Assessing Collaborative Physical Tasks Via Gestural Analysis

Edgar Rojas-Muñoz, *Member, IEEE*, and Juan Wachs, *Senior Member, IEEE*

***Abstract*—Recent studies have shown that gestures are useful indicators of understanding, learning, and memory retention. However, and specially in collaborative settings, current metrics that estimate task understanding often neglect the information expressed through gestures. This work introduces the physical instruction assimilation (PIA) metric, a novel approach to estimate task understanding by analyzing the way in which collaborators use gestures to convey, assimilate, and execute physical instructions. PIA estimates task understanding by inspecting the number of necessary gestures required to complete a shared task. PIA is calculated based on the multiagent gestural instruction comparer (MAGIC) architecture, a previously proposed framework to represent, assess, and compare gestures. To evaluate our metric, we collected gestures from collaborators remotely completing the following three tasks: block assembly, origami, and ultrasound training. The PIA scores of these individuals are compared against two other metrics used to estimate task understanding: number of errors and amount of idle time during the task. Statistically significant correlations between PIA and these metrics are found. Additionally, a Taguchi design is used to evaluate PIA's sensitivity to changes in the MAGIC architecture. The factors evaluated the effect of changes in time, order, and motion trajectories of the collaborators' gestures. PIA is shown to be robust to these changes, having an average mean change of 0.45. These results hint that gestures, in the form of the assimilation of physical instructions, can reveal insights of task understanding and complement other commonly used metrics.**

***Index Terms*—Collaboration, gestures, knowledge representation, task understanding.**

I. Introduction

EFFECTIVE team collaboration is often necessary to perform shared physical tasks correctly. For example, effective collaboration is paramount in training involving tool manipulation [1], [2]. A key for effective collaboration is to achieve a common understanding of the shared tasks [3]. Specifically, research has shown that gesturing during collaborative task performance can help attaining shared task understanding. For instance, gestures have been found to be crucial indicators of task understanding and learning in educational environments [4]–[6]. Moreover, gestures are linked to problem solving abilities, can decrease working memory load, and help recalling information [7], [8].

Despite of the relevance of gestures in collaborative tasks, the assessment of task understanding is usually performed without considering gestures. Instead, task understanding is typically measured through indirect metrics such as task completion time and number of errors [9], [10], or subjective techniques such as interviews and concept mappings [11]. We argue and provide evidence in this article that gesture performance-related metrics (e.g., morphological and semantic similarities) can provide insights about task understanding.

In our previous work, we introduced the multiagent gestural instructions comparer (MAGIC) architecture, an approach to represent and compare gestures' morphology, semantics, and pragmatics [12]. Using MAGIC, the current work defines an approach to estimate task understanding from the gestures used by the collaborators. This notion of understanding is obtained from how the collaborators use gestures to convey and execute physical instructions. Therefore, our work: 1) introduces the physical instruction assimilation (PIA) metric, a proxy for task understanding based on gestures; 2) compares PIA against other task understanding proxy metrics; and 3) evaluates the sensitivity of our metric to changes in the way gestures are represented by the MAGIC architecture.

The rest of this article is organized as follows. Section II reviews prior work on the importance of gestures, and metrics for task understanding. Section III details the MAGIC architecture and the PIA metric. Section IV details our experimental design to acquire gestures from collaborators performing a physical task remotely, presents two other task understanding proxy metrics to compare PIA against, and details our sensitivity analysis. Section V presents the results of our experimental apparatus, further discussed in Section VI. Section VII concludes this article and expands on future work.

II. Related Work

Gestures are an essential component in knowledge development and in human interactions related to knowledge development. Theories of embodied learning have shown that students can understand scientific concepts better when they use gestures as part of their learning process [13]. Functional magnetic resonance imaging (fMRI) techniques have also shown that gestures can promote learning and knowledge retention of students by mapping their problem-solving strategies to different areas of the brain [14], [15]. Studies have shown that individuals who gestured as they performed memory tasks were able to

Manuscript received December 3, 2019; revised August 1, 2020 and December 10, 2020; accepted December 13, 2020. Date of publication February 8, 2021; date of current version March 12, 2021. This work was supported by the Office of the Assistant Secretary of Defense for Health Affairs under Award W81XWH-14-1-0042. This article was recommended by Associate Editor V. Fuccella. *(Corresponding author: Juan Wachs.)*

The authors are with the School of Industrial Engineering, Purdue University, West Lafayette, IN 47907-2050 USA (e-mail: emuoz@purdue.edu; jpwachs@purdue.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/THMS.2021.3051305.

Digital Object Identifier 10.1109/THMS.2021.3051305

recall more information during the tasks: gesturing allowed students to create lasting and generalizable associations of the concepts to memorize [8], [16]. Gestures can also be used to share knowledge and perspectives of a shared problem, therefore increasing situation awareness and common grounding [17]. Finally, gestures play a role in the speech production process, as both the timing and meaning of them can shape the ideas being conveyed [18].

Additionally, gestures play a major role in collaborative task understanding. Studies have pointed that observing the gestures performed by others can help achieving a better task understanding, as gestures can reveal others' goals and intentions [6], [19]. Moreover, observing others' gestures can also affect how our subsequent actions are planned and executed [20]–[22]. Gesture observation is particularly important in collaborative tasks where mismatches between the information conveyed through gestures and through speech occur [23], [24]. For instance, gestures coproduced with speech tend to include information not present in speech, and ignoring this information can hinder task understanding [25].

Despite of the relevance of gestures in collaborative tasks, gestures are mostly neglected while assessing task understanding. Instead, task understanding is traditionally assessed using subjective and objectives approaches. Subjective approaches encompass interviews, concept mappings, and inquiries [11], and have been employed extensively in areas such as medicine and aviation [26], [27]. These approaches, however, depend on prone-to-biases behavioral observations [28]. Contrarily, objective approaches rely on indirect quantitative measurements of task understanding. Such metrics include task completion time, number of errors, and task completion percentage [29], [30]. While these approaches are broadly adopted, they do not consider the nonverbal information conveyed by the collaborators, which is crucial for proper understanding.

Finally, approaches to abstract and communicate gestures among team members have been studied, especially in tasks were individuals are not colocated. The goal of those approaches is to increase the performance in collaborative physical tasks by representing and transmitting rich and compact gesture abstractions [31]. Some of these approaches include using two-dimensional (2-D) [31], [32] or 3-D [33], [34] gesture visualizations at the collaborator's workspace. However, these approaches share a limiting factor. These systems confine gestures to an illustrative role: a representation of the gestures is generated and transmitted, but is not used to obtain insights regarding the quality of the collaboration. In the rest of this work, we elaborate on how our approach can be used to generate task understanding insights based on how the gestures are performed and assimilated by individuals collaborating remotely.

III. Methodology

This section presents an approach estimate task understanding using gesture analysis. The section is divided into the following two main sections. The first section explains the MAGIC architecture. MAGIC represents gestures with a data structure containing the gesture's morphology, semantics, and pragmatics.

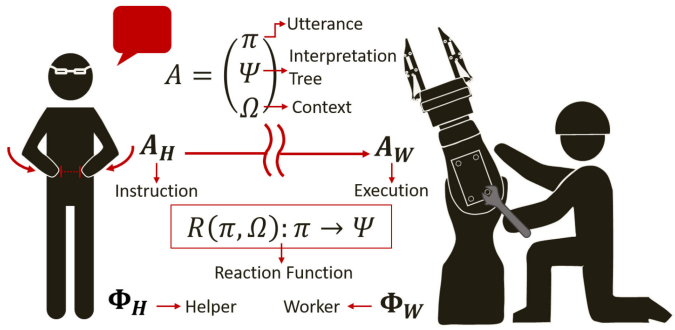

Fig. 1. MAGIC framework. A helper instructs a worker on how to give maintenance to a robotic arm. The elements of an action performed by an agent are linked via the reaction function, a relation in terms of a specific utterance and a given context.

