, 해당 시점에 알맞은 제어량을 출력하는 방식으로 swing-up 제어를 수행한다.

상기된 방식의 학습이 이루어지기 위해서는 Sim-to-Real 학습 기법이 필수적으로 요구된다. 앞서 비유를 들었던 방식의 학습을 위해선 강화학습 에이전트가 최대한 다양한 상태를 경험하는 것이 요구되는데, 실물 시스템만을 사용한 환경에서는 물리적인 제약조건이 존재하기 때문이다. 현실의 실물 시스템에서는 중력에 의해 모든 진자가 바닥을 향한 상태 외에는, 연구자가 각 진자들의 각도와 각속도를 임의로 초기화 할 수 없다. 이로 인해 강화학습 에이전트가 경험할 수 있는 상태의 범위에는 한계가 생기고, 경험해보지 못한 상태에 대해서는 학습이 이루어지지 않는다. 따라서 모든 상황에 대응할 수 없게 되며, 이는 결국 외란이 인가된 상황에도 완벽하게 대처할 수 없게 되는 결과를 야기한다.

하지만 시뮬레이션으로 구성된 환경은 그러한 물리적인 제약으로부터 자유로워진다. 이러한 환경에서는 매번 시뮬레이션이 시작할 때 마다 두 진자의 각도와 각속도, 나아가 대차의 위치와 가속도까지. 상태 정보에 해당하는 모든 값을 연구자가 임의로 설정할 수 있다. 이러한 환경의 특성을 활용하게 되면, 강화학습 에이전트가 실물 시스템에서는 한번도 겪지

못했을법한 상황에 대해서도 시뮬레이션에서는 자유롭게 학습을 진행할 수 있게 된다. 이러한 기법을 통해 강화학습 에이전트는 광범위한 상태 정보를 축적하고, 그에 대한 행동을 학습하는 과정이 매우 용이해진다. 이를 통해 강한 외란이 인가된 상황을 맞이하더라도 이미 시뮬레이션 상에서 경험했던 상태 정보에 해당할 확률이 높기 때문에, 제어 불능 상태에 빠지지 않고 성공적인 제어를 수행할 수 있게 되는 것이다. 본 논문에서는 이러한 특성을 갖는 강화학습 기반의 제어기가 'Recovery 특성'을 갖는다고 명명한다. 이는 기존의 제어기에서는 강한 외란을 인가했을 시 불안정한 상태로 천이되어 제어가 불능해지는데 반해, Sim-to-Real 기법을 활용해 구현된 제어기는 불안정한 상태에 이른 뒤에도 다시 안정한 상태로 '회복'할 수 있는 특성을 가지기 때문이다.

하지만 상기된 Sim-to-Real 학습 기법의 독특한 이점을 제대로 활용하기 위해서는, 서론에서 언급된 현실 격차를 최소화시킬 수 있도록 실제 시스템의 모델 정합성이 우수하다는 전제가 강력하게 요구된다. 시뮬레이션 환경에서 학습이 완벽하게 이루어졌다고 하더라도, 실제 시스템에서는 동적 특성이 다르게 나타난다면 이는 전혀 효용성이 없는 제어기가 되기 때문이다. 이를 위해 후술될 3장에서는 실제 시스템의 수학적 모델 방정식을 구하고, 그와 정합성이 우수하도록 기구적 구조를 설계하는 방안을 제안한다.

3. 모델 정합성이 높은 2단 도립진자 구조

3.1 2단 도립진자의 수학적 모델방정식

그림 1은 실험에 사용된 2단 도립진자의 기구적 개념도를 나타내며, 그림에서 사용되는 변수들은 SI 단위계를 사용함을 가정하고, 세부적인 의미는 다음과 같다. M은 대차(cart)의 질량, m_1, m_2는 각각 1단 진자와 2단 진자의 질량을 의미하며 l_1, l_2는 각각 1단 진자와 2단 진자의 회전축으로부터 무게 중심까지의 길이를 나타낸다. θ_1은 1단 진자의 회전 변위로써 지면에 대한 법선과 이루는 각이며, θ_2는 2단 진자가 1단 진자와 이루는 상대적인 회전변위를 나타내고, L_1은 1단 진자의 회전축부터 2단 진자의 회전축까지의 길이를 의미한다. 그리고 c_1과 c_2는 1단 진자와 2단 진자의 회전축에서 발생하는 마찰계수를 의미하며, y는 대차의 초기위치로부터의 변위, u는 대차의 가속도를 나타낸다. 또한, i,j,k는 레일의 중심점을 원점으로 한 직각좌표계의 각 좌표축을 의미한다.

