Mechatronics 85 (2022) 102833

Contents lists available at ScienceDirect

Mechatronics

journal homepage: www.elsevier.com/locate/mechatronics

Control framework for collaborative robot using imitation learning-based teleoperation from human digital twin to robot digital twin☆

Hyunsoo Lee [a], Seong Dae Kim [b,*], Mohammad Aman Ullah Al Amin [c]

[a] Kumoh National Institute of Technology, School of Industrial Engineering, South Korea
[b] University of Tennessee at Chattanooga, Department of Engineering Management & Technology, Chattanooga, TN, USA
[c] University of Texas at Arlington, Department of Industrial Engineering, Arlington, TX, USA

ARTICLE INFO

Keywords:
Collaborative robot
Teleoperation framework
Imitation learning
Digital twin
Bezier curve-based smooth pose mapping
Convolutional encoder-decoder

ABSTRACT

Despite the deployment of collaborative robots for various industrial processes, their teaching and control remain comparatively difficult tasks compared with general industrial robots. Various imitation learning methods involving the transfer of human poses to a collaborative robot have been proposed. However, most of these methods depend heavily on deep learning-based human recognition algorithms that fail to recognize complicated human poses. To address this issue, we propose an automated/semi-automated vision-based teleoperation framework using human digital twin and a collaborative robot digital twin models. First, a human pose is recognized and reasoned to a human skeleton model using a convolution encoder-decoder architecture. Next, the developed human digital twin model is taught using the skeletons. As human and collaborative robots have different joints and rotation architectures, pose mapping is achieved using the proposed Bezier curve-based smooth approximation. Then, a real collaborative robot is controlled using the developed robot digital twin. Furthermore, the proposed framework works successfully using a human digital twin in the case of recognition failures of human poses. To verify the effectiveness of the proposed framework, transfers of several human poses to a real collaborative robot are tested and analyzed.

1. Introduction

The deployment of Industry 4.0 technologies in various industrial fields has resulted in the development of innovative robot controls and relevant processes. One of the groundbreaking fields in robot industries is collaborative robots. A general collaborative robot resembles a human in terms of the movements and poses as well as utility of hands and arms. The introduction of collaborative robots has proven to be a breakthrough for delicate process applications from medical surgeries [1] to various manufacturing operations such as assembly [2], welding [3] and inspection [4].

Several relevant research studies and applications have been proposed for teaching and controlling a general robot. Gilmore et al. [5] introduced a stochastic process approach to describe the pallet-filling process of a robot. Although the stochastic algebra modeled the movement of a robot with a predefined task, it was limited by the fact that the task was comparatively simple. Toquica et al. [6] applied deep learning approaches, e.g., deep neural network (DNN), long short-term memory (LSTM), and gated recurrent unit (GRU), to obtain a mapping from the end-effect position to the joint angles considering the structure of a parallel robot.

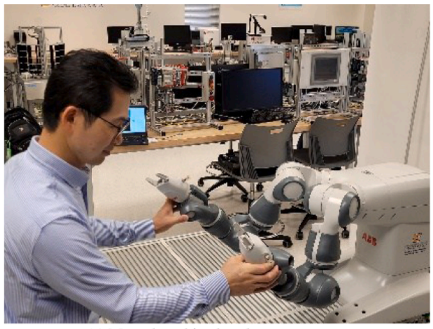
However, the teaching and control of a collaborative robot have been considered more difficult than the control of general industrial robots with an arm. In general, a collaborative robot has around 10 degrees of freedom (DoF). For instance, *ABB Yumi*© has 14 DoFs, 7 in each arm. Although a higher number of DoFs enables the robot to carry out delicate manufacturing processes, it is difficult to teach and to control their poses as quid pro quo. Most conventional collaborative robots support manual teaching methods that involve the robot memorizing joints' position and rotation information with the help of a human operator, as shown in Fig. 1 (a).

A possible alternative to the aforementioned manual teaching methods is imitation learning [7–9]. Imitation learning is quick, and the recent advances made in deep learning techniques have accelerated its adoption. Most imitation learning frameworks use convolutional neural networks (CNNs) to recognize human poses or use 3D depth-based

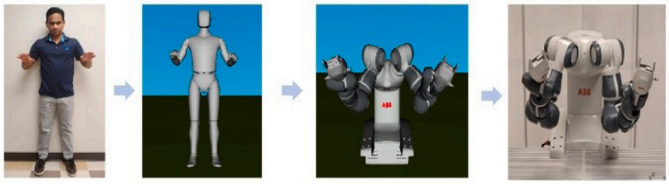
☆ This paper was recommended for publication by Associate Editor Dr. Tsu-Chin Tsao.

* Corresponding author at: Address: 615 McCallie Ave, Chattanooga, TN 37403, USA.
E-mail address: Seongdae-Kim@utc.edu (S.D. Kim).

https://doi.org/10.1016/j.mechatronics.2022.102833
Received 5 October 2021; Received in revised form 18 March 2022; Accepted 19 May 2022
Available online 24 May 2022



(a) Manual teaching by a human operator.



(b) Teaching of a collaborative robot using the proposed digital twin-based imitation learning.

Fig. 1. A human CPS model-based collaborative robot teaching framework

devices such as *Microsoft Kinect*©. Detailed reviews of these approaches are provided in the following sections.

However, the performance of existing imitation learning techniques depends heavily on the quality of embedding deep learning modules. Although some success has been achieved in the detection of relatively simpler human motion using contemporary deep learning applications, attempts at detecting complicated human arm poses and movements have posed significant challenges, and more often than not, have failed [10].

To overcome this issue, this study proposes a new and efficient control framework for a collaborative robot using the two digital twin models: human cyber-physical system (CPS) / digital twin and collaborative robot digital twin. Fig. 1 (b) shows the teaching of a collaborative robot using the proposed framework.

One of the advantages of the proposed control framework is the semi-automatic teleoperation-based control using the human digital twin model. In the case of recognition failures of complicated human poses, it is possible to control the robot using the developed teleoperation framework using human and collaborative robot digital twin models.

The following section provides the background knowledge and literature reviews of relevant robotic fields. Section 3 explains the digital twin-based control framework and detailed mechanisms, including convolutional deep learning, human digital twin model, collaborative robot digital twin, and their pose mapping theorem. To demonstrate the effectiveness of the proposed framework, several human poses and their transfers were tested and analyzed as described in Section 4.

2. Background and literature review

While manual teaching has been a traditional and popular method for collaborative robots, the complicated structures of cooperative robots and emerging technologies including machine visions and deep learning have accelerated new and intelligent control methods for collaborative robots. In this manner, an imitation learning-based teleoperation framework is proposed. Generally, in imitation learning, a human pose is reasoned using the provided deep learning model and converted into a reasoned human skeleton model.

The primary goals of imitation learning for robots are human-pose recognition and pose transfer to humanoid robots. Over the years, multiple cameras-based computer vision techniques or 3D camera vision devices such as *Microsoft Kinect* © have been used for human-pose recognition. Human skeletons are obtained comparatively easily using these devices. Another human-pose acquisition trend is the use of deep learning methods [11–16]. Human pose images or motion frames are used as input for deep learning architectures, following which the joints' information [13] is captured and mapped to a real robot. Table 1 summarizes several imitation learning studies and their applications.

As shown in Table 1, most of the existing imitation learning frameworks have several limitations and issues because it is difficult to recognize poses as imitation inputs and deliver poses considering the difference between the skeleton architectures of humans and target robots. For these reasons, several research studies have used direct skeleton signals from conventional devices and considered the teaching of a target humanoid robot using the same skeleton configuration. While several research studies and applications introduced comparatively recent deep-learning techniques, this trend is limited because these

Table 1
Characteristics of existing imitation learning studies and their applications

Research studies	Input	Output	Used methods	Characteristics and limitation
[17]	Simulation data of a predefined robot	Well-trained robot poses	Inverse reinforcement learning (IRL) using adversarial network	- Robot-to-robot imitation learning - The same joint architecture
[18]	IoT sensor based signals	Robotic poses	Deep Reinforcement Learning	- No human-robot imitation learning
[19]	Manual data for human and robot information	Trained human-robot poses	Random forest (RF) method	- No inference in human motion capture using sensors
[20]	Kinect-based human skeleton signals	Humanoid robot poses	Denavit-Hartenberg (DH)-based rotation mapping	- No inference in human motion capture using sensors - The same architecture between human and a robot
[21]	Kinect-based human skeleton signals	Humanoid robot poses	Direct mapping	- No inference in human motion capture using sensors - The same architecture between human and a robot
[22]	Robot's infant data	Well-trained robot poses	Graph-based Bayesian learning	- robot-to-robot imitation learning - The same joint architecture
[23]	Human pose using sensors	A humanoid pose	Direct pose mapping	- No inference in human motion capture using sensors - The same joint architecture
[24]	A single camera image of human motion	Upper body poses of a humanoid robot	PCA-processed Expectation Maximization (EM) method	- The same architecture of human and a humanoid robot
[25]	Marker based human motions	Upper body poses of a humanoid robot	Hidden Markov model (HMM)	- Traditional marker-based human motion capture
[26]	A 3D camera-based single image / simulation data of human arm	One-hand robot arm pose	Network learning	- Limited on one-handed robot arm

depend heavily on the efficiency of the embedded deep learning modules.

