Abstract—Gestures are a key aspect of communication during collaboration: through gestures we can express ideas, inquire and formalize instructions as we collaborate. Nevertheless, gesture analysis is not currently used to assess quality of task collaboration. One possible reason for this is that there is no consensus on how to represent and compare gestures from the semantic standpoint. To address this, this paper introduces three novel approaches to compare gestures performed by individuals as they collaborate to complete a physical task. Our approach relies on solving three variations of an integer optimization assignment problem, i.e. based on gesture similarity, based on temporal synchrony, and based on a combination of both. We collected the gestures of 40 participants (divided into 20 pairs) as they performed two collaborative tasks, and generated a human baseline that compared and matched their gestures. Afterwards, our gesture comparison approach was evaluated against other gestures comparison approaches based on how well they replicated the human baseline. Our approach outperformed the other approaches, agreeing with the human baseline over 85% of the times. Thus, the obtained results support the proposed technique for gesture comparison. This in turn can lead to the development of better methods to evaluate collaborative physical tasks.

I. INTRODUCTION

Gestures are key components of human interaction, collaboration and fundamental communication. For example, several studies have pointed to the benefits of using gestures to facilitate learning [1], for generating novel problem-solving abilities [2], and for offloading working memory [3], among other benefits. The impact of gesture use has also been studied in the context of collaborative tasks, as people tend to perform gestures as they collaborate during physical tasks (e.g. a construction task) [4]. For instance, gestures can either be used to solve the task (e.g. assembling parts) or as a means of inquiry or instruction (e.g. asking for clarification) [5]. Additionally, gestures have been found to facilitate the exchange of ideas and express intentions during teamwork for task completion [6], which leads to better collaborations. Moreover, gestures are considered so significant to support idealization, teamwork and collaboration, that there are functional systems that allow users to convey gestures during remote and distributed collaboration [7]–[9].

While gestures are considered beneficial for proper human interaction and collaboration, they are often treated only as illustrational aids. In other words, although gestures encompass a great deal of information related to the collaborative task, this information is commonly ignored. This issue is aggravated by the fact that there is currently no consensus on how to represent and compare gestures from the semantic standpoint. Such metric is necessary as a proxy to evaluate the quality of the collaboration. The reason for this is that gestures could reveal whether the instructions and actions performed by the collaborators are properly being understood and executed. While techniques to represent and compare gestures have been studied [10]–[12], none of these approaches has tackled the problem of comparing gestures in a comprehensive way (i.e. encompassing the gestures’ shape, movement, meaning and context). Such comprehensive approach is necessary to compare the gestures generated in, for example, Helper-Worker collaborative scenarios (i.e. a collaborator guiding another one through a shared task), where the gestures can differ greatly visually.

A recent work by Rojas-Muñoz and Wachs attempted to address this gap by introducing the Multi-Agent Gestural Instruction Comparer (MAGIC) framework, an architecture to abstract and compare gestures based on their morphology, semantics and pragmatics [13]. Although their work had promising results, their approach required a priori information to compare the gestures effectively. Moreover, their approach had to exhaustively compare all gestures, which was computationally inefficient. Finally, a reduced number of participants and tasks limited their generalizability.

In this paper, we present an approach to compare the gestures performed in Helper-Worker collaborative scenarios that addresses the gaps in the framework presented in [13]. We introduce three different methods to compare gestures based on solving integer optimization assignment problems. The first approach matches gestures based on the similarity between the abstractions representing the gestures (MAGIC-based approach; similar to [13]). The second method matches the gestures based on their temporal synchrony (time-based approach). The third approach combines the previous two into a method that considers both gesture similarity and temporal synchrony (hybrid approach). Additionally, our approach’s generalizability was evaluated by introducing a larger number of participants and two collaborative tasks.

The contributions of this work include: (1) introducing three novel approaches to compare gestures in Helper-Worker collaborative scenarios based on solving integer optimization assignment problems; (2) evaluating the generalizability with two different collaborative tasks and a larger pool of subjects; and (3) analyzing if the perceived difficulty of a task impacts how gestures are compared.
The paper proceeds as follows: Section II reviews prior work related to the importance of gestures in collaborative tasks and current approaches for gesture representation and comparison. Section III summarizes the MAGIC architecture and introduces our new gesture comparison approach. Section IV describes our setup to acquire gestures during collaboration. Section V presents and discusses our results, and Section VII concludes the paper.

II. BACKGROUND

Gestures are an integral component of how humans learn and interact. Whether voluntarily [14] or involuntarily [15], gestures are extensively used to communicate intentions, express emotions, convey interpersonal attitudes, and support speech and other cultural expression forms [16]. Moreover, numerous studies have found a connection between the use of gestures and the acquisition and development of knowledge and problem-solving skills [17]. For example, children that were allowed to gesture as they learnt mathematical concepts (e.g. addition) retained the knowledge better and came up with new strategies to tackle problems, as opposed to those that were not allowed to gesture. In those studies, gestures allowed the students to create associations between concepts and their own movements, which revealed new solving strategies that were not evident before. Similar findings were seen for memory retention and language learning [1], [18].

Gestures have also been studied in the context of collaborative tasks. Research has shown that collaborators can better understand each other’s intentions by looking at each other’s gestures [19]. Moreover, observing those gestures can influence the planning and execution of subsequent actions: people can better coordinate their actions based on the gestures performed by others [20]. Additionally, a systematic use of gestures while performing a task can allow collaborators to assess task status, increase situation awareness and achieve a common grounding [4], [21]. Furthermore, gestures are considered so significant to support teamwork that there are systems to convey gestures during remote collaboration. Such approaches include using projectors to display the gestures [22], creating shared visual spaces via head-mounted cameras [21] and displaying a representation of the collaborators’ gestures via mixed reality [7].

In spite that gestures are highly significant for shaping the task performance during collaboration, gestures have not been used to quantitatively assess quality of collaboration. Furthermore, there exists no consensus about how to represent and compare gestures in a way that encompasses more than just the gestures’ morphological characteristics. In the context of a collaboration where there is a Helper and a Worker, as in [4], these comparisons are necessary to evaluate whether the instructions are properly understood and executed. Current approaches for gesture representation and comparison follow three main directions: morphology-based, semantic-based and elicitation-based approaches. Morphology-based representations compare gestures based on their physical appearance, often relying on the positions of the hands, motion orientations, among others [23]–[25]. Albeit effective, these approaches struggle to find similarities between gestures performed in Worker-Helper collaborative scenarios. In such settings, the visual appearance of gestures referring to the same concept (functionally equivalent gestures [19, p.101]) can differ significantly. Conversely, semantic-based approaches leverage linguistics frameworks to construct logical abstractions representing the gestures’ meaning and context [26]–[28]. These approaches, however, have not been used to quantitatively compare gestures. Finally, elicitation-based approaches are common when creating lexicons of gestures to control devices (e.g. UAVs, medical imaging software [12], [29]). By using a set of descriptors (e.g. movement planes, parts of body being referenced), high-level comparisons between the gestures can be performed. However, these abstractions seek to group gestures into a predefined sets of actions or categories, which cannot be directly adapted to the large variety of actions that are generated in collaborative tasks.

Rojas-Muñoz and Wachs tackled the issue of gesture representation, comparison and assessment in the Helper-Worker collaborative scenario by combining the three previously described gesture comparison directions into a single framework [13]. Their Multi-Agent Gestural InstructionComparer (MAGIC) framework leveraged a gestural taxonomy classification, a dynamic semantics framework and a constituency parsing to represent and compare gestures in a way that considered the gestures’ morphology (i.e. shape and movement), semantics (i.e. meaning, timing), and pragmatics (e.g. context). Nonetheless, their gesture matching approach greatly depended on which aspects of their gestural abstraction were used to compare the gestures: selecting the wrong subsection of their representation cause a significant drop in their results. Moreover, finding what part of the representation was suitable for comparison required a-priori knowledge. Additionally, their approach had to exhaustively compare all the gestures, leading to computational overhead in tasks with larger number of gestures. Finally, their framework was evaluated over a limited set of participants and tasks, making it not as generalizable. Our work leverages and extends the framework in [13] by proposing three different approaches to compare gestures and evaluating them in more generalizable settings.

III. METHODOLOGY

Our work compares gestures performed in collaborative physical scenarios. Specifically, scenarios where one individual is remotely guiding another to complete a task, a challenging setup for standard gesture comparison approaches [30]. Let the $\Phi_H$ Helper be the individual giving instructions about how to complete the task. Similarly, let the $\Phi_W$ Worker be the individual receiving the instructions and executing them. Both the $\Phi_H$ Helper and the $\Phi_W$ Worker will interact via speech and gesture until the task is completed. Additionally, let $W$ be a set containing all the $w_i$ Worker-authored gestures, and let $H$ be a set containing all the $h_j$ Helper-authored gestures. These definitions will be used throughout the rest of the paper.
A. Multi-Agent Gestural Instruction Comparator

An approach to represent and obtain similarity metrics is required to compare gestures. MAGIC was a recently introduced framework for gesture representation, comparison and assessment that tackles this need [13]. The framework’s goal was to abstract gestures into a representation that considered aspects of the gestures such as their shape, movement, meaning and context. These abstractions were named Interpretation Trees [13]. These tree data structures were generated for each gesture after applying a process consisting of a gestural taxonomy classification, a dynamic semantics framework and a constituency parsing. By representing gestures as trees, the authors were able to encompass the information conveyed by the gestures while preserving the hierarchical relation between the gestures’ components (e.g. a finger is part of the hand, which is consequently part of the arm). Additionally, their tree configuration made it possible to group similar aspects of the gestures into subtrees. For example, aspects related to the gestures’ movement (e.g. plane of motion, motion orientation) can be extracted as a subtree, isolated from other aspects from the gesture (e.g. shape, meaning). Fig. 1 presents an example of an Interpretation Tree and subtrees extracted from it. In our work, we leverage MAGIC’s Interpretation Trees to represent gestures and obtain quantitative measurements of gesture similarity.

B. Gesture Matching through Integer Optimization Problems

The last stage of the MAGIC framework introduced a gesture matching routine that compared gestures based on their Interpretation Trees [13]. Once Interpretation Trees were generated for each gesture, the intersection of each tree against each other was computed. The number of elements in the intersection was used as a measure of gesture similarity: the higher the number of elements in the intersection between two Interpretation Trees, the higher the similarity between the gestures. Additionally, the comparisons could be performed either over the entire Interpretation Tree or against specific subtrees (represented as $\Psi^{X}$, the subtree of the Interpretation Tree that has the node $X$ as its root node). Therefore, using [13], the Helper-authored gestures that is the most functionality equivalent to the $w_i$. Worker-authored gesture is found by computing the following relation:

$$\text{max} \left( \sum_{\text{nodes} \in \Psi^{X}}^{\Psi^{X_i} \cap \Psi^{X_j}} \right) ; \quad i = 1, 2, \ldots, |W| ; \quad j = 1, 2, \ldots, |H|$$

The result after applying this process to all the gestures was a gesture matching solution describing how the gestures performed by the Helper and the Worker were matched, i.e. which gestures were the most functionality equivalent. This approach, however, has two limitations. First, the outcome of the gesture matching process heavily depends on the subtree selected to perform the comparisons. Their process of selecting the appropriate subtree was done based on a priori knowledge, making it not necessarily generalizable to other collaborative tasks. Additionally, comparing every $w_i$ Worker-authored gesture against every $h_j$ Helper-authored gestures can lead to computational overhead in scenarios with a larger number of gestures. To address these limitations, we propose three approaches to match gestures based on solving integer optimization assignment problems.

Instead of exhaustively comparing gestures to find a gesture matching solution, our gesture matching approach conducts the search in such a way that the matching is determined based on an optimality criterion among the collaborators’ gestures. Let an $e_{ij}$ edge weight the representation of a matching between two gesturers. Each $e_{ij}$ edge weight will take a value of 1 if the $h_j$ Helper-authored gesture matches (i.e. is the most functionality equivalent) to the $w_i$ Worker-authored, and 0 otherwise. Our approach represents each gesture matching solution with a $E$ matrix of $e_{ij}$ edge weights of size $|W| \times |H|$. Finally, the goal of our approach is to solve the integer optimization assignment problems to find the $E$ matrices that describes the matching of the gestures performed by the collaborators such that the overall cost of the matching is optimized.

1) MAGIC-based optimization: Our initial approach rewrites the gesture matching approach from [13], presented in Equation (1), as an integer optimization assignment problem. Instead of exhaustively comparing all the gestures to find the best gesture matching solution (i.e. the best $E$ matrix), a single $B$ matrix of distance costs (of size $|W| \times |H|$) is computed beforehand and integrated to the cost function of our optimization problem. Each of the $b_{ij}$ distance costs is generated by computing the number of nodes in the intersection between the Interpretation Trees or subtrees representing different gestures:

$$b_{ij} = \left( \sum_{\text{nodes} \in \Psi^{X}}^{\Psi^{X_i} \cap \Psi^{X_j}} \right) ; \quad i = 1, 2, \ldots, |W| ; \quad j = 1, 2, \ldots, |H|$$

After computing the $B$ matrix of distance costs using Equation (2), the optimal gesture matching solution will be found by solving the following problem:
maximize \[
\sum_{j=1}^{[H]} \sum_{i=1}^{[W]} b_{ij} c_{ij}
\]
subject to \[
\sum_{j=1}^{[H]} e_{ij} = 1, \forall i
\]
\[
e_{ij} \in \{0,1\}
\]
\[
i=1,2,\ldots,[W]; j=1,2,\ldots,[H]
\]

The cost function is maximized whenever each \(w_i\)Worker-authored gesture is matched to the \(h_j\) Helper-authored gesture with the highest number of common nodes (based on the subtrees being compared). This formulation is constrained to each \(w_i\)Worker-authored gesture only being matched to one \(h_j\) Helper-authored gesture. This is because, in most cases, each \(w_i\) Worker-authored gesture is performed in reaction to a single \(h_j\) Helper-authored gesture. While multiple \(w_i\) Worker-authored gestures can be required to perform an instruction conveyed with a single \(h_j\) Helper-authored gesture, it is not possible to fulfill instructions given with different \(h_j\) Helper-authored gestures with only one \(w_i\) Worker-authored gesture. In other words, each new \(h_j\) Helper-authored gesture “outscopes” the previous one: even though the information encompassed by an gesture might be transferred to a new one, it is not possible to react to a specific old gesture once a new one has been made [26]. This MAGIC-based approach, however, does not address the issue of how the gesture matching significantly varies with respect to the subtrees selected to compare against.

2) Time-based optimization: A feature of collaborative tasks is that gestures performed by the \(\Phi_W\) Worker are performed in reaction to gestures performed by the \(\Phi_H\) 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 \(w_i\) Worker-authored gesture and the \(h_j\) Helper-authored gesture are, the more likely they are related.

The MAGIC-based optimization from Equation (3) 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 \(C\) matrix of time costs (of size \([W] \times [H]\)) is computed with respect to the time in which the gestures were performed. The time in which each gesture was performed is stored in two vectors \(\vec{t}_W\) and \(\vec{t}_H\), respectively for \(w_i\) Worker-authored gestures and \(h_j\) Helper-authored gestures. Afterwards, the vectors are expanded into a matrix (of size \([W] \times [H]\)) by multiplying the \(\vec{t}_W\) and \(\vec{t}_H\) by a vector of ones (of size \([H] \times 1\) and \([W] \times 1\), respectively). Finally, the \(C\) matrix of time costs is computed by solving the following relation:

\[
C = \vec{t}_W \vec{t}_H^T - \vec{1} \vec{1}^T
\]

After computing the \(C\) matrix of time costs using Equation (4), the optimal gesture matching solution will be found by solving the following problem:

minimize \[
\sum_{j=1}^{[H]} \sum_{i=1}^{[W]} c_{ij} e_{ij}
\]
subject to \[
\sum_{j=1}^{[H]} e_{ij} = 1, \forall i
\]
\[
e_{ij} \in \{0,1\}
\]
\[
i=1,2,\ldots,[W]; j=1,2,\ldots,[H]
\]

The cost function is minimized whenever each \(w_i\) Worker-authored gesture is matched to the most recently performed \(h_j\) Helper-authored gesture. As in Equation (3), a constraint regulates that each \(w_i\) Worker-authored gesture can only be matched to one \(h_j\) 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 \(w_i\) Worker-authored gesture is compared against a \(h_j\) 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) Hybrid optimization: Our final formulation combines the previous two approaches so that both temporal synchrony and gesture similarity are considered. 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 against. The approach computes a \(D\) matrix of hybrid costs (of size \([W] \times [H]\)) from the previous \(B\) matrix of distance costs and \(C\) matrix of time costs. We propose a function based on the signum function [31] to regulate the effect of the \(b_{ij}\) and \(e_{ij}\) input costs in the \(d_{ij}\) hybrid costs. This function 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:

\[
d_{ij} = \left(\frac{-e^{-\alpha c_{ij}} - e^{-\alpha c_{ij}} - \beta}{-e^{-\alpha c_{ij}} - \beta} - e^{-\alpha c_{ij}}\right) + \frac{b_{ij}}{\sum_{j=1}^{[H]} b_{ij}}
\]

The \(\alpha\) damping constant regulates the damping caused by the \(e_{ij}\) time costs, and the \(\beta\) translation constant regulates when the damping begins. Fig. 2 showcases the effect of these constants over the time damping section of Equation (6). The x-axes represent the values of the \(c_{ij}\) time-based costs from Equation (4). The y-axes represent the resulting values after applying the time damping section of Equation (6). The left graph of Fig. 2 showcases effect of the \(\alpha\) constant. Lower \(\alpha\) values increase the importance of gestures performed less recently, while higher \(\alpha\) values emphasize the most recent gestures. Typical values for this constant are \(0.05 \leq \alpha \leq 0.1\) to balance the importance between recent and older performed gestures. The right graph of Fig. 2 showcases effect of the \(\beta\) constant. The \(\beta\) constant regulates
when Equation (6) activates. Typical values for this constant are $0.9 \leq \beta \leq 1.1$. For $\beta$ 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 (6) reaches its maximum value when the $h_j$ Helper-authored gesture and the $w_i$ Worker-authored gesture were performed at the same time. After computing the $D$ matrix of hybrid costs using Equation (6), the optimal gesture matching solution will be found by solving the following assignment problem:

$$\text{maximize} \sum_{j=1}^{[H]} \sum_{i=1}^{[W]} d_{ij} e_{ij}$$

subject to $\sum_{j=1}^{[H]} e_{ij} = 1, \forall i$

$$e_{ij} \in \{0, 1\}, \quad i = 1, 2, \ldots, [W]; \quad j = 1, 2, \ldots, [H]$$

The cost function is maximized whenever a balance between two conditions is achieved: 1) each $w_i$ Worker-authored gesture is matched to the most recently performed $h_j$ Helper-authored gesture, and 2) each $w_i$ Worker-authored gesture is matched to the $h_j$ Helper-authored gesture that has the highest number of common nodes (based on the subtrees being compared). As in the Equations (3) and (5), the constraint that establishes that each $w_i$ Worker-authored gesture can only be matched to one $h_j$ Helper-authored gesture is preserved. There is no need for an additional temporal constraint as in Equation (5), as the effect of time is considered in the time damping section of Equation (6).

Fig. 3 exemplifies how our gesture matching approaches match the gestures performed by a $\Phi_H$ Helper and a $\Phi_W$ Worker. The first row showcases ten gestures performed by a $\Phi_H$ Helper to convey instructions ($h_j; 1 \leq j \leq 10$). The second row showcases ten gestures performed by a $\Phi_W$ Worker to execute the received instructions ($w_i; 1 \leq i \leq 10$). An arrow connecting the $w_i$ Worker-authored gesture and the $h_j$ Helper-authored gestures represents that these gestures match. The color of the arrows represent 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 $\Phi_W$ 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 $\Phi_H$ 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 will match gesture $h_1$ with gesture $w_1$, represented by a green arrow connecting the two gestures.

The next section describes our experimental setup to evaluate our three gesture matching approaches against baseline approaches for gesture comparisons.

IV. EXPERIMENTAL APPARATUS

A. Data Collection

A user study was conducted in which 40 participants divided into 20 Helper-Worker pairs (graduate students, 22 males and 18 females; age mean of 25.52 ± 3.5 years old) completed a collaborative task. The pairs were randomly assigned to one of two collaborative tasks: a block assembly task in which the collaborators used blocks to assemble a structure; or an origami folding task in which participants had to make a hat using a sheet of paper. The tasks were 24 and 11 steps long, respectively. These two tasks are common in experiments evaluating the use of gestures during collaborative physical tasks [13], [22], [32]. Once the task was selected, the participants were randomly assigned to either the $\Phi_W$ Worker or the $\Phi_H$ Helper role. The $\Phi_W$ Worker and $\Phi_H$ Helper were assigned to different rooms. Both rooms included an RGB-D camera and a screen connected to a computer. Additionally, the $\Phi_W$ Worker room had a table in which blocks or a sheet of paper (depending on which task the pair was assigned to) were placed. On the other hand, the $\Phi_H$ Helper room had an instruction booklet detailing the steps to complete the collaborative task. Therefore, the participants had to collaborate to complete the steps detailed in the booklet (available only to the $\Phi_H$ Helper) using the blocks/sheet of paper (available only to the $\Phi_W$ Worker). The Helper-Worker pairs were able to interact with each other at all time through a Skype video call, displayed in the room’s screen. This setup allowed participants to use both speech and gestures to communicate. Through this setup, 2211 different gestures were collected (1316 performed by the $\Phi_H$ Helper, 895 performed by the $\Phi_W$ Worker) over more than 3 hours and a half of video. Written consent was acquired from each participant.
B. Ground Truth Annotation

Our gesture matching results were evaluated against a human-annotated baseline. A member of the research team manually annotated which gestures were functionally equivalent. For example, if the $\Phi_H$ Helper made a gesture to indicate how to connect two blocks, all the gestures performed by the $\Phi_W$ Worker to fulfill this instruction were matched to the $\Phi_H$ Helper’s gesture. This generated a $\hat{E}$ matrix of ground truth gesture matchings for each pair of participants. Afterwards, the optimal $E$ matrices obtained with Equations (3), (5) and (7) were compared against this $\hat{E}$ ground truth matrix. The gesture matching score was obtained by computing the F1-Score between the two matrices:

$$\text{Matching Score} = \frac{2TP}{2TP + FN + FP}$$  (8)

where TP are the True Positives (i.e. $\hat{E}_{ij} = E_{ij} = 1$), FN are the False Negatives (i.e. $\hat{E}_{ij} = 1; \hat{E}_{ij} = 0$); and FP are the False Positives (i.e. $\hat{E}_{ij} = 0; \hat{E}_{ij} = 1$) [33]. This score represents the percentage of agreement of the gesture matching approaches with the human-annotated baseline.

C. Gesture Comparison Baselines

Following [13], we compared our approaches against two gesture matching baselines: morpho-semantic descriptors (MSD) [34], and naïve time synchronization (NTS). MSD are Boolean vectors representing characteristics of the gestures (e.g. is there rightward movement? Is the hand pointing at the head?). Each gesture was represented using a $48 \times 1$ vector (a subset of 48 descriptors from [34]). The vectors were compared using Hamming distance and cosine similarity. NTS compared gestures based on their temporal occurrence. Each gesture was represented as a timestamp in seconds, with 0 and 1 as the start and end of the video, respectively. For each $h_j$ Helper-authored gesture, a time window before and after its execution was created. Every $w_i$ Worker-authored gesture inside this window was associated to the given $h_j$ Helper-authored gesture. To obtain these time windows, a $t_k$ timestamp was assigned to each gesture. Afterwards, the bounds of the time windows between consecutive timestamps were calculated as $t_{k+1} - t_k$. This process was performed for all the gestures in the $H$ set, until the whole completion time was divided into time windows.

D. Evaluating the Perceived Task Load

After completing the task, each participant assessed their perceived task load using the NASA TLX [35]. The NASA TLX is a self-assessment questionnaire to evaluate the load experienced while performing a task. The questionnaire is divided into 6 sections: mental demand, physical demand, temporal demand, perceived performance, effort required for completion, and generated frustration. Higher TLX scores indicate that the task imposed a higher load.

V. RESULTS AND DISCUSSION

The optimization problems were solved using the IBM’s CMPLX Optimizer from the NEOS Server [36]–[38]. Fig. 4 showcases the matching scores of the gesture matching approaches, using different subtrees to compare the gestures [13]. Namely, the comparisons were made against: 1) the subtree representing the gesture’s context (Context Subtree);
2) the subtree representing if the gesture is being used to represent an object (Exemplify Subtree); 3) the subtree representing the gesture’ shape (Shape Subtree); 4) the subtree encompassing all aspects of the gesture except its context and variable declarations (Predicative Subtree); 5) the entire $Ψ_H$ Helper Interpretation Tree; and 6) the combination of subtrees representing the gesture’s meaning (Meaning Subtree). The hybrid optimization approach obtained higher matching scores for every comparison. Additionally, the matching scores were lower for the origami task.

Our MAGIC-based optimization replicates the findings in [13]: the approach outperformed the two gesture matching baselines (MSD and NTS), but the results fluctuate significantly based on which subtrees are used to compare the gestures. The highest matching score was found when comparing the $Ψ_W$ Worker Interpretation Trees (specifically, the subtree representing the context), against the subtree of the $Ψ_H$ Helper Interpretation Trees representing the gesture’s meaning. The results reaffirm the importance of meaning and context when comparing gestures. In contrast, comparisons based only on the gestures’ physical appearance did not obtain high matching scores: the gestures performed by the $Φ_H$ Helper and the $Φ_W$ Worker were visually very distinct. These findings are congruent with the results in [13].

The results of our time-based approach show that gestures performed one after the other are likely to be related. This approach differs from the NTS approach since it gives higher importance to recently performed gestures instead of creating a time window in which each gesture has equal importance. However, both approaches find temporal relations between gestures without comparing them content-wise.

The hybrid optimization approach presents a new alternative that integrates advantages from the MAGIC-based and 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. 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 or the one in [13]. Therefore, our hybrid approach is a better and more stable option for gesture comparison.

When comparing the block and origami tasks, the TLX scores showed that participants perceived the origami task as more mentally ($0.281 \pm 0.183$ vs. $0.437 \pm 0.251$), physically ($0.133 \pm 0.097$ vs. $0.188 \pm 0.135$) and temporally ($0.348 \pm 0.259$ vs. $0.354 \pm 0.207$) demanding. Moreover, the origami task was considered harder ($0.314 \pm 0.221$ vs. $0.426 \pm 0.271$) and more frustrating ($0.271 \pm 0.228$ vs. $0.310 \pm 0.228$). Finally, participants considered they performed worse in the origami task ($0.148 \pm 0.096$ vs. $0.354 \pm 0.270$).

Initially, the block assembly task was expected to be more demanding as it requires more than double the number of steps than the origami task. However, the results demonstrate the opposite. 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 error 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.

VI. CONCLUSION

This work introduced three novel approaches to compare gestures performed by individuals as they performed a collaborative physical task. By solving three different variations of an integer optimization assignment problem, our approach matched gestures based on how functionally equivalent they are. The three variations included optimization based on gesture similarity (MAGIC-based), based on temporal synchrony (time-based), and based on a combination of both (hybrid). An experimental setup was designed to evaluate the performance of the proposed methods against a human baseline and two other gesture matchings approaches. Our proposed gesture matching approach agreed with the human baseline over 85% of the times, outperforming the other gesture matching baselines. Our MAGIC-based approach was able to compare gestures without requiring exhaustive computations that could lead to computational overhead. Additionally, our time-based approach reduced the number of computations and demonstrated the importance of integrating a temporal aspect when comparing gestures performed in a collaborative settings. Finally, our hybrid approach out-
performed all other approaches and baselines, demonstrating that both gesture similarity and time synchrony are necessary for meaningful gesture comparison in collaborative settings. The results obtained with our hybrid approach did not change drastically based on how the gesture similarity was obtained, an improvement from previous research in this field. Also, a workload assessment questionnaire indicated that the level of demand and frustration caused by the task affected the gesture matching scores. Overall, these results demonstrate that our approach can be used to compare gestures in a similar way a human would do, which could lead to better methods to evaluate the performance of collaborative tasks.

VII. ACKNOWLEDGMENTS

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

REFERENCES