This work presents a summarized explanation of the elements of the MAGIC framework; in-depth explanations can be found in our previous work [12]. The second section introduces the PIA metric for task understanding estimation.

A. Multiagent Gestural Instruction Comparer Architecture

Inspired by two semiotics frameworks [35], [36], MAGIC is an architecture to represent and compare gestures at the morphological (e.g., trajectories, shapes), semantical (e.g., meaning, timing), and pragmatical (e.g., context, environmental elements) levels. In doing so, MAGIC allows us to perform comparisons between gestures that consider more information than just the gestures' appearance. Table I presents the formal definitions introduced in the MAGIC architecture. Fig. 1 presents a schematic of the MAGIC architecture.

B. Reaction Function

MAGIC represents gestures into Ψ interpretation trees, generated from a specific π utterance under a given Ω context. This mapping is modeled via the $R()$ reaction function, a three-stage module that receives π utterances as input and outputs Ψ interpretation trees. The three stages of the $R()$ reaction function include a taxonomy classification of the given gestures to reveal high-level semantics and pragmatics of the gesture; a dynamic semantics framework to represent each gesture as a logical form; and a constituency parsing to generate the Ψ interpretation trees from these logical forms.

1) Gestural Taxonomy Classification: Gestural taxonomies express classification criteria that differentiate gestures from one another [37]. These criteria can include the gestures' communicative intent, expressiveness, iconicity, among other factors. These classifications reveal high-level information regarding the gestures' meaning and context. The first stage of the $R()$ reaction function module leverages a gestural taxonomy to obtain a η classification for each π utterance. These η classifications can be used to identify gestures that either introduce redundancy to the collaboration or are involuntarily performed. MAGIC's gestural taxonomy combines well-known taxonomies, i.e., McNeil, Goodwin, and Poggi's taxonomies [38]–[40], into a single,

TABLE I
ELEMENTS AND NOTATION OF THE MAGIC ARCHITECTURE

MAGIC Element	Definition	Example
Φ_W Worker	Collaborator directly manipulating the environment to perform a task	A person executing actions
Φ_H Helper	Collaborator communicating the commands required to perform a task	A person instructing what actions to perform
A_H Instructions	Action communicating how to perform a task	A verbal command saying "Take a wrench", accompanied by a gesture pointing at a wrench
A_W Executions	Action performed to complete a task	A gesture performed to reach for the wrench
π Utterance	The smallest unit of speech or gesture that communicates a complete idea	A gesture to pinpoint a wrench; a verbal command saying "Stop!"
Ψ Interpretation Tree	MAGIC's data structure representing π Utterances	See Fig. 3
Ω Context	Elements and concepts introduced in the current π_i Utterance, which can be referenced by future π_{i+j} Utterances	Let "Take a wrench" and "Use the wrench to remove that bolt" be the first and second Utterances (π_1 and π_2). The Ω Context of π_2 would include all the elements introduced in π_1, such as the "Wrench" concept.

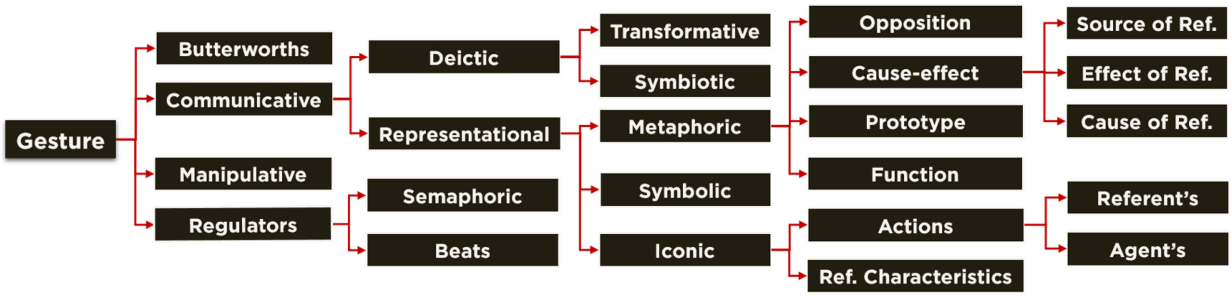

Fig. 2. MAGIC's gestural taxonomy. Classification labels closer to the root provide information about the gestures' symbolical expressiveness, while labels closer to the leaves present fine-grained information such as iconicity.

hierarchy-based model: each η classification is represented as a node in a tree. In this hierarchical classification model, nodes closer to the tree's root reveal coarse information related to the gestures' communicative intent and symbolical expressiveness (e.g., is the gesture communicating a message, or is not a meaningful gesture?). Conversely, nodes closer to the tree's leaves reveal fine-grained information such as iconicity and intention (e.g., is the gesture referring to the Φ_W worker or the Φ_H helper?).

Fig. 2 presents MAGIC's gestural taxonomy. An in-depth explanation of each node in our gestural taxonomy can be found in the referenced literature, as well in Adam Kendon's book [37, p. 84]. As an example, a pointing gesture would be assigned to the "*communicative*" and "*deictic*" η classifications. These labels reveal that the gesture is associated with the transmission of a specifiable message (the "*communicative*" component), and with the determination in space and/or time of a given element (the "*deictic*" component).

2) Extended Segmented Discourse Representation Structure: The second stage of the $R()$ reaction function module abstracts the gesture into a logical form that represents its morphology, semantics, and pragmatics. An example of a framework that accomplishes this goal is segmented discourse representation structure (SDRS), a formal dynamic semantics framework that represents verbal utterances using logical forms [41]. SDRS represents the meaning of utterances through SDRS-formulae. These formulae describe how each utterance modifies the discourse's context (where the discourse is a set containing all the utterances). In their follow-up work, Lascarides and Stone expanded SDRS by integrating attribute-pair tables that described high-level morphological features of the gestures, following Kopp, Tepper, and Cassell's typed feature structures [42], [43]. For example, a pointing gesture would be expressed with the following table:

$$\begin{bmatrix} \textbf{pointing-gesture} \\ \text{right-hand-shape} : \textit{asl-1} \\ \text{right-finger-direction} : \textit{forward} \\ \text{right-palm-direction} : \textit{forward} \\ \text{right-location} : \vec{f} \end{bmatrix}$$

where $\vec{f}$ is the location of the tip of the right index finger. A close inspection of this structure reveals that is not scalable. For example, the structure describes shapes using predefined lexicons (e.g., American Sign Language), making it nonmodular to the variety of gestures humans can generate. Therefore, we propose an extension of the SDRS framework [Extended SDRS, henceforth extended segmented discourse representation structure (ESDRS)] that creates ϕ ESDRS-formulae to describe the gestures' morphology, semantics, pragmatics. These ϕ ESDRS-formulae can represent both verbal and gestural utterances (denoted by the $[\mathcal{G}]$ operator). The atomic elements introduced in

TABLE II
ATOMIC COMPONENTS OF THE ESDRS FRAMEWORK

ESDRS Element	Definition	Example
i Individual Variable	An element introduced by an π Utterance	A wrench
e Eventuality Variable	An event happening during the π Utterance	Grabbing a wrench
p Spatiotemporal Locality	Four-dimensional vectors (x, y, z, t) representing positions in time and space	The location of a wrench
v Virtual Mapping	Transformation from world space to gesture space	When pointing at a wrench, the position of the wrench in world space is mapped to gesture space
Predicate	Clauses describing the interactions between the other atomic components	$Wrench(i)$ assigns the concept of "Wrench" to the i individual variable

the ESDRS framework to construct these ϕ ESDRS-formulae are presented in Table II. Therefore, the second stage of $R()$ reaction function module generates ϕ ESDRS-formulae by receiving a π utterance, a Ω context, and the η classification as inputs.