To overcome these issues, this study proposes an imitation learning-based teleoperation and control framework from a human pose (X) to the motion of a real collaborative robot (Y) using different skeleton architectures for human poses and the robot. To transfer poses, skeleton architecture and joints' information of a human digital twin ($X^{'}$) and a robot digital twin ($Y^{'}$) are developed, as shown in Eq. (1).

$$X \rightarrow X^{'} \rightarrow Y^{'} \rightarrow Y \tag{1}$$

The most representative issues for imitation learning are the automatic/semi-automatic recognition of human poses and contrasting skeletal architectures between a human and a collaborative robot. For instance, human arms have at least six DoFs(the left shoulder, left elbow, left wrist, right shoulder, right elbow, and right wrist joints). By contrast, a collaborative robot (e.g., *ABB Yumi*©) has 14 DoFs. Eqs. (2) and (3) denote the different joints' information of a human and a robot, respectively.

$$\mathrm{X} = (\mathrm{X}_1, \cdots, X_m), \quad \mathrm{X_i} \in R \tag{2}$$

where m is the number of joints in human arms

$$\mathrm{Y} = (\mathrm{Y}_1, \cdots, Y_n), \quad Y_i \in R \tag{3}$$

where n is the number of joints in a collaborative robot's arms $X^{'}$ and $Y^{'}$ share the same space as Y and X, respectively. Then, three mappings (Eqs. (4)–(6)) are required for human-collaborative robot pose mappings.

$$\mathrm{f} : \mathrm{X} \rightarrow \mathrm{X}^{'} \tag{4}$$

$$\mathrm{g} : \mathrm{X}^{'} \rightarrow Y^{'} \tag{5}$$

$$\mathrm{h} : \mathrm{Y}^{'} \rightarrow Y \tag{6}$$

Although most of the existing imitation learning elaborated on Eq. (4) using several deep-learning techniques, this study proposes a new and effective motion transfer framework that supports mapping Eq. (7).

$$\mathrm{MT} : \mathrm{h} \circ \mathrm{g} \circ \mathrm{f} \tag{7}$$

Owing to the different structures between humans and a collaborative robot, the motion transition (MT in Eq. (7)) from f to h fails to guarantee accurate movements of a robot in most of the relevant research studies and applications provided in Table 1. In particular, one of the challenging issues is mapping g in Eq. (5). For $\mathrm{m} < \mathrm{n}$, the number of output variables was larger than the number of input variables. This indicates that general deep learning and function approximation are a part of this imitation learning-based teleoperation. To solve this issue, a pose similarity-based motion transfer method is proposed. The proposed motion transfer method is based on a single vision for 3D human-pose extraction. There have been a number of research methods and applications for capturing human poses. Table 2 summarizes several human-pose recognition methods.

The detailed 3D human pose extraction of the proposed method is provided in Section 3.2. The following section elaborates on the overall framework and detailed modules of the proposed control framework.

3. Imitation learning-based teleoperation framework using human digital twin and a collaborative robot digital twin

As mentioned in the previous sections, this study considers the transfer of human poses to a collaborative robot using the two digital twin frameworks ($X^{'}$ and $Y^{'}$). As a human skeleton model is different from that of a target robot, as shown in Eqs. (8) and (9), in general, $\mathrm{n} > \mathrm{m}$ is assumed.

Table 2
Research studies for 3D human-pose extraction.

Research studies	Input source	Output	Used methods	Characteristics
[27]	A 2D single image	3D skeleton model	CNN Encoder/ Decoder architecture	Mapping with human skeleton model and CNN-based feature
[28]	Top/front (multiple) images with depth maps In ITOP (2018) and UBC3V (2016)	3D skeleton model	CNN	Linear combination of predefined poses
[29]	a single image with depth information in EVAL(2012) and ITOP (2018)	3D human pose	CNN & RNN (LSTM)	Consideration of temporal poses
[30]	Single image	3D human pose with a camera	CNN & 3D pose matching	Pretrained pose-based matching
[31]	a 2D single image	3D skeleton model	CNN encoder -decoder architecture	Application to image frames using CNN and optical flow-based tracking

$$\mathrm{X}^{'} = \left(\mathrm{X}_1^{'}(\mathrm{t}), \cdots, \mathrm{X}_\mathrm{m}^{'}(\mathrm{t})\right), \ \mathrm{m} \in \mathrm{N}, \ \mathrm{X}_\mathrm{i}^{'} \in R \tag{8}$$

$$\mathrm{Y}^{'} = \left(\mathrm{Y}_1^{'}(\mathrm{t}), \cdots, \mathrm{Y}_\mathrm{n}^{'}(\mathrm{t})\right), \ \mathrm{n} \in \mathrm{N}, \ \mathrm{Y}_\mathrm{i}^{'} \in R \tag{9}$$

Fig. 2 shows the proposed imitation learning-based teleoperation framework. The framework consists of the two digital twin models, human posture estimation model, pose adjusting algorithm, and mapping algorithm for the transfer of human poses to a collaborative robot.

The proposed framework uses a single 2D human-pose image. The input image is reasoned to a 3D skeleton model using a modified deep learning of ResNet-based convolutional encoder/decoder [27]. Then, the human digital twin model corrects its human poses using the reasoned skeletons and the theorem, as provided in Section 3.2. The poses of the developed robot digital twin were corrected with a criterion of pose similarity using the Bezier curve approximation. Finally, the extracted robot joints' information (e.g., rotation axis and angle in each joint) is transmitted to a connected real collaborative robot using a PLC program. Robot Operating System (ROS) 1 or ROS 2 can be good tools for teaching and controlling a cooperative robot. However, the objective of the proposed framework is to equip a vision-based teleoperation ability using several deep learning architectures. As these commercial software programs don't have these functionalities, the custom tool is developed.

3.1. Human digital twin model for imitation learning-based teleoperation

In this subsection, we elaborate on the developed human digital twin model, illustrated in Fig. 3.

A number of artificial human models have been proposed to date, most of which have been designed using anthropometric data, and then used for operation or safety tests. [32] implemented an artificial human arm model using a *virtual reality modeling language* (VRML). This model controls an anthropometric data-based anthropomorphic joint model in which each arm has seven DOFs. Overall, the joints are controlled using Denavit-Hartenberg (DH) parameters [33] and Levenberg-Marquardt based inverse kinematic positions.

The human model used in the present study was modeled using the anthropometric data [34] recorded in The Korean Agency for Technology and Standards' 7th Korean Anthropometric Data Survey Report.

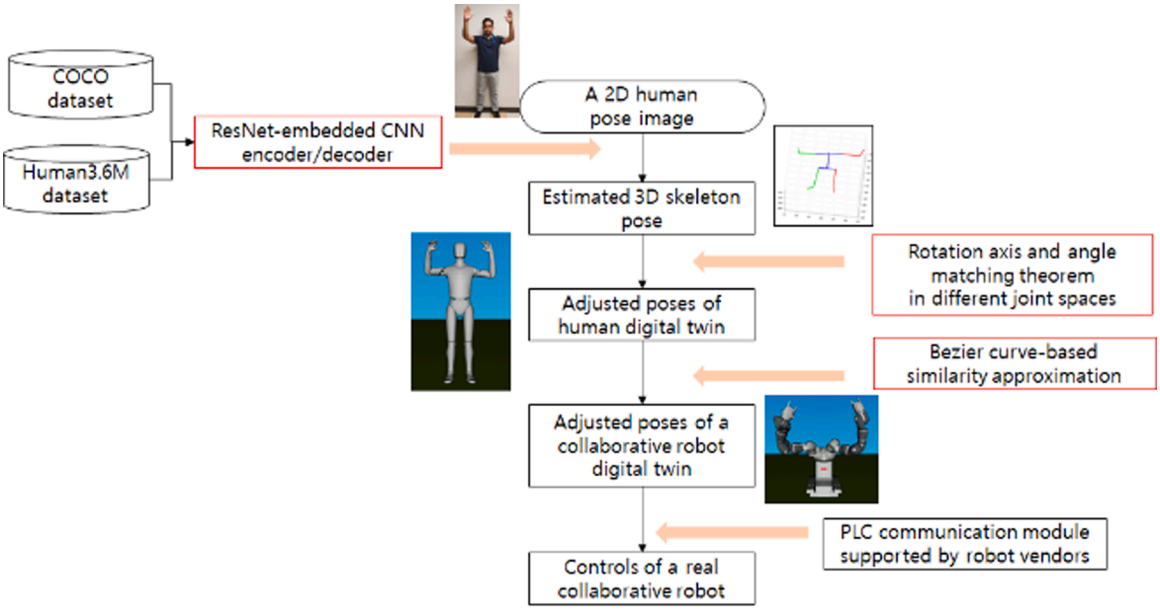


Fig. 2. The overall imitation learning-based teleoperation framework using both digital twin models.