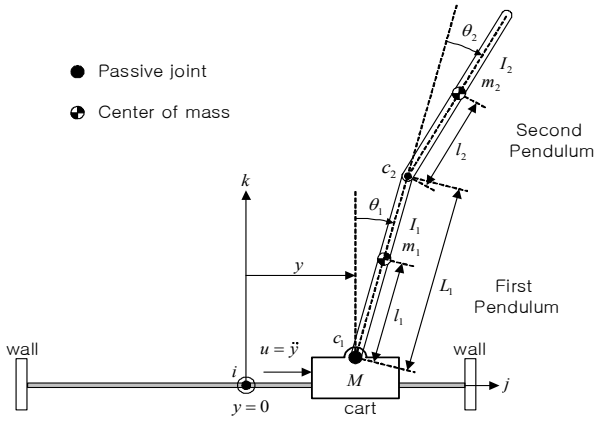


그림 1 2단 도립진자 기구적 개념도
Fig. 1 Mechanical conceptual diagram of a double inverted pendulum

2단 도립진자의 수학적 모델은 Euler-Lagrange equation을 이용하여 유도하면 다음과 같이 식 (1)로 나타낼 수 있다.

$$\begin{bmatrix} n_1 \\ n_2 \end{bmatrix} \ddot{y} + \begin{bmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{bmatrix} \begin{bmatrix} \ddot{\theta}_1 \\ \ddot{\theta}_2 \end{bmatrix} + \begin{bmatrix} r_1 \\ r_2 \end{bmatrix} = 0. \tag{1}$$

위 식의 각 요소는 식(2)와 같다.

$$\begin{aligned} n_1 &= h_1\cos(\theta_1) + h_2\cos(\theta_1+\theta_2), \\ n_2 &= h_2\cos(\theta_1+\theta_2), \\ m_{11} &= h_3 + h_6 + 2h_4\cos(\theta_2), \\ m_{12} &= h_6 + h_4\cos(\theta_2), \\ m_{21} &= h_6 + h_4\cos(\theta_2), \\ m_{22} &= h_6, \\ r_1 &= -h_4\sin(\theta_2)(2\dot{\theta}_1\dot{\theta}_2 + \dot{\theta}_2^2) - h_5\sin\theta_1 \\ &\quad - h_7\sin(\theta_1+\theta_2) + c_1\dot{\theta}_1, \\ r_2 &= h_4\sin(\theta_2)\dot{\theta}_1^2 - h_7\sin(\theta_1+\theta_2) + c_2\dot{\theta}_2. \end{aligned} \tag{2}$$

$h_1 \sim h_7$은 식 (3)의 형태로 정의되고, 여기서 g는 중력가속도 9.81[m/s^2]를 나타낸다.

$$\begin{aligned} h_1 &= m_1 l_1 + m_2 L_1, \\ h_2 &= m_2 l_2, \\ h_3 &= I_1 + m_1 l_1^2 + m_2 L_1^2, \\ h_4 &= m_2 L_1 l_2, \\ h_5 &= g(m_1 l_1 + m_2 L_{1)}, \\ h_6 &= I_2 + m_2 l_2^2, \\ h_7 &= g m_2 l_2. \end{aligned} \tag{3}$$

식 (1)을 재배열 하면 식 (4)의 형태로 다시 표기할 수 있고,

$$\begin{bmatrix} \ddot{\theta}_1 \\ \ddot{\theta}_2 \end{bmatrix} = -\begin{bmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{bmatrix}^{-1} \left\{ \begin{bmatrix} n_1 \\ n_2 \end{bmatrix} \ddot{y} + \begin{bmatrix} r_1 \\ r_2 \end{bmatrix} \right\}. \tag{4}$$

위 식을 전개하게 되면 식 (5)로 표현할 수 있다. 이때, 식 (5)에서 상태 벡터를 $x_1 = y$, $x_2 = \theta_1$, $x_3 = \theta_2$, $x_4 = \dot{y}$, $x_5 = \dot{\theta}_1$, $x_6 = \dot{\theta}_2$로 정의하고 $\ddot{y}$을 가속도 u로 나타내면, 최종적으로 2단 도립진자의 모델방정식은 식 (6)과 같은 비선형 상태방정식으로 나타낼 수 있다.

$$\ddot{\theta}_1 = \frac{(-m_{22}n_1 + m_{12}n_2)\ddot{y} + (-m_{22}r_1 + m_{12}r_2)}{\Phi}, \quad (5)$$
$$\ddot{\theta}_2 = \frac{(m_{21}n_1 - m_{11}n_2)\ddot{y} + (m_{21}r_1 - m_{11}r_2)}{\Phi},$$
$$\Phi = m_{11}m_{22} - m_{12}m_{21}.$$

$$\underbrace{\begin{bmatrix}\dot{x}_1\\ \dot{x}_2\\ \dot{x}_3\\ \dot{x}_4\\ \dot{x}_5\\ \dot{x}_6\end{bmatrix}}_{\dot{x}} = \underbrace{\begin{bmatrix} x_4 \\ x_5 \\ x_6 \\ u \\ \dfrac{(-m_{22}n_1 + m_{12}n_2)u + (-m_{22}r_1 + m_{12}r_2)}{\Phi} \\ \dfrac{(m_{21}n_1 - m_{11}n_2)u + (m_{21}r_1 - m_{11}r_2)}{\Phi} \end{bmatrix}}_{f(x,u)} \quad (6)$$

상기된 모델방정식에서 1단 진자와 2단 진자는 각 중심점에서 i축 방향의 회전축을 중심으로 하는 회전만이 존재한다는 것을 가정한다. 또한 대차는 j축 방향의 수평운동만이 발생할 수 있고, 그 이외의 수평운동과 회전운동은 발생하지 않는 것을 가정한다. 후술 될 2절에서의 기구부 구조는 위 가정에 최대한 부합할 수 있도록 설계함으로써 모델의 정합성을 최대화 시키는 것을 목적으로 한다.