Consider the following verbal and gestural utterances:

π_1: "Take that wrench"

π_2: *The right arm is extended forward. The right-hand's fingers make a fist shape except the index finger, which is extended and points forward. The hand stays in place.*

The ϕ ESDRS-formulae corresponding to these utterances is presented. $\phi_{2\,T}$, ϕ_{2S}, and ϕ_{2M} are elements within ϕ_2, but presented separate for ease of reading as (*) shown at the bottom of the page.

3) Constituency Parsing: The last stage of the $R()$ reaction function represents the ϕ ESDRS-formulae as data structures to perform quantitative comparisons. MAGIC leverages a constituency parsing to generate these representations, known as Ψ interpretation trees, from the ϕ ESDRS-formulae. Generating tree data structures from logical forms has been studied extensively [44], [45], as trees can represent the attributes of the logical forms and the hierarchy between them.

MAGIC's constituency parsing approach introduces several nested constituents to represent the elements of the ϕ ESDRS-formulae. These constituents can be separated into the following two subgroups: predicative and nonpredicative. The nonpredicative subgroup includes all the elements from the ϕ ESDRS-formulae that are not predicates: discourse referents, spatiotemporal localities, etc. Contrarily, the predicative subgroup includes the predicates, organized based on specific aspects of the gestures. For instance, predicates related to shape will be grouped separate from those related to movement. Table III briefly describes the nested constituents in our approach.

Fig. 3 exemplifies how different gestures generate different Ψ interpretation trees. The nodes highlighted in green represent the atomic components of the ESDRS' framework in the tree, while the nodes highlighted in blue represent the nested constituents of our parsing approach. Creating Ψ interpretation trees is a semiautonomous process. The ShG and MvG constituents are built by approximating the position and orientation in 3-D space of the person's body using a Microsoft Kinect camera. An approximation of the trajectory is constructed by extracting the zero-velocity points in a motion trajectory (points in 3-D space where $\frac{\partial x}{\partial t} = \frac{\partial y}{\partial t} = \frac{\partial z}{\partial t} = 0$). This was later adjusted by an annotator, who approximated this trajectory manually. Other aspects of the gestures such as their meaning and context were labeled manually.

C. Gesture Matching Through Integer Optimization

The last stage of the MAGIC architecture involves matching the collaborators' gestures. Our approach describes gesture similarity by counting the number of shared nodes between the gestures' subtress. First, we construct a B matrix of b_{ij} similarity scores (of size $|W| \times |H|$), where each score describes the similarity between the Ψ_{W_i} worker interpretation tree and the Ψ_{H_j} helper interpretation tree. Let H be a set containing

$$\phi_1 : \exists\, i_1 \left[\begin{array}{c}\text{Wrench}(i_1) \wedge \text{Take}(e_1, i_1) \wedge \\ \text{Loc}(e_1, i_1, v_w(\vec{p_w}))\end{array}\right] \qquad \phi_2 : [\mathcal{G}] \exists\, i_2 \left[\begin{array}{c}\text{Gesture}(i_2) \wedge \text{TaxClass}(i_2) \wedge \text{Shape}(i_2) \wedge \\ \text{Movement}(i_2) \wedge \text{Synchro}(e_1)\end{array}\right]$$

$$\phi_{2\,T} : [\mathcal{G}] \Big[\text{Communicative}(i_2) \wedge \text{Deictic}(i_2)\Big] \qquad \phi_{2\,M} : [\mathcal{G}] \exists i_{M_1} \left[\begin{array}{c}\text{Trajectory}(i_{M_1}, v_I(\vec{p}_1), v_I(\vec{p}_2)) \wedge \\ \text{Main Plane Coronal}(i_{M_1}) \wedge \\ \text{Direction Static}(i_{M_1}) \wedge \\ \text{Component}(i_{M_1}, i_2)\end{array}\right]$$

$$\phi_{2S} : [\mathcal{G}] \exists_{\substack{i_{S_1}, i_{S_2}, i_{S_3}, \\ i_{S_4}, i_{S_5}, i_{S_6}, i_{S_7}}} \left[\begin{array}{c}\text{Arm}(i_{S_1}) \wedge \text{Hand}(i_{S_2}) \wedge \text{Thumb Finger}(i_{S_3}) \wedge \text{Ring Finger}(i_{S_4}) \wedge \\ \text{Middle Finger}(i_{S_5}) \wedge \text{Index Finger}(i_{S_6}) \wedge \text{Little Finger}(i_{S_7}) \wedge \\ \text{Pose Extended}(i_{S_1}) \wedge \text{Orientation Forward}(i_{S_1}) \wedge \text{Pose Not Extended}(i_{S_2}) \wedge \\ \text{Orientation Forward}(i_{S_2}) \wedge \text{Pose Not Extended}(i_{S_3}) \wedge \text{Pose Not Extended}(i_{S_4}) \wedge \\ \text{Pose Not Extended}(i_{S_5}) \wedge \text{Pose Extended}(i_{S_6}) \wedge \text{Orientation Forward}(i_{S_6}) \wedge \\ \text{Pose Not Extended}(i_{S_7}) \wedge \text{Component}(i_{S_1}, i_2) \wedge \text{Component}(i_{S_2}, i_{S_1}) \wedge \\ \text{Component}(i_{S_3}, i_{S_2}) \wedge \text{Component}(i_{S_4}, i_{S_2}) \wedge \text{Component}(i_{S_5}, i_{S_2}) \wedge \\ \text{Component}(i_{S_6}, i_{S_2}) \wedge \text{Component}(i_{S_7}, i_{S_2})\end{array}\right] \qquad (*)$$