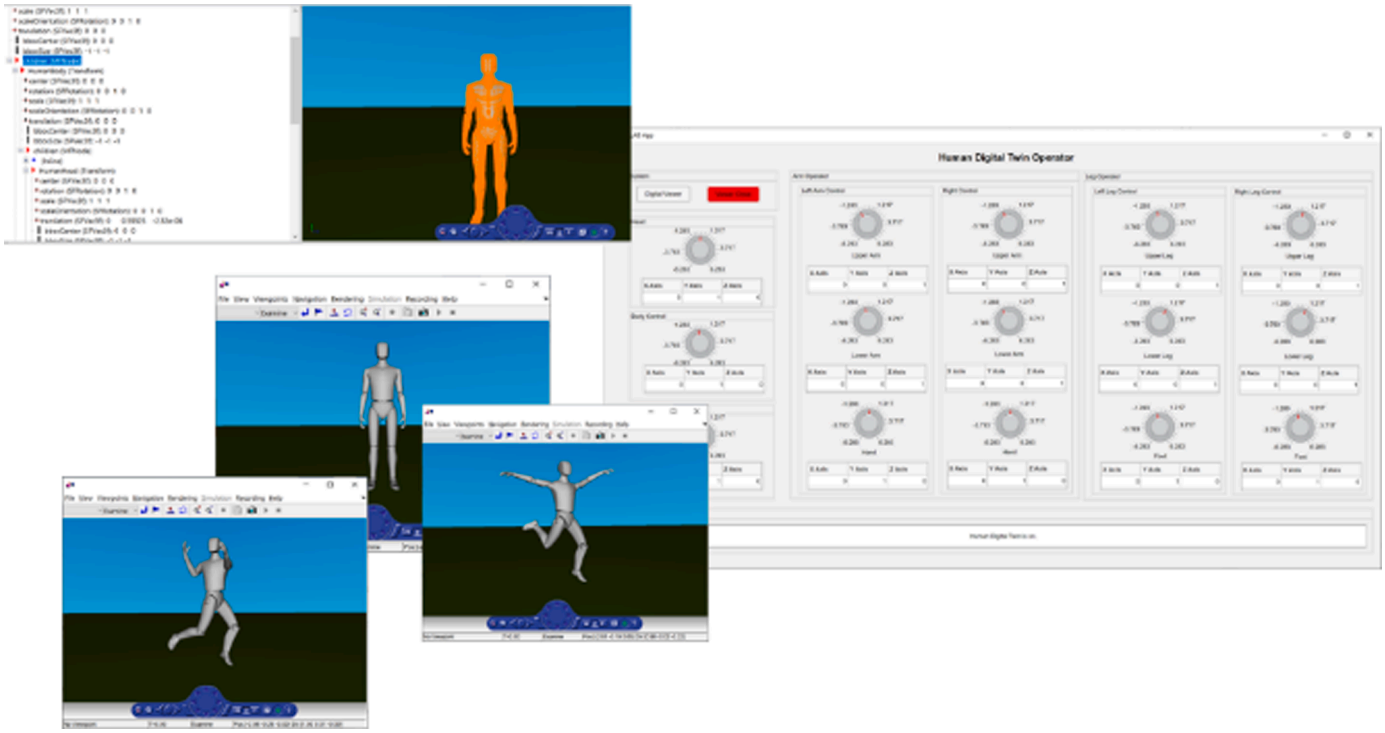


Fig. 3. The developed human digital twin.

Then the model was modified to align it with our aim of the pose transfer to humanoid robots. The uses of the full-body pose in humans and the developed human digital twin model can cover the controls of various humanoid robots as well as arm-based robots. In this manner, the human digital twin model is developed with full-body motions. As shown in Fig. 3., the human model has 15 joints: head/neck, shoulder, waist, three left arm joints (shoulder arm, elbow, and wrist), three right arm joints, three left leg joints, and three right leg joints. Each joint, guided by anthropometric data, facilitates movements around relevant rotation axes and angles. Each human part was modeled using conventional CAD software and converted into a corresponding VRML file. Furthermore, an integrated human model was constructed considering a human skeletal scene graph.

To control each joint with respect to the joint axis and angles, quaternion-based knob handlers were implemented using *MATLAB R2021a* and C++. The human digital twin operator is shown in Fig. 3. The functionalities of the developed human digital twin include direct human-pose control and pose transfer from recognized human poses.

3.2. *Human-pose recognition and transfer to Human digital twin*

More complicated controls considering both arms in a collaborative

robot require another well-defined mapping framework. This study proposes mapping from a human action to a bi-arm operation of a collaborative robot. As a prerequisite task, a human-pose recognition and its transfer to the developed human digital twin model are required. As shown in Table 2, most human recognitions use deep-learning techniques, particularly CNNs. [27] used a human skeleton model and graph convolution networks (GCN) to recognize human actions. The study applied the *OpenPose* library [27] to obtain a human skeleton set. The library represents a body motion identification-based software library [35,36] using a bottom-up strategy. The bottom-up strategy [37] identifies parts of human bodies and then classifies multiple humans using a probability-based heatmap and part affinity fields (PAF).

In the present study, we use a modified version of this method. *OpenPose* library-based CNN encoder-decoder architecture is trained using several human poses. Then, convolution filters and the relevant weights are readjusted. Fig. 4. shows the CNN encoder-decoder architecture for extracting a 3D human skeleton model.

As shown in Fig 4., this study recognizes a 3D human skeleton model from a 2D human picture captured by a camera. The architecture used consists of five segments: input segment, CNN-encoder, hidden feature layer, CNN-decoder, and human skeleton estimator. A 2D human-pose image enters the input layer. Then, pose features are extracted in the CNN encoder, which is a combination of three convolution layers and three maxpooling layers (six layers). The reasoned features pass through the hidden feature layers (six layers) for further reasoning. To construct a human skeleton model, a CNN-decoder architecture is applied, and then an estimated human skeleton model is obtained using the output layers (regression layer). As shown in Fig. 2, this model is modified from a ResNet-embedded CNN encoder/decoder architecture [27]. The overall deep learning architecture was implemented using the *OpenPose* library [37] and *MATLAB© Deep Learning Toolbox*.

However, this study focuses on the transfer of a reasoned pose, while a number of related research studies have focused on accurate human-pose reasoning. Therefore, the human-pose estimation module can be replaced with a more accurate estimator.

After the human skeleton model is estimated, the subsequent task is to adjust the developed human digital twin model. Fig. 5 (a) shows the upper body skeleton (a target pose) of an estimated human model, and Fig. 5 (b) shows the upper skeletal architecture (a subject) of the human digital twin model.

While a matching of a current human digital pose to a target pose is needed, it must be considered that both pose spaces are different. As both poses exist in different geometric spaces, pose matching requires the following procedures: (1) definition of reference plains, (2) space rotation with respect to a reference plain, and (3) a line rotation (joint) with respect to a matched reference plain. Fig. 5 (c) shows the process for human-pose matching, which is achieved using the following theorem.

Theorem: Rotation axis ($R_{A,Line}$) and angle (θ_{Line}) of a line connected to a plain, with respect to another line connected to another plain.

$$R_{A,Line} = \frac{\overline{qp} \times \overline{b'a'}}{|\overline{qp} \times \overline{b'a'}|} \tag{10}$$

$$\theta_{Line} = \cos^{-1}\left(\frac{\overline{qp} \cdot \overline{b'a'}}{|\overline{qp}||\overline{b'a'}|}\right) \tag{11}$$

Proof: In a digital twin model, a point q is a joint in the human model and line $\overline{qp}$ is a part of the arm that is connected to joint q. Both lines $\overline{rq}$ and $\overline{rs}$ are on a plane. The normal vector (n_{cps}) of the plane was calculated using Eq. (12).

$$n_{cps} = \overline{rq} \times \overline{rs} \tag{12}$$

where $\times$ is a cross product with two vectors. The corresponding normal vector (n_{target}) in the target skeleton is calculated using Eq. (13).

$$n_{target} = \overline{cb} \times \overline{cd} \tag{13}$$

To extract the rotation axis and angle, the prerequisite condition is the matching between planes: $\overline{qrs}$ and $\overline{bcd}$. The plane ($\overline{bcd}$) is matched onto $\overline{qrs}$, with respect to the criterion: the joint q's location is the same as

