MAGIC: A Fundamental Framework for Gesture Representation, Comparison and Assessment

Edgar Rojas-Muñoz and Juan P. Wachs
School of Industrial Engineering, Purdue University, USA

Abstract— **Gestures play a fundamental role in instructional processes between agents. However, effectively transferring this non-verbal information becomes complex when the agents are not physically co-located. Recently, remote collaboration systems that transfer gestural information have been developed. Nonetheless, these systems relegate gestures to an illustrative role: only a representation of the gesture is transmitted. We argue that further comparisons between the gestures can provide information of how well the tasks are being understood and performed. While gesture comparison frameworks exist, they only rely on gesture's appearance, leaving semantics and pragmatical aspects aside. This work introduces the Multi-Agent Gestural Instructions Comparer (MAGIC), an architecture that represents and compares gestures at the morphological, semantical and pragmatical levels. MAGIC abstracts gestures via a three-stage pipeline based on a taxonomy classification, a dynamic semantics framework and a constituency parsing; and utilizes a comparison scheme based on subtrees intersections to describe gesture similarity. This work shows the feasibility of the framework by assessing MAGIC's gesture matching accuracy against other gesture comparison frameworks during a mentor-mentee remote collaborative physical task scenario.**

I. INTRODUCTION

Gestures play a fundamental role in instructional processes between people: whenever individuals with distinct knowledge backgrounds (e.g. mentors and mentees) transfer and reproduce instructions during face-to-face interactions, a substantial portion of this exchange is done through non-verbal means [1]. However, effectively transferring this non-verbal information has become more complex as the modern workforce became more and more distributed and remote collaborative settings became more common. Remote collaborative systems were proposed as means to make up for such lack of physical co-presence between agents. Consequently, new research focused on determining how to properly represent and transfer such gestures using these remote collaborative systems. In spite that several studies show that gestures reveal meaningful information about learning, memory and other cognitive processes [1]–[3], most collaborative systems relegate gestures to an illustrational role. This means that although gestures are transmitted between the collaborators, insights about their impact in the cognitive processes occurring between the parties are not obtained. We argue that gesture performance-related metrics (e.g. morphological and semantic similarities) provide useful information when evaluating physical instructional outcomes between mentors and mentees. This information is comparable to that obtained from conventional task-related metrics (e.g. completion time, completion rate, reaction time, number of errors, etc.). Therefore, a remote collaborative framework is proposed to include semantic and morphological aspects when comparing between the agents' gesture. The proposed solution would bridge the gap in currently available gestural comparison frameworks that solely rely in the gesture's morphology, while ignoring the gesture meaning (semantics) or the context in which it was performed (pragmatics).

This paper presents preliminary work done towards the development of the Multi-Agent Gestural Instructions Comparer (MAGIC) framework; an architecture capable of representing and comparing gestures from agents at the morphological, semantical and pragmatical levels. For this work, we focus on a remote physical collaboration scenario because: (1) is an acceptable setup to evaluate understanding from task performance; and (2) the gestures exchanged between mentors and mentees will be physically different and cannot be correlated via morphological comparisons. MAGIC abstracts gestures' morphology, semantics and pragmatics via a three-stage pipeline based on a taxonomy classification, a dynamic semantics framework and a constituency parsing approach. Finally, a comparison scheme based on subtree intersections is applied to these trees to measure gesture similarity.

The contributions of this work include: (1) introducing MAGIC, an architecture to compare the gestures of agents with distinct knowledge bases at the morphological, semantical and pragmatical levels; (2) creating a gestural taxonomy for remote collaboration; (3) extending a dynamic semantics framework to represent morphological gestural information; and (4) defining a constituency parsing approach to represent gestures as tree structures. Thereby, the MAGIC architecture could act as a first step towards assessing task understanding in mentor-mentee scenarios through the analysis of gestures.

The paper is organized as follows: Section II reviews prior work related to gestural remote collaboration systems, and morphological and semantical comparisons between gestures. Section III describes the overall framework. Section IV introduces the experimental setup to acquire gestures from agents performing a physical collaborative task remotely, as well as the metric selected to compare among the trees representing gestures. Section V evaluates and discusses our approach with respect to two other methods used to compare gestures. Section VI concludes the paper and discusses options for future improvement of our work.

This work was partly supported by the Office of the Assistant Secretary of Defense for Health Affairs under Award No. W81XWH-14-1-0042. Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the funders.

978-1-7281-0089-0/19/$31.00

II. BACKGROUND

The importance of gestures in remote collaborative tasks is well-known and has been studied extensively: gestures facilitate knowledge acquisition, offer content redundancy and complement abstract concept representations in dialogue [4], [5]. Additionally, gestures facilitate the creation of mutual knowledge, beliefs and assumptions, which leads to better conversational grounding [6], [7]. This conversational grounding involves sharing mental models among the parties [8], [9], as well as the physical means in which the models become tangible actions (e.g. the gestures) [10], [11]. Some of the approaches that have been explored to promote physical communication between parties include: shared visual spaces between agents via head-mounted cameras [12], [13]; communication of pointing gestures via a laser pointer attached to a robotic platform [14], [15]; and projected representation of gestures onto the agents' workspaces, either in 2D [16], [17] or 3D [18], [19]. Nonetheless, most of such systems assess task performance through metrics such as completion time, reaction time upon receiving an instruction, perceived workload, and task completion percentage. In this work we argue that gesture-related metrics such as morphological and semantical similarities should be considered in addition to the aforementioned ones when assessing how knowledge is imparted in a remote collaborative setting. This idea is consistent with findings presented in the literature exploring the role gestures play in knowledge exchange between agents [12], [16], [20].

Gesture representation and comparison methodologies follow two main trends: morphology-based and semantic-based representations. The morphology-based view is intrinsically linked to the gesture appearance, and thus metrics such as statistical modeling techniques [21], neural networks embeddings [22] or distance metrics [23], among others, are constructed with respect of how the gestural data structure was constructed [24]. On the other hand, the semantic-based view constructs logical abstractions of gestures based on linguistics frameworks [25]–[27]. Nonetheless, comparisons between the gestures are not explored in these frameworks.

Fundamental morphology-based representations are based on raw data, directly dependent on the gesture capturing device. Non-optical sensors capture and represent gestures as an arrangement of joint positions and angles over time [28], [29]. Conversely, vision-based models tend to explore distinct representations such as succession of motion signatures [30], filter-extracted features [31], [32], neural network-generated embeddings [22], [33], among others. More recently, richer gestural representations can be obtained via word embeddings [34], and morpho-semantic descriptors [35]. Our work implements morpho-semantic descriptors to represent the gesture's shape and movement. The problem with the aforementioned appearance-based approaches is that the bodily actions performed by non-co-located agents may differ significantly without necessarily implying lack of conversational grounding, as often shown in the remote collaborative systems literature.