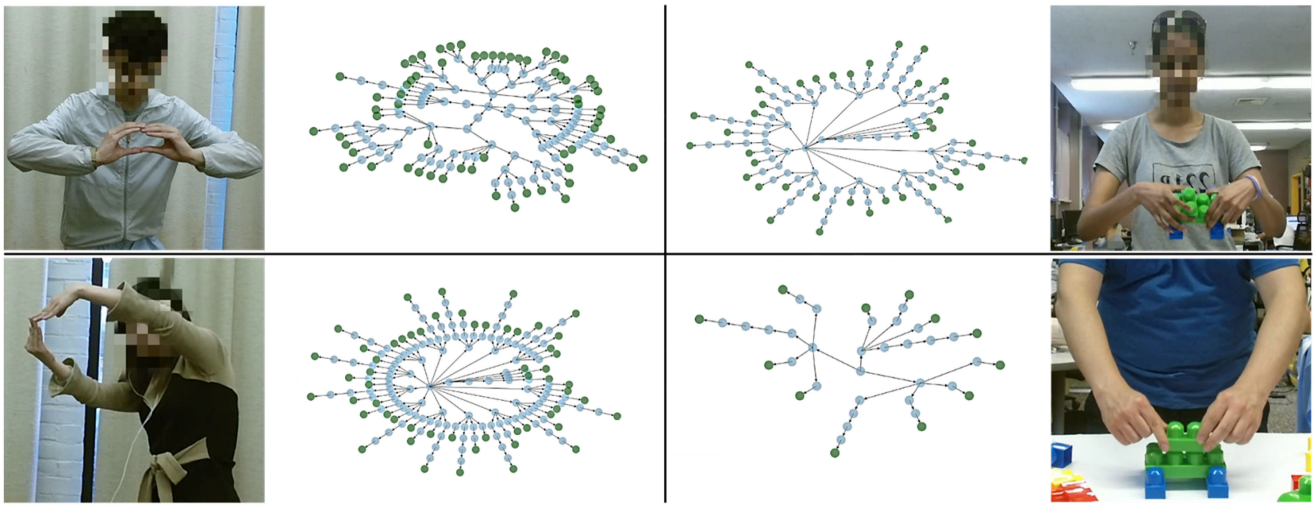

Fig. 3. Gestures and subcomponents of their associated Ψ interpretation trees, generated through the MAGIC architecture. Comparisons between these structures will be performed to calculate a matching between the gestures of the collaborators.

TABLE III
NESTED CONSTITUENTS OF MAGIC'S PARSING APPROACH

Nested Constituent	Type	Contains
Variable Group (VG)	Non-Predicative	i individual and e eventuality variables
Spatiotemporal Group (SG)	Non-Predicative	s spatiotemporal localities
Mapping Group (MG)	Non-Predicative	v virtual mappings
Context Group (CG)	Non-Predicative	References to the discourse referents and predicates from previous π Utterances
Large Predicate Group (LPG)	Predicative	All predicates
Shape Group (ShG)	Predicative	Predicates related to the gesture's shape
Loc Group (LoG)	Predicative	$Loc()$ predicates, related to localizing i individual variables
Exemplifies Group (ExG)	Predicative	$Exemplifies()$ predicates, related to depicting i individual variables with a gesture
TaxClass Group (TaG)	Predicative	Predicates related to the gesture's taxonomy classification
Synchro Group (SyG)	Predicative	$Synchro()$ predicates, related to whether a gesture was performed in synchrony with a given e eventuality variable
Movement Group (MvG)	Predicative	Predicates related to the gesture's movement
Extra Predicates	Predicative	Predicates that do not fit into any of the previous categories

all the h_j helper-authored gestures, and W a set containing all the w_i worker-authored gestures. Additionally, let $\mathcal{X}$ be a nested constituent (e.g., CG, SyG) from a Ψ interpretation tree, and let $\Psi^{\mathcal{X}}$ be the subtree of Ψ with $\mathcal{X}$ as its root (e.g., Ψ^{CG} is a context subtree). Each b_{ij} is computed via the following relation:

$$b_{ij} = \text{num\_nodes}\, \Psi^{\mathcal{X}}_{W_i} \cap \Psi^{\mathcal{X}}_{H_j}; \begin{matrix} i = 1, 2, \ldots, |W| \\ j = 1, 2, \ldots, |H|. \end{matrix}$$

Our previous work described gesture similarity adequately by comparing the Ψ^{CG}_W worker context subtrees against the helper subtrees representing meaning ($\Psi^{\text{ExG}}_H \cup \Psi^{\text{LoG}}_H \cup \Psi^{\text{SyG}}_H$) [12]. The idea behind this comparison was that all A_W executions were performed in response to an A_I Instruction. Therefore, common information should exist between a gesture communicating an instruction, i.e., in its meaning, and a gesture performed in response to it, i.e., in its context. Henceforth, we will use the B similarity matrix of b_{ij} similarity scores obtained from this particular comparison to perform the matching between the gestures.

The b_{ij} similarity scores from the previous step are used as cost coefficients in an integer linear problem (ILP). This ILP finds an E matrix of e_{ij} edge weights (of size $|W| \times |H|$) that optimally describes how the w_i worker-authored gestures match the h_j helper-authored gestures. Each e_{ij} will take a value of 1 if the w_i worker-authored gesture matches the h_j helper-authored gesture are functionally equivalent, and 0 otherwise. Given the W and H gesture sets and the B similarity matrix, the process of finding the optimal E gesture matching matrix is defined as

$$\begin{aligned} \text{maximize} \quad & \sum_{i=1}^{|W|} \sum_{j=1}^{|H|} b_{ij} e_{ij} \\ \text{subject to} \quad & \sum_{j=1}^{|H|} e_{ij} = 1 \quad \forall i \\ & e_{ij} \in \{0, 1\} \\ & i = 1, 2, \ldots, |W|;\ j = 1, 2, \ldots, |H|. \end{aligned}$$

An optimal E matrix will be one that matches the w_i worker-authored gestures and h_j helper-authored gestures that have

higher number of common nodes, as described by the B similarity matrix. Since all A_W executions were generated in response to a specific A_H instruction, our problem constrains each w_i worker-authored gesture to be matched to one and only one h_j helper-authored gesture.

D. Physical Information Assimilation Metric

The PIA score estimates task understanding from the quality of assimilation of the physical instructions exchanged between collaborators. PIA was created based on two theories that describe how mapping others' actions and gestures to our own can enable understanding [5], [46]. According to these theories, the exchange of utterances happens in two phases. The speaker (i.e., the Φ_H helper) presents an utterance to the receiver (i.e., the Φ_W worker). If the receiver generates enough understanding evidence $\mathcal{E}$, the speaker can assume that the receiver understood the utterance. The process in which the receiver provides this evidence $\mathcal{E}$ can be divided into four states, ranging from not noticing the initial utterance (State 0) to the correct understanding of it (State 3). If the receiver did not provide enough evidence $\mathcal{E}$ to support correct understanding (State 3), either the speaker needs to elaborate the utterance, or the receiver needs to generate different evidence $\mathcal{E}$.

We part from the assumption that Clark's framework can be applied to gestural utterances. Our approach assumes that perfect task understanding happens when every helper-authored gesture is mapped to one and only one worker-authored gesture. In other words, the Φ_W worker generated evidence $\mathcal{E}$ that supported correct understanding (State 3) for every utterance given by the Φ_H helper. Contrarily, a task where one helper-authored gesture is mapped to several worker-authored gestures represents poor understanding (i.e., State 3 was not reached for every utterance).

The last stage of our approach approximates the behavior described by Clark and colleagues by generating the PIA score from the E gesture matching matrix found in the previous stage. The score is calculated with the following formula:

$$\text{PIA} = \frac{1}{|H|}\sum_{j=1}^{|H|}\left(\sum_{i=1}^{|W|} e_{ij}\right)^{-1}.$$

The maximum PIA score (100) will be found when no h_j helper-authored gesture is matched with more than one w_i worker-authored gesture. Contrarily, the PIA score will be less than 100 when: 1) multiple w_i worker-authored gestures are matched to the same h_j helper-authored gesture; or 2) at least one h_j helper-authored gesture was not matched to any w_i worker-authored gesture. As previously mentioned, the PIA score will approximate how well the physical instructions given by the collaborators being are received and executed. This, according to Clark and colleagues' framework, could be considered as a way to approximate task understanding between collaborators. The following section describes our experimental apparatus to validate our PIA score as an estimator of task understanding, comparing it against other metrics commonly used to estimate task understanding.

IV. Experimental Apparatus

A. Participants

Three user studies were conducted with a total of 60 participants (graduate students, 34 males, and 26 females, age 26.36 $\pm$ 4.4 years old). The participants were divided into 30 helper–worker pairs to collaboratively complete a task.