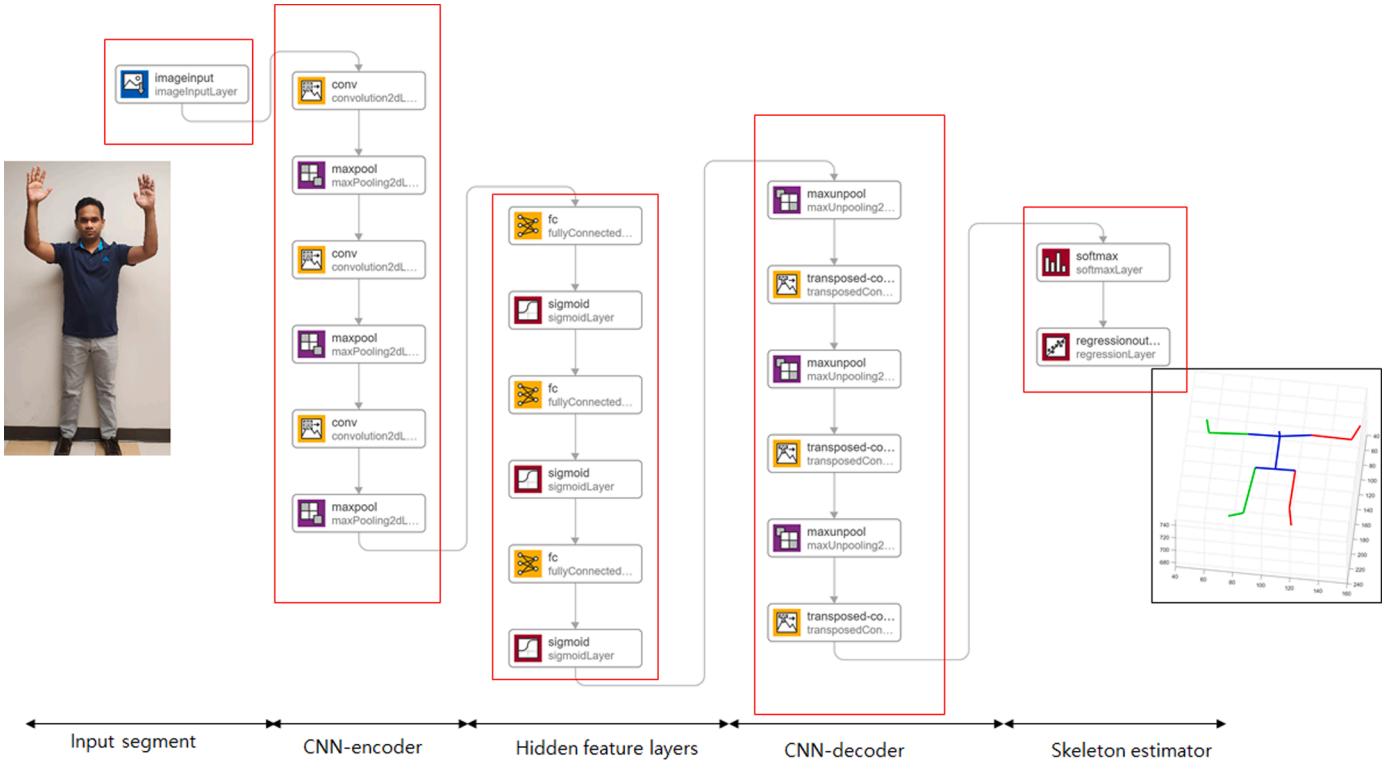
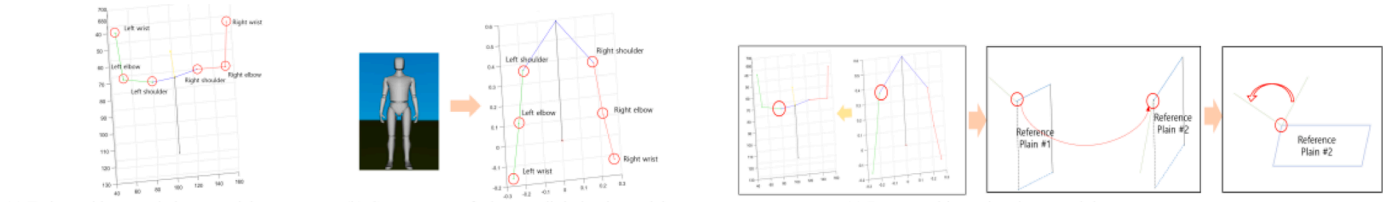


Fig. 4. CNN encoder /decoder architecture for human-pose extraction.



(a) Estimated human skeleton model as a target (b) Current pose of a human digital twin model (c) Pose matching using the prosed theorem.

Fig. 5. Target skeleton and subject skeleton (human digital twin) for pose transfer.

the location of the corresponding joint b. Then, the rotation axis ($R_{A,plain} = (R_{A,X}, R_{A,Y}, R_{A,z})$) and angle ($\theta_{plain}$) are calculated using Eq. (14) and Eq. (15), respectively.

$$R_{A,plain} = \frac{n_{target} \times n_{cps}}{|n_{target} \times n_{cps}|} \tag{14}$$

$$\theta_{plain} = \cos^{-1}\left(\frac{n_{target} \cdot n_{cps}}{|n_{target}||n_{cps}|}\right) \tag{15}$$

As the joint angle and the related axis exist on $\overline{qrs}$, $\overline{bcd}$ is rotated and matched on $\overline{qrs}$ as shown in Fig. 5(c). θ_{plain} is the radius angle. Then, the plain $\overline{bcd}$ and the line $\overline{ba}$ are transformed using the new line $\overline{b'a'}$ using Eqs. (16) and (17). The plain $\overline{bcd}$ is rotated with the θ_{plain} and the rotation matrix in Eq. (16).

$$R_{plain}(:,1) = \begin{bmatrix} cos\theta_{plain} + R_{A,X}^2(1 - cos\theta_{plain}) \\ R_{A,Y}R_{B,X}(1 - cos\theta_{plain}) - R_{A,Z}sin\theta_{plain} \\ R_{A,Z}R_{B,X}(1 - cos\theta_{plain}) - R_{A,Y}sin\theta_{plain} \end{bmatrix}$$

$$R_{plain}(:,2) = \begin{bmatrix} R_{A,X}R_{B,Y}(1 - cos\theta_{plain}) - R_{A,Z}sin\theta_{plain} \\ cos\theta_{plain} + R_{A,Y}^2(1 - cos\theta_{plain}) \\ R_{A,Z}R_{B,Y}(1 - cos\theta_{plain}) - R_{A,X}sin\theta_{plain} \end{bmatrix} \tag{16}$$

$$R_{plain}(:,3) = \begin{bmatrix} R_{A,X}R_{B,Z}(1 - cos\theta_{plain}) + R_{A,Y}sin\theta_{plain} \\ R_{A,Y}R_{B,Z}(1 - cos\theta_{plain}) - R_{A,X}sin\theta_{plain} \\ cos\theta_{plain} + R_{A,Z}^2(1 - cos\theta_{plain}) \end{bmatrix}$$

$$\overline{b'a'} = R_{plain} \cdot \overline{ba} \tag{17}$$

Using the rotation matrix R_{plain}, the plain $\overline{b'c'd'}$ is on $\overline{qrs}$. Finally, the rotation axis and angle from $\overline{qp}$ to $\overline{b'a'}$, with respect to the origin (Point q) on $\overline{qrs}$ are calculated using Eqs. (10), (11).Q.E.D.

Finally, the overall human digital model is adjusted and corrected using the estimated human-pose skeletons and the proposed theorem. The subsequent task is to transfer the poses in a human digital twin model to a collaborative robot digital twin. Before elaborating on this,