Semantic-based representations are the second view for gesture comparisons. Gianluca Giorgolo introduced the concept of iconic semantics, a framework to extract the gestures' semantics from iconic gestures based on the meaning they co-express when aligned with speech [36], [37]. Additionally, co-speech gesture projection has gained attention lately, which analyzes gesture meaning in terms of whether the gesture accompanies or supplements the spoken information [38], [39]. Finally, Lascarides and Stone expanded their dynamic semantics framework [25] to describe how gestures modify or extend the discourse's context, effectively providing a description of the gesture's semantics and pragmatics [40]. Our architecture builds on top of Lascarides and Stone work, but extends it to provide a quantitative metric to measure semantical similarity.

III. METHODOLOGY

A. Multi-Agent Gestural Instruction Comparer Architecture

Consider a scenario where two agents collaborate remotely to repair a robotic arm. In this scenario, one agent does not have the knowledge to perform the repairs alone but is physically present in the plant, whereas the other agent has the knowledge but is remotely located. Conventionally, the agents would exchange instructions through a remote collaboration system and receive feedback of their performance in the form of task-related metrics (e.g. task completion rate, completion time). MAGIC introduces the concept of gesture-based evaluation, which provides a measurement of task understanding by analyzing the gestures performed by both parties. With MAGIC, the gestures performed by the agents can be represented and compared at the morphological (e.g. trajectories, shapes), semantical (e.g. meaning, timing) and pragmatical (e.g. context, environmental elements) levels.

Following Charles Morris' Theory of Signs (ToS) [41], MAGIC defines the most basic elements in a collaborative task scenario: Actions and Agents. Let an $\boldsymbol{\Phi}$ Agent be a person/robot/avatar involved in the collaborative task process. In this work, we define $\boldsymbol{\Phi_W}$ as the **Worker** - the agent that directly manipulates the environment, and $\boldsymbol{\Phi_H}$ as the **Helper** - the agent who guides the **Worker** throughout the task. MAGIC's $\boldsymbol{\Phi}$ Agents are inspired by ToS Interpreters, as both create interpretations from information. Furthermore, let an Action $\boldsymbol{A}$ be the verbal and physical processes by which the agents communicate between each other. This work treats the **Helper**-authored actions as $\boldsymbol{A_H}$ Instructions (e.g. instructing to grab a piece of the robotic arm), and the **Worker**-authored actions as $\boldsymbol{A_W}$ Executions (e.g. actually grabbing the piece).

Each $\boldsymbol{A}$ Action is a three-element tuple. Let the first element of an $\boldsymbol{A}$ Action tuple be a $\boldsymbol{\pi}$ Utterance, either verbal (e.g. "Grab the piece by its corner") or gestural (e.g. gesturing the shape of the piece). Utterances are defined as the smallest unit of speech or gesture that communicates a complete idea. MAGIC's $\boldsymbol{\pi}$ Utterances are inspired by ToS Sign Vehicles, as both are the mediums to exchange information. In addition, let the $\boldsymbol{\mathcal{D}}$ Discourse be a set containing all the utterances, such that $\forall \boldsymbol{\pi}, \boldsymbol{\pi} \in \boldsymbol{\mathcal{D}}$.

Let the second element of an $\boldsymbol{A}$ Action tuple be an $\boldsymbol{\Psi}$ Interpretation, the expected reaction to a certain $\boldsymbol{\pi}$ Utterance. $\boldsymbol{\Psi}$ Interpretations are related to ToS Interpretants, as both represent the agents' disposition to react in a certain way after receiving a stimulus. MAGIC abstracts and represents these expected reactions as a tree data structure (hereafter, $\boldsymbol{\Psi}$ Interpretation Trees). Examples of these trees will by provided in the following sections. Finally, let the third element of an $\boldsymbol{A}$ Action tuple be a $\boldsymbol{\Omega}$ Context, the conditions that motivated an $\boldsymbol{\Phi}$ Agent to generate a certain $\boldsymbol{\Psi}$ Interpretation Tree: the context contains all the elements that could influence an agent into creating an interpretation from a certain utterance. MAGIC's $\boldsymbol{\Omega}$ Context can be viewed as a subset of the more general ToS Context: both encompass information generated by elements in the surroundings. However, MAGIC's context only encompasses elements referenced in previous Utterances ($\boldsymbol{\pi_{t-1}, \pi_{t-2}, \ldots, \pi_{t-|\mathcal{D}|}}$). For example, if a utterance introduced "gripper" as one of the components of the robotic arm, this concept will become part of the available context of future utterances.

The MAGIC architecture allows to compare gestures by generating and matching the $\boldsymbol{\Psi}$ Interpretation Trees. MAGIC models this transition from a gesture to an interpretation with the $\boldsymbol{R()}$ Reaction Function, an approximation of how a $\boldsymbol{\pi}$ Utterance produces an $\boldsymbol{\Psi}$ Interpretation Tree under a given $\boldsymbol{\Omega}$ Context. The elements of the $\boldsymbol{R()}$ Reaction Function will be further explained in the next subsection. Note that while MAGIC's focus is to compare gestures, verbal utterances are considered since they belong to the gestures' context. Fig. 1 presents a schematic of MAGIC's approach and definitions.

B. The Reaction Function

The $\boldsymbol{R()}$ Reaction Function is represented with a three-stage pipeline that receives a $\boldsymbol{\pi}$ Utterance and a $\boldsymbol{\Omega}$ Context as inputs and outputs an $\boldsymbol{\Psi}$ Interpretation Tree. Comparing between the generated interpretations provides a measure of similarity between the gestures.

1) Gestural Taxonomy Classification: The first stage of the $\boldsymbol{R()}$ Reaction Function involves the use a gestural taxonomy to obtain a $\boldsymbol{\eta}$ Classification label for each $\boldsymbol{\pi}$ Utterance. By combining elements from taxonomies presented by McNeil, Goodwin and Poggi [2], [42], [43], MAGIC's gestural taxonomy leverages a tree configuration to assign a $\boldsymbol{\eta}$ Classification to each gesture. Labels close to the taxonomy tree's root contain coarse information regarding the gesture's symbolical expressiveness, whereas labels close to the taxonomy tree's leaves represent fine-grained information such as iconicity. This $\boldsymbol{\eta}$ Classification can be used to assign a weighted importance to specific sections of the final $\boldsymbol{\Psi}$ Interpretation Tree to be created. Currently, the classification labels are assigned by a member of the research team via video recordings of each gesture. The future work section will discuss efforts to automate the process by obtaining these classification labels from a larger pool of subjects.

