ASSESSING COLLABORATIVE PHYSICAL TASKS VIA GESTURAL ANALYSIS USING THE "MAGIC" ARCHITECTURE

by

Edgar Javier Rojas Muñoz

A Dissertation

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

School of Industrial Engineering

West Lafayette, Indiana

August 2020

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Juan P. Wachs, Chair
School of Industrial Engineering

Dr. Myrdene Anderson
Department of Anthropology

Dr. Dan Goldwasser
Department of Computer Science

Dr. Mario Ventresca
School of Industrial Engineering

Approved by:
Dr. Abhijit Deshmukh

This dissertation is dedicated to my two families: Rojas-Muñoz and Borror.

ACKNOWLEDGMENTS

Foremost, I would like to express my heartfelt gratitude to my advisor, Prof. Juan P. Wachs, for his continuous support of my graduate study and research. His encouragement, patience, motivation, and immense knowledge helped me in all the time of research and writing of this dissertation. Besides my advisor, I would like to thank the rest of my thesis committee: Prof. Myrdene Anderson, Prof. Dan Goldwasser, and Prof. Mario Ventresca for providing many valuable insights and comments that improved the content of this dissertation.

I also appreciate the support and help from my friends and lab mates: Daniel Andersen, Juan Antonio Barragan, Maria Eugenia Cabrera, Daniela Chanci, Margie Cubillo, Mariela Fernandez, Glebys Gonzalez, Shyam Sundar Kannan, Chengyuan Lin, , Melissa Paniagua, Dr. Voicu Popescu, Oscar Rincon, Natalia Sánchez, Teerachart (Taa) Soratana, Dr. Thomas Sors, Uttara Tipnis, Anirudh Tunga, Hector Ulate, Daniel Vasquez, Joselyn Walsh, Chenxi Xiao, Ting Zhang, Xingguang Zhang, Tian Zhou, and René Zúñiga, Special thanks to Marisol De La Vega Solano and Naveen Madapana for being such a crucial support during this journey.

This dissertation is 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.

TABLE OF CONTENTS

LIST OF TABLES ... 7

LIST OF FIGURES ... 8

ABSTRACT ... 10

1. INTRODUCTION ... 11

1.1 Significance ... 12

1.2 Definitions ... 13

1.3 Research Problem ... 13

1.3.1 Research Question 1 (RQ1) ... 13

1.3.2 Research Question 2 (RQ2) ... 14

1.3.3 Research Question 3 (RQ3) ... 14

1.4 Overview of the Document’s Structure ... 14

2. LITERATURE REVIEW ... 15

2.1 Importance and Challenges of Gestures in Human Interactions ... 15

2.2 Gestures in Remote Collaborative Tasks ... 21

2.2.1 Remote collaborative systems ... 22

2.3 Gesture Representation, Comparison and Assessment ... 24

2.3.1 Morphology-based gesture representations ... 25

2.3.2 Semantics-based gesture representations ... 27

2.4 Approaches to Estimate Task Understanding ... 28

2.4.1 Subjective approaches to estimate understanding ... 28

2.4.2 Objective approaches to estimate understanding ... 29

2.5 Summary ... 30

3. METHODOLOGY ... 32

3.1 Multi-Agent Gestural Instructions Comparer ... 33

3.2 The Reaction Function ... 36

3.2.1 Gestural Taxonomy Classification ... 37

3.2.2 Extended Segmented Discourse Representation Structure ... 38

3.2.3 ESDRS Constituency Parsing ... 42

3.3 Gesture Matching through Integer Optimization Problems ... 47

3.3.1 MAGIC-based Optimization 48

3.3.2 Time-based Optimization 54

3.3.3 Hybrid Optimization 56

3.4 Physical Information Assimilation Metric 60

3.5 Summary 62

4. EXPERIMENTS AND RESULTS 63

4.1 Data Collection 63

4.2 Ground Truth Annotation 67

4.3 Evaluating Gesture Comparison Approaches 68

4.3.1 Gesture Comparison Results 70

4.4 Evaluating Task Understanding Estimation Approaches 77

4.4.1 Task Understanding Estimation Results 77

4.5 Evaluating Perceived Workload 83

4.6 Summary 88

5. CONCLUSIONS AND FUTURE WORK 89

5.1 Applications of Task Understanding Evaluation via Gestures 91

5.2 Applications beyond Human-to-Human Collaboration 91

5.3 Applications in Artificial Intelligence 92

REFERENCES 93

VITA 113

PUBLICATIONS 114

LIST OF TABLES

Table 4.1 Morpho-Semantic Descriptors. 68

Table 4.2 Comparison of Task Understanding Proxy Metrics on the Block Assembly Task. Results indicate percentages [%]. 80

Table 4.3 Comparison of Task Understanding Proxy Metrics on the Origami Task. Results indicate percentages [%]... 80

Table 4.4 Comparison of Task Understanding Proxy Metrics on the Ultrasound Training Task. Results indicate percentages [%]. 80

LIST OF FIGURES

Figure 2.1 Gestures aiding to conversational grounding. Through gestures, individuals can indicate a) quantities, b) shapes, c) distances, and d) locations.. 19

Figure 2.2 Examples of remote gesture collaboration systems. The image includes remote collaboration systems based on: a) projectors (D. Kirk et al., 2007), b) head-mounted cameras (Kraut et al., 2003), c) laser pointers (Kuzuoka et al., 2004), and d) AR (Sodhi et al., 2013). 24

Figure 2.3 Examples of morphology-based gesture representation techniques. The image includes gesture representation approaches based on: a) glove-based systems (Dipietro et al., 2008), b) motion signatures (Chiu & Marsella, 2014), c) filter-extracted features (Konečnỳ & Hagara, 2014), d) network-generated embeddings (Stergiopoulou & Papamarkos, 2009). 26

Figure 2.4 Functionally equivalent gestures with different visual appearance. 27

Figure 2.5 Reaction time vs. number of correct responses characteristic curve. Adapted from (Pachella, 1973). 30

Figure 3.1 Platform architecture and interactions between its components. 34

Figure 3.2 Multi-Agent Gestural Instructions Comparer framework. 36

Figure 3.3 Three-stage pipeline of the $\boldsymbol{R}()$ Reaction Function. 37

Figure 3.4 MAGIC's gestural taxonomy. 38

Figure 3.5 Example of MAGIC's Interpretation Tree. The black-circled nodes indicate the five main constituents of an Interpretation Tree. The numbered regions indicate the main nested constituents of tree's predicative group. 46

Figure 3.6 Comparison between the Interpretation Trees generated from the gestures performed by different agents.. 47

Figure 3.7 Example of subtree intersection similarity. Similar subtrees will have a higher number of common nodes. 49

Figure 3.8 Example of subtree intersection similarity. Similar subtrees will have a higher number of common nodes. 51

Figure 3.9 Effects of α and β in time damping effect, for $cij \in$ [-50,500]. The α constant regulates the degree in which time attenuates the final hybrid cost. The β constant regulates when the damping starts. 57

Figure 3.10 Example of how the collaborators' gestures are matched according to the different optimization problems. 59

Figure 3.11 **E** matrices representing the gesture matching solutions showcased in Figure 3.10.. 59

Figure 3.12 Example of a graph representing the **E** matrix from Figure 3.11, obtained using the MAGIC-based optimization approach. The $\boldsymbol{eij}$ edge weights in the image have a value of “1” in the **E** matrix. 61

Figure 4.1 Example of participants collaborating to complete a shared task. *Left*: Steps of the block assembly task are depicted. *Right*: Steps of the block assembly task are depicted. 64

Figure 4.2 Ultrasound training task. The complete ultrasound simulator is shown in a). The ultrasound phantom is shown in b). The vessel lumens as seen in the ultrasound are shown in c). The process of extracting blood is shown in d). The foreign body as seen in the ultrasound is shown in e). 65

Figure 4.3 Worker station setup. 66

Figure 4.4 Helper station setup. 67

Figure 4.5 Temporal synchronization gesture matching approach. 70

Figure 4.6 Number of gestures performed by the Helper-Worker pairs. 72

Figure 4.7 Gesture matching scores for the block assembly, origami, and ultrasound training tasks. The scores represent the percentage of agreement of the gesture matching approaches with the human baseline. 73

Figure 4.8 Results of gesture matching approaches (subtree intersection, tags tree kernel, subpaths tree kernels, combined) in the matching scores. 76

Figure 4.9 Task completion time and Idle time per Helper-Worker pair. 78

Figure 4.10 Number of actions and errors per Helper-Worker pair. 79

Figure 4.11 Pearson correlation matrix for the task understanding proxy metrics. All metrics reported significant correlations in all the tasks (block assembly correlations = $p \leq 0.05$; origami and ultrasound training correlations = $p \leq 0.04$). 81

Figure 4.12 Normalized understanding assessment questionnaire answers, divided based on the participants’ role. 85

Figure 4.13 Normalized NASA TLX answers, divided based on the participants’ role. 86

Figure 4.14 Correlations between the PIA metric and all the UAQ and TLX questions, divided based on the participants’ role and the task performed. An asterisk represents statistical significance between PIA and the respective criterion ($p \leq 0.05$) 87

ABSTRACT

Effective collaboration in a team is a crucial skill. When people interact together to perform physical tasks, they rely on gestures to convey instructions. This thesis explores gestures as means to assess physical collaborative task understanding. This research proposes a framework to represent, compare, and assess gestures' morphology, semantics, and pragmatics, as opposed to traditional approaches that rely mostly on the gestures' physical appearance. By leveraging this framework, functionally equivalent gestures can be identified and compared. In addition, a metric to assess the quality of assimilation of physical instructions is computed from gesture matchings, which acts as a proxy metric for task understanding based on gestural analysis. The correlations between this proposed metric and three other task understanding proxy metrics were obtained. Our framework was evaluated through three user studies in which participants completed shared tasks remotely: block assembly, origami, and ultrasound training. The results indicate that the proposed metric acts as a good estimator for task understanding. Moreover, this metric provides task understanding insights in scenarios where other proxy metrics show inconsistencies. Thereby, the approach presented in this research acts as a first step towards assessing task understanding in physical collaborative scenarios through the analysis of gestures.

1. INTRODUCTION

Effective group collaboration is often necessary to perform shared physical tasks. For example, effective collaboration is of paramount importance in training programs that involve interacting with the physical environment, tool use and manipulation (Gauglitz et al., 2012; Kurillo et al., 2008). One of the keys for effective collaboration is achieving a common understanding of the shared tasks (Kieras & Bovair, 1984). Consequently, a factor that has proven beneficial to attain this shared understanding is a constant use of gestures between the collaborators as they interact. Gestures have been found to be key indicators of task understanding and learning (Church & Goldin-Meadow, 1986; Goldin-Meadow, 2005; Goldin-Meadow & Sandhofer, 1999; Knoblich & Sebanz, 2006; Ping et al., 2014). Moreover, gestures are linked to learning and problem solving, working memory allocation, and information recall (Cook et al., 2012; Frick-Horbury, 2002). In spite of the relevance of gestures in collaborative tasks, assessment of task understanding is usually done without considering gestures themselves but based on the outcomes directly. Furthermore, when gestures are transmitted between parties in such systems, is usually for illustrational purposes (Fussell et al., 2004; Kirk, Rodden, & Fraser, 2007).

Three fundamental problems need to be addressed to analyze collaborative process via gestures. First, there is a need to create a rich gesture representation that encompasses the gestures' morphology, meaning, and context. Although several approaches already exist (Asher & Lascarides, 2003; Madapana & Wachs, 2017; Parvini et al., 2009), these approaches either do not consider key aspects of the gesture's meaning and context, and most often are not presented in an analytical form. For example, some of these approaches are intrinsically linked to the gesture's appearance, which may differ significantly without necessarily implying lack of mutual understanding. This leads to the second fundamental problem: there is currently no approach to compare gesture representations. Finally, a process to analyze collaborative process via gestures is necessary. Traditionally, task understanding is assessed using subjective techniques such as interviews and concept mappings (Kemp et al., 2008; White & Gunstone, 2014), or indirect objective metrics such as completion time, number of errors, or conversational analyses (Barmby et al., 2007; Skemp, 1976). However, these metrics have shown to be often inconsistent when assessing task understanding (Pachella, 1973; Wickelgren, 1977). We argue and provide evidence throughout this thesis that gesture performance-related metrics (e.g. morphological and semantic similarities) provide useful insights to evaluate collaborative interactions between agents in physical tasks.

1.1 Significance

Gestures are key components in human interaction. When people perform a physical task, they constantly gesture, either as means of task completion (e.g. assembling parts) or as means of inquire or instruction. These gestures contain information related to the what, why and how of the task being performed. Analyzing these gestures can reveal information of how well the tasks are being understood and performed, “assimilated” in brief. For example, the hands’ physical motions (i.e. gestures) sonography students perform when learning how to operate an ultrasound device are an integral part of the assessment of their training process. Therefore, comparing the gestures performed by trainees to those of their instructors while performing tasks can reveal meaningful insights about cognitive processes they are experiencing, the knowledge gain and the overall learning process.

These insights can be extended to the assessment of collaborative process between individuals. Collaborating to perform shared tasks is a crucial skill: the ability to tackle problems with others is key in a world with a dynamic and ever-growing workforce. Nonetheless, this collaboration needs to be effective for the tasks to be completed in a successful way. Evaluating how well the team members understand the goals and instructions of the task through their actions is a natural way to estimate if the task is being performed correctly. A gesture-based approach to estimate task understanding can complement currently available task understanding metrics while alleviating their subjectivity.

Nonetheless, this work is not limited to human-human collaboration. Physical interactions are also used when interacting with robots and virtual agents, which may interact back. Although the kinematics of such systems may different significantly from that of a human, the interactions between them could be described using a paradigm similar to the one describing human-human collaboration. By defining a framework to represent, compare, and assess gestures, better coaching and assessment protocols can be implemented for human-human and human-machine collaboration.

1.2 Definitions

In this section, some important terms used throughout this document are defined.

1. **Collaborative Task**: An assignment in which two or more individuals work together to achieve a common goal.
2. **Task Understanding**: Acquiring the knowledge of the necessary steps to solve a problem.
3. **Semantics**: Aspects describing the meaning of an object or event.
4. **Pragmatics**: Aspects describing the context in which an object is, or an event happens.
5. **Gestural Analysis**: Evaluating the appearance and content of gestures using quantitative and qualitative tools.
6. **Gesture Similarity**: Metric showing the degree in which two gestures are functionally equivalent. Such resemblance can represent the morphology, semantics or pragmatics.

1.3 Research Problem

The design and implementation of an approach to assess collaborative physical tasks via gestural analysis includes: (a) representing gestures encompassing their morphology, semantics, and pragmatics; (b) comparing such representations quantitatively to obtain gesture similarity; and (c) estimating how well is the task being understood based on the aforementioned gestural comparisons.

1.3.1 Research Question 1 (RQ1)

How to analytically and compactly represent gestures including their morphology, meaning, and context?

Gestures encapsulate a large amount of information. Representing this information and representing it will allow for comparisons and assessments between the gestures performed by different individuals. This question surveys what are the aspects of gestures that should be captured, and how to represent them in an analytical yet compact manner.

1.3.2 Research Question 2 (RQ2)

How to measure gesture similarity and compare between gestures?

Identifying which gestures are functionally equivalent can provide benefits in task assessment. Having represented the information present in gestures (RQ1), an approach to compare this information is developed. These comparison strategies should consider the gestures' morphology, semantics, and pragmatics. In doing so, this question explores how to quantify the level in which two gestures resemble each other. Experiments will be conducted to measure the similarity of the gestures performed by individuals collaborating to complete a common task.

1.3.3 Research Question 3 (RQ3)

How can gesture similarity lead to estimate task understanding?

Once we represent (RQ1) and compare (RQ2) gestures from different individuals performing a common task, this question addresses how well the collaborators understand the instructions they are receiving by analyzing the gestures they display.

1.4 Overview of the Document's Structure

Chapter 1 introduces the research problem and the motivation behind it. Chapter 2 provides a review of the literature on the topics related to this research. Chapter 3 explains the proposed methodology used to tackle the research problem. Chapter 4 discusses experiments and the obtained results. Finally, in Chapter 5, conclusions and future work is discussed.

2. LITERATURE REVIEW

This chapter gives an overview of the state-of-the-art on the topics related to this proposal, namely gestures, collaboration, and task understanding. First, the importance and challenges of gestures in human-human interaction are discussed. The use of gestures is linked to a large variety of benefits: it can help grounding new concepts and learning new tasks, and can reduce the cognitive load experienced during task performance. Research has shown that gestures are linked to improved task understanding. The second section of the literature review explores the importance of gesture in collaborative settings, particularly those in which the collaborators are not co-located. Protocols to represent and communicate gestures remotely in such situations have been defined and tested, and will be detailed in this section. In the third section, the current techniques to represent and compare gestures will be presented. There are currently two avenues to represent gestures: morphology-based approaches and semantics-based approaches. Examples of these approaches will be given. Finally, the state-of-the-art approach to measure task understanding are discussed. Measuring task understanding can be performed through subjective and objective metrics. The benefits and challenges of both these traditional approaches will be presented and discussed.

2.1 Importance and Challenges of Gestures in Human Interactions

Gestures are an essential component in human interactions and development. Several theorists point that non-vocal signing preceded the evolution of oral speech (Hewes et al., 1973). Whether consciously or not, humans actively gesture to inform one another about their intentions and ideas (Kendon, 2004). For example, whenever individuals interact face-to-face or mediated through devices, a substantial portion of this exchange is done through non-verbal means such as gesture (Argyle, 2013). In this exchange, gestures are used to express emotions, communicate interpersonal attitudes, and to accompany and support speech and other culture-specific symbolisms (Argyle, 2013, Ch. 1). Additionally, gestures facilitate the acquisition of new knowledge and offer content redundancy (McNeill, 1985). McNeill elaborates by describing how gestures and speech have parallel semantic and pragmatic functions: gestures and speech share a computational stage in the brain, and therefore one helps in the generation of the other (McNeill, 1985, p. 370).

The use of gestures, however, goes beyond face-to-face interactions: people engage in nonverbal communication behaviors even when they are not co-located (Wei, 2006). For example, Lee reports how people gesture when speaking over the phone, even if the other person was not able to visualize those gestures (Humphreys, 2005). This example demonstrates how fundamental gestures are in the everyday routine.

The study of gestures presents implicit challenges. For instance, gestures do not necessarily communicate meaning: they can be used as an aid for recall, or be performed involuntarily with no particular meaning (Kendon, 2004, Ch. 6). Moreover, studies on gesture agreement reveal that gestures are not universal: people might perform different gestures to refer to the same concept, or even perform the same gesture to refer to different concepts (Gonzalez et al., 2018; Vatavu & Wobbrock, 2015; Wobbrock et al., 2005). Furthermore, gestures are culture-dependent: when gestures are used to express abstract concepts, the mapping between the gesture and the concept might not be present in very culture (McNeill, 1985, p. 370; Molnar-Szakacs, Wu, Robles, & Iacoboni, 2007). Moreover, gestures might still differ even when such concept mappings exist. Albeit these challenges, gestures persist as an active field of study. In this subsection, we present a detailed overview of specific evidence of how gestures are beneficial to human interactions.

2.1.1 Gestures for cognitive offload, learning and problem solving

Gestures can indicate readiness for learning new concepts, and can reveal and stimulate the learners' thoughts. Roth and colleagues have studied the impact of gesture usage in students learning scientific concepts (Roth, 2001; Roth & Welzel, 2001). Their observational studies reported that students relied on gestures in instructional settings where concepts were hard to explain only using speech. This behavior not only allowed students to learn the concepts better, but also lowered the cognitive loads students experienced when explaining complex constructs.

In the context of learning, research has shown that individuals who gesture had better problem-solving abilities and retain the knowledge better. Cook and colleagues analyzed the impact gestures had in the children's abilities to solve mathematical problems (Broaders et al., 2007; Cook et al., 2008). Their results concluded that children who gestured were able to retain the knowledge better, as compared with children that were not allowed to gesture. This finding was confirmed with a follow-up test performed 4 weeks after the learning session: the relation between instruction and learning was weak for children that were not allowed to gesture. Moreover, students that were told to gesture while explaining their thought process and answers came up with new problem-solving strategies expressed only through gestures. These findings corroborate that gestures play an important role in the learning process: the process of gesturing reinforces the assimilation of new ideas. Radford also evaluated the impact of gestures in the context of learning mathematics, and found that gestures allowed students to solve problems that they were not able to do so before: gestures were part of the students' abstract thinking (2009). Along these lines, Chu and Kita analyzed the impact of gestures in individuals performing spatial rotation tasks (2011). Their study revealed that individuals who were allowed to gesture performed more problems correctly. The problem-solving strategies developed by this group persisted even after the same group was not allowed to gesture anymore. Their study concluded that gestures allowed participants to offload part of the mental calculations that were required to perform the task.

Gestures are also linked to increased information recall. For example, a study showed that people who gestured were able to recall more words when compared to individuals that did not gesture, both immediately and after a two-week interval (Frick-Horbury, 2002). Gestures allow creating meaningful associations with the words, both in the case of concrete words (e.g. oven, hammer) and abstract words (e.g. moral, development). Such associations can reduce the demands on individuals' working memory. Cook, Yip, and Goldin-Meadow explored this idea by asking participants to remember letters while explaining their solutions to math problems and producing different types of movements (i.e. gestures, no gestures, meaningless movements) (2012). Participants recalled significantly more letters when gesturing, which was attributed to four possible mechanisms: 1) gestures represent information in a format that complements speech; 2) gestures did not complement speech, but introduced redundancy by conveying the same information in multiple formats; 3) gestures did not help with the way information was presented, but in the way attention was directed; or 4) gestures externalized the speakers' ideas. However, regardless of the mechanism, gestures effectively led students to a better recall process.

Finally, gestures are linked to the process of speech production. Several theorists point that non-vocal signing preceded and guided the evolution of oral speech (Hewes et al., 1973). Although different authors dispute the way in which this process happens, the role of gestures in speech production process is not disputed (Butterworth & Hadar, 1989; Feyereisen, 1987; McNeill, 1987, 1989). Given this relation, gestures play a cognitive role in conceptualizing the messages that will be verbalized. Therefore, gestures play a critical factor in human performance by helping speakers organize their ideas (Alibali et al., 2000).

2.1.2 Gestures for common grounding

Clark and Brennan describe grounding in communication as a process thorough which mutual knowledge, beliefs, and assumptions between agents in a conversation (or interaction) is created (1991). This collective process includes contributing to a conversation, confirmation of understanding, creation a shared pool of concepts, among other examples. As part of conversational grounding, mental models are shared among the participants, leading to an improvement of team performance and decision-making (Converse et al., 1993; Espinosa et al., 2001, 2002; Mathieu et al., 2000).

Gestures facilitate the process of common grounding. Fussell and colleagues extensively reviewed the role of gestures in conversational grounding (Fussell et al., 2004). Different types of gestures are used to convey different aspects of the information among the team, such as pointing, iconic, spatial, and kinetic gestures. For example, a pointing gesture is used to highlight a specific object in the environment, and an iconic gesture can emphasize unique object’s attributes. Figure 2.1 provides an example of how gestures can aid to the conversational grounding between agents. Alibali et al. reported that, in some cases, people reveal aspects of their mental processes and their representation through gestures instead of speech (1999). Such mental representations include visual or perceptual information that individual might find easier to communicative using gestures.

Finally, gestures can be the channel by which those mental models translate into tangible actions (Clark, 1996; Clark & Marshall, 2002). For example, the authors illustrated how pointing gestures can be used to describe the mental representation of a required book in a library, allowing the librarian to create associations and eventually identify the referred book.

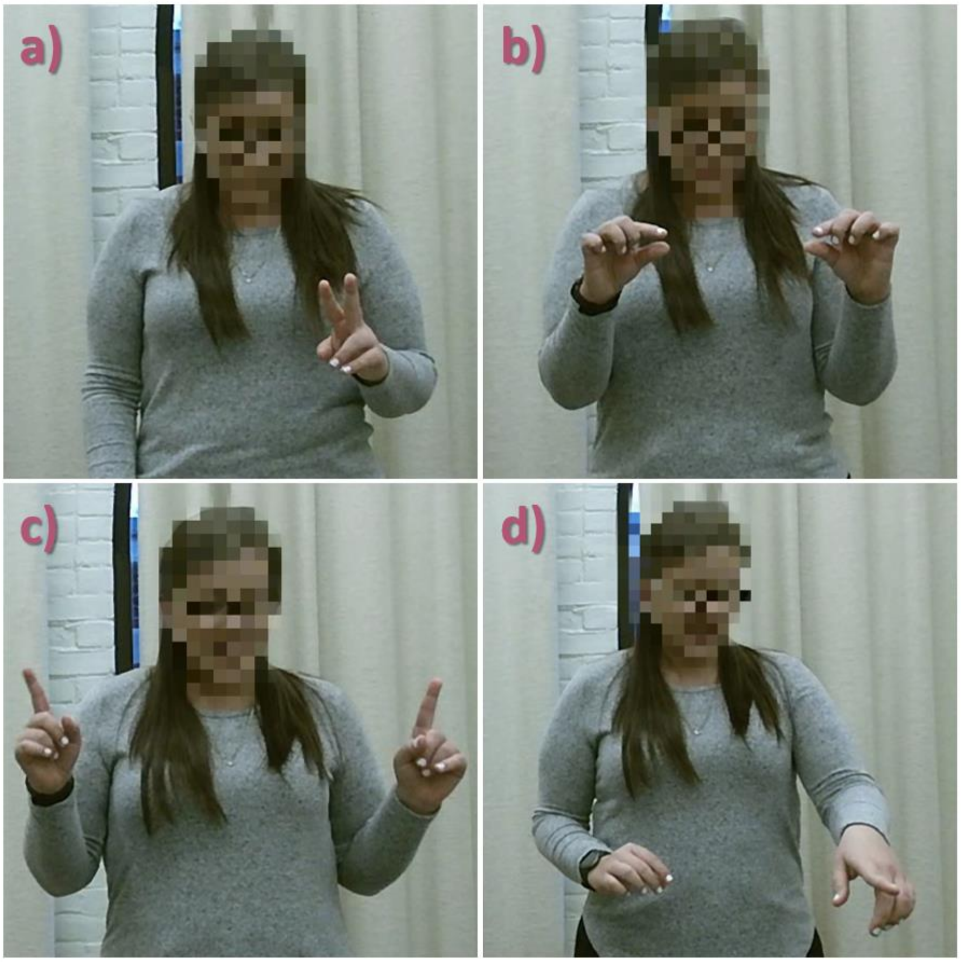

Figure 2.1 Gestures aiding to conversational grounding. Through gestures, individuals can indicate a) quantities, b) shapes, c) distances, and d) locations.

2.1.3 Gestures for understanding

When individuals interact to complete a shared task or refer to a common concept, a crucial element in the common grounding process is to develop a shared understanding of the task being referred. Achieving understanding of a task involves acquiring the knowledge of the necessary steps used to complete the task successfully (Skemp, 1976). This understanding can include the use of a tool, grasping implicit social cues, or simply acknowledging information received during a conversation. Additionally, research reveals that achieving a proper understanding of a task can lead to an increased social awareness and performance. For example, children can develop a greater social awareness whenever they acquire an understanding of the language, social cues, and emotions in the environment that surrounds them (Carpendale & Lewis, 2006). In another example, adults' motivation in their work environment increased whenever they acquired a better understanding of the tasks they were in charge of (Porter, Bigley, & Steers, 2003). Finally, Kieras and Bovair provided evidence of how learning can be increased whenever individuals understand the task and tools they are interacting with (1984).

A large corpus of studies has focused on the importance of gestures in task understanding. For instance, nonverbal cues can be beneficial in the development of understanding of social expressions and symbols by children (Leekam et al., 2010). In the case of adults, observing individuals gesturing while performing a task can help understand their goals and intentions, which can translate into a better understanding of the task (Goldin-Meadow & Beilock, 2010). This improvement in understanding by observing others gesturing has been linked to how gestures can activate the self's motor system (Calvo-Merino et al., 2004). For example, Sebanz and colleagues illustrated how the areas in the brain controlling motor responses were activated as people observed their teammates performing gestures (Sebanz, Knoblich, et al., 2006). In their study, participants solved a spatial task either alone or with partner, and were refrained from acting at certain moments. During these interruptions, the activation in the areas of the brain related to motor responses and action perception were recorded. The amplitude in the brain signals was significantly higher for participants working with a partner: their brain activated more as they observed the actions of their partner. Ping et al. also reported evidence on the relation between activations in the motor system and observing others performing gestures (2014). Finally, Reed and McGoldrick conducted a study in which participants were asked to determine whether two pictures of body postures were equal (primary task) (2007). Simultaneously, participants were presented with body postures they needed to replicate (secondary task) while performing the primary task. Their results showed that participants were able to understand and memorize the body postures from the primary task better when the body parts involved the both the primary and secondary tasks were congruent (e.g. when both tasks involved movements of the right arm). Therefore, when the gestures performed and observed by a participant were congruent, better task understanding and performance was achieved.

Observing others' gestures not only shapes the understanding of the task, but how subsequent actions are planned and executed. Sebanz, Bekkering, and Knoblich performed an extensive review of how individuals coordinate their actions based on the actions and gestures performed by others (2006). As the authors detailed: "*action coordination is achieved by integrating the 'what' and 'when' of others' actions in one's own action planning*" (2006, p. 75). The impact of others' gestures in problem solving has also been analyzed. Similar experiments demonstrated that the understanding and performance of a task is affected by the gestures performed by others while completing the same task (Beilock & Goldin-Meadow, 2010; A. Hamilton et al., 2004). Cook and Tanenhaus also analyzed this phenomena and concluded that the perceptual-motor information conveyed by others' hand gestures affects the actions performed by other by foreshadowing subsequent actions (2009).

Finally, gesture observation is particularly important during tasks in which there is a mismatch between the information conveyed through gesture and through speech. For example, Goldin-Meadow and colleagues have led several studies about the "gesture-speech" mismatch and their effect in cognitive processes experienced in children (Goldin-Meadow, 2005). Such mismatches can also reveal 1) whether children are misunderstanding mathematical concepts, or/and 2) whether children are more receptive to receive further instruction in those particular mathematical concepts in it (Church & Goldin-Meadow, 1986). Adults observing such children gesturing were able to understand better the thought processes that the children experienced, even in situations where the information was not explicitly provided through speech (Goldin-Meadow & Sandhofer, 1999). Other works have provided evidence regarding the different roles gestures have depending on whether they are produced with or without speech (Goldin-Meadow, 2006; McNeill et al., 1994). For example, gestures co-produced with speech tend to include information that is not present in speech, whereas gestures produced in isolation often present characteristics that resemble natural language (Goldin-Meadow, 2006, p. 35).

2.2 Gestures in Remote Collaborative Tasks

Achieving effective collaboration is necessary for team work: it can lead individuals to develop of mutual understanding, reduce errors, and enable a sense of progress and engagement (Martinez-Moyano, 2006). Consequently, a factor that has proven beneficial to sustain this collaboration is the use of gestures between the collaborators as they interact during a task to support projectability of actions (Fussell et al., 2004). This alludes to the capacity gestures have to assist in the projectability of actions. Projectability refers to the way an action or part of it foreshadows another one (Auer, 2005). Gestures assist in this process by allowing the individuals collaborating to predict, anticipate, and prefigure the unfolding of their teammates' actions (Kuzuoka et al., 2004). Nonetheless, this projectability of actions is possible when the team has a shared visual space available, i.e. they are able to see the gestures performed by the others (2004; p. 478). These shared visual spaces allow teams to interact via gestures, which translates into a better ability to assess comprehension and task state (Kraut et al., 2003).

2.2.1 Remote collaborative systems

Clark and Brennan defined eight constraints that the communication medium imposes in the individuals collaborating (1991). One of these eights constraints deals particularly with the copresence of the collaborators: whenever collaborators are not co-located, more effort is required for collaboration. This relates back to the concept of projectability: whenever the projectability of actions, and of gestures in particular, gets hindered, the individuals' situation awareness and common grounding also gets reduced (Luff et al., 2003). Remote collaborative systems have been developed to enable work when the team members are not co-located. The idea is to use such systems to transfer a representation of the collaborators' physical actions and messages (through gestures). While such systems can maintain the projectability of actions (Heath & Luff, 1991), they are not a replacement for face-to-face interactions since the communication mutual grounding is limited by the technological meditation (Brennan, 1998). Without the possibility of direct interaction, remote collaboration systems offer the best alternative to allow natural interaction between individuals in distributed settings (Kuzuoka et al., 2004). For example, Tang demonstrated that transmitting gestures remotely allowed collaborators to regulate turn taking, negotiate shared spaces, store information, and express ideas (Tang, 1991). In the following, we will focus on different approaches to transmit gestures among participants in a collaborative task.

Two types of systems can facilitate the projectability of actions and gestures: linked gesture systems and mediated gesture systems (Kirk, Crabtree, & Rodden, 2005). Linked systems create shared space in which the gestures of the collaborators can be visualized. For example, Kirk and colleagues used projectors (Figure 2.2a) to display the gestures of the remote collaborators (Kirk et al., 2005, 2007; Kirk & Fraser, 2005, 2006) and found that although such systems have to ability to promote awareness, the difference between the viewpoint of the collaborators can hinder the projectability of actions and gestures. Mixed reality has also been leveraged to create linked gesture systems. For example, two groups leveraged mixed reality 2D displays to visualize gestures remotely (Kirk, Crabtree, & Rodden, 2005; Kirk & Fraser, 2005; Yamashita, Kaji, Kuzuoka, & Hirata, 2011) and found that such system can improve performance, however, at a cost of making the collaboration process more impersonal. Shared visual spaces between the collaborators have also been established via head-mounted cameras (Figure 2.2b) (Alem et al., 2011; Kraut et al., 2003, 1996) which allowed collaborators to complete shared tasks but introduced issues related to video latency and motion sickness.

The second type of platforms is the mediated gesture systems. This type of system employs a representation of the gesture that may not directly resemble the visual appearance of the gesture. The goal is not necessarily to transmit the shape and movement of gestures, but to represent them in a way that facilitates the task performance and understanding. For example, the DOVE drawing platform allows a remote collaborator (i.e helper) to create lines over a live video feed of the local collaborator's (i.e. worker) workspace (Fussell et al., 2004; Ou et al., 2003). Platforms that leverage laser pointers and gaze information to represent gestures conveying spatial information have also been explored (Figure 2.2c) (Akkil et al., 2016; Kuzuoka et al., 2004; Luff et al., 2003) to convey gesture related information. For example, a laser attached to a mobile robotic platform allowed a remote helper to point at specific locations in the worker's workspace. Finally, three-dimensional Augmented Reality (AR) representation of the collaborators' hands (Figure 2.2d) (Huang et al., 2019; Huang & Alem, 2011, 2013; Sodhi et al., 2013; Tecchia et al., 2012; Wang et al., 2019; Zenati-Henda et al., 2014). In these AR-based systems, the helper's hands are captured and appear as virtual objects visible to the worker.

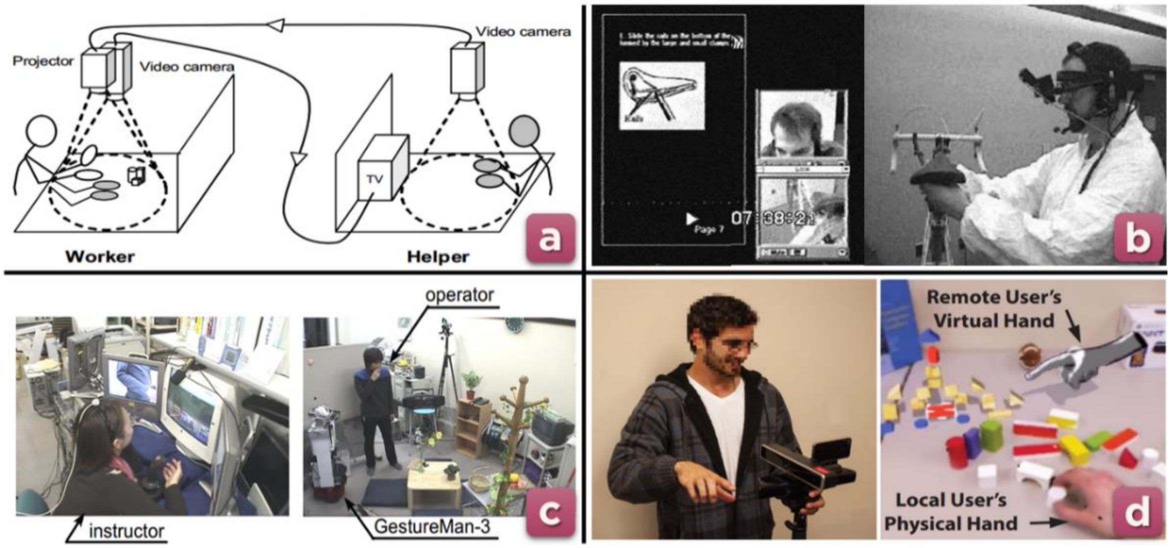

Figure 2.2 Examples of remote gesture collaboration systems. The image includes remote collaboration systems based on: a) projectors (D. Kirk et al., 2007), b) head-mounted cameras (Kraut et al., 2003), c) laser pointers (Kuzuoka et al., 2004), and d) AR (Sodhi et al., 2013).

2.3 Gesture Representation, Comparison and Assessment

This chapter has discussed the importance of gestures in collaborative process and how are gestures conveyed in scenarios where individuals are not co-located. The previous remote collaboration systems share a limiting factor. These systems relegate gestures to an illustrational role: a representation of the gestures is communicated, but there have been no attempts to discover insights regarding the quality of the collaboration through the gestures. Nonetheless, assessing the quality of collaborations through gestures would require finding a way to represent and compare the gestures performed by the as they collaborate. The next chapter presents in detail our approach to represent and compare gestures to perform these assessments. Therefore, this subsection provides an overview of the state-of-the-art in terms of techniques to tackle this problem.

Gesture representation and comparison methodologies follow two main avenues: morphology-based and semantic-based representations. The morphology-based view is intrinsically linked to the gesture appearance. Comparisons performed over these structures will typically depend on the location of the hands with respect to the body (relative or absolute), motion, orientation, and visual appearance among others (Mitra & Acharya, 2007). Techniques to compare this type of representations include statistical modeling (Caridakis et al., 2010), neural networks embeddings (Ge et al., 2008), and distance metrics (Zhao et al., 2013). Conversely, the semantic-based view constructs logical representations of gestures based on linguistics frameworks (Asher & Lascarides, 2003; Montague, 1970; Potts, 2005). The focus of these representations is to encompass information related to the meaning and the context of the gesture; methods to obtain quantitative insights from these structures are yet to be explored.

2.3.1 Morphology-based gesture representations

Fundamental morphology-based representations leverage trajectory or appearance based information obtained directly on the gesture-capturing device. Whenever non-optical capture devices are employed, gestures are usually represented as an arrangement of joint positions and angles over time. Glove-based systems are the most common non-optical capture device to obtain motion and trajectory related data (Figure 2.3a) (Dipietro et al., 2008; Parvini et al., 2009). These systems leverage sensors attached to a glove that is worn in the hand. The position of these sensors can then be acquired via three-dimensional tracking or force-feedback actuators. Conversely, vision-based models capture the gestures using color and infrared cameras. After acquiring the gestures, they can be represented with a variety of approaches such as succession of motion signatures (Figure 2.3b) (Chiu & Marsella, 2014), filter-extracted features (Figure 2.3c) (Konečnỳ & Hagara, 2014; Rautaray & Agrawal, 2015), neural network-generated embeddings (Figure 2.3d) (Ge et al., 2008; Stergiopoulou & Papamarkos, 2009), among others. Finally, recent one-shot learning techniques leveraging in neurological features have been used to represent gesture classes (Cabrera et al., 2017), and morpho-semantic binary descriptors (Madapana & Wachs, 2017).

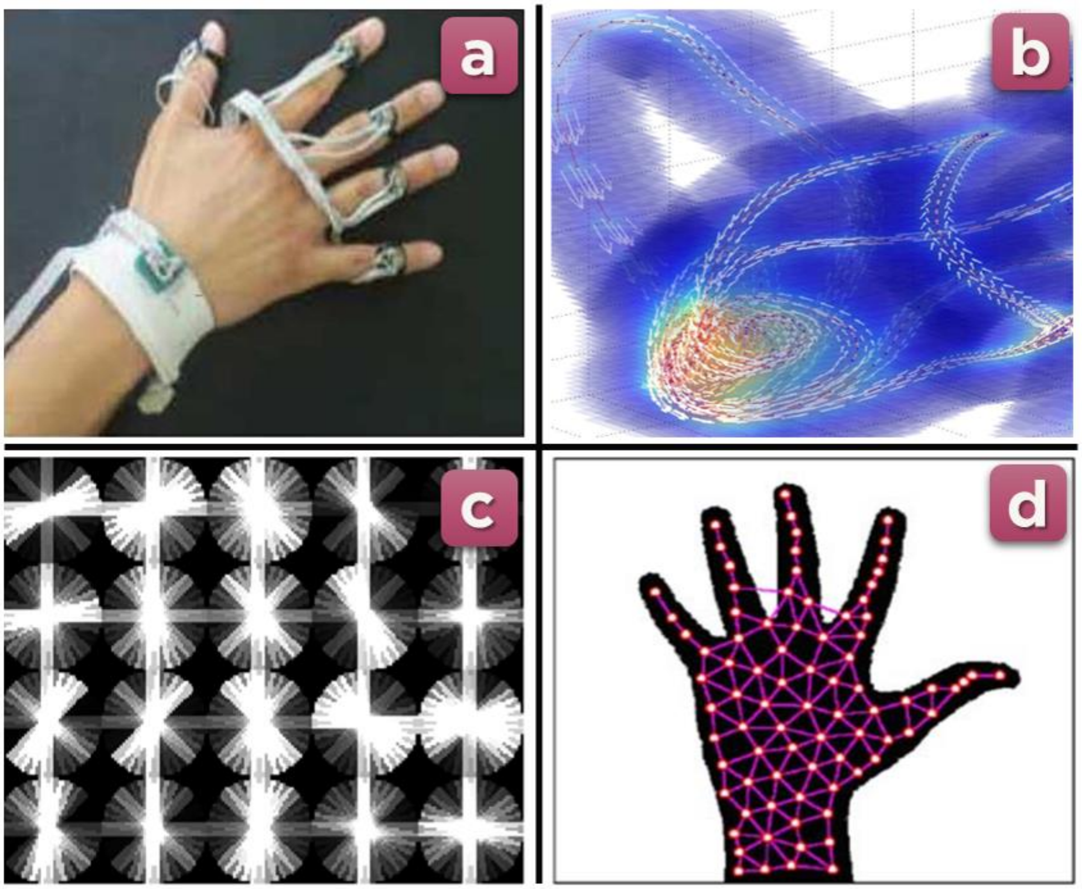

Figure 2.3 Examples of morphology-based gesture representation techniques. The image includes gesture representation approaches based on: a) glove-based systems (Dipietro et al., 2008), b) motion signatures (Chiu & Marsella, 2014), c) filter-extracted features (Konečný & Hagara, 2014), d) network-generated embeddings (Stergiopoulou & Papamarkos, 2009).

Nonetheless, these approaches are intrinsically linked to the gestures' appearance, and therefore gestures need to be physically similar to be considered equivalent. This is usually not the case in remote scenarios where agents have different culture, backgrounds, and expertise levels (e.g. mentor-mentee scenarios). In those cases, the gestures performed by the team members may differ significantly without necessarily implying lack of conversational grounding. Figure 2.4 presents an example of such situation: the gestures of all the agents are functionally equivalent, but their physical appearance differs.

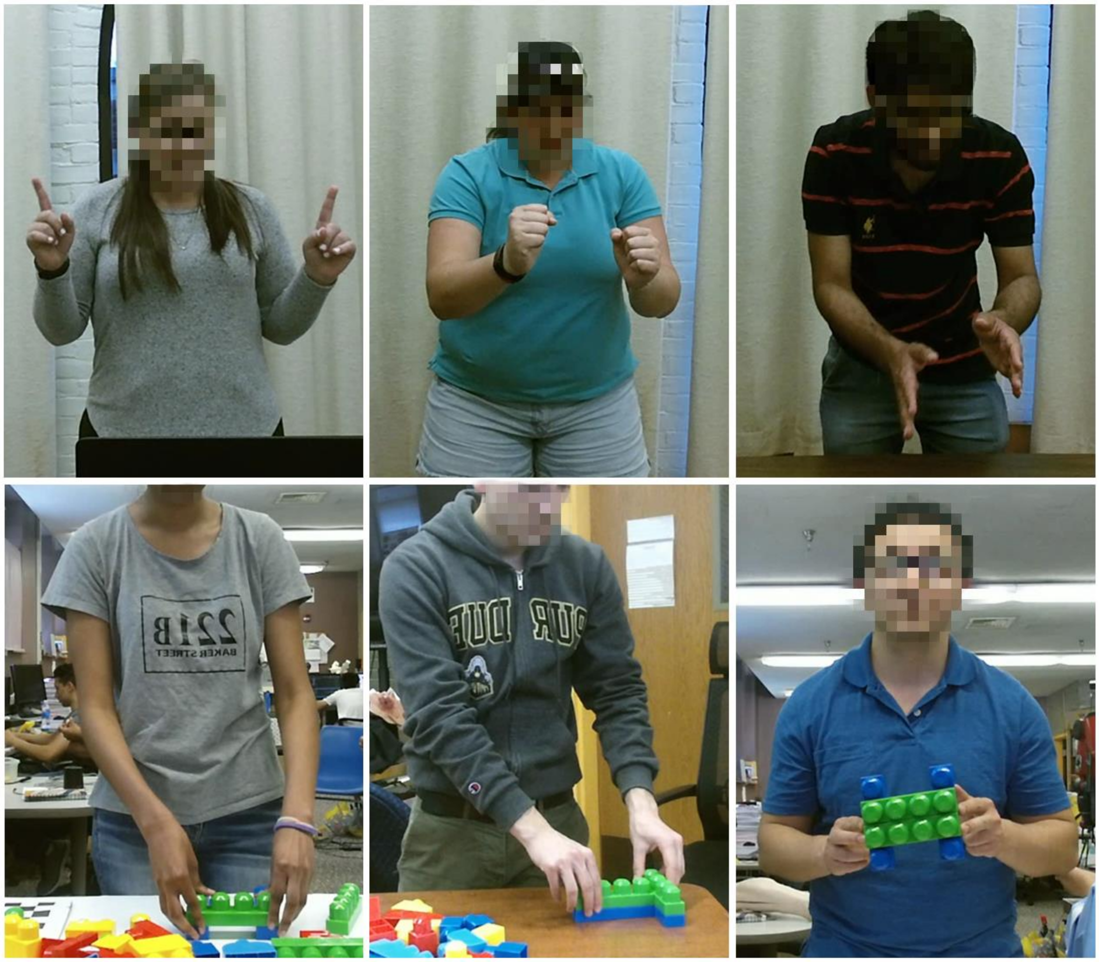

Figure 2.4 Analogue gestures with different visual appearance.

2.3.2 Semantics-based gesture representations

Semantic-based representations are the second approach for gesture comparisons. The goal of these representations is to create formal representation of the gesture's meaning and context. 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 (2010, 2011). Focused specifically on iconic gestures, Giorgolo defines the concept of the iconic space of the gesture to describe the gestures' spatial configuration. This configuration is then expressed with an additional table representing the movement and shape of the gesture. Similarly, Lascarides and Stone described gestures' semantics (i.e. aspects describing the meaning of the gesture) and pragmatics (i.e. aspects describing the context in which the gesture was generated) with a table that complements their dynamic semantics framework (2009). Their framework will be explained in detail in their next chapter, as our approach to represent gestures builds on top of their framework. Finally, co-speech gesture projection is another approach to represent gestures that has gained attention lately (Ebert & Ebert, 2014; Schlenker, 2018). The co-speech gesture projection approach creates a formal representation of gestures, paying special attention to whether the gestures accompany or supplement the spoken information.

However, these different tables or formal clauses were conceived only as representation structures. This means that while all these approaches successfully represent the gestures' semantics and pragmatics, comparison and assessment of gestures using these structures have not been explored to the best of our knowledge.

2.4 Approaches to Estimate Task Understanding

Despite gestures' relevance in conveying information during collaborative tasks, gestures are mostly neglected while assessing task understanding. Instead, task understanding is traditionally assessed using subjective and objectives approaches related to task performance. Examples of these approaches will be explained in this subsection, providing insights about how to measure understanding.

2.4.1 Subjective approaches to estimate understanding

White and Gunstone presented subjective approaches to measure understanding in an educational setting (2014). Students' understanding was probed using techniques such as interviews, concept mappings, relational diagrams, and inquiries. The idea of these techniques is to elicit responses in the individuals that display the degree in which a concept is being grasped.

These approaches are used in a variety of settings. For example, Kemp et al. studied the impact of different inquiry methods to assess the understanding of patients about their own medical condition (2008). Several studies have reported the use of subjective approaches to measure understanding in the context of formal education. Watson et al. used a questionnaire with a coding scale to measure the understanding that elementary and middle school students had about the concept of statistical variation (Watson et al., 2003). Their findings allowed them to divide understanding into different tiers that provided information to teachers evaluating the students' progress. Kannemeyer also utilized a coding strategy to categorize the responses given to a questionnaire by calculus students (2005). These coding approaches are widely adopted, as they enable quantitative measures of understanding from unstructured qualitative data sources. Concept mappings have also been studied as a reliable way to assess the understanding of science and mathematics concepts (Brinkmann, 2003; Mintzes et al., 2005). By representing concepts graphically in the form of a network, the underlying thought processes experienced when understanding a concept can be tracked and reinforced with ease. Subjective approaches to measure understanding have also been widely studied in the aviation domain. Different frameworks such as the Model for Assessing Pilots' Performance (MAPP) and the Line Operational Evaluation (LOE) have been used to evaluate the understanding of crew members about the different tasks to perform during flight (Baker & Dismukes, 2002; Mavin & Dall'Alba, 2011). These approaches rely both on the judgements of instructors and self-reported scores to evaluate how much understanding and situation awareness the person has regarding a flight task.

Albeit effective, subjective approaches to assess understanding are highly sensitive to aspects extraneous to understanding. For example, the capacity of the individuals to contextualize the received information can play a role while assessing for understanding (Gumperz, 1992). Gumperz elaborated this idea by explaining how aspects such as situated interpretation and inferencing capacities can impact the way understanding is perceived and measured. Additionally, approaches like LOE are dependent on accurate observation of the people's behavior, which requires the evaluators to be trained on how to assess the people's behavior adequately (Baker & Dismukes, 2002, p. 211).

2.4.2 Objective approaches to estimate understanding

Conversely, objective approaches rely on indirect metrics to measure understanding. Examples of these metrics include completion and idle time (Hoffman, 2013), number of errors (Barmby et al., 2007; Skemp, 1976), reaction time (Ping et al., 2014), conversational analyses (Kirk et al., 2007), among others. These approaches are common in several disciplines, such as medical training (Ju et al., 2012; Mohan et al., 2017), aviation (Estival & Molesworth, 2016; Molesworth & Estival, 2015; Wu et al., 2019), human-robot collaboration (Hoffman, 2013; Hoffman & Breazeal, 2007; Nikolaidis & Shah, 2013), and education (Barmby et al., 2007; Harries & Barmby, 2008; Mintzes et al., 2001). Nonetheless, inconsistencies between different objective approaches can be found. For example, a high completion time does not necessarily implies a high number of errors: the trade-off between speed and accuracy has shown that spending more time understanding and executing a task can lead to reduced error rates (Chien et al., 2010; Wickelgren, 1977). A similar link has been encountered between error rates and reaction times: whenever participants attempted to have shorter reaction times, their accuracy was impacted (Pachella, 1973). The relation between reaction time and percentage of correct responses will usually follow a characteristic curve as the one presented in Figure 2.5 (1973, p. 38).

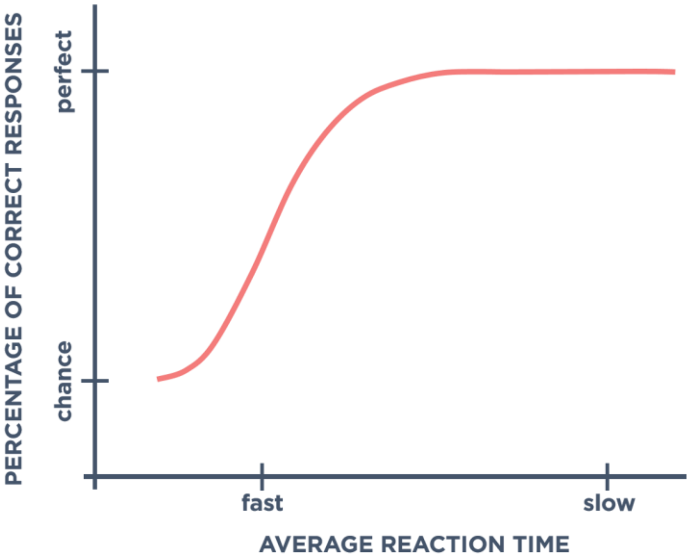

Figure 2.5 Reaction time vs. number of correct responses characteristic curve. Adapted from (Pachella, 1973).

2.5 Summary

This chapter presented the state-of-the-art in importance of gestures for human collaboration. Gestures can lead to improved learning outcomes, higher recall, and better learning and understanding. Additionally, the chapter surveys the use of gestures in collaborative systems for both co-located and remotely located collaborators. When the collaborators are not co-located, there is a need to remotely transfer instructions conveyed using gestures. This chapter discussed the way in which gestures are represented and compared. Nonetheless, current gesture representation and comparison approaches have intrinsic limitations. We hypothesize and present evidence in the next chapter that gestures should be represented and compared considering morphology, semantics and pragmatics. Finally, a new approach to measure task assimilation is discussed. The next chapter will leverage on some existing indirect metrics for task understanding and will be extended a novel metric based on gestural analysis.

3. METHODOLOGY

This chapter presents the proposed approaches to address the previously introduced research questions. These approaches are described in detail. The proposed solution is divided in two main elements. The first element is the Multi-Agent Gestural Instruction Comparer (MAGIC) architecture, a framework to represent and compare gestures' morphology (e.g. trajectories, shapes), semantics (e.g. meaning, timing), and pragmatics (e.g. meaning in context, environmental elements). This element addresses RQ1, related to defining the components of a rich gesture representation. Having created a rich gesture representation, RQ2 is addressed. RQ2 focuses on how to measure gesture similarity, which is tackled by gesture matching processes performed as the last step of the MAGIC architecture. RQ3 is addressed by the second element of our solution: the calculation of the Physical Instruction Assimilation (PIA) metric. This metric provides a score representing the quality of assimilation of physical instructions in scenarios where agents collaborate to solve a shared physical task. The system architecture is illustrated in Figure 3.1, which we will elaborate in detail later in this chapter. A comprehensive overview of the components of these architecture can be found in our previous work (Rojas-Muñoz & Wachs, 2019).

Consider the following vignette describing a collaborative task. Walter is a recently hired engineer in an engine manufacturing plant, and one of his main tasks will be maintaining the production line's robotic arms. Hanna, the company's senior engineer, will train Walter remotely on how to provide such maintenance. After establishing audiovisual communication, Hanna would be able to see Walter's workspace and instruct him on the specific steps involved in maintenance of robotic arms. For example, Hanna would tell "Take that wrench", followed by a gesture pointing at a particular wrench in Walter's workspace. After receiving this command, Walter would reach for the wrench and prepare for the next instruction. As Walter continues following these instructions, Hanna would correct him, as appropriate. These corrections would be done mostly based on the errors he makes and presented in the form of gestures mostly. Similarly, Walter may use gestures to inquire about particular specific steps. This iterative process would continue until the completion of the maintenance task.

Following the previous example, The MAGIC architecture and the PIA metric would provide another approach to assess Walter's understanding and performance by comparing the similarities between the gestures performed by Hanna and Walter. By adding another layer of knowledge concerning gesture performance together with other objective metrics, such as completion time and the number of errors, a more comprehensive evaluation of Walter’s performance can be obtained. This vignette will be referred throughout the thesis to exemplify the different elements of our architecture. Figure 3.1 presents a schematic of the architecture used to tackle these research questions. The elements will be explained in depth throughout the chapter.

3.1 Multi-Agent Gestural Instructions Comparer

MAGIC is a framework that can be used to represent and compare gestures’ morphology (e.g. trajectories, shapes), semantics (e.g. meaning, timing), and pragmatics (e.g. meaning in context, environmental elements). MAGIC introduces the concept of gesture-based evaluation, which provides a measurement of task understanding by analyzing the gestures performed by different parties. In doing so, MAGIC allows to perform comparisons between gestures that consider more information than just gesture appearance.

MAGIC draws inspiration from two semiotics frameworks to create a structure that represent gestures: Charles Morris' Theory of Signs (ToS), and Sebeok and Danesi's Modeling Systems Theory (MST) (Morris, 1938; Sebeok & Danesi, 2000). ToS is a framework to represent how humans perform the meaning creation process, and it inspires the main theoretical elements of MAGIC’s architecture, to be explained in this section (Morris, 1938). **Φ** *Agents* and ***A*** *Actions* are the first elements represented in the MAGIC framework. Every ***A*** *Action* represents a command that is conveyed, and each **Φ** *Agent* represents an individual generating and performing these commands. MAGIC's **Φ** *Agents* resemble ToS Interpreters, as both create interpretations from information. In our previous vignette, Hanna and Walter are the **Φ** *Agents*. Moreover, a distinction is done between a $\boldsymbol{\Phi}_{\mathbf{W}}$ *Worker*: the **Φ** *Agent* that directly manipulates the environment to accomplish the task (e.g. Walter); and the $\boldsymbol{\Phi}_{\mathbf{H}}$ *Helper*: the **Φ** *Agent* that communicates the commands required to perform the task (e.g. Hanna). The same distinction applies to the ***A*** *Actions*: Helper-generated actions are known as $\boldsymbol{A}_{\mathbf{H}}$ *Instructions*, and Worker-generated actions are known as $\boldsymbol{A}_{\mathbf{W}}$ *Executions*. Hanna’s verbal command “Take that wrench”, accompanied by a pointing gesture targeting a specific wrench in Walter's workspace is an example of an $\boldsymbol{A}_{\mathbf{H}}$ *Instruction*. Likewise, the gesture Walter performed to reach and grab the wrench is an example of an $\boldsymbol{A}_{\mathbf{W}}$ *Execution*.

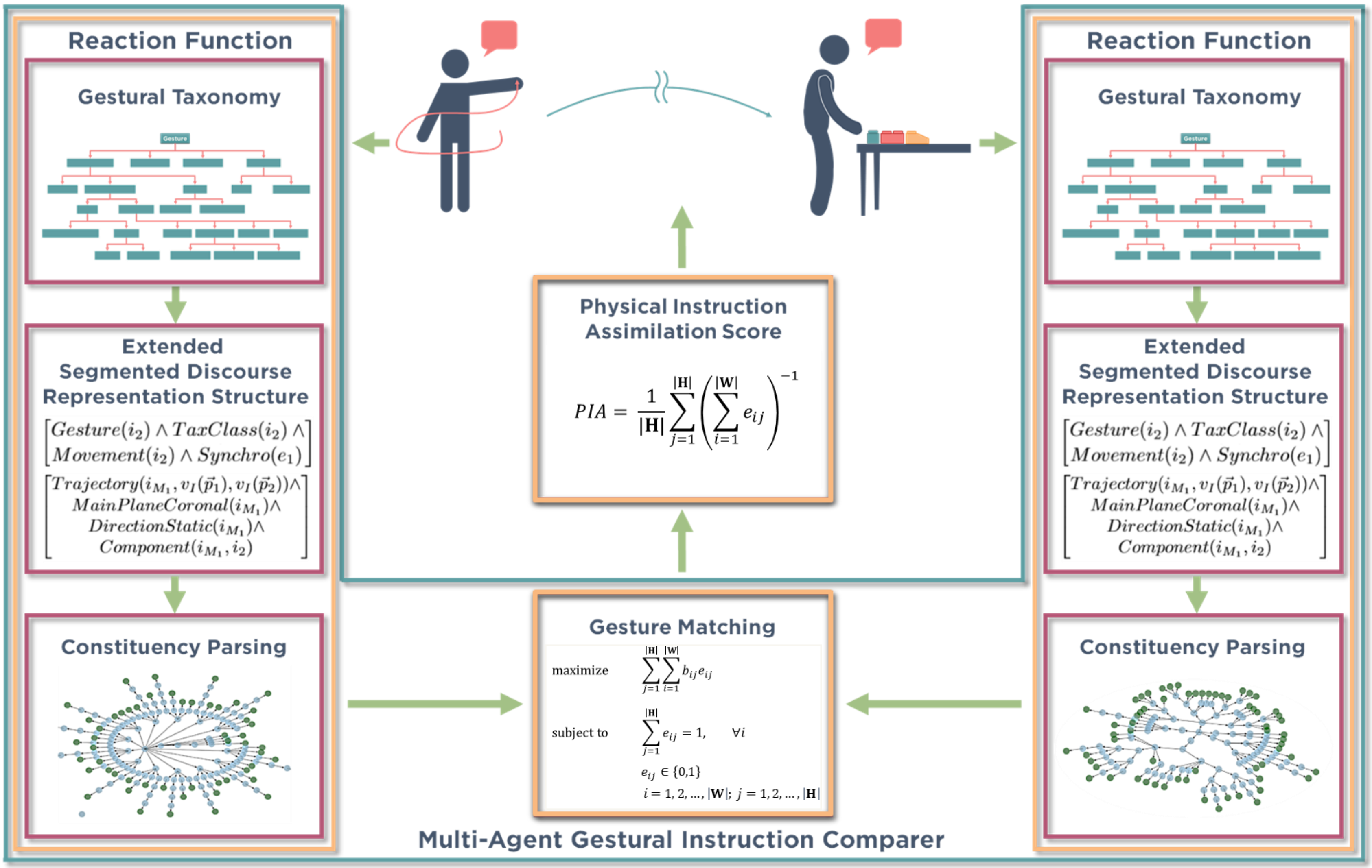
Reaction Function
Gestural Taxonomy
Gesture
Extended
Segmented Discourse
Representation Structure
$\begin{bmatrix} Gesture(i_2) \wedge TaxClass(i_2) \wedge \\ Movement(i_2) \wedge Synchro(e_1) \end{bmatrix}$
$\begin{bmatrix} Trajectory(i_{M_1}, v_I(\vec{p}_1), v_I(\vec{p}_2)) \wedge \\ MainPlaneCoronal(i_{M_1}) \wedge \\ DirectionStatic(i_{M_1}) \wedge \\ Component(i_{M_1}, i_2) \end{bmatrix}$
Constituency Parsing
Physical Instruction Assimilation Score
$PIA = \frac{1}{|\mathbf{H}|} \sum_{j=1}^{|\mathbf{H}|} \left(\sum_{i=1}^{|\mathbf{W}|} e_{ij} \right)^{-1}$
Gesture Matching
maximize $\sum_{j=1}^{|\mathbf{H}|} \sum_{i=1}^{|\mathbf{W}|} b_{ij} e_{ij}$
subject to $\sum_{j=1}^{|\mathbf{H}|} e_{ij} = 1, \quad \forall i$
$e_{ij} \in \{0,1\}$
$i = 1, 2, \ldots, |\mathbf{W}|;\ j = 1, 2, \ldots, |\mathbf{H}|$
Reaction Function
Gestural Taxonomy
Gesture
Extended
Segmented Discourse
Representation Structure
$\begin{bmatrix} Gesture(i_2) \wedge TaxClass(i_2) \wedge \\ Movement(i_2) \wedge Synchro(e_1) \end{bmatrix}$
$\begin{bmatrix} Trajectory(i_{M_1}, v_I(\vec{p}_1), v_I(\vec{p}_2)) \wedge \\ MainPlaneCoronal(i_{M_1}) \wedge \\ DirectionStatic(i_{M_1}) \wedge \\ Component(i_{M_1}, i_2) \end{bmatrix}$
Constituency Parsing
Multi-Agent Gestural Instruction Comparer

Figure 3.1 Platform architecture and interactions between its components.

MAGIC represents each $\boldsymbol{A}$ *Action* as a three-element tuple $(\boldsymbol{\pi}, \boldsymbol{\Psi}, \boldsymbol{\Omega})$. The first element of MAGIC's $\boldsymbol{A}$ *Action* tuple is an $\boldsymbol{\pi}$ *Utterance*, the smallest unit of speech or gesture that communicates a complete idea. MAGIC $\boldsymbol{\pi}$ *Utterances* can be expressed in terms of MST's forms. Sebeok and Danesi defined a $\boldsymbol{\xi}$ *form* as a human-created model capable of conveying meaning (2000). Following this theory, each $\boldsymbol{\pi}$ *Utterance* in MAGIC can be expressed as either a $\boldsymbol{\xi}_{\mathbf{I}}$ *singularized form* (i.e. a meaning representation communicating a singular concept); or a $\boldsymbol{\xi}_{\mathbf{II}}$ *composite form* (i.e. a combination of $\boldsymbol{\xi}_{\mathbf{I}}$ *singularized forms* that communicates a concept). Additionally, these forms can be either verbal ($\boldsymbol{\xi}_{\mathbf{I}}^{\boldsymbol{v}}$) or gestural ($\boldsymbol{\xi}_{\mathbf{I}}^{\boldsymbol{g}}$) forms. Examples of $\boldsymbol{\xi}_{\mathbf{I}}$ *singularized forms* as $\boldsymbol{\pi}$ *Utterances* include the command "Stop!" (as a $\boldsymbol{\xi}_{\mathbf{I}}^{\boldsymbol{v}}$ verbal *singularized form*), or Hanna's gesture to pinpoint the wrench (as a $\boldsymbol{\xi}_{\mathbf{I}}^{\boldsymbol{g}}$ gestural *singularized form*). Examples of $\boldsymbol{\xi}_{\mathbf{II}}$ *composite forms* as $\boldsymbol{\pi}$ *Utterances* include Hanna's verbal command "Take that wrench" (as a $\boldsymbol{\xi}_{\mathbf{II}}^{\boldsymbol{v}}$ verbal *composite form*), or a "Repeat" gesture in which the gesture of tracing a circle in the air is performed multiple times in a row (as a $\boldsymbol{\xi}_{\mathbf{II}}^{\boldsymbol{g}}$ gestural *composite form*). MAGIC's $\boldsymbol{\pi}$ *Utterances* resemble ToS Sign Vehicles, as both are the elements utilized to exchange information during the collaboration. In addition, let the $\boldsymbol{\mathcal{D}}$ *Discourse* be a set containing all the utterances, such that $\forall \boldsymbol{\pi}, \boldsymbol{\pi} \in \boldsymbol{\mathcal{D}}$.

The second element of MAGIC's $\boldsymbol{A}$ *Action* tuple is an $\boldsymbol{\Psi}$ *Interpretation Tree,* a data structure representing an $\boldsymbol{\pi}$ *Utterance*. $\boldsymbol{\Psi}$ *Interpretation Trees* are a simplification of ToS' Interpretants, which represent the disposition to react in a certain manner after receiving a stimulus. Morris himself elaborated: "*Such a disposition can, if one wishes, be interpreted in probabilistic terms, as the probability of reacting in a certain way under certain conditions because of the appearance of the sign.*" (1964, p. 3). For example, after Walter received Hanna's instruction requesting grabbing a wrench, he could have responded in a number of ways, e.g. grabbed a correct wrench or a different tool, or asked Hanna to repeat the instruction, among others. Each of these possible responses would be associated to a different ToS' Interpretant. The way in which MAGIC synthesizes the ToS' Interpretant concept in the $\boldsymbol{\Psi}$ *Interpretation Trees* will be explained in the following subsections of this chapter.

The last element of MAGIC's $\boldsymbol{A}$ *Action* tuple is a $\boldsymbol{\Omega}$ *Context*. ToS' Context is introduced as a container of all elements that could influence a particular individual to generate a ToS' Interpretant. Under this description, the ToS' Context of Hanna and Walter's collaboration comprehends the elements in their respective environments, their prior knowledge and preconceptions, among other factors. MAGIC's **Ω** *Context* can be viewed as a subset of the more general ToS' Context: both encompass information generated by elements in the surroundings. However, a container to all the elements capable of influencing a particular response in an individual cannot be concretely defined. Instead, this work follows an approach where only the significant surround at a moment and with respect to an angle is represented, as represented by Jakob von Uexküll's *Umwelt* (1937/2001). Following this idea, MAGIC's **Ω** *Context* contains only those elements that were introduced in previous utterances ($\boldsymbol{\pi_{t-1}}$, $\boldsymbol{\pi_{t-2}}$, ..., $\boldsymbol{\pi_{t-|\mathcal{D}|}}$). For example, consider Hanna's "Take that wrench" instruction as the first utterance ($\boldsymbol{\pi_1}$), and her next command "Use the wrench to remove the nut on the left" as the second utterance ($\boldsymbol{\pi_2}$). The **Ω** *Context* of $\boldsymbol{\pi_2}$ would include all the elements introduced in $\boldsymbol{\pi_1}$, such as the "Wrench" concept introduced in $\boldsymbol{\pi_1}$. Figure 3.2 presents a schematic of the MAGIC framework's elements: a $\mathbf{\Phi_H}$ *Helper* instructs a $\mathbf{\Phi_W}$ *Worker* on how to give maintenance to a robotic arm. The elements of an ***A*** *Action* performed by the agents are linked via the ***R*()** Reaction Function, a relation in terms of a specific **π** *Utterance* and a given **Ω** *Context*.

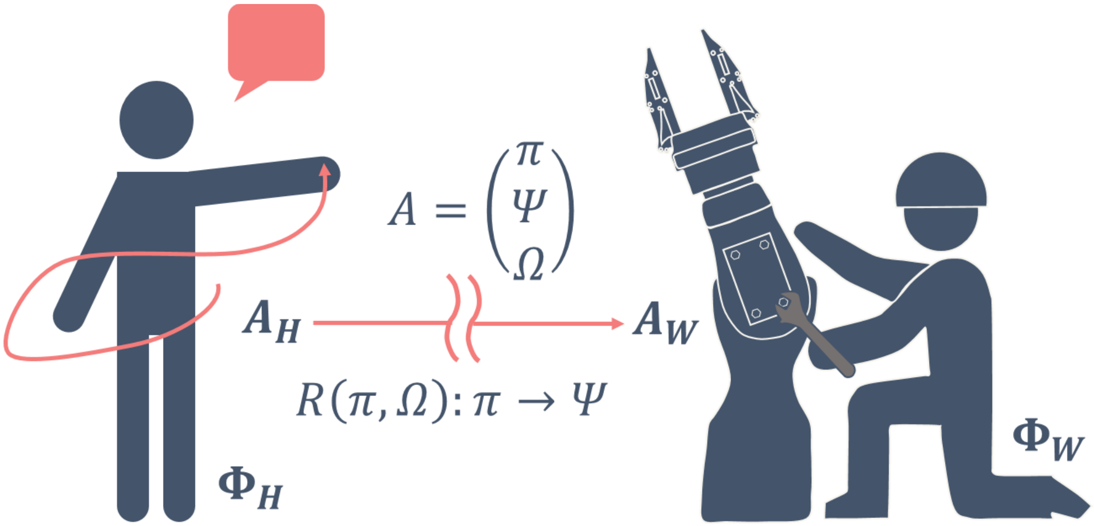
$A = \begin{pmatrix} \pi \\ \Psi \\ \Omega \end{pmatrix}$
A_H
A_W
$R(\pi, \Omega): \pi \rightarrow \Psi$
Φ_H
Φ_W

Figure 3.2 Multi-Agent Gestural Instructions Comparer framework.

3.2 The Reaction Function

MAGIC represents the gestures from the **Φ** *Agents* into an **Ψ** *Interpretation Tree*, 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 inputs and returns **Ψ** *Interpretation Trees* as outputs. The three stages of the ***R()*** *Reaction Function* include a taxonomy classification to reveal high level semantics and pragmatics uses of the gestures; 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. Figure 3.3 presents the three-stage pipeline of the ***R()*** *Reaction Function* module. The elements of this module will be explained in this subsection.

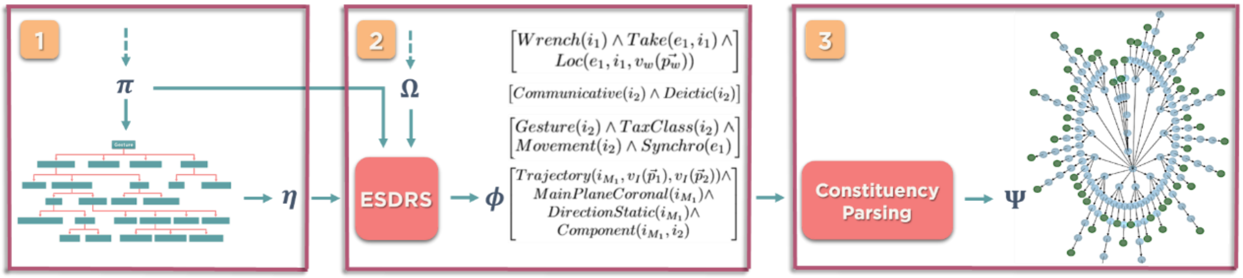
1
π
Gesture
η
2
Ω
ESDRS
ϕ
$\begin{bmatrix} Wrench(i_1) \wedge Take(e_1, i_1) \wedge \\ Loc(e_1, i_1, v_w(\vec{p_w})) \end{bmatrix}$
$[Communicative(i_2) \wedge Deictic(i_2)]$
$\begin{bmatrix} Gesture(i_2) \wedge TaxClass(i_2) \wedge \\ Movement(i_2) \wedge Synchro(e_1) \end{bmatrix}$
$\begin{bmatrix} Trajectory(i_{M_1}, v_I(\vec{p}_1), v_I(\vec{p}_2)) \wedge \\ MainPlaneCoronal(i_{M_1}) \wedge \\ DirectionStatic(i_{M_1}) \wedge \\ Component(i_{M_1}, i_2) \end{bmatrix}$
3
Constituency Parsing
Ψ

Figure 3.3 Three-stage pipeline of the ***R()*** Reaction Function.

3.2.1 Gestural Taxonomy Classification

Gestural taxonomies express classification criteria that differentiate gestures from one another (Kendon, 2004). These criteria can include the gestures' communicative intention, expressiveness, and 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*. Currently, these ***η*** *Classifications* are being used to identify gestures that either introduce redundancy to the collaboration or are involuntarily performed. Under the rationale that redundant and involuntary gestures represent a less effective information transfer (Hsia, 1977). MAGIC's gestural taxonomy combines well-known taxonomies, i.e. McNeil, Goodwin, and Poggi's taxonomies (Goodwin, 2003; McNeill, 1992; Poggi, 2008); into a single, 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 nor 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 $\mathbf{\Phi}_{\mathbf{W}}$ *Worker* or the $\mathbf{\Phi}_{\mathbf{H}}$ *Helper*?).

Figure 3.4 presents MAGIC's gestural taxonomy. An in-depth explanation of each our gestural taxonomy nodes can be found in the referenced literature, as well in Adam Kendon's book (2004, p. 84). As an example, Hanna's pointing gesture would be assigned to the "*Communicative*" and "*Deictic*"' $\boldsymbol{\eta}$ *Classifications*. These labels reveal that her 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).

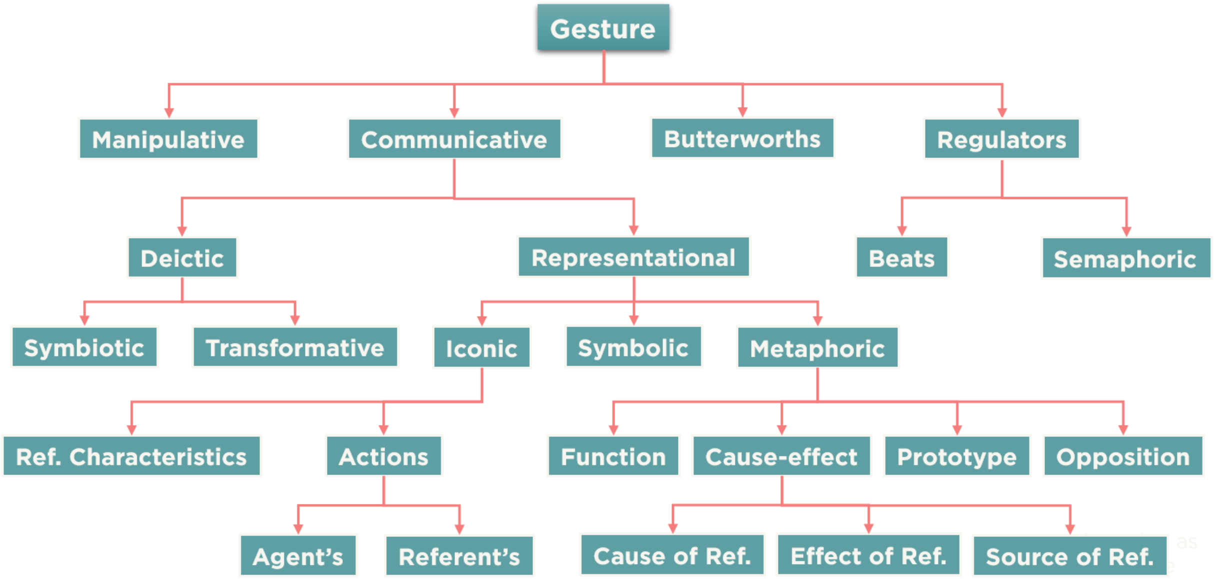

Figure 3.4 MAGIC's gestural taxonomy.

3.2.2 Extended Segmented Discourse Representation Structure

The second stage of the ***R*()** *Reaction Function* module represents 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 (Asher & Lascarides, 2003). 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 a follow-up work, Lascarides and Stone expanded SDRS by integrating attribute-pair tables that describe high-level morphological features of the gestures

(2009), following Kopp, Tepper, and Cassell's Typed Feature Structures (2004). Following Kopp, Tepper, and Cassell's approach, Hanna’s pointing gesture would be expressed with the following table:

$$\begin{bmatrix} \textbf{pointing\_gesture} & \\ \text{right\_hand\_shape:} & asl-1 \\ \text{right\_finger\_direction:} & forward \\ \text{right\_palm\_direction:} & forward \\ \text{right\_location:} & \vec{f} \end{bmatrix},$$

where $asl-1$ represent the hand shape of the number 1 in American Sign Language (U.S. Department of Health and Human Services, 2019), $forward$ represents the direction of the motion (straight away from the body of the person), and $\vec{f}$ 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 pre-defined lexicons (e.g. American Sign Language), making it non-modular to the variety of gestures humans can generate. Therefore, we propose an extension of the SDRS framework (Extended SDRS, henceforth ESDRS) that: (1) integrates the affixed SDRS attribute-pair table as part of $\boldsymbol{\phi}$ *ESDRS-formulae*; and (2) defines a standard set of components to describe gestures' morphology. These $\boldsymbol{\phi}$ *ESDRS-formulae* can represent both verbal and gestural utterances (denoted by the $[\mathcal{G}]$ operator). Therefore, the second stage of $\boldsymbol{R}()$ Reaction Function module involves generating $\boldsymbol{\phi}$ *ESDRS-formulae* after receiving an $\boldsymbol{\pi}$ *Utterance*, a $\boldsymbol{\Omega}$ *Context*, and the $\boldsymbol{\eta}$ *Classification* from the prior stage as inputs.

ESDRS describes the meaning of an $\boldsymbol{\pi}$ *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 four main atomic components: discourse referents, spatiotemporal localities, virtual mappings, and predicates. 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, the wrench). An $\boldsymbol{e}$ *eventuality variable* is a temporal event in the discourse (e.g. connecting pieces, grabbing the wrench). These discourse referents are introduced in the $\boldsymbol{\phi}$ *ESDRS-formulae* via the $\exists$ operator, and represent the elements of the $\boldsymbol{\mathcal{D}}$ *Discourse* that will be added to the $\boldsymbol{\Omega}$ *Context* of the following $\boldsymbol{\pi}$ *Utterances*. A $\boldsymbol{p}$ *spatiotemporal locality* represents the position in time of a specific individual variable. Each $\boldsymbol{p}$ *spatiotemporal locality* is a 4-dimensional vector $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{t})$, where $\boldsymbol{x}$, $\boldsymbol{y}$, and $\boldsymbol{z}$ represents the position in space of an $\boldsymbol{i}$ *individual variable* and $\boldsymbol{t}$ represents a moment in time. A

$\boldsymbol{v}$ *virtual mapping* represents a transformation over a $\boldsymbol{p}$ *spatiotemporal locality* mapping a point from world space into a point in gesture space. These $\boldsymbol{v}$ *virtual mappings* are required when the absolute position of an element in world space is represented by a relative position gesture space. For example, Hanna's pointing gesture exemplifies a virtual mapping: the world position of the wrench in Walter's workspace (represented as by a spatiotemporal location) is mapped to Hanna's gesture space (her finger and the direction in which she is pointing).

Finally, predicates are clauses describing the interactions between the atomic components. Predicates can be seen as tests over the $\boldsymbol{\mathcal{M}}$ *model* components: the components will move from $\boldsymbol{\Omega_i}$ into $\boldsymbol{\Omega_o}$ 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 (Stokke, 2014). For example, the $\boldsymbol{A}_{\mathbf{W}}$ Execution in which Walter grabs the wrench can be expressed by the predicate *Take*($\mathbf{e_1}$, $\mathbf{i_1}$, $\mathbf{i_2}$), where $\mathbf{i_1}$, $\mathbf{i_2}$, and $\mathbf{e_1}$ are discourse referents that respectively represent the wrench, Walter's hand, and the event of grabbing the wrench. An explanation of the elements of the $\boldsymbol{\phi}$ *ESDRS-formulae* is presented in (Asher & Lascarides, 2003). All the predicates included in SDRS (e.g. *Loc*(), *Exemplifies*(), *Component*()) are also included in ESDRS.

Consider the following verbal and gestural utterances:

π_1: "Take that wrench."

π_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 $\boldsymbol{\phi}$ ESDRS-formula is given by:

$$\pi_1 : \exists i_1 \begin{bmatrix} Wrench(i_1) \wedge Take(e_1, i_1) \wedge \\ Loc\big(e_1, i_1, v_w(\overrightarrow{p_w})\big) \end{bmatrix},$$

where $\boldsymbol{i_1}$ are individual variables (introduced into the discourse via the ∃ operator), $\boldsymbol{e_1}$ is an eventuality variable, and $\boldsymbol{v_w}$ is a *virtual mapping* over the $\overrightarrow{\boldsymbol{p_w}}$ *spatiotemporal locality*. This $\boldsymbol{\phi}$ *ESDRS-formula* includes the predicates *Wrench*(), *Take*(), and *Loc*(), which are conditioning the $\boldsymbol{\mathcal{M}}$ *model*'s components and therefore updating the discourse context.

To generate gestural $\boldsymbol{\phi}$ *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 $\boldsymbol{\eta}$ *Classifications* will be translated into ESDRS-formulae predicates. Additionally, ESDRS translates Kopp, Tepper, and Cassell's Typed Feature Structures 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 defined by Madapana and Wachs (2017). Finally, EDSRS introduces the *Synchro*() predicate to describe whether the gesture was performed in synchronicity with a specific event in the discourse (i.e. an eventuality variable). For example, if $\boldsymbol{e_1}$ represents the event of grabbing a piece, *Synchro*($\boldsymbol{e_1}$) represents that the gesture was performed during $\boldsymbol{e_1}$.

Consequently, the $\boldsymbol{\phi}$ *ESDRS-formulae* of corresponding to these utterances are presented. $\boldsymbol{\pi_{2T}}$, $\boldsymbol{\pi_{2S}}$, and $\boldsymbol{\pi_{2M}}$ are elements within $\boldsymbol{\pi_2}$, but are presented separate for ease of reading:

$$\pi_2 \colon \exists i_2 \begin{bmatrix} Gesture(i_2) \land TaxClass(i_2) \land Shape(i_2) \land \\ Movement(i_2) \land Synchro(e_1) \end{bmatrix}$$

$$\pi_{2T} \colon \exists i_2 [Communicative(i_2) \land Deictic(i_2)]$$

$$\pi_{2M} \colon \exists i_{M_1} \begin{bmatrix} Trajectory\left(i_{M_1}, v_{\mathrm{I}}(\overrightarrow{p_1}), v_{\mathrm{I}}(\overrightarrow{p_2})\right) \land \\ MainPlaneCoronal\left(i_{M_1}\right) \land \\ DirectionStatic\left(i_{M_1}\right) \land \\ Component(i_{M_1}, i_2) \end{bmatrix}$$

$$\pi_{2S}\colon \exists \begin{array}{c} i_{S_1}, i_{S_2}, \\ i_{S_3}, i_{S_4}, \\ i_{S_5}, i_{S_6}, i_{S_7} \end{array} \left[\begin{array}{c} Arm(i_{S_1}) \wedge Hand(i_{S_2}) \wedge ThumbFinger(i_{S_3}) \wedge \\ RingFinger(i_{S_4}) \wedge MiddleFinger(i_{S_5}) \wedge \\ IndexFinger(i_{S_6}) \wedge LittleFinger(i_{S_7}) \wedge \\ PoseExtended(i_{S_1}) \wedge OrientationForward(i_{S_1}) \wedge \\ PoseNotExtended(i_{S_2}) \wedge OrientationForward(i_{S_2}) \wedge \\ PoseNotExtended(i_{S_3}) \wedge PoseNotExtended(i_{S_4}) \wedge \\ PoseNotExtended(i_{S_5}) \wedge PoseExtended(i_{S_6}) \wedge \\ OrientationForward(i_{S_6}) \wedge PoseNotExtended(i_{S_7}) \wedge \\ Component(i_{S_1}, i_2) \wedge Component(i_{S_2}, i_{S_1}) \wedge \\ Component(i_{S_3}, i_{S_2}) \wedge Component(i_{S_4}, i_{S_2}) \wedge \\ Component(i_{S_5}, i_{S_2}) \wedge Component(i_{S_6}, i_{S_2}) \wedge \\ Component(i_{S_7}, i_{S_2}) \end{array}\right]$$

3.2.3 ESDRS Constituency Parsing

The last stage of the ***R()*** *Reaction Function* module represents the ***ϕ*** *ESDRS-formulae* as data structure representing the gestures' morphology, semantics, and pragmatics. ***ϕ*** *ESDRS-formulae* are not designed for comparisons, and an additional data structure to represent these formulae is needed. These representations are known as **Ψ** *Interpretation Trees*, and they are used to obtain gesture similarity insights. MAGIC leverages a constituency parsing approach to generate these tree structures, which capture both the value of the ***ϕ*** *ESDRS-formulae* elements and the relationship between them. Figure 3.5 presents an example of an **Ψ** *Interpretation Tree*. The components of this tree will be explained in detailed in this subsection.

MAGIC’s constituency parsing approach introduces several nested constituents to represent the different elements of the ***ϕ*** *ESDRS-formulae*. These constituents can be separated into two main subgroups: predicative and non-predicative. As the name implies, the non-predicative subgroup includes all the elements from the ***ϕ*** *ESDRS-formulae* that are not predicates: discourse referents, spatiotemporal localities, among others. Alternatively, the predicative subgroup includes the different predicates from the formulae, grouping them accordingly to how they are used to describe similar aspects of the gesture. For example, predicates related to the gesture's shape will be grouped together, and will be separate from predicates related to the gesture's movement.

The first constituent of the non-predicative subgroup is the Variable Group (VG, black-circled in Figure 3.5), which includes all the discourse referents from the current $\boldsymbol{\phi}$ *ESDRS-formula* (e.g a nut Walter needs to remove). The second constituent is the Spatiotemporal Group (SG, black-circled in Figure 3.5), which contains all spatiotemporal localities from the current $\boldsymbol{\phi}$ *ESDRS-formula* (e.g. the position of the nut Walter needs to remove). The third constituent is the Mapping Group (MG, black-circled in Figure 3.5), which contains all the virtual mappings from the current $\boldsymbol{\phi}$ *ESDRS-formula* (e.g. Hanna's performed a pointing gesture to indicate the position of the nut. A virtual mapping that transforms the location of the nut in Walter's workspace to Hanna's gesture space will be created). Finally, the last constituent of the non-predicative subgroup is the Context Group (CG, black-circled in Figure 3.5), which contains a reference to all the discourse referents and predicates that were introduced in the previous utterances that are being utilized present in the current utterance (e.g. Hanna's instructed Walter to remove the nut with the wrench. The wrench Hanna is referring to was introduced in a previous $\boldsymbol{\pi}$ *Utterance*, making it accessible as a member of the current $\boldsymbol{\pi}$ *Utterance*'s $\boldsymbol{\Omega}$ *Context*).

The predicate subgroup, represented in the $\boldsymbol{\Psi}$ *Interpretation Trees* as the Large Predicate Group (LPG, black-circled in Figure 3.5) contains all the predicates from the current $\boldsymbol{\phi}$ *ESDRS-formulae*. Moreover, each LPG can be subdivided into seven different constituents: each grouping predicates dealing with different characteristics of the gestures. The Shape Group (ShG, Region 1 in Figure 3.5) includes all the discourse referents and predicates related to the gesture's shape. The Loc Group (LoG, Region 2 in Figure 3.5) includes all the *Loc*() predicates, related to localizing individual variables. For example if $\boldsymbol{i_1}$ and $\boldsymbol{e_1}$ are discourse referents that respectively represent the wrench and the event of grabbing the wrench, then the predicate *Loc*($\boldsymbol{e_1}$, $\boldsymbol{i_1}$, $\boldsymbol{v_1}(\overrightarrow{\boldsymbol{p_1}})$) implies that the wrench was physically and temporally located at $\overrightarrow{\boldsymbol{p_1}}$ during the time spanned by entire $\boldsymbol{e_1}$. The Exemplifies Group (ExG, Region 3 in Figure 3.5) includes all the *Exemplifies*() predicates, related to depicting a specific individual variable with a gesture. For example, if Hanna performs a fist gesture to convey the idea of the nut, then we can respectively represent $\boldsymbol{i_1}$ and $\boldsymbol{i_2}$ as the nut and Hanna's hand and the predicate *Exemplifies*($\boldsymbol{i_2}$, $\boldsymbol{i_1}$) to represent the meaning of Hanna's gestural utterance.

The TaxClass Group (TaG, Region 4 in Figure 3.5) includes all the predicates related to the gesture's taxonomy classification. For example, a member of the TaG is the predicate *Deictic*($\boldsymbol{i_1}$), which represents that $\boldsymbol{i_1}$ is a gesture classified as deictic. The Synchro Group (SyG, Region 5 in Figure 3.5) includes all the Synchro predicates, related to whether a gesture was performed in synchronicity with a given discourse referent. For example, if $\boldsymbol{e_1}$ represents the event of removing the nut, then *Synchro*($\boldsymbol{e_1}$) implies that the current gestural utterance was performed in synchronicity with the discourse referent $\boldsymbol{e_1}$. The Movement Group (MvG, Region 7 in Figure 3.5): contains the discourse referents and predicates related to the gesture's movement. Finally, all the other predicates that do not fit into any of the previous categories will be grouped together (Extra Predicates, Region 6 in Figure 3.5).

Using $\boldsymbol{\Psi}$ *Interpretation Trees* to represent gestures introduce several advantages. The most noticeable advantage is that key elements representing the gestures are encompassed into a single structure. For instance, the $\boldsymbol{\Psi}$ *Interpretation Trees* include information regarding to the time in which the gestures were performed with respect to verbal instructions, the semantic content encompassed in the gesture, iconicity, expressiveness, movement, among others. Therefore, this structure allows for computational comparisons to be performed that can simultaneously inspect a gesture’s morphology, semantics, and pragmatics. Another advantage of the $\boldsymbol{\Psi}$ *Interpretation Trees* comes from the arrangement of the structure into subtrees. As mentioned, predicates related to specific aspects of the gesture are group together into specific groups or subtrees. This allows for more specific comparisons to be performed. For instance, if we only want to compare gestures based in their physical execution, then the subtrees representing shape and movement could be inspected in isolation instead of the entire $\boldsymbol{\Psi}$ *Interpretation Tree*. Moreover, the structure is robust to aspects such as time misalignments when annotating the gestures. Specifically, the predicates encompassed in the Synchro Group represent whether there is relation between the time a gesture was authored and the meaning it conveys. Therefore, time incongruencies that might introduce noise during the gesture comparison stage can be addressed by consulting the information in the Synchro Group. Finally, $\boldsymbol{\Psi}$ *Interpretation Trees* are easily scalable: other subtrees could be added to consider information currently not encompassed in the trees, such as gaze, face gestures, body posture, among others.

This parsing approach is the last part of the ***R()*** *Reaction Function* module: gestures are represented into **Ψ** *Interpretation Trees* that encapsulate morphology, semantics and pragmatics. Comparisons between **Ψ** *Interpretation Trees* and their subcomponents will be performed to calculate a matching between the gestures of the agents collaborating. Figure 3.6 provides examples of how different gestures are represented with different **Ψ** *Interpretation Trees*. Specifically, the figure showcases the Exemplifies Group constituent from the trees.

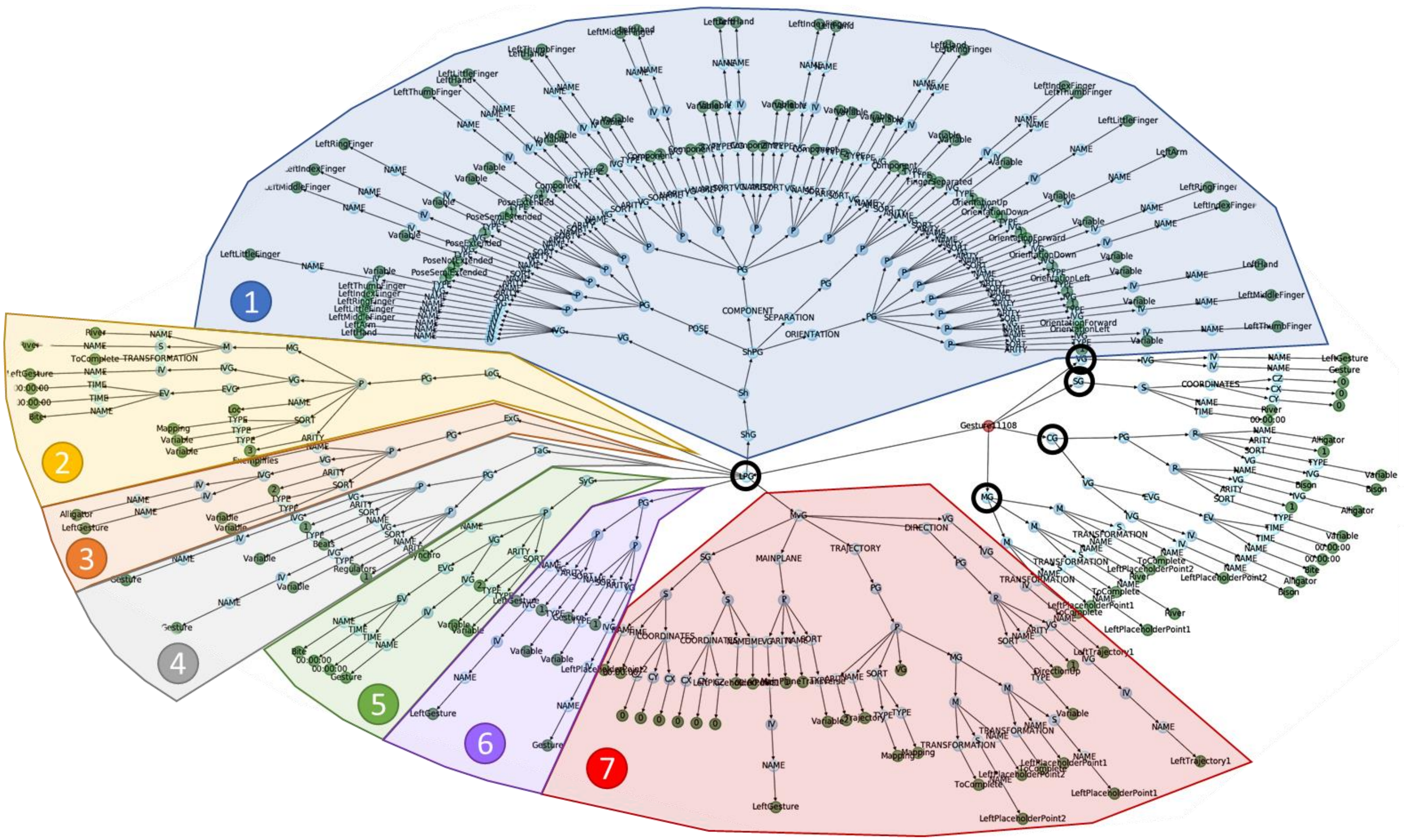

Figure 3.5 Example of MAGIC's Interpretation Tree. The black-circled nodes indicate the five main constituents of an Interpretation Tree. The numbered regions indicate the main nested constituents of tree’s predicative group.

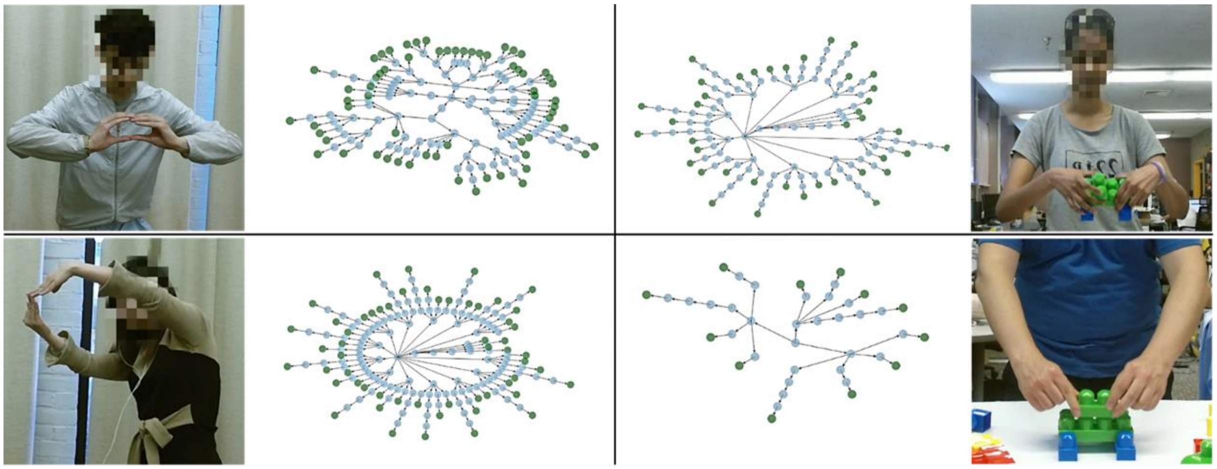

Figure 3.6 Comparison between the Interpretation Trees generated from the gestures performed by different agents.

3.3 Gesture Matching through Integer Optimization Problems

Sebanz and colleagues studied how perception and action are linked in social interactions (Knoblich & Sebanz, 2006; Sebanz et al., 2003). Their work reviewed studies on how mapping the actions performed by others to self-performed actions can enable action understanding and identification. Specifically, their work gave light on how effective collaboration can be enabled by situations in which the self-performed actions are functionally equivalent to the actions performed by others (2006, p. 101). Following the link between functionally equivalent actions and effective collaboration, the last stage of the MAGIC architecture performs a gesture matching process to identify functionally equivalent gestures. Following our vignette, this process analyzes all the gestures generated by Hanna and Walter, and computes which of all the gestures performed by Walter can be associated with particular gestures generated by Hanna.

Using the aforementioned approach, the search is conducted in such a way that the matching is determined based on an optimality criterion among the collaborators' gestures. Let an $\boldsymbol{e_{ij}}$ edge weight the representation of a matching between two gestures. Each $\boldsymbol{e_{ij}}$ edge weight will take a value of 1 if the $\boldsymbol{h_j}$ Helper-authored gesture matches (i.e. is the most functionally equivalent) to the $\boldsymbol{w_i}$ Worker-authored, and 0 otherwise. Our approach represents each gesture matching solution with a $\mathbf{E}$ matrix of $\boldsymbol{e_{ij}}$ edge weights of size $|\mathbf{W}| \times |\mathbf{H}|$. The goal of our approach is to solve the integer optimization assignment problems to find the $\mathbf{E}$ matrices that describes the matching of the gestures performed by the collaborators such that the overall cost of the matching is optimized. We define three different integer optimization approaches to obtain these $\mathbf{E}$ matrices: MAGIC-based optimization, Time-based optimization, and Hybrid optimization. These approaches are detailed in our previous work (Rojas-Muñoz & Wachs, 2020).

3.3.1 MAGIC-based Optimization

The first optimization approach leverages similarity scores based on the $\boldsymbol{\Psi}$ *Interpretation Trees* obtained from the $\boldsymbol{R()}$ *Reaction Function* module. Let $\boldsymbol{\mathcal{X}}$ be a nested constituent from an $\boldsymbol{\Psi}$ *Interpretation Tree* (e.g. CG, LPG, SyG). Then, $\boldsymbol{\Psi}^{\boldsymbol{\mathcal{X}}}$ is the subtree of $\boldsymbol{\Psi}$ that has $\boldsymbol{\mathcal{X}}$ as its root (e.g. $\boldsymbol{\Psi}^{\boldsymbol{CG}}$ represents a context subtree). These subtrees can also be combined, in the form $\boldsymbol{\Psi}^{\boldsymbol{\mathcal{X}_1}} \cup \boldsymbol{\Psi}^{\boldsymbol{\mathcal{X}_2}}$. Our approach consists in comparing the $\boldsymbol{\Psi}_{\mathbf{W}}$ Worker *Interpretation Trees* and their respective subtrees against the $\boldsymbol{\Psi}_{\mathbf{H}}$ Helper *Interpretation Trees* and their respective subtrees. Constructing these subtress with main constituents of the $\boldsymbol{\Psi}$ *Interpretation Trees* as their root (e.g. LGP, ShG, CG) allows to compare specific aspects of the gestures between each other (e.g. shape, movement, context). For these comparisons, the entire $\boldsymbol{\Psi}$ *Interpretation Tree* was considered as another subtree.

The first part in this approach is to generate a $\mathbf{B}$ coefficient matrix of distance costs (of size $|\mathbf{W}| \times |\mathbf{H}|$) that summarizes how similar is each $\boldsymbol{\Psi}_{\mathbf{W}}$ Worker *Interpretation Tree* to each $\boldsymbol{\Psi}_{\mathbf{H}}$ Helper *Interpretation Tree.* Each $\boldsymbol{b_{ij}}$ coefficient in this matrix summarizes the similarity between the $\boldsymbol{\Psi}$ *Interpretation Trees* representing the $\boldsymbol{w_i}$ Worker-authored gesture ($\boldsymbol{\Psi}_{\mathbf{w_i}}$) and the one representing the $\boldsymbol{h_j}$ Helper-authored gesture ($\boldsymbol{\Psi}_{\mathbf{h_j}}$). We use two approaches to generate these $\mathbf{B}$ matrices: intersection between $\boldsymbol{\Psi}$ *Interpretation Trees*, and two variations of a string tree kernel.

Interpretation Tree matching via subtree intersections

The first approach to generate the $\mathbf{B}$ coefficient matrices of distance costs involves intersecting the $\boldsymbol{\Psi}_{\mathbf{h_j}}$ *Helper Interpretation Tree* with the $\boldsymbol{\Psi}_{\mathbf{w_i}}$ *Worker Interpretation Tree*. Each of the $\boldsymbol{b_{ij}}$ distance cost coefficients is generated by computing the number of nodes in the graph generated after intersecting the $\boldsymbol{\Psi}$ *Interpretation Trees* or subtrees representing different gestures. This formulation is depicted in Equation (3-1):

$$\boldsymbol{b_{ij}} = \left(num\_nodes\ \boldsymbol{\Psi}_{\mathbf{H_j}}^{\mathcal{X}_1} \cap \boldsymbol{\Psi}_{\mathrm{W}_i}^{\mathcal{X}_2}\right);\ \begin{array}{l} i = 1,2,\dots,|\mathbf{W}| \\ j = 1,2,\dots,|\mathbf{H}| \end{array} \qquad (3\text{-}1)$$

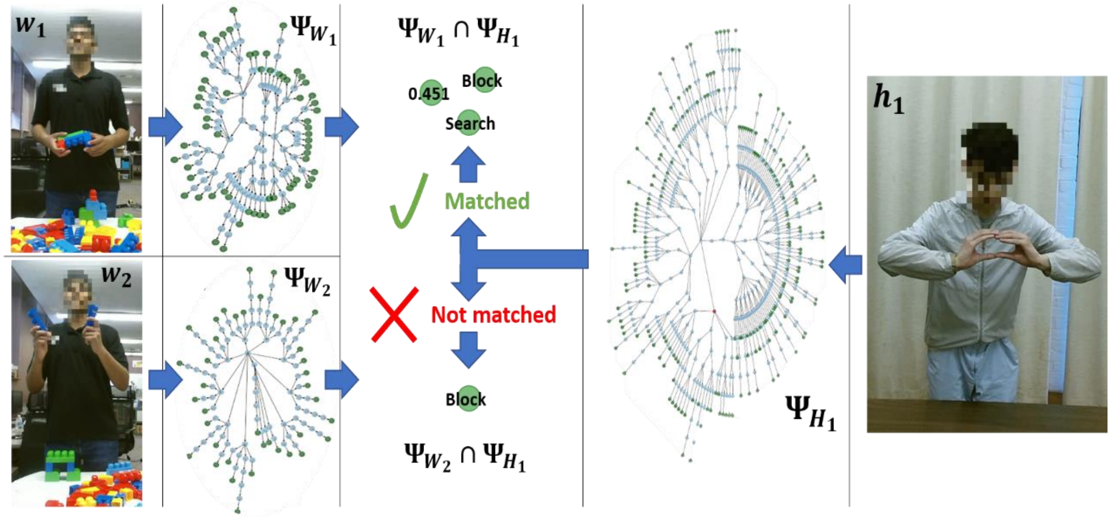

Figure 3.7 Example of subtree intersection similarity. Similar subtrees will have a higher number of common nodes.

Figure 3.7 presents a visual example of the intersection between two sets of $\mathbf{\Psi}$ *Interpretation Trees* leaves. In the figure, the gesture performed by the $\mathbf{\Phi_H}$ *Helper* ($\boldsymbol{h_1}$) needs to be matched to the most functionally equivalent gesture from those performed by the $\mathbf{\Phi_W}$ *Worker* ($\boldsymbol{w_1}$ or $\boldsymbol{w_2}$). These gestures will be respectively represented by the $\mathbf{\Psi_{H_1}}$, $\mathbf{\Psi_{W_1}}$ and $\mathbf{\Psi_{W_2}}$ *Interpretation Trees*. The $\mathbf{\Psi_{H_1}}$ Helper *Interpretation Tree* will be intersected with both the $\mathbf{\Psi_{W_1}}$ and $\mathbf{\Psi_{W_2}}$ Worker *Interpretation Trees* ($\mathbf{\Psi_{W_1}} \cap \mathbf{\Psi_{H_1}}$ and $\mathbf{\Psi_{W_2}} \cap \mathbf{\Psi_{H_1}}$, respectively). By acquiring the number of nodes in these intersections, the most functionally equivalent gesture to $\boldsymbol{h_1}$ can be identified. In this example, the $\mathbf{\Psi_{W_1}} \cap \mathbf{\Psi_{H_1}}$ intersection had three common nodes (with values “0,451”, “Search”, and “Block”). On the other hand, the $\mathbf{\Psi_{W_2}} \cap \mathbf{\Psi_{H_1}}$ intersection had only one common node (with value “Block”). Therefore, the $\boldsymbol{w_1}$ Worker-authored gesture will be selected as the most functionally equivalent gesture to the $\boldsymbol{h_1}$ Helper-authored gesture.

Interpretation Tree matching via tree kernels

The previous approach has the limitation of not considering aspects of the $\mathbf{\Psi}$ *Interpretation Trees* topology such as the paths connecting the nodes trees and the relation between the parent and child nodes. This is necessary to perform comprehensive comparisons between structured data, such as trees. To address this, the second approach to generate the $\mathbf{B}$ coefficient matrices of distance costs involves representing the $\mathbf{\Psi}$ *Interpretation Trees* as $\mathbf{S_\Psi}$ strings and applying a string tree kernel to compare them. Two approaches are followed to create these string representations, depicted in Figure 3.8:

- *Interpretation Trees* as tag strings: This approach represents the relation between parents and child nodes in a tree. A recursive approach is used to traverse the trees and store the name attributes of the nodes in a single string of characters. The pseudocode of the approach is presented in Algorithm 3.1.
- *Interpretation Trees* as subpath strings: This approach represents the expanded set of subpaths connecting the nodes of the tree. First, all the simple paths connecting the tree's root node and the leaves are obtained using the approach presented in Algorithm 3.2. Afterwards, these paths are expanded to include the subpaths within each of them using the approach presented in Algorithm 3.3.

Algorithm 3.1 Recursive representation of $\boldsymbol{\Psi}$ Interpretation Trees as a tag strings.

```
Input: $\boldsymbol{n}$ --- Current parent node. Initially, the root of the $\boldsymbol{\Psi^{\mathcal{X}}}$ Interpretation Tree.
Output: $\mathbf{S}_{\boldsymbol{\Psi}}$ --- String representing of the $\boldsymbol{\Psi^{\mathcal{X}}}$ Interpretation Tree using tags.

children ← Children of $\boldsymbol{n}$ node          # Obtain the list of children of the node
name ← label of the $\boldsymbol{n}$ node              # Obtain the node's name attribute
if len(children) == 0:
    return [name]                   # $\boldsymbol{S}_{\boldsymbol{\Psi}}$ will get assembled via the recursion calls
end if
else:
    tag ← [name
    for child in children:
        tag ← tag + recursiveCall(child)          # Call function for each child
    end for
    return tag]
end else
```

Algorithm 3.2 Recursive approach to obtain all paths between root and leaves in an $\boldsymbol{\Psi}$ Interpretation Tree.

Input: $\boldsymbol{n}$ **---** Current parent node. Initially, the root of the $\boldsymbol{\Psi}^{\boldsymbol{\mathcal{X}}}$ Interpretation Tree.
$\boldsymbol{path}$ **---** Incremental path from parent node. Initialized as an empty array.
Output: ***AllPaths*** **---** Array containing all paths between root and leaves.

```
global AllPaths
$\boldsymbol{children}$ ← Children of $\boldsymbol{n}$ node        # Obtain the list of children of the node
$\boldsymbol{name}$ ← label of the $\boldsymbol{n}$ node              # Obtain the node's name attribute
if len($\boldsymbol{children}$) == 0:
    $\boldsymbol{path}$.append($\boldsymbol{name}$)                 # Append node's name to current path
    AllPaths.append($\boldsymbol{path}$)             # Append current path to global list of paths
    return 0
end if
else:
    for child in $\boldsymbol{children}$:
        $\boldsymbol{path}$.append($\boldsymbol{name}$)             # Append node's name to current path
        recursiveCall(child, $\boldsymbol{path}$)       # Call function for each child, with
                                                    updated path
    end for
end else
```

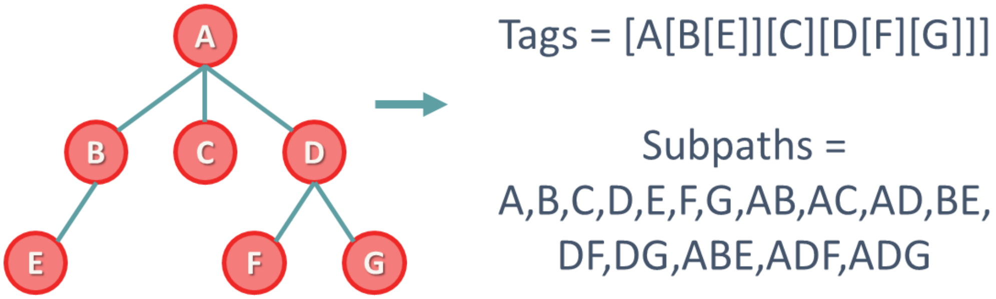

Figure 3.8 Example of subtree intersection similarity. Similar subtrees will have a higher number of common nodes.

Algorithm 3.3 Approach to expand tree paths into subpaths.

Input: ***AllPaths* ---** Current parent node. Initially, the root of the $\boldsymbol{\Psi}^{\mathcal{X}}$ Interpretation
Output: $\mathbf{S}_{\boldsymbol{\Psi}}$ **---** String representing of the $\boldsymbol{\Psi}^{\mathcal{X}}$ Interpretation Tree using subapths.

```
expandedPaths ← Empty array
for path in AllPaths:                                # Iterate through all the found paths
    depthOfPath ← 1                          # Number of elements each subpath will have
    while depthOfPath <= len(path):
        initialPosition ← 0                          # Initial index to slice current path
        finalPosition ← depthOfPath                   # Final index to slice current path
        while initialPosition <= len(path) - depthOfPath:
            expandedPaths.append(path[initialPosition:finalPosition])    # Slice
            initialPosition ← initialPosition + 1
            finalPosition ← finalPosition + 1
        end while
       depthOfPath ← depthOfPath + 1
    end while
end for
uniquePaths ← removeRepeatedPaths(expandedPaths) # Remove repeated paths
$S_{\Psi}$ ← string(uniquePaths)                 # Convert array of expanded paths to string
```

After representing the $\boldsymbol{\Psi}$ *Interpretation Trees* with $\mathbf{S}_{\boldsymbol{\Psi}}$ strings, a string tree kernel can be applied to obtain a measure of similarity between the trees. These scores, calculated as depicted in Equation (3-2), will be used as the $\boldsymbol{b_{ij}}$ distance cost coefficients that describe the similarity between the $\boldsymbol{\Psi}_{\mathbf{h_j}}$ *Helper Interpretation Tree* with the $\boldsymbol{\Psi}_{\mathbf{w_i}}$ *Worker Interpretation Tree*.

$$\sum_{k=0}^{|F|} num_{s_k}\left(\mathrm{S}_{\Psi_\mathrm{i}}\right) num_{s_k}\left(\mathrm{S}_{\Psi_\mathrm{j}}\right) w_{|s_k|}\,, \forall s_k \in F;\ \begin{array}{l} i = 1, 2, \dots, |\mathbf{W}| \\ j = 1, 2, \dots, |\mathbf{H}| \\ k = 1, 2, \dots, |F| \end{array} \qquad (3\text{-}2)$$

where $\mathbf{S}_{\boldsymbol{\Psi}_1}$ and $\mathbf{S}_{\boldsymbol{\Psi}_2}$ are the string representations of two **Ψ** *Interpretation Trees* ($\boldsymbol{\Psi}_\mathbf{1}$ and $\boldsymbol{\Psi}_\mathbf{2}$, respectively), F is a set containing all common substrings between $\mathbf{S}_{\boldsymbol{\Psi}_1}$ and $\mathbf{S}_{\boldsymbol{\Psi}_2}$, and s_k is the common substring in the set F at index k. Algorithm 3.4 showcases the pseudocode of how to find all common substrings between $\mathbf{S}_{\boldsymbol{\Psi}_1}$ and $\mathbf{S}_{\boldsymbol{\Psi}_2}$. Additionally, $num_{s_k}(\mathrm{S}_{\Psi_1})$ is a function that counts the number of appearances of the s_k substring in the S_{Ψ_1} string, and $w_{|s_k|}$ is a weight parameter that determines the importance of the s_k substring. In our case, $w_{|s_k|}$ is calculated as the ratio of the length of s_k and the maximum length between $\mathbf{S}_{\boldsymbol{\Psi}_1}$ or $\mathbf{S}_{\boldsymbol{\Psi}_2}$, as depicted in Equation (3-3):

$$w_{|s_k|} = \left(\frac{|s_k|}{\max\left(\left|\mathbf{S}_{\boldsymbol{\Psi}_1}\right|, \left|\mathbf{S}_{\boldsymbol{\Psi}_2}\right|\right)}\right) \tag{3-3}$$

After obtaining the **B** coefficient matrix of distance costs, the **E** matrix (that describes the optimal matching between the gestures performed by the collaborators) is found by solving the integer optimization problem depicted in Equation (3-4):

$$\begin{array}{ll} \text{maximize} & \displaystyle\sum_{j=1}^{|\mathbf{H}|}\sum_{i=1}^{|\mathbf{W}|} b_{ij}e_{ij} \\ \text{subject to} & \displaystyle\sum_{j=1}^{|\mathbf{H}|} e_{ij} = 1, \ \ \forall i \\ & e_{ij} \in \{0,1\} \\ & i = 1, 2, \dots, |\mathbf{W}|; \ j = 1, 2, \dots, |\mathbf{H}| \end{array} \tag{3-4}$$

The cost function is maximized whenever each $\boldsymbol{w_i}$ Worker-authored gesture is matched to the $\boldsymbol{h_j}$ Helper-authored gesture with the highest number of common nodes (based on the subtrees being compared). This formulation is constrained to each $\boldsymbol{w_i}$ Worker-authored gesture only being matched to one new $\boldsymbol{h_j}$ Helper-authored gesture. This constrain is explained in subsection 3.4, which that details the formulation of the PIA metric.

Algorithm 3.4 Approach to find all common substrings in two strings.

Input: $\mathbf{S}_{\mathbf{\Psi}_1}$ --- String representing the $\mathbf{\Psi_1}$ Interpretation Tree
$\mathbf{S}_{\mathbf{\Psi}_2}$ --- String representing the $\mathbf{\Psi_2}$ Interpretation Tree
Output: ***Substrings* ---** Array containing all common substrings between $\mathbf{S}_{\mathbf{\Psi}_1}$ and $\mathbf{S}_{\mathbf{\Psi}_2}$

```
Substrings ← Empty array
for i in range(len($\mathbf{S}_{\mathbf{\Psi}_1}$)):                      # Iterate through $\boldsymbol{S}_{\boldsymbol{\Psi}_1}$
    match ← Empty string
    for j in range(len($\mathbf{S}_{\mathbf{\Psi}_2}$)):                  # Iterate through $\boldsymbol{S}_{\boldsymbol{\Psi}_2}$
        if i + j < len($\mathbf{S}_{\mathbf{\Psi}_1}$) and $\mathbf{S}_{\mathbf{\Psi}_1}$[i + j] == $\mathbf{S}_{\mathbf{\Psi}_2}$[j]):   # Substring in given range is common
            match ← match + $\mathbf{S}_{\mathbf{\Psi}_2}$[j]            # Add current character
        end if
        else:
            Substrings.append(match)          # Add current substring to array
            match ← Empty string              # Reset matching string
        end else
    if match not empty:
        Substrings.append(match)              # Add current substring to array
```

3.3.2 Time-based Optimization

A feature of collaborative tasks is that gestures performed by the $\mathbf{\Phi_W}$ *Worker* are performed in response to gestures performed by the $\mathbf{\Phi_H}$ *Helper* (e.g. tying a nut after being instructed to do so). This shows a link between the time in which a gesture was authored and the time in which a response to it is provided. Hence, gestures can be compared based on time proximity: the closer in time the $\boldsymbol{w_i}$ Worker-authored gesture and the $\boldsymbol{h_j}$ Helper-authored gesture are, the more likely they are related.

The MAGIC-based optimization from Equation (3-4) does not consider time in its formulation. Therefore, our second approach was formulated to explore the relevance of time when comparing gestures in collaborative settings. A $\mathbf{C}$ coefficient matrix of time costs (of size $|\mathbf{W}| \times |\mathbf{H}|$) is computed with respect to the time in which the gestures were performed. The time in which each gesture is performed is stored in two vectors $\overrightarrow{t_{\mathbf{W}}}$ and $\overrightarrow{t_{\mathbf{H}}}$, respectively for $\boldsymbol{w_i}$ Worker-authored gestures and $\boldsymbol{h_j}$ Helper-authored gestures. Afterwards, the vectors are expanded into matrices (of size $|\mathbf{W}| \times |\mathbf{H}|$) by multiplying $\overrightarrow{t_{\mathbf{W}}}$ and $\overrightarrow{t_{\mathbf{H}}}$ by vectors of ones (of size $|\mathbf{H}| \times \mathbf{1}$ and $|\mathbf{W}| \times \mathbf{1}$, respectively). Finally, the $\mathbf{C}$ coefficient matrix of time costs is computed by subtracting these expanded matrices. Equation (3-5) summarizes the previous formulation, where T represents the transpose operation:

$$\mathbf{C} = \overrightarrow{t_{\mathbf{W}}}\overrightarrow{\mathbf{1}}^{\mathrm{T}} - \overrightarrow{\mathbf{1}}\overrightarrow{t_{\mathbf{H}}}^{\mathrm{T}} \tag{3-5}$$

After computing the $\mathbf{C}$ coefficient matrix of time costs using Equation (3-5), the optimal gesture matching solution will be found by solving the integer optimization problem depicted in Equation (3-6):

$$\begin{aligned} \text{minimize} \quad & \sum_{j=1}^{|\mathbf{H}|}\sum_{i=1}^{|\mathbf{W}|} c_{ij}e_{ij} \\ \text{subject to} \quad & \sum_{j=1}^{|\mathbf{H}|} e_{ij} = 1, \ \ \forall i \\ & \sum_{j=1}^{|\mathbf{H}|} c_{ij}e_{ij} \geq 0, \ \ \forall i \\ & e_{ij} \in \{0,1\} \\ & i = 1, 2, \ldots, |\mathbf{W}|; \ j = 1, 2, \ldots, |\mathbf{H}| \end{aligned} \tag{3-6}$$

The cost function is minimized whenever each $\boldsymbol{w_i}$ Worker-authored gesture is matched to the most recently performed $\boldsymbol{h_j}$ Helper-authored gesture. Similarly to the problem depicted in Equation (3-4), a constraint regulates that each $\boldsymbol{w_i}$ Worker-authored gesture can only be matched to one $\boldsymbol{h_j}$ Helper-authored gesture. An additional constraint is imposed to prevent negative costs from being considered in the minimization problem. A negative cost will be obtained whenever a $\boldsymbol{w_i}$ Worker-authored gesture is compared against a $\boldsymbol{h_j}$ Helper-authored gesture performed in a later time. This non-negativity constraint prevents always selecting the gesture with the most negative cost (the last gesture performed during the task).

3.3.3 Hybrid Optimization

Besides not considering time in its formulation, the MAGIC-based optimization approach from Equation (3-4) has another limitation: the outcomes of the gesture matching process can heavily depend on the subtrees selected to perform the comparisons. For example, comparing gestures only by analyzing their movement subtrees can result in incorrect gesture matchings. Therefore, a hybrid optimization approach was formulated so that both temporal synchrony and gesture similarity are considered when comparing the gestures. Integrating the temporal aspect to the MAGIC-based formulation keeps the gesture matching results from varying significantly with respect to the subtrees selected to compare.

The approach computes a $\mathbf{D}$ coefficient matrix of hybrid costs (of size $|\mathbf{W}| \times |\mathbf{H}|$) from the previous $\mathbf{B}$ coefficient matrix of distance costs and $\mathbf{C}$ coefficient matrix of time costs. We propose a function based on the signum function to regulate the effect of the b_{ij} and c_{ij} input costs in the d_{ij} hybrid costs (Bracewell & Bracewell, 1986). This function, depicted in Equation (3-7) combines a time damping section that reduces the importance of gestures based on the time they were performed, and a distance averaging section that normalizes the b_{ij} costs:

$$d_{ij} = \left[\left(-e^{-\alpha c_{ij}} \frac{-e^{-\alpha c_{ij}} - \beta}{|-e^{-\alpha c_{ij}} - \beta|} - e^{-\alpha c_{ij}}\right) + \gamma \frac{b_{ij}}{\sum_{j=1}^{|\mathbf{W}|} b_{ij}}\right] \tag{3-7}$$

The α damping constant regulates the damping effects of the c_{ij} time costs, and the β translation constant regulates when the damping begins. Figure 3.9 showcases the effect of these constants over the time damping section of Equation (3-7). The x-axes represent the values of the c_{ij} time-based costs from Equation (3-5). The y-axes represent the resulting values after applying the time damping section of Equation (3-7). The left graph of Figure 3.9 showcases effect of the α constant. Lower α values increase the importance of gestures performed less recently, while higher α values emphasize the most recent gestures. Typical values are $0.05 \leq \alpha \leq 0.1$ to balance the importance between recent and older performed gestures. The right graph of Figure 3.9 showcases effect of the β constant. The β constant regulates when Equation (3-7) activates. Typical values are $0.9 \leq \beta \leq 1.1$. For β values lower than 0.9, emphasis is given to gestures not yet performed, and values higher than 1.1 ignore the most recently performed gestures. The time damping section of Equation (3-7) reaches its maximum value when the $\boldsymbol{h_j}$ Helper-authored gesture and the $\boldsymbol{w_i}$ Worker-authored gesture were performed at the same time. Finally, the formulation introduces the γ distance constant, which regulates the importance that will be given to the normalized b_{ij} costs. This constant was set to the value of 2, which gives equal importance to both the time damping section and the distance averaging section.

After computing $\mathbf{D}$ coefficient matrix of hybrid costs using Equation (3-7), the optimal gesture matching solution will be found by solving the integer optimization problem depicted in Equation (3-8):

$$\begin{aligned} &\text{maximize} && \sum_{j=1}^{|\mathbf{H}|}\sum_{i=1}^{|\mathbf{W}|} d_{ij}e_{ij} \\ &\text{subject to} && \sum_{j=1}^{|\mathbf{H}|} e_{ij} = 1, \;\; \forall i \\ & && e_{ij} \in \{0,1\} \\ & && i = 1, 2, \dots, |\mathbf{W}|;\; j = 1, 2, \dots, |\mathbf{H}| \end{aligned} \tag{3-8}$$

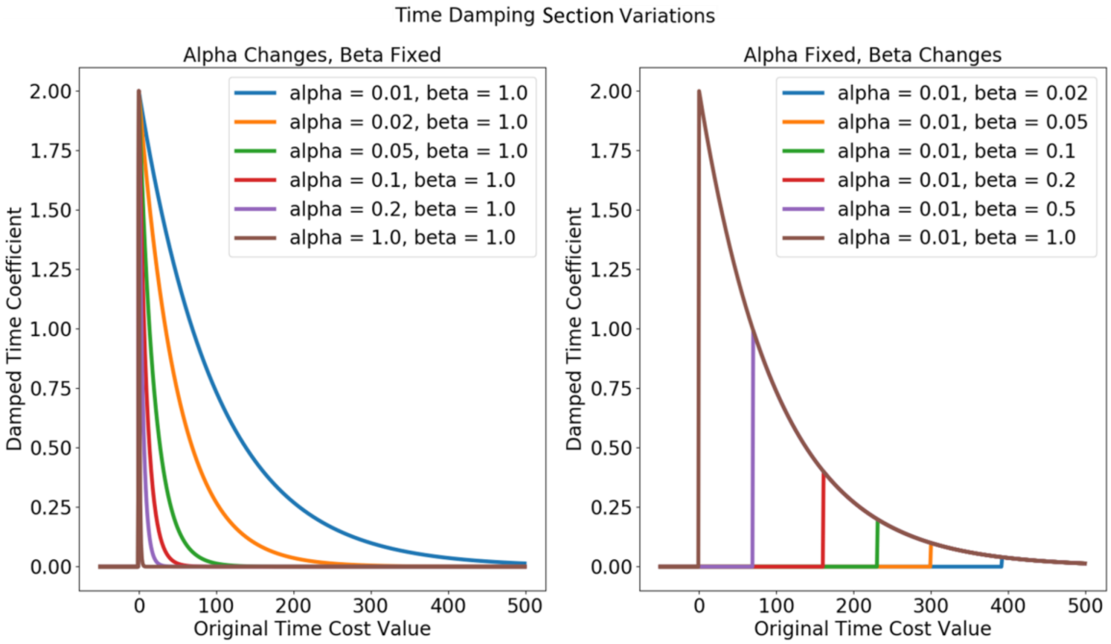

Figure 3.9 Effects of α and β in time damping effect, for c_{ij} ∈ [-50,500]. The α constant regulates the degree in which time attenuates the final hybrid cost. The β constant regulates when the damping starts.

The cost function is maximized whenever a balance between two conditions is achieved: 1) each $\boldsymbol{w_i}$ Worker-authored gesture is matched to the most recently performed $\boldsymbol{h_j}$ Helper-authored gesture, and 2) each $\boldsymbol{w_i}$ Worker-authored gesture is matched to the $\boldsymbol{h_j}$ Helper-authored gesture that has the highest number of common nodes (based on the subtrees being compared). As in the Equations (3-4) and (3-6), the constraint that establishes that each $\boldsymbol{w_i}$ Worker-authored gesture can only be matched to one $\boldsymbol{h_j}$ Helper-authored gesture is preserved. There is no need for an additional temporal constraint as in Equation (3-6), as the effect of time is considered in the time damping section of Equation (3-7).

Figure 3.10 exemplifies how our gesture matching approaches match the gestures performed by a $\boldsymbol{\Phi_{\mathrm{H}}}$ *Helper* and a $\boldsymbol{\Phi_{\mathrm{W}}}$ *Worker*. The first row showcases ten gestures performed by a $\boldsymbol{\Phi_{\mathrm{H}}}$ *Helper* to convey instructions ($\boldsymbol{h_j}$; $j = 1, 2, \dots, 10$). The second row showcases ten gestures performed by a $\boldsymbol{\Phi_{\mathrm{W}}}$ *Worker* to execute the received instructions ($\boldsymbol{w_i}$; $i = 1, 2, \dots, 10$). An arrow connecting the $\boldsymbol{w_i}$ Worker-authored gesture and the $\boldsymbol{h_j}$ Helper-authored gestures represents that these gestures match. The color of the arrows represents a different gesture matching approach: green, blue, and red for the MAGIC-based, time-base, and hybrid approach, respectively. The gestures are matched based on the cost coefficients from the **B**, **C**, and **D** matrices, showcased in the last three rows. Each column (one per each gesture performed by the $\boldsymbol{\Phi_{\mathrm{W}}}$ *Worker*) will have 30 costs associated, clustered into three groups representing each gesture matching approach (green for MAGIC-based, blue for time-based, red for hybrid). Each of these three groups contains 10 cost coefficients, one per each gesture performed by the $\boldsymbol{\Phi_{\mathrm{H}}}$ *Helper*. Within each group of 10 cost coefficients, one cost coefficient is emphasized with a rectangle. The emphasized cost coefficients are the ones that optimize the objective functions. Therefore, the optimal gesture matching will be found when only the e_{ij} edge weights corresponding to the emphasized cost coefficients have a value of 1 in the **E** matrix. For example, in the third row and first column, the b_{11} distance cost is emphasized. This means that the MAGIC-based approach matched the gesture $\boldsymbol{h_1}$ with the gesture $\boldsymbol{w_1}$, represented by a green arrow connecting the two gestures. Additionally, Figure 3.11 depicts the **E** matrices generated from the example in Figure 3.10.

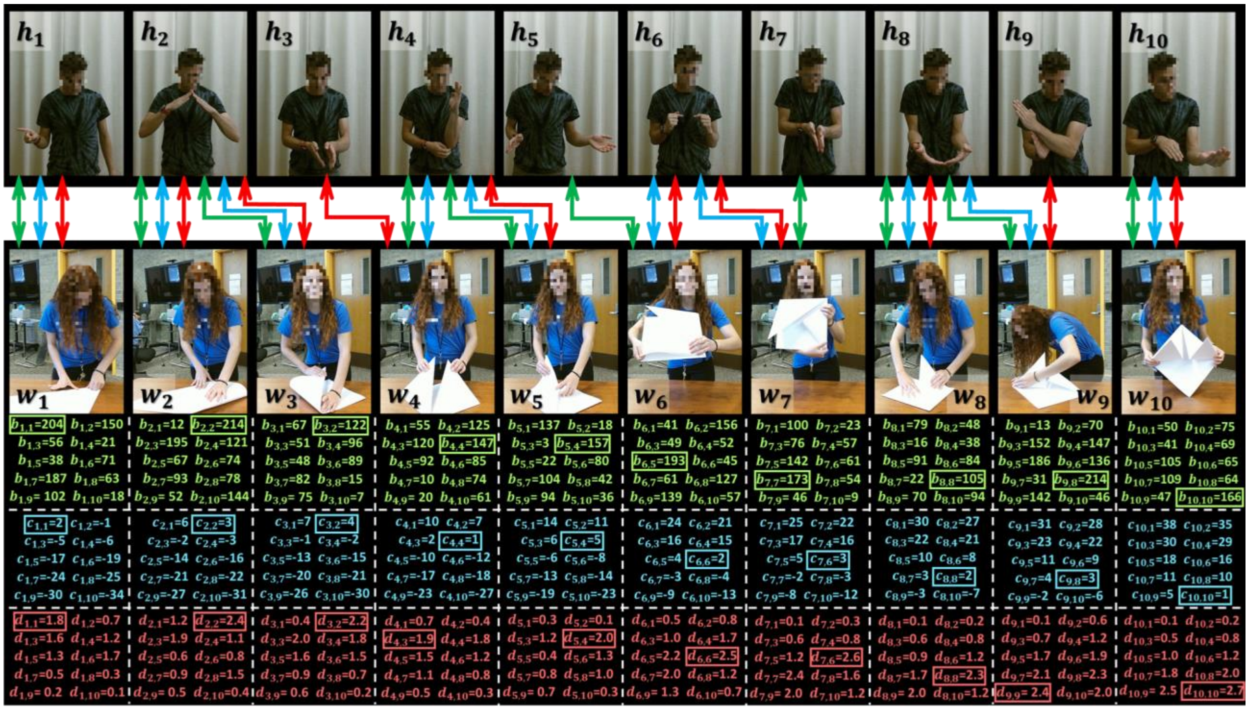
h_1
h_2
h_3
h_4
h_5
h_6
h_7
h_8
h_9
h_{10}
w_1
w_2
w_3
w_4
w_5
w_6
w_7
w_8
w_9
w_{10}
$b_{1,1}$=204 $b_{1,2}$=150 $b_{1,3}$=56 $b_{1,4}$=21 $b_{1,5}$=38 $b_{1,6}$=71 $b_{1,7}$=187 $b_{1,8}$=63 $b_{1,9}$= 102 $b_{1,10}$=18
$c_{1,1}$=2 $c_{1,2}$=-1 $c_{1,3}$=-5 $c_{1,4}$=-6 $c_{1,5}$=-17 $c_{1,6}$=-19 $c_{1,7}$=-24 $c_{1,8}$=-25 $c_{1,9}$=-30 $c_{1,10}$=-34
$d_{1,1}$=1.8 $d_{1,2}$=0.7 $d_{1,3}$=1.6 $d_{1,4}$=1.2 $d_{1,5}$=1.3 $d_{1,6}$=1.7 $d_{1,7}$=0.5 $d_{1,8}$=0.3 $d_{1,9}$= 0.2 $d_{1,10}$=0.1
$b_{2,1}$=12 $b_{2,2}$=214 $b_{2,3}$=195 $b_{2,4}$=121 $b_{2,5}$=67 $b_{2,6}$=74 $b_{2,7}$=93 $b_{2,8}$=78 $b_{2,9}$= 52 $b_{2,10}$=144
$c_{2,1}$=6 $c_{2,2}$=3 $c_{2,3}$=-2 $c_{2,4}$=-3 $c_{2,5}$=-14 $c_{2,6}$=-16 $c_{2,7}$=-21 $c_{2,8}$=-22 $c_{2,9}$=-27 $c_{2,10}$=-31
$d_{2,1}$=1.2 $d_{2,2}$=2.4 $d_{2,3}$=1.9 $d_{2,4}$=1.1 $d_{2,5}$=0.6 $d_{2,6}$=0.8 $d_{2,7}$=0.9 $d_{2,8}$=1.5 $d_{2,9}$= 0.5 $d_{2,10}$=0.4
$b_{3,1}$=67 $b_{3,2}$=122 $b_{3,3}$=51 $b_{3,4}$=96 $b_{3,5}$=48 $b_{3,6}$=89 $b_{3,7}$=82 $b_{3,8}$=15 $b_{3,9}$= 75 $b_{3,10}$=7
$c_{3,1}$=7 $c_{3,2}$=4 $c_{3,3}$=-1 $c_{3,4}$=-2 $c_{3,5}$=-13 $c_{3,6}$=-15 $c_{3,7}$=-20 $c_{3,8}$=-21 $c_{3,9}$=-26 $c_{3,10}$=-30
$d_{3,1}$=0.4 $d_{3,2}$=2.2 $d_{3,3}$=2.0 $d_{3,4}$=1.8 $d_{3,5}$=1.6 $d_{3,6}$=1.5 $d_{3,7}$=0.9 $d_{3,8}$=0.7 $d_{3,9}$= 0.6 $d_{3,10}$=0.2
$b_{4,1}$=55 $b_{4,2}$=125 $b_{4,3}$=120 $b_{4,4}$=147 $b_{4,5}$=92 $b_{4,6}$=85 $b_{4,7}$=10 $b_{4,8}$=74 $b_{4,9}$= 20 $b_{4,10}$=61
$c_{4,1}$=10 $c_{4,2}$=7 $c_{4,3}$=2 $c_{4,4}$=1 $c_{4,5}$=-10 $c_{4,6}$=-12 $c_{4,7}$=-17 $c_{4,8}$=-18 $c_{4,9}$=-23 $c_{4,10}$=-27
$d_{4,1}$=0.7 $d_{4,2}$=0.4 $d_{4,3}$=1.9 $d_{4,4}$=1.8 $d_{4,5}$=1.5 $d_{4,6}$=1.2 $d_{4,7}$=1.1 $d_{4,8}$=0.8 $d_{4,9}$=0.5 $d_{4,10}$=0.3
$b_{5,1}$=137 $b_{5,2}$=18 $b_{5,3}$=3 $b_{5,4}$=157 $b_{5,5}$=22 $b_{5,6}$=80 $b_{5,7}$=104 $b_{5,8}$=42 $b_{5,9}$= 94 $b_{5,10}$=36
$c_{5,1}$=14 $c_{5,2}$=11 $c_{5,3}$=6 $c_{5,4}$=5 $c_{5,5}$=-6 $c_{5,6}$=-8 $c_{5,7}$=-13 $c_{5,8}$=-14 $c_{5,9}$=-19 $c_{5,10}$=-23
$d_{5,1}$=0.3 $d_{5,2}$=0.1 $d_{5,3}$=1.2 $d_{5,4}$=2.0 $d_{5,5}$=0.4 $d_{5,6}$=1.3 $d_{5,7}$=0.8 $d_{5,8}$=1.0 $d_{5,9}$= 0.7 $d_{5,10}$=0.3
$b_{6,1}$=41 $b_{6,2}$=156 $b_{6,3}$=49 $b_{6,4}$=52 $b_{6,5}$=193 $b_{6,6}$=45 $b_{6,7}$=61 $b_{6,8}$=127 $b_{6,9}$=139 $b_{6,10}$=57
$c_{6,1}$=24 $c_{6,2}$=21 $c_{6,3}$=16 $c_{6,4}$=15 $c_{6,5}$=4 $c_{6,6}$=2 $c_{6,7}$=-3 $c_{6,8}$=-4 $c_{6,9}$=-9 $c_{6,10}$=-13
$d_{6,1}$=0.5 $d_{6,2}$=0.8 $d_{6,3}$=1.0 $d_{6,4}$=1.7 $d_{6,5}$=2.2 $d_{6,6}$=2.5 $d_{6,7}$=2.0 $d_{6,8}$=1.2 $d_{6,9}$= 1.3 $d_{6,10}$=0.7
$b_{7,1}$=100 $b_{7,2}$=23 $b_{7,3}$=76 $b_{7,4}$=57 $b_{7,5}$=142 $b_{7,6}$=61 $b_{7,7}$=173 $b_{7,8}$=54 $b_{7,9}$= 46 $b_{7,10}$=9
$c_{7,1}$=25 $c_{7,2}$=22 $c_{7,3}$=17 $c_{7,4}$=16 $c_{7,5}$=5 $c_{7,6}$=3 $c_{7,7}$=-2 $c_{7,8}$=-3 $c_{7,9}$=-8 $c_{7,10}$=-12
$d_{7,1}$=0.1 $d_{7,2}$=0.3 $d_{7,3}$=0.6 $d_{7,4}$=0.8 $d_{7,5}$=1.2 $d_{7,6}$=2.6 $d_{7,7}$=2.4 $d_{7,8}$=1.6 $d_{7,9}$= 2.0 $d_{7,10}$=1.2
$b_{8,1}$=79 $b_{8,2}$=48 $b_{8,3}$=16 $b_{8,4}$=38 $b_{8,5}$=91 $b_{8,6}$=84 $b_{8,7}$=22 $b_{8,8}$=105 $b_{8,9}$= 70 $b_{8,10}$=94
$c_{8,1}$=30 $c_{8,2}$=27 $c_{8,3}$=22 $c_{8,4}$=21 $c_{8,5}$=10 $c_{8,6}$=8 $c_{8,7}$=3 $c_{8,8}$=2 $c_{8,9}$=-3 $c_{8,10}$=-7
$d_{8,1}$=0.1 $d_{8,2}$=0.2 $d_{8,3}$=0.6 $d_{8,4}$=0.8 $d_{8,5}$=0.9 $d_{8,6}$=1.2 $d_{8,7}$=1.7 $d_{8,8}$=2.3 $d_{8,9}$= 2.0 $d_{8,10}$=1.2
$b_{9,1}$=13 $b_{9,2}$=70 $b_{9,3}$=152 $b_{9,4}$=147 $b_{9,5}$=186 $b_{9,6}$=136 $b_{9,7}$=31 $b_{9,8}$=214 $b_{9,9}$=142 $b_{9,10}$=46
$c_{9,1}$=31 $c_{9,2}$=28 $c_{9,3}$=23 $c_{9,4}$=22 $c_{9,5}$=11 $c_{9,6}$=9 $c_{9,7}$=4 $c_{9,8}$=3 $c_{9,9}$=-2 $c_{9,10}$=-6
$d_{9,1}$=0.1 $d_{9,2}$=0.6 $d_{9,3}$=0.7 $d_{9,4}$=1.2 $d_{9,5}$=1.7 $d_{9,6}$=1.9 $d_{9,7}$=2.1 $d_{9,8}$=2.3 $d_{9,9}$= 2.4 $d_{9,10}$=2.0
$b_{10,1}$=50 $b_{10,2}$=75 $b_{10,3}$=41 $b_{10,4}$=69 $b_{10,5}$=105 $b_{10,6}$=65 $b_{10,7}$=109 $b_{10,8}$=64 $b_{10,9}$=47 $b_{10,10}$=166
$c_{10,1}$=38 $c_{10,2}$=35 $c_{10,3}$=30 $c_{10,4}$=29 $c_{10,5}$=18 $c_{10,6}$=16 $c_{10,7}$=11 $c_{10,8}$=10 $c_{10,9}$=5 $c_{10,10}$=1
$d_{10,1}$=0.1 $d_{10,2}$=0.2 $d_{10,3}$=0.5 $d_{10,4}$=0.8 $d_{10,5}$=1.0 $d_{10,6}$=1.2 $d_{10,7}$=1.8 $d_{10,8}$=2.0 $d_{10,9}$= 2.5 $d_{10,10}$=2.7

Figure 3.10 Example of how the collaborators' gestures are matched according to the different optimization problems.

Obtaining the **E** matrices representing gesture matching completes the MAGIC architecture. Gestures are successfully represented into data structures encompassing morphology, semantics, and pragmatics. Moreover, these structures are leveraged to compare which gestures are more functionally equivalent to each other. The next subsection explains our approach to use these **E** matrices to generate the PIA metric, a score that estimates task understanding by representing how well are the physical instructions used by the participants being assimilated.

$$\begin{array}{c|cccccccccc} & h_1 & h_2 & h_3 & h_4 & h_5 & h_6 & h_7 & h_8 & h_9 & h_{10} \\ \hline w_1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ w_2 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ w_3 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ w_4 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ w_5 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ w_6 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ w_7 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ w_8 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ w_9 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ w_{10} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}$$

MAGIC-based Approach

$$\begin{array}{c|cccccccccc} & h_1 & h_2 & h_3 & h_4 & h_5 & h_6 & h_7 & h_8 & h_9 & h_{10} \\ \hline w_1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ w_2 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ w_3 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ w_4 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ w_5 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ w_6 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ w_7 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ w_8 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ w_9 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ w_{10} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}$$

Time-based Approach

$$\begin{array}{c|cccccccccc} & h_1 & h_2 & h_3 & h_4 & h_5 & h_6 & h_7 & h_8 & h_9 & h_{10} \\ \hline w_1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ w_2 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ w_3 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ w_4 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ w_5 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ w_6 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ w_7 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ w_8 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ w_9 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ w_{10} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}$$

Hybrid Approach

Figure 3.11 **E** matrices representing the gesture matching solutions showcased in Figure 3.10.

3.4 Physical Information Assimilation Metric

The PIA metric is a score representing the quality of assimilation of the physical instructions exchanged between agents collaborating to solve a shared physical task. Clark and colleagues developed a framework that describes the process of understanding when individuals communicate (Clark & Brennan, 1991; Clark & Schaefer, 1987; Sacks et al., 1978). The exchange of utterances happens in two phases. The speaker (i.e. the $\boldsymbol{\Phi}_{\mathbf{H}}$ *Helper*) presents an utterance to the receiver (i.e. the $\boldsymbol{\Phi}_{\mathbf{W}}$ *Worker*). If the receiver generates enough understanding evidence $\boldsymbol{\mathcal{E}}$, the speaker can assume that the receiver understands the utterance. The process in which the receiver provides this evidence $\boldsymbol{\mathcal{E}}$ can be divided into 4 states, ranging from not noticing the initial utterance (State 0) to the correct understanding of it (State 3). If the evidence $\boldsymbol{\mathcal{E}}$ provided by the receiver does not support a State 3 understanding, either the speaker will need to elaborate the conveyed utterance, or the receiver will need to generate different evidence $\boldsymbol{\mathcal{E}}$. We part from the assumption that Clark's framework can be applied to the gestural utterances. Given this framework, our approach assumes perfect task understanding of a collaboration process happens when every $\boldsymbol{h_j}$ Helper-authored gesture is mapped to one and only one $\boldsymbol{w_i}$ Worker-authored gesture. In other words, the $\boldsymbol{\Phi}_{\mathbf{W}}$ *Worker* generated evidence $\boldsymbol{\mathcal{E}}$ that supported correct understanding (State 3) for every utterance given by the $\boldsymbol{\Phi}_{\mathbf{H}}$ *Helper*. Contrarily, a task where one $\boldsymbol{h_j}$ Helper-authored gesture is mapped to several $\boldsymbol{w_i}$ Worker-authored gestures represents poor understanding: the State 3 was not reached for every utterance.

The PIA metric represents this behavior by analyzing the optimal gesture matching solutions generated in the previous stage. Given the $\mathbf{W}$ and $\mathbf{H}$ sets and $\mathbf{E}$ matrix of $\boldsymbol{e_{ij}}$ edge weights from the previous stage, the gesture matching for a particular task can be represented with a bipartite graph $\mathbf{T} = (\mathbf{H}, \mathbf{W}, \mathbf{E})$. $\mathbf{H}$ and $\mathbf{W}$ are disjoint and independent sets, and $\mathbf{E}$ contains the matchings between their vertices. Figure 3.12 presents a graphic example of the graph representing the gesture matching from Figure 3.11.

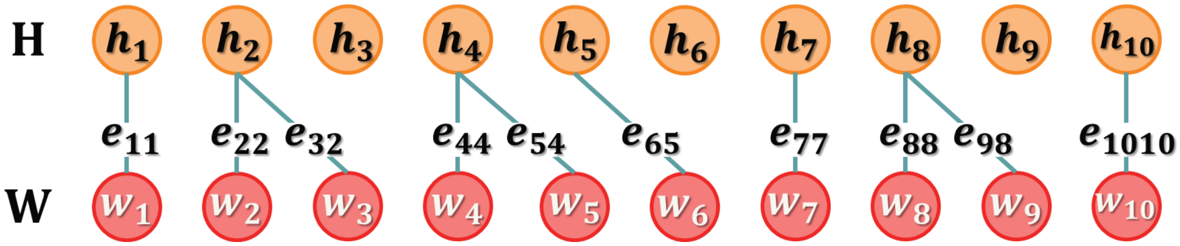

Figure 3.12 Example of a graph representing the **E** matrix from Figure 3.11, obtained using the MAGIC-based optimization approach. The $\boldsymbol{e_{ij}}$ edge weights in the image have a value of “1” in the **E** matrix.

Given the characteristics of the gesture matching performed in the previous stage, MAGIC ensures (except in one situation that will be further elaborated) that a matching of **H** can be found in every bipartite graph **T** representing a task. A matching of **H** is a set of the edges chosen in such a way that no two edges share endpoint vertices, as described in Equation (3-9):

$$\left|N\left(\widetilde{H}\right)\right| \geq \left|\widetilde{H}\right|, \forall\, \widetilde{H} \subseteq \mathbf{H} \tag{3-9}$$

where $N\left(\widetilde{H}\right)$ represents the neighborhood of $\widetilde{H}$, the set of vertices in **W** that are connected to vertices of $\widetilde{H}$. Moreover, the scenario in which PIA is highest (i.e. one-to-one gesture matching) has a matching of **H** that satisfies the formulation in Equation (3-10):

$$\left|N\left(\widetilde{H}\right)\right| = \left|\widetilde{H}\right|, \forall\, \widetilde{H} \subseteq \mathbf{H} \tag{3-10}$$

This scenario describes a maximum bipartite matching of **H**, constrained by |**H**| and |**W**| ≥ 1, and $deg(\boldsymbol{w_i}) = 1$, where $deg(\boldsymbol{w_i})$ represents the number of edges of the $\boldsymbol{w_i}$ vertex (Glover, 1967). Following this formulation, the PIA score can be computed as given in Equation (3-11):

$$\text{PIA} = \frac{1}{|\mathbf{H}|}\sum_{j=1}^{|\mathbf{H}|}\left(\sum_{i=1}^{|\mathbf{W}|} e_{ij}\right)^{-1} \tag{3-11}$$

A PIA score of 100, will be found whenever no $\boldsymbol{h_j}$ Helper-authored gesture is matched to more than one $\boldsymbol{w_i}$ Worker-authored gesture (a maximum matching of $\mathbf{H}$). Alternatively, the PIA score will be less than 100 whenever a maximum matching of $\mathbf{H}$ is not achieved. This will be the case when: 1) multiple $\boldsymbol{w_i}$ Worker-authored gestures are associated to the same $\boldsymbol{h_j}$ Helper-authored gesture; or 2) at least one $\boldsymbol{h_j}$ Helper-authored gesture was not associated with any $\boldsymbol{w_i}$ Worker-authored gesture. This latter scenario is the previously mentioned condition in which a matching of $\mathbf{H}$ cannot be found. To address such cases, a pre-processing step is applied to the **E** matrices in which their column rank is inspected. If a particular **E** matrix does not have full column rank (e.g. a column with only zeros, representing an unmatched $\boldsymbol{h_j}$ Helper-authored gesture), the matrix reduction process is performed in which the linearly dependent column is removed. This process is only performed over the **E** matrices to prevent the PIA from running into undefined scenarios (e.g. division by zero). This preprocessing step does not affect the $\frac{1}{|\mathbf{H}|}$ term in the PIA calculation, and therefore penalties will still be introduced in the calculation, inversely proportional to the amount of $\boldsymbol{h_j}$ Helper-authored gestures.

3.5 Summary

This chapter explains in detail the framework proposed to assess collaborative physical tasks through gestural analysis. An architecture to represent and compare gestures' morphology, semantics and pragmatics is proposed. In doing so, MAGIC allows to perform comparisons between gestures that consider more information than just gestures' appearance. Additionally, a metric to estimate task understanding based on the gesture used by the individuals collaborating is proposed. It should be noted that the PIA metric does not replace the common proxy metrics for task understanding. Instead, the PIA metric complements these metrics by analyzing task understanding from a different perspective, the physical one.

4. EXPERIMENTS AND RESULTS

This chapter explains the experimental design and the results obtained based on the proposed methodology. A data collection protocol was established to elicit and collect gestures from individuals completing a collaborative task remotely. The data from participants collaborating to complete shared tasks was used to generate representations of gestures. Participants completed either a block assembly task, an origami task, or an ultrasound training task. The representations were generated using MAGIC's approach, and two other approaches to represent gestures: morpho-semantic descriptors (Madapana & Wachs, 2017) and a temporal synchronization approach. RQ1 was addressed by comparing the quality of the representation structure. This was done by evaluating the matching scores of the three different gesture representation structures. Moreover, the gesture matching for MAGIC was computed using both the subtree intersections and the tree kernels approaches, detailed in the previous chapter, which addresses RQ2 (*How to measure gesture similarity?).* Finally, the data from the gestures was also used to obtain insights regarding task understanding. For this, we obtained estimates of task understanding with the PIA metric and three other common proxy metrics: number of errors, idle time, and task completion percentage. We address RQ3 (*How can gesture similarity lead to estimate task understanding?*) by obtaining insights of task understanding via gestural analysis. Additional insights of the participants' understanding and performance as they completed the task were obtained via an understanding assessment questionnaire and the NASA Task Load Index (TLX) (Hart, 2006).

4.1 Data Collection

Three user studies were conducted in which a total of 60 participants (graduate students, 34 males and 26 females, aged 26.36 ± 4.4 years old). The participants were divided into 30 Helper-Worker pairs to collaboratively complete a shared task. To explore the generalizability of our approach, the shared task to be completed was different for each user study. The tasks performed are described as follows:

1 Block Assembly Task: Participants had to complete a block assembly task similar to those in Fussell et al. and Kirk et al. (2004; 2007). The objective of the Helper-Worker pairs was to assemble a helicopter using the blocks, depicted in Figure 4.1. The task consisted of 24

different steps, in which blocks needed to be connected in specific ways to assemble the model.

2 Origami Task: Participants had to complete an origami assembly task similar to those in (Fakourfar et al., 2016; Kim et al., 2019). The objective of the Helper-Worker pairs was to assemble a samurai hat by folding a paper sheet, depicted in Figure 4.1. The task consisted of 11 different steps, in which the paper needed to fold onto itself in specific ways.

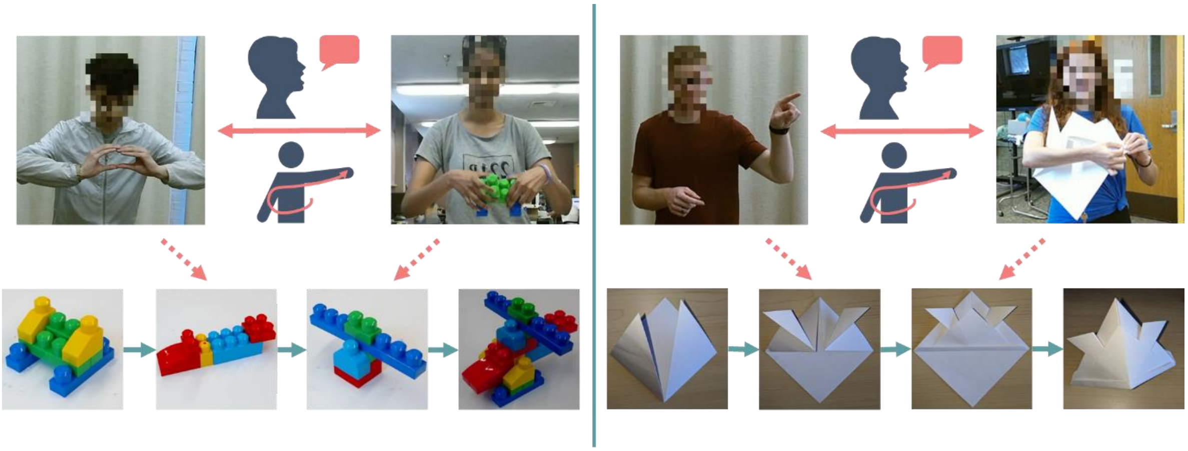

Figure 4.1 Example of participants collaborating to complete a shared task. *Left*: Steps of the block assembly task are depicted. *Right*: Steps of the block assembly task are depicted.

3 Ultrasound Training Task: Participants had to complete an ultrasound task comprised of three different subtasks. For this purpose, an ultrasound phantom with seven vessel lumens was created using ballistic gel, following (Amini et al., 2015). Each vessel was filled with water, each of a different color. Afterwards, the model was coated with a mixture of silicone and paint to emulate skin and to prevent the water to be seen at plain sight. Additionally, a wooden object (6cm length × 1.5cm width × 1.5mm height) was inserted to simulate a foreign body. Participants had to use an ultrasound probe (Telemed MicrUs MC10-5R10S-3) to complete three common tasks in ultrasound training curricula. First, participants had 10 minutes to detect the position of the vessel lumens, as in (Amini et al., 2015). Afterwards, participants had 10 minutes to extract 2cc of water from the vessels with a syringe, similar to (Thorn et al., 2016). Finally, participants had 5 minutes to identify the position, shape, and orientation of the foreign body inside the ultrasound phantom, similar to (Schlager et al., 1991). Figure 4.2 showcases the ultrasound phantom and the

tasks participants had to perform. The user study describing the ultrasound training task can be found explained in detail in (Rojas-Muñoz & Wachs, 2020).

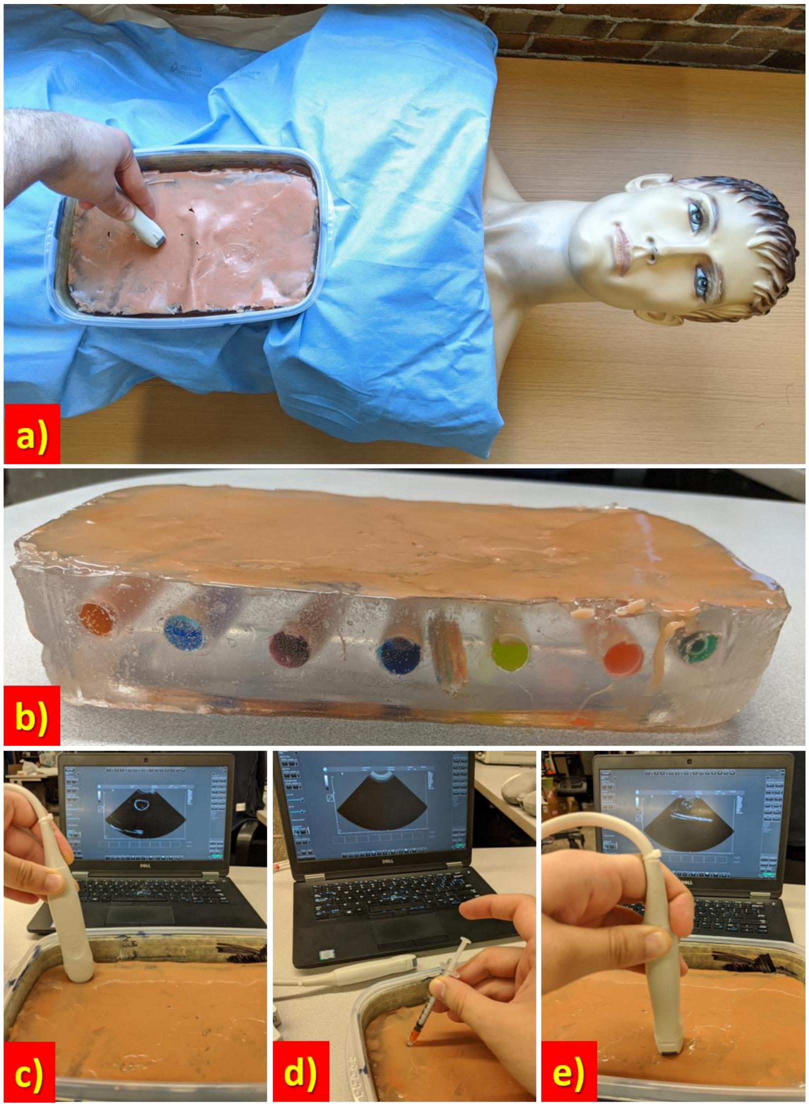

Figure 4.2 Ultrasound training task. The complete ultrasound simulator is shown in a). The ultrasound phantom is shown in b). The vessel lumens as seen in the ultrasound are shown in c). The process of extracting blood is shown in d). The foreign body as seen in the ultrasound is shown in e).

After signing a written consent form, participants were randomly assigned to either the $\mathbf{\Phi_W}$ *Worker* or the $\mathbf{\Phi_H}$ *Helper* role, and were directed to different stations according to their role. The $\mathbf{\Phi_W}$ *Worker* station (Figure 4.3) included elements to perform the task (e.g. blocks, paper sheet, ultrasound probe and phantom) placed on a table; color and depth cameras installed to record the participant's activity, and a display connected a computer hosting a Skype video call with the Helper station. The $\mathbf{\Phi_H}$ *Helper* station (Figure 4.4) had the same setup, but replaced the elements to perform the task with printed instructions on how to complete the task. Therefore, only the $\mathbf{\Phi_H}$ *Helper* knew the steps to proceed with the task, and only the $\mathbf{\Phi_W}$ *Worker* could interact with the elements to complete it. The $\mathbf{\Phi_H}$ *Helper* conveyed the instructions in the booklet to the $\mathbf{\Phi_W}$ *Worker* through verbal instructions, gestures, facial expressions, among others.

The generated $\boldsymbol{w_i}$ Worker-authored gestures were divided into 2 groups: responses to verbal utterances, and responses to gestural utterances. Since this dissertation focuses in the assimilation of the physical instructions, only responses to gestural utterances were consider for our calculations. A total of 3498 gestures were extracted, 2101 performed by the $\mathbf{\Phi_H}$ *Helper* and 1397 performed by the $\mathbf{\Phi_W}$ *Worker*, over the span of 14 hours of video. Elements introduced by verbal instructions (e.g. discourse referents, predicates) were considered as part of the $\mathbf{\Omega}$ *Context* of the gestural utterances. Finally, although non-verbal cues such as face expressions and posture also communicate information (Argyle, 2013), our work does not consider this information as part of the $\mathbf{\Omega}$ *Context* of the gestural utterances.

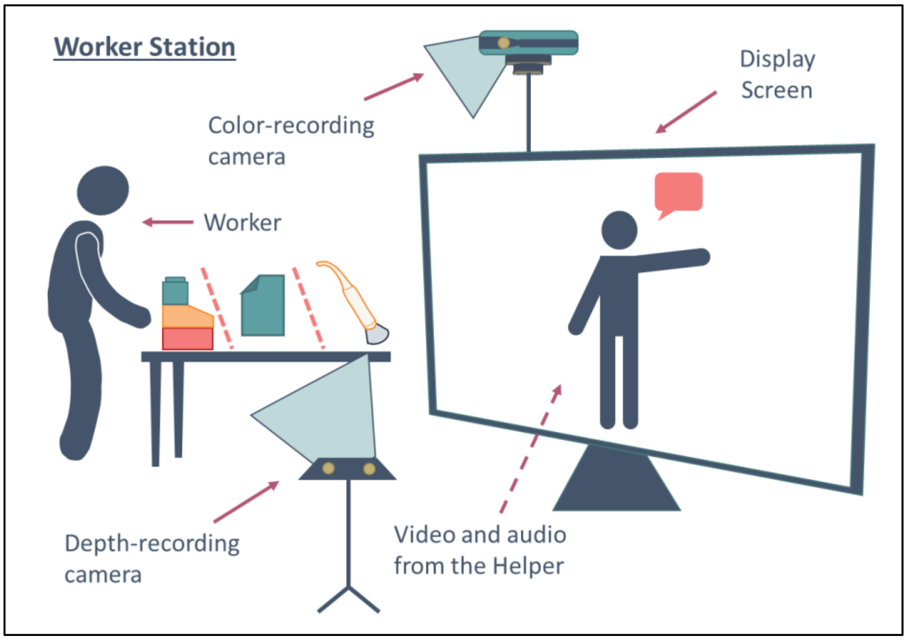

Figure 4.3 Worker station setup.

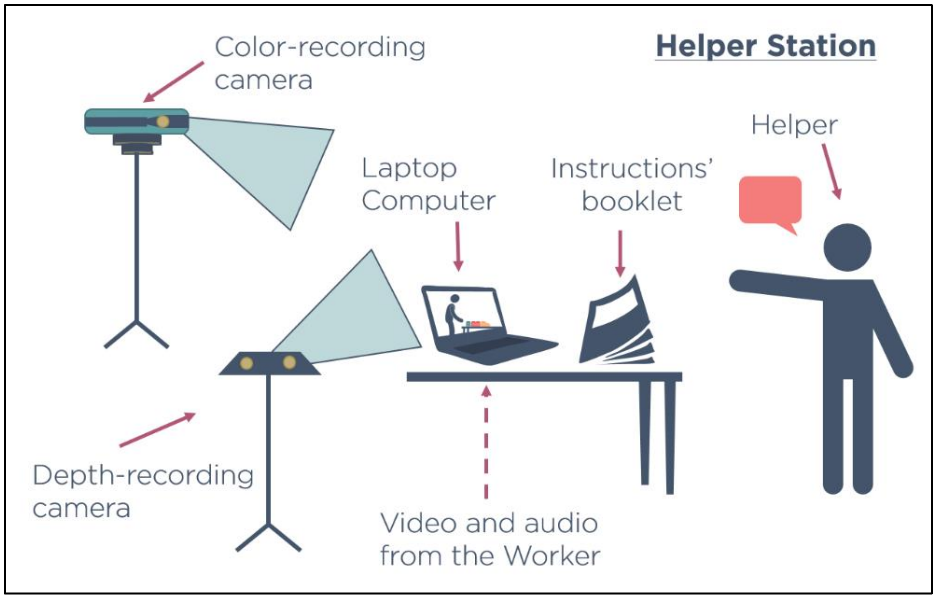

Figure 4.4 Helper station setup.

Having obtained the participants' gestures, a member of the research team watched the videos to create a gesture matching ground truth. Every $\boldsymbol{w_i}$ Worker-authored gesture was matched to a $\boldsymbol{h_j}$ Helper-authored gesture, based on the verbal and contextual information exchanged by the participant. The output were $\widehat{\mathbf{E}_{\boldsymbol{\ell}}}$ matrices of $\boldsymbol{e_{ij}}$ edge weights of gesture matchings representing human-annotated ground truth, one per Helper-Worker pair of $(\boldsymbol{\ell} = 1, 2, \dots, 30)$.

4.2 Ground Truth Annotation

The objective of this experiment was to analyze how well MAGIC $\boldsymbol{\Psi}$ *Interpretation Trees* were in capturing the information from the gestures (e.g. shape, movement, context-related information). This was performed by: 1) obtaining a $\mathbf{E}_{\boldsymbol{\ell}}^{\mathbf{k}}$ matrix of gestures matches for each of the Helper-Worker pairs; and 2) comparing these $\mathbf{E}_{\boldsymbol{\ell}}^{\mathbf{k}}$ matrices against the $\widehat{\mathbf{E}_{\boldsymbol{\ell}}}$ matrix of ground truth gestures matches. The later comparison is represented as a matching score (4-1), the harmonic average of precision and recall of different $\mathbf{E}_{\boldsymbol{\ell}}^{\mathbf{k}}$ matrices when compared with the ground truth gesture matches:

$$Matching\ Score = \frac{2\text{TP}}{2\text{TP} + \text{FN} + \text{FP}} \tag{4-1}$$

where TP are the True Positives (i.e. $\widehat{\mathbf{E}_{\ell}} = \mathbf{E}_{\ell}^{\mathbf{k}} = 1$), FN are the False Negatives (i.e. $\widehat{\mathbf{E}_{\ell}} = 1;\ \mathbf{E}_{\ell}^{\mathbf{k}} = 0$); and FP are the False Positives (i.e. $\widehat{\mathbf{E}_{\ell}} = 0;\ \mathbf{E}_{\ell}^{\mathbf{k}} = 1$) (Goutte & Gaussier, 2005).

4.3 Evaluating Gesture Comparison Approaches

MAGIC's gesture matching scores were evaluated against the gesture matching scores of two baseline approaches: morpho-semantic descriptors (MSD) vectors (Madapana & Wachs, 2017), and a naïve temporal synchronization (NTS) approach. Thus, our work uses a total of $\mathbf{k} = 3$ different gesture matching approaches. Moreover, MAGIC's gesture matching scores can be obtained both via subtree intersections and two different tree kernels. Therefore, we report the results of comparing the gestures using the MAGIC-based optimization and Hybrid optimization approaches for each of these three methods of generating the **B** coefficient matrices of distance costs.

MSD are Boolean vectors representing physical (e.g. is the hand moving to the right?) and semantic (e.g. is the hand referring to the head?) characteristics of the gestures. Each gesture was therefore represented with a 48×1 Boolean vector, in which each row represents whether one of the selected 48 descriptors described in Table 4.1 is presented or not for that gesture. MSD vectors were included as a gesture representation baseline based on physical similarity.

Table 4.1 Morpho-Semantic Descriptors.

#	**High-level descriptor**	**Low-level descriptor**
1	Left Hand Trajectory	Right, Up, Left, Down, Forward, Backward, Clockwise, Counter-clockwise, Iterative
2	Right Hand Trajectory	
3	Left Hand Orientation	Right, Left, Up, Down, Forward, Backward
4	Right Hand Orientation	
5	Left Arm Pose	
6	Right Arm Pose	
7	Left Hand Motion Plane	Sagittal, Frontal, Transverse
8	Right Hand Motion Plane	

Let $\vec{x}$ be the MSD vector representing a $\boldsymbol{h_j}$ Helper-authored gesture. Similarly, let $\vec{y}$ be MSD vector representing a $\boldsymbol{w_i}$ Worker-authored gesture. To compare these MSD vectors and obtain a notion of gesture similarity, the Hamming Distance (4-2) and Cosine Similarity (4-3) metrics were used.

$$Hamming\ Distance(\vec{x},\vec{y}) = \sum_{i=1}^{|\vec{x}|} \vec{x}_{\mathrm{i}} == \vec{y}_{\mathrm{i}} \tag{4-2}$$

$$Cosine\ Similarity(\vec{x},\vec{y}) = \frac{\vec{x}\cdot\vec{y}}{\|\vec{x}\|\|\vec{y}\|} \tag{4-3}$$

NTS was the second baseline approach. This 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 $\boldsymbol{w_i}$ Worker-authored gesture performed right after a $\boldsymbol{h_j}$ Helper-authored 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 response to the other one). For each $\boldsymbol{h_j}$ Helper-authored gesture, a time window before and after its execution was created. Every $\boldsymbol{w_i}$ Worker-authored gesture inside this time window was associated to the given $\boldsymbol{h_j}$ Helper-authored gesture. To obtain these time windows, the $\boldsymbol{t_j}$ timestamp in which each Helper-authored gesture was performed is recorded. Afterwards, the time range between consecutive timestamps was calculated as, for example, $\boldsymbol{t_2 - t_1}$. This time ranges will be split to represent the bounds of the time windows (e.g. $\boldsymbol{\frac{t_2-t_1}{2}}$ will be both the final bound of the $\boldsymbol{h_2}$ Helper-authored gesture, and the initial bound of the $\boldsymbol{h_1}$ Helper-authored gesture). This process will be performed for all the gestures in the **H** set. The initial bound for the first Helper-authored gesture will be the beginning of the video, and the final bound for the last Helper-authored gesture will be the end of the video. Figure 4.4 depicts this approach.

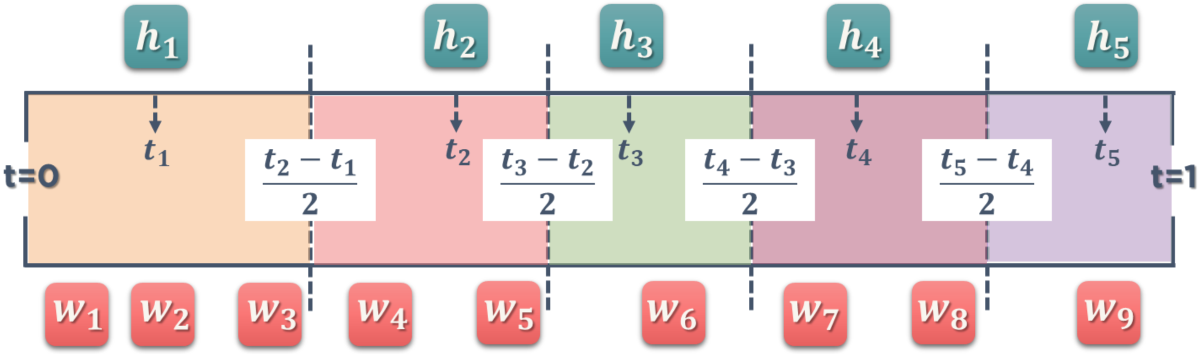

Figure 4.5 Temporal synchronization gesture matching approach.

4.3.1 Gesture Comparison Results

The optimization problems were solved using the IBM's CPLX Optimizer from the NEOS Server (Czyzyk et al., 1998; Dolan, 2001; Gropp & Moré, 1997). Additionally, the hybrid cost coefficients were calculating by setting the constants to $\alpha = 0.01, \beta = 1.01, and\ \gamma = 2$. Figure 4.6 summarizes the gesture usage of the participants in the different tasks. Moreover, Figure 4.7 showcases the matching scores of the gesture matching approaches. The MAGIC-based optimization and Hybrid optimization results presented in Figure 4.7 were obtained using the subtree intersection approach. Namely, the comparisons were made against: 1) the subtree representing the gesture's context (Context Subtree); 2) the subtree representing whether the gesture is being used to represent an object (Exemplify Subtree); 3) the subtree representing the gesture's shape (Shape Subtree); 4) the subtree encompassing all aspects of the gesture except its context and variable declarations (Predicative Subtree); 5) the entire $\mathbf{\Psi_H}$ *Helper Interpretation Tree*; and 6) the combination of subtrees representing the gesture's meaning (Meaning Subtree). These subtrees were selected empirically based on isolating specific properties of the gestures (e.g. only shape, only semantics, only movement).

When focusing on the difference between the tasks, the matching scores for block assembly task and the ultrasound training tasks were the highest and lowest, respectively. The lower scores for the ultrasound training task were expected, as the task was considerably more complex than the other two: more gestures had to be performed to complete each of subtasks. Although the block assembly task was expected to be more demanding than the origami tasks since as it requires more than double the number of steps to be completed, the results demonstrate the opposite. Nonetheless, a closer examination shows that each step in the block assembly task had relatively the same difficulty. Contrarily, two steps in the origami task were considerably more difficult than the others. Participants performed more errors during these two steps, and were noticeably more frustrated. This increased cognitive demand can be linked to the lower gesture matching scores obtained for the origami task. The collaborators performed more gestures during these two steps. Moreover, the meaning and context of these gestures were similar because the collaborators kept referring to the same instruction in different ways. This increased number of functionally similar gestures in reduced time spans could have led the gesture matching approaches to underperform.

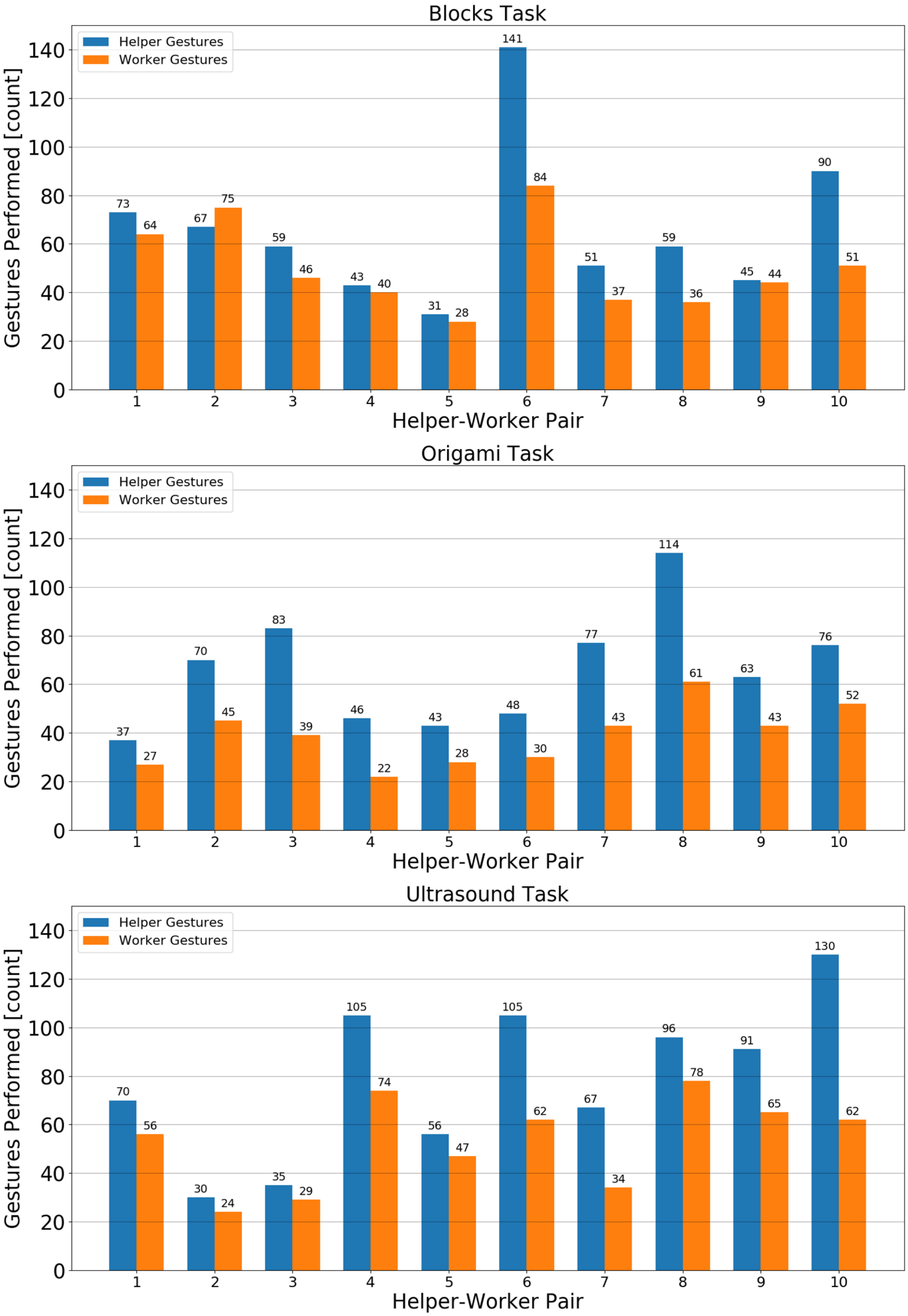

Figure 4.6 Number of gestures performed by the Helper-Worker pairs.

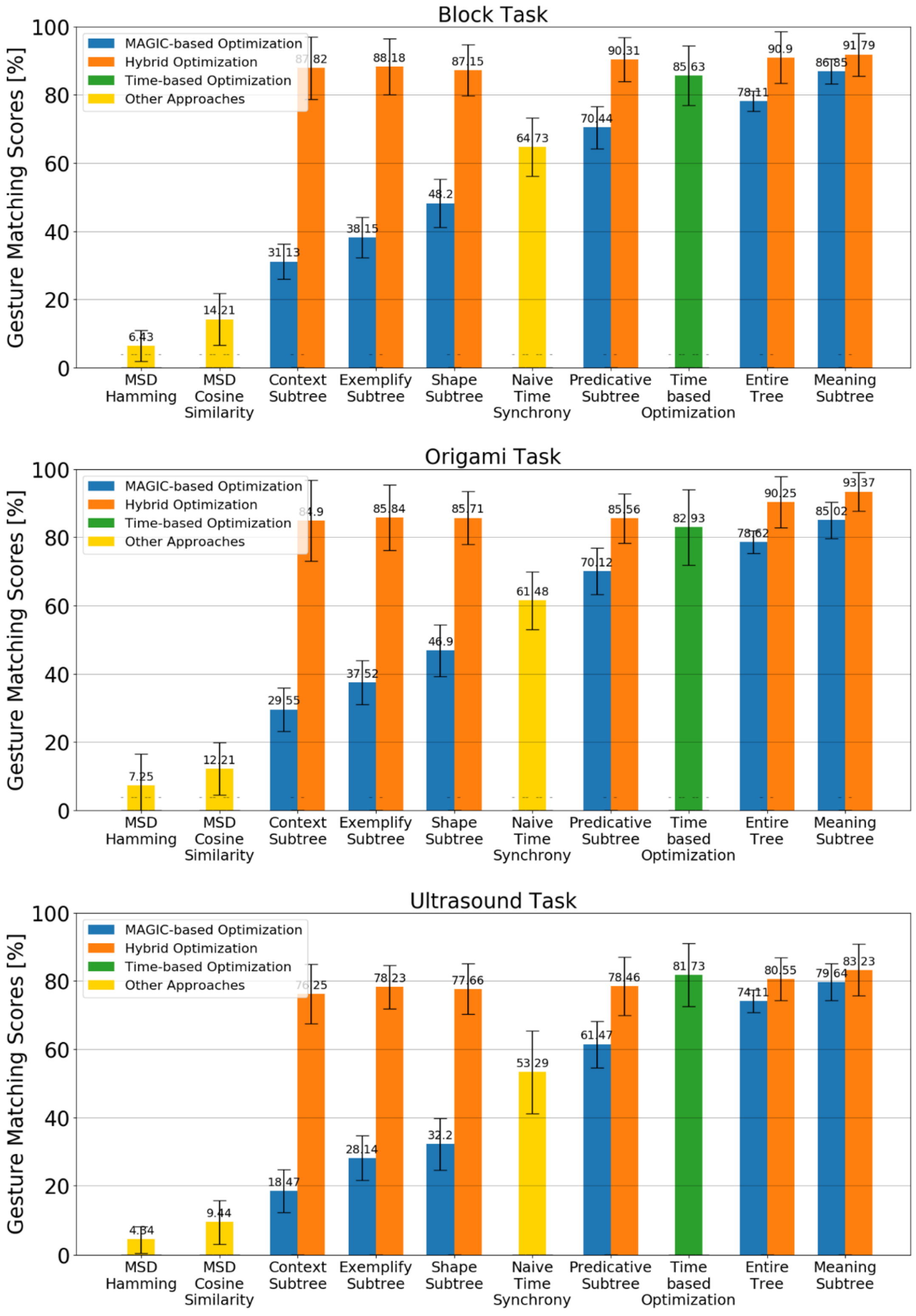

Figure 4.7 Gesture matching scores for the block assembly, origami, and ultrasound training tasks. The scores represent the percentage of agreement of the gesture matching approaches with the human baseline.

The results from Figure 4.7 demonstrate that our MAGIC-based gesture comparison approaches outperform the MSD and NTS gesture comparison baselines in most cases. The results, however, fluctuate significantly based on which subtrees are used to compare the gestures. This demonstrates that selecting which subtree to compare against is key to obtain the highest matching scores when using MAGIC-based optimization, 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 combination of subtrees that presented the highest matching scores was the union between the ExG, LoG, and SyG subtrees of the $\mathbf{\Psi_H}$ Helper *Interpretation Trees*, i.e. $\mathbf{\Psi_H^{ExG \cup LoG \cup SyG}}$; with the CG subtrees of the $\mathbf{\Psi_W}$ Worker *Interpretation Trees*, i.e. $\mathbf{\Psi_W^{CG}}$. The combination of the ExG, LoG, and SyG subtrees contains most of the semantic information of the $\mathbf{\Psi_H}$ Helper *Interpretation Tree*. Further, the information encompassed in the $\mathbf{\Psi_W^{CG}}$ Worker context subtree can be seen as a response to a specific gesture performed by the $\mathbf{\Phi_H}$ *Helper*. Therefore, the information present in a specific context subtree should also be present in another gesture's meaning subtree. In contrast, comparisons based only on the gestures' physical appearance did not obtain high matching scores: the gestures performed by the $\mathbf{\Phi_H}$ *Helper* and the $\mathbf{\Phi_W}$ *Worker* were visually very distinct.

As expected, the results show that gestures performed one after the other are likely to be related. This is confirmed both by our Time-based optimization approach and the NTS approach. Our approach, however, outperforms NTS since it gives higher importance to recently performed gestures instead of creating a time window in which each gesture has equal importance. This is important because, as denoted by Lascarides and Stone (2009), when gestures are performed in succession, the most recent "outscopes" the previous ones: even though the information encompassed by an gesture might be transferred to a new one, it is not possible to respond to a specific old gesture once a new one has been made. However, both approaches find temporal relations between gestures without comparing them content-wise, which is why the MAGIC-based and Hybrid approaches are required.

The hybrid optimization approach presents a new alternative that integrates advantages from the MAGIC-based and Time-based optimization methods. The hybrid optimization approach outperforms the other approaches in almost every case, agreeing with the human baseline over 85% of the times for the block assembly and origami tasks, and over 76% of the times for the ultrasound training task. This approach addresses the problem of requiring a priori knowledge (i.e. which subtree to select) to compare the gestures. The results do not fluctuate significantly based on which subtrees are selected, as opposed to the MAGIC-based approach. Therefore, our hybrid approach is a better and more stable option for gesture comparison.

Finally, Figure 4.8 compares the effect of using our three different methods of generating the $\mathbf{B}$ coefficient matrices of distance costs (i.e. subtree intersections and two different tree kernels). The matrices are computed from the gestures matching scores obtained via the MAGIC-based and Hybrid optimization approaches. The results show that, when using the MAGIC-based optimization, tree kernels led to better gesture matching scores when the size of the subtree selected to compare the gestures was smaller. Contrarily, the scores were worse when comparing against the entire $\boldsymbol{\Psi}$ *Interpretation Tree* and against the $\boldsymbol{\Psi}^{\mathbf{LPG}}$ Predicative Subtree. However, the reduced scores should not be attributed to the number of nodes of the subtrees, but to their attributes and links between them. Smaller subtrees that encompass specific aspects of the gestures (e.g. shape, context) have unique arrangements of nested constituents within them (e.g. a finger can move to upward, forward, rightward). However, when these smaller subtrees are combined, the arrangements of their nested constituents are no longer unique. For instance, the nodes “Direction” and “Forward” can now be referring to the orientation of the hand or its movement. Therefore, the tree kernel approach would give a higher similarity score to gestures that have “Direction Forward” inside their nested constituents, even if they are referring to distinct aspects of the gesture. Thus, the number of False Positives from Equation (4-1) were mostly higher on larger subtrees with common nested constituents. This, subsequently, had a negative impact in the matching scores in the entire $\boldsymbol{\Psi}$ *Interpretation Tree* and the $\boldsymbol{\Psi}^{\mathbf{LPG}}$ Predicative Subtree scenarios. The results, nonetheless, show no evident difference between the approaches when the Hybrid optimization approach is used to calculate the optimal gesture matching.

To improve the gesture matching results, a *combined* approach that selects which gesture matching approach to use based on the size of the subtrees was implemented. A t threshold based on the subtrees’ number of nodes was defined to determine what gesture matching approach should be used. Empirically, a value of $t = 300$ was determined to perform this selection. The green bars in Figure 4.8 represent the results obtained using the *combined* approach. The gesture matching results were equal to the best alternative version, acting as a MAX function without the need to test the other alternatives.

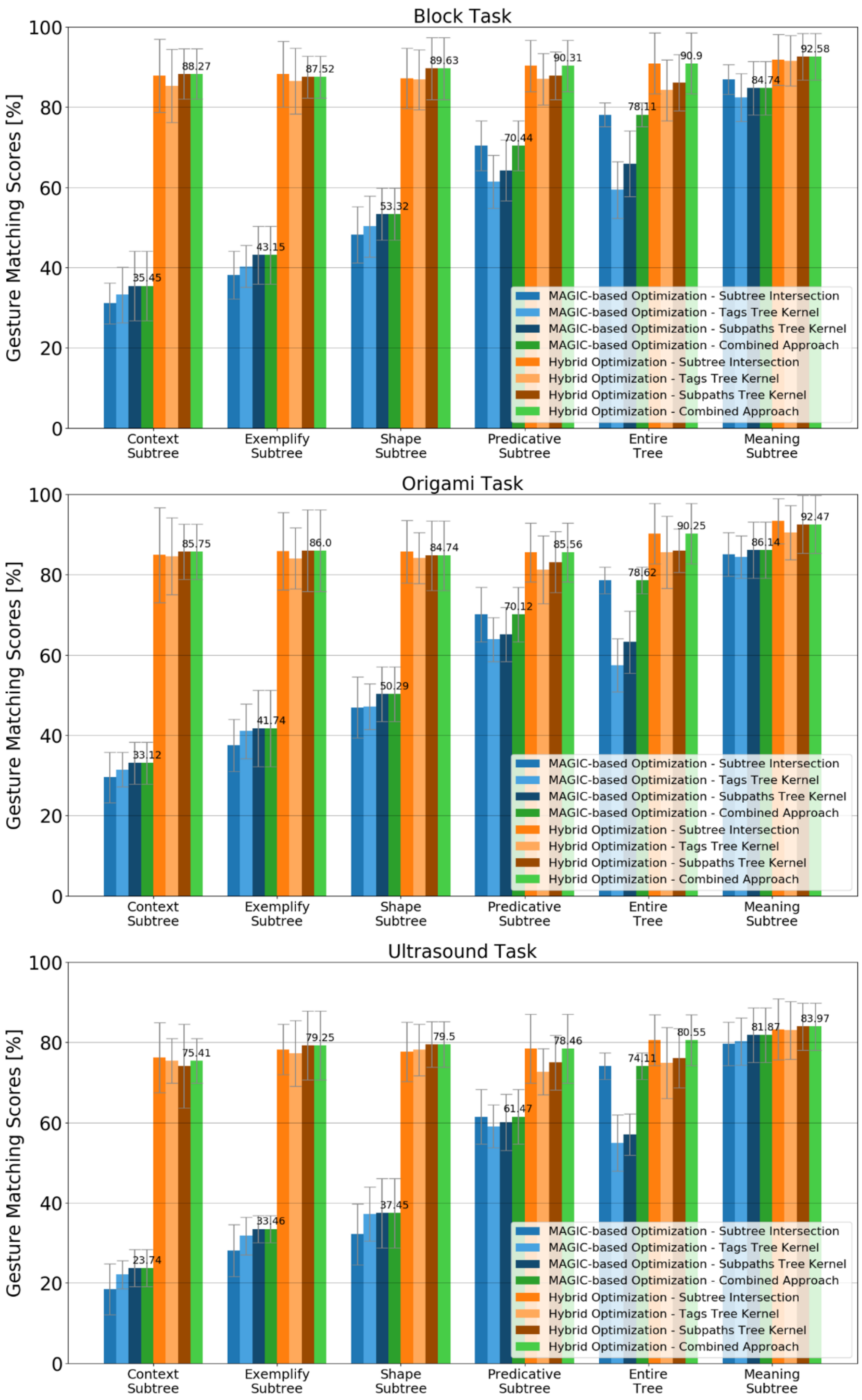

Figure 4.8 Results of gesture matching approaches (subtree intersection, tags tree kernel, subpaths tree kernels, combined) in the matching scores.

4.4 Evaluating Task Understanding Estimation Approaches

Our experimental setup evaluated the hypothesis of whether the quality of assimilation of physical instructions (in the form of our PIA metric) can estimate task understanding. This hypothesis was evaluated by comparing PIA against three other proxy metrics for task understanding: error rate, idle time rate, and task completion percentage (Hoffman, 2013; Martinez-Moyano, 2006). Error rate was calculated as the rate between the instructions in which the $\mathbf{\Phi_W}$ *Worker* performed errors and the total number of instructions. An error was counted each time the $\mathbf{\Phi_W}$ *Worker* picked an incorrect block or assembled blocks incorrectly. Idle time rate is defined as the rate between the time in which the $\mathbf{\Phi_W}$ *Worker* does not perform an action related to the task (e.g. the time spent thinking or asking questions), and the total task completion time, in seconds. Listening to the $\boldsymbol{h_j}$ Helper-authored instructions was not considered as idle time.

Additionally, task completion percentage was calculated in two ways: 1) the rate between the number of vessels found and the total number of vessels in the model (Vessel Detection Completion Percentage; VDCP); and 2) the rate between the number of vessels from which blood was successfully extracted and the total number of vessels in the model (Blood Extraction Completion Percentage; BECP). These metrics were annotated by members of the research team. Afterwards, the Pearson correlation coefficient was computed to analyze the relation between the metrics. Members of the research team annotated the metrics. Afterwards, the Pearson product moment correlation (*r*) was computed to analyze the relationship between the different metrics (Pearson, 1895).

4.4.1 Task Understanding Estimation Results

Figure 4.9 reports the task completion time for each Helper-Worker pair and the total time spent idle or asking for clarification. For example, the first Helper-Worker pair in the origami task completed the task in 4 minutes and 55 seconds (295 seconds total), from which 13 seconds were spent without having a proper understanding of the task. Additionally, Figure 4.10 reports the total number of actions and errors performed by the Helper-Worker pairs. Moreover, Table 4.2, Table 4.3 Table 4.4 present the obtained results for each of the proxy metrics for task understanding.

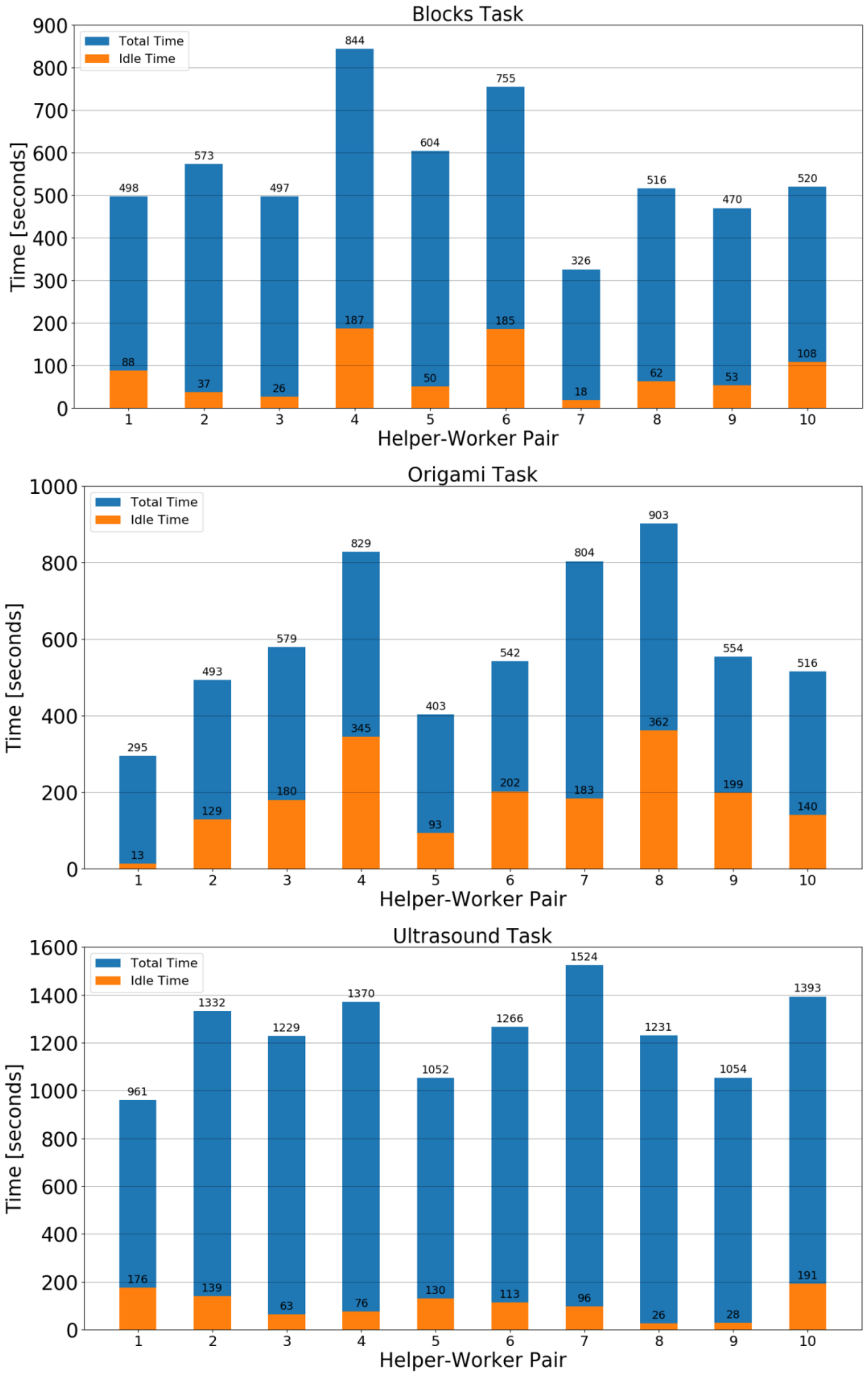

Figure 4.9 Task completion time and Idle time per Helper-Worker pair.

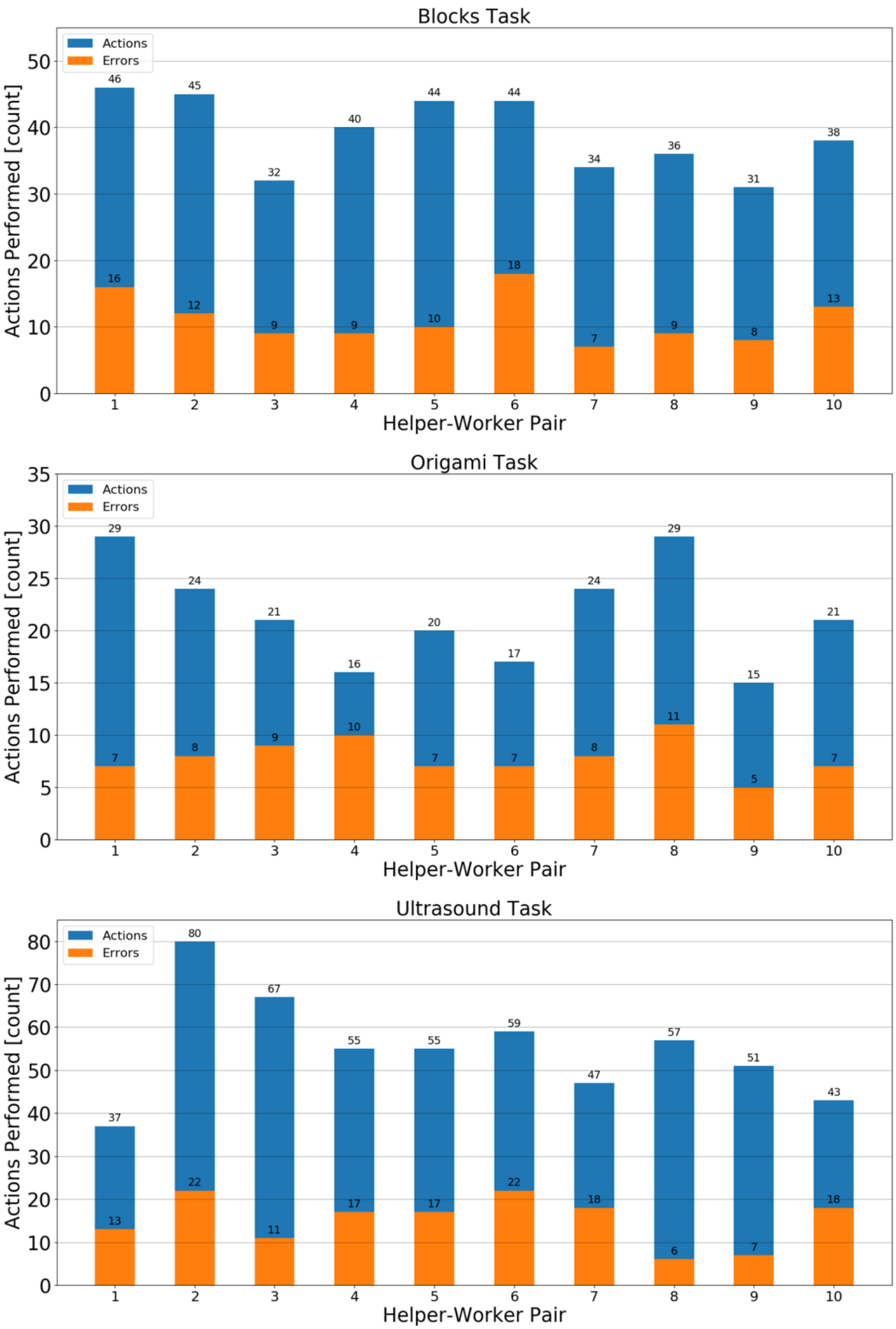

Figure 4.10 Number of actions and errors per Helper-Worker pair.

Table 4.2 Comparison of Task Understanding Proxy Metrics on the Block Assembly Task. Results indicate percentages [%].

H-W Pair	PIA	Error Rate	Idle Time Rate
1	61.30	34.78	17.67
2	72.14	26.67	6.46
3	59.04	28.13	5.23
4	55.43	22.50	22.16
5	66.13	22.73	8.28
6	43.38	40.91	24.50
7	72.55	20.59	5.52
8	51.69	25.00	12.02
9	57.78	25.81	11.28
10	52.04	34.21	20.77

Table 4.3 Comparison of Task Understanding Proxy Metrics on the Origami Task. Results indicate percentages [%].

H-W Pair	PIA	Error Rate	Idle Time Rate
1	72.45	24.13	4.40
2	51.43	33.33	26.16
3	43.37	42.85	31.09
4	44.56	62.50	41.61
5	61.28	35.00	23.08
6	53.12	41.18	37.27
7	53.90	33.33	22.76
8	43.27	37.93	40.09
9	53.97	33.33	35.92
10	64.47	33.33	27.13

Table 4.4 Comparison of Task Understanding Proxy Metrics on the Ultrasound Training Task. Results indicate percentages [%].

H-W Pair	PIA	VDCP	BECP	Error Rate	Idle Time Rate
1	52.86	28.57	28.57	35.14	18.31
2	65.00	85.71	85.71	27.50	10.44
3	75.24	100.00	100.00	16.42	5.13
4	62.22	100.00	100.00	30.91	5.55
5	60.71	42.86	14.29	30.91	12.36
6	55.71	42.86	0.00	37.29	8.93
7	52.24	57.14	42.86	38.30	6.30
8	79.69	100.00	100.00	10.53	2.11
9	72.53	85.71	71.43	13.73	2.66
10	47.69	14.29	14.29	41.86	13.71

Additionally, Figure 4.11 shows the Pearson correlation matrix for the compared metrics. Every comparison reported statistically significant correlations (block assembly correlations = p ≤ 0.05; origami and ultrasound training correlations = p ≤ 0.04) between PIA and all the other proxy metrics. This means that when participants spent more time understanding instructions or performed more errors, their assimilation of physical instructions was poor, i.e. performed a higher number of unnecessary gestures. This indicates that gestures can be used to estimate how well a remote physical task is being understood and performed. This finding reinforces the idea that observing someone else's actions can increase understanding in shared tasks (Dekker, 2017). Based on this idea, task performance and understanding can be estimated based on the number of gestures performed by the collaborators: low PIA scores indicate that many unnecessary gestures were performed.

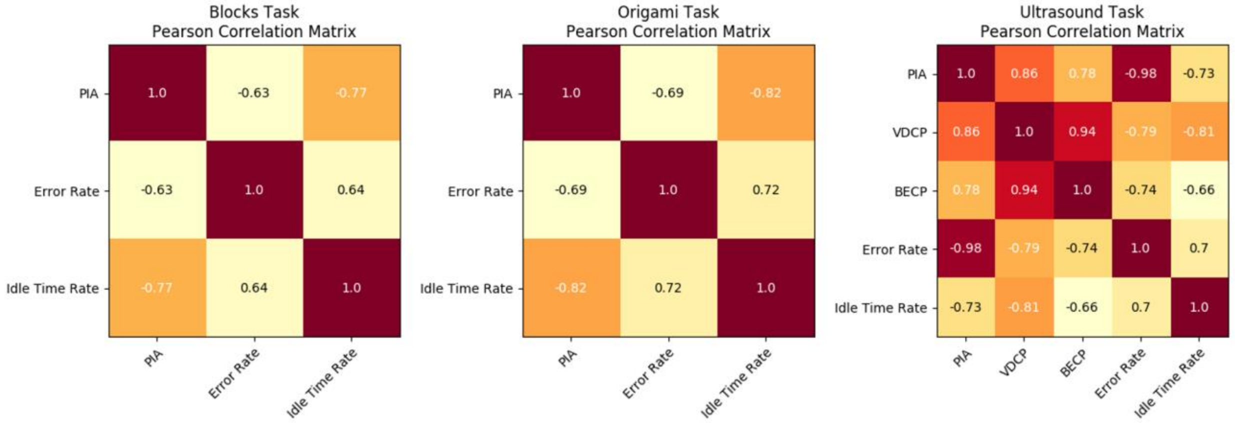

Figure 4.11 Pearson correlation matrix for the task understanding proxy metrics. All metrics reported significant correlations in all the tasks (block assembly correlations = p ≤ 0.05; origami and ultrasound training correlations = p ≤ 0.04).

Moreover, PIA metric was seen to be useful in scenarios where inconsistencies between the other proxy metrics were found. For example, consider the scores obtained by the Helper-Worker pair #4 in the block assembly task (error rate = 22.50, idle time rate = 22.16). This pair reported the lowest error rate, but the second to highest idle rate. This implies that relying only on time-based and error-based task understanding estimators is not enough to make a decision regarding the quality of the understanding of this pair. This is congruent with research showing that spending more time performing and understanding a task can, in some cases, lead to a lower error rate (i.e. higher accuracy) (Albinsson & Zhai, 2003; Chien et al., 2010; Pachella, 1973;

Wickelgren, 1977). However, this pair reported one of the lowest PIA scores. The low PIA score implies that even if the Helper-Worker pair performed very few errors, they did not have proper understanding of the task during their execution: unnecessary physical instructions were exchanged between the collaborators that reflected poor understanding. Another example can be seen: Helper-Worker pair #4 in the ultrasound training task had perfect completion percentages, but an average error rate and PIA. A closer inspection revealed a critical mistake made by this pair: they reported finding 9 vessels, even though the model only had 7 vessels. Afterwards, this pair performed several unnecessary gestures and errors trying to locate these extra vessels again during the blood extraction task. PIA, therefore, can be considered as a tiebreaker approach in scenarios in which other proxy metrics are inconsistent: the task understanding from the Helper-Worker pair #4 can be classified as low.

It should be noted, however, that although the PIA metric provides a novel way to estimate task understanding, this metric should not be used in isolation from the other proxy metrics. We envision PIA to be a complementary metric rather than a replacement for common proxy metrics. For example, consider a hypothetical example of a Helper-Worker pair that achieved a 100 PIA score: every $\boldsymbol{A}_{\mathbf{H}}$ *Instruction* translated into one and only one $\boldsymbol{A}_{\mathbf{W}}$ *Execution*. While this represents perfect assimilation of physical instructions, it does not capture whether the $\boldsymbol{A}_{\mathbf{W}}$ *Executions* were correct: the pair could have performed errors along the way and therefore have a non-zero error rate. This example showcases how these proxy metrics should be used in combination, as each of them captures different aspects of the understanding experienced by the individuals.

Additionally, the PIA metric offers interesting insights regarding the overall quality of the understanding. For example, distinctions between “good”, “decent”, and “bad” assimilation of physical instructions (hence, task understanding) could be obtained by setting thresholds based in the PI+ scores. Elaborating, if a threshold of 70 were to be established as an indicator of “decent” task understanding, only a sixth of the Helper-Worker pairs would have achieved this goal. Another advantage of the PIA metric is its generalizability. The PIA metric requires a **E** matrix of $\boldsymbol{e_{ij}}$ edge weights to generate a proxy score for task understanding, but is agnostic to the framework utilized to create the **E** matrices. In this work, we used the MAGIC architecture to generate this **E** matrix. Nonetheless, the PIA scores could also be generated from the **E** matrices generated from other gesture matching approaches.

One known limitation of the PIA metric has to do with instructions that require more than one gesture to be completed. For instance, if the wrench that Hanna pointed to would have been under a pile of tools, Walter would have been required to perform the action of moving the other tools first before taking the wrench. 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. Therefore, the previous example would have been penalized in the PIA metric due to not having a one-to-one gesture matching. While this assumption can be understood in the light of Clark and colleagues' framework (Clark & Brennan, 1991; Clark & Schaefer, 1987; Sacks et al., 1978), there are scenarios where more than one gesture is necessary to complete an instruction.

4.5 Evaluating Perceived Workload

As mentioned, additional insights of the participants' understanding and performance were obtained via an understanding assessment questionnaire (UAQ) and the NASA TLX. After compiling the answers to these questionnaires, the Pearson correlation coefficient was computed to analyze the relationship between the PIA metric and the answers to both the understanding assessment questionnaire and the NASA TLX. Correlations were calculated separately for both the $\boldsymbol{\Phi}_{\mathbf{H}}$ *Helper* and the $\boldsymbol{\Phi}_{\mathbf{W}}$ *Worker*.

The understanding assessment questionnaire consisted of 8 Likert scale questions (5 = Strongly Agree to 1 = Strongly Disagree) evaluating the overall understanding participants had during the task. Two out of these 8 questions varied depending on the participant's role (i.e. $\boldsymbol{\Phi}_{\mathbf{H}}$ *Helper* and $\boldsymbol{\Phi}_{\mathbf{W}}$ *Worker*) and the performed task (e.g. block assembly, origami, ultrasound training). The questions included: "I was able to understand the task" (UAQ1), "I was able to understand the verbal instructions given by the other person" (UAQ2); "I was able to understand the gestural instructions given by the other person" (UAQ3), "I was able to understand the actions performed by the other person" (UAQ4); "I was able to determine if the other person was understanding me" (UAQ7), and "I feel I could guide the task again with minimal to no challenge" (UAQ8). Questions 5 and 6 differed with respect to the participant's role. For the $\boldsymbol{\Phi}_{\mathbf{H}}$ *Helper* role UAQ5 was "I was able to understand the questions that were asked to me, if any", and UAQ6 was "I was able to understand when mistakes happened along the task, if any". For the $\boldsymbol{\Phi}_{\mathbf{W}}$ *Worker* role, questions UAQ5 and UAQ5 differed based on the type of task. For the block assembly task, UAQ5 was "I was able to understand which blocks the other person was referring to", and UAQ6 was "I was able to understand how to connect the blocks". For the origami task, UAQ5 was "I was able to understand which part of the paper the other person was referring to", and UAQ6 was "I was able to understand how to fold the paper". For the ultrasound training task, UAQ5 was "I was able to understand which part of the ultrasound phantom the other person was referring to", and UAQ6 was "I was able to understand how to find the vessels".

Subsequently, the NASA TLX evaluates perceived workload using six criteria: mental demand (TLX1), physical demand (TLX2), temporal demand (TLX3), perceived performance (TLX4), effort required (TLX5), and generated frustration (TLX6). Each of these criteria is represented with a 21-level Likert Scale question. Higher TLX scores indicate higher task load. The participants' answers to both the understanding assessment questionnaire and the NASA TLX are presented as a normalized value.

Figure 4.12 presents the normalized UAQ answers for the three user experiments, divided based on the participants' role. Likewise, Figure 4.13 presents the normalized TLX answers for the tasks (block assembly, origami, ultrasound training), divided based on the participants' role. Finally, Figure 4.14 presents the correlations between the PIA metric and all the UAQ and TLX questions, divided based on the participants' role and the task performed.

The correlations between the PIA metric and questionnaires are depicted in Figure 4.14 revealed interesting insights. With respect to the ultrasound training task, significant negative correlations were found between the $\mathbf{\Phi_W}$ *Workers*' PIA scores and their perceived performance and frustration. This means that when $\mathbf{\Phi_W}$ *Workers* reported being frustrated and unsatisfied with their performance, they tended to receive lower PIA scores. A possible explanation for this is how frustration levels can increase due to performing errors and correcting them (Évain et al., 2016). Taking heed to this trend can be beneficial during collaborative tasks: if unnecessary gestures in a physical task indicate frustration, bad performances could be predicted in advance (Fillauer et al., 2020; Haraldsen et al., 2019). Moreover, significant positive correlations were found between the PIA metric and several questions in the understanding assessment questionnaire. These correlations also hinted at the relation between the assimilation of physical instructions and task understanding: $\mathbf{\Phi_W}$ *Workers* that reported an overall better understanding of the task tended to receive higher PIA scores. Contrarily, no correlations were found between the $\mathbf{\Phi_H}$ *Helpers*' PIA scores and their responses to the questionnaires. Reduced levels of engagement due to the remote nature of the task could be a possible explanation for the lack of correlations (Fruchter & Cavallin, 2011). A similar correlation was seen in the origami task in the case of the $\mathbf{\Phi_H}$ *Helpers*: they tended to receive lower PIA scores whenever they reported higher levels of frustration. Additionally, the Helper-Worker pairs received higher PIA scores whenever they considered that were able to understand their collaborators' questions better (UAQ5). There were, however, no significant correlations between PIA and any of the questionnaire answers for the block assembly tasks. The reduced cognitive demand that the participants reported during this task, as reported in Figure 4.12 and Figure 4.13, is a possible explanation for this trend.

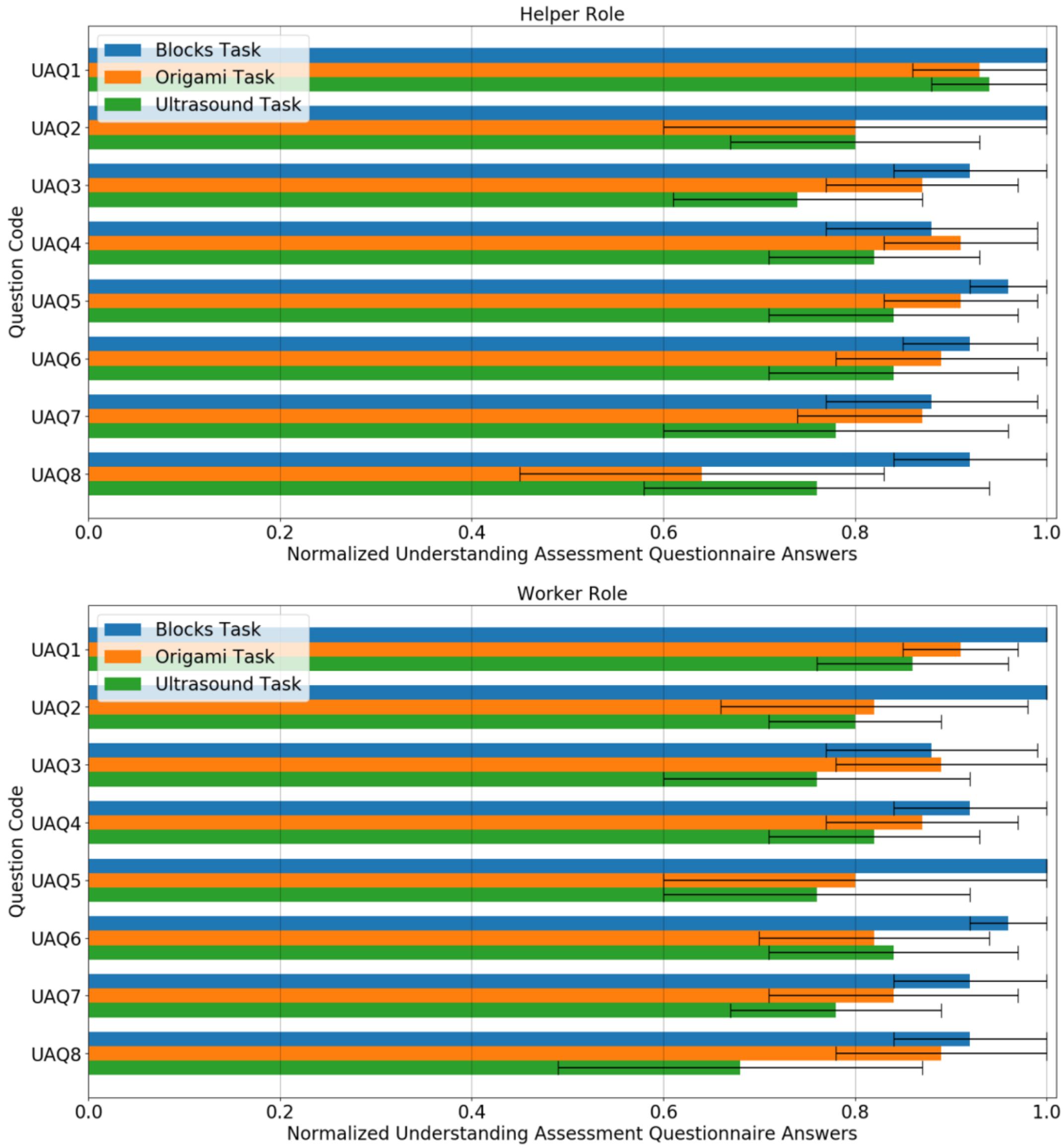

Figure 4.12 Normalized understanding assessment questionnaire answers, divided based on the participants' role.

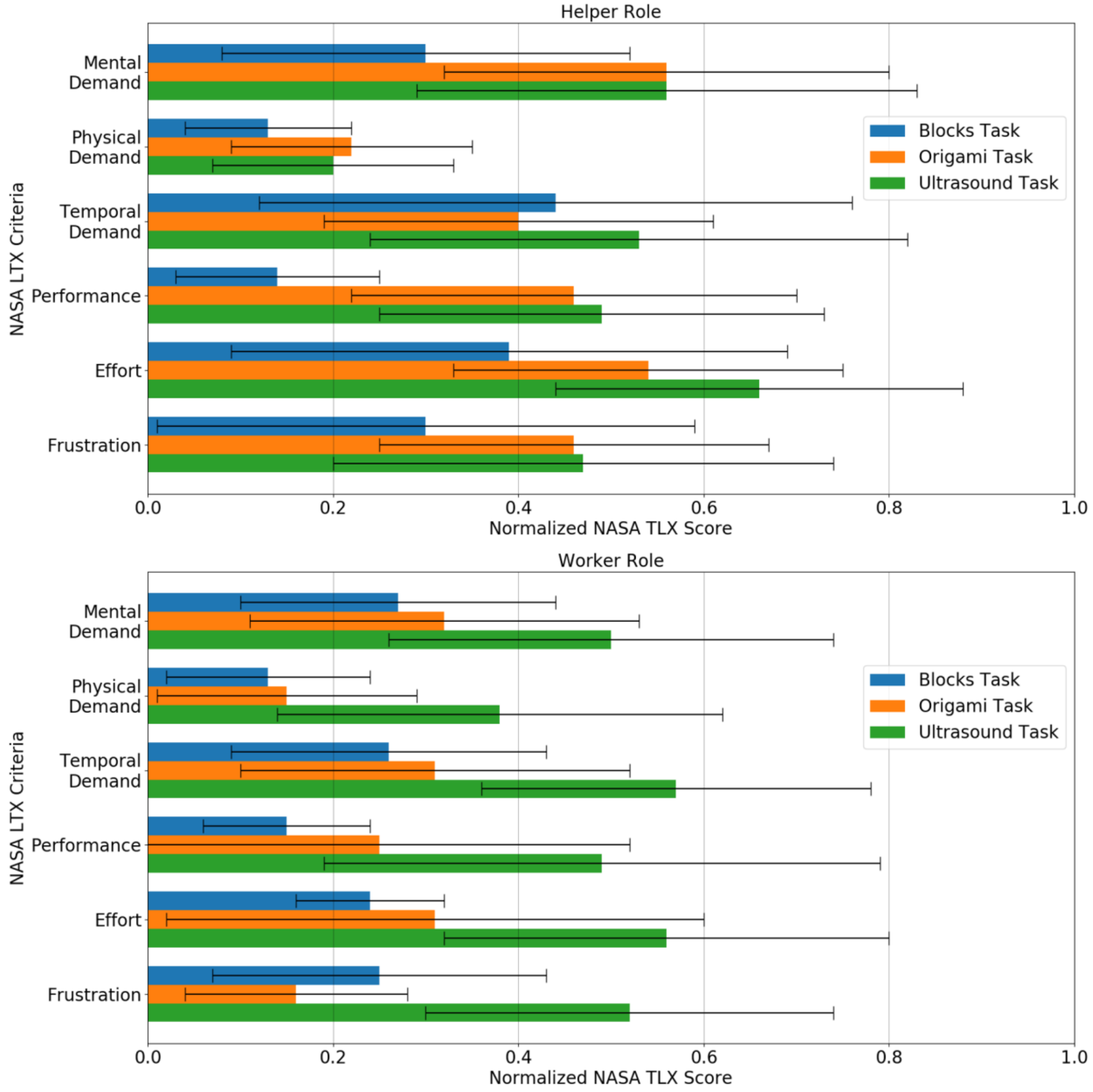

Figure 4.13 Normalized NASA TLX answers, divided based on the participants' role.

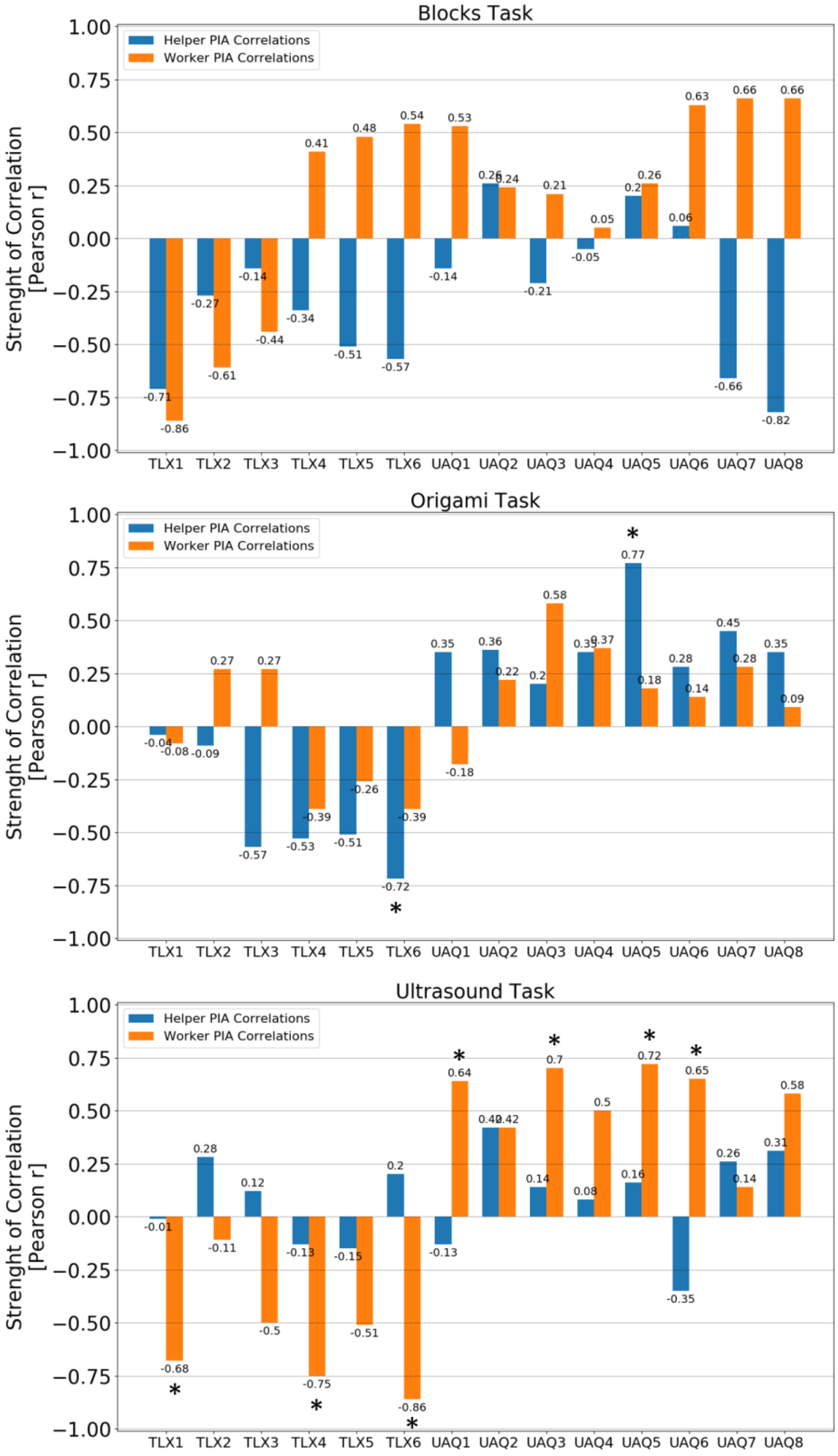

Figure 4.14 Correlations between the PIA metric and all the UAQ and TLX questions, divided based on the participants' role and the task performed. An asterisk represents statistical significance between PIA and the respective criterion ($p \leq 0.05$)

4.6 Summary

This chapter presented experiments and data collection methods to validate the proposed approach to assess collaborative physical via gestural analysis. Results addressing each of the research questions indicate the potential of the proposed research approach. RQ1 and RQ2 are answered by the gesture matching scores obtained in our experiment. RQ3 is also answered by our metric proxy metric for task understanding based on gestural analysis.

5. CONCLUSIONS AND FUTURE WORK

This research led to the development of a novel approach to estimate task performance and understanding through gesture-based analyses. To accomplish this, a two-stage approach was developed. The first stage of this research leveraged MAGIC, a framework to represent, compare, and assess gestures' morphology, semantics, and pragmatics, as opposed to traditional approaches that rely mostly on the gestures' physical appearance. The framework relies on a gestural taxonomy classification, a dynamic semantics framework, and a constituency parsing to express the gestures performed by collaborators as a data structure. Based on these data structures, the second stage of this research defined a metric to assess the quality of assimilation of physical instructions. This metric can act as a proxy metric for task understanding based on gestural analysis.

Our framework was evaluated through three user studies in which participants remotely completed one of three shared tasks: block assembly, paper folding, and ultrasound training. These user studies evaluated both our approach to compare gestures, and our approach to estimate task understanding based on the assimilation of physical instructions. The results revealed that our gesture matching approach reflected human-annotated gesture matchings better than two other gesture matching techniques. Moreover, we found significant correlations between our PIA metric and three other standard metrics to estimate task understanding. These correlations indicate that our proposed metric can act as a good task understanding estimator. Thereby, the approach presented in this research acts as a first step towards assessing task understanding in physical collaborative scenarios through the analysis of gestures.

The findings in this research have some limitations. First, our metric does not consider verbal-only utterances. For instance, a verbal instruction indicating where to connect a block will not be considered when computing similarity if the instruction was not accompanied by a gesture. This implies that our approach will not act as an adequate proxy metric to measure understanding in tasks where the $\mathbf{\Phi_H}$ *Helper* decides not to accompany the instructions with gestures, or in tasks that do not involve physical instructions. Examples of such tasks can include mathematical problem-solving tasks or memory tasks, where the performance does not necessary depends on the physical actions (although embodiment theories indicate that even in those cases, physical action leads to better task performance (Aussems & Kita, 2019; Yeo & Tzeng, 2020).

Another limitation of the approach has to do with redundancy in gestures. While redundancy can be useful in communication, it represents a less effective information transfer medium (Hsia, 1977). Hsia expanded on this issue: "*The ideal* (information transfer) *would be the elimination of redundancy, so that information processed through any physical or physiological channel could be maximized to the limit of the capacity, thereby minimizing the effort and cost involved in information transference*." (1977, p. 64). Although Hsia also comments that the complete elimination of redundancy in communication is practically unattainable (1977, p. 64), our approach was designed to introduce penalties where redundant or unnecessary gestures were used.

The approach is also limited by the range of non-verbal cues that are considered to generate the **Ψ** *Interpretation Trees*. Currently, this methodology only considers hand gestures to estimate task understanding. While this makes sense under the light of the current work, there are other non-verbal cues that may have an effect in collaboration. For instance, intent and emotion can be inferred from body movements and posture (Solanas et al., 2020); engagement can be estimated via facial expressions (Zhang et al., 2020); and attention and interest from gaze (A. F. de C. Hamilton, 2016). Nonetheless, the integration of body, face and gaze cues in our approach only requires the addition of supplementary subtrees into the **Ψ** *Interpretation Tree* structure.

In addition, the approach currently does not work in real-time, as it depends on manual annotation of the data. This is critical for its widespread adoption and introduces a bottleneck in the possible applications of this framework. Currently, significant time is spent in the manual annotation of all the gestures performed by the participants as they perform the shared tasks. Machine learning techniques could be integrated to our framework to address such constraints. For instance, image captioning techniques with context attention could be incorporated to automatically obtain the gestures' pragmatics (Cornia et al., 2018). Likewise, speech-to-text routines could be included to obtain the verbal context of the task (Chung et al., 2019).

Finally, fuzzy logic could be integrated to assist with the extension of MAGIC to more realistic settings (Novák et al., 2012). For instance, real values between 0 and 1 could be used to represent gesture matchings in **E** matrices instead of the Boolean values of 0 or 1, describing the probability a gesture has of being matched to another. Although implementing fuzzy logic would require changing our approaches to compare, match, and use gestures to estimate task understanding, it would allow us to represents human interactions in a more naturalistic fashion (i.e. gestures could be matched in several possible ways, instead of in an unique way).

5.1 Applications of Task Understanding Evaluation via Gestures

The results of this research can have a positive impact in the way remote tasks are performed and assessed. For instance, ultrasound training can be improved by including gesture-related metrics to assess trainees. Portable ultrasound devices are being integrated into telementoring platforms to provide remote assistance in austere regions (Kirkpatrick et al., 2017; McBeth et al., 2010, 2011). In these contexts, the users are expected to perform ultrasound tasks fast and accurately (i.e. low error rates and the idle time rates). A gesture-based criterion to assess ultrasound tasks could capture performance aspects that are ignored by other metrics, leading to more reliable assessments of the training.

Additionally, recent crises such as the COVID-19 pandemic have revealed the necessity of developing more reliable approaches to perform work remotely. This, however, poses a challenge in the way task understanding is currently assessed. For instance, the collaborators will not be able to see their non-verbal cues, which will hinder the collaboration process. Therefore, developing a novel approach to estimate task understanding based on gestures can alleviate this lack of physical co-presence by evaluation task understanding from currently not consider by other common metrics of assessment.

5.2 Applications beyond Human-to-Human Collaboration

This research explored how gestures can be used to assess the collaboration between human agents. However, an extension could be performed to apply this framework to other types of agents, namely robotic assistants and virtual avatars. Such an extension should encompass a redefinition of the morphology section of our **Ψ** *Interpretation Trees*, as robots and virtual avatars morphology could greatly differ from that of a human. Moreover, the range of communication cues (e.g. gaze, facial expressions, turn-taking intention) that need to be included to represent the collaboration's context in situations with non-human agents should also be specified (de Coninck et al., 2019; Zhou & Wachs, 2018).

For example, gestures have been used to allow diver to communicate with underwater robots (Islam et al., 2019). In these contexts, machine learning techniques are used to recognize the diver's commands based on the shape of the hands. Alternatively, MAGIC could be used to model the diver's actions. Specifically, the robot could leverage the **Ψ** *Interpretation Trees* to represent the meaning and context of the diver's gestures. This would allow to predict the diver's instructions even when the gestures used to convey them were not seen by the robots during its training process. Additionally, Saunderson and Nejat surveyed how the nonverbal behaviors of robots can influence human behavior (2019). Although their studies show that the robot's gaze, gestures, facial expressions, and body movements do influence humans, the effect of such nonverbals cues in human performance has not been explored in depth. Instead, the PIA metric could be leveraged to estimate understanding in such tasks.

5.3 Applications in Artificial Intelligence

As discussed in the previous subsection, one of the possible applications of the MAGIC framework is the classification of unseen gestures by a machine learning algorithm. Similar problems have been tackled lately under the zero-shot learning paradigm (Madapana & Wachs, 2020; Thomason & Knepper, 2016). Such approaches represent the gestures using semantic descriptors such as direction of motion, shape of the hand, among others. Since some of the predicates in the MAGIC architecture were inspired on semantics descriptors, the **Ψ** *Interpretation Trees* could be used to perform zero-shot learning gestural classification.

The area of robot coaching could also benefit from this research. This field studies a robot's performance of a task via human physical correction. Recently, approaches to modify the kinematic behavior of a robot via Dynamic Movement Primitives (DMP) have been proposed (Papageorgiou et al., 2020; Talignani Landi et al., 2019). Such DMPs are extracted from direct contact between robot and human. Instead, our MAGIC architecture could be used to encode the information present in the DMPs, allowing coaching without direct contact with the robot. This can in smart factories where the presence of humans is minimized.

REFERENCES

Akkil, D., James, J. M., Isokoski, P., & Kangas, J. (2016). GazeTorch: Enabling gaze awareness in collaborative physical tasks. *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, 1151–1158.

Albinsson, P.-A., & Zhai, S. (2003). High precision touch screen interaction. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 105–112.

Alem, L., Tecchia, F., & Huang, W. (2011). HandsOnVideo: Towards a gesture based mobile AR system for remote collaboration. In *Recent trends of mobile collaborative augmented reality systems* (pp. 135–148). Springer.

Alibali, M. W., Bassok, M., Solomon, K. O., Syc, S. E., & Goldin-Meadow, S. (1999). Illuminating mental representations through speech and gesture. *Psychological Science*, *10*(4), 327–333.

Alibali, M. W., Kita, S., & Young, A. J. (2000). Gesture and the process of speech production: We think, therefore we gesture. *Language and Cognitive Processes*, *15*(6), 593–613.

Amini, R., Kartchner, J. Z., Stolz, L. A., Biffar, D., Hamilton, A. J., & Adhikari, S. (2015). A novel and inexpensive ballistic gel phantom for ultrasound training. *World Journal of Emergency Medicine*, *6*(3), 225.

Argyle, M. (2013). *Bodily communication*. Routledge.

Asher, N., & Lascarides, A. (2003). *Logics of conversation*. Cambridge University Press.

Auer, P. (2005). Projection in interaction and projection in grammar. *Text-Interdisciplinary Journal for the Study of Discourse*, *25*(1), 7–36.

Aussems, S., & Kita, S. (2019). Seeing iconic gestures while encoding events facilitates children's memory of these events. *Child Development*, *90*(4), 1123–1137.

Baker, D. P., & Dismukes, R. K. (2002). A framework for understanding crew performance assessment issues. *The International Journal of Aviation Psychology*, *12*(3), 205–222.

Barmby, P., Harries, T., Higgins, S., & Suggate, J. (2007). How can we assess mathematical understanding. *Proceedings of the 31st Conference of the International Group for the Psychology of Mathematics Education*, *2*, 41–48.

Beilock, S. L., & Goldin-Meadow, S. (2010). Gesture changes thought by grounding it in action. *Psychological Science*, *21*(11), 1605–1610.

Bracewell, R. N., & Bracewell, R. N. (1986). *The Fourier transform and its applications* (Vol. 31999). McGraw-Hill New York.

Brennan, S. E. (1998). The grounding problem in conversations with and through computers. *Social and Cognitive Approaches to Interpersonal Communication*, 201–225.

Brinkmann, A. (2003). Graphical knowledge display–mind mapping and concept mapping as efficient tools in mathematics education. *Mathematics Education Review*, *16*(4), 35–48.

Broaders, S. C., Cook, S. W., Mitchell, Z., & Goldin-Meadow, S. (2007). Making children gesture brings out implicit knowledge and leads to learning. *Journal of Experimental Psychology: General*, *136*(4), 539.

Butterworth, B., & Hadar, U. (1989). *Gesture, speech, and computational stages: A reply to McNeill.*

Cabrera, M. E., Novak, K., Foti, D., Voyles, R., & Wachs, J. P. (2017). What makes a gesture a gesture? Neural signatures involved in gesture recognition. *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, 748–753.

Calvo-Merino, B., Glaser, D. E., Grèzes, J., Passingham, R. E., & Haggard, P. (2004). Action observation and acquired motor skills: An FMRI study with expert dancers. *Cerebral Cortex*, *15*(8), 1243–1249.

Caridakis, G., Karpouzis, K., Drosopoulos, A., & Kollias, S. (2010). SOMM: Self organizing Markov map for gesture recognition. *Pattern Recognition Letters*, *31*(1), 52–59.

Carpendale, J., & Lewis, C. (2006). *How children develop social understanding*. Blackwell Publishing.

Chien, J. H., Tiwari, M. M., Suh, I. H., Mukherjee, M., Park, S.-H., Oleynikov, D., & Siu, K.-C. (2010). Accuracy and speed trade-off in robot-assisted surgery. *The International Journal Of Medical Robotics and Computer Assisted Surgery*, *6*(3), 324–329.

Chiu, C.-C., & Marsella, S. (2014). Gesture generation with low-dimensional embeddings. *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems*, 781–788.

Chu, M., & Kita, S. (2011). The nature of gestures' beneficial role in spatial problem solving. *Journal of Experimental Psychology: General*, *140*(1), 102.

Chung, Y.-A., Weng, W.-H., Tong, S., & Glass, J. (2019). Towards unsupervised speech-to-text translation. *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7170–7174.

Church, R. B., & Goldin-Meadow, S. (1986). The mismatch between gesture and speech as an index of transitional knowledge. *Cognition*, *23*(1), 43–71.

Clark, H. H. (1996). *Using language*. Cambridge university press.

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. *Perspectives on Socially Shared Cognition*, *13*(1991), 127–149.

Clark, H. H., & Marshall, C. R. (2002). Definite reference and mutual knowledge. *Psycholinguistics: Critical Concepts in Psychology*, *414*.

Clark, H. H., & Schaefer, E. F. (1987). Collaborating on contributions to conversations. *Language and Cognitive Processes*, *2*(1), 19–41.

Converse, S., Cannon-Bowers, J., & Salas, E. (1993). Shared mental models in expert team decision making. *Individual and Group Decision Making: Current Issues*, *221*.

Cook, S. W., Mitchell, Z., & Goldin-Meadow, S. (2008). Gesturing makes learning last. *Cognition*, *106*(2), 1047–1058.

Cook, S. W., & Tanenhaus, M. K. (2009). Embodied communication: Speakers' gestures affect listeners' actions. *Cognition*, *113*(1), 98–104.

Cook, S. W., Yip, T. K., & Goldin-Meadow, S. (2012). Gestures, but not meaningless movements, lighten working memory load when explaining math. *Language and Cognitive Processes*, *27*(4), 594–610.

Cornia, M., Baraldi, L., Serra, G., & Cucchiara, R. (2018). Paying more attention to saliency: Image captioning with saliency and context attention. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, *14*(2), 1–21.

Czyzyk, J., Mesnier, M. P., & Moré, J. J. (1998). The NEOS Server. *IEEE Journal on Computational Science and Engineering*, *5*(3), 68–75.

de Coninck, F., Yumak, Z., Sandino, G., Veltkamp, R., & CleVR, B. (2019). Non-Verbal Behavior Generation for Virtual Characters in Group Conversations. *2019 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*, 41–418.

Dekker, S. (2017). *The field guide to understanding'human error'*. CRC press.

Dipietro, L., Sabatini, A. M., & Dario, P. (2008). A survey of glove-based systems and their applications. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *38*(4), 461–482.

Dolan, E. D. (2001). *The NEOS Server 4.0 Administrative Guide* (Technical Memorandum ANL/MCS-TM-250). Mathematics and Computer Science Division, Argonne National Laboratory.

Ebert, C., & Ebert, C. (2014). Gestures, demonstratives, and the attributive/referential distinction. *Handout of a Talk given at Semantics and Philosophy in Europe (SPE 7)*.

Espinosa, A., Kraut, R., Lerch, J., Slaughter, S., Herbsleb, J., & Mockus, A. (2001). Shared mental models and coordination in large-scale, distributed software development. *ICIS 2001 Proceedings*, 64.

Espinosa, A., Kraut, R., Slaughter, S., Lerch, J., Herbsleb, J., & Mockus, A. (2002). Shared mental models, familiarity, and coordination: A multi-method study of distributed software teams. *ICIS 2002 Proceedings*, 39.

Estival, D., & Molesworth, B. (2016). Native English speakers and EL2 pilots. *Aviation English: A Lingua Franca for Pilots and Air Traffic Controllers*, 140.

Évain, A., Argelaguet, F., Strock, A., Roussel, N., Casiez, G., & Lécuyer, A. (2016). Influence of error rate on frustration of BCI users. *Proceedings of the International Working Conference on Advanced Visual Interfaces*, 248–251.

Fakourfar, O., Ta, K., Tang, R., Bateman, S., & Tang, A. (2016). Stabilized annotations for mobile remote assistance. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 1548–1560.

Feyereisen, P. (1987). *Gestures and speech, interactions and separations: A reply to McNeill (1985).*

Fillauer, J. P., Bolden, J., Jacobson, M., Partlow, B. H., Benavides, A., & Shultz, J. N. (2020). Examining the effects of frustration on working memory capacity. *Applied Cognitive Psychology*, *34*(1), 50–63.

Frick-Horbury, D. (2002). The use of hand gestures as self-generated cues for recall of verbally associated targets. *The American Journal of Psychology*.

Fruchter, R., & Cavallin, H. (2011). Attention and engagement of remote team members in collaborative multimedia environments. In *Computing in Civil Engineering (2011)* (pp. 875–882).

Fussell, S. R., Setlock, L. D., Yang, J., Ou, J., Mauer, E., & Kramer, A. D. (2004). Gestures over video streams to support remote collaboration on physical tasks. *Human-Computer Interaction*, *19*(3), 273–309.

Gauglitz, S., Lee, C., Turk, M., & Höllerer, T. (2012). Integrating the physical environment into mobile remote collaboration. *Proceedings of the 14th International Conference on Human-Computer Interaction with Mobile Devices and Services*, 241–250.

Ge, S. S., Yang, Y., & Lee, T. H. (2008). Hand gesture recognition and tracking based on distributed locally linear embedding. *Image and Vision Computing*, *26*(12), 1607–1620.

Giorgolo, G. (2010). A formal semantics for iconic spatial gestures. In *Logic, Language and Meaning* (pp. 305–314). Springer.

Giorgolo, G. (2011). Integration of gesture and verbal language: A formal semantics approach. *International Gesture Workshop*, 216–227.

Glover, F. (1967). Maximum matching in a convex bipartite graph. *Naval Research Logistics Quarterly*, *14*(3), 313–316.

Goldin-Meadow, S. (2005). *Hearing gesture: How our hands help us think*. Harvard University Press.

Goldin-Meadow, S. (2006). Talking and thinking with our hands. *Current Directions in Psychological Science*, *15*(1), 34–39.

Goldin-Meadow, S., & Beilock, S. L. (2010). Action's influence on thought: The case of gesture. *Perspectives on Psychological Science*, *5*(6), 664–674.

Goldin-Meadow, S., & Sandhofer, C. M. (1999). Gestures convey substantive information about a child's thoughts to ordinary listeners. *Developmental Science*, *2*(1), 67–74.

Gonzalez, G., Madapana, N., Taneja, R., Zhang, L., Rodgers, R., & Wachs, J. P. (2018). Looking Beyond the Gesture: Vocabulary Acceptability Criteria for Gesture Elicitation Studies. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *62*, 997–1001.

Goodwin, C. (2003). The body in action. In *Discourse, the body, and identity* (pp. 19–42). Springer.

Goutte, C., & Gaussier, E. (2005). A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. *European Conference on Information Retrieval*, 345–359.

Gropp, W., & Moré, J. J. (1997). Optimization Environments and the NEOS Server. In M. D. Buhman & A. Iserle (Eds.), *Approximation Theory and Optimization* (pp. 167–182). Cambridge University Press.

Gumperz, J. J. (1992). Contextualization and understanding. *Rethinking Context: Language as an Interactive Phenomenon*, *11*, 229–252.

Hamilton, A. F. de C. (2016). Gazing at me: The importance of social meaning in understanding direct-gaze cues. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1686), 20150080.

Hamilton, A., Wolpert, D., & Frith, U. (2004). Your own action influences how you perceive another person's action. *Current Biology*, *14*(6), 493–498.

Haraldsen, H. M., Solstad, B. E., Ivarsson, A., Halvari, H., & Abrahamsen, F. E. (2019). Change in Basic Need Frustration in Relation to Perfectionism, Anxiety and Performance in Elite Junior Performers. *Scandinavian Journal of Medicine & Science in Sports*.

Harries, T., & Barmby, P. (2008). Representing multiplication. *MATHEMATICS TEACHING-DERBY-*, *206*, 37.

Hart, S. G. (2006). NASA-task load index (NASA-TLX); 20 years later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *50*(9), 904–908.

Heath, C., & Luff, P. (1991). Disembodied conduct: Communication through video in a multi-media office environment. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 99–103.

Hewes, G. W., Andrew, R., Carini, L., Choe, H., Gardner, R. A., Kortlandt, A., Krantz, G. S., McBride, G., Nottebohm, F., Pfeiffer, J., Rumbaugh, D. G., Steklis, H. D., Raliegh, M. J., Stopa, R., Suzuki, A., Washburn, S., & Wescott, R. W. (1973). Primate communication and the gestural origin of language [and comments and reply]. *Current Anthropology*, *14*(1/2), 5–24.

Hoffman, G. (2013). Evaluating fluency in human-robot collaboration. *International Conference on Human-Robot Interaction (HRI), Workshop on Human Robot Collaboration*, *381*, 1–8.

Hoffman, G., & Breazeal, C. (2007). Cost-based anticipatory action selection for human–robot fluency. *IEEE Transactions on Robotics*, *23*(5), 952–961.

Hsia, H. (1977). Redundancy: Is it the lost key to better communication? *AV Communication Review*, *25*(1), 63–85.

Huang, W., & Alem, L. (2013). HandsinAir: A wearable system for remote collaboration on physical tasks. *Proceedings of the 2013 Conference on Computer Supported Cooperative Work Companion*, 153–156.

Huang, W., & Alem, L. (2011). Supporting hand gestures in mobile remote collaboration: A usability evaluation. *Proceedings of the 25th BCS Conference on Human-Computer Interaction*, 211–216.

Huang, W., Kim, S., Billinghurst, M., & Alem, L. (2019). Sharing hand gesture and sketch cues in remote collaboration. *Journal of Visual Communication and Image Representation*, *58*, 428–438.

Humphreys, L. (2005). Cellphones in public: Social interactions in a wireless era. *New Media & Society*, *7*(6), 810–833.

Islam, M. J., Ho, M., & Sattar, J. (2019). Understanding human motion and gestures for underwater human–robot collaboration. *Journal of Field Robotics*, *36*(5), 851–873.

Ju, R., Chang, P. L., Buckley, A. P., & Wang, K. C. (2012). Comparison of Nintendo Wii and PlayStation2 for enhancing laparoscopic skills. *JSLS: Journal of the Society of Laparoendoscopic Surgeons*, *16*(4), 612.

Kannemeyer, L. (2005). Reference framework for describing and assessing students—Understanding in first year calculus. *International Journal of Mathematical Education in Science and Technology*, *36*(2–3), 269–285.

Kemp, E. C., Floyd, M. R., McCord-Duncan, E., & Lang, F. (2008). Patients prefer the method of “tell back-collaborative inquiry” to assess understanding of medical information. *J Am Board Fam Med*, *21*(1), 24–30.

Kendon, A. (2004). *Gesture: Visible action as utterance*. Cambridge University Press.

Kieras, D. E., & Bovair, S. (1984). The role of a mental model in learning to operate a device. *Cognitive Science*, *8*(3), 255–273.

Kim, S., Lee, G., Huang, W., Kim, H., Woo, W., & Billinghurst, M. (2019). Evaluating the Combination of Visual Communication Cues for HMD-based Mixed Reality Remote Collaboration. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–13.

Kirk, D., Crabtree, A., & Rodden, T. (2005). Ways of the hands. *ECSCW 2005*, 1–21.

Kirk, D., Rodden, T., & Fraser, D. S. (2007). Turn it this way: Grounding collaborative action with remote gestures. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1039–1048.

Kirk, D. S., & Fraser, D. S. (2005). The effects of remote gesturing on distance instruction. *Proceedings of Th 2005 Conference on Computer Support for Collaborative Learning: Learning 2005: The next 10 Years!*, 301–310.

Kirk, D., & Stanton Fraser, D. (2006). Comparing remote gesture technologies for supporting collaborative physical tasks. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1191–1200.

Kirkpatrick, A. W., McKee, J. L., McBeth, P. B., Ball, C. G., LaPorta, A., Broderick, T., Leslie, T., King, D., Beatty, H. E. W., Keillor, J., & others. (2017). The Damage Control Surgery in Austere Environments Research Group (DCSAERG): A dynamic program to facilitate real-time telementoring/telediagnosis to address exsanguination in extreme and austere environments. *Journal of Trauma and Acute Care Surgery*, *83*(1), S156–S163.

Knoblich, G., & Sebanz, N. (2006). The social nature of perception and action. *Current Directions in Psychological Science*, *15*(3), 99–104.

Konečnỳ, J., & Hagara, M. (2014). One-shot-learning gesture recognition using hog-hof features. *The Journal of Machine Learning Research*, *15*(1), 2513–2532.

Kopp, S., Tepper, P., & Cassell, J. (2004). Towards integrated microplanning of language and iconic gesture for multimodal output. *Proceedings of the 6th International Conference on Multimodal Interfaces*, 97–104.

Kraut, R. E., Fussell, S. R., & Siegel, J. (2003). Visual information as a conversational resource in collaborative physical tasks. *Human–Computer Interaction*, *18*(1–2), 13–49.

Kraut, R. E., Miller, M. D., & Siegel, J. (1996). Collaboration in performance of physical tasks: Effects on outcomes and communication. *Proceedings of the 1996 ACM Conference on Computer Supported Cooperative Work*, 57–66.

Kurillo, G., Bajcsy, R., Nahrsted, K., & Kreylos, O. (2008). Immersive 3d environment for remote collaboration and training of physical activities. *2008 IEEE Virtual Reality Conference*, 269–270.

Kuzuoka, H., Kosaka, J., Yamazaki, K., Suga, Y., Yamazaki, A., Luff, P., & Heath, C. (2004). Mediating dual ecologies. *Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work*, 477–486.

Lascarides, A., & Stone, M. (2009). A formal semantic analysis of gesture. *Journal of Semantics*, *26*(4), 393–449.

Leekam, S. R., Solomon, T. L., & Teoh, Y.-S. (2010). Adults' social cues facilitate young children's use of signs and symbols. *Developmental Science*, *13*(1), 108–119.

Luff, P., Heath, C., Kuzuoka, H., Hindmarsh, J., Yamazaki, K., & Oyama, S. (2003). Fractured ecologies: Creating environments for collaboration. *Human-Computer Interaction*, *18*(1), 51–84.

Madapana, N., & Wachs, J. (n.d.). Feature Selection for Zero-Shot Gesture Recognition. *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)(FG),* 309–313.

Madapana, N., & Wachs, J. (2017). ZSGL: zero shot gestural learning. *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, 331–335.

Martinez-Moyano, I. (2006). Exploring the dynamics of collaboration in interorganizational settings. *Creating a Culture of Collaboration: The International Association of Facilitators Handbook*, *4*, 69.

Mathieu, J. E., Heffner, T. S., Goodwin, G. F., Salas, E., & Cannon-Bowers, J. A. (2000). The influence of shared mental models on team process and performance. *Journal of Applied Psychology*, *85*(2), 273.

Mavin, T. J., & Dall'Alba, G. (2011). Understanding complex assessment: A lesson from aviation. *Proceedings of the 4th International Conference of Education, Research and Innovation*, 6563–6570.

McBeth, P. B., Crawford, I., Blaivas, M., Hamilton, T., Musselwhite, K., Panebianco, N., Melniker, L., Ball, C. G., Gargani, L., Gherdovich, C., & others. (2011). Simple, almost anywhere, with almost anyone: Remote low-cost telementored resuscitative lung ultrasound. *Journal of Trauma and Acute Care Surgery*, *71*(6), 1528–1535.

McBeth, P. B., Hamilton, T., & Kirkpatrick, A. W. (2010). Cost-effective remote iPhone-teathered telementored trauma telesonography. *Journal of Trauma and Acute Care Surgery*, *69*(6), 1597–1599.

McNeill, D. (1985). So you think gestures are nonverbal? *Psychological Review*, *92*(3), 350.

McNeill, D. (1987). *So you do think gestures are nonverbalp Reply to Feyereisen (1987).*

McNeill, D. (1989). *A straight path—To where? Reply to Butterworth and Hadar*.

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. University of Chicago press.

McNeill, D., Cassell, J., & McCullough, K.-E. (1994). Communicative effects of speech-mismatched gestures. *Research on Language and Social Interaction*, *27*(3), 223–237.

Mintzes, J. J., Wandersee, J. H., & Novak, J. D. (2001). Assessing understanding in biology. *Journal of Biological Education*, *35*(3), 118–124.

Mintzes, J. J., Wandersee, J. H., & Novak, J. D. (2005). *Assessing science understanding: A human constructivist view*. Academic Press.

Mitra, S., & Acharya, T. (2007). Gesture recognition: A survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *37*(3), 311–324.

Mohan, D., Farris, C., Fischhoff, B., Rosengart, M. R., Angus, D. C., Yealy, D. M., Wallace, D. J., & Barnato, A. E. (2017). Efficacy of educational video game versus traditional educational apps at improving physician decision making in trauma triage: Randomized controlled trial. *Bmj*, *359*, j5416.

Molesworth, B. R., & Estival, D. (2015). Miscommunication in general aviation: The influence of external factors on communication errors. *Safety Science*, *73*, 73–79.

Molnar-Szakacs, I., Wu, A. D., Robles, F. J., & Iacoboni, M. (2007). Do you see what I mean? Corticospinal excitability during observation of culture-specific gestures. *PLoS One*, *2*(7), e626.

Montague, R. (1970). Pragmatics and intensional logic. *Synthese*, *22*(1–2), 68–94.

Morris, Charles W. (1964). *Signification and Significance a Study of the Relations of Signs and Values*.

Morris, Charles William. (1938). Foundations of the Theory of Signs. In *International encyclopedia of unified science* (pp. 1–59). Chicago University Press.

Nikolaidis, S., & Shah, J. (2013). Human-robot cross-training: Computational formulation, modeling and evaluation of a human team training strategy. *Proceedings of the 8th ACM/IEEE International Conference on Human-Robot Interaction*, 33–40.

Novák, V., Perfilieva, I., & Mockor, J. (2012). *Mathematical principles of fuzzy logic* (Vol. 517). Springer Science & Business Media.

Ou, J., Fussell, S. R., Chen, X., Setlock, L. D., & Yang, J. (2003). Gestural communication over video stream: Supporting multimodal interaction for remote collaborative physical tasks. *Proceedings of the 5th International Conference on Multimodal Interfaces*, 242–249.

Pachella, R. G. (1973). *The interpretation of reaction time in information processing research*. MICHIGAN UNIV ANN ARBOR HUMAN PERFORMANCE CENTER.

Papageorgiou, D., Kastritsi, T., & Doulgeri, Z. (2020). A passive robot controller aiding human coaching for kinematic behavior modifications. *Robotics and Computer-Integrated Manufacturing*, *61*, 101824.

Parvini, F., McLeod, D., Shahabi, C., Navai, B., Zali, B., & Ghandeharizadeh, S. (2009). An approach to glove-based gesture recognition. *International Conference on Human-Computer Interaction*, 236–245.

Pearson, K. (1895). VII. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, *58*(347–352), 240–242.

Ping, R. M., Goldin-Meadow, S., & Beilock, S. L. (2014). Understanding gesture: Is the listener's motor system involved? *Journal of Experimental Psychology: General*, *143*(1), 195.

Poggi, I. (2008). Iconicity in different types of gestures. *Gesture*, *8*(1), 45–61.

Porter, L. W., Bigley, G. A., Steers, R. M., & others. (2003). *Motivation and work behavior*.

Potts, C. (2005). *The logic of conventional implicatures*. Oxford University Press on Demand.

Radford, L. (2009). Why do gestures matter? Sensuous cognition and the palpability of mathematical meanings. *Educational Studies in Mathematics*, *70*(2), 111–126.

Rautaray, S. S., & Agrawal, A. (2015). Vision based hand gesture recognition for human computer interaction: A survey. *Artificial Intelligence Review*, *43*(1), 1–54.

Reed, C. L., & McGoldrick, J. E. (2007). Action during body perception: Processing time affects self–other correspondences. *Social Neuroscience*, *2*(2), 134–149.

Rojas-Munoz, E., & Wachs, J. P. (n.d.). Beyond MAGIC: Matching Collaborative Gestures using an Optimization-based Approach. *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)(FG)*, 296–303.

Rojas-Muñoz, E., & Wachs, J. P. (n.d.). The MAGIC of E-Health: A Gesture-Based Approach to Estimate Understanding and Performance in Remote Ultrasound Tasks. *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)(FG)*, 314–318.

Rojas-Munoz, E., & Wachs, J. P. (2019). MAGIC: A Fundamental Framework for Gesture Representation, Comparison and Assessment. *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, 1–8.

Roth, W.-M. (2001). Gestures: Their role in teaching and learning. *Review of Educational Research*, *71*(3), 365–392.

Roth, W.-M., & Welzel, M. (2001). From activity to gestures and scientific language. *Journal of Research in Science Teaching*, *38*(1), 103–136.

Sacks, H., Schegloff, E. A., & Jefferson, G. (1978). A simplest systematics for the organization of turn taking for conversation. In *Studies in the organization of conversational interaction* (pp. 7–55). Elsevier.

Saunderson, S., & Nejat, G. (2019). How robots influence humans: A survey of nonverbal communication in social human–robot interaction. *International Journal of Social Robotics*, *11*(4), 575–608.

Schlager, D., Sanders, A. B., Wiggins, D., & Boren, W. (1991). Ultrasound for the detection of foreign bodies. *Annals of Emergency Medicine*, *20*(2), 189–191.

Schlenker, P. (2018). Gesture projection and cosuppositions. *Linguistics and Philosophy*, *41*(3), 295–365.

Sebanz, N., Bekkering, H., & Knoblich, G. (2006). Joint action: Bodies and minds moving together. *Trends in Cognitive Sciences*, *10*(2), 70–76.

Sebanz, N., Knoblich, G., & Prinz, W. (2003). Representing others' actions: Just like one's own? *Cognition*, *88*(3), B11–B21.

Sebanz, N., Knoblich, G., Prinz, W., & Wascher, E. (2006). Twin peaks: An ERP study of action planning and control in coacting individuals. *Journal of Cognitive Neuroscience*, *18*(5), 859–870.

Sebeok, T. A., & Danesi, M. (2012). *The forms of meaning: Modeling systems theory and semiotic analysis* (Vol. 1). Walter de Gruyter.

Skemp, R. R. (1976). Relational understanding and instrumental understanding. *Mathematics Teaching*, *77*(1), 20–26.

Sodhi, R. S., Jones, B. R., Forsyth, D., Bailey, B. P., & Maciocci, G. (2013). BeThere: 3D mobile collaboration with spatial input. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 179–188.

Solanas, M. P., Vaessen, M. J., & de Gelder, B. (2020). The role of computational and subjective features in emotional body expressions. *Scientific Reports*, *10*(1), 1–13.

Stergiopoulou, E., & Papamarkos, N. (2009). Hand gesture recognition using a neural network shape fitting technique. *Engineering Applications of Artificial Intelligence*, *22*(8), 1141–1158.

Stokke, A. (2014). Truth and context change. *Journal of Philosophical Logic*, *43*(1), 33–51.

Talignani Landi, C., Ferraguti, F., Fantuzzi, C., & Secchi, C. (2019). A Passivity-Based Strategy for Manual Corrections in Human-Robot Coaching. *Electronics*, *8*(3), 320.

Tang, J. C. (1991). Findings from observational studies of collaborative work. *International Journal of Man-Machine Studies*, *34*(2), 143–160.

Tecchia, F., Alem, L., & Huang, W. (2012). 3D helping hands: A gesture based MR system for remote collaboration. *Proceedings of the 11th ACM SIGGRAPH International Conference on Virtual-Reality Continuum and Its Applications in Industry*, 323–328.

Thomason, W., & Knepper, R. A. (2016). Recognizing Unfamiliar Gestures for Human-Robot Interaction through Zero-Shot Learning. *International Symposium on Experimental Robotics*, 841–852.

Thorn, S., Gopalasingam, N., Bendtsen, T. F., Knudsen, L., & Sloth, E. (2016). A technique for ultrasound-guided blood sampling from a dry and gel-free puncture area. *The Journal of Vascular Access*, *17*(3), 265–268.

U.S. Department of Health and Human Services. (2019). *American Sign Language*. National Institutes of Health. https://www.nidcd.nih.gov/sites/default/files/Documents/health/hearing/NIDCD-American-Sign-Language-2019.pdf

Vatavu, R.-D., & Wobbrock, J. O. (2015). Formalizing agreement analysis for elicitation studies: New measures, significance test, and toolkit. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 1325–1334.

von Uexküll, J. (1937/2001). The new concept of Umwelt: A link between science and the humanities. *Semiotica*, *2001*(134), 111–123.

Wang, P., Zhang, S., Bai, X., Billinghurst, M., He, W., Sun, M., Chen, Y., Lv, H., & Ji, H. (2019). 2.5 DHANDS: a gesture-based MR remote collaborative platform. *The International Journal of Advanced Manufacturing Technology*, 1–15.

Watson, J. M., Kelly, B. A., Callingham, R. A., & Shaughnessy, J. M. (2003). The measurement of school students' understanding of statistical variation. *International Journal of Mathematical Education in Science and Technology*, *34*(1), 1–29.

Wei, C. Y. (2006). Not crazy, just talking on the phone: Gestures and mobile phone conversations. *2006 IEEE International Professional Communication Conference*, 299–307.

White, R., & Gunstone, R. (2014). *Probing understanding*. Routledge.

Wickelgren, W. A. (1977). Speed-accuracy tradeoff and information processing dynamics. *Acta Psychologica*, *41*(1), 67–85.

Wobbrock, J. O., Aung, H. H., Rothrock, B., & Myers, B. A. (2005). Maximizing the guessability of symbolic input. *CHI'05 Extended Abstracts on Human Factors in Computing Systems*, 1869–1872.

Wu, Q., Molesworth, B. R., & Estival, D. (2019). An Investigation into the Factors that Affect Miscommunication between Pilots and Air Traffic Controllers in Commercial Aviation. *The International Journal of Aerospace Psychology*, 1–11.

Yamashita, N., Kaji, K., Kuzuoka, H., & Hirata, K. (2011). Improving visibility of remote gestures in distributed tabletop collaboration. *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work*, 95–104.

Yeo, L.-M., & Tzeng, Y.-T. (2020). Cognitive Effect of Tracing Gesture in the Learning from Mathematics Worked Examples. *International Journal of Science and Mathematics Education*, *18*(4), 733–751.

Zenati-Henda, N., Bellarbi, A., Benbelkacem, S., & Belhocine, M. (2014). Augmented reality system based on hand gestures for remote maintenance. *2014 International Conference on Multimedia Computing and Systems (ICMCS)*, 5–8.

Zhang, Z., Li, Z., Liu, H., Cao, T., & Liu, S. (2020). Data-driven Online Learning Engagement Detection via Facial Expression and Mouse Behavior Recognition Technology. *Journal of Educational Computing Research*, *58*(1), 63–86.

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

Zhou, T., & Wachs, J. P. (2018). Early prediction for physical human robot collaboration in the operating room. *Autonomous Robots*, *42*(5), 977–995.

VITA

Edgar Javier Rojas Muñoz

School of Industrial Engineering, Purdue University

Education

Licenciatura*, Computer Engineering, 2016, Instituto Tecnológico de Costa Rica. Cartago, Costa Rica.

Topic: Research, Design and Construction of a Telementoring Tool for the Recognition and Capture of Real-Time Touch Gestures

PhD, Industrial Engineering, 2020, Purdue University, West Lafayette, Indiana, United States

Topic: Assessing Collaborative Physical Tasks via Gestural Analysis using the MAGIC Architecture

Research Interests

Human-Computer Interaction, Semiotics, Augmented Reality, Virtual Reality, Assistive Technologies

** A 'Licenciatura' is a degree technically higher than a Bachelor, but technically lower than a Master.*

PUBLICATIONS

Journals

1. **Rojas-Muñoz, E.**, Lin, C., Sanchez-Tamayo, N., Cabrera, M., Andersen, D., Popescu, V., Barragan, J., Zarzaur, B., Murphy, P., Anderson, K., Douglas, T., Griffis, C., McKee, J., Kirkpatrick, A., Wachs, J. (2020). Evaluation of an augmented reality platform for austere surgical telementoring: a randomized controlled crossover study in cricothyroidotomies. Nature Digital Medicine, 3(1), 1-9.
2. **Rojas-Muñoz, E.**, Cabrera, M. E., Lin, C., Andersen, D., Popescu, V., Anderson, K., Zarzaur, B., Mullis, B., Wachs, J. (2020). The System for Telementoring with Augmented Reality (STAR): A head-mounted display to improve surgical coaching and confidence in remote areas. Surgery.
3. **Rojas-Muñoz, E.**, Cabrera, M. E., Lin, C., Sanchez-Tamayo, N., Andersen, D., Popescu, V., Anderson, K., Zarzaur, B., Mullis, B., Wachs, J. (2020). Telementoring in Leg Fasciotomies via Mixed-Reality: Clinical Evaluation of the STAR Platform. Military Medicine, 185(Supplement_1), 513-520.
4. **Rojas-Muñoz, E.**, Cabrera, M. E., Andersen, D., Popescu, V., Marley, S., Mullis, B., Zarzaur, B., Wachs, J. (2019). Surgical telementoring without encumbrance: a comparative study of see-through augmented reality-based approaches. Annals of surgery, 270(2), 384-389.
5. **Rojas-Muñoz, E.**, Andersen, D., Cabrera, M., Popescu, V., Marley, S., Zarzaur, B., Mullis, B., Wachs, J. (2019). Augmented Reality as a Medium for Improved Telementoring. Military medicine, 184(Supplement_1), 57-64.
6. Andersen, D., Cabrera, M., **Rojas-Muñoz, E.**, Gonzalez, G., Popescu, V., Mullis, B., Marley, S., Zarzaur, B., Wachs, J. (2019). Augmented Reality Future Step Visualization for Robust Surgical Telementoring. Simulation in Healthcare, 14(1), 59-66.

Conferences

1. **Rojas-Muñoz, E.**, & Wachs, J. P. (2020). Beyond MAGIC: Matching Collaborative Gestures using an Optimization-based Approach. 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), 296–303.

2. **Rojas-Muñoz, E.**, & Wachs, J. P. (2020). The MAGIC of E-Health: A Gesture-Based Approach to Estimate Understanding and Performance in Remote Ultrasound Tasks. 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), 314–318.
3. **Rojas-Muñoz, E.**, Couperus, K., & Wachs, J. (2020). DAISI: Database for AI Surgical Instruction. arXiv preprint arXiv:2004.02809.
4. Lin, C., **Rojas-Muñoz, E.**, Cabrera, M., Sanchez-Tamayo, N., Andersen, D., Popescu, V., Barragan, J., Zarzaur, B., Murphy, P., Anderson, K., Douglas, T., Griffis, C., Wachs, J. (2020How about the mentor? Effective Workspace Visualization in AR Telementoring. In 2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR) (pp. 212-220). IEEE.
5. **Rojas-Muñoz, E.**, & Wachs, J. P. (2019). MAGIC: A Fundamental Framework for Gesture Representation, Comparison and Assessment. 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), 1–8.
6. Lin, C., **Rojas-Muñoz, E.**, Cabrera, M., Sanchez-Tamayo, N., Andersen, D., Popescu, V., Barragan, J., Zarzaur, B., Murphy, P., Anderson, K., Douglas, T., Griffis, C., Wachs, J. (2019). Robust High-Level Video Stabilization for Effective AR Telementoring. In 2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR) (pp. 1038-1039). IEEE.
7. Andersen, D., Lin, C., Popescu, V., **Rojas-Muñoz, E.**, Cabrera, M., Mullis, B., Zarzaur, B., Marley, S., Wachs, J. (2018). Augmented Visual Instruction for Surgical Practice and Training. In 2018 IEEE Workshop on Augmented and Virtual Realities for Good (VAR4Good) (pp. 1-5). IEEE.
8. Lin, C., Andersen, D., Popescu, V., **Rojas-Muñoz, E.**, Cabrera, M., Mullis, B., Zarzaur, B., Marley, S., Anderson, K., Wachs, J. (2018). A first-person mentee second-person mentor AR interface for surgical telementoring. In 2018 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct) (pp. 3-8). IEEE.
9. Dey, A., Billinghurst, M., Welch, G., **Rojas-Muñoz, E.** (2018, October). 3rd Virtual and Augmented Reality for Good (VAR4Good) Workshop. In 2018 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct) (pp. 364-364). IEEE.