3.2 2단 도립진자의 기구부 및 구동부

Sim-to-Real 학습에서 가장 중요한 것은 Sim에 해당하는 시뮬레이션 환경을 구현하는데 사용되는 수학적 모델 방정식과 Real에 해당하는 실제 진자 시스템의 동적 특성 간 정합성이 우수하도록 설계해야 한다는 것이다. 둘 간의 정합성이 좋지 않은 경우에는, 시뮬레이션 상에서의 학습이 성공적이더라도 실물 시스템에서 그 성능을 제대로 내지 못할 가능성이 매우 높아지기 때문이다. 실제 시스템을 모델 방정식과의 정합성이 우수하도록 설계하기 위해서는, 실제 시스템의 동작이 모델 방정식에서 사용된 가정과 부합하는 움직임만을 갖도록 설계해야한다. 모델 방정식에서 고려하지 않은 요소가 발생하는 경우, 시뮬레이션 환경과 실물 시스템의 동적 응답간에 차이가 발생하기 때문이다.

저자들이 속한 연구실에서는 오랜 기간 다양한 도립진자 시스템을 직접 제작하며 모델 방정식과 정합성이 우수한 기구적 구조를 제안한 바 있다[13]. 본 논문에서는 해당 구조에서 더 개선된 형태의 기구부와 구동부 구조를 설계함으로써 모델과 실제 시스템의 응답 정합성을 향상시키고, 이를 통해 시뮬레이션과 실제 시스템의 현실 격차를 줄일 수 있는 방안을 제시한다. 제안되는 2단 도립진자 시스템의 기구적 구조는 그림 2에 보이는 바와 같다.

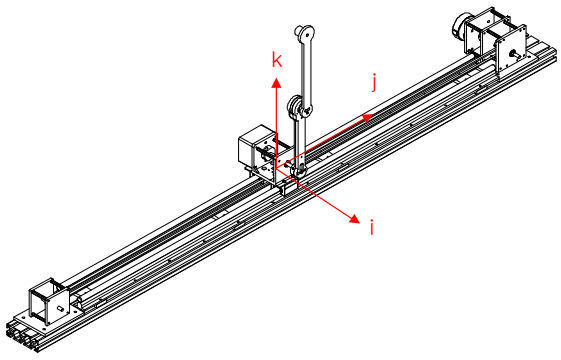


그림 2 2단 도립진자 기구적 구조
Fig. 2 The mechanical structure of a double inverted pendulum

3.2.1 구동부 설계

그림 3은 기존 [13]에서 제안했던 구조로, 풀리와 구동부가 결합된 형태를 나타낸다. 해당 구조는 감속기를 사용하지 않은 BLDC 모터를 이용하여 직접 풀리를 구동함으로써 백래시를 제거하고, 이를 통해 백래시로 인한 limit cycle 문제를 해결한 형태이다. 또한, 풀리에 장착된 타이밍 벨트의 장력을 극복하기 위해 2개의 베어링을 사용해 2중으로 지지하여 벨트의 장력이 풀리를 관통하는 축에만 전달되게 함으로써 모터에 가해지는 부하를 제거하도록 설계되었다.

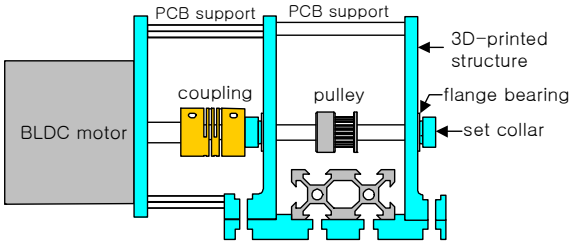


그림 3 3D 프린트 구조물을 사용한 구동부 구조
Fig. 3 Driving structure using 3D printed framework

하지만 해당 구조는 전체적인 구동부를 감싸고 있는 소재가 3D 프린터에서 사용되는 PLA 소재로, 상대적으로 낮은 강도로 인해 파손이 발생하거나 변형된다는 문제점이 발생하였다. 이를 방지하기 위해 그림 4와 같이 해당 구조물 전체를 강성

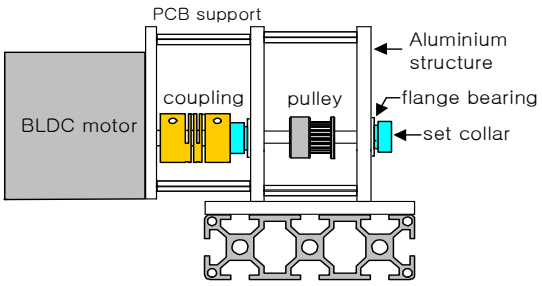


그림 4 알루미늄 합판 구조물을 사용한 구동부 구조
Fig. 4 Driving structure using an aluminum composite panel

이 높은 알루미늄 합금 소재(알루미늄 6061) 판으로 대체하여 구조물의 손상 및 변형을 제거함으로써 모델 방정식에서 고려하지 않은 요소가 발생할 가능성을 배제하였다.