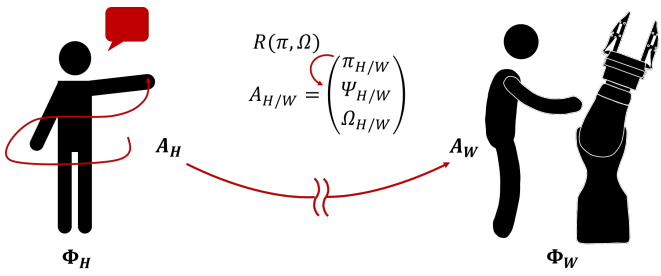

Fig. 1. Multi-Agent Gestural Instructions Comparer scheme. Two agents (**Helper** and **Worker**) collaborate to assemble a robotic arm. The elements of an $\boldsymbol{A}$ Action performed by an $\boldsymbol{\Phi}$ Agent are linked via the $\boldsymbol{R()}$ Reaction Function, a relation describing how a specific $\boldsymbol{\pi}$ Utterance and a given $\boldsymbol{\Omega}$ Context produce a specific $\boldsymbol{\Psi}$ Interpretation Tree.

Fig. 2 presents MAGIC's gestural taxonomy. A proper description of each node label can be found in the referenced literature [2], [42], [43]. MAGIC's framework emphasizes the role of communicative gestures (used when transmitting a message) and manipulative gestures (performed while physically interacting with the environment). Most of the **Helper** gestures will fall into one of the communicative categories, whereas most of the gestures performed by the **Worker** fall into the manipulative category.

2) Extended Segmented Discourse Representation Structure: Segmented Discourse Representation Structure (SDRS) is a formal dynamic semantics framework introduced by Asher and Lascarides that represents semantic and pragmatic information from utterances [25]. SRDS represents the meaning of utterances via SDRS-formulae, logical forms that describe how each utterance updates the discourse's context. Lascarides and Stone extended the SDRS framework to represent gestural utterances via a list of attribute-value pairs that characterize the gesture's physical performance [40]. Unfortunately, this feature structure is represented as a separate table, divided from the rest of the SDRS language. We propose an extension to the SDRS framework (Extended SDRS, henceforth ESDRS) that: (1) represents morphological aspects of gestures as part of $\boldsymbol{\phi}$ ESDRS-formulae; and (2) defines a standard set of components to describe gestures' morphology. Therefore, this stage of the $\boldsymbol{R()}$ Reaction Function pipeline takes a $\boldsymbol{\pi}$ Utterance, a $\boldsymbol{\Omega}$ Context, and the $\boldsymbol{\eta}$ Classification from the previous stage as inputs, and outputs $\boldsymbol{\phi}$ ESDRS-formulae.

ESDRS represents the meaning of an utterance by how its $\boldsymbol{\phi}$ ESDRS-formula transforms an $\boldsymbol{\Omega_i}$ input Context into $\boldsymbol{\Omega_o}$ output Context, under a specific $\boldsymbol{\mathcal{M}}$ model. This $\boldsymbol{\mathcal{M}}$ model contains the distinct elements by which ESDRS expresses an utterance's content, namely Discourse Referents, Spatiotemporal Localities and Virtual Mappings. Discourse referents come in two types: individual variables and eventuality variables. An $\boldsymbol{i}$ individual variable represents elements of the discourse (e.g. a gripper, peace). An $\boldsymbol{e}$ eventuality variable is a temporal event in the discourse (e.g. connecting pieces, tightening screws). A $\boldsymbol{p}$ spatiotemporal locality represents the position in time of a specific individual variable. Each spatiotemporal locality is a 4-dimensional vector $\boldsymbol{(x, y, z, t)}$, where $\boldsymbol{x}$, $\boldsymbol{y}$ and $\boldsymbol{z}$ represents the position in space of the individual variable and $\boldsymbol{t}$ represents a moment in time. Finally, a $\boldsymbol{v}$ virtual mapping represents a transformation over

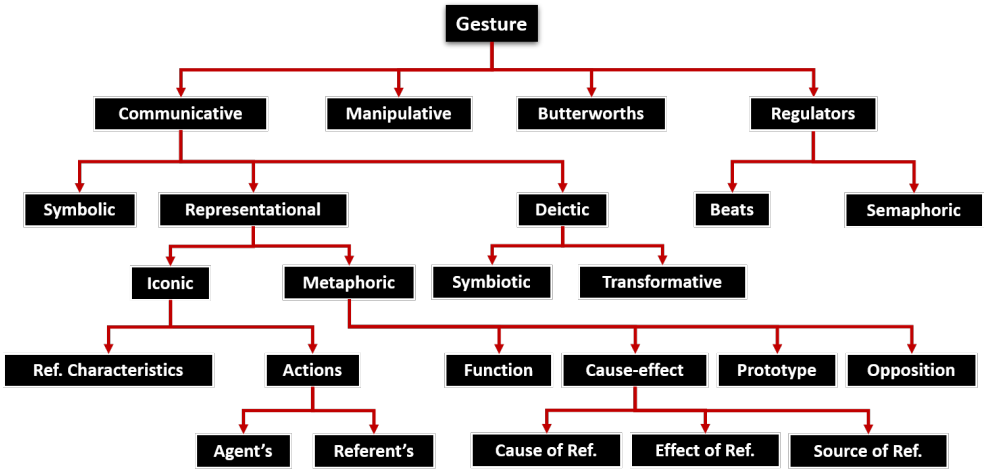

Fig. 2. MAGIC's gestural taxonomy. Classification labels closer to the root provide information about the gestures' symbolical expressiveness, while labels closer to the leaves present fine-grained information such as iconicity.

a spatiotemporal location mapping a point from world space into a point in gesture space. For example, virtual mappings are required when the absolute position of an element in world space (e.g. position of the robotic arm) is represented by a relative position gesture space (e.g position of the hand).