B. Procedure

To explore the generalizability of our approach, the task to complete was different in each user study. The first task was block assembly, similar to [31] and [47], were participants followed 24 steps to assemble a block helicopter. The second task was origami, similar to [48], were participants followed 11 steps to fold a paper sheet into a paper hat. The third task was ultrasound training, similar to [49], [50], where participants used an ultrasound probe to locate veins, extract blood and find a foreign body using an ultrasound phantom.

Our experimental setup evaluated whether the quality of assimilation of physical instructions (in the form of our PIA metric) can estimate task understanding. This was evaluated by comparing PIA against two other proxy metrics for task understanding: error rate and idle-time rate [9], [51]. Error rate was calculated as the rate between the instructions in which the Φ_W worker made errors over the total number of instructions received. For instance, an error was counted each time the Φ_W worker picked an incorrect block or connected blocks in an incorrect manner. Conversely, the idle-time rate was defined as the rate between the time in which the Φ_W worker did not perform any action (e.g., the time spent thinking or asking questions), and the total task completion time, in seconds. Listening to the h_j helper-authored instructions was not considered idle time. Members of the research team annotated these metrics after the completion of all trails.

Additionally, we performed a sensitivity analysis of the PIA metric. Four factors were selected to inspect how changes in timing and meaning of the gestures represented with MAGIC could affect the PIA scores. The first factor, *delay*, introduced a time delay of d seconds in the time in which the gestures of the Φ_H helper were performed. For instance, if t_1 represents the time of the h_1 helper-authored gesture, the *delay* factor will change it to be $t_1 + d$. We set d to be a random value between 0 and the total task completion time. The second factor, *stretch*, stretched the timeline of w_i worker-authored gestures by an increasing St value. For example, the *stretch* factor will change the time of the h_1 helper-authored gesture to be $t_1 \times St$. We set St as a random value between 1 and 100.

The third factor, *trajectory*, performed an affine transformation in the 3-D positions of the body joints acquired with the Kinect. For example, if the point $p = (x, y, z)$ represents the 3-D position of a body joint, the *trajectory* factor will change it to be $p' = Ap$, where A is an affine transformation matrix. We set the A matrix to be the combination of three rotation matrices (θ° rotations over each axis, with θ between 0 and 90) and a translation matrix with random translation coefficients for each axis (between 0 and half the axis length). Finally,

the fourth factor, *order*, randomly changed the order of the w_i worker-authored gestures in the W set. For instance, if the W set has three gestures, w_1, w_2, w_3, an output of the *order* factor could be a W' set with gestures ordered as w_2, w_3, w_1. With the exception of the *trajectory* factor, these factors can be used interchangeably between worker-authored or helper-authored gestures. The evaluation of each of these factors with both worker-authored and helper-authored gestures is not included in the manuscript for conciseness purposes.

C. Apparatus

After signing a written consent form (IRB Protocol #1810021222), participants were randomly assigned to either the Φ_W worker or the Φ_H helper role, and were directed to different stations according to their role. The Φ_W worker station included the elements to complete the task placed on a table (e.g., blocks, paper sheet, ultrasound probe, and phantom), color and Kinect cameras to record the participant's activity, and a display connected a computer hosting a Skype video call with the helper station. The Φ_H helper station had the same setup, but replaced the elements to complete the task with a booklet containing written instructions on how to complete the task. Therefore, only the Φ_H helper knew the steps to complete the task, and only the Φ_W worker could interact with the elements to complete it. The Φ_H helper conveyed the instructions in the booklet to the Φ_W worker through verbal instructions, gestures, facial expressions, among others.

A total of 3498 gestures (2101 performed by the Φ_H helper, 1397 performed by the Φ_W worker) were extracted over the span of 14 h of video. Each of these gestures was represented using an Ψ interpretation tree, and was considered in the PIA calculation. Our work does not generate Ψ interpretation trees from verbal utterances. Instead, the information introduced by them (e.g., discourse referents, predicates) is considered part of the Ω_W context of the gestural utterances. The reason for this is that PIA is not used to compare verbal utterances, as they inherently lack the morphological aspects of gestures (e.g., shape, movement).

D. Design

A correlation analysis using the Pearson product moment correlation (r) was performed to analyze the relationship between PIA and the other task understanding proxy metrics [52]. The PIA score was treated as the dependent variable, while the error rate and idle-time rate scores were treated as independent variables. For the sensitivity analysis, a two-level four-factor Taguchi design was performed to evaluate PIA's sensitivity to these factors [53]. The PIA score was treated as the dependent variable, while the *delay*, *stretch*, *trajectory*, and *order* factors were treated as the independent factors. We report both the change in the means and the noise-to-signal (SN) ratio after changing the levels of the four independent factors. For each helper–worker pair, 16 (2^4) combinations of factors were evaluated (e.g., only *stretch*, only *trajectory*, and *order*, all at once). These 16 combinations were evaluated for each of the 30 helper–worker pairs in the three tasks, for a total of 480 permutations. To keep the number of tests as reduced as possible, we do not present multiple permutations of each of the combinations over the helper–worker pairs. Additionally, all random values in the sensitivity analysis follow a uniform distribution.

V. Results

Table IV summarizes the results of the task understanding proxy metrics. Higher PIA values estimate better task understanding, while lower error and idle-time rates estimate better task understanding. Fig. 4 shows the Pearson correlation matrix. PIA reported significant correlations with all other metrics ($p \leq 0.05$). In terms of error rate across tasks, the average number of actions per helper–worker pair was 38.57 ± 15.76, and the average number of errors was 11.36 ± 4.81. In terms of idle time across tasks, the average task completion time per helper–worker pair was 13:18 $\pm$ 5:54 min, and the average idle-time rate was 2:03 $\pm$ 1:26 min.

The sensitivity analysis revealed that PIA was robust to changes in the MAGIC architecture. The *delay* factor caused a mean change of 0.96 in PIA ($p \leq 0.01$), the *stretch* factor caused a mean change of 0.52 in PIA ($p = 0.09$); the *trajectory* factor caused a mean change of 0.31 in PIA ($p = 0.3$), and the *order* factor caused a mean change of 0.03 in PIA ($p = 0.9$). Similarly, the *delay* factor caused an SN ratio change of 0.15 ($p = 0.04$), the *stretch* factor caused an SN ratio change of 0.04 ($p = 0.3$); the *trajectory* factor caused an SN ratio change of 0.05 ($p = 0.2$), and the *order* factor did not cause an SN ratio change ($p = 0.9$). Therefore, only the *delay* factor had a significant effect in the calculation of PIA.

VI. Discussion

People heavily rely on gestures to convey physical instructions. Our results reported that participants performed more than 10 gestures per minute, corroborating the importance of gestures in collaborative tasks [6], [31]. This systematic use of gestures and bodily actions reinforces the idea that they act as a proxy for understanding: people's task understanding can become clear by observing someone's actions [54]. The correlation matrices revealed statistically significant correlations between PIA and all the other proxy metrics. For example, when understanding was poor due to more time spent cognitively processing instructions or high error rates, understanding was also poor due to low assimilation of physical instructions (in the form of PIA). These correlations indicate that PIA can act as a proxy to estimate task understanding: PIA showed the same trends found via idle-time rate and number of errors.