3.2.2 대차 및 레일부 설계

대차의 병진운동을 위한 도립진자의 레일 구조를 그림 5에 보이는 V-slot형 2040 프로파일에서, 그림 6에 나타난 2개의 선형 가이드 레일을 사용하는 구조로 개선하였다. 그림 5의 구조에서는 2단 도립진자의 swing-up 제어를 위해 대차에 힘을 인가하면, 대차에 결합된 프로파일이 α만큼의 각도로 비틀림을 겪게 된다. 이러한 레일의 비틀림 각도로 인해 진자가 그림 1에서의 j축을 중심으로 회전하게 되는데, 이는 앞서 서술 했던 모델 방정식의 가정에 전혀 부합하지 않는 요소로 작용하게 된다. 해당 구조를 그림 6와 같이 선형 가이드 레일로 변경한 구조에서는 상기된 α와 같은 레일의 비틀림 요소가 전혀 발생하지 않게 된다. 더불어, 가이드 레일에 결합된 대차 또한 j축 방향의 수평 운동만이 발생하게 되고, 비틀림에 의한 그 외의 수평운동과 회전운동을 제거하여 3.1절에서 서술한 모델 방정식과의 정합성을 크게 향상시킨다.

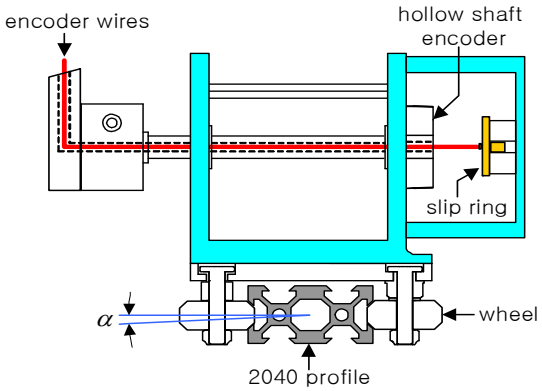


그림 5 2040 알루미늄 프로파일을 이용한 레일 및 대차 구조
Fig. 5 The structure of the rail and cart constructed using 2040 aluminum profile

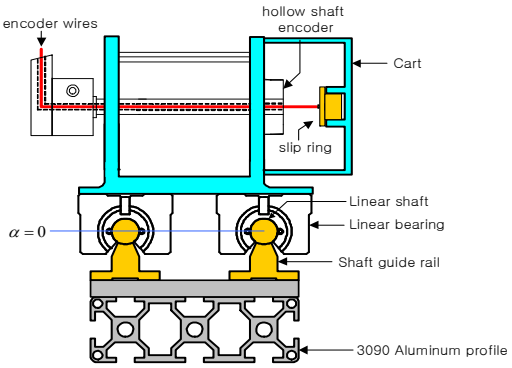


그림 6 3090 알루미늄 프로파일과 이중 선형 가이드 레일을 이용한 레일 및 대차 구조
Fig. 6 The structure of the rail and cart using 3090 aluminum profile and dual linear guide rails

이외에도 회전 조인트에 복렬 베어링 구조를 사용하여 진자의 i축 중심 회전을 강제시키고, 베어링의 고체 상태 윤활제를 제거하여 대차의 정지 마찰과 쿨롱 마찰을 저감시키는 등의 추가적인 설계가 사용되었다. 해당 내용들은 참고문헌 [13]에서 기술된 바와 동일하기에 본 논문에서는 자세히 다루지 않기로 한다.

4. 실험 및 결과

본 장에서는 앞서 서술한 모델 방정식과 해당 모델에 정합성이 우수한 2단 도립진자 시스템을 이용하여, Sim-to-Real 기법을 활용한 Recovery 특성을 갖는 강화학습 기반 제어기를 설계하는 실험을 진행한다. 이러한 강화학습 기반 제어기를 설계하기 위한 개발 및 실험 환경으로는 저자가 이전에 작성한 문헌 [14]에서 사용한 환경을 일부 변형하여 사용하였다. 이때 강화학습 에이전트가 학습을 위해 직접적으로 상호작용하는 환경은 3장에서 서술한 수학적 모델을 바탕으로 Python 상에 시뮬레이션으로 구현하였다. 해당 시뮬레이션 환경을 구축하는데 사용된 2단 도립진자의 물리적 파라미터는 표 1에 나열되어 있으며, 상미분 방정식의 해를 구하기 위한 솔버로는 Runge-kutta 방법을 선택하였다.