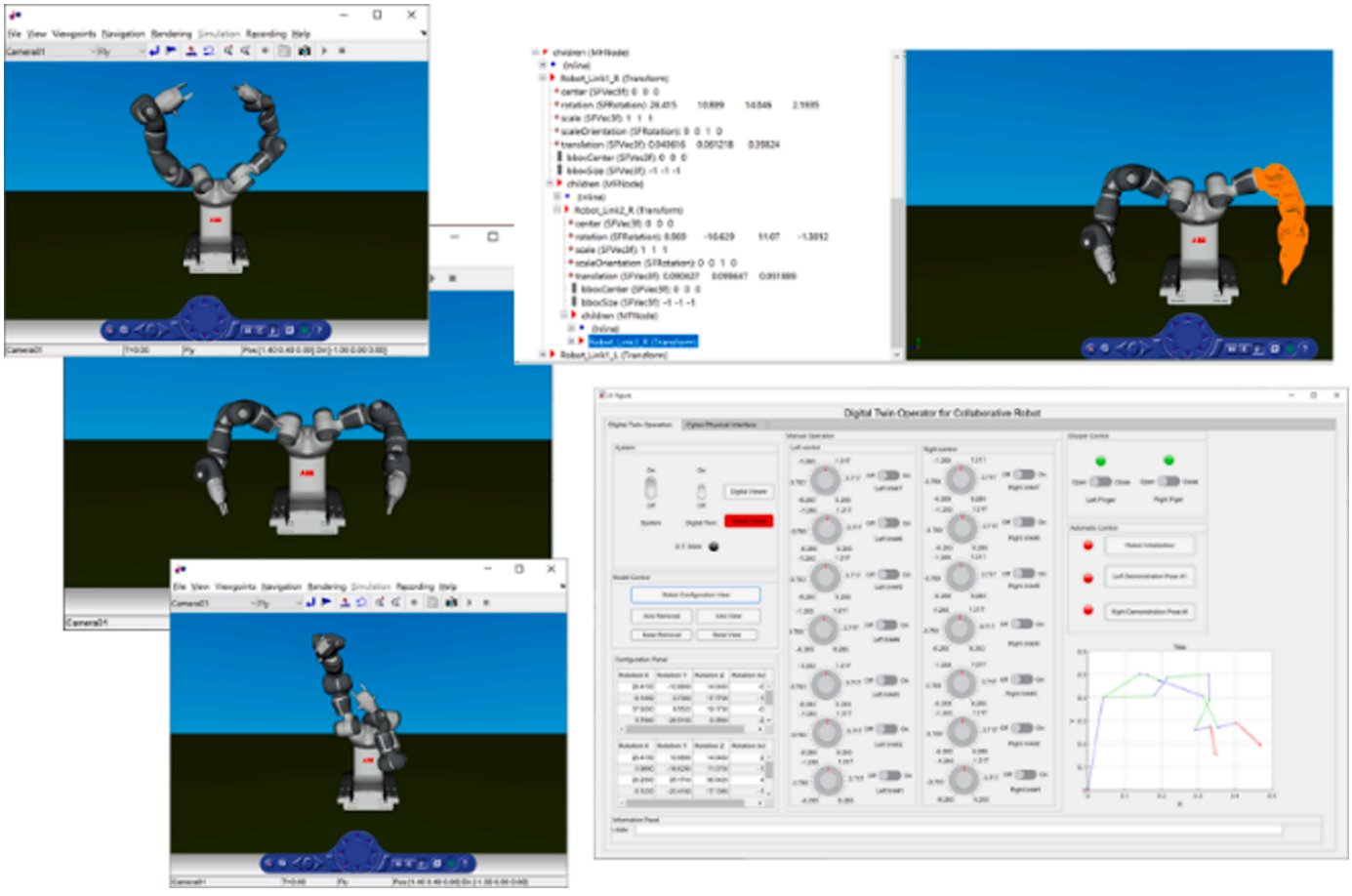


Fig. 6. The developed digital twin of a collaborative robot.

the following subsection explains the developed digital twin model of a collaborative robot.

3.3. Digital twin model for collaborative robot

As shown in Fig. 1(b), controls of a collaborative robot are achieved using its digital twin model. To develop a collaborative robot digital twin, this study refers to an *ABB Yumi*© robot that has 14 DoF in its arms. Each arm has seven joints. Each robot joints and parts are modeled using the CAD models of *ABB*©. Then, each CAD model is converted into a corresponding VRML file like a human digital twin model.

Then, overall VRML files are assembled using the robot architecture. Fig. 6. Shows a scene graph of the implemented collaborative robot and its control model. Then, each control knob of a corresponding joint is implemented considering the joint specification including joint's h-value, rotation axis and angles. Like the developed human digital twin model, it is implemented using *Matlab R2021a* and C++.

One of the functions is to control directly the connected real collaborative robot using the PLC program. The PLC program is supported by a robot controller maker (Allen-Bradley©). The most important function is to transfer the human pose in a developed human digital twin model to a real collaborative robot. The following subsection elaborates on its transfer.

3.4. Motion transfer from human digital twin to collaborative robot digital twin

As described in Section 3.2., a human pose is reasoned using the provided deep learning model and converted into a reasoned human skeleton model. Then, imitation learning is applied to the motion transfer from the developed human digital twin to a collaborative digital twin.

However, imitation learning requires similarity criteria between humans and robot poses. Lei et al. [7] considered four features (hip roll angle, hip pitch angle, knee pitch angle, and ankle pitch angle) to check the similarities between human poses and poses of a humanoid robot. Their research checked the difference between two objects, and all the sum of the differences was compared with a predefined constant. Maeda et al. [38] introduced a couple of oscillators for imitation learning between a human and a robot. Additional oscillating signals are used as low-level features for passing the basketball. Wang et al. [39] introduced a virtual joint-based approach that uses a predefined robot arm–human arm mapping. In their study, a robot was trained using a performance analysis that compared the angular gaps between the human arms and robotic arms. Zuo et al. [17] applied a Q-learning method to train a humanoid robot using a well-trained robot. However, their imitation learning is based on robots with the same configuration. The matching is achieved accurately because a real human skeleton architecture is the same as that of the human digital model. Thus, this coincidence makes it possible to joints' rotation-based mappings.

In the proposed framework, the motion transfer from a human digital twin a robot digital twin is difficult because the two skeleton architectures are different. For instance, while the arm of a human has three joints, a collaborative robot (i.e., *ABB Yumi* Robot, as shown in Fig. 6.) has seven joints. Eq. (18) denotes a function $g(\cdot)$ considering both arms.

$$\left(Y'_{R,1}, \cdots, Y'_{R,7}, Y'_{L,1}, \cdots, Y'_{L,7}\right) = g\left(X'_{R,1}, \cdots, X'_{R,3}, X'_{L,1}, \cdots, X'_{L,3}\right) \tag{18}$$

One simple solution to this problem is to match the number of variables. This indicates that the four output variables are fixed in the robotic arm. However, this solution is limited by the fact that the lengths between the two adjacent joints and joint angles are different. To overcome this issue, a similarity between the human arm pose and the robotic arm is considered. In addition, the removal of several variables among the 14 variables may result in linear movements without smoothness, which, however, may harm the structure of the robot arm.

To overcome this issue, this study introduces a new and effective imitation learning-based teleoperation method using Bezier curve-based similarity. A number of research studies have provided several curve fitting methods from a set of points. While curve approximation [40] is a traditional research subject in geometry, it has been recently applied in various deep learning fields. Lee [41] applied a *non-uniform rational B-spline* (NURBS)-based curve approximation to the development of a sketch-based discrete simulator, where a user's sketch is converted to a stochastic Petri net model using a self-organizing map. While NURBS-based curve approximation enables accurate control, this study uses the Bezier curve-based approximation with the criteria of the lower burden of computation considering human arms' joint structures. As human arms have fixed joint structures and movements, it is determined empirically that the Bezier curve-based approximation is a reasonable motion approximation with the test of the anthropometric data [34]. The Bezier curve is defined by Eq. (19) where CP_i is the i^{th} control point of C(u), where $B_{n,i}(u) = \binom{n}{i} u^i \cdot (1-u)^{n-i}$ and, $0 \le u \le 1$.

$$C(u) = \sum_i^N B_{n,i}(u) \cdot CP_i \tag{19}$$

As shown in Fig. 5(b), both arms of the developed human digital twin consist of linear connections. The mapping from the human CPS model to a collaborative robot CPS model is constructed using a Bezier curve-based approximation, as shown in Fig. 7.

The objective of this Bezier curve-based pose approximation is to obtain the rotation axis and rotation angle (Radian value) of each robot arm joint. The target is the corrected arm pose of the human CPS model, as shown in Fig. 5. To provide more smoothness in a robotic arm, it is assumed that the control points in a Bezier curve ($N = 6$) exist on the linear segments on a human CPS hand. As each robot arm has seven DoFs, the sixth-degree Bezier curve is considered. Then, more control points (CP_i) are generated by considering an equidistance [39] between two adjacent joints.

Then, an approximated Bezier curve is generated using the reasoned control points [42]. The generated Bezier curve was segmented into six curves using the curve parameter ($u, 0 \le u \le 1$). As the arm of the robot CPS consists of seven joints, the curve is segmented into six parts: $\overline{C\left(\frac{1}{6}\right)C(0)}$, $\overline{C\left(\frac{2}{6}\right)C\left(\frac{1}{6}\right)}$, $\overline{C\left(\frac{3}{6}\right)C\left(\frac{2}{6}\right)}$, $\overline{C\left(\frac{4}{6}\right)C\left(\frac{3}{6}\right)}$, $\overline{C\left(\frac{5}{6}\right)C\left(\frac{4}{6}\right)}$, and$\overline{C(1)C\left(\frac{5}{6}\right)}$, Each curve is then linearized to calculate the joint axis and angle, respectively.