Moreover, PIA helped in scenarios where the other metrics were inconsistent. For instance, consider the scores obtained by the B-4 helper–worker pair (error rate = 22.50, idle-time rate = 22.16). This pair reported the lowest error rate within this task, but the second-to-highest idle rate. However, their low PIA score reveals that they performed several unnecessary gestures while completing the task. This PIA score implies that although this pair performed very few errors, they did not have proper understanding of the task. PIA, therefore, provides a gesture-based perspective to estimate task understanding, which is currently not encompassed in other approaches.

TABLE IV
COMPARISON OF TASK UNDERSTANDING PROXY METRICS ON THREE COLLABORATIVE TASK (B = BLOCK ASSEMBLY; O = ORIGAMI; U = ULTRASOUND TRAINING)

H-W Pair	PIA	Error Rate	Idle Time Rate	H-W Pair	PIA	Error Rate	Idle Time Rate	H-W Pair	PIA	Error Rate	Idle Time Rate
B-1	61.30	34.78	17.67	O-1	72.45	24.13	4.40	U-1	52.86	35.14	18.31
B-2	72.14	26.67	6.46	O-2	51.43	33.33	26.16	U-2	65.00	27.50	10.44
B-3	59.04	28.13	5.23	O-3	43.37	42.85	31.09	U-3	75.24	16.42	5.13
B-4	55.43	22.50	22.16	O-4	44.56	62.50	41.61	U-4	62.22	30.91	5.55
B-5	66.13	22.73	8.28	O-5	61.28	35.00	23.08	U-5	60.71	30.91	12.36
B-6	43.38	40.91	24.50	O-6	53.12	41.18	37.27	U-6	55.71	37.29	8.93
B-7	72.55	20.59	5.52	O-7	53.90	33.33	22.76	U-7	52.24	38.30	6.30
B-8	51.69	25.00	12.02	O-8	43.27	37.93	40.09	U-8	79.69	10.53	2.11
B-9	57.78	25.81	11.28	O-9	53.97	33.33	35.92	U-9	72.53	13.73	2.66
B-10	52.04	34.21	20.77	O-10	64.47	33.33	27.13	U-10	47.69	41.86	13.71

Results indicate percentages over 100 [%].

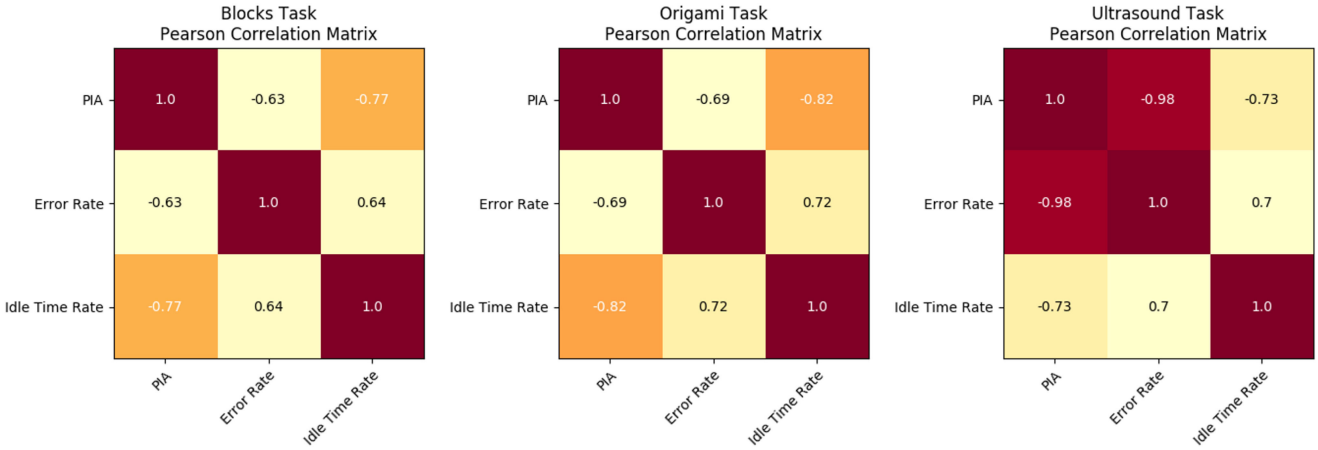

Fig. 4. Pearson correlation matrices for the task understanding proxy metrics. All metrics reported significant correlations in all the tasks ($p \leq 0.05$).

The sensitivity analysis revealed that PIA was robust to changes in the MAGIC architecture. Specifically, we evaluated the effect of changes in time, order, and motion trajectories of the collaborators' gestures. It is important to note that since the comparisons are mostly dependent on the subtrees representing meaning and context, the effects of time, order, and trajectories are limited to nodes within subtress encompassing time, order, and trajectories information. The effect of time (i.e., *delay* and *stretch* factors) led to changes of almost 1 and 0.5 point in the PIA score. This effect was the highest among the factors, reporting statistical significance for the *delay* factor. The importance of time has to do with how it is represented in the Ψ interpretation trees. Time is considered as part of the Synchro Group to describe whether the gesture was performed in synchrony with an e eventuality variable, and as part of the movement group to describe the trajectories of the gestures. Since the Ψ_H^{SyG} subtree representing the synchro group is part of the subtrees representing meaning, *delay* and *stretch* led to the highest effect among the sensitivity factors.

Moreover, since the Ψ_H^{MvG} subtree representing the movement group was not an integral part of the calculation of gesture similarity for this work, the PIA score was robust against the changes in the *trajectory* factor. However, this effect could be higher if evaluating similarity only against the gestures' morphology. Finally, the effect of the *order* factor was negligible. This demonstrates that PIA is not affected by how the W set is ordered. This happens because: 1) gesture similarity is described as the number of common nodes within the Ψ interpretation trees, and 2) the ILP solves for all e_{ij} edge weights at the same time, instead of solving iteratively.

Additionally, PIA can offer significant insights regarding the overall quality of the understanding. For example, thresholds can be set to distinguish between "good," "decent," and "bad" assimilation of physical instructions. Elaborating, if a threshold of 70 were to be empirically set as an indicator of "decent" task understanding, only 20% of the helper–worker pairs would have achieved this goal.

Another advantage of PIA is its generalizability. We evaluated PIA using three physical tasks, and found significant correlations with the other metrics. This hints that, as long as the task has a high number of physical instructions, PIA can be a useful task understanding estimator. Moreover, PIA is independent of

the framework used to compute the E matrices required for its calculation. Our work leveraged MAGIC to generate these E matrices [12], but other gesture matching approaches could also be used to generate the E matrices.

However, the PIA metric should not be used in isolation from the other proxy metrics: we envision PIA as a complementary metric to assess understanding in physical tasks. For instance, consider a hypothetical example of a helper–worker pair that achieved a 100 PIA score: every h_j helper-authored gesture was matched to one and only one w_i worker-authored gesture, and no h_j helper-authored gesture remained unmatched. While this represents perfect assimilation of physical instructions, it does not capture whether the A_W executions were correct: errors could have been performed along the way and, therefore, have a nonzero error rate. This example showcases how these metrics should be used in combination, as each captures different aspects of task understanding.

The PIA metric presents some limitations. First, the metric does not consider verbal-only utterances. For example, the metric does not take into account a verbal instruction indicating where to connect a block if the instruction was not accompanied by a gesture. Therefore, the PIA metric will not act as an adequate proxy metric to measure understanding in tasks where the Φ_H helper decides not to accompany the instructions with gestures, or in tasks that do not involve physical instructions. Moreover, the manual annotation introduces a bottleneck in our approach. Although we follow a semiautonomous process to annotate the gestures, annotating each gesture can take up to 1 min on average. Machine learning techniques could be integrated into our framework to address such constraints. For instance, image captioning techniques with context attention could be incorporated to automatically obtain the gestures' pragmatics [55]. Likewise, speech-to-text routines could be used to obtain information from verbal instructions autonomously [56].