표 1 실험에 사용된 2단 도립진자의 파라미터
Table 1 Parameters of the double inverted pendulum used in the experiment

Parameter	Link	
	$i=1$	$i=2$
m_i	0.2351 [kg]	0.1452 [kg]
I_i	0.0012 [kgm^2]	0.0010 [kgm^2]
l_i	0.0667 [m]	0.1288 [m]
L_i	0.1645 [m]	-
c_i	4.5116e-04	2.9198e-04

강화학습 에이전트는 연속적인 행동 공간을 갖는 시스템에서 많이 사용되는 SAC(Soft Actor Critic)알고리즘을 통해 구현하였다. 해당 알고리즘은 연속적인 행동 공간을 지닌 복잡한 환경에서 높은 효율성과 안정성이 입증되었으며, 최대 엔트로피 항을 학습 과정에 추가함으로써 탐험을 통한 행동 정책의 다양성과 안정성을 향상시킬 수 있는 특징을 가지고 있다[15]. 해당 알고리즘을 구현하는데 있어 사용된 하이퍼 파라미터의 값들은 표 2에서 확인할 수 있으며, 두 개의 히든 레이어의 유닛 수가 400, 300으로 변경된 점 이외에는 참고문헌 [15]의 저자들이 사용한 파라미터 값을 모두 동일하게 사용하였다. 해당 알고리즘을 사용하여 구현된 강화학습 에이전트는 Python상의 2단 도립진자 시뮬레이션 환경과 지속적으로 상호작용하며 swing-up을 하기위한 제어 기법을 학습하게 된다.

표 2 SAC 알고리즘에 사용된 하이퍼 파라미터
Table 2 Hyperparameters used in the SAC algorithm

Parameter	Value
optimizer	Adam[16]
learning rate	3e-04
discount factor (γ)	0.99
replay buffer size	1e6
number of hidden layer	2
number of hidden units per 1^{st} layer	400
number of hidden units per 2^{nd} layer	300
nonlinearity	ReLU
target smoothing coefficient (τ)	0.005

2단 도립진자의 시뮬레이션 환경에서 관측 가능한 환경의 상태 정보는 3장에서 기술된 상태 방정식에 따라 $<y, \theta_1, \theta_2, \dot{y}, \dot{\theta}_1, \dot{\theta}_2>$ 로 이루어진 6개의 데이터로 구성된다. 이때, θ_1과 θ_2는 추후 원활한 보상함수의 설계를 위해 나머지 연산을 적용하여 $-\pi < \theta < \pi$의 범위로 제한한다. 추가적으로 θ_1과 θ_2는 학습 과정에서 정규화와 연속성의 이점을 얻기 위해 $\sin(\theta_i)$, $\cos(\theta_i)$의 형태로 재구성하여 사용한다. 결과적으로 강화학습 에이전트에게 전달되는 상태 정보의 형태는 $<y, \sin(\theta_1), \cos(\theta_1), \sin(\theta_2), \cos(\theta_2), \dot{y}, \dot{\theta}_1, \dot{\theta}_2>$로 구성된 8개의 데이터 묶음이 된다. 강화학습 에이전트는 해당 상태 정보를 입력으로 받아 자신의 행동 정책에 따른 행동, 즉 제어량을 출력한다. 이때 출력되는 제어량은 모터의 가속도 값 u에 해당하며, 실제 시스템 구동기의 작동 능력을 고려하여 $-15 < u < 15$의 값으로 제한한다.

시뮬레이션상의 학습 환경에서 한 에피소드의 길이는 10초로 설정하였고 시뮬레이션은 1ms 주기로 업데이트 되며, 학습 과정은 10ms마다 이루어진다. 따라서 에이전트는 환경과 한 에피소드당 최대 1000번 상호작용을 하게 되고, 상호작용이 일어나는 순간마다 그 시점의 보상 값에 기반하여 자신의 행동 정책을 개선한다. 보상 값을 산출하기 위한 보상함수는 식 (8)의 형태로 사용하였다.

$$\begin{aligned} R_u &= \exp(-0.03|u|) \\ R_y &= \exp(-0.001y^2) \\ R_{\theta_1} &= 0.5+0.5\cos(\theta_1) \\ R_{\theta_2} &= 0.5+0.5\cos(\theta_2) \\ R_{\dot{\theta}_1} &= 0.4+0.6\exp(-0.09|\dot{\theta}_1|) \\ R_{\dot{\theta}_2} &= 0.4+0.6\exp(-0.09|\dot{\theta}_2|) \end{aligned} \tag{7}$$

$$Reward = R_u \times R_y \times R_{\theta_1} \times R_{\theta_2} \times R_{\dot{\theta}_1} \times R_{\dot{\theta}_2} \tag{8}$$

상기된 보상함수를 이루는 각각의 요소는 그림 7에서 확인할 수 있으며, 모든 항은 0에 수렴할수록 보상 값이 증가하는 특성을 나타낸다. 이를 통해 두 개의 진자가 모두 도립된 상태, 즉 swing-up에 성공한 상태에서 최소한의 움직임만을 유지하는 방향으로 행동 정책을 학습하게 된다.