Additionally, ϕ ESDRS-formulae are built via predicates, tests over the $\mathcal{M}$ model elements. These tests provide propositional information of how the elements interact between each other. These elements will move from $\mathbf{\Omega_i}$ into $\mathbf{\Omega_o}$ by satisfying these predicative tests, effectively updating the discourse's context. This process is known as Context Change Potential, and characterizes the meaning of the utterances [25]. For example, the predicate $Connect(i_1, i_2)$ represents that the individual variables i_1 and i_2 are *connected*. All the predicates included in SDRS (e.g. $Loc()$, $Exemplifies()$, $Component()$) are also included in ESDRS.

Consider the following verbal and gestural utterances:

$\boldsymbol{\pi_1}$: "Grab the piece by its corner"

$\boldsymbol{\pi_2}$: *The speaker puts both hands in front of her. The left hand makes a fist shape. The right hand faces the left hand with the index finger extended, pointing at the left hand. Other fingers are not extended. Both hands stay in place.*

The verbal utterance ϕ ESDRS-formula is given by:

$$\pi_1 : \exists\, i_1, i_2 \begin{bmatrix} Piece(i_1) \wedge Corner(i_2) \wedge Grab(e_1, i_2) \wedge \\ Component(i_2, i_1) \wedge Loc(e_1, i_2, v_c(\vec{p_c})) \end{bmatrix},$$

where p and c are individual variables (introduced into the discourse via the $\exists$ operator), e_1 is an eventuality variable, and v_c is a virtual mapping over the spatiotemporal location p_c. This ϕ ESDRS-formula includes the predicates $Piece()$, $Corner()$, $Component()$, $Grab()$ and $Loc()$, which are conditioning the $\mathcal{M}$ model elements and therefore updating the discourse context. For a more detailed explanation of SDRS, refer to Asher and Lascarides work [25].

To generate gestural ϕ ESDRS-formulae, additional elements introduced in ESDRS must be defined. ESDRS introduces the $TaxClass()$ predicative group, which contains predicates related to the gestures' taxonomy classification. The $\boldsymbol{\eta}$ Classification (and the parent nodes) will be translated into ESDRS-formulae predicates. Additionally, ESDRS translates the feature table from [40] into two predicative groups: $Shape()$ and $Movement()$. The $Shape()$ group introduces individual variables describing the fine-grained components of a gesture's morphology (i.e. arms, hands, fingers), as well as predicates referring to their relative pose, orientation and separation. The $Movement()$ group treats each zero-velocity point in a motion trajectory (points in 3D space where $\frac{\partial x}{\partial t} = \frac{\partial y}{\partial t} = \frac{\partial z}{\partial t} = 0$) as spatiotemporal locations, and each trajectory (hand motions between two zero-velocity points) as individual variables. In addition, the $Movement()$ group introduces predicates to describe the gesture's main plane of motion and the motion trajectories' direction. The predicates in both these groups are inspired from the morpho-semantic descriptors in [35]. Finally, ESDRS introduces the $Synchro()$ predicate, describing whether the gesture was performed in synchrony with a specific event in the discourse (i.e. an eventuality variable). For example, if e_1 represents the event of grabbing a piece, $Synchro(e_1)$ represents that the gesture was performed during e_1.

Consequently, the gestural utterance ϕ ESDRS-formula (with π_{2T}, π_{2S} and π_{2M} as elements of π_2) is given by:

$$\pi_2 : [\mathcal{G}] \exists\, i_3, i_4, i_5 \begin{bmatrix} Gesture(i_3) \wedge TaxClass(i_3) \wedge \\ Shape(i_3) \wedge Movement(i_3) \wedge \\ Synchro(e_1) \wedge Exemplifies(i_3, i_1) \end{bmatrix}$$

$$\pi_{2T} : [\mathcal{G}] \left[Communicative(i_3) \wedge Deictic(i_3) \right]$$

$$\pi_{2M} : [\mathcal{G}] \exists\, i_{M_1}, i_{M_2} \begin{bmatrix} Trajectory(i_{M_1}, v_I(\vec{p_1}), v_I(\vec{p_2})) \wedge \\ Trajectory(i_{M_2}, v_I(\vec{p_3}), v_I(\vec{p_4})) \wedge \\ MainPlaneCoronal(i_{M_1}) \wedge \\ DirectionStatic(i_{M_1}) \wedge \\ MainPlaneCoronal(i_{M_2}) \wedge \\ DirectionStatic(i_{M_2}) \wedge \\ Component(i_{M_1}, i_3) \wedge \\ Component(i_{M_2}, i_3) \end{bmatrix}$$

$$\pi_{2S} : [\mathcal{G}]\exists i_{S_1}, i_{S_2}, i_{S_3}, i_{S_4}, i_{S_5}, i_{S_6}, i_{S_7}, i_{S_8}, i_{S_9}, i_{S_{10}}, i_{S_{11}}, i_{S_{12}}, i_{S_{13}}, i_{S_{14}} \left[\begin{array}{c} Arm(i_{S_1}) \wedge Arm(i_{S_2}) \wedge Hand(i_{S_3}) \wedge Hand(i_{S_4}) \wedge \\ ThumbFinger(i_{S_5}) \wedge RingFinger(i_{S_6}) \wedge MiddleFinger(i_{S_7}) \wedge \\ IndexFinger(i_{S_8}) \wedge LittleFinger(i_{S_9}) \wedge ThumbFinger(i_{S_{10}}) \wedge \\ RingFinger(i_{S_{11}}) \wedge MiddleFinger(i_{S_{12}}) \wedge IndexFinger(i_{S_{13}}) \wedge \\ LittleFinger(i_{S_{14}}) \wedge PoseSemiExtended(i_{S_1}) \wedge OrientationLeft(i_{S_1}) \wedge \\ PoseNotExtended(i_{S_3}) \wedge OrientationForward(i_{S_3}) \wedge PoseNotExtended(i_{S_5}) \wedge \\ PoseNotExtended(i_{S_6}) \wedge PoseNotExtended(i_{S_7}) \wedge PoseNotExtended(i_{S_8}) \wedge \\ PoseNotExtended(i_{S_9}) \wedge PoseSemiExtended(i_{S_2}) \wedge OrientationLeft(i_{S_2}) \wedge \\ PoseNotExtended(i_{S_4}) \wedge OrientationForward(i_{S_4}) \wedge PoseNotExtended(i_{S_{10}}) \wedge \\ PoseExtended(i_{S_{11}}) \wedge PoseNotExtended(i_{S_{12}}) \wedge PoseNotExtended(i_{S_{13}}) \wedge \\ PoseNotExtended(i_{S_{14}}) \wedge OrientationForward(i_{S_{11}}) \wedge Separated(i_{S_{10}}, i_{S_{11}}) \wedge \\ Separated(i_{S_{11}}, i_{S_{12}}) \wedge Component(i_{S_1}, i_3) \wedge Component(i_{S_2}, i_3) \wedge \\ Component(i_{S_3}, i_{S_1}) \wedge Component(i_{S_5}, i_{S_3}) \wedge Component(i_{S_6}, i_{S_3}) \wedge \\ Component(i_{S_7}, i_{S_3}) \wedge Component(i_{S_8}, i_{S_3}) \wedge Component(i_{S_9}, i_{S_3}) \wedge \\ Component(i_{S_4}, i_{S_2}) \wedge Component(i_{S_{10}}, i_{S_4}) \wedge Component(i_{S_{11}}, i_{S_4}) \wedge \\ Component(i_{S_{12}}, i_{S_4}) \wedge Component(i_{S_{13}}, i_{S_4}) \wedge Component(i_{S_{14}}, i_{S_4}) \end{array} \right]$$