In terms of future work, the assumption that high PIA scores are linked to a one-to-one gesture matching can be further inspected. Currently, the PIA metric penalizes the cases in which more than one worker-authored gesture is required to perform the instruction conveyed with one helper-authored gesture. This can be an issue in scenarios where more than one gesture is necessary to complete an instruction. For instance, if the pointed wrench would have been under a pile of tools, the Φ_W worker would need to perform gestures to move the other tools before taking the wrench. This would have been penalized by the PIA score, as they count as extra gestures. Furthermore, studies should be conducted with different population groups to prevent biases. A population of only graduate students can introduce biases because of the age of the participants and their level of education [57], [58]. Furthermore, gestures are culture-dependent, and they can have different meaning based on the collaborators' culture [59]. Subsequently, future experimental designs could be performed with a time limit to generate other proxy metrics such as the task completion percentage. Finally, the PIA scores should be compared against human-reported understanding evaluations.

VII. Conclusion

This work presented the PIA metric, a novel approach to estimate task understanding via gestural analysis. The metric obtains insights of task understanding by analyzing the way in which collaborators use gestures to convey, assimilate, and execute physical instructions. The gestures of helper–worker pairs of individuals collaborating to complete shared tasks were abstracted and compared using the MAGIC architecture's gesture representation routines. After this, an integer optimization problem was solved to obtain the PIA scores of each helper–worker pair. These PIA scores were compared against two other metrics used to estimate task understanding: number of errors performed and amount of idle time during the task. The results showed that PIA had statistically significant correlations with the two other metrics. In addition, a Taguchi design was used to evaluate PIA's sensitivity to changes in the MAGIC architecture. This analysis revealed that the PIA metric was robust to changes in time, order, and motion trajectories of the collaborators' gestures. Overall, the results reported in this work indicate that the assimilation of physical instructions can reveal insights of task understanding. Moreover, PIA is agnostic to the task being performed and to the framework utilized to match the gestures, making it more generalizable. This implies that gestures can provide a new perspective to estimate task understanding that is currently not encompassed in other common approaches.

References

[1] G. Kurillo, R. Bajcsy, K. Nahrsted, and O. Kreylos, "Immersive 3D environment for remote collaboration and training of physical activities," in *Proc. IEEE Virtual Reality Conf.*, 2008, pp. 269–270.

[2] S. Gauglitz, C. Lee, M. Turk, and T. Höllerer, "Integrating the physical environment into mobile remote collaboration," in *Proc. 14th Int. Conf. Hum.-Comput. Interact. Mobile Devices Serv.*, 2012, pp. 241–250.

[3] D. E. Kieras and S. Bovair, "The role of a mental model in learning to operate a device," *Cogn. Sci.*, vol. 8, no. 3, pp. 255–273, 1984.

[4] S. Goldin-Meadow and S. L. Beilock, "Action's influence on thought: The case of gesture," *Perspectives Psychol. Sci.*, vol. 5, no. 6, pp. 664–674, 2010.

[5] G. Knoblich and N. Sebanz, "The social nature of perception and action," *Current Directions Psychol. Sci.*, vol. 15, no. 3, pp. 99–104, 2006.

[6] R. M. Ping, S. Goldin-Meadow, and S. L. Beilock, "Understanding gesture: Is the listener's motor system involved?," *J. Exp. Psychol.: Gen.*, vol. 143, no. 1, pp. 195–204, 2014.

[7] D. Frick-Horbury, "Use of hand gestures as self-generated cues for recall of verbally associated targets," *Amer. J. Psychol.*, vol. 115, no. 1, pp. 1–20, 2002.

[8] S. W. Cook, T. K. Yip, and S. Goldin-Meadow, "Gestures, but not meaningless movements, lighten working memory load when explaining math," *Lang. Cogn. Process.*, vol. 27, no. 4, pp. 594–610, 2012.

[9] G. Hoffman, "Evaluating fluency in human-robot collaboration," in *Proc. Int. Conf. Hum.-Robot. Interact. Workshop Hum. Robot. Collaboration*, 2013, vol. 381, pp. 1–8.

[10] P. Barmby, T. Harries, S. Higgins, and J. Suggate, "How can we assess mathematical understanding," in *Proc. 31st Conf. Int. Group Psychol. Math. Educ.*, 2007, vol. 2, pp. 41–48.

[11] R. White and R. Gunstone, *Probing Understanding*. Evanston, IL, USA: Routledge, 2014.

[12] E. Rojas-Muñoz and J. P. Wachs, "MAGIC: A fundamental framework for gesture representation, comparison and assessment," in *Proc. 14th Int. Conf. Autom. Face Gesture Recognit.*, 2019, pp. 1–8.

[13] L. Rueckert, R. B. Church, A. Avila, and T. Trejo, "Gesture enhances learning of a complex statistical concept," *Cogn. Res.: Princ. Implications*, vol. 2, no. 1, pp. 1–6, 2017.