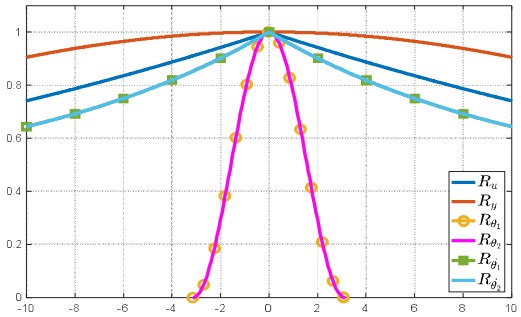


그림 7 보상함수 그래프
Fig. 7 Reward function graph

추가적으로 y의 값이 0.4[m]를 초과하는 경우에는 해당 에피소드는 학습에 도움이 되지 않기 때문에 해당 시점에서 조기 종료시킨다. 이는 추후 학습된 제어기를 실제 시스템에서 사용하는 상황에서, 실제 시스템이 동작할 수 있는 레일의 범위를 초과하는 상황을 방지하기 위함이다.

상기된 조건의 실험 환경에서 에피소드를 반복하여 실험을 진행하였고, 그 결과는 그림 8에서 확인할 수 있다. 약 1800회의 학습이 경과한 시점부터 에피소드 10개의 보상값 평균이 학습에 수렴하였음을 확인할 수 있다. 그러나 학습이 수렴한 이후에도 일부 에피소드에서는 보상이 현저히 낮게 나타나는 현상이 관측되는데, 이는 해당 실험에서 사용되는 제어기가 Recovery 특성을 가질 수 있도록 무작위한 초기 조건에서 실행되었기 때문이다. 각 에피소드가 시작될 때 $<y, \theta_1, \theta_2, \dot{y}, \dot{\theta}_1, \dot{\theta}_2>$로 구성된 환경의 상태 정보는 무작위성을 갖도록 초기화하여, 에이전트가 광범위한 상태 정보를 경험할 수 있도록 학습 환경을 설정하였다. 각 상태 정보가 따르는 난수의 범위는 식 (9)와 같다.

$$\begin{aligned} y &\sim U(-0.2, 0.2) \\ \theta_1 &\sim U(-\pi, \pi) \\ \theta_2 &\sim U(-\pi, \pi) \\ \dot{y} &\sim U(-1, 1) \\ \dot{\theta}_1 &\sim U(-10, 10) \\ \dot{\theta}_2 &\sim U(-20, 20) \end{aligned} \tag{9}$$

하지만 무작위성을 갖는 6개의 데이터가 모여 하나의 상태 정보를 형성하기 때문에, 결합된 상태정보가 물리법칙을 따르지 않는 상황이 발생할 수 있다. 이런 상태 정보를 초기 조건으로 가지고 모델방정식의 연산이 이루어지면, 물리법칙에 위배되어 물리적 의미가 없는 결과를 내게 된다. 강화학습 에이전트의 입장에서는 지금까지 한번도 경험해보지 못한 상태정

보를 입력으로 받게 되므로, 학습된 행동정책이 아닌 무작위성이 짙은 행동을 수행하게 된다. 이로 인해 발생하는 의미없는 대차의 이동은 시뮬레이션의 조기 종료 조건에 빠르게 도달하게 만든다. 따라서 해당 에피소드는 조기에 종료되며, 이런 현상은 그림 8에서 최종 보상이 낮은 특정 에피소드들로 나타나게 된다.

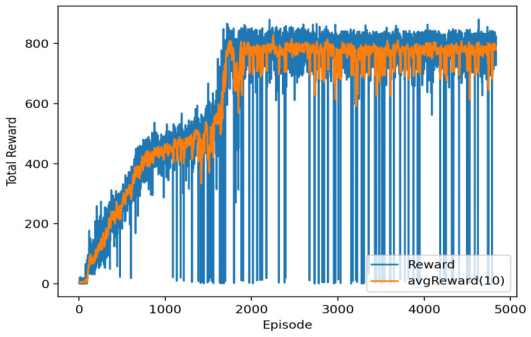


그림 8 학습 결과 그래프
Fig. 8 Learning results graph

그러나 이와 같은 문제는 학습이 완료된 제어기를 실제 시스템에 적용할 때에는 전혀 고려하지 않아도 되는 요소가 된다. 실제 시스템에서는 물리적인 법칙을 위배하는 현상이 발생할 수가 없기 때문이다. 따라서 강화학습 에이전트가 그러한 상태 정보를 관측하는 상황 자체가 발생하지 않기 때문에, 정상적인 상태 정보만을 가지고 학습에 기반한 정확한 제어량을 산출하여 swing-up 제어를 수행하게 된다.