The joint axis and angle are calculated using the theorem proposed in Section 3.2. Finally, the robot CPS arms are corrected using the human CPS arms. The overall process is shown in Fig. 7.

Then, the real collaborative robot is controlled using an embedded robot-connection program supported by a robot maker. The following section shows several transferred poses from a human to a collaborative robot, and the proposed framework is compared with existing methods.

4. Experimental results and comparison analysis

This section provides the experimental results of the proposed framework and compares it with existing methods. To check the effectiveness of the proposed framework, several human poses were tested for their transfer to a collaborative robot. Fig. 8. shows several human poses and their transfer to the *ABB Yumi* robot. Most of the errors for pose transfer from a human to a collaborative robot occur due to pose recognition using the modified convolutional encoder-decoder.

Thus, to measure the accuracies of the proposed imitation learning-based teleoperation framework, two accuracy measurements were introduced: accuracy type I and accuracy type II. These accuracies are used for measuring how human motion is transferred to a motion of a cooperative robot closely. Both accuracies are defined in Eqs. (20)–(22). The first accuracy is measured for the gap between two curves: an approximated curve from a human arm and an approximated curve from a human skeleton. Each joint location was checked, and an arm's pose

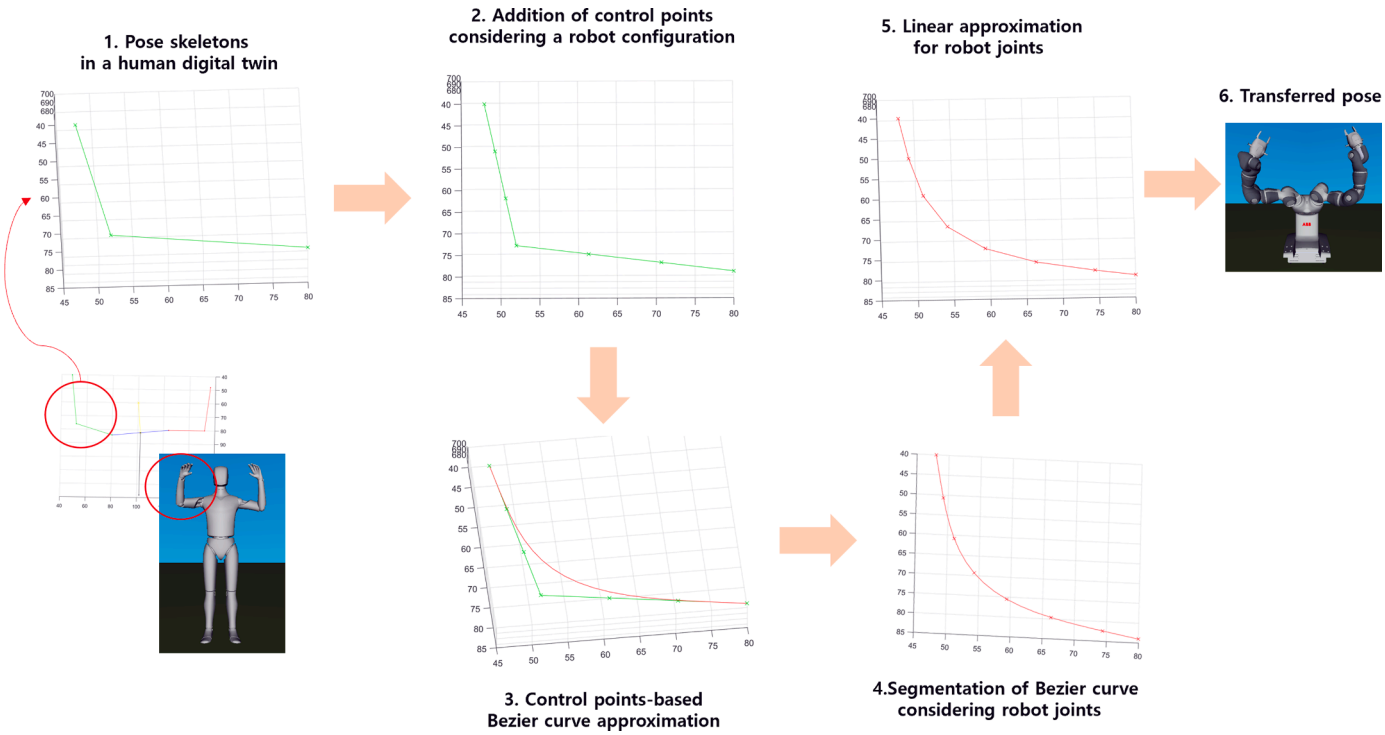


Fig. 7. Bezier curve-based pose approximation of a left arm

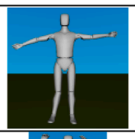
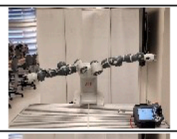
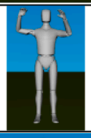
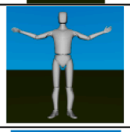
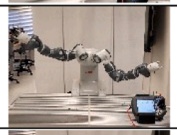
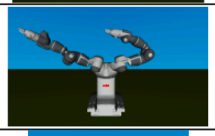
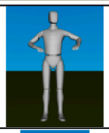
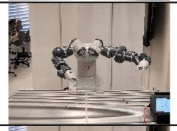
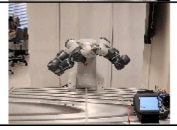
Human poses (2D)	Reasoned human skeleton (3D)	Human digital twin model	Collaborative robot digital twin	Real Robot pose	Accuracy type I (%)
					82.35
					79.96
					81.46
					72.39
					80.25
					86.37

Fig. 8. Pose transfer tests using the proposed framework.

was approximated using the proposed deep learning architecture. Then, the reasoned skeleton using the convolutional deep learning framework was approximated into a Bezier curve, as shown in Eq. (19). The gap between the approximated curves was measured using the gap between two control points. Pande et al. [43] proposed an error between two Bezier curves using control points from both curves. However, their measurement is limited by the fact that their curves are simple cubic Bezier curves. As this study approximates an arm pose with a six-degree

Bezier curve, the error measurement is extended with a higher-degree Bezier curve. In addition, it is normalized with the arm length for comparative analysis. Eq. (20) can be used to measure a comparative gap between a human arm pose and a recognized human skeleton using a deep learning framework. d is a normalization constant for the comparative gap and it can be different with each application. In this study, d is set with 30 using empirical tests.

$$\mathrm{E}(\widetilde{C(u)}, \widetilde{C'(u)}) = \frac{1}{\mathrm{d}}\sum_{i=0}^{6}|P_i - P_i'|\cdot 100(\%) \tag{20}$$

As denoted in Eq. (20), $\widetilde{C(u)}$ and P_i are an curve from a real human pose figure and the i^{th} control point of the human arm pose. Similarly, $\widetilde{C'(u)}$ and P_i' are an approximated curve and the i^{th} control point of the human CPS arm pose. Accuracy type I is an average value considering both arms, as shown in Eq. (21).

$$\text{Accuracy type I}(\%) = 100 - \frac{\mathrm{E}(\widetilde{C(u)}, \widetilde{C'(u)})_L + \mathrm{E}(\widetilde{C(u)}, \widetilde{C'(u)})_R}{2} \tag{21}$$

where $\mathrm{E}(\cdot)_\mathrm{L}$= the error of a left arm pose, and $\mathrm{E}(\cdot)_\mathrm{R}$ = the error of a right arm pose

In order to measure the pose transition accuracy, a real human pose is measured on the spot. And then, the pose is taken on a single vision. The 2D image is used with the proposed framework. The both accuracy types are measured using the real human pose (3D), 2D human pose, approximated human CPS pose (3D), robot digital twin pose (3D) and final real robot pose (3D).

As shown in Fig. 8., the accuracies are comparatively low owing to the applied convolutional deep learning module. While the embedding of more accurate deep learning may improve the accuracy, existing deep learning methods are still limited in their ability for accurate skeleton extraction from a single image.

To solve this issue, a number of cooperative frameworks [44,45] integrating deep learning and human adaptive evaluation have been proposed.

This study applies the direct control of the developed human CPS model to pose transfer. As it is possible to control a human pose directly using the developed human digital twin model, a more accurate pose transfer is achieved. Fig. 9. shows the pose transfer results using the human CPS-collaborative robot CPS mapping. Accuracy type II is defined, as denoted in Eq. (22), where $\widetilde{C'(u)}$ and $\widetilde{C''(u)}$ are approximated six-degree Bezier curves of an arm pose in the developed human CPS model and in the real collaborative robot CPS, respectively.

$$\text{Accuracy type II}(\%) = 100 - \frac{\mathrm{E}(\widetilde{C'(u)}, \widetilde{C''(u)})_L + \mathrm{E}(\widetilde{C'(u)}, \widetilde{C''(u)})_R}{2} \tag{22}$$

where $\mathrm{E}(\cdot)_\mathrm{L}$ = the error of a left arm pose, and $\mathrm{E}(\cdot)_\mathrm{R}$= the error of a right arm pose

As provided in Section 3.4, theoretically, the accuracy type II has to be 100%. However, the control transform from the developed cooperative CPS framework to a real collaborative robot (*ABB Yumi*) has several communication and control errors. These errors result in slight pose transfer errors as shown in Fig. 9. However, it is considered that the use of the developed human CPS and collaborative robot CPS models contributes to more accurate pose imitation.