3) ESDRS Constituency Parsing: ϕ ESDRS-formulae are not designed for comparisons: an additional data structure to abstract them needs to be generated to assess gesture similarity. MAGIC leverages tree structures to do this as they can capture both the value of the ϕ ESDRS-formulae elements and the relationship between them. Therefore, the third stage of the $\boldsymbol{R}()$ Reaction Function pipeline applies a constituency parsing to the ϕ ESDRS-formulae to generate $\boldsymbol{\Psi}$ Interpretation Trees. Fig. 3 presents an $\boldsymbol{\Psi}$ Interpretation Tree, with its root colored in red. The leaves of the tree (green-colored) depict the values of the ESDRS elements. The nested constituents of the ESDRS-formulae are represented in blue. MAGIC's parsing introduces different nested constituents to represent the ESDRS-formulae's structure. Following the tree's hierarchy, each root node can include five main constituents (the black-circled nodes in Fig. 3):

1) Variable Group (VG): contains discourse referents.
2) Spatiotemporal Group (SG): contains spatiotemporal localities.
3) Mapping Group (MG): contains virtual mappings.
4) Context Group (CG): contains the discourse referents and predicates (introduced in the previous utterances) that are referred to in the current utterance.
5) Large Predicate Group (LPG): contains the predicates.

Moreover, the LPG is divided into seven constituents (denoted with the colored bounding areas in Fig. 3), each of them having their respective nested constituents:

1) Shape Group (ShG): contains the discourse referents and predicates related to the gesture's shape.
2) Loc Group (LoG): contains all the $Loc()$ predicates, used whenever individual variables are spatially contained in a spatiotemporal locality at each moment in time spanned by an eventuality variable.
3) Exemplifies Group (ExG): contains all the $Exemplifies()$ predicates, representing whenever a gesture is used to depict a specific individual variable.
4) TaxClass Group (TaG): contains the predicates related to the gesture's taxonomy classification.
5) Synchro Group (SyG): contains the Synchro predicate.
6) Extra Predicates: contains every predicate that is not contained in any of the other groups.
7) Movement Group (MvG): contains the discourse referents and predicates related to the gesture's movement.

This parsing approach encapsulates all the information represented by the ϕ ESDRS-formulae into a tree structure that takes morphological, semantical and pragmatical information into account. Measuring the similarity between the generated $\boldsymbol{\Psi}$ Interpretation Trees will describe how similar are the gestures from the agents collaborating. The next section explains how MAGIC's trees can be compared between them and against other baseline gesture data structures.

IV. Experimental Apparatus

A. Data Collection

Two participants were recruited to collaboratively complete a block assembly task (comparable to the tasks presented in previous works [16], [20]). Participants were randomly assigned to a role (e.g. **Worker** or **Helper**) and were situated in different rooms. The **Worker** was provided with blocks to assemble a model, whereas the **Helper** was provided with instructions detailing how to assemble it. Participants were able to see and communicate with each other via a Skype call. No restrictions were imposed on the type of information participants were able to exchange (verbal commands, gestures, facial expressions, etc.). Participants were recorded with a Microsoft Kinect and a RGB camera. A total of 102 gestures were manually segmented from these recordings, 13 performed by the **Helper** and 89 performed by the **Worker**. **Worker** gestures were divided into 3 groups: responses to verbal utterances, responses to gestural utterances, and those not performed as a response. Because the aim of this work, only the group consisting of responses to gestural utterances was consider for the analysis. This reduced the dataset to 13 **Helper** gestures and 36 **Worker** gestures, over a span of 14 minutes of video (a ratio of approximately 3 **Worker** gestures for each **Helper** gesture). The obtained gestures were compared using the approaches described in the next subsection.