[14] E. M. Wakefield, E. L. Congdon, M. A. Novack, S. Goldin-Meadow, and K. H. James, "Learning math by hand: The neural effects of gesture-based instruction in 8-year-old children," *Attention, Perception, Psychophys.*, vol. 81, no. 7, pp. 2343–2353, 2019.
[15] M. Aldugom, K. Fenn, and S. W. Cook, "Gesture during math instruction specifically benefits learners with high visuospatial working memory capacity," *Cogn. Res.: Princ. Implications*, vol. 5, no. 1, pp. 1–12, 2020.
[16] E. L. Congdon, M. A. Novack, N. Brooks, N. Hemani-Lopez, L. O'Keefe, and S. Goldin-Meadow, "Better together: Simultaneous presentation of speech and gesture in math instruction supports generalization and retention," *Learn. Instruction*, vol. 50, pp. 65–74, 2017.
[17] C. Hilliard and S. W. Cook, "Bridging gaps in common ground: Speakers design their gestures for their listeners," *J. Exp. Psychol.: Learn., Memory, Cogn.*, vol. 42, no. 1, pp. 91–103, 2016.
[18] Y. Gao, Y. Liu, and C. Zhou, "Production and interaction between gesture and speech: A review," *Int. J. English Linguistics*, vol. 6, no. 2, pp. 131–138, 2016.
[19] S. Kang and B. Tversky, "From hands to minds: Gestures promote understanding," *Cogn. Res.: Principles Implications*, vol. 1, no. 1, pp. 1–15, 2016.
[20] S. Walther and V. A. Mittal, "Why we should take a closer look at gestures," *Schizophr Bull.*, vol. 42, pp. 259–261, 2016.
[21] N. Sebanz, H. Bekkering, and G. Knoblich, "Joint action: Bodies and minds moving together," *Trends Cogn. Sci.*, vol. 10, no. 2, pp. 70–76, 2006.
[22] E. De Stefani and D. De Marco, "Gesture, language and emotional communication: An embodied view of social interaction," *Front. Psychol.*, vol. 10, 2019, Art. no. 2063.
[23] W. Pouw, T. van Gog, R. A. Zwaan, and F. Paas, "Are gesture and speech mismatches produced by an integrated gesture-speech system? A more dynamically embodied perspective is needed for understanding gesture-related learning," *Behav. Brain Sci.*, vol. 40, 2017, Art. no. e68.
[24] T. Koschmann, "Why would the discovery of gestures produced by signers jeopardize the experimental finding of gesture-speech mismatch?," *Behav. Brain Sci.*, vol. 40, 2017, Art. no. e60.
[25] D. J. Gibson, E. A. Gunderson, E. Spaepen, S. C. Levine, and S. Goldin-Meadow, "Number gestures predict learning of number words," *Devlop. Science*, vol. 22, no. 3, 2019, Art. no. e12791.
[26] M. DeJonckheere *et al.*, "Using text messaging, social media, and interviews to understand what pregnant youth think about weight gain during pregnancy," *J. Med. Internet Res. Formative Res.*, vol. 3, no. 2, 2019, Art. no. e11397.
[27] T. J. Mavin and G. Dall'Alba, "Understanding complex assessment: A lesson from aviation," in *Proc. 4th Int. Conf. Educ., Res. Innov.*, 2011, pp. 6563–6570.
[28] E. Q. Toy, "Gender bias in phone interviews: When male voices pay off," Ph.D. dissertation, Dept. Psychol., San Francisco State Univ., San Francisco, CA, USA, 2019.
[29] K. Wilcox, J. Laran, A. T. Stephen, and P. P. Zubcsek, "How being busy can increase motivation and reduce task completion time," *J. Personality Social Psychol.*, vol. 110, no. 3, 2016, Art. no. 371.
[30] T. Li, C. J. Manns, C. North, and K. Luther, "Dropping the baton? Understanding errors and bottlenecks in a crowdsourced sensemaking pipeline," in *Proc. ACM Hum.-Comput. Interact.*, 2019, vol. 3, pp. 1–26.
[31] D. Kirk, T. Rodden, and D. S. Fraser, "Turn it this way: Grounding collaborative action with remote gestures," in *Proc. Special Int. Group Comput.-Hum. Interact. Conf. Hum. Factors Comput. Syst.*, 2007, pp. 1039–1048.
[32] N. Yamashita, K. Kaji, H. Kuzuoka, and K. Hirata, "Improving visibility of remote gestures in distributed tabletop collaboration," in *Proc. Conf. Comput. Supported Cooperative Work*, 2011, pp. 95–104.
[33] W. Huang and L. Alem, "Handsinair: A wearable system for remote collaboration on physical tasks," in *Proc. Conf. Comput. Supported Cooperative Work Companion*, 2013, pp. 153–156.
[34] W. Huang, S. Kim, M. Billinghurst, and L. Alem, "Sharing hand gesture and sketch cues in remote collaboration," *J. Vis. Commun. Image Representation*, vol. 58, pp. 428–438, 2019.
[35] C. W. Morris, "Foundations of the theory of signs," in *International Encyclopedia of Unified Science*. Berlin, Germany: Springer, 1938, pp. 1–59.
[36] T. A. Sebeok and M. Danesi, *Forms of Meaning: Modeling Systems Theory and Semiotic Analysis*. Roslyn, NY, USA: Walter de Gruyter, 2012.
[37] A. Kendon, *Gesture: Visible Action As Utterance*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[38] D. McNeill, *Hand and Mind: What Gestures Reveal About Thought*. Chicago, IL, USA: Univ. Chicago Press, 1992.
[39] C. Goodwin, "The body in action," in *Discourse, the Body, and Identity*. Berlin, Germany: Springer, 2003, pp. 19–42.
[40] I. Poggi, "Iconicity in different types of gestures," *Gesture*, vol. 8, no. 1, pp. 45–61, 2008.
[41] N. Asher and A. Lascarides, *Logics of Conversation*. Cambridge, U.K.: Cambridge Univ. Press, 2003.
[42] S. Kopp, P. Tepper, and J. Cassell, "Towards integrated microplanning of language and iconic gesture for multimodal output," in *Proc. 6th Int. Conf. Multimodal Interfaces*, 2004, pp. 97–104.
[43] A. Lascarides and M. Stone, "A formal semantic analysis of gesture," *J. Semantics*, vol. 26, no. 4, pp. 393–449, 2009.
[44] M. Purver, A. Eshghi, and J. Hough, "Incremental semantic construction in a dialogue system," in *Proc. 9th Int. Conf. Comput. Semantics*, 2011, pp. 365–369.
[45] A. Okhotin, "Hardest languages for conjunctive and Boolean grammars," *Inf. Comput.*, vol. 266, pp. 1–18, 2019.
[46] H. H. Clark and E. F. Schaefer, "Collaborating on contributions to conversations," *Lang. Cogn. Processes*, vol. 2, no. 1, pp. 19–41, 1987.
[47] S. R. Fussell, L. D. Setlock, J. Yang, J. Ou, E. Mauer, and A. D. Kramer, "Gestures over video streams to support remote collaboration on physical tasks," *Hum.-Comput. Interact.*, vol. 19, no. 3, pp. 273–309, 2004.
[48] S. Kim, G. Lee, W. Huang, H. Kim, W. Woo, and M. Billinghurst, "Evaluating the combination of visual communication cues for HMD-based mixed reality remote collaboration," in *Proc. Conf. Hum. Factors Comput. Syst.*, 2019, pp. 1–13.
[49] R. Amini, J. Z. Kartchner, L. A. Stolz, D. Biffar, A. J. Hamilton, and S. Adhikari, "A novel and inexpensive ballistic gel phantom for ultrasound training," *World J. Emerg. Med.*, vol. 6, no. 3, pp. 225–228, 2015.
[50] S. Thorn, N. Gopalasingam, T. F. Bendtsen, L. Knudsen, and E. Sloth, "A technique for ultrasound-guided blood sampling from a dry and gel-free puncture area," *J. Vascular Access*, vol. 17, no. 3, pp. 265–268, 2016.
[51] I. Martinez-Moyano, "Exploring the dynamics of collaboration in interorganizational settings," *Creating a Culture of Collaboration: The International Association of Facilitators Handbook*, vol. 4. San Francisco, CA, USA: Jossey-Bass, 2006, pp. 69–85.
[52] K. Pearson, "Note on regression and inheritance in the case of two parents," *Proc. Roy. Soc. London*, vol. 58, pp. 240–242, 1895.
[53] G. S. Peace, *Taguchi Methods: A Hands-On Approach*. Reading, MA, USA: Addison-Wesley, 1993.
[54] S. Dekker, *Field Guide to Understanding 'Human Error'*. Boca Raton, FL, USA: CRC Press, 2017.
[55] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, "Paying more attention to saliency: Image captioning with saliency and context attention," *Trans. Multimedia Comput., Commun., Appl.*, vol. 14, no. 2, pp. 1–21, 2018.
[56] Y.-A. Chung, W.-H. Weng, S. Tong, and J. Glass, "Towards unsupervised speech-to-text translation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 7170–7174.
[57] L. Schubotz, A. Özyürek, and J. Holler, "Age-related differences in multimodal recipient design: Younger, but not older adults, adapt speech and co-speech gestures to common ground," *Lang., Cogn. Neurosci.*, vol. 34, no. 2, pp. 254–271, 2019.
[58] J. Cheng, W. Zhou, X. Lei, N. Adamo, and B. Benes, "The effects of body gestures and gender on viewer's perception of animated pedagogical agent's emotions," in *Proc. Int. Conf. Hum.-Comput. Interact.*, 2020, pp. 169–186.
[59] I. Molnar-Szakacs, A. D. Wu, F. J. Robles, and M. Iacoboni, "Do you see what i mean? Corticospinal excitability during observation of culture-specific gestures," *PLoS One*, vol. 2, no. 7, 2007, Art. no. e626.