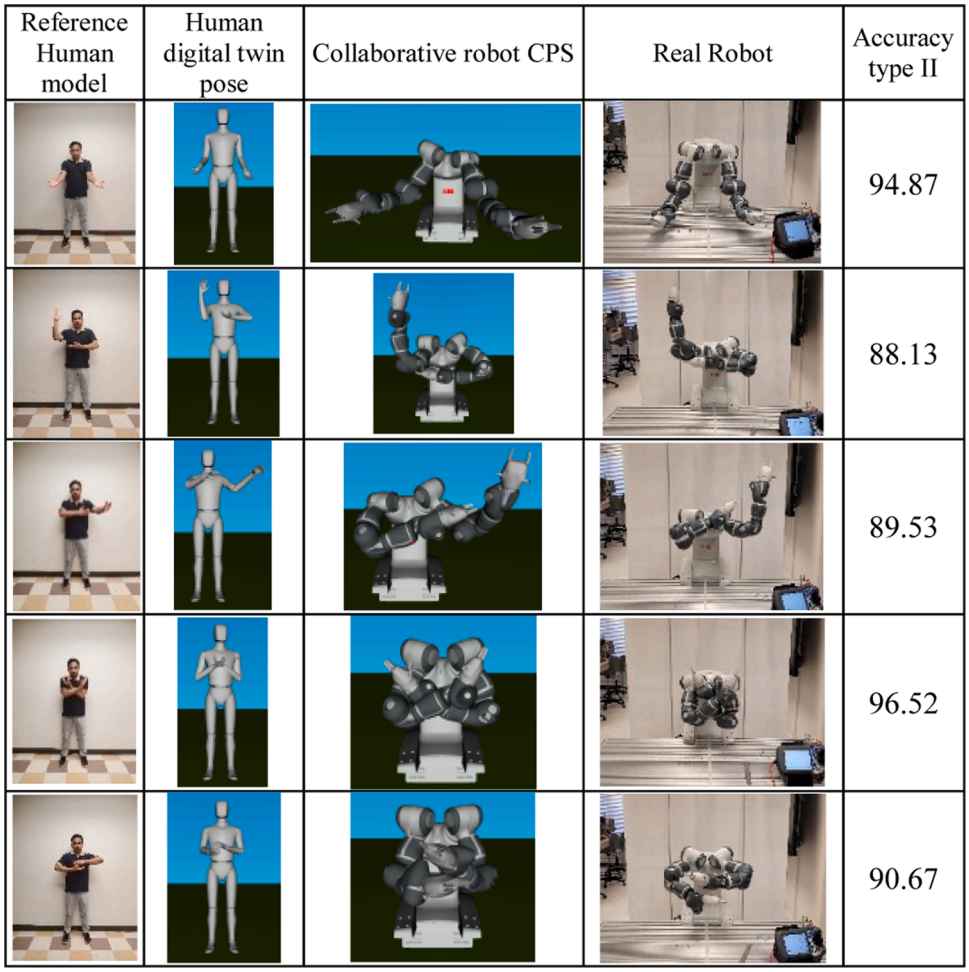


Fig. 9. Pose transfer tests using the semi-automatic control of the proposed framework.

The proposed framework is considered a new and efficient cooperative control framework that uses the two digital twin models. While most of the existing imitation learning methods use direct mapping from human pose to a collaborative robot using several deep learning methods, their pose transfers are limited because these can only transfer comparatively simple poses and their accuracies depend heavily on the efficiency of the reasoning method used. The proposed framework overcomes these issues and is considered an easier control framework for guiding a collaborative human robot.

5. Conclusion and further studies

We proposed a new framework for efficiently transferring human poses to collaborative robots using a human digital twin and a collaborative robot digital twin. A 2D human pose was taken as the input. First, a human figure was reasoned to a human skeletal model using the embedded convolutional encoder-decoder architecture. The estimated skeletal model was used for the pose correction of a developed human digital model. To match the human joints in the digital twin model with the estimated human pose, we provide the rotation axis and angle-related theorem using which a human pose was transferred to a human digital twin. Then, the pose was mapped to the pose of a collaborative robot digital twin. To consider the gap between the DoF of human pose and DoF of a collaborative robot, pose similarity was considered. A human pose was approximated to a set of Bezier curves, and then the robot poses were generated using the Bezier curve-based pose adjustment. Finally, a real collaborative robot was controlled using a PLC program in the implemented robot digital twin.

One advantage of the proposed framework is the direct control of the robot using a human digital twin. While existing imitation learning methods attempt to incorporate more accurate deep learning modules for human-pose capture, inaccurate deep learning modules may mislead the control of a collaborative robot. The proposed framework overcomes this issue by digital twin models. To demonstrate the effectiveness of the proposed framework, various human poses and their pose transfers were tested and analyzed. In terms of the motion transition speed, the proposed framework is developed with a speed target for industrial purposes. However, the proposed framework is limited with a motion transition considering each human motion. The value of this research can be more helpful with the transition from a sequence/video of human poses to a real-time teleoperation system. Even though the provided framework is limited for the full range real-time teleoperation controls, it is valuable for the fundamental framework for real-time teleoperation controls of cooperative robots.

In further studies, consecutive motion transfers should be considered. Furthermore, as more industrial processes are expected to be managed by collaborative robots, a real-time motion transfer technique should be explored. The proposed framework can be expanded for consecutive motion transfers considering various industrial processes. In addition, intelligent control which takes into account unexpected events can be realized using the framework proposed in the present study.

CRediT authorship contribution statement

Hyunsoo Lee: Conceptualization, Methodology, Software, Validation, Formal analysis, Writing – original draft. **Seong Dae Kim:** Writing – review & editing, Supervision, Project administration, Resources. **Mohammad Aman Ullah Al Amin:** Data curation, Visualization.

Declaration of Competing Interest

The authors whose names are listed immediately below certify that they have NO affiliations with or involvement in any organization or entity with any financial interest (such as honoraria; educational grants; participation in speakers' bureaus; membership, employment, consultancies, stock ownership, or other equity interest; and expert testimony or patent-licensing arrangements), or non-financial interest (such as personal or professional relationships, affiliations, knowledge or beliefs) in the subject matter or materials discussed in this manuscript.

Acknowledgment

This research was supported by The Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (grant number: NRF-2021R1A2C1008647) and the Ministry of Education (grant number: NRF-2018R1D1A3B07047113), Republic of Korea. The authors appreciate the help of Dr. Erkan Kaplanoglu with the use of ABB Yumi robot at the University of Tennessee at Chattanooga.