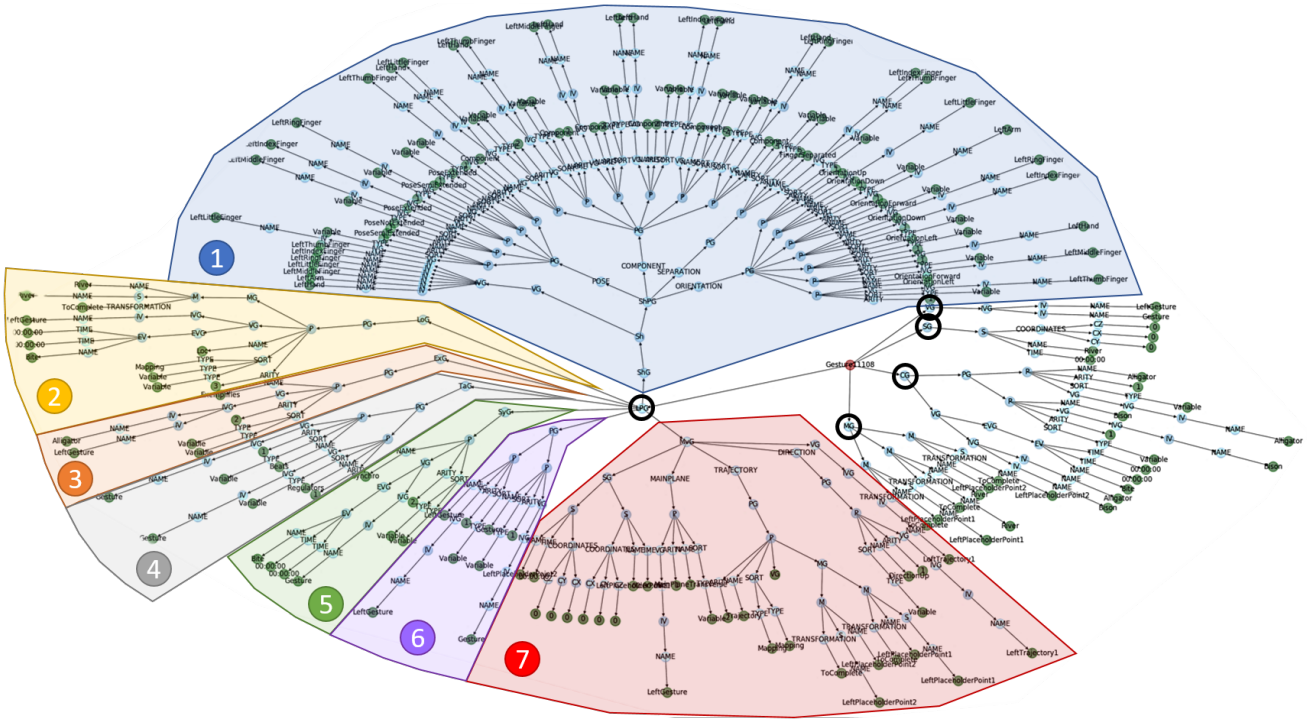

Fig. 3. Representation of one gesture using MAGIC's Interpretation Tree structure. The black-circled nodes indicate the five main constituents of an Interpretation Tree. The numbered regions indicate the main nested constituents of the tree: Shape Group (ShG, 1), Loc Group (LoG, 2), Exemplifies Group (ExG, 3), TaxClass Group (TaG, 4), Synchro Group (SyG, 5), Extra Predicates (6), and Movement Group (MvG, 7).

B. Gesture Comparison

The intersection between subtrees is used to compare between Ψ Interpretation Trees. An important advantage of Ψ Interpretation Trees is that they store context (in the form of the CG Context Group). Our similarity approach consists in obtaining the **Worker**'s context subtree and comparing it against other subtrees in the Ψ_H **Helper** Interpretation Trees. The intuition behind this approach is that the context of each Ψ_W **Worker** Interpretation Tree can be tracked back to elements in a matching Ψ_H **Helper** Interpretation Tree. Therefore, let $\mathcal{X}$ be a nested constituent from an Ψ Interpretation Tree (e.g. CG, LPG, SyG). Then, $\Psi^{\mathcal{X}}$ is the subtree of Ψ that has $\mathcal{X}$ as its root (e.g. Ψ^{CG} represents a context subtree). These subtrees can also be combined, in the form $\Psi^{\mathcal{X}_1} \cup \Psi^{\mathcal{X}_2}$. By following this comparison approach, the Ψ_{W_j} **Worker** Interpretation Tree that matches the Ψ_{H_i} **Helper** Interpretation Tree will satisfy:

$$max\left(num\_nodes\, \Psi^{\mathcal{X}}_{H_i} \cap \Psi^{CG}_{W_j}\right); \begin{matrix} 1 \leq i \leq I \\ 1 \leq j \leq J \end{matrix},$$

where I is the total number of **Helper** gestures and J is the total number of **Worker** gestures. Fig. 4 presents a visual example of the intersection between two sets of Ψ Interpretation Trees leaves.

MAGIC's gesture matching performance was evaluated against two baseline metrics: morpho-semantic descriptors (MSD) vectors [35], and a temporal synchronization (TS) approach. The MSD vectors were compared between each other via Hamming distance and cosine similarity. MSD vectors were included as a gesture comparison baseline based on physical similarity. To prevent bias, these vectors were manually annotated by a member extraneous to the research team. The TS approach represented each gesture as a normalized timestamp in seconds (0 and 1 being the start and end of the video, respectively). The approach compared gestures based on their temporal occurrence. A **Worker** gesture performed right after a **Helper** gesture is likely to be associated with the same concept, and thus representing "similar" meaning. In other words, their $\boldsymbol{A}$ Actions are synchronized (as one tends to be the reaction to the other one). For each **Helper** gesture, a time window before and after its execution was created. Every **Worker** gesture inside this time window was associated to the given **Helper** gesture.

During the data collection phase, a team member paired each **Worker** gesture with its corresponding **Helper** gesture. These matches provided a ground truth of correspondences for each gesture. Afterwards, the three gesture comparison approaches were evaluated in terms of their matching accuracy. That is, for each data structure representing a **Worker** gesture, find the **Helper** gesture with the highest similarity score. The matching accuracy is given by the ratio between the amount positive gesture matchings (with respect to the ground truth) and the total number of gesture comparisons.

V. RESULTS AND DISCUSSION

Fig. 5 summarizes the approaches' matching accuracies. The results demonstrate that MAGIC's Ψ Interpretation Trees can match gestures from different agents with a higher accuracy than other gesture representation approaches.