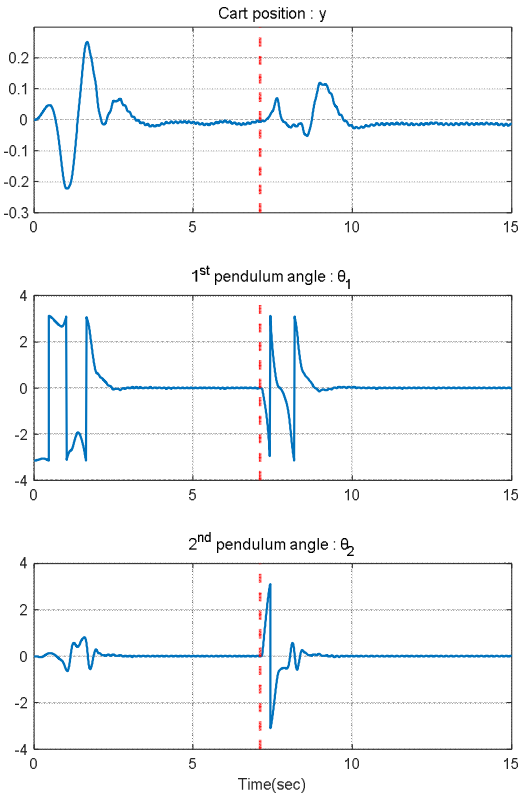


그림 9 외란 인가 실험 결과
Fig. 9 Results of Disturbance Injection Experiment

그림 9는 실제 2단 도립진자 시스템에 구현된 강화학습 기반 제어기를 적용하여 수행한 swing-up 제어의 결과를 나타내는 그래프이다. 초기 swing-up 제어에 성공한 뒤 선형 상태를 유지하고 있는 모습을 확인할 수 있고, 약 5초 이후 외부에서 강한 외란을 인가하였다. 그림 9에서 점선으로 표시된 시점이 외란을 인가한 순간을 나타낸다. 이로 인해 시스템이 불안정한 상태로 천이되었지만, 곧바로 다시 swing-up 제어를 시도하여 도립상태로 회복하는 모습을 확인할 수 있다. 이러한 결과를 통해 2장에서 언급했던 Recovery 특성을 갖는 강화학습 기반 제어기의 성능을 확인할 수 있다.

실험의 결과를 좀 더 명확하게 확인하기 위해, 실험 과정을 영상으로 기록하여 연구실 Youtube 채널에 업로드하였다. 해당 영상은 https://youtu.be/4ELdGB9UYZo 에서 확인할 수 있다. (영상 제목 : Reinforcement learning control of a double inverted pendulum with good recovery performance, 채널명 : Embedded Control Lab). 해당 영상에서 2단 도립진자 시스템에 어떠한 외란을 인가한 경우에도, 강화학습 기반 제어기가 성공적으로 swing-up 제어를 수행하는 모습을 확인할 수 있다.

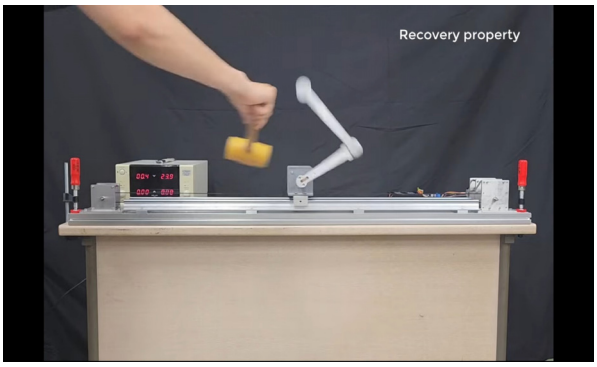


그림 10 실험 과정을 기록한 Youtube 영상
Fig. 10 YouTube video of the experiment procedure

5. Conclusion

본 논문에서는 Sim-to-Real 학습 기법을 활용하여 강화학습 기반의 제어기를 설계하고 검증하였다. 특히, 강한 외란에 의해 불안정해진 상태에서도 swing-up을 수행할 수 있는 능력을 확인할 수 있었다. 이는 강화학습 기반의 제어기가 전통적인 제어 기법의 한계를 극복할 수 있고, 복잡한 제어 문제에 있어 효과적인 해결 방안이 될 수 있음을 보여준다.

또한, 시뮬레이션과 실제 환경 간의 정합성을 높이기 위한 설계 방안을 제시하였다. 이를 통해 Sim-to-Real 학습 기법의

주요 도전과제인 현실 격차를 줄이는 방법론을 구체화하였으며, 실제 시스템에서의 실험을 통해 그 유효성을 입증하였다.

본 논문에서는 2단 도립진자의 swing-up 제어에 초점을 맞추었지만, 이를 확장하여 다양한 제어 문제에 적용하는 연구를 생각해볼 수 있다. 최근에는 다단 도립진자의 독특한 특성을 활용한 천이제어와 같은 새로운 제어방식이 제시되었으며[17], 3단 도립진자와 같이 더 난도 높은 시스템에 대한 연구도 진행되고 있다[12]. 이러한 확장된 문제에 대해서도 본 연구에서 사용된 접근 방식이 유용하게 활용될 수 있을 것으로 기대된다.

References

[1] N. Muskinja and B. Tovornik, “Swinging Up and Stabilization of a Real Inverted Pendulum,” in IEEE Transactions on Industrial Electronics, vol. 53, no. 2, pp. 631-639, April 2006.

[2] Y. Otani, T. Kurokami, A. Inoue, and Y. Hirashima, “A Swingup Control of an Inverted Pendulum with Cart Position Control,” IFAC Proceedings, vol. 34, pp. 395-400, 2001.

[3] K. Graicehn, M. Treuer, and M. Zeitz, “Swing-up of the Double Pendulum on a Cart by Feedforward and Feedback Control with Experimental Validation,” Automatica, vol. 43, pp. 63-71, 2007.

[4] J. Kober, J. A. Bagnell, and J. Peters, “Reinforcement Learning in Robotics: A Survey,” The International Journal of Robotics Research, vol. 32, pp. 1238-1274, 2013.

[5] S. Israilov, L. Fu, J. Sánchez-Rodríguez, F. Fusco, G. Allibert,C. Raufaste, and A. Médéric, “Reinforcement Learning Approach to Control an Inverted Pendulum: A General Framework for Educational Purposes,” PLoS ONE, vol. 18, e0280071, 2023.

[6] J. Baek, C. Lee, Y. S. Lee, S. Jeon, and S. Han, “Reinforcement Learning to Achieve Real-time Control of Triple Inverted Pendulum,” Engineering Applications of Artificial Intelligence, vol. 128, 107518, 2024.

[7] Y. Gil, J. H. Park, J. Baek, and S. Han, “Quantization-aware Pruning Criterion for Industrial Applications,” IEEE Transactions on Industrial Electronics, vol. 69, no. 3, pp. 3203-3213, 2022.

[8] J. Baek, H. Jun, J. Park, H. Lee, and S. Han, “Sparse Variational Deterministic Policy Gradient for Continuous Real-time Control,” IEEE Transactions on Industrial Electronics, vol. 68, no. 10, pp. 9800-9810, 2021.

[9] G. Dulac-Arnold, D. Mankowitz and T. Hester, “Challenges of Real-world Reinforcement Learning,” arXiv preprint arXiv:1904.12901, 2019.

[10] W. Zhao, J. P. Queralta, and T. Westerlund, “Sim-to-Real Transfer in Deep Reinforcement Learning for Robotics: a Survey,” 2020 IEEE Symposium Series on Computational Intelligence (SSCI), pp. 737-744, 2020.

[11] N. Jakobi, P. Husbands, and I. Harvey. “Noise and the Reality Gap: The Use of Simulation in Evolutionary Robotics,” Advances in Artificial Life: Third European Conference on Artificial Life Granada, pp. 704-720, 1995.

[12] T. Glück, A. Eder, and A. Kugi,, “Swing-up Control of a Triple Pendulum on a Cart with Experimental Validation,” Automatica, vol. 49, pp. 801-808, 2013.

[13] D. Ju, C. Choi, J. Jeong and Y. S. Lee, “Design and Parameter Estimation of a Double Inverted Pendulum for Model-based Swing-up Control,” Journal of Institute of Control, Robotics and Systems (in Korean), vol. 28, no. 9, pp. 793-803, 2022.

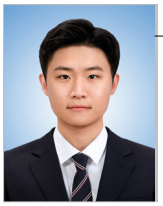
[14] T. Lee, D. Ju, and Y. S. Lee, “Development Environment of Reinforcement Learning-based Controllers for Real-world Physical Systems Using LW-RCP,” Journal of Institute of Control, Robotics and Systems (in Korean), vol. 29, no. 7, pp. 543-549, 2023.

[15] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine. “Soft Actor-critic: Off-policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor,” International conference on machine learning. PMLR, pp. 1861-1870, 2018.

[16] D, P. Kingma. and J. Ba. “Adam: A Method for Stochastic Optimization,” arXiv preprint arXiv:1412.6980, 2014.

[17] J. Jeong, D. Ju, Y. Fujiyama, and Y. S. Lee, “Transition Control of a Double Inverted Pendulum Using an LW-RCP,” Journal of Institute of Control, Robotics and Systems (in Korean), vol. 29, no. 9, pp. 694-703, 2023.

저자소개



이태건 (Taegun Lee)

He received B.S. degree in electrical engineering from Inha university in 2023. He is now a M.S. candidate in electrical and computer engineering at Inha university. His research interests include reinforcement learning, embedded systems and optimal control.

주도윤 (Doyoon Ju)

He received M.S. degree in electrical and computer engineering from Inha university in 2023. He is now a Ph.D. candidate in electrical and computer engineering at Inha university. His research interests include optimal control, embedded systems and reinforcement learning.

이영삼 (Young Sam Lee)

He received B.S. and M.S. degrees in electrical engineering from Inha University, Incheon, South Korea, in 1999, and the Ph.D. degree in electrical engineering from Seoul National University, South Korea, in 2003. From 2003 to 2004, he was a Senior Researcher with Samsung Electronics Co. Since 2004, he has been with the Department of Electrical and Computer Engineering, Inha University. He is the author of four books and more than 60 articles. His research interests include computer-aided control system designs, rapid control prototyping, control and instrumentation, robot engineering, and embedded systems.