References

[1] Heunis CM, Barata BF, Furtado GP, Misra S. Collaborative surgical robots: optimal tracking during endovascular operations. IEEE Robotics & Automation magazine 2020;27(3):29–44.
[2] Weckenborg C, Kieckhafer K, Muller C, Grunewald M, Spengler TS. Balancing of assembly lines with collaborative robots. Business Research 2020;13:93–132.
[3] Canfield SL, Owens JS, Zuccaro SG. Zero moment control for lead-through teach programming and process monitoring of a collaborative welding robot. Journal of Mechanisms and Robotics 2021;13(3):1–10.
[4] Cai H, Mostofi Y. Human-robot collaborative site inspection under resource constraints. IEEE Transactions on Robotics 2019;35(1):200–15.
[5] Gilmore S, Hillston J, Holton R, Rettelbach M. Specifications in stochastic process algebra for a robot control problem. International Journal of Production Research 1996;34:1065–80. https://doi.org/10.1080/00207549608904950.
[6] Toquica JS, Oliveira PS, Souza WSR, Motta JMST, Borges DL. An analytical and a deep learning model for solving the inverse kinematic problem of an industrial parallel robot. Computers & Industrial Engineering 2021;151:106682.
[7] Lei J, Song M, Li Z, Chen C. Whole-body humanoid robot imitation with pose similarity evaluation. Signal Processing 2015;108:136–46.
[8] Lin J, Hwang K-S. Balancing and reconstruction of segmented postures for humanoid robots in imitation of motion. IEEE Access 2017;5:17534–42.
[9] Schaal S. Is imitation learning the route to humanoid robots? Trends in Cognitive Sciences 1999;3:233–42.
[10] Hussein A, Gaber MM, Elyan E, Jayne C. Imitation learning: a survey of learning methods. ACM Computing Surveys 2017;50(2):1–21.
[11] Ganapathi V, Plagemann C, Koller D, Thrun S. Real-time human pose tracking from rage data. Springer; 2012.
[12] Ionescu C, Papava D, Olaru V, Sminchisescu C. Human3.6M: large scale datasets and predictive methods for 3D human sensing in natural environments. IEEE Transactions on Pattern Analysis and Machine Intelligence 2014;36(7):1–15.
[13] Newell, A., Yang, K., & Deng, J. (2016) Stacked hourglass networks for human pose estimation. arXiv:1603.06937 [Cs], March 22, 2016. https://arxiv.org/abs/1603.06937.
[14] Shafaei, A., & Little, J.J. (2016). Real-time human motion capture with multiple depth cameras. Proceedings of the 13th Conference on Computer and Robot Vision. British Columbia, Canada.
[15] Wei, S.-E., Ramakrishna, V., Kanade, T., & Sheikh, Y. (2016). Convolutional pose machines, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 4724-4732). Las Vegas, NV, USA.
[16] Weiss I, Ray M. Model-based recognition of 3D objects from single images. IEEE Transactions on Pattern Analysis and Machine Intelligence 2001;23(2):116–28.
[17] Zuo G, Zhao Q, Chen K, Li J, Gong D. Off-policy adversarial imitation learning for robotic tasks with low-quality demonstrations 2020;97(106794):1–10.
[18] Liu Y, Zhang W, Pan S, Li Y, Chen Y. Analyzing the robotic behavior in a smart city with deep enforcement and imitation learning using IoRT. Computer Communications 2020;150(2020):346–56.
[19] Al-Yacoub A, Zhao YC, Eaton W, Goh YM, Lohse N. Improving human robot collaboration through force/torque based learning for object manipulation. Robotics and Computer-Integrated Manufacturing 2021;69(102111):1–15.
[20] Lin C, Chen P, Pan Y, Chang C, Huang K. The manipulation of real-time Kinect-based robotic arm using double hand gestures. Journal of Sensor 2020;2020:1–9.
[21] Kwon, D.H., & Gebhardt, R. (2021) An affordable, accessible human motion controlled interactive robot and simulation through ROS and Azure Kinect, Proceedings of Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction, Boulder, Colorado, USA.
[22] Chung MJ, Friesen AL, Fox D, Meltzoff AN, Rao RPN. A Bayesian developmental approach to robotic goal-based imitation learning. Plos One 2015;10(11):1–18.
[23] Koenemann, J., & Bennewitz, M. (2012) Whole-body imitation of human motions with a Nao humanoid, Proceedings of HRI 2012 Conference, Boston, Massachusetts, USA.
[24] Sabbaghi, E., Bahrami, M., & Ghidary, S.S. (2014) Learning of gestures by imitation using a monocular vision system on a humanoid robot. Proceedings of the 2nd RSI/ISM International Conference on Robotics and Mechanics, Tehran, Iran.
[25] Lee, D., Ott, C., Nakamura, Y., & Hirzinger, G. (2011) Physical human robot interaction in imitation learning, Proceedings of 2011 International Conference on Robotics and Automation. Shanghai, China.

[26] Bonardi, A., James, S., & Davison, A.J. (2019) Learning one-shot imitation from humans without humans, arXiv:1911.01103 [Cs], Nov 04, 2019. https://arxiv.org/abs/1911.01103.
[27] Shi, l., Zhang, Y., Cheng, J., & Lu, H. (2019). Skeleton-based action recognition with multi-stream adaptive graph convolutional networks. arXiv:1912.06971. Dec 15,2019. https://arxiv.org/abs/1912.06971.
[28] Marin-Jimenez MJ, Romero-Ramirez FJ, Munoz-Salinas R, Medina-Carnicer R. 3D human pose estimation form depth maps using a deep combination of poses. Journal of Visual Communication and Image Representation 2018;55(1):627–39.
[29] Haque, A., Peng, B., Luo, Z., Alahi, A., Yeung, S., & Fei-Fei, L. (2016). Towards viewpoint invariant 3D human pose estimation. Proceedings of the European Conference on Computer Vision (ECCV) (pp. 160-177). Amsterdam, Netherlands.
[30] Chen, C.-H., & Ramana, D. (2017) 3D human pose estimation = 2D pose estimation + matching. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 7035-7043). Honolulu, HI, USA.
[31] Xiao, B., Wu, H., & Wei, Y. (2018). Simple baselines for human pose estimation and tracking, Proceedings of the European Conference on Computer Vision (ECCV) (pp. 1-16), Munich, Germany.
[32] Miry A, Ali AA. Simulation of inverse kinetic solution for artificial human arm using hybrid algorithm in virtual reality. AI-Mustansiriyah J. Sci. 2013;24(5):425–36.
[33] Denavit J, Hartenbery RS. A kinematic notation for lower-pair mechanisms based on matrices. Trans ASME J. Appl. Mech. 1955;23:215–21.
[34] Lee H, Cha WC. Virtual reality-based ergonomic modeling and evaluation framework for nuclear power plant operation and control. Sustainability 2019;11(9):1–16.
[35] COCO. (2018). COCO Common Objects in Context. Retrieved from https://cocodataset.org. Accessed January 20, 2021.
[36] Kikawada, T. (2021). Human pose estimation with deep learning. Retrieved from https://github.com/matlab-deep-learning/Human-Pose-Estimation-with-Deep-Learning/releases/tag/v1.0.3. Accessed February 15, 2021.
[37] Cao, Z., Hidalgo, G., Simon, T., Wei, S., & Sheikh, Y. (2019) OpenPose: real time multi-person 2D pose estimation using part affinity fields, ArXiv:1812.08008 [Cs], May 30, 2019. https://arxiv.org/abs/1812.08008.
[38] Maeda G, Koc O, Morimot J. Phase portraits as movement primitives for fast humanoid robot control. Neural Network 2020;129:109–22.
[39] Wang Z, Liang R, Chen Z, Liang B. Fast and intuitive kinematics mapping for human-robot motion imitating: a virtual-joint-based approach. IFAC PapersOnLine 2020;53(2):10011–8.
[40] Piegl L, Tiller W. The NURBS Book. Berlin: Springer; 1995.
[41] Lee H. Development of real-time sketch-based on-the-spot process modeling and analysis system. Journal of Manufacturing Systems 2020;54(C):215–26.
[42] Lee H. Cooperative NURBS surface modeling framework using partial control algorithm and concurrent protocol. International Journal of Collaborative Enterprise 2014;4(4):320–36.
[43] Pande, S.D., Patil, U.A., Chinchore, R., & Chetty, M. (2019). Precise approach for modified 2 stage algorithm to find control points of cubic Bezier Curve. Proceedings of the 5th International Conference on Computing Communication Control and Automation. Maharastra, India.
[44] Kim J, Lee H. Cooperative multi-agent interaction and evaluation framework considering competitive networks with dynamic topology changes. Applied Sciences 2020;10(17):1–16.
[45] Kim J, Lee H. Adaptive human-machine evaluation framework using stochastic gradient descent-based reinforcement learning for dynamic competition network. Applied Science 2020;10(7):1–15.

Hyunsoo Lee is a Professor in the School of Industrial Engineering at Kumoh National Institute of Technology, S. Korea, where he has been since 2011. From 2021, he serves as Department Chair. He received his Ph.D. in Industrial and Systems Engineering from the Texas A&M University, College Station, TX, USA in 2010. He received a B.S. from Sungkyunkwan University in 1997, and an M.S. from the POSTECH in 2002. His research interests span nonlinear control, stochastic optimization, intelligent system design and smart manufacturing. He is a director of Virtual Intelligence and Data Optimization / Engineering (V.I.D.E.O.) laboratory (http://kit.kumoh.ac.kr/~hsl).

Seong Dae Kim is Associate Professor in Engineering Management at the University of Tennessee at Chattanooga. He earned Ph.D. in Industrial Engineering at Texas A&M University. His credentials include Project Management Professional (PMP), Associate Certified Analytics Professional (aCAP), and Lean Six Sigma Black Belt. His research and teaching areas include characterizing and identifying hidden risks, decision & risk analysis for energy technologies, data analytics application to transportation, sports, and decision making, technology forecast using TRIZ method, creative problem solving methods for system improvement, Lean Six sigma process improvement, project management, digital twin modeling, and computer simulation of systems.

Mohammad Aman Ullah Al Amin is currently doing my Ph.D. in Industrial Engineering at the University of Texas Arlington started in Fall 2021. He graduated with Masters in Engineering Management from the University of Tennessee at Chattanooga in 2021 where he worked as a research assistant for the Department of Engineering Management and Technology (2019-2021). He also earned a Post-Baccalaureate Certificate in Computational and Applied Statistics from there. Before that, he did his bachelor's in Industrial & Production Engineering from Khulna University of Engineering & Technology back in 2016.