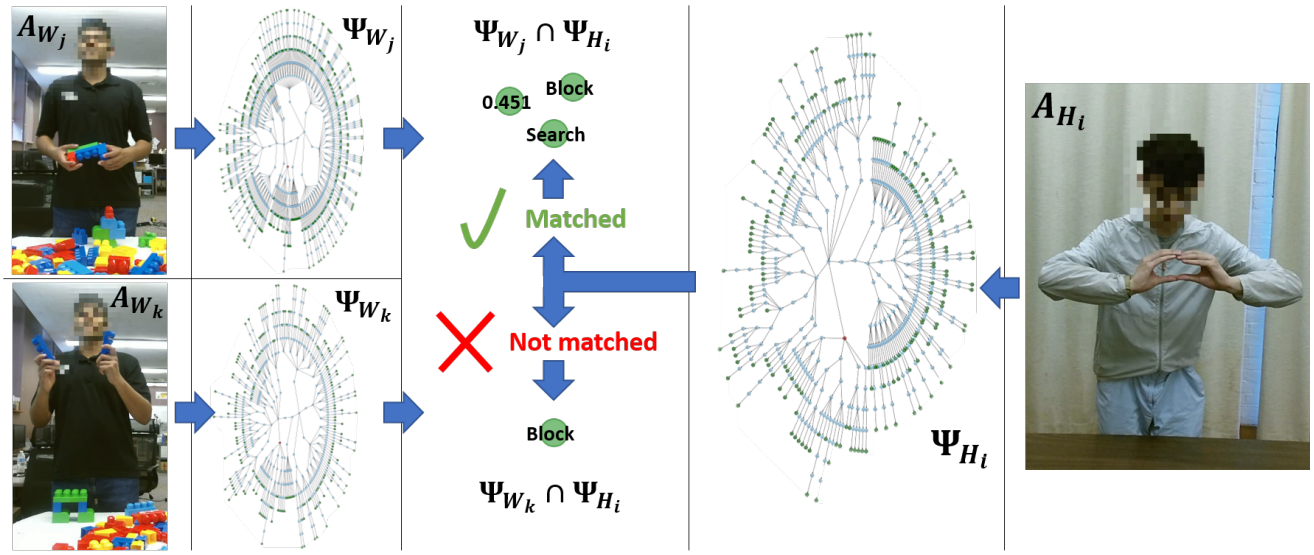

Fig. 4. Subtree intersection similarity applied on two different sets of Interpretation Trees. A **Helper** performs a gesture and an Ψ Interpretation Tree is generated from it. Similarly, the Ψ Interpretation Tree of each **Worker**-authored gesture are generated. A notion of the similarity between the gestures can be obtained by intersecting the subtress of these Ψ Interpretation Trees. Subtrees that are similar will have a higher number of common nodes.

Selecting which subtree to compare against is key to obtain proper matching accuracies, as information unrelated to the given comparison can be introduced when the wrong subtree is selected (e.g. comparing shape against meaning). Based on our experiments, the subtree that presented the highest matching accuracy is the union between the ExG, LoG and SyG subtrees. This is because these three subgroups contain most of the semantical information of the Ψ_H **Helper** Interpretation Trees. In other words, because the **Worker** gesture was generated in response to a **Helper** gesture, the information encompassed in the Ψ_W^{CG} **Worker** context subtree had similar information to the one in its corresponding **Helper** meaning subtree ($\Psi_H^{ExG} \cup \Psi_H^{LoG} \cup \Psi_H^{SyG}$). Comparisons against the Ψ_H^{CG} and Ψ_H^{LPG} **Worker** subtrees and against the entire Ψ_H **Worker** Interpretation Tree were also included to demonstrate how the selection of a wrong subtree can lead to poor matching accuracies. The information contained in the Shape and Movement Groups was considered not as relevant for the comparison shown in this work, which was mostly based on the relation between meaning and context. Nonetheless, these subgroups are still relevant for the structure, as most of the current gesture recognition algorithms perform their classification based on the physical aspects represented in these subtrees.

MAGIC's high matching accuracy results can be traced back to SDRS, as the framework was designed to represent meaning and context with relative ease. Nonetheless, both MAGIC and SDRS point at an important fact: properly obtaining the meaning and context of a gesture is key to matching success. For this work, the meaning and context elements were manually annotated by members of the research team. How to automatically segment and assign meaning and context of a gesture is an open research problem that will be explored in future versions of the MAGIC's architecture.

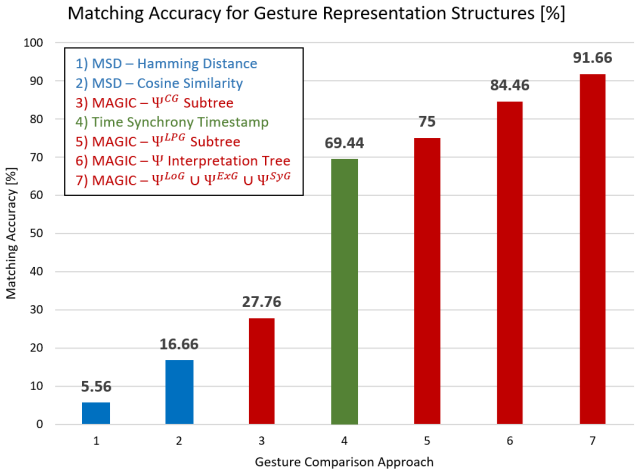

Fig. 5. Matching accuracies for different gesture representation structures. Results can be interpreted as how well were the ground truth matches replicated by the gesture matching approaches. Most of MAGIC's subtree intersection similarities provided better-than-baseline matching accuracies.

VI. CONCLUSIONS

This work introduced the MAGIC framework, an architecture to represent and compare gestures at the morphological, semantical and pragmatical levels. By leveraging a gestural taxonomy, a dynamic semantics framework and a constituency parsing, MAGIC creates a generalizable abstraction of gestures through a tree data structure. The gestures of two agents performing a collaborative task were captured and represented with two gesture representation baselines and with MAGIC's gesture representation structure. After obtaining a human-annotated ground truth describing how the gestures of these agents were matched, matching accuracies for the different gesture representation structures were calculated.

MAGIC's gesture matching accuracies were higher than those obtained with the baseline structures. A limitation of the presented experiment has to do with the manual annotation of semantical and pragmatical information. Future work will investigate how MAGIC's gesture matching capabilities can be used to assess task understanding and performance in mentor-mentee scenarios through the analysis of gestures.

VII. ACKNOWLEDGMENTS

This work was partly supported by the Office of the Assistant Secretary of Defense for Health Affairs under Award No. W81XWH-14-1-0042. Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the funders. We also thank Mr. Madapana for his assistance regarding the MSD vectors.

REFERENCES

[1] M. Argyle, *Bodily communication*. Routledge, 2013.

[2] D. McNeill, *Hand and mind: What gestures reveal about thought*. University of Chicago press, 1992.

[3] L. Radford, "Why do gestures matter? sensuous cognition and the palpability of mathematical meanings," *Educational Studies in Mathematics*, vol. 70, no. 2, pp. 111–126, 2009.

[4] A. Kendon, *Gesture: Visible action as utterance*. Cambridge University Press, 2004.

[5] D. McNeill, "So you think gestures are nonverbal?," *Psychological review*, vol. 92, no. 3, p. 350, 1985.

[6] S. E. Brennan, "The grounding problem in conversations with and through computers," *Social and cognitive approaches to interpersonal communication*, pp. 201–225, 1998.

[7] H. H. Clark, S. E. Brennan, *et al.*, "Grounding in communication," *Perspectives on socially shared cognition*, vol. 13, no. 1991, pp. 127–149, 1991.

[8] J. E. Mathieu, T. S. Heffner, G. F. Goodwin, E. Salas, and J. A. Cannon-Bowers, "The influence of shared mental models on team process and performance.," *Journal of applied psychology*, vol. 85, no. 2, p. 273, 2000.

[9] S. Converse, J. Cannon-Bowers, and E. Salas, "Shared mental models in expert team decision making," *Individual and group decision making: Current issues*, vol. 221, 1993.

