HHS Public Access

Author manuscript

IEEE Trans Hum Mach Syst. Author manuscript; available in PMC 2021 October 01.

Published in final edited form as:

IEEE Trans Hum Mach Syst. 2020 October ; 50(5): 434–443. doi:10.1109/THMS.2020.2992216.

Agreement Study Using Gesture Description Analysis

Naveen Madapana [graduate students],
School of Industrial Engineering, Purdue University, Indiana, USA

Glebys Gonzalez [graduate students],
School of Industrial Engineering, Purdue University, Indiana, USA

Lingsong Zhang [Associate Professor],
Department of Statistics, Purdue University, Indiana, USA

Richard Rodgers [Associate Professor],
Department of Clinical Neurosurgery, Indiana University School of Medicine, Indiana, USA

Juan Wachs [Associate Professor]
School of Industrial Engineering, Purdue University, Indiana, USA

Abstract

Choosing adequate gestures for touchless interfaces is a challenging task that has a direct impact on human-computer interaction. Such gestures are commonly determined by the designer, ad-hoc, rule-based or agreement-based methods. Previous approaches to assess agreement grouped the gestures into equivalence classes and ignored the integral properties that are shared between them. In this work, we propose a generalized framework that inherently incorporates the gesture descriptors into the agreement analysis (GDA). In contrast to previous approaches, we represent gestures using binary description vectors and allow them to be partially similar. In this context, we introduce a new metric referred to as Soft Agreement Rate ($\mathcal{SAR}$) to measure the level of agreement and provide a mathematical justification for this metric. Further, we performed computational experiments to study the behavior of $\mathcal{SAR}$ and demonstrate that existing agreement metrics are a special case of our approach. Our method was evaluated and tested through a guessability study conducted with a group of neurosurgeons. Nevertheless, our formulation can be applied to any other user-elicitation study. Results show that the level of agreement obtained by $\mathcal{SAR}$ is 2.64 times higher than the previous metrics. Finally, we show that our approach complements the existing agreement techniques by generating an artificial lexicon based on the most agreed properties.

I. Introduction

GESTURAL modalities offer an intuitive and natural mode of interaction with machines that resembles human-to-human communication. Nowadays, gesture-based interfaces are part of handheld devices, smart TVs, autonomous vehicles and gaming consoles [1], [2]. Further, several orthodox disciplines such as surgery have benefited from such interfaces, since it

nmadapan@purdue.edu.

allows surgeons to control medical systems in a touchless manner while maintaining total asepsis [3]-[5].

One critical step when developing gestural systems is the selection of gesture lexicons that are in compliance with the preferences of the users. This can be achieved through participatory design methodologies such as user elicitation or guessability studies [6]. Nevertheless, the common denominator is that all methods recommend involving end-users in the early stages of the design process to obtain high acceptability interfaces [7]. A commonly used methodology to understand user preferences consists of elicitation studies followed by agreement analysis [8]. The latter study is especially beneficial in domains requiring a particular expertise (i.e. radiology, urology, neurology, and aviation) since these populations have intrinsic knowledge about the environment that shapes the gestures they commonly use. In particular, agreement analysis quantifies the degree of preference among the users. Several methods exist for measuring agreement [9]-[11] as they allow interface designers to create standardized lexicons based on the majority's preferences [10], [12].

Existing agreement approaches are based on the formation of equivalence groups among the elicited proposals. Human judgment or intuition is often used to group those proposals that are similar but not identical. We refer to these methodologies as "hard classification" approaches (1 - identical or 0 - nonidentical). This rigid interpretation of gestures (based on appearance only and leaving aside their properties) ignores the common properties that the gestures may have, producing an apparent low agreement rate [13], [14]. This can be particularly problematic if the goal of the elicitation study is to produce a standard lexicon based on the agreement, since the final lexicon may not reflect user preferences. In such cases, the designers may benefit from a more granular concept of similarity that allows participants to partially agree on proposed gestures. Thus, our work aims to quantify the degree of similarity between the proposals instead of assigning binary labels. In order to measure the similarity, we propose to decompose the proposals into finite atomic properties. We refer to these similarity-based methodologies as "soft classification" approaches. In the "soft" approaches, the proposals are represented as a combination of high-level properties (soft representations) which allows having partial similarity between the proposals.

Consider the following toy example to contrast these two approaches. Say, there are four gestures for the command *pan up* as shown in Figure 1. These four gestures are clearly dissimilar and will not be grouped into an equivalence group according to the former approaches. Hence the "hard" approaches do not consider the fact that these gestures share a common property known as *upward motion* though there are other properties that are different. In this way, the "hard" gesture classification scheme fails to identify the common attributes in the gestures that do not belong to the same equivalence group.

Another limitation of "hard classification" approaches lies in the interpretation of the agreement rate. In other words, the level of agreement does not directly represent the number of users that agreed on a particular proposal. In this regard, Vatavu et. al [9] proposed threshold values to determine low, medium and high agreement rates. However, these thresholds do not have a qualitative interpretation in terms of the average percentage of participants that agreed on a proposal (η). Hence, we propose an empirical relation between

the level of agreement and η. Furthermore, when the agreement rates are low, it is extremely difficult for system designers to select/design the final gestures based on a majority vote. These limitations can be addressed using soft representations as the designers can instead rely on the descriptors with the highest agreement to develop the final gestures.

In our previous work [15], we proposed to incorporate gesture properties into agreement analysis. However, this approach was based on intuition rather than on a thorough mathematical grounding. Our work builds on this idea and provides a rigorous mathematical foundation for the agreement approaches that are concerned with soft representations. Next, we established a quantitative relationship between our formulation and the existing agreement techniques [9], and showed that our approach complements the current methodologies for agreement assessment [9], [16].

The key contributions of this paper are as follows: 1. A mathematical argumentation to perform agreement analysis by incorporating gesture descriptors, which we refer to as Gesture Description Analysis (GDA), 2. A mathematical proof showing that our approach can also generalize to the “hard” classification of gestures. 3. An empirical relationship between the agreement metric and the average percentage of participants that agreed on a particular gesture.

II. Background

Gesture-based systems have become a rapidly growing technology in the last two decades [17], [18]. This popularity is grounded in the fact that gestures are natural and easy to use, in fact, they are an integral part of our conversational repertoire. They are not only intuitive to the user but also there is a cultural aspect to them, such as in *emblems* [19]. Additionally, they can be used as a safe and aseptic alternative to traditional interfaces in the particular case of medical environments [3], [5].

Development of gestural interfaces usually adheres to the following workflow: 1. Determine the number of commands for the task at hand, 2. Generate a gesture lexicon, and 3. obtain gesture instances and train a classifier that can recognize those gestures. A gesture lexicon is defined as a mapping that relates each command to a particular gesture. The lexicon generation is a critical step to achieve a high usability in a a system [18]. The literature shows three main approaches to this problem: arbitrary or authoritarian, technology-driven [20] and user elicited [21]. In the first two approaches, the system developers can design the gestures based on their own expertise [2] or the gestures are chosen based on the technology used for interaction [22]. The last approach is based on the participatory design (for instance, user elicitation studies) that involves end-users of the system at the early stages of the design process [10], [12].

The work done in gesture elicitation can be classified into two categories: 1. Constrained [14], and 2. Unconstrained elicitation [23], [24]. In the first category, a finite gesture set is created either from intuition or from the previous experience, and then the participants pick a gesture for each command. The main problem with these approaches is that it may cause forced consensus since the users have a smaller pool of options [25]. The second category

consists of a completely unconstrained setup, where the users can draw, perform or describe the gestures that they think would best fit each command [23], [24]. Gestures elicited in such studies fall under the category of iconic gestures as they are representative of their respective commands. Further, this experimental design has the advantage of including the user's knowledge directly into the design of a gestural lexicon. This is particularly important in areas where the users are domain experts, since they are naturally aware of the characteristics, requirements, and limitations of the systems that are being used.

After the gesture elicitation procedure is complete, agreement analysis is commonly conducted to determine an optimal lexicon [13] or to create a set of guidelines for interface designs [14]. The agreement is defined as the ratio of the number of agreed pairs to all possible pairs [24]. One of the most popularly used agreement metrics was proposed by Wobbrock et al. [16]. This work defined an agreement index $\mathscr{A}_r$ for guessability studies with symbolic inputs. Since then, other agreement proposals including participatory studies have emerged. Morris et al. [26] made an agreement proposal that could handle different group sizes for each referent. This is the case when each user can generate more than one proposal per referent.

More recently, a new metric called "Agreement rate" ($\mathscr{AR}_r$) was proposed by Vatavu et al. to address a critical issue associated with the metric ($\mathscr{A}_r$) [9]: When there is no agreement between users, the value of the metric is non-zero, which can produce an overestimation of the agreement. Other works have found formulae that are very similar to $\mathscr{AR}_r$. For example, the work in [12] found a very similar metric when attempting to find a rule that would describe the general agreement among user-elicited gestures.

Our "soft" agreement proposal is based on the premise that gestures can be described by a set of descriptors; and therefore, the user preference can also be studied in terms of these descriptors. There are several ways of decomposing and annotating a gesture: structural, descriptive, functional and categorical [27]. In a categorical transcription, a finite number of categories is defined for the gesture and then the gestures are systematically annotated according to the categories [27]. These types of transcriptions can be easily represented as binary vectors, making them relevant to our approach. The work in [28] divided the proposals into phases, which is usually done in the structural transcription, and then proceeded to annotate each phase according to a discrete set of possible hand configurations (categories).

III. Methodology

A. Notations

Let us start by defining the notations. Let C be the total number of commands or referents. Let P_r be the set of all proposals or gestures for the command r, where $r = 1, \ldots, C$. Let $u_r \leq |P_r|$ be the number of unique gestures for the command r. Let P_r^i be a subset of gestures for command r that are considered identical. Thus, $|P_r^i|$ would be the number of identical gestures in the set i for the command r. We use $|.|$ to denote the number of elements in a set and use $\|.\|$ to denote the L2 norm of a vector. The total number of gesture examples (N_r) for

the command r can be represented as the following (Eq. 1). Note that all referents have same number of gesture examples.

$$N = N_r = \sum_{i=1}^{u_r} | P_r^i | = | P_r | \tag{1}$$

The two most commonly used agreement metrics were defined by Wobbrock et al. [9], [24]. These formulations were given in equations 2 and 3 respectively.

$$\mathscr{A}_r = \sum_{i=1}^{u_r} \left(\frac{| P_r^i |}{| P_r |} \right)^2 \tag{2}$$

$$\mathscr{AR}_r = \frac{1}{N(N-1)} \sum_{i=1}^{u^r} | P_r^i | (| P_r^i | - 1) \tag{3}$$

Where $\mathscr{A}_r$ and $\mathscr{AR}_r$ are the level of agreement for command r. While the former metric results in a value of non-zero when there is no agreement, the latter metric takes a value of zero when all the subjects suggested a different proposal.

B. Gesture Representations: Hard vs Soft

A gesture descriptor is an entity that measures a particular property or a characteristic of a gesture. For example, in the case of hand gestures, these properties could include the direction of motion, shape of the trajectory, orientation of the hand, number of open fingers, etc. A finite set of descriptors can be used to distinctly represent a wide range of gestures. Each descriptor is assigned a binary value of one when a property is present and zero otherwise. In this regard, let us define the *gesture description* as a collection of binary values that can be used to represent a gesture as a binary vector in higher dimensions.

For instance, say we have the following descriptors: leftward, rightward, upward, downward, and circular motion of the hand. In this case, each gesture will be represented as a 5-dimensional binary vector, where each element indicates whether an attribute is present or not (refer to Figure 2a). The corresponding hard representation for these gestures is shown in Figure 2b

The main advantage of gesture descriptors is that they allow a “soft” representation of the gestures. The meaning of “soft” in this context is associated with how similarity is measured between two gestures. A “hard” representation only allows gestures to be considered as either identical (equivalence class) or completely different. Whereas, the soft representation of gestures allows measuring the similarity as a continuous value. While the approaches proposed by Wobbrock et al. [9], [24] fall under the category of hard representations, our approach utilizes soft representations to analyze the agreement between participants.

C. Gesture Description Analysis (GDA)

Given the *gesture descriptions,* (binary vectors) of elicited gesture proposals for command r, the objective of GDA is to determine the level of participants' agreement. To define an agreement index, we need a similarity metric that describes the distance between two vectors (binary in our case). Popular similarity metrics in the area of pattern recognition and data mining include cosine similarity, Jaccard similarity, and Hamming distance, which vary between zero and one (one when the vectors are equivalent and zero when they are orthogonal) [29]. Note that, the Jaccard similarity does not consider zero-zero as an agreement but only considers one-one as an agreement. It complies with the context of GDA because when two gestures lack a descriptor (value of the descriptor = 0), it does not imply agreement. For instance, if two gestures lack a circular motion descriptor, it indicates subjects agreed on not selecting this descriptor. However, it does not imply that the gestures are similar when those two subjects did not select a particular descriptor. Given the sparse nature of the descriptions, the cosine distance offers a good alternative to measure gesture similarity. The Hamming distance was discarded as it considers zero-zero (a descriptor being absent in both the gestures) as an agreement.

We propose an agreement metric referred to as Soft Agreement Rate ($\mathcal{SAR}$). Let a gesture proposal for referent r be represented as a binary vector S_r^i, where $i = 1, 2, \ldots, N_r$. The $\mathcal{SAR}$ is defined as a mean of the Jaccard similarity applied to all possible pairwise combinations of binary vectors corresponding to gestures in P_r (Equation 4). The overall $\mathcal{SAR}$ is defined as a mean of $\mathcal{SAR}$ of individual commands (Equation 5). The mathematical representation of $\mathcal{SAR}$ relates to ($\mathcal{AR}$) in terms of considering all possible pairwise combinations. Similar to the ($\mathcal{AR}$), the proposed metric $\mathcal{SAR}$ takes a value of 0 where there is no agreement and takes a value of 1 when all participants agree on a proposal.

$$\mathcal{SAR}_r = \frac{2}{N(N-1)} \sum_{j=k+1}^{N} \sum_{k=1}^{N} J(S_r^j, S_r^k) \tag{4}$$

$$\mathcal{SAR}_{overall} = \frac{2}{CN(N-1)} \sum_{r=1}^{C} \sum_{j=k+1}^{N} \sum_{k=1}^{N} J(S_r^j, S_r^k) \tag{5}$$

Where, S_r^j and S_r^k represent the binary vectors of j^{th} and k^{th} gesture proposal for the command r. Since the Jaccard similarity (J) between two zero vectors is not defined, we propose a conditional definition for this metric (Eq. 6 and 7). Table I shows the pairwise Jaccard similarity of the gestures depicted in Figure 2a. Similarly, a conditional definition for cosine similarity is proposed as this metric is not defined when one of the vectors is zero vector. Our methodology can be easily generalized to other similarity metrics such as cosine, Hamming, etc.

$$J(a,b) = \begin{cases} 0, & \text{if } \|a\| + \|b\| = 0 \\ \dfrac{a\,.\,b}{\|a\|^2 + \|b\|^2 - a\,.\,b}, & \text{otherwise} \end{cases} \tag{6}$$

$$cos(a,b) = \begin{cases} 0, & \text{if } \|a\|\,\|b\| = 0 \\ \dfrac{a\,.\,b}{\|a\|\,\|b\|}, & \text{otherwise} \end{cases} \tag{7}$$

Where a and b are the binary vectors, and $a.b$ denotes the dot product between the vectors.

For the purpose of completeness, we redefine other metrics such as disagreement rate ($\mathscr{DR}$) and coagreement rate ($\mathscr{CR}$) in the context of soft representations, referred as $\mathscr{SDR}$ and $\mathscr{SCR}$ respectively (Eq. 8 and 9). These metrics were initially introduced by Vatavu et al. [9] using hard representations.

$$\mathscr{SDR}_r = 1 - \mathscr{SAR}_r \tag{8}$$

$$\mathscr{SCR}(r_1, r_2, \ldots r_q) = \frac{2}{N\,(N-1)} \sum_{j=k+1}^{N} \sum_{k=1}^{N} \prod_{m=1}^{q} J\,(S_{r_m}^{j}, S_{r_m}^{k}) \tag{9}$$

D. Relation to Existing Metrics

In this section, we show that the metric $\mathscr{AR}$ proposed by Vatavu et al. and Wachs et al. [9], [12] is a special case of our approach. In other words, our methodology can be generalized to the case of "hard" representations. Let the $\mathscr{SAR}^{hard}$ denote the soft agreement rate when the gestures are treated as rigid entities and are grouped into equivalence classes instead of as a combination of descriptors. In machine learning, it is common to represent a distinct equivalence class (gestures in this case) as a one hot (OH) vector, which is a unit vector with only one value equal to unity and rest of the values equal to zero [30].

For the command r, there are u^r distinct equivalence classes or unique gestures. Each unique gesture is assigned to a distinct OH vector of length u^r. This implies that all the gestures in the set P_r^i are assigned to the same OH vector. The Jaccard similarity between two identical OH vectors is unity and between two distinct OH vectors is 0. This nullifies all of the distinct pairwise OH vectors. The resulting nonzero combinations are obtained from the pairwise combinations within the subset P_r^i.

$$J(S_r^j, S_r^k; j \neq k) = 0;\ J(S_r^j, S_r^k; j = k) = 1$$
$$\Rightarrow \sum_{j=k+1}^{N} \sum_{k=1}^{N} J(S_r^j, S_r^k) = \frac{1}{2} \sum_{i=1}^{u^r} |\,P_r^i\,|\,(\,|\,P_r^i\,| - 1)$$

Therefore, $\mathcal{SAR}_r^{hard}$ is given as follows:

$$\mathcal{SAR}_r^{hard} = \frac{1}{N_r(N_r-1)} \sum_{i=1}^{u^r} |P_r^i|(|P_r^i| - 1)$$
$$\mathcal{SAR}_r^{hard} = \mathcal{AR}_r \tag{10}$$

Interestingly enough, the resulting equation is exactly equal to the one proposed by Vatavu et al. [9]. This proves that our approach is general enough to adapt to both soft and hard representations of gestures. Hence, we conclude that the existing metrics that are widely used in the literature are a special case of our approach.

E. AR Metric Interpretation

It is clear that the agreement rate is proportional to the number of participants that agreed on a particular proposal. However, there is no direct relation between the level of agreement and the average percentage of participants that agreed on a proposal (η). For example, a value of $\eta = 0.40$ indicates that 40% of participants agreed on a proposal. In this section, we present an elegant way of interpreting the level of agreement. Given the agreement rate, the main idea is to estimate the average number of subjects that agreed on a particular gesture.

For instance, assume we have 10 proposals for a referent r in the following three scenarios: 1. 4 proposals are equivalent and rest are different ($\mathcal{AR} = 0.13$, $\eta = 0.40$), 2. 6 proposals are equivalent ($\mathcal{AR} = 0.33$, $\eta = 0.60$), 3. 8 proposals are equivalent ($\mathcal{AR} = 0.62$, $\eta = 0.80$). For each of these cases, it is unclear how the $\mathcal{AR}$ is related to η. In more complex scenarios such as when 3 out of 10 subjects picked a particular gesture, 2 out of 10 picked another gesture, and rest of subjects picked different gestures, it is very difficult to quantify η.

Thus, we propose an empirical relation between η and level of agreement, which is a numerical approximation of η in terms of the agreement value. It was found through computational experiments that the value of η is very close to the square root of the $\mathcal{AR}$ (Eq. 11) as shown in the Figure 3. In the case of hard representations, η indicates the average number of subjects that agreed on a gesture.

In addition, we propose the notion of η for $\mathcal{SAR}$ for the purpose of completeness. Equation 12 shows the extension of the numerical interpretation for $\mathcal{SAR}$. In the case of soft representations, η gives the average number of subjects that agreed on a set of gesture descriptors.

$$\eta_r^{\mathcal{AR}} = \sqrt{\mathcal{AR}_r} = \sqrt{\frac{1}{N(N-1)} \sum_{i=1}^{u^r} |P_r^i|(|P_r^i| - 1)} \tag{11}$$

$$\eta_r^{\mathcal{SAR}} = \sqrt{\mathcal{SAR}_r} = \sqrt{\frac{2}{N\,(N-1)} \sum_{j=k+1}^{N} \sum_{k=1}^{N} J\left(S_r^j, S_r^k\right)} \tag{12}$$

IV. EXPERIMENTAL METHODS

A. Case of Study : Gesture Elicitation with Neurosurgeons

In our previous work, we conducted a guessability study with a group of nine neurosurgeons to obtain their gestural preferences [15]. Each surgeon was asked to create gestures (proposals) for each of the 28 commands (referents) present in a radiology image browser. A total of 252 (28 x 9) gestures were considered. These elicited gestures will be used as a part of a gesture recognition system that controls a medical image browser in the operating room. In this work, we utilize these elicited gestures to evaluate and test our agreement assessment methodology.

Further, we utilized the gesture descriptors proposed in the literature [15], [31], [32] to create soft representations of the gestures. These descriptors were classified into three categories as shown in Table II: Motion – describes the movement of hands, Orientation – the mean direction in which the hand is pointing and State – number of fingers that are open [15]. By definition, the descriptors are binary and orthogonal i.e. they can be either present or absent in a gesture and there are no interdependencies between the descriptors. Hence, the descriptors are carefully chosen so that they are orthogonal to each other. In addition, descriptors related to the head and leg movements are not considered as the gestures elicited by surgeons did not contain such movements. Thus, agreement studies based on descriptors must have a selection step, where the relevant gestural properties are identified. Overall, a set of 55 descriptors that were particular to this case of the study were used in our analysis (refer to Table II).

B. GDA: Annotations and Analysis

We conducted two experiments in order to compute the level of agreement using two metrics: $\mathcal{AR}$ and $\mathcal{SAR}$. It was ensured that participants of these experiments were naïve to the gestures in the elicitation study i.e. they were not neurosurgeons. In the first experiment, six participants (four women and two men of age: 28 ± 4) were asked to group the gestures for each referent into equivalence classes, where each class consisted of gestures that were physically similar. This procedure was repeated for all 28 referents. These groupings were utilized to compute the level of agreement using $\mathcal{AR}$ formulation (hard representations, see Eq. 3). An average standard deviation of 0.09 was found among the agreement values obtained from these six participants. This value quantified the disagreement between the participants.

In the second experiment, the same group of six participants was asked to annotate descriptors for each of the 252 gestures. The final descriptor annotations were obtained by taking the majority vote among the participants. We developed software that facilitates annotating descriptors for each gesture. This software consists of 1. A window that plays the

gesture video and 2. A set of questions asking the participants to annotate whether a particular descriptor is present in the gesture shown. On an average, 5 out of 6 participants (83%) agreed with respect to their binary annotations. Furthermore, their annotations were used to determine a set of ambiguous descriptors i.e. the descriptors that were either confusing or difficult to identify. We found that the properties related to the *overall flow* (inward and outward) and *combined movement* (circular and rectangular) of hand were ambiguous and participants tend to disagree on their annotations.

The annotations were automatically parsed by the software to generate a vector consisting of 55 binary values (*zero* when the descriptor is absent and *one* when the descriptor is present). Furthermore, the annotation process disregarded every movement that was an outcome of a transition. For example, if a surgeon brought a hand up to be able to move it down, then the upward motion would be discarded in the annotation. The rationale for this is that those motions are not an intrinsic part of the gesture and they should be ignored in the annotation process [33]. We used these annotations to compute agreement using $\mathcal{SAR}$ metric.

C. Statistical Significance Tests

In this section, we propose a statistical testing methodology to compare the agreement rates of a pair of referents. This assessment helps determine the differences between the agreement rates of referents within the same group (i.e. <*zoom in, zoom out*>, <*flip horizontal, flip vertical*>, etc.). If such differences are statistically significant, the system designer needs to redesign the gestures. For example, if the agreement rates of *zoom in* and *zoom out* are significantly different, the designer can potentially choose the proposal of the referent with higher agreement rate and redesign a complementary or a mirror gesture for the referent with a lower agreement rate.

Our experimental design resembles the repeated measures design as the subject population remains constant across all referents. A *two-tailed paired t-test* was used to identify the pairs of referents that have significantly different agreement rates. To avoid multiple comparisons, controlling the False Discovery Rate (FDR) approach was used to find the final list of significant results [34].

Let r_i and r_j for $i \neq j$ be two distinct referents. In a paired t-test, observations are nothing but an entity measured in the two different test conditions. In this case, the entity that is being measured is the degree of similarity between the proposals (a value between 0 and 1) and the two referents act as the conditions. Given $N = 9$ subjects, there are $\frac{1}{2}N(N-1) = 36$ observations. The observation $v^i_{m,n}$, corresponds to the Jaccard similarity between the gesture descriptions of m^{th} and n^{th} proposal for referent r_i, where, $n \in \{1, 2, \ldots, N\}$ and $m \in \{n+1, 2, \ldots, N\}$. Given that a $DOF = 35$ and a level of confidence $\alpha = 0.95$, we determine the t-statistic (threshold probability) i.e. the probability that there is a significant difference between the two referents. This t-statistic is compared against the p-value obtained through a paired t-test conducted on the given data.

In this regard, we define the null hypothesis (H_0) and alternate hypothesis (H_a) as the following:

H_0: A pair of referents have equal agreement rates.

H_a: There is a difference between the agreement rates of the given pair of referents.

Given that there are $C = 28$ referents, this procedure is repeated for all possible pairs of referents ($C \times (C-1)/2 = 378$). Hence, we treat this problem as multiple hypotheses testing with 378 statistical tests. Given that the level of confidence for each test is 5%, there is a random chance that 5% of these tests would show significant differences for multiple independent tests. Therefore, we use the False Discovery Rate (FDR) method proposed by Benjamini and Hochberg [34] to screen out some false positives (the test that showed significance when it is not). The 378 p-values obtained through the aforementioned t-tests act as an input to control the FDR. Finally, the q-values (i.e. FDR adjusted p-values) that are less than 0.05 are considered significantly different. The algorithm 1 shows the pseudocode for this statistical test.

Algorithm 1 Significance Testing

```
S ← 9, C ← 28                          ▷ No. of subjects and referents
d_m^r ← Binary vector                  ▷ Description of gesture m of
    referent r . Where, m ∈ {1, ..., S}
Z ← C(C − 1) / 2                       ▷ No. of t-tests
prob_list ← []                         ▷ p-values obtained from t-tests
cmd_pairs ← []                         ▷ List of IDs of pairs of referents
for r in 1 : C do                      ▷ Loop over all commands
    v(r) ← []                          ▷ observations for referent r
    for i in 1 : S do                  ▷ Loop over all pairs of subjects
        for j in 1 : i do
            v(r) . add(J(d_i^r, d_j^r))   ▷ J : Jaccard similarity
        end for
    end for
end for
for i in 1 : C do                      ▷ Loop over all pairs of commands
    for j in 1 : i do
        p_value ← t_test(v(i), v(j))
        prob_list . add(p_value)
        cmd_pairs . add((i, j))
    end for
end for
q_values ← FDR(prob_list)              ▷ Apply FDR on the list
    of probability values to obtain modified p-values.
significance_ids ← arg(q_values < 0.05)   ▷ Find out
    the pairs of commands that are significantly different.
i*, j* ← cmd_pairs . get(significance_ids)   ▷
    (i*, j*) contains the pairs of command ids that passed the
    significance test.
```

V. RESULTS AND DISCUSSION

A. Agreement Analysis

Once the gestures corresponding to each referent are annotated with respect to their equivalence classes and gesture descriptors, agreement analysis was conducted using the $\mathscr{AR}$ and $\mathscr{SAR}$ formulations. Table III depicts the level of agreement for each referent using both of these metrics. Note that the referents with the same context are considered as a group. For example, *scroll up* and *scroll down* are grouped together, as they share the same context, namely *scroll*. Based on intuition, referents within the same group are expected to have similar agreement rates as their respective gestures are complementary to each other. For instance, *zoom in* and *zoom out* have similar agreement rates. However, subjects tend to choose completely different gestures for some referents in the same group as noticed in the case of *ruler measure* and *ruler delete*.

Table III is interpreted in the following manner. Consider the command *scroll up*, on an average, $\eta^{\mathscr{AR}} = 28.87\,\%$ of the participants agreed on a particular gesture corresponding to that command and $\eta^{\mathscr{SAR}} = 54.5\,\%$ of the participants agreed on a set of descriptors. Though η is an empirical and approximate measure of an average percentage, it is meant to provide a qualitative interpretation of the agreement values.

B. Results of Significance Tests

The statistical testing methodology described in Section IV-C was utilized to obtain the pairs of commands that have significantly different agreement rates among the participants. Overall, we conducted 378 hypotheses tests, each test corresponding to a pair of referents. Prior to applying FDR, 124 t-tests satisfied the significance criteria ($p < 0.05$). However, removing the probable false positives using the FDR technique resulted in 75 significant pairs. Figure 4 shows those pairs of commands that were significant (lighter color - indicates significance $q < 0.05$). Note that the matrix in Figure 4 is symmetric.

It was found that the majority of the 75 significant tests (64 tests) correspond to the pair of referents belonging to a different group. For instance, *scroll up* was significantly different from other commands but not *scroll down* as the gestures corresponding to *scroll up* and *scroll down* were complementary to each other. Similarly, *zoom in* and *zoom out* had similar agreement rates. However, there were few exceptions such as *pan up* whose agreement rate was significantly different from *pan left* and *pan right*. For such referents, system designers need to either re-conduct the gesture elicitation study to collect more proposals or design the gesture for *pan up* so that it is compatible with *pan left* and *pan right*.

C. Distribution of $\mathscr{SAR}$ Metric

In this section, we present the probability distribution function (PDF) $\mathscr{PDF}(S, Z)$ of the $\mathscr{SAR}$ metric by varying the number of subjects (S) and the number of descriptors (Z). This involved forming binary gesture description vectors of dimension Z for each of the gesture proposals elicited by the S subjects. These description vectors were sampled from a Bernoulli distribution with probability $P(1) = 0.5$. First, the random descriptions were generated and then, the level of agreement using $\mathscr{SAR}$ was computed. This procedure is

repeated for 10^7 iterations and the normalized histogram of agreement values was constructed using 100 bins of equal intervals in [0,1]. Figure 5a shows the PDF of $\mathcal{SAR}$ when the no. of descriptors remain constant and the no. of subjects vary ($\mathcal{PDF}(S \mid D = 55)$). The shape of the distribution resembles the bell curve with the peak occurring at 0.33 approximately. For $S = 9$ and $D = 55$, the cumulative probability $P(\mathcal{SAR} \leq 0.35) = 0.88$ while $P(\mathcal{SAR} \leq 0.40) = 0.999$. Similarly, Figure 5b shows the PDF of $\mathcal{SAR}$ when the no. of subjects remain constant and the no. of descriptors vary ($\mathcal{PDF}(D \mid S = 9)$). This distribution also resembles a bell curve and the peak occurs at 0.31 approximately.

Note that these PDFs were constructed assuming that the input data resembles a Bernoulli distribution with $P(1) = 0.5$. However, the actual gesture description data is sparse with zeroes occurring more frequently than ones $P(1) = 0.07$. Hence, we conducted a new set of computational experiments to construct the PDF of $\mathcal{SAR}$ when we feed the data that resembles the real data (Bernoulli distribution with $P(1) = 0.07$). Figures 6a and 6b show the PDF of our metric when the parameters S and Z are varied. Note that the shape of this distribution does not look like a bell curve anymore and the peak occurs between 0.0 and 0.1. For $S = 9$ and $D = 55$, the cumulative probability $P(\mathcal{SAR} \leq 0.04) = 0.84$ while $P(\mathcal{SAR} \leq 0.07) = 0.99$. These PDFs can be used to determine if the computed $\mathcal{SAR}$ values were occurring by chance.

The class priors (i.e. the probability at which zeros and ones occur) was measured from the actual gesture data. Thus, we hypothesized that the real data resembles a Bernoulli distribution with $P(1) = 0.07$. Furthermore, it was assumed that the gestures elicited by the participants are independent of each other. However, the elicited gestures are affected by the prior experience and expertise of the participants, which is popularly known as legacy bias [35], [36]. This would make the gesture proposals related to each other, which is reflected very well in the agreement rates shown in the Table III.

D. Qualitative Interpretation and Comparison

GDA considers the properties of gestures in the agreement analysis and hence produces higher agreement rates in comparison to hard representations. In other words, participants are more likely to agree on some high-level properties of the gestures even when they do not agree on the entire gesture. These results are intuitive and expected, considering that $\mathcal{SAR}$ takes into account the partial similarity between two gestures. Indirectly, the $\mathcal{SAR}$ metric focuses on what experts emphasize in the gestures rather than on their plain spatio-temporal appearance.

We further argue that $\mathcal{SAR}$ and $\mathcal{AR}$ complement each other. Consider the commands *zoom in* and *zoom out*. The values of $\mathcal{SAR}$ and $\mathcal{AR}$ are very close to each other. In this case, $\mathcal{SAR}$ determines the mostly agreed descriptors which in turn helps identify the physical gesture that contains these descriptors. It is not surprising that the mostly agreed gesture that is determined by $\mathcal{AR}$ is likely to contain those properties. While the GDA provides information about the properties of the most agreed gesture, the $\mathcal{AR}$ aids to determine the final gesture itself.

GDA is particularly advantageous when the values of $\mathscr{AR}$ are very low i.e. $\mathscr{AR} < 0.1$ [9]. Such a low value of $\mathscr{AR}$ indicates that there is very little agreement between the participants. In these cases, it is hard to determine the final gesture as the most agreed gesture is chosen by few participants. In such scenarios, the gesture interface designers will be greatly benefited from GDA as it allows them to determine the properties of the final gesture. This is due to the fact that designers can create a gesture that contains these highly agreed properties. Similarly, when the difference between $\mathscr{SAR}$ and $\mathscr{AR}$ is large as in the case of *scroll up*, the interface designers are recommended to utilize both the methodologies when determining the final gesture lexicon.

E. Lexicon Generation

Finally, we created a gesture lexicon using the top three most popular descriptors for each referent. The gestures were artificially constructed so they would be in compliance with the descriptors. This top ranking was obtained by summing all the gesture description vectors S_r^i corresponding to command r, and then finding the descriptors with the highest sum. Figure 7 shows the generated gestures for the 12 commands. One gesture from each group (see Table III for grouping) was randomly selected for this depiction. The descriptors are assumed to be for the right hand unless it is specified otherwise.

Each cell in Figure 7 is interpreted in the following manner. Consider the command *scroll up*; the popularity of the mostly agreed gesture [12] is 2/9 indicating that 2 out of 9 surgeons chose the gesture depicted on the *right* while the gesture obtained using GDA is depicted on the *left*. The popularity of 1/9 implies that all surgeons chose a different gesture. In such cases, the gesture obtained using agreement was not illustrated. Instead, we showed a gesture that complies with the top three descriptors. Overall, the gestures obtained by both the methods are the same for 6 commands, different for 15 commands, and the comparison is not possible for 7 commands as their popularity is 1/9.

The gestures shown in Figure 7 can be used throughout the entire lexicon or they can be used as an alternative to the elicited gestures whenever the agreement is too low. For instance, consider the *information window open* command, 22% of surgeons chose the gesture consisting of both the hands moving away from each other. However, 88%, 77% and 44% of the surgeons' gestures consisted of open palm, upward motion and an outward flow (closing of the hand) respectively. In this regard, we argue that the gesture obtained using GDA is a better choice, especially when the agreement is low, as surgeons are more likely to agree on the descriptors and not necessarily agree on a particular gesture. Moreover, the commands corresponding to the *rotate, ruler* and *layout* groups (see groups in Table III) can greatly benefit from the artificial proposals as their agreement is extremely low.

VI. Limitations

This section discusses the limitations of the proposed method and the potential solutions to tackle them. The first limitation lies in the nature of the descriptors i.e. they are binary and can take only two distinct values (0 or 1). In other words, the descriptors that are partially present in the gesture are considered as either present or absent in a gesture depending on

the annotation protocol. This issue can be addressed by treating the descriptors as continuous values between 0 and 1 in order to obtain a more granular score for agreement. While binary descriptors are easy to annotate, annotating the descriptors in a continuous manner is subjective (How to determine the difference between 0.41 and 0.42 regarding the upward motion?). Second, we illustrated our approach using hand gestures elicited from a guessability study conducted with a group of neurosurgeons. The results reported in this work are specific to the gestures developed by neurosurgeons. However, this framework can be extended to full body gestures for gaming consoles and touch gestures for smartphones by modifying the list of gesture descriptors. In other words, the current list of descriptors consists of properties related to hand motion/shape/configuration. However, full-body gestures require descriptors that explain the properties related to torso movements, leg motions, etc.

Each descriptor summarizes the entire gestural utterance by measuring a particular property of the gesture. Hence, the order in which the descriptors appear in the gesture was not considered in our analysis. For instance, consider two gestures that have the same set of descriptors, however, the order in which they appear in the gesture is different. Our approach would assign the same description vector to both the gestures. Hence, our method is limited to the scenarios where the order of occurrence of the descriptors does not alter the meaning of the gesture. The advantage of this is that it allows for a very compact representation which is time-invariant, with the downside of loosing temporal information.

This issue can be addressed by developing complex agreement formulations that can incorporate the sequence of descriptors. In this paper, the Jaccard distance is used as a metric to evaluate the similarity between two description vectors. However, if we were to use sequential descriptors, the Levenshtein distance could be used to measure the similarity between two sequences. This metric was popularly used for computing the similarity between two strings. Lastly, the η metric proposed in this paper is empirical and it is important to include a future investigation of its mathematical properties.

VII. Conclusions

Previous approaches to assess the level of agreement ignored the integral properties leading to a rigid representation of gestures. This work is primarily concerned with representing gestures as a combination of their high-level properties and thereby integrating the gesture descriptions into the agreement methodologies. Few previous works have proposed agreement formulae that incorporate the description of gestures into the agreement analysis. However, those formulae were based on intuition rather than on a thorough mathematical foundation. In this regard, we propose a generalized approach to measure agreement by incorporating Gesture Description Analysis (GDA) and provide mathematical justification for the agreement metric that we refer to as Soft Agreement Rate ($\mathcal{SAR}$). Next, we prove that the existing agreement metric $\mathcal{AR}$ is a special case of our approach. Furthermore, we developed an empirical relation between the level of agreement and the average number of participants that agreed on a particular gesture or a set of descriptors. This numerical approximation provides a qualitative interpretation of the agreement values.

Acknowledgment

This work is supported by the Agency for Healthcare Research and Quality (AHRQ), National Institute of Health (NIH) - under the Project No. 1R18HS024887-01. Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by NIH.

References

[1]. Istance H, Hyrskykari A, Immonen L, Mansikkamaa S, and Vickers S, “Designing Gaze Gestures for Gaming: An Investigation of Performance,” in Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications, ser. EtRA ’10. New York, NY, USA: ACM, 2010, pp. 323–330. [Online]. Available: http://doi.acm.org/10.1145/1743666.1743740

[2]. Wang Y, Yu T, Shi L, and Li Z, “Using human body gestures as inputs for gaming via depth analysis,” in 2008 IEEE International Conference on Multimedia and Expo, 6 2008, pp. 993–996.

[3]. Hettig J, Mewes A, Riabikin O, Skalej M, Preim B, and Hansen C, “Exploration of 3d medical image data for interventional radiology using myoelectric gesture control,” in Proceedings of the Eurographics Workshop on Visual Computing for Biology and Medicine Eurographics Association, 2015, pp. 177–185.

[4]. Jacob MG, Wachs JP, and Packer RA, “Hand-gesture-based sterile interface for the operating room using contextual cues for the navigation of radiological images,” Journal of the American Medical Informatics Association : JAMIA, vol. 20, no. e1, pp. e183–e186, 6 2013 [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3715344/ [PubMed: 23250787]

[5]. Wachs JP, Stern HI, Edan Y, Gillam M, Handler J, Feied C, and Smith M, “A Gesture-based Tool for Sterile Browsing of Radiology Images,” Journal of the American Medical Informatics Association : JAMIA, vol. 15, no. 3, pp. 321–323, 2008 [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2410001/ [PubMed: 18451034]

[6]. Muller MJ and Kuhn S, “Participatory design,” Commun. ACM, vol. 36, no. 6, pp. 24–28, 6 1993 [Online]. Available: http://doi.acm.org/10.1145/153571.255960

[7]. Spinuzzi C, “The methodology of participatory design,” Technical communication, vol. 52, no. 2, pp. 163–174, 2005.

[8]. Dong H, Danesh A, Figueroa N, and El Saddik A, “An elicitation study on gesture preferences and memorability toward a practical hand-gesture vocabulary for smart televisions,” IEEE Access, vol. 3, pp. 543–555, 2015.

[9]. Vatavu R-D and Wobbrock JO, “Formalizing Agreement Analysis for Elicitation Studies: New Measures, Significance Test, and Toolkit,” in Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, ser. CHI ’15. New York, NY, USA: ACM, 2015, pp. 1325–1334. [Online]. Available: http://doi.acm.org/10.1145/2702123.2702223

[10]. Wobbrock JO, Morris MR, and Wilson AD, “User-defined Gestures for Surface Computing,” in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ser. CHI ’09. New York, NY, USA: ACM, 2009, pp. 1083–1092. [Online]. Available: http://doi.acm.org/10.1145/1518701.1518866

[11]. Vatavu R-D and Wobbrock JO, “Between-subjects elicitation studies: Formalization and tool support,” in Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems ACM, 2016, pp. 3390–3402.

[12]. Stern HI, Wachs JP, and Edan Y, “Optimal Consensus Intuitive Hand Gesture Vocabulary Design,” in 2008 IEEE International Conference on Semantic Computing, 8 2008, pp. 96–103.

[13]. Arefin Shimon SS, Lutton C, Xu Z, Morrison-Smith S, Boucher C, and Ruiz J, “Exploring Non-touchscreen Gestures for Smartwatches,” in Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, ser. CHI ’16. New York, NY, USA: ACM, 2016, pp. 3822–3833. [Online]. Available: http://doi.acm.org/10.1145/2858036.2858385

[14]. Chan E, Seyed T, Stuerzlinger W, Yang X-D, and Maurer F, “User Elicitation on Single-hand Microgestures,” in Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, ser. CHI ’16. New York, NY, USA: ACM, 2016, pp. 3403–3414. [Online]. Available: http://doi.acm.org/10.1145/2858036.2858589

[15]. Madapana N, Gonzalez G, Rodgers R, Zhang L, and Wachs JP, “Gestures for picture archiving and communication systems (pacs) operation in the operating room: Is there any standard?” PLOS ONE, vol. 13, no. 6, pp. 1–13, 06 2018 [Online]. Available: 10.1371/journal.pone.0198092

[16]. Wobbrock JO, Aung HH, Rothrock B, and Myers BA, “Maximizing the guessability of symbolic input,” in CHI’05 extended abstracts on Human Factors in Computing Systems. ACM, 2005, pp. 1869–1872.

[17]. Garber L, “Gestural technology: Moving interfaces in a new direction [technology news],” Computer, vol. 46, no. 10, pp. 22–25, 2013.

[18]. Nielsen M, Störring M, Moeslund TB, and Granum E, “A procedure for developing intuitive and ergonomic gesture interfaces for hci,” in International gesture workshop. Springer, 2003, pp. 409–420.

[19]. Wachs JP, Kölsch M, Stern H, and Edan Y, “Vision-based hand-gesture applications,” Communications of the ACM, vol. 54, no. 2, pp. 60–71, 2011 [Online]. Available: http://dl.acm.org/citation.cfm?id=1897838 [PubMed: 21984822]

[20]. Jacob MG, Li YT, and Wachs JP, “Gestonurse: A multimodal robotic scrub nurse,” in 2012 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 3 2012, pp. 153–154.

[21]. Piumsomboon T, Clark A, Billinghurst M, and Cockburn A, “User-defined Gestures for Augmented Reality,” in CHI ’13 Extended Abstracts on Human Factors in Computing Systems, ser. CHI EA ’13. New York, NY, USA: ACM, 2013, pp. 955–960. [Online]. Available: http://doi.acm.org/10.1145/2468356.2468527

[22]. Ren Z, Yuan J, Meng J, and Zhang Z, “Robust Part-Based Hand Gesture Recognition Using Kinect Sensor,” IEEE Transactions on Multimedia, vol. 15, no. 5, pp. 1110–1120, 8 2013.

[23]. Connell S, Kuo P-Y, Liu L, and Piper AM, “A Wizard-of-Oz Elicitation Study Examining Child-defined Gestures with a Whole-body Interface,” in Proceedings of the 12th International Conference on Interaction Design and Children, ser. IDC ’13. New York, NY, USA: ACM, 2013, pp. 277–280. [Online]. Available: http://doi.acm.org/10.1145/2485760.2485823

[24]. Vatavu R-D, “User-defined Gestures for Free-hand TV Control,” in Proceedings of the 10th European Conference on Interactive TV and Video, ser. EuroITV ’12. New York, NY, USA: ACM, 2012, pp. 45–48. [Online]. Available: http://doi.acm.org/10.1145/2325616.2325626

[25]. Ruiz J, Li Y, and Lank E, “User-defined Motion Gestures for Mobile Interaction,” in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ser. CHI ’11. New York, NY, USA: ACM, 2011, pp. 197–206. [Online]. Available: http://doi.acm.org/10.1145/1978942.1978971

[26]. Morris MR, “Web on the wall: insights from a multimodal interaction elicitation study,” in Proceedings of the 2012 ACM international conference on Interactive tabletops and surfaces ACM, 2012, pp. 95–104.

[27]. Kipp M, Gesture generation by imitation: From human behavior to computer character animation. Universal-Publishers, 2005.

[28]. Kipp M, Neff M, and Albrecht I, “An annotation scheme for conversational gestures: how to economically capture timing and form,” Language Resources and Evaluation, vol. 41, no. 3-4, pp. 325–339, 2007.

[29]. Niwattanakul S, Singthongchai J, Naenudorn E, and Wanapu S, “Using of Jaccard coefficient for keywords similarity,” in Proceedings of the International MultiConference of Engineers and Computer Scientists, vol. 1, 2013.

[30]. Chren WA, “One-hot residue coding for low delay-power product CMOS design,” IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing, vol. 45, no. 3, pp. 303–313, 3 1998.

[31]. Giorgolo G, “A Formal Semantics for Iconic Spatial Gestures,” in Logic, Language and Meaning, ser. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, 2010, pp. 305–314, dOI: 10.1007/978-3-642-14287-1_31. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-642-14287-1_31

[32]. Lascarides A and Stone M, “A Formal Semantic Analysis of Gesture,” Journal of Semantics, vol. 26, no. 4, p. 393, 2009 [Online]. Available: 10.1093/jos/ffp004

[33]. Paquin V and Cohen P, "A Vision-Based Gestural Guidance Interface for Mobile Robotic Platforms," in Computer Vision in Human-Computer Interaction, ser. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, 5 2004, pp. 39–47. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-540-24837-8_5

[34]. Benjamini Y and Hochberg Y, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," Journal of the Royal Statistical Society. Series B (Methodological), vol. 57, no. 1, pp. 289–300, 1995 [Online]. Available: http://www.jstor.org/stable/2346101

[35]. Morris MR, Danielescu A, Drucker S, Fisher D, Lee B, schraefel m. c., and Wobbrock JO, "Reducing legacy bias in gesture elicitation studies," interactions, vol. 21, no. 3, pp. 40–45, 5 2014 [Online]. Available: http://doi.acm.org/10.1145/2591689

[36]. Köpsel A and Bubalo N, "Benefiting from legacy bias," interactions, vol. 22, no. 5, pp. 44–47, 8 2015 [Online]. Available: http://doi.acm.org/10.1145/2803169

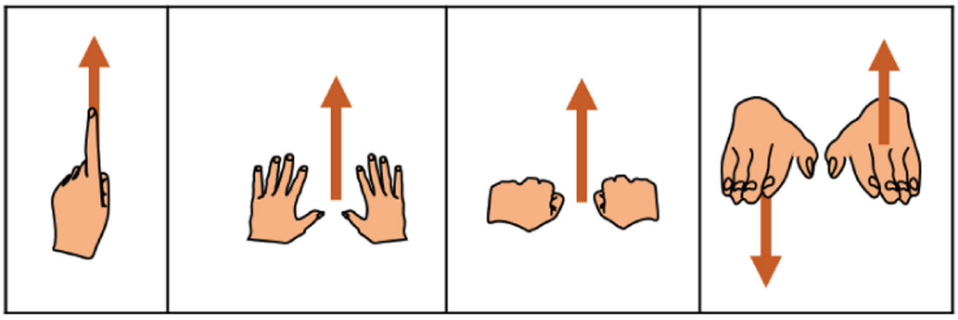
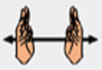


Fig. 1:
Illustration of similar gestures that are nonidentical.

Gesture ID	Gestures/ Property	1. Leftward motion	2. Rightward motion	3. Upward motion	4. Downward motion	5. Circular motion
G1		1	1	1	0	1
G2		1	1	0	1	1
G3		0	0	1	0	0
G4		1	1	0	0	0

(a) Soft representation.

Gesture ID	Gestures/ Property	G1	G2	G3	G4
G1		1	0	0	0
G2		0	1	0	0
G3		0	0	1	0
G4		0	0	0	1

(b) Hard representation.

Fig. 2:
Hard and soft representation of gestures.

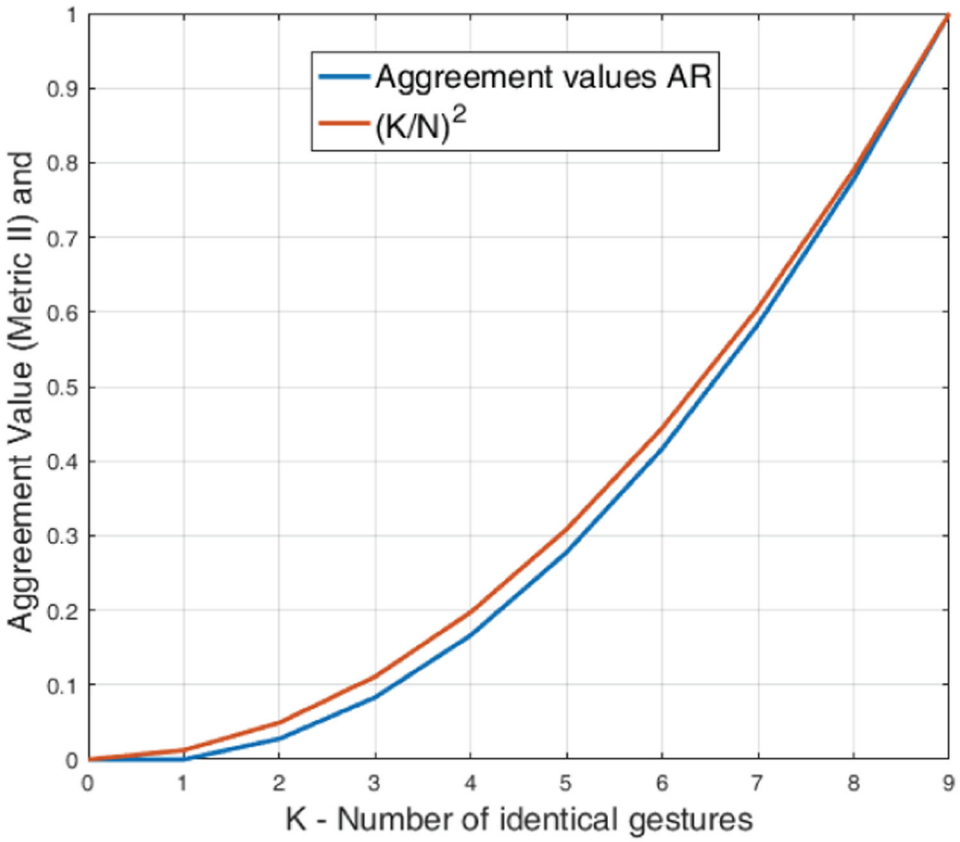


Fig. 3:

Illustration of the relation between the agreement values and η_r^{AR}. In this example, K indicates the number of proposals that are considered identical and $N - K$ proposals are treated completely different.

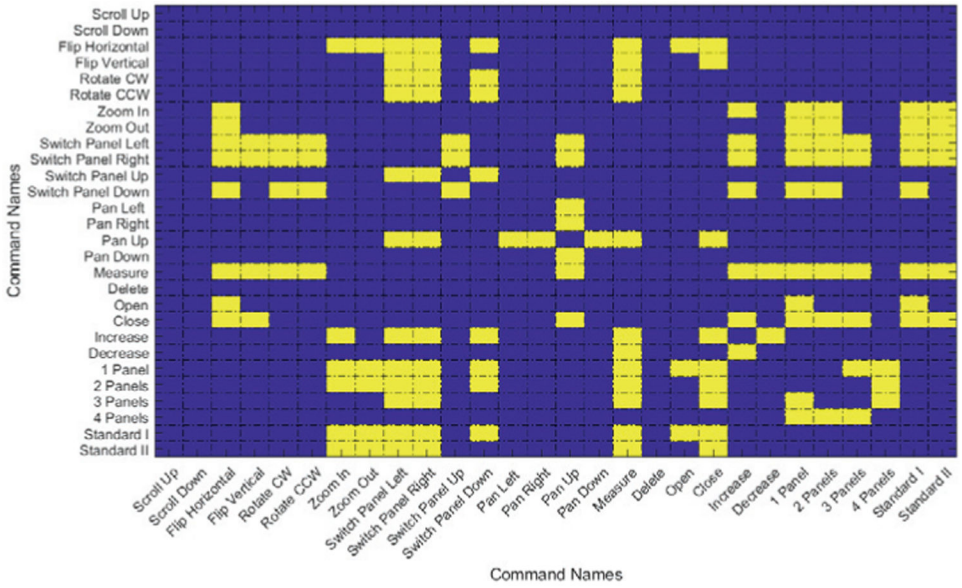
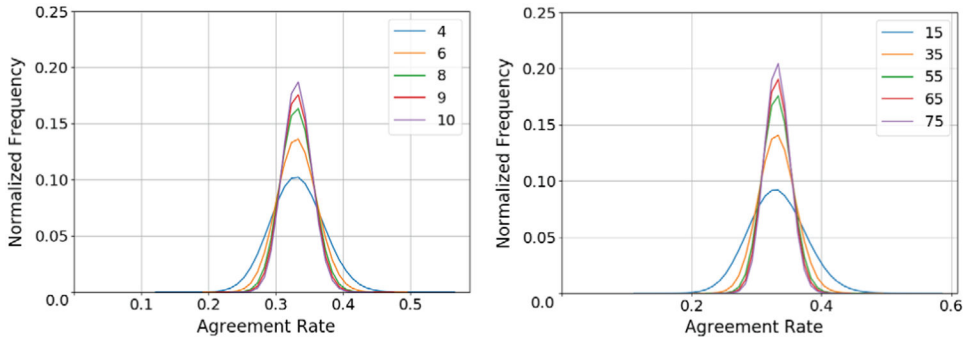


Fig. 4:

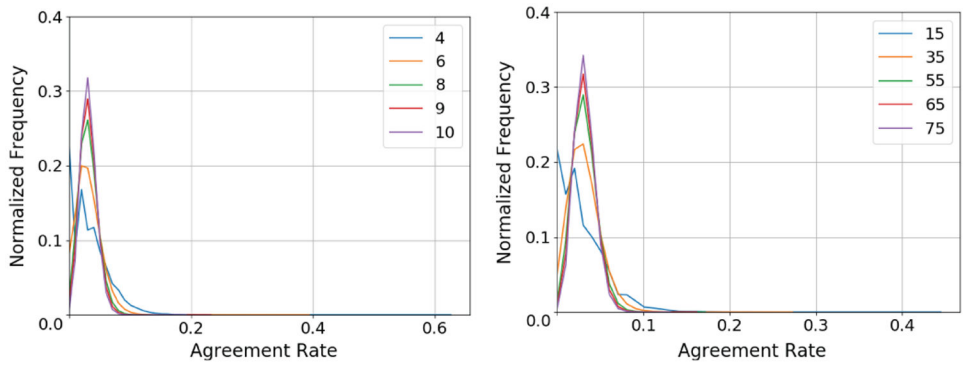
Commands that have significantly different agreement among participants. Lighter color indicates significance.



(a) $D = 55$ and Varying S. (b) $S = 9$ and Varying D.

Fig. 5:

Probability distribution of $\mathscr{SAR}$ with varying number of subjects and descriptors when the input gesture description data is sampled from a Bernoulli distribution with p(1) = 0.50.



(a) $D = 55$ and Varying S. (b) $S = 9$ and Varying D.

Fig. 6:

Probability distribution of $\mathcal{SAR}$ with varying number of subjects and descriptors when the input gesture description data is sampled from a Bernoulli distribution with p(1) = 0.93.

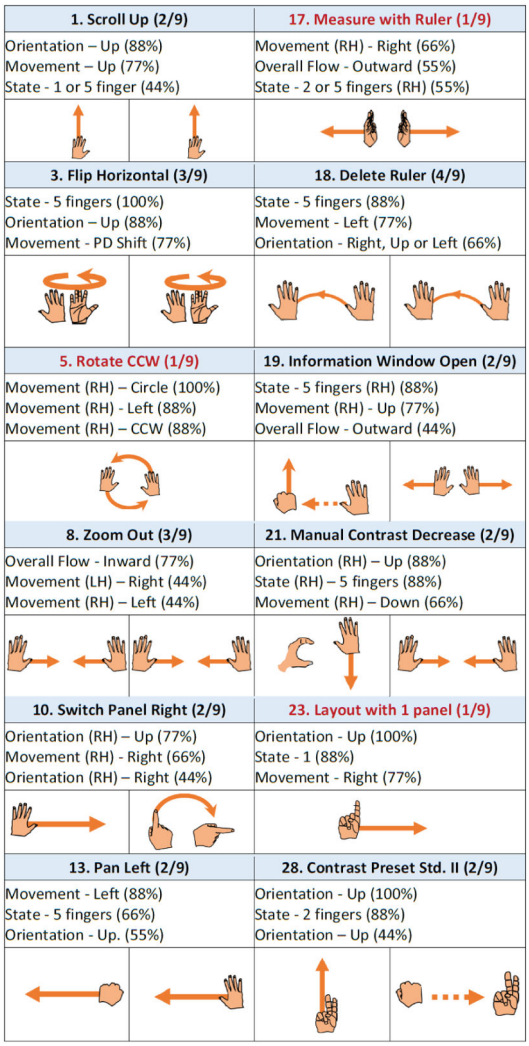


Fig. 7:

Illustration of the final gesture lexicon obtained from the **soft** (left) and the **hard representations** (right). Note that *LH* and *RH* represent left and right hands respectively, and the dotted line indicates that there is no motion while the solid line indicates the presence of motion.

TABLE I:

Illustration of pairwise Jaccard similarity.

	***G*1**	***G*2**	***G*3**	***G*4**
*G*1	1.0	0.5	0.3	0.7
*G*2	0.5	1.0	0.0	0.7
*G*3	0.3	0.0	1.0	0.0
*G*4	0.7	0.7	0.0	1.0

TABLE II:

Gesture Descriptors in our case of study.

Category	Sub-Category	Descriptors
Motion	Right hand motion	Right, up, left, down, forward, backward, clockwise, counterclockwise, iterative, circular, rectangular
	Left hand motion	
	Shifts	Palmar and dorsal Shifts
	Overall flow	Inward and Outward
	Combined movement	Circular and Rectangular
Orientation	Right hand orientation	Right, up, left, down, forward and backward
	left hand orientation	
State	Right hand state	Closed fist, one, two, three, four, five fingers, V shape and C shapes
	Left hand state	

TABLE III:

Values of agreement ($\mathcal{AR}$ and $\mathcal{SAR}$) and estimated average number of subjects that agreed on particular command or a set of descriptors (η^{AR} and η^{SAR})

Command	$\mathcal{AR}$	$\eta^{\mathcal{AR}}(\%)$	$\mathcal{SAR}$	$\eta^{\mathcal{SAR}}(\%)$
Scroll Up	0.08 ± 0.07	28.8	0.30 ± 0.05	54.8
Scroll Down	0.10 ± 0.09	31.6	0.28 ± 0.02	53.8
Flip Horizontal	0.13 ± 0.10	36.5	0.39 ± 0.02	63.1
Flip Vertical	0.10 ± 0.14	32.4	0.33 ± 0.02	57.5
Rotate CW	0.20 ± 0.16	44.7	0.33 ± 0.02	58.2
Rotate CCW	0.18 ± 0.14	42.8	0.32 ± 0.03	57.0
Zoom In	0.19 ± 0.09	44.0	0.20 ± 0.01	44.7
Zoom Out	0.22 ± 0.06	47.7	0.19 ± 0.01	44.5
Panel Left	0.15 ± 0.16	38.7	0.23 ± 0.02	48.6
Panel Right	0.14 ± 0.15	38.0	0.23 ± 0.03	48.5
Panel Up	0.16 ± 0.19	40.8	0.31 ± 0.03	55.8
Panel Down	0.15 ± 0.19	39.4	0.28 ± 0.05	53.7
Pan Left	0.13 ± 0.06	36.5	0.34 ± 0.05	58.4
Pan Right	0.12 ± 0.06	34.9	0.34 ± 0.04	59.0
Pan Up	0.11 ± 0.07	33.3	0.37 ± 0.03	60.8
Pan Down	0.11 ± 0.06	34.1	0.34 ± 0.04	59.0
Ruler Measure	0.12 ± 0.06	34.9	0.23 ± 0.03	48.2
Ruler Delete	0.19 ± 0.07	44.0	0.32 ± 0.02	56.7
Window Open	0.06 ± 0.03	25.8	0.24 ± 0.03	49.2
Window Close	0.03 ± 0.02	19.7	0.21 ± 0.02	46.8
Inc Contrast	0.06 ± 0.07	24.7	0.32 ± 0.03	57.1
Dec Contrast	0.05 ± 0.04	23.5	0.30 ± 0.02	54.7
Layout 1	0.09 ± 0.13	30.7	0.33 ± 0.02	57.6
Layout 2	0.07 ± 0.06	26.8	0.31 ± 0.03	55.9
Layout 3	0.08 ± 0.09	29.8	0.30 ± 0.02	55.1
Layout 4	0.06 ± 0.06	24.7	0.29 ± 0.01	53.9
Preset 1	0.06 ± 0.08	25.8	0.30 ± 0.03	54.9
Preset 2	0.07 ± 0.07	27.8	0.30 ± 0.03	54.7
Mean ± Std	0.12 ± 0.05	33.7 ± 7.41	0.30 ± 0.05	54.4 ± 5.0