[10] H. H. Clark, "Using language. 1996," *Cambridge University Press: Cambridge*, vol. 952, pp. 274–296, 1996.

[11] H. H. Clark and C. R. Marshall, "Definite reference and mutual knowledge," *Psycholinguistics: critical concepts in psychology*, vol. 414, 2002.

[12] R. E. Kraut, S. R. Fussell, and J. Siegel, "Visual information as a conversational resource in collaborative physical tasks," *Human–Computer Interaction*, vol. 18, no. 1-2, pp. 13–49, 2003.

[13] R. E. Kraut, D. Gergle, and S. R. Fussell, "The use of visual information in shared visual spaces: Informing the development of virtual co-presence," in *Proceedings of the 2002 ACM conference on Computer supported cooperative work*, pp. 31–40, ACM, 2002.

[14] P. Luff, C. Heath, H. Kuzuoka, J. Hindmarsh, K. Yamazaki, and S. Oyama, "Fractured ecologies: creating environments for collaboration," *Human-Computer Interaction*, vol. 18, no. 1, pp. 51–84, 2003.

[15] H. Kuzuoka, J. Kosaka, K. Yamazaki, Y. Suga, A. Yamazaki, P. Luff, and C. Heath, "Mediating dual ecologies," in *Proceedings of the 2004 ACM conference on Computer supported cooperative work*, pp. 477–486, ACM, 2004.

[16] D. Kirk, T. Rodden, and D. S. Fraser, "Turn it this way: grounding collaborative action with remote gestures," in *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pp. 1039–1048, ACM, 2007.

[17] N. Yamashita, K. Kaji, H. Kuzuoka, and K. Hirata, "Improving visibility of remote gestures in distributed tabletop collaboration," in *Proceedings of the ACM 2011 conference on Computer supported cooperative work*, pp. 95–104, ACM, 2011.

[18] W. Huang and L. Alem, "Handsinair: a wearable system for remote collaboration on physical tasks," in *Proceedings of the 2013 conference on Computer supported cooperative work companion*, pp. 153–156, ACM, 2013.

[19] R. S. Sodhi, B. R. Jones, D. Forsyth, B. P. Bailey, and G. Maciocci, "Bethere: 3d mobile collaboration with spatial input," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 179–188, ACM, 2013.

[20] 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," *Human-Computer Interaction*, vol. 19, no. 3, pp. 273–309, 2004.

[21] G. Caridakis, K. Karpouzis, A. Drosopoulos, and S. Kollias, "Somm: Self organizing markov map for gesture recognition," *Pattern Recognition Letters*, vol. 31, no. 1, pp. 52–59, 2010.

[22] S. S. Ge, Y. Yang, and T. H. Lee, "Hand gesture recognition and tracking based on distributed locally linear embedding," *Image and Vision Computing*, vol. 26, no. 12, pp. 1607–1620, 2008.

[23] X. Zhao, T. Feng, and W. Shi, "Continuous mobile authentication using a novel graphic touch gesture feature," in *Biometrics: Theory, Applications and Systems (BTAS), 2013 IEEE Sixth International Conference on*, pp. 1–6, IEEE, 2013.

[24] S. Mitra and T. Acharya, "Gesture recognition: A survey," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 37, no. 3, pp. 311–324, 2007.

[25] N. Asher and A. Lascarides, *Logics of conversation*. Cambridge University Press, 2003.

[26] R. Montague, "Pragmatics and intensional logic," *Dialectica*, vol. 24, no. 4, pp. 277–302, 1970.

[27] C. Potts, *The logic of conventional implicatures*. No. 7, Oxford University Press on Demand, 2005.

[28] F. Parvini, D. McLeod, C. Shahabi, B. Navai, B. Zali, and S. Ghandeharizadeh, "An approach to glove-based gesture recognition," in *International Conference on Human-Computer Interaction*, pp. 236–245, Springer, 2009.

[29] L. Dipietro, A. M. Sabatini, P. Dario, *et al.*, "A survey of glove-based systems and their applications," *IEEE Trans. Systems, Man, and Cybernetics, Part C*, vol. 38, no. 4, pp. 461–482, 2008.

[30] C.-C. Chiu and S. Marsella, "Gesture generation with low-dimensional embeddings," in *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pp. 781–788, International Foundation for Autonomous Agents and Multiagent Systems, 2014.

[31] S. S. Rautaray and A. Agrawal, "Vision based hand gesture recognition for human computer interaction: a survey," *Artificial Intelligence Review*, vol. 43, no. 1, pp. 1–54, 2015.

[32] J. Konečnỳ and M. Hagara, "One-shot-learning gesture recognition using hog-hof features," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 2513–2532, 2014.

[33] E. Stergiopoulou and N. Papamarkos, "Hand gesture recognition using a neural network shape fitting technique," *Engineering Applications of Artificial Intelligence*, vol. 22, no. 8, pp. 1141–1158, 2009.

[34] W. Thomason and R. A. Knepper, "Recognizing unfamiliar gestures for human-robot interaction through zero-shot learning," in *International Symposium on Experimental Robotics*, pp. 841–852, Springer, 2016.

[35] N. Madapana and J. Wachs, "Zsgl: zero shot gestural learning," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pp. 331–335, ACM, 2017.

[36] G. Giorgolo, "A formal semantics for iconic spatial gestures," in *Logic, Language and Meaning*, pp. 305–314, Springer, 2010.

[37] G. Giorgolo, "Integration of gesture and verbal language: a formal semantics approach," in *International Gesture Workshop*, pp. 216–227, Springer, 2011.

[38] C. Ebert and C. Ebert, "Gestures, demonstratives, and the attributive/referential distinction," *Handout of a talk given at Semantics and Philosophy in Europe (SPE 7)*, 2014.

[39] P. Schlenker, "Gesture projection and cosuppositions," *Linguistics and Philosophy*, vol. 41, no. 3, pp. 295–365, 2018.

[40] A. Lascarides and M. Stone, "A formal semantic analysis of gesture," *Journal of Semantics*, vol. 26, no. 4, pp. 393–449, 2009.

[41] C. W. Morris, "Foundations of the theory of signs," in *International encyclopedia of unified science*, pp. 1–59, Chicago University Press, 1938.

[42] C. Goodwin, "The body in action," in *Discourse, the body, and identity*, pp. 19–42, Springer, 2003.

[43] I. Poggi, "Iconicity in different types of gestures," *Gesture*, vol. 8, no. 1, pp. 45–61